Econ 607 Jonathan Wright
Fall 2023

Basic Time Series

Definitions

A time series is **(weakly) stationary** if its first two moments are finite and are independent of time.

A time series is **strictly stationary** if the distribution of $(x_k,...x_{T+k})$ is independent of k.

A time series of the form $y_t = \Sigma_{j=1}^{p} a_j y_{t-j} + \Sigma_{j=0}^{q} c_j \varepsilon_{t-j}$ is an **ARMA(p,q)**.

AR and MA processes are obvious special cases.

The **autocovariance** at lag j is $\gamma_j = E(x_t x_{t-j})$.

A moving average process is **invertible** if it can be written as an absolutely summabble autoregression.

For an invertible process, we can back out the errors from the data if we know the parameters. For a non-invertible process we can't.

Autocovariances

For an AR(1) $\gamma_j = \sigma^2 \frac{\phi^j}{1-\phi^2}$.

For an MA(1) $\gamma_0 = (1+\theta^2)\sigma^2$, $\gamma_1 = \theta\sigma^2$, $\gamma_j = 0 \forall j > 1$.

Estimating the Autocovariances

It is easy to estimate the autocovariances by their sample counterparts $\hat{\gamma}_j = T^{-1}\Sigma(x_t - \overline{x})(x_{t-j} - \overline{x})$.

Moreover, if $\gamma = (\gamma_0, \gamma_1, ...\gamma_h)'$ and $E(\varepsilon_t^4) = \eta\sigma^4$ then

$$T^{1/2}(\hat{\gamma} - \gamma) \rightarrow_d N(0,V)$$

where $Cov(\hat{\gamma}_p, \hat{\gamma}_q) = T^{-1}\{(\eta-3)\gamma(p)\gamma(q) + \Sigma_{k=-\infty}^{\infty}[\gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p)]\}$.

Estimation of AR and MA processes

Estimation can take place by maximizing the likelihood function. But that must be done numerically. For an AR(p) conditioning on the first p observations makes it much easier.

$$\log(f(y_{p+1}, y_{p+2}...y_T \mid y_1,...y_p, \theta)) = -\frac{(T-p)}{2}\log(2\pi) - \frac{(T-p)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} ... - \phi_p y_{t-p})^2$$

which can clearly be maximized in closed form. This is the way that at an AR is usually estimated, although this is not the exact likelihood. The exact likelihood includes the density of the first p observations and that can make a substantive difference. For the AR(1) the exact log likelihood can be written as:

$$-\frac{T}{2}\log(2\pi) - \frac{(T-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma_{t=2}^{T}(y_t - c - \phi_1 y_{t-1})^2 - \frac{1}{2}\log V - \frac{1}{2V}(y_1 - \frac{c}{1-\phi_1})^2$$

where $V = \frac{\sigma^2}{1-\phi^2}$ is the variance of y_1.

We can also estimate an MA(q) more simply by conditioning on $\varepsilon_0 = \varepsilon_{-1} ... = \varepsilon_{1-q} = 0$. Given this assumption, the errors can all be recovered uniquely given a parameter value and the conditional likelihood is then

$$-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{t=1}^{T}\varepsilon_t^2$$

The spectral density

Consider a stationary time series. There are a number of ways to represent its persistence properties. It can be represented by

1. The AR representation, $A(L)y_t = \varepsilon_t$ where $A(L) = I - A_1 L - A_2 L^2 ... - A_p L^p$
2. The MA representation, $y_t = C(L)\varepsilon_t$ where $C(L) = I + C_1 L + C_2 L^2 ... + C_p L^p$
3. The autocovariance function, $\gamma(j) = E(y_t y_{t-j})$, or
4. The spectral density. The spectral density is

$$f(\omega) = \frac{1}{2\pi}\Sigma_{j=-\infty}^{\infty}\gamma(j)e^{-i\omega j} = \frac{Var(\varepsilon_t)}{2\pi}C(e^{i\omega})^2 = \frac{Var(\varepsilon_t)}{2\pi}\frac{1}{A(e^{i\omega})^2}$$

for $\omega = [-\pi, \pi]$. Note that $f(\omega) = f(-\omega)$.

Once you have any one of these, you can always work out the others.

MA representation to ACF

$\gamma(j) = E[(\Sigma_{i=0}^{\infty}c_i\varepsilon_{t-i})(\Sigma_{k=0}^{\infty}c_k\varepsilon_{t-j-k})] = \sigma^2\Sigma_{i=0}^{\infty}c_i c_{i+j}$

Spectrum to ACF

$\gamma_k = \int_{-\pi}^{\pi} f(\omega)e^{i\omega k}d\omega$

Stationarity and Invertibility

For an AR process, the condition for stationarity is that $A(L) = 0$ has only solutions such that $|L| > 1$. For an MA process, the condition for invertibility is that $C(\mathrm{L}) = 0$ has only solutions such that $|L| > 1$.

Examples of spectral density

AR(1): $f(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{(1-\alpha e^{i\omega})(1-\alpha e^{-i\omega})} = \frac{\sigma^2}{2\pi}\frac{1}{(1+\alpha^2 - 2\alpha\cos\omega)}$

MA(1): $f(\omega) = \frac{\sigma^2}{2\pi}(1+ce^{i\omega})(1+ce^{-i\omega}) = \frac{\sigma^2}{2\pi}(1+c^2+2c\cos(\omega))$

The Periodogram

The periodogram of a time series $\{x_t\}_{t=1}^{T}$ is defined as

$$I(\lambda_j) = T^{-1} \| \Sigma_{t=1}^{T} x_t e^{-it\lambda_j} \|^2$$

for $\lambda_j = 2\pi j / T, j = 0,1,...T/2$. The periodogram is an asymptotically unbiased but not consistent estimate of the spectrum. In fact for $j \neq 0, T/2$

$$\frac{I(\lambda_j)}{f(\lambda_j)} \rightarrow_d \frac{\chi^2(2)}{2}$$

while for $j = 0, T/2$

$$\frac{I(\lambda_j)}{f(\lambda_j)} \rightarrow_d \chi^2(1)$$

Estimating the Spectral Density

There are two ways of proceeding.

1. Fit an AR to the time series in question. The spectral density is $\frac{Var(\varepsilon_t)}{2\pi} \frac{1}{A(e^{i\omega})^2}$. Just replace the parameters with estimates.
2. Smooth the periodogram.

$$\hat{f}(\lambda_j) = \Sigma W(k) I(\lambda_{j+k})$$

where the weights sum to 1 and $W(k) = W(-k)$. As long as the spectrum is smooth, this is consistent.

Beveridge-Nelson Decomposition

Suppose that $x_0 = 0$ and $\Delta x_t = C(L)\varepsilon_t = \Sigma_{j=0}^{\infty} c_j \varepsilon_{t-j}$. Then

$x_t = C(1) z_t + C^*(L)\varepsilon_t$

where $z_t = z_{t-1} + \varepsilon_t = \Sigma_{s=1}^{t} \varepsilon_s$ and $c_j^* = -\Sigma_{i=j+1}^{\infty} c_i$ (Permanent-Transitory Decomposition).

Proof:

$x_t = \Sigma_{s=1}^{t} \Delta x_s = \Sigma_{s=1}^{t} C(L)\varepsilon_s$

$= c_0\varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2}...$

$\quad + c_0\varepsilon_{t-1} + c_1\varepsilon_{t-2}...$

$= c_0\varepsilon_t + \Sigma_{j=0}^{1} c_j \varepsilon_{t-j}...$

$= (C(1) - \Sigma_{j=1}^{\infty} c_j)\varepsilon_t + (C(1) - \Sigma_{j=2}^{\infty} c_j)\varepsilon_{t-1}...$

$= C(1)\Sigma_{s=1}^{t} \varepsilon_s + C^*(L)\varepsilon_t$

The Central Limit Theorem for Dependent Data

If u_t is iid with $2+\delta$ moments for some $\delta > 0$, then the central limit theorem says that

$$T^{-1/2} \Sigma_{t=1}^{T} u_t \rightarrow_d N(0, Var(u_t))$$

Now suppose instead that u_t is stationary with spectral density $f(.)$ and covariance function $\gamma(.)$, but that it is not necessarily iid. The process must satisfy some other conditions. The central limit theorem for dependent data now says that

$$T^{-1/2}\Sigma_{t=1}^{T}u_t \rightarrow_d N(0, 2\pi f(0))$$

and of course we can write $2\pi f(0) = \Sigma_{j=-\infty}^{\infty}\gamma(j) = Var(u_t)C(1)^2 = Var(u_t)/A(1)^2$.

The regression model with Dependent Errors

Now think of a linear regression model of the form

$$y_t = \beta' x_t + \varepsilon_t$$

Suppose that ε_t is stationary but may be dependent. Then from the law of large numbers:

$$T^{-1}\Sigma_{t=1}^{T}x_t x_t' \rightarrow_p E(x_t x_t') \equiv M$$

and from the central limit theorem for dependent data:

$$T^{-1/2}\Sigma_{t=1}^{T}x_t\varepsilon_t \rightarrow_d N(0, J)$$

where J is 2π times the spectral density of $x_t\varepsilon_t$ at frequency zero. Or, if $\Gamma(j) = E(z_t z_{t-j}')$ is the autocovariance function of $z_t = x_t\varepsilon_t$ then $J = \Sigma_{j=-\infty}^{\infty}\Gamma(j)$.

Combining these gives

$$T^{1/2}(\hat{\beta} - \beta) = (T^{-1}\Sigma x_t x_t')^{-1}T^{-1/2}\Sigma x_t\varepsilon_t \rightarrow_d N(0, M^{-1}JM^{-1}) \quad (1)$$

and this is a general form for the asymptotic distribution of an OLS estimator with possibly dependent data. Note that M is a symmetric matrix.

Now if ε_t were iid (homoskedastic and serially uncorrelated), $E(x_t\varepsilon_t x_{t-j}'\varepsilon_{t-j}) = 0$ for $j \neq 0$ and $E(x_t x_t'\varepsilon_t^2) = E(x_t x_t' E(\varepsilon_t^2 \mid x_t)) = E(\varepsilon_t^2)E(x_t x_t')$. So in this case, $J = \Gamma(0) = Var(\varepsilon_t)M$ and

$$T^{1/2}(\hat{\beta} - \beta) \rightarrow_d N(0, Var(\varepsilon_t)M^{-1})$$

But alas we are rarely so lucky in time series and so we have to resort to the formula in (1). To estimate M is easy, as we just use $\hat{M} = T^{-1}\Sigma_{t=1}^{T}x_t x_t'$. Estimating J is harder but it is clearly a special case of the general problem of estimating a spectral density at frequency zero.

Estimating the spectral density at zero frequency

The problem of estimating the spectral density at frequency zero comes up quite a bit, particularly in getting standard errors for time series. Of course it is a special case of estimating the spectral density as above. Here are three ways of doing it.

1. Fit an AR to the time series in question. The spectral density (ignoring the 2π) is $A(1)^{-1}\Sigma A(1)^{-1}'$. Just replace the parameters with estimates.

2. Hansen and Hodrick (1980) proposed $\Sigma_{-l_T}^{l_T}\hat{\Gamma}(j)$ where $\hat{\Gamma}(j) = T^{-1}\Sigma z_t z_{t-j}'$ and l_T is a truncation parameter. It is intuitive that if $l_T \rightarrow \infty$ as $T \rightarrow \infty$ at the right rate that $\Sigma_{j-l_T}^{l_T}\hat{\Gamma}(j) \rightarrow_p J$.

3. Unfortunately, the Hansen and Hodrick estimator is not necessarily positive definite. This could mean that you would have negative variance estimates, and so imaginary standard errors. The proposal of Newey and West (1987) avoids this problem. It uses $\Sigma_{j=-l_T}^{l_T}(1-\frac{|j|}{l_T+1})\hat{\Gamma}(j)$. It so happens that this estimator is not only consistent for J but is guaranteed to be positive definite. As practical advice, the truncation parameter l_T is set to the largest distance apart that two observations might be expected to be substantially correlated with each other. More formal answers to the question of choosing this parameter (also known as the "bandwidth") are provided in Andrews (1991) and Andrews and Monohan (1992). Lazarus, Lewis, Stock and Watson (2018) recommend $l_T = 1.3T^{1/2}$.

Kiefer-Vogelsang Asymptotics

It turns out that t- and F-tests using Newey-West standard errors can have poor size, especially if the truncation parameter is big. Kiefer and Vogelsang (2005) assume that $l_T = bT$ for some b between 0 and 1. Consider the Wald test of the hypothesis that $\beta = \beta_0$. Instead of the usual χ^2 distribution, this has a distribution

$$B(1)'[\frac{2}{b}\int_0^1 B_b(r)B_b(r)'dr - \frac{1}{b}\int_0^{1-b}(B_b(r+b)B_b(r)'+B_b(r+b)B_b(r)')dr]^{-1}B(1)$$

where $B(r)$ is a standard Brownian motion and $B_b(r) = B(r) - rB(1)$. The advice of Lazarus, Lewis, Stock and Watson (2018) includes using Kiefer-Vogelsang critical values.

Clustered Standard Errors

Dependent errors come up in cross-sectional and panel data too. The general result is
$n^{1/2}(\hat{\beta}-\beta) \rightarrow_d N(0, M^{-1}JM^{-1})$
where $n^{-1}\Sigma_{i=1}^n x_i x_i' \rightarrow_p M$ and $J = n^{-1}Var(\Sigma_{i=1}^n z_i) = n^{-1}\Sigma_{i=1}^n\Sigma_{j=1}^n E(z_i z_j)$ and $z_i = x_i\varepsilon_i$. However, we cannot estimate J consistently by $n^{-1}\Sigma_{i=1}^n\Sigma_{j=1}^n z_i z_j$.

"Clustering" uses $n^{-1}\Sigma_{i=1}^n\Sigma_{j=1}^n w_{ij} z_i z_j$ where w_{ij} is 1 for pairs of observations within a cluster, and 0 otherwise. Clustered standard errors are commonly used in panel data applications. But clustered standard errors and spectral density estimators are all just ways of weighting the elements being summed in J so as to avoid estimating too many free parameters.

Information Criteria

Consider the problem of fitting the model

$$y_t = \Sigma_{j=0}^p \beta_j x_{j,t} + \varepsilon_t$$

and we want to pick how many RHS variables to include (this includes the lag order for an AR). This is what information criteria do. The residual sample variance is $\hat{\sigma}_p^2 = \frac{1}{T}\Sigma e_t(p)^2$. Obviously adding more regressors always reduces this, so information criteria are of the form

$$\ln \hat{\sigma}_p^2 + pg(T)$$

for different penalty functions, $g(T)$. These are

AIC $g(T) = \frac{2}{T}$

BIC $g(T) = \frac{\ln T}{T}$

Theorem (Consistency) If $g(T) \to 0$ but $Tg(T) \to \infty$ and there is a true model with p_0 regressors then $\hat{p} \to p_0$. (BIC satisfies this; AIC doesn't).

Outline of Proof. Suppose $p < p_0$. Then

$$s_p^2 - s_{p_0}^2 = \ln \hat{\sigma}_p^2 + pg(T) - \ln \hat{\sigma}_{p_0}^2 - p_0 g(T) = \ln(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p_0}^2}) + (p - p_0)g(T)$$

$\therefore \lim_{T\to\infty} s_p^2 - s_{p_0}^2 > 0$

$\therefore \lim_{T\to\infty} P(s_p^2 - s_{p_0}^2 < 0) = 0$

Suppose $p > p_0$

$$T(s_p^2 - s_{p_0}^2) = T\ln(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p_0}^2}) + T(p - p_0)g(T) = O_p(1) + T(p - p_0)g(T)$$

$\therefore \lim_{T\to\infty} T(s_p^2 - s_{p_0}^2) = \infty$

$\therefore \lim_{T\to\infty} P(s_p^2 - s_{p_0}^2 < 0) = 0$

BIC has a Bayesian justification and should be thought of as selecting the true model. AIC is justified for picking the model that minimizes mean square error. These two objectives are fundamentally at odds with each other.

Suppose that the true model is $y_t = \varepsilon_t$, but we estimate $y_t = \beta' x_t + \varepsilon_t$ (over-specified)

The estimated error variance is $\sigma^2(1 - \frac{k}{T})$

If forecast out-of-sample, forecast variance is $\sigma^2(1 + \frac{k}{T})$

The intuition of AIC is that it corrects for this difference. There are other closely related information criteria, such as final prediction error

$$FPE = \ln(\hat{\sigma}_p^2) + \ln(\frac{T + p}{T - p})$$

Econ 607 Jonathan Wright
Fall 2023

Handout on Panel Data

Suppose that we have a model of the form

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it} \quad (1)$$

which is a very common occurrence with macro as well as micro datasets. In micro applications, n has typically been large and T is small, but there are many macro applications where that is not the case. If the constant were the same for all i, then this would be a pooled regression, but the power of panel data is that we can allow for unobserved heterogeneity as long as it does not change over time.

The existence of fixed effects does not avoid the need to worry about the dependence properties in the errors. It is possible to assume that the errors are iid in both the cross-section and the time-series, but that is usually too strong an assumption. Clustering will allow for correlation within blocks (e.g. regions of the country, firms in an industry).

Driscoll and Kraay (1998) have an approach to standard errors that are robust to both cross-sectional dependence and some autocorrelation. It is a panel version of Newey West. We can take the model and rewrite it as

$$y_{it} - \overline{y}_i = \beta(x_{it} - \overline{x}_i) + \varepsilon_{it} - \overline{\varepsilon}_i \quad (2)$$

or $\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\varepsilon}_{it}$ for short. Now consider running a pooled regression for all n and T and get the OLS estimate $\hat{\beta} = (X'X)^{-1}X'Y$. Define $h_{it} = \tilde{x}_{it}(\tilde{y}_{it} - \hat{\beta}\tilde{x}_{it})$. Then take the cross sectional averages, $h_t = \frac{1}{N}\Sigma_{i=1}^{N} h_{it}$. Define

$$\hat{\Omega}_j = \Sigma_{t=j+1}^{T} h_t h'_{t-j}$$

Then the Driscoll-Kraay variance-covariance matrix is

$$(X'X)^{-1}[\hat{\Omega}_0 + \Sigma_{j=1}^{l} \frac{1}{l+1-j}(\hat{\Omega}_j + \hat{\Omega}'_j)](X'X)^{-1}$$

The Driscoll-Kraay approach makes mild assumptions and works well. Petersen (2009) gives a discussion of methods for standard errors with panel data that is now a classic reference especially in finance applications.

A situation that often arises is that we have a panel where we think that the slope coefficients may differ over the cross-section. If it is as simple as:

$$y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it} \quad (3)$$

then there is no real panel data estimation element. It is SUR which is OLS equation by equation unless the errors are correlated cross-sectionally. But there may be some more structure to the slope coefficients. Suppose that there is some variable z_i and we are willing to assume that:

$$\beta_i = \gamma_0 + \gamma_1 z_i$$

Then, by substitution, we have:

$$y_{it} = \alpha_i + \gamma_0 z_i + \gamma_1 x_{it} z_i + \varepsilon_{it}$$

which is now back to being a standard fixed effects model.

Recently, methods have been proposed for estimating group structure in panel data models. You know that the cross-sectional observations belong to G groups. Cross-section i belongs to group $g(i) \in \{1,...G\}$. The groups can differ in fixed effects and/or in slope coefficients. Consider the least squares estimator

$$\arg\min_{\alpha,\beta,g} \Sigma_{i=1}^{N}\Sigma_{t=1}^{T}(y_{it} - \alpha_{g(i)} - \beta_{g(i)}x_{it})^2$$

This has two advantages. One is that there is an efficiency gain from reducing the number of fixed effects that are estimated. The other is that there may be interesting stories about which cross-sections have higher and lower slope coefficients. This leaves unadressed the question of what the number G is. Lu and Su (2017) have a testing procedure where they determine the number G.

Econ 607 Jonathan Wright
Fall 2023

Handout on the Bootstrap

The bootstrap

The bootstrap is a simulation method for forming confidence intervals and obtaining standard errors using only information from the sample. Like a Monte-Carlo simulation, but uses only the data.

Advantages
1. Applicable in a wide range of contexts
2. Easy.
3. The bootstrap in some cases may produce better approximations (knocks out an extra term in the Edgeworth expansion).

Basic idea of bootstrap

Suppose that $X_1, X_2, ...X_n$ is iid with a parameter θ and an estimator $\hat{\theta}$.

We can resample from $X_1, X_2, ...X_n$ with replacement (otherwise just reordering) and get a new estimate $\hat{\theta}_i^{boot}$.

Can repeat n_B times.

Idea is that θ is to $\hat{\theta}$ as $\hat{\theta}$ is to $\hat{\theta}_i^{boot}$.

Hall justifies the bootstrap with a "Russian Dolls" analogy
- Doll zero: population that we do not get to see
- Doll one: sample we observe
- Doll two: bootstrap sample

A simple use of the bootstrap is to estimate the bias in $\hat{\theta}$.

$$E(\hat{\theta} - \theta) \approx \frac{1}{n_B} \Sigma_{i=1}^{n} (\hat{\theta}_i^{boot} - \hat{\theta})$$

Also, say I want a 95 percent confidence interval for θ.
Sort the bootstrap estimates $\hat{\theta}_i^{boot}$.
Let F_α^{boot} denote the percentiles of the distribution of $\hat{\theta}_i^{boot}$.
Three ways of forming the bootstrap CI.

<u>Method 1</u>

A confidence interval for θ is [$F^{boot}_{0.025}, F^{boot}_{0.975}$].
Called "other percentile"

Method 2

Pretend that we know the distribution of $\hat{\theta}-\theta$ and assume that this is pivotal (does not depend on θ). Call it G.

If we knew G, a 1-2α confidence interval for θ would be

$\{\theta : G_\alpha \le \hat{\theta}-\theta \le G_{1-\alpha}\}$

$= \{\theta : -G_{1-\alpha} \le \theta-\hat{\theta} \le -G_\alpha\}$

$= \{\theta : \hat{\theta}-G_{1-\alpha} \le \theta \le \hat{\theta}-G_\alpha\}$

But we don't know G

We do however know the distribution of $\hat{\theta}^{boot}_i$ and can assume that the distribution of $\hat{\theta}^{boot}_i-\hat{\theta}$ is the same as the distribution of $\hat{\theta}-\theta$ (assuming pivotalness again).

$G_\alpha = F^{boot}_\alpha - \hat{\theta}$ and $G_{1-\alpha} = F^{boot}_{1-\alpha} - \hat{\theta}$

The confidence interval is then

$\{\theta : \hat{\theta}-(F^{boot}_{1-\alpha}-\hat{\theta}) \le \theta \le \hat{\theta}-(F^{boot}_\alpha-\hat{\theta})\}$

$= \{\theta : 2\hat{\theta}-F^{boot}_{1-\alpha} \le \theta \le 2\hat{\theta}-F^{boot}_\alpha\}$.

Call this the "percentile" interval.

Method 3

But alas: in most cases $\hat{\theta}-\theta$ is not pivotal

Suppose that $\dfrac{\hat{\theta}-\theta}{s_{\hat{\theta}}}$ is pivotal (at least asymptotically).

Let $\tilde{F}$ be its distribution.

If $\tilde{F}$ is known, then

$\{\theta : \hat{\theta}-s_{\hat{\theta}}\tilde{F}_{1-\alpha} \le \theta \le \hat{\theta}-s_{\hat{\theta}}\tilde{F}_\alpha\}$

is a $1-2\alpha$ confidence interval for θ.

Now let $\tilde{F}^{boot}$ denote the bootstrap distribution of $\dfrac{\hat{\theta}^{boot}_i-\hat{\theta}}{s_{\hat{\theta}}}$. This should be the same (asymptotically) as $\tilde{F}$

CI: $\{\theta : \hat{\theta}-s_{\hat{\theta}}\tilde{F}^{boot}_{1-\alpha} \le \theta \le \hat{\theta}-s_{\hat{\theta}}\tilde{F}^{boot}_\alpha\}$

This is the percentile-t interval.

Simple example of the bootstrap

Suppose that these are the speeds of a sample of seven cars on a stretch of highway:
79 73 68 77 86 71 69

Form a 95 percent confidence interval for the population mean (a) asymptotically, and (b) via the bootstrap.

$\bar{X} = 74.71,\ s = 6.40,\ n = 7$

The 95 percent asymptotic confidence interval is

$$74.71 \pm \frac{2.45 * 6.4}{\sqrt{7}}$$

which is from 68.78 to 80.64.

Bootstrap: Simple Matlab code

```
x=[79  73  68  77  86  71  69]';
xbar=mean(x); se=std(x)/sqrt(7);
nboot=1000;
for imc=1:nboot;
        xboot=x(ceil(7*rand(7,1)));
        xbarboot(imc)=mean(xboot);
        xse=std(xboot)/sqrt(7);
        tstatboot(imc)=(mean(xboot)-xbar)/xse;
end;
xbarboot=sort(xbarboot); tstatboot=sort(tstatboot);
```

[xbarboot(25) xbarboot(975)]

[xbar-(tstatboot(975)*se) xbar-(tstatboot(25)*se)]

The last two lines give the OP and Percentile t CIs for mean

	Confidence Intervals	
	Lower Bound	Upper Bound
Asymptotic	68.78	80.64
Other Percentile	70.57	79.71
Percentile-t	69.71	78.86

The bootstrap in a regression model

Suppose I have a linear regression model

$y_i = \beta' x_i + \varepsilon_i$

The most standard implementation of the bootstrap entails the following steps:

(i) Estimate the parameter vector β and work out the residuals $e_i = y_i - \hat{\beta}' x_i$

(ii) Resample from the residuals with replacement and from the regressors with replacement.

(iii) Build up a new dataset of the dependent variables as $y_i^{BOOT} = \hat{\beta}' x_i^{BOOT} + e_i^{BOOT}$

(iv) Work out the quantity of interest in this new dataset

(v) Repeat (ii)-(iv) many times.

Other percentile, percentile or percentile-t confidence intervals can then be worked out. In part (ii), one could hold the regressors fixed. This amounts to making all statements conditional on the observed regressors.
The bootstrap described here as the undesirable feature that it imposes that the residuals and regressors are independent; not just that $E(e_i \mid x_i) = 0$. There is likely however to be conditional heteroskedasticity that this bootstrap is going to destroy. A way around this is to use the wild bootstrap, or heteroskedasticity-robust bootstrap. To create the bootstrap samples, for each observation, take the following distribution:

$$P(e_i^{BOOT} = (\frac{1+\sqrt{5}}{2})e_i) = \frac{\sqrt{5}-1}{2\sqrt{5}}$$

$$P(e_i^{BOOT} = (\frac{1-\sqrt{5}}{2})e_i) = \frac{\sqrt{5}+1}{2\sqrt{5}}$$

This ensures that $E(e_i^{BOOT} \mid x_i) = 0$, $E((e_i^{BOOT})^2 \mid x_i) = e_i^2$ and $E((e_i^{BOOT})^3 \mid x_i) = e_i^3$. Intuitively, the idea is to randomize the sign of the residuals in repeated bootstrap samples. That would be sufficient to allow for conditional heteroskedasticity, but not for skewness.

The bootstrap in time series models

Now, the idea of the bootstrap is that all observations are independent of each other. That doesn't work with time series. There are three ways of proceeding, that we'll illustrate in the context of the AR(1) time series model $y_t = \alpha y_{t-1} + u_t$.

Method 1: Parametric Bootstrap

1. Estimate α by $\hat{\alpha}$.
2. Form the residuals $\hat{u}_t = y_t - \hat{\alpha} y_{t-1}$
3. Resample from the residuals with replacement, and form new data $\{y_t\}$.
4. Form a confidence interval for α, or a function of α, on the bootstrap data as before.

The problem is that this doesn't work if $\alpha = 1$ because the distribution is not pivotal.

Method 2: Bias-Adjusted parametric Bootstrap (Kilian (1998))

1. Estimate α by $\hat{\alpha}$.
2. Form the residuals $\hat{u}_t = y_t - \hat{\alpha} y_{t-1}$
3. Resample from the residuals with replacement, and form new data $\{y_t\}$.
4. Use the bootstrap samples to estimate the bias in $\hat{\alpha}$, $B(\hat{\alpha})$.
5. Now form a new set of residuals as $u_t^* = y_t - \alpha^* y_{t-1}$ where $\alpha^* = \hat{\alpha} - B(\hat{\alpha})$.
6. Resample from the residuals with replacement, and form new data.
7. Use the bootstrap sample to estimate α and then adjust each of these by subtracting off the (negative) bias $B(\hat{\alpha})$. Take any required function of these estimates of α.
8. Save the percentiles of the bootstrap sample in (7). This is a bias-adjusted confidence interval.

Method 3: Block Bootstrap

Instead of resampling individual observations, resample blocks of observations. The blocks should be sufficiently short that there are many blocks; yet sufficiently long that there will be little dependence across blocks. In an extension, the blocks can have random length (stationary bootstrap). This idea, due to Politis and Romano (1994) can be summarized with the following algorithm:

1. Randomly choose one observation from the sample.
2. With probability *p*, pick the next observation, otherwise go back to 1 (and go back to 1 in any case if you are at the final observation in the series).
3. Repeat until you have a bootstrap sample with the same number of observations as the original sample

This algorithm produces blocks of data of random length, but the expected length is $\frac{1}{1-p}$. Clearly, the more persistent the series, the longer you want the blocks and so the higher *p* should be.

Method 4: Grid Bootstrap

Hansen (1999) proposes instead the grid bootstrap.

Consider testing the hypothesis $\alpha = \alpha_0$

1. Form the implied errors $u_t = y_t - \alpha_0 y_{t-1}$
2. Resample from these errors with replacement.
3. Find the upper and lower percentiles of the bootstrap distribution of the t-statistic testing the hypothesis that $\alpha = \alpha_0$.
4. Comparing the actual t-statistic with these critical values, decide to accept or reject.

A confidence interval for α can be formed by inverting the acceptance region of this test.

See Berkowitz and Kilian (2000) for more discussion of the bootstrap in time series. We'll return to the bootstrap in the context of the estimation of VARs.

Econ 607 Jonathan Wright
Fall 2023

Handout on Markov Chain Monte Carlo (MCMC) Methods

The potential for Bayesian analysis has been greatly increased by algorithms for simulating from posterior distributions. Two (related) widely used algorithms are Gibbs sampling and Metropolis Hastings.

Gibbs sampling concerns the case when you can easily draw from conditional distributions, but not from joint distributions. Say you know the distribution of Y|X and X|Y. Gibbs sampling simply says that you can start at any X, take a draw of Y given that X, and then iterate back and forth between the two conditionals. If you discard some initial values, then the draws of X and Y are draws from their joint distribution.

Metropolis-Hastings simulates samples from the full joint density function. Suppose that $f(Y)$ is the density that I want to draw from. I cannot directly draw from that density. Here is the algorithm.

1. Take an arbitrary value of Y, Y_t.

2. Take a proposal distribution, $q(z|Y_t)$ that I can draw from.

3. With probability $\min(1, \frac{f(z)q(Y_t|z)}{f(Y_t)q(z|Y_t)})$, I accept the proposal, meaning that I set $Y_{t+1} = z$. Otherwise, I reject the proposal meaning that I set $Y_{t+1} = Y_t$.

4. Repeat many times.

The first many draws are discarded, and the remaining draws are draws from the density $f(Y)$. It is important to note that they are not independent of each other, but the chain nonetheless converges to the required density.

A common proposal distribution is the random walk distribution, giving random-walk Metropolis Hastings. In this, the proposed draw is $z = Y_t + \varepsilon$, where ε is a normal random variable. This has the feature that the proposal distribution is symmetric ($q(a|b) = q(b|a)$) and so the acceptance probability in step 3 reduces to $\min(1, \frac{f(z)}{f(Y_t)})$. This still leaves the question of the variance of ε. A small variance will give a high acceptance probability, but the chain will be slow to move to new parts of the support. A big variance will give a lower acceptance probability. The rule of thumb is to target an acceptance rate of around 35%, and pick the variance of ε to achieve this.

There is a further feature of Metropolis-Hastings in the context of drawing from posterior densities. The basic starting point of all Bayesian analysis is Bayes rule, which says that with data X and a parameter θ,

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)d\theta}$$

The denominator can be tricky to compute. But if I want to simulate from the posterior density using random-walk Metropolis Hastings, the acceptance probability is:

$$\min(1, \frac{p(z \mid X)}{p(\theta_t \mid X)}) = \min(1, \frac{p(X \mid z)p(z)}{p(X \mid \theta_t)p(\theta_t)})$$

That involves the prior and the likelihood, but the messy integral cancels out. (This is incidentally the key reason why we prefer this to the inverse cdf method which evaluates the inverse cdf at uniform random variables on the unit interval).

Let's take a trivial example. I want to draw from an exponential density with parameter 1, but I don't know how. But I do know the exponential probability density (exppdf in Matlab). Here is code for a random walk Metropolis Hastings.

```
randn('seed',123); rand('seed',123);
y(1)=0;
for t=2:1100;
   y0=y(t-1);
   yp=y0+(2*randn(1,1));
   f0=exppdf(y0,1);
   fp=exppdf(yp,1);
   if rand(1,1)<min(fp/f0,1); accept(t-1,1)=1; y(t,1)=yp; else; accept(t-1,1)=0; y(t,1)=y0; end;
end;
disp(mean(accept))
hist(y(101:end))
```

If I run this, the accept rate is 0.34, which is good.

A different version of Metropolis-Hastings is independence Metropolis-Hastings, where each proposal is an independent draw from the same proposal distribution, so $q(z \mid Y_t) = q(z)$. In this case, of course the proposal distribution is not symmetric and so the simplification to the acceptance probability does not go through.

Handout on the Equity Risk Premium Puzzle

Consider a representative agent economy in which investors trade assets so as to maximize

$$E_t(\Sigma_{s=0}^{\infty}\beta^s u(C_{t+s}))$$

in the usual notation. The investor's first order condition in a consumption CAPM can be written as

$$P_t = E_t(\beta\frac{u'(c_{t+1})}{u'(c_t)}(P_{t+1}+D_{t+1}))$$

which can also be rewritten as

$$E_t(\beta\frac{u'(C_{t+1})}{u'(C_t)}R_{t+1})=1$$

or

$$E_t(M_{t+1}R_{t+1})=1$$

where $M_{t+1}=\beta\frac{u'(C_{t+1})}{u'(C_t)}$ is the stochastic discount factor or pricing kernel. This means that the gross riskfree rate is $R_{f,t}=1/E_t(m_{t+1})$. And it also means that we can write

$$E_t(M_{t+1}R_{t+1}^e)=0$$

where $R_{t+1}^e=R_{t+1}-R_{f,t}$ is the excess return. The two equations highlighted in red are different ways of writing the basic building blocks of modern finance. The expected product of the stochastic discount factor and *gross returns* is one; the expected product of the stochastic discount factor and *excess returns* is zero.

A natural choice for the instantaneous utility function is $u(c_t)=\frac{C_t^{1-\gamma}}{1-\gamma}$ and γ is the coefficient of relative risk aversion, so this is called the constant relative risk aversion (CRRA) utility function.
So the Euler equation is now of the form

$$E_t(\beta\frac{C_{t+1}^{-\gamma}}{C_t^{-\gamma}}R_{t+1})=1$$

Assume consumption-growth and returns are log-normal. Then

$$E_t(\beta\frac{C_{t+1}^{-\gamma}}{C_t^{-\gamma}}R_{t+1}) = E_t(\exp(\log(\beta\frac{C_{t+1}^{-\gamma}}{C_t^{-\gamma}}R_{t+1})))$$

$$= E_t(\exp(\log(\beta)+r_{f,t}-\gamma\log(\frac{C_{t+1}}{C_t}))) = \exp(\log(\beta)+r_{f,t}-\gamma g+\frac{\gamma^2\sigma_c^2}{2})$$

where $r_{f,t}=\log(R_{f,t})$, $g=E_t\log(\frac{C_{t+1}}{C_t})$ and $\sigma_c^2=Var_t\log(\frac{C_{t+1}}{C_t})$. So

$$\log(\beta)+r_{f,t}-\gamma g+\frac{\gamma^2\sigma_c^2}{2}=0$$

. Also

$$\log(\beta) - \gamma g + E_t r_{t+1} + \frac{\gamma^2}{2}\sigma_c^2 + \frac{1}{2}\sigma_r^2 - \gamma\sigma_{c,r} = 0$$

where $r_{t+1} = \log(R_{t+1})$, $\sigma_r^2 = Var_t r_{t+1}$ and $\sigma_{c,r} = Cov_t(r_{t+1}, \log(C_{t+1}/C_t))$. Hence

$$E_t r_{t+1} - r_{f,t} = \gamma\sigma_{c,r} - \frac{1}{2}\sigma_r^2$$

Now let us suppose that the correlation between returns and consumption growth is 1—the highest it could possibly be. Then the equity risk premium is equal to $\gamma\sigma_c\sigma_r - \frac{1}{2}\sigma_r^2$. This is an upper bound.

Now, from the data we know that $\sigma_r = 0.16$ and $\sigma_c = 0.01$. Here is the implied upper bound for different coefficients of risk aversion.

Coefficient of Relative Risk Aversion	Upper Bound on Equity Risk Premium
3	-0.008
5	-0.0048
10	0.0032
30	0.0352
50	0.0672

The actual observed equity risk premium is about 7 percent. This is the equity premium puzzle of Mehra and Prescott (1985). Intuitively, what makes stocks risky is that they lose value in bad times when marginal utility is high. But consumption is smooth, and so unless agents are very risk-averse, marginal utility cannot be flapping around that much. The basic puzzle is robust across countries and across time (Campbell (1999)).

Another way of making the same point is to note that

$$E_t(M_{t+1}R_{t+1}) = Cov(M_{t+1}, R_{t+1}) + E(M_{t+1})E(R_{t+1})$$

$$1 = Cov(M_{t+1}, R_{t+1}) + \frac{1}{R_{f,t}}E(R_{t+1})$$

$$E(R_{t+1}) - R_{f,t} = -\frac{1}{R_{f,t}}Cov(M_{t+1}, R_{t+1})$$

So you need lots of negative covariance between the pricing kernel and returns to explain a big risk premium.

Here are a few of the explanations that have been proposed in the literature:

1. Long-run risks (Bansal and Yaron (2004)).

In the consumption CAPM with intertemporally separable preferences, the coefficient γ does "double duty". It is both the coefficient of risk aversion and it is the reciprocal of the intertemporal elasticity of

substitution $\frac{d\ln(C_{t+1}/C_t)}{d\log(R_{f,t})}$. It seems intuitive that the highly risk-averse agent will also want to smooth consumption over time. Epstein and Zin (1991) proposed an alternative utility function that breaks this link. The utility function is

$$U_t = \{(1-\beta)(C_t)^{\frac{1-\gamma}{\theta}} + \beta E_t[U_{t+1}^{1-\gamma}]^{\frac{1}{\theta}}]\}^{\frac{\theta}{1-\gamma}}$$

where γ is the coefficient of relative risk aversion, $\theta = \frac{1-\gamma}{1-1/\psi}$ and ψ is the intertemporal elasticity of substitution. This reduces to the standard utility function if $\gamma = 1/\psi$ as in this case, we can write

$$U_t = \{(1-\beta)(C_t)^{1-\gamma} + \beta E_t[U_{t+1}^{1-\gamma}]]\}^{\frac{1}{1-\gamma}}$$
$$\therefore U_t^{1-\gamma} = (1-\beta)(C_t)^{1-\gamma} + \beta E_t[U_{t+1}^{1-\gamma}]]$$

Substituting $V_t = \frac{U_t^{1-\gamma}}{1-\gamma}$ gives $V_t = \frac{1-\beta}{1-\gamma}C_t^{1-\gamma} + \beta E_t V_{t+1}$ and so the two utility functions are equivalent. But this restriction is not imposed.

The Euler equation with these preferences turns out to be

$$E_t[\beta^\theta (\frac{C_{t+1}}{C_t})^{-\theta/\psi} (\frac{W_{t+1}}{W_t - C_t})^{\theta-1} R_{t+1}] = 1$$

where W_t is the total wealth of the representative agent.

Let $g_{t+1} = \log(\frac{C_{t+1}}{C_t})$ and $g_{d,t+1} = \log(\frac{D_{t+1}}{D_t})$ denote the growth rates of consumption and dividends, respectively. The innovation of Bansal and Yaron is to allow both of these to have a small but very persistent component. So

$$g_{t+1} = \mu + x_t + \varepsilon_{t+1}$$
$$g_{d,t+1} = \mu_d + \phi x_t + \varepsilon_{d,t+1}$$
$$x_{t+1} = \rho x_t + v_{t+1}$$

The model is solved numerically. They calibrate $\gamma = 10$ and $\psi = 1.5$ (which would not be possible without Epstein-Zin preferences), and can get close to the observed equity risk premium in the data. Intuitively, the idea is that what makes equities so risky is that they lose value in precisely the state-of-the world in which consumption growth is going to be low for decades, and so marginal utility is high.

2. Habit formation.

Abel (1990) and Campbell and Cochrane (1999) use habit formation as a way of accounting for the equity risk premium. Suppose that the representative agent has a utility function

$$E_t(\Sigma_{s=0}^{\infty}\beta^s \frac{(C_{t+s} - X_{t+s})^{1-\gamma}}{1-\gamma})$$

where X_t is a "habit" level of consumption. So what matters is the difference between your actual consumption, and what you have got used to. This difference cannot go negative; if it did, then there would be negative infinite utility. Define the "surplus consumption ratio"

$$S_t = \frac{C_t - X_t}{C_t}$$

and assume that that

$$\log(s_{t+1}) = (1-\phi)\overline{s} + \phi s_t + \lambda(s_t)(c_{t+1} - c_t - g)$$

where $s_t = \log(S_t)$, $c_t = \log(C_t)$ and λ is a nonlinear sensitivity function. Again, this model can match the equity risk premium with reasonable values of risk aversion.

The model written down here treats habit as "external". That is, each individual's habit is affected by everyone else's consumption. The effect of my own consumption on future habits is therefore negligible. If, instead, I care about my own habit level ("internal habits"), then things get more complicated since I have to think about how my consumption today affects habits in the future. But that has been looked at too. In fact, there is a taxonomy of four possible habit models: internal v. external and where habit enters as a difference v. habit entering as a ratio.

3. The consumption of the rich. Most individuals do not participate in the stock market, at least not directly. Those that do are very rich, and may have more volatile consumption. Aït-Sahalia, Parker and Yogo (2004) back out the coefficient of risk aversion implied by a consumption-CAPM using the volatility of consumption of luxury goods (Tiffany sales, BMW sales etc.). The implied coefficients of risk aversion are typically around 6, which seems much more reasonable.

4. Multiple goods (related to 3). Following Piazzesi et al. (2007), suppose that there are preferences over two goods: all non-housing consumption and consumption of housing services. The representative agent has the utility function

$$E_t(\Sigma_{s=0}^{\infty}\beta^s \frac{C_{t+s}^{1-\gamma}}{1-\gamma})$$

but where consumption is $C_t = [c_t^{(\varepsilon-1)/\varepsilon} + \omega H_t^{(\varepsilon-1)/\varepsilon}]^{\varepsilon/(\varepsilon-1)}$, with c_t and H_t denoting non-housing and housing consumption, respectively. Utility is not separable over non-housing and housing consumption.[1] The pricing kernel is

$$M_{t+1} = \beta(\frac{c_{t+1}}{c_t})^{-\gamma}(\frac{1+\omega\alpha_{t+1}^{(\varepsilon-1)/\varepsilon}}{1+\omega\alpha_t^{(\varepsilon-1)/\varepsilon}})^{\frac{1-\varepsilon\gamma}{(\varepsilon-1)}}$$

where $\alpha_t = H_t / c_t$. As long as $\varepsilon \neq 1/\gamma$, this puts an extra term in the stochastic discount factor that may help generate a higher equity risk premium. The idea is that the level of housing services directly affects the marginal utility of non-housing consumption. For a reasonable parameterization, with $\gamma = 5$, they get an equity risk premium of 3.5 percent (not as high as in the data-but going in the right direction).

5. Rare disasters. An old idea that recent events just made popular once again (Rietz (1988), Barro (2006)). Suppose that in each period, consumption growth follows the process

$$\Delta \log C_t = \mu + \sigma\varepsilon_t \text{ with probability } 1-p$$

$$\Delta \log C_t = \mu + \sigma\varepsilon_t + \log(1-b) \text{ otherwise}$$

[1] Notice that if $\gamma = 1/\varepsilon$ then utility is separable for then $C_t = [c_t^{1-\gamma} + \omega H_t^{1-\gamma}]^{\frac{1}{1-\gamma}} \Rightarrow C_t^{1-\gamma} = c_t^{1-\gamma} + \omega H_t^{1-\gamma}$.

So each period, there is a probability p of a disaster which permanently lowers consumption by a factor b. If stocks represent claim to the consumption stream and a representative agent has CRRA preferences in the consumption CAPM, then the equity risk premium is

$$\sigma^2\gamma + pb[(1-b)^{-\gamma} - 1]$$

For example, let us suppose that $b = 0.3$, so that a disaster is a permanent 30 percent reduction in consumption. Let the volatility of consumption growth be $\sigma = 0.01$ and let risk aversion be $\gamma = 5$. Here is the equity risk premium for different choices of p.

p	Equity Risk Premium
0.00	0.05%
0.01	1.53%
0.02	3.02%
0.05	7.47%

So it is quite easy to get a big risk premium with only a modest probability of disaster. These probabilities are so small that you may not actually observe a disaster at all in the sample (and, moreover, the countries and periods for which you can get data are the ones which did not have disasters). On the other hand, the disaster scenario is a bit extreme, especially when you note that the economy never recovers from the disaster at least in this baseline model.

The Hansen-Jagganathan Bound

These different methods are all really tricks to try to get more volatility in the stochastic discount factor. Hansen and Jagganathan (1991) have a nice framework for explaining the need for this. Go back to the basic asset pricing equation

$$E_t(M_{t+1}R_{t+1}) = 1$$

From the Cauchy-Schwarz inequality, we have

$$|Cov(M_{t+1}, R_{t+1})| \leq \sigma_m\sigma_r$$

where σ_m^2 and σ_r^2 are the variances of m_{t+1} and R_{t+1}, respectively. Also,

$$Cov(M_{t+1}, R_{t+1}) = E(M_{t+1}R_{t+1}) - E(M_{t+1})E(R_{t+1}) = 1 - E(M_{t+1})E(R_{t+1})$$

$$\therefore \sigma_m^2\sigma_r^2 \geq (1 - E(M_{t+1})\,\mathrm{E}(\mathrm{R}_{t+1}))^2$$

$$\therefore \sigma_m^2 \geq \frac{(1 - E(M_{t+1})\,\mathrm{E}(\mathrm{R}_{t+1}))^2}{\sigma_r^2}$$

The multivariate extension of this (with a vector of asset returns) is:

$$\therefore \sigma_m^2 \geq (i - E(M_{t+1})\mu)'\Sigma^{-1}(i - E(M_{t+1}))$$

Where i is a vector of ones and μ and Σ are the mean and variance-covariance matrix of returns.

So this is a lower bound on the volatility of the stochastic discount factor. This is the Hansen-Jagganathan bound. Given a candidate stochastic discount factor and the expected return and volatility of returns on any asset, it is easy to check this. For example, using the simple consumption CAPM with CRRA preferences, the right-hand side will be far too big (unless the coefficient of risk aversion is about 50). All of the resolutions to the equity risk premium puzzle are really ways of increasing σ_m so that this bound is satisfied.

Handout on GMM

Part 1: The basics

GMM nests all the familiar estimators as special cases. The idea is that there is a moment condition $E(h(Y_i,\theta_0))=0$ where $\{Y_i\}_{i=1}^n$ denotes data and θ_0 is the true value of a parameter θ. The moment condition is a kx1 vector, the parameter is a px1 vector.

"Just identified" case

First suppose that k=p. We can find θ so as to solve

$n^{-1}\Sigma_{i=1}^n h(Y_i,\hat{\theta})=0$

Under standard conditions, $\hat{\theta}\to_p \theta$ and here is a sketch of the derivation of the asymptotic distribution:

Suppose that $n^{-1}\Sigma_{i=1}^n \frac{\partial h(Y_i,\theta_0)}{\partial\theta}\to_p D=E(\frac{\partial h(Y_i,\theta_0)}{\partial\theta})$ and $n^{-1/2}\Sigma_{i=1}^n h(Y_i,\theta_0)\to_d N(0,E)$.

$$0=n^{-1}\Sigma_{i=1}^n h(Y_i,\hat{\theta})=n^{-1}\Sigma_{i=1}^n h(Y_i,\theta_0)+(\hat{\theta}-\theta_0)n^{-1}\Sigma_{i=1}^n \frac{\partial h(Y_i,\theta_0)}{\partial\theta}$$

$$\therefore n^{1/2}(\hat{\theta}-\theta_0)n^{-1}\Sigma_{i=1}^n \frac{\partial h(Y_i,\theta_0)}{\partial\theta}=-n^{-1/2}\Sigma_{i=1}^n h(Y_i,\theta_0)$$

$$\therefore n^{1/2}(\hat{\theta}-\theta_0)\to_d N(0,D^{-1}ED'^{-1})$$

"Overidentified case"

If k>p, we can find θ so as to solve

$\hat{\theta}=\arg\min_\theta (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))'W(n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))$

for some symmetric positive-definite kxk weight matrix W. Again, under standard conditions, $\hat{\theta}\to_p\theta$. Here is a sketch of the derivation of the asymptotic distribution of the GMM estimator.

FOC: $\frac{\partial n^{-1}\Sigma_{i=1}^n h(Y_i,\hat{\theta})}{\partial\theta}'Wn^{-1}\Sigma_{i=1}^n h(Y_i,\hat{\theta})=0$

$$h(Y_i,\hat{\theta})=h(Y_i,\theta_0)+\frac{\partial h(Y_i,\theta_0)}{\partial\theta}(\hat{\theta}-\theta_0)$$

$$\therefore \frac{\partial n^{-1}\Sigma_{i=1}^n h(Y_i,\hat{\theta})}{\partial\theta}'W[n^{-1}\Sigma_{i=1}^n (h(Y_i,\theta_0)+\frac{\partial h(Y_i,\theta_0)}{\partial\theta}(\hat{\theta}-\theta_0))]=0$$

$$\therefore D'W[n^{-1}\Sigma_{i=1}^n h(Y_i,\theta_0)+D(\hat{\theta}-\theta_0)]=0$$

$$\therefore D'W[n^{-1/2}\Sigma_{i=1}^n h(Y_i,\theta_0)+Dn^{1/2}(\hat{\theta}-\theta_0)]=0$$

$$\therefore n^{1/2}(\hat{\theta}-\theta_0)=-(D'WD)^{-1}D'Wn^{-1/2}\Sigma_{i=1}^n h(Y_i,\theta_0)$$

$\therefore n^{1/2}(\hat{\theta}-\theta_0) \to_d N(0,(D'WD)^{-1}D'WEW(D'WD)^{-1})$

Notes:

● If $W = E^{-1}$ (or a consistent estimate thereof), $n^{1/2}(\hat{\theta}-\theta_0) \to_d N(0,(D'E^{-1}D)^{-1})$.

● Any weight matrix other than E^{-1} will give a less efficient GMM estimator.

● Chamberlain (1987) shows that this estimator is efficient in the family of all estimators that are based on the moment condition $E(h(Y_i,\theta_0)) = 0$.

● If the moment conditions are not serially correlated, then $\hat{E} = \frac{1}{n}\Sigma_{i=1}^n h(Y_i,\theta)h(Y_i,\theta)'$ is a consistent estimate of E. Or there can be serial correlation, in which case a zero-frequency spectral density estimator is needed.

● Typically E will depend on θ; so we can do a two-step estimator starting with the identity weight matrix.

$$\hat{\theta}_{(1)} = \arg\min_\theta (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))' I (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))$$

$$\hat{\theta}_{(2)} = \arg\min_\theta (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))' E(\hat{\theta}_{(1)})^{-1} (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))$$

This is called the two-step GMM estimator. Or we can use "continuously updated GMM"

$$\hat{\theta}_{(cu)} = \arg\min_\theta (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))' E(\theta)^{-1} (n^{-1}\Sigma_{i=1}^n h(Y_i,\theta))$$

for which the FOC has an extra term

● If k=p, the weight matrix doesn't matter. Moreover D is a square matrix and so

$$n^{1/2}(\hat{\theta}-\theta_0) \to_d N(0,(D'E^{-1}D)^{-1}) \Rightarrow n^{1/2}(\hat{\theta}-\theta_0) \to_d N(0,D^{-1}ED'^{-1})$$

<u>Everything is a special case of GMM</u>

Here's how different estimators are special cases of GMM.

(i) OLS is GMM.

$y_i = \beta' x_i + \varepsilon_i$

$E(x_i\varepsilon_i) = 0$ so $E(x_i(y_i - \beta' x_i)) = 0$

k=p; so just identified.

GMM estimator solves $\Sigma_{i=1}^n x_i(y_i - \beta' x_i) = 0 \Rightarrow \hat{\beta} = (\Sigma_{i=1}^n x_i x_i')^{-1}\Sigma_{i=1}^n x_i y_i$

which is just OLS.

(ii) IV is GMM

$y = X\beta + \varepsilon$

$X = Z\pi + v$

$E(Z'\varepsilon) = 0 \Rightarrow E(Z'(y - X\beta)) = 0$

If k=p, GMM solves

$Z'(y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (Z'X)^{-1}Z'y$

which is the usual IV formula with no surplus instruments.

If k>p, the two-step GMM estimator solves

$\hat{\beta} = \arg\min_\beta (y - X\beta)'Z(Z'Z)^{-1}Z'(y - X\beta)$

which means that

$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$
which is the usual 2SLS estimator. This simplifies to the just-identified case when k=p, but in this case only. The continuous updating GMM estimator reduces to LIML.

The general 2SLS formula can be written as $\hat{\beta} = (\hat{X}'X)^{-1}\hat{X}'y$ where $\hat{X} = Z(Z'Z)^{-1}Z'X$ which are the fitted values in a regression of X on Z.

It can also be written as $\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$, which gives the estimator the "two stage least squares" interpretation

- First regress X on Z
- The regress y on the fitted values.

(iii) Maximum likelihood is GMM.
$\hat{\theta} = \arg\max_{\theta} \Sigma_{i=1}^{n} \log f(\theta)$
$\Sigma_{i=1}^{n} \frac{\partial \log f(\theta)}{\partial \theta} = 0$ is a px1 vector of scores.
The maximum likelihood estimator is a just-identified GMM estimator solving
$n^{-1}\Sigma_{i=1}^{n} \frac{\partial \log f(\hat{\theta})}{\partial \theta} = 0$
Suppose that $n^{-1}\Sigma_{i=1}^{n} \frac{\partial^2 \log f(\theta)}{\partial \theta^2} \to_p D$ and $n^{-1/2}\Sigma_{i=1}^{n} \frac{\partial \log f(\theta)}{\partial \theta} \to_d N(0,E)$
Then $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, D^{-1}ED^{-1})$
Importantly, this formula doesn't even require that the density that is specified is correct. If it is not, then this is called a pseudo-maximum likelihood estimator (it's setting the moment condition to zero, but that moment condition isn't actually maximizing the likelihood). If it is correct, then the information equality is $D = -E$ and so things simplify to $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, I^{-1})$ where $I = D = -E$. So these give two formulas for the distribution of MLE; one that is robust to misspecification and one that is not. Both require derivatives and in many cases these derivatives have to be calculated numerically.
All of this carries over to the time series context. Here we factor the density
$f(y_1, \ldots y_T) = \Pi_{t=2}^{T} f(y_t \mid y_1 \ldots y_{t-1}) f(y_1)$
and so write the log likelihood as $\Sigma_{t=1}^{T} \log f(y_t \mid Y_{t-1})$ where Y_{t-1} denotes the history up to time t-1 and Y_0 is empty. This is a again a just-identified GMM estimator based on the condition $E(\frac{\partial \log f(y_t \mid Y_{t-1})}{\partial \theta}) = 0$.

<u>Testing overidentifying restrictions</u>
If k=p, we can see if all the moment conditions are jointly satisfied (if k=p, we cannot, because the conditions will be satisfied at the estimated parameter value by construction). The test statistic is

$$(n^{-1}\Sigma_{i=1}^{n} h(Y_i,\hat{\theta}))' W (n^{-1}\Sigma_{i=1}^{n} h(Y_i,\hat{\theta}))$$

where W is an estimate of E^{-1}. If there is in fact some θ_0 such that $E(h(Y_i,\theta_0)) = 0$ then this test statistic is asymptotically $\chi^2(k-p)$ distributed.

Part 2: Identification

GMM requires not only that $E(h(Y_i,\theta_0)) = 0$ at the "true" parameter value, θ_0 but also that $E(h(Y_t,\theta)) \neq 0$ for $\theta \neq \theta_0$. Otherwise, there is no way to tell apart the true parameter and an imposter.

Moreover, the asymptotic distribution of GMM that we derived clearly requires that $D = E(\frac{\partial h(Y_i,\theta_0)}{\partial \theta})$ is of full column rank.

Definitions: A model is <u>locally identified</u> if $D = E(\frac{\partial h(Y_i,\theta_0)}{\partial \theta})$ is of full column rank.
A model is <u>globally identified</u> if $E(h(Y_t,\theta)) \neq 0$ for all $\theta \neq \theta_0$

The results on consistency and the limiting distribution of GMM require local and global identification.

Now consider the linear IV model

$$y = X\beta + \varepsilon$$
$$X = Z\pi + v$$

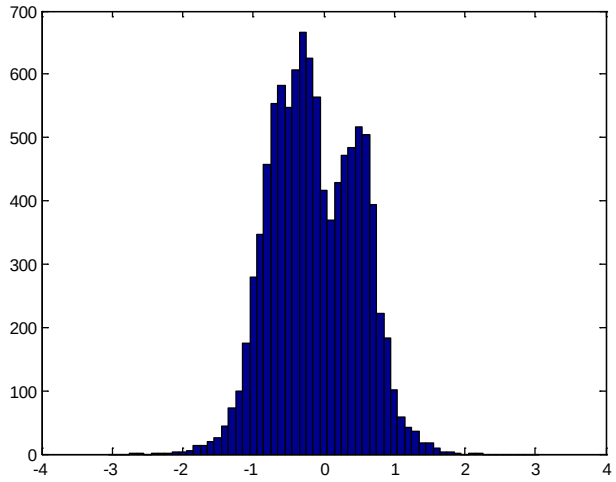
In this case $h(\theta) = Z'(y - X\beta)$ and so $D = E(Z'X) = Z'Z\pi$.
Without loss of generality, let's assume that $Z'Z = I$. So identification (local and global) requires that π is of full column rank (and therefore nonzero).

<u>Complete lack of identification</u>
Suppose that $\pi = 0$ and $k = p = 1$.
Then $\hat{\beta}_{IV} - \beta = (Z'X)^{-1}Z'\varepsilon = (Z'v)^{-1}Z'\varepsilon = (n^{-1/2}Z'v)^{-1}n^{-1/2}Z'\varepsilon$
Now suppose that $n^{-1/2}Z'v \to_d N(0,\sigma_v^2)$ and $n^{-1/2}Z'\varepsilon \to_d N(0,\sigma_\varepsilon^2)$ and the correlation between these two is ρ. Then

$$\hat{\beta}_{IV} - \beta \to_d \frac{\sigma_\varepsilon(\rho + \xi_1)}{\sigma_v \xi_2}$$

where ξ_1 and ξ_2 are independent standard normals. We can see three things:
(i) The estimator is not consistent.
(ii) The estimator is biased, and it is biased in the same direction as OLS.
(iii) The distribution is a Cauchy-type distribution.

Here is the simulated distribution of the 2SLS t-statistic with $\pi = 0$ and $k = p = 1$ and $\rho = -0.3$ in a sample size of 100…it's very nonnormal, centered a bit below zero and actually bimodal!



This basic intuition applies for other values of k and p and even when π is small but nonzero.

Detecting a lack of identification

The classic test is the F-test testing the hypothesis that $\pi = 0$ in a regression of X on Z.
In the case $p = 1$ this is

$$\hat{\pi}'(\hat{\Sigma}_v(Z'Z)^{-1})^{-1}\hat{\pi}/k = \hat{\Sigma}_v^{-1}X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X/k = \hat{\Sigma}_v^{-1}X'Z(Z'Z)^{-1}Z'X/k$$

where $\hat{\Sigma}_v$ is the estimated variance-covariance matrix from the regression.

But identification requires π to have full column rank.

Cragg and Donald (1993) proposed the following test

$$g_{\min} = \text{mineval}(G)$$

where

$$G = \hat{\Sigma}_v^{-1/2}X'Z(Z'Z)^{-1}Z'X\hat{\Sigma}_v^{-1/2}/k$$

The null limiting distribution of G is a Wishart distribution, divided by k. Thus the limiting distribution of $g_{\min}$ is the smallest eigenvalue of a Wishart distribution, divided by k.

(Note: If X_i are iid $N(0, I)$ px1 random vectors, then $\Sigma_{i=1}^n X_i X_i'$ has a Wishart distribution.)

Weak identification

The problem is that even if π is nonzero, it may be so small that the IV estimator is biased and the associated t-statistic is nonnormal in relevant sample sizes. And "relevant" sample sizes may mean very big sample sizes. Stock, Wright and Yogo (2002) gives a review of some of the issues that arise with small π.

Staiger and Stock (1997) derive the limiting distribution of instrumental variables estimators under a "local to zero" asymptotics. As in the near unit root asymptotics, the idea is that this gives a better approximation to the small sample properties of the estimators and test statistics. They assumed that
$\pi = Cn^{-1/2}$. Here is the derivation of the limiting distribution of 2SLS in this case with $k = p$.

Assume that $n^{-1/2}Z'v \rightarrow_d N(0, \Phi_{zv})$, $n^{-1/2}Z'\varepsilon \rightarrow_d N(0, \Phi_{z\varepsilon})$ (where these may be correlated with each other) and $n^{-1}Z'Z \rightarrow_p Q_{zz}$. Then

$$\hat{\beta}_{IV} - \beta = (Z'X)^{-1}Z'\varepsilon = (Z'Z\pi + Z'v)^{-1}Z'\varepsilon = ((n^{-1}Z'Z)c + n^{-1/2}Z'v)^{-1}n^{-1/2}Z'\varepsilon$$

$$\therefore \hat{\beta}_{IV} - \beta \rightarrow_d (Q_{zz}C + \Phi_{Zv})^{-1}\Phi_{Z\varepsilon}$$

This provides a good approximation to the very unusual properties of 2SLS estimators in small samples when π is quite small.

But there is a solution to the weak instruments problem that goes back to work of Anderson and Rubin (1949) and relies on the relationship between tests and confidence sets. Suppose I give up on estimation and just want to test the hypothesis that $\beta = \beta_0$. Then, if the hypothesis is true, $u_0 = Y - X\beta_0$ will be orthogonal to Z. We can test this with the statistic

$$AR = \frac{u_0'Z(Z'Z)^{-1}Z'u_0}{u_0'[I - Z(Z'Z)^{-1}Z']u_0 / (n-k)}$$

which is asymptotically $\chi^2(k)$ distributed (and actually has an exact F-distribution under normality).

What happens if the hypothesis is false? Then $u_0 = u - X(\beta_0 - \beta)$ and this will be correlated with Z, provided that the instruments are relevant. If $\pi = 0$ then the test will not be consistent, but it shouldn't be (how can you reject a claim about β with completely useless instruments?)

What happens if the instruments aren't orthogonal to the error? Then the test will asymptotically reject for all parameter values. So it is a joint test that the instruments are valid and that we have found the right parameter value.

Typically we want to construct a confidence set, not test a hypothesis. But the confidence set can be formed as the inverse of the acceptance region of the test (from a grid). If the instruments are irrelevant, the confidence set will be big, but it should be. For a more formal treatment of this point, see Dufour (1997).

What makes this confidence set work is the fact that the test statistic, AR, is pivotal. Other pivotal statistics may give more powerful tests and smaller confidence sets. One was proposed by Kleibergen (2002) for the case $k > p$

$$K = \frac{u_0'Z\tilde{\pi}(\tilde{\pi}'Z'Z\tilde{\pi})^{-1}\tilde{\pi}'Z'u_0}{u_0'[I - Z(Z'Z)^{-1}Z']u_0 / (n-k)}$$

where

$$\tilde{\pi} = (Z'Z)^{-1}Z'(X - \hat{\Sigma}_{uu}^{-1}\hat{\Sigma}_{ux}u_0)$$

$$\hat{\Sigma}_{xu} = \frac{1}{n-k}X'(I - Z(Z'Z)^{-1}Z')u_0$$

and

$$\hat{\Sigma}_{uu} = \frac{1}{n-k}u_0'(I - Z(Z'Z)^{-1}Z')u_0 .$$

Notes: (i) In the just-identified case, the $\tilde{\pi}$s would cancel, and we are back to AR.
(ii) Conditional on $\tilde{\pi}$, the null limiting distribution of K is $\chi^2(p)$ and is independent of $\tilde{\pi}$ so this holds unconditionally as well. So the degrees of freedom are smaller than in the AR test.
(iii) Intuitively, the K statistic is projecting u_0 onto a smartly chosen subspace of the instruments and so should be (and in small samples usually is) more efficient. The sense in which it is wisely chosen is that $\tilde{\pi}$ is forced to be asymptotically independent of u_o.

The GMM Case

In the GMM context, we can also use a weak-identification robust confidence set. The set is

$$\{\theta : n^{-1/2}\Sigma_{i=1}^{n}h(Y_i,\theta)'E(\theta)^{-1}n^{-1/2}\Sigma_{i=1}^{n}h(Y_i,\theta) \le F(k)\}$$

where $F(k)$ is the upper percentile of a $\chi^2(k)$ distribution and $E(\theta)$ is an estimate of the asymptotic variance of $n^{-1/2}\Sigma_{i=1}^{n}h(Y_i,\theta)$. Concretely, in the case without serial correlation,

$$E(\theta) = n^{-1}\Sigma_{i=1}^{n}h(Y_i,\theta)h(Y_i,\theta)'$$

This is the analog of the AR test and makes no assumptions of local or global identification. It was proposed by Stock and Wright (2000). Note that the weight matrix is evaluated at the *hypothesized* θ, not an *estimate* of this parameter, and this is actually crucial. Kleibergen (2005) proposes an alternative confidence set that may be smaller.

As discussed in Stock, Wright and Yogo (2002), the bias of 2SLS depends on the concentration parameter, which is $\pi'\pi / k$. It is the *average strength* of the instruments that matters; adding to the number of instrument can be harmful if it reduces the average strength. An alternative way of looking at identification problems thinks of π as fixed, but the number of instruments as large. Bekker (1994) shows that if π is fixed but $k = \alpha n$ then

$$\hat{\beta}_{2SLS} \to_p \beta + (Q + \alpha\Omega)^{-1}\Omega\Sigma_{vu}$$

where $\Sigma_{vu} = E(u_i v_i)$ and $\Omega = E(v_i v_i')$. Accordingly, 2SLS is not consistent.

Part 3: Lots of applications

(i) New-Keynesian Phillips curve.

$$\pi_t = \beta_0 + \beta_1\pi_{t-1} + \beta_2 E_t(\pi_{t+1}) + \beta_3 s_t + \varepsilon_t$$

Let $v_{t+1} = \pi_{t+1} - E_t(\pi_{t+1})$ be the one-step ahead forecast error. Under rational expectations, this should be orthogonal to any variable in the information set at time t. So we can substitute

$$\pi_t = \beta_0 + \beta_1\pi_{t-1} + \beta_2[\pi_{t+1} - v_{t+1}] + \beta_3 s_t + \varepsilon_t$$

$$\therefore \pi_t = \beta_0 + \beta_1 \pi_{t-1} + \beta_2 \pi_{t+1} + \beta_3 s_t + u_t$$

where $\varepsilon_t - \beta_2 v_{t+1}$. But this equation cannot be consistently estimated by OLS, because π_{t+1} is correlated with the error. It can however be estimated by IV, using any variable in the information set at time t as an instrument.

The initial work on this (Gali and Gertler (1999)) used many lags as instruments. More recently, authors have used smaller numbers of instruments. Identification here is about the predictability of inflation and that is quite doubtful. Generally, the coefficient β_2 is found to be less than 1, but imprecisely estimated.

Note also that this New Keynesian Phillips curve comes from a log-linearization about a steady state inflation rate. Cogley and Sbordone (2007) give a version with I(1) inflation.

Barnichon and Mesters (2020) have a clever alternative way of estimating this forward looking New-Keynesian Phillips curve. They take the above equation, and instrument it with a vector of lagged monetary policy shocks. If ξ_t is a monetary policy shock (Romer and Romer (2004), high frequency data or something else), then define $z_t = (\Sigma_{h=0}^{H} \xi_{t-h}, \Sigma_{h=0}^{H} h\xi_{t-h}, \Sigma_{h=0}^{H} h^2 \xi_{t-h})'$ and then use that as the instruments. As long as monetary policy shocks affect future inflation, they are relevant instruments. They may be weak, but it is OK to use Anderson-Rubin methods to be robust to weak identification.

(ii) Forward-looking Taylor rule

$$i_t = \beta_0 + \beta_1 E_t(\pi_{t+1}) + \beta_2 E_t(y_{t+1}) + \varepsilon_t$$

Likewise, the expectations can be replaced by future realized values, appealing to rational expectations.

(iii) Angrist and Krueger (1991). This paper relaunched the interest in weak instruments. A classical regression in labor economics is of income on education. It suffers from endogeneity bias. Angrist and Krueger revisited this regression, but used quarter-of-birth interacted with state-of-residence as dummies. They therefore used many weak instruments.

Later Bound, Baker and Jaeger (1995) pointed out that generating random quarter-of-birth dummies (useless instruments by construction) could generate similar results. The Angrist and Krueger paper had in effect used many weak instruments, which is precisely the situation in which the small sample properties of IV estimators are worst.

(iv) Regressions of growth on measures of financial intermediation, using legal origin as dummies.

(v) The consumption CAPM

$E_t(\delta R_{t+1}(\frac{C_{t+1}}{C_t})^{-\gamma}) = 1$ is the Euler equation with CRRA prefernces

$$\therefore E([\delta R_{t+1}(\frac{C_{t+1}}{C_t})^{-\gamma} - 1] \otimes Z_t) = 0$$

where Z_t is any vector in the information set at time t. Unlike the linear examples, the optimization in this case has to be numerical. Letting Z_t denote lags of returns and consumption growth, the point estimates of δ are close to 1; those of γ are close to 0. Weak-identification robust confidence sets however indicate large values of γ. The required derivatives for standard errors are available in closed form as the derivative of the moment condition with respect to δ is $R_{t+1}(\frac{C_{t+1}}{C_t})^{-\gamma} \otimes Z_t$ while that respect to γ is $\delta R_{t+1} \ln(\frac{C_{t+1}}{C_t}) \otimes Z_t$.

(vi) Treatment of generated regressors. Consider the model

$$y_i = \beta z_i^* + \varepsilon_i$$
$$z_i^* = \delta w_i$$
$$z_i = z_i^* + u_i$$

where ε_i and u_i are iid with mean zero and variance 1 and are mutually independent and w_i is independent of ε_i and u_i. Suppose also that $E(w_i^2) = 1$

Suppose that we observe y_i, z_i and w_i. A tempting procedure is the following two step estimator. In step 1, regress z_i on w_i and get an estimate $\hat{\delta}$. In step 2, regress y_i on $\hat{\delta} w_i$. This is a problem with a *generated regressor*. The uncertainty in δ affects our standard error. But we can control for this by thinking of it as a just-identified GMM model. The numerical parameter estimates we get are the same as using two moment conditions:

(i) $E((y_i - \beta\delta w_i)w_i) = 0$

(ii) $E((z_i - \delta w_i)w_i) = 0$

Under the stated (special) assumptions the variance-covariance matrix of the moment conditions is just the identity matrix and the matrix of derivatives is $\begin{pmatrix} -\delta & -\beta \\ 0 & -1 \end{pmatrix}$. GMM gives a trick for solving a hard problem.

(vii) Factor models in finance. Suppose that we have a scalar factor model:

$R_{it} = \alpha_i + \beta_i f_t + \varepsilon_{it}$

$E(R_{it}) = \beta_i \lambda$

We have moment conditions

$E(R_t - \alpha - \beta f_t) = 0$

$E((R_t - \alpha - \beta f_t) f_t) = 0$

$E(\beta' R_t) = \beta' \beta \lambda$

and there are a total of $2n+1$ moment conditions where n is the number of test assets, so the system is just identified. GMM is equivalent to estimating the model by the Fama-MacBeth procedure. GMM has the advantage that it delivers standard errors that automatically take account of the fact that the βs are estimated.

To be concrete, if Σ is the variance-covariance matrix of the errors, then the variance-covariance matrix of the moment conditions is:

$$\begin{pmatrix} \Sigma & E(f)\Sigma & \Sigma\beta \\ E(f)\Sigma & E(f^2)\Sigma & E(f)\Sigma\beta \\ \beta'\Sigma & E(f)\beta'\Sigma & \beta'\Sigma\beta + (\beta'\beta)^2 Var(f) \end{pmatrix}$$

and the matrix of derivatives is available in closed form as:

$$\begin{pmatrix} -I_n & -E(f)I_n & 0 \\ -E(f)\mathrm{I}_n & -E(f^2)I_n & 0 \\ 0 & -\beta'\lambda & -\beta'\beta \end{pmatrix}$$

An aside on the subject of factor models in finance. Fama and MacBeth (1973) proposed an alternative methodology for standard errors. It is to first run n time series regressions to estimate the β_i s, then run T cross sectional regressions to get an estimate of λ for each time period, $\hat{\lambda}_t$ and then the overall estimator of λ is $T^{-1}\Sigma_{t=1}^{T}\hat{\lambda}_t$ with an estimated variance of $\frac{1}{T}\Sigma_{t=1}^{T}\frac{(\hat{\lambda}_t-\lambda)^2}{T-1}$. A variant on this in the second step is to regress average returns on the estimated betas and for this latter estimator, GMM will give the correct standard errors taking account of the fact that betas were estimated in the first stage.

(vii) Simulated method of moments (SMM). SMM is a special case of GMM where the theoretical moments cannot be worked out directly, and so have to be obtained through simulation. It arises in structural Industrial Organization.

Here is the setup. We have some moments of data, $E(Y_i)$, $E(Y_i^2)$ etc.. Write these in the form $E(f(Y_i))$. These depend on an unknown parameter vector θ . The moment condition is that $E(f(Y_i)-m(\theta))=0$. The problem is that the function is not observed, but can be simulated. Thus we replace $m(\theta)$ by $\frac{1}{S}\Sigma_{j=1}^{S}m_j(\theta)$ where $m_j(\theta)$ is the j th simulation. We then define $e_i(\theta)=Y_i-\frac{1}{S}\Sigma_{j=1}^{S}m_j(\theta)$ and solve the GMM problem:

$$\hat{\theta}=\arg\min_{\theta}\Sigma_{i=1}^{n}e_t(\theta)'W\Sigma_{i=1}^{n}e_t(\theta)$$

This is a standard GMM problem, always involving a numerical optimization. The limiting distribution will take account of the fact that $m(\theta)$ is estimated, not known. With the usual efficient weight matrix, the limiting distribution is:

$$n^{1/2}(\hat{\theta}-\theta_0)\rightarrow_d N(0,(1+k)D^{-1}ED'^{-1})$$

where $n^{-1}\Sigma_{i=1}^{n}\frac{\partial^2 e_i(\theta)}{\partial\theta^2}\rightarrow_p D$, E is the variance of $e_i(\theta)$ and $k=S/n$. A few things to note:

(a) With enough draws, the simulation is perfect, and k goes to zero. In simple applications, you can and should take enough draws that the simulation error can be neglected. In some applications, that may not be feasible.

(b) When simulating the moment condition, the same seed must be used for each candidate θ. More precisely, for the simulations, you draw one set of shocks and use this same set for each parameter.
(c) The version of SMM described here is for iid data, but standard extensions are available for the case of time series dependence.
(d) If moments are in very different scales and two-step GMM is used, the first step can be very sensitive to scaling (which of course feeds through to the whole GMM procedure). This problem can be avoided by defining $e_i(\theta) = (Y_i - \frac{1}{S}\Sigma_{j=1}^{S} m_j(\theta)) / \frac{1}{S}\Sigma_{j=1}^{S} m_j(\theta)$.

Here is a very simple illustration of the idea. Suppose that $X_i \sim iidN(\mu,1)$ and we observe the average of n observations on $Y_i = \exp(X_i)$. Call this $\bar{Y}$. We know that $E(Y_i)$ is a function of μ, $f(\mu)$, but let's pretend that we don't know a closed form for this function (actually of course we do). We have a moment condition $E(Y_i - f(\mu)) = 0$. This is a case of just identified GMM, solving the equation $\bar{Y} = f(\mu)$, if we know the function f. But for any given μ we can easily simulate m normal random variables, and exponentiate and average them to approximate $f(\mu)$.

(viii) Indirect inference. This is a similar idea to simulated method of moments. Suppose that I have data from a model and I can simulate data from this model with a px1 vector of parameters θ. Suppose that there is an auxiliary model that approximates the data with a kx1 parameter vector β. For this model, we can write down a log-likelihood function

$$l = \Sigma_{i=1}^{n} \log(f(y_i \mid x_i, \beta))$$

This is not the actual log-likelihood of the data, so it is a pseudo-log-likelihood. Next find the value of β that maximizes this function, $\hat{\beta}$. Next, for any candidate θ, simulate data from the true model---let these data be $\{\tilde{y}_{i,m}(\theta)\}_{i=1}^{n}$ on the mth of M simulations. Define

$$\tilde{\beta}_m(\theta) = \arg\max_{\beta} \Sigma_{i=1}^{n} \log(f(\tilde{y}_{i,m}(\theta) \mid x_i, \beta))$$

Lastly, our indirect inference estimator is:

$$\hat{\theta} = \arg\min_{\theta} (\frac{1}{M}\Sigma_{m=1}^{M} \tilde{\beta}_m(\theta) - \hat{\beta})' W (\frac{1}{M}\Sigma_{m=1}^{M} \tilde{\beta}_m(\theta) - \hat{\beta})$$

Defining $D = \operatorname{plim} n^{-1}\Sigma_{i=1}^{n} \frac{\partial^2 \log f(y_i \mid x_i, \beta(\theta))}{\partial \beta^2}$ and $E = \operatorname{Var}(\frac{\partial \log f(y_i \mid x_i, \beta(\theta))}{\partial \beta})$, the optimal weight matrix is $DE^{-1}D$ and the asymptotic distribution of the indirect inference estimator is

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, (1 + \frac{1}{M})[\frac{\partial b(\theta)'}{\partial \theta} DE^{-1} D \frac{\partial b(\theta)}{\partial \theta}]^{-1})$$

where $b(\theta)$ is the probability limit of $\tilde{\beta}_m(\theta)$.

(ix) Bartik instruments. This is a recipe for a widely applicable instrumental variables strategy. Suppose that we have a cross-sectional regression of wage growth on employment growth in region i:

$$w_i = \beta_0 + \beta_1 e_i + u_i$$

and we want to interpret this as a labor supply curve. To do this we need an instrument that is correlated with labor demand but not supply. The instrument that Bartik (1991) proposed is to take the initial share of industry k in employment in location i, w_{ik}, and then to instrument e_i by $\Sigma_k w_{ik} g_k$ where g_k is national employment growth in industry k. The idea is that employment is growing in the region just because it happened to have started out with a large exposure to a growth industry and that is correlated with labor demand but not labor supply. It is a widely used approach in many contexts (see Goldsmith-Pinkham, Sorkin and Swift (2020)).

(x) Granular instrumental variables. This is another recipe for a widely applicable instrumental variables strategy in macro, recently proposed by Gabaix and Koijen. Here is the idea. Let S_i be the share of country i in the world market. Demand by country i at time t is

$$D_{it} = S_i y_{it}$$

where

$$y_{it} = \phi p_t + \eta_t + u_{it}$$

Global demand is $y_{St} = \Sigma_{i=1}^n S_i y_{it}$. Define $y_{Et} = \Sigma_{i=1}^n \frac{1}{n} y_{it} = \Sigma_{i=1}^n \frac{1}{n}\frac{1}{S_i} D_{it}$, $u_{St} = \Sigma_{i=1}^n S_i u_{it}$ and $u_{Et} = \Sigma_{i=1}^n n^{-1} u_{it}$. Supply is:

$$s_t = \lambda p_t + v_t$$

We want to regress y_{St} on p_t but OLS identified a mix of demand and supply shocks. To see this,

$$y_{St} = \phi p_t + \eta_t + u_{St} = \lambda p_t + v_t$$

$$\therefore p_t = \frac{\eta_t + u_{St} - v_t}{\lambda - \phi}$$

and so in a regression of global output on price, the explanatory variable will be correlated with η_t.

But the regression of y_{St} on p_t with instrument $Z_t = y_{St} - y_{Et} = \Sigma_{i=1}^n (S_i - \frac{1}{n}) u_{it}$ identifies λ in the supply curve. This is called the granular instrument. It works because

$$E(Z_t v_t) = E(\Sigma_{i=1}^n (S_i - \frac{1}{n}) u_{it} v_t) = 0$$

Moreover, the regression

$$y_{Et} = \phi p_t + \eta_t + u_{Et}$$

estimated by IV using the same granular instrument identifies ϕ in the demand curve. This works because

$$E(Z_t(\eta_t + u_{Et})) = E(\Sigma_{i=1}^n (S_i - \frac{1}{n}) u_{it}(\eta_t + u_{Et})) = E(\Sigma_{i=1}^n (S_i - \frac{1}{n}) u_{it} \frac{1}{n} \Sigma_{j=1}^n u_{jt}) = \Sigma_{i=1}^n \frac{\sigma_u^2}{n}(S_i - \frac{1}{n}) = 0$$

Handout on VARs

Suppose that Y_t is a px1 vector of time series such that $A(L)Y_t = u_t$ where the innovations u_t have variance-covariance matrix Σ and $A(L)$ is a matrix lag polynomial of order n. We assume that the VAR is stationary (all solutions to the equation $|A(L)| = 0$ lie outside the unit circle).

- MA representation: $Y_t = A(L)^{-1}u_t = C(L)u_t$
- Companion form: (for the case $n = 2$) $\begin{pmatrix} Y_t \\ Y_{t-1} \end{pmatrix} = A\begin{pmatrix} Y_{t-1} \\ Y_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \end{pmatrix}$ where $A = \begin{pmatrix} A_1 & A_2 \\ I & 0 \end{pmatrix}$. .
  This trick means that a univariate AR can be written as a VAR.

The VAR can be estimated by OLS equation-by-equation (same as SUR estimator, if there are no parameter restrictions). We'll come back to estimation below.

A VAR can be used for forecasting in a rather atheoretical way. But the errors are one-step-ahead prediction errors for each element of Y_t, that lack a substantive macroeconomic interpretation. Sims (1980) proposed an approach for using a VAR to address more structural macroeconomic questions.

Let's assume that there are underlying structural errors ε_t such that $u_t = R\varepsilon_t$. We assume that the structural shocks are independent and have variance 1 (a question of the units that the shocks are measured in). So the variance-covariance matrix of ε_t is the identity matrix and $\Sigma = RR'$. Another approach would be to normalize the diagonal elements of R to unity instead.

We can write
$Y_t = C(L)R\varepsilon_t = A(L)^{-1}R\varepsilon_t = D(L)\varepsilon_t$
where $D(L) = D_0 + D_1L + D_2L^2 ....$ and can then figure out "structural impulse responses".. $\frac{\partial y_{t+h}}{\partial \varepsilon_t} = D_h$.
We can also work out a "forecast error variance decomposition", because
$Y_{t+k} - Y_{t+k|t} = D_0\varepsilon_{t+k} + D_1\varepsilon_{t+k-1} ... + D_{k-1}\varepsilon_{t+1}$
Taking, without loss of generality, the first element of Y_t, the variance of its k-period ahead forecast errors is $\Sigma_{l=0}^{k-1}\Sigma_{j=1}^{p} d_l(1,j)^2$ where the ij th element of the matrix D_l is written as $d_l(i,j)$.
So the fraction of the variance of the forecast error in the first element of Y_t that is due to the first structural shock is $\frac{\Sigma_{l=0}^{k-1} d_l(1,1)^2}{\Sigma_{l=0}^{k-1}\Sigma_{j=1}^{p} d_l(1,j)^2}$

and all the forecast errors can be decomposed in this way.
We can estimate $A(L)$ and Σ and know that $\Sigma = RR'$. But we need to know R to work out structural impulse responses, forecast error variance decompositions and so on. Identification is about solving for R from Σ. R has p^2 elements; Σ has $\frac{p(p+1)}{2}$ elements, so we need more restrictions to solve for R from Σ. These are the identifying restrictions of the VAR.

A common question that a VAR tries to figure out (though by no means the only one) is what are the effects of a monetary policy shock (others include effects of fiscal policy shocks and technology shocks). It's no use regressing macroeconomic outcomes on shocks to interest rates, because the most obvious reason why interest rates would be tightened is that the Fed expects inflation or growth to be high in the future. Monetary policy is endogenous. The structural VAR though is trying to figure out the effects of exogenous monetary policy shocks—the FOMC changing monetary policy not because of differences in the outlook for growth or inflation

1. Direct measurement of monetary policy shocks. This approach has been developed in particular by David and Christina Romer. It's not really a structural VAR identification, but I cover it here in any case.

One approach, the so-called "Narrative approach" (Romer and Romer (1989)) involves reading FOMC minutes/records to identify 6 occasions when there was a contractionary monetary policy shock that owed to a change in the apparent preferences of policy makers. Let D_t be the dummy that is 1 in the month of such a shock and 0 otherwise. Then estimate the impulse responses by running the regression

$$y_t = a_o + \Sigma_{j=1}^{24} b_j y_{t-j} + \Sigma_{k=0}^{36} c_k D_{t-k} + \text{error} \qquad (1)$$

and the impulse responses can then be solved out. The paper found that activity was substantially reduced, reaching a trough after about a couple of years. These shocks also accounted for much of the variability in output growth.

Romer and Romer (2004) propose another direct measure of monetary policy shocks. Before each FOMC meeting, the Federal Reserve staff prepare a forecast, known as the Greenbook. Romer and Romer regress the change in the Federal Funds target announced at each FOMC meeting on: the old fed funds rate, the Greenbook growth forecasts for quarters $t-1, t, t+1, t+2$, the revisions to those growth forecasts, the inflation forecasts, the revisions to the invlation forecasts and the current quarter unemployment forecast. The residuals from this regression are then treated as estimates of monetary policy shocks. They then estimate a regression of the same form as (1) to estimate the impulse responses. Again large real effects of monetary policy shocks are found, peaking after a couple of years.

2. Cholesky restrictions. Let's consider a toy model for concreteness. $Y_t = (g_t, \pi_t, r_t)'$: growth, inflation and short term interest rates. Let the structural shocks be ε_t^g, ε_t^π and ε_t^r. If we assume that a shock to ε_t^π has no effect on π_t (contemporaneously) and that a shock ε_t^r has no effect on

g_t or π_t then R must be lower triangular. The solution to the equation $\Sigma = RR'$ is unique (up to a flipping of signs).

Often applying this to estimating the effects of monetary policy shocks, the effect of shocks on growth is substantial and peaks after a few quarters (seems reasonable), but the effect of a monetary policy tightening is initially to raise prices (counter to the conventional wisdom). This is the "price puzzle" (Intuitively, what might be going wrong here?).

Other "short-run" restrictions can be imposed. For example, Blanchard and Perotti identify a VAR in which the contemporaneous effects of output and inflation shocks on the deficit are not assumed to be zero, but rather are imposed from reading spending and tax code rules.

Short-run restrictions may also be used to identify the effect of the monetary policy shock alone. Suppose that the monetary policy rule is

$$r_t = \gamma' I_t + \varepsilon_t^r$$

where I_t are variables in the information set at time t that the Fed observes. Suppose that there is no contemporaneous feedback from the monetary policy shock to I_t. Then we can run the regression of r_t on I_t and get an estimate of ε_t^r. By the "no feedback" assumption OLS is consistent. This involves fewer restrictions than the Cholesky ordering, but also only allows us to identify one particular shock (monetary policy). The impulse responses can then be estimated by regressing Y_t on $\varepsilon_t^r, \varepsilon_{t-1}^r, \varepsilon_{t-2}^r, \ldots$. This was the approach used by Christiano, Eichenbaum and Evans (1996). Their VAR had real GDP, the GDP deflator, commodity prices, nonborrowed reserves, the federal funds rate and total reserves. They argued that inclusion of commodity prices made the "price puzzle" disappears.

3. Long-run restrictions (Blanchard and Quah). Suppose that Δy_t is output growth and u_t is the unemployment rate. The vector of time series is $Y_t = (u_t, \Delta y_t)'$. Suppose that $A(L)Y_t = u_t$ is the reduced form VAR and that

$$Y_t = C(L)u_t = D(L)\varepsilon_t$$

where the structural shocks are $\varepsilon_t = (\varepsilon_t^D, \varepsilon_t^S)'$ ("transitory" and "permanent" or "demand" and "supply" shocks). By the Beveridge-Nelson decomposition, we can write

$$y_t = D_{21}(1)\Sigma_{s=1}^t \varepsilon_s^D + D_{21}^*(L)\varepsilon_t^D + D_{22}(1)\Sigma_{s=1}^t \varepsilon_s^S + D_{22}^*(L)\varepsilon_t^S$$

The identifying assumption of Blanchard and Quah is that a demand shock has no permanent effect on output, and so $D_{21}(1) = 0$.

$$D(L)\varepsilon_t = C(L)u_t \Rightarrow D(1)D(1)' = C(1)\Sigma C(1)' \quad (2)$$

$$D(L) = C(L)R \Rightarrow D(1) = C(1)R \quad (3)$$

From (1), $D(1)$ is the Cholesky factor of $C(1)\Sigma C(1)'$.

From (2), $R = C(1)^{-1}D(1)$.

So the algorithm is $R = C(1)^{-1}Chol\{C(1)\Sigma C(1)'\} = A(1)^{-1}Chol\{A(1)^{-1}\Sigma A(1)^{-1}'\}$.

4. Agnostic identification/sign restrictions (Faust (1998)). The restriction that $\Sigma = RR'$, coupled with restrictions on the signs of the contemporaneous impulse responses (e.g. a monetary policy tightening does not increase inflation this period) isn't enough to identify a unique R, but it is enough to identify a set of possible values of R. We say that R is set-identified, not point-identified. Suppose that R^* denotes the set of possible values of R. Then the impulse responses are given by $(\inf_{R\in R^*} C_j R, \sup_{R\in R^*} C_j R)$.

The idea of identification through sign restrictions has been active in recent years. There is also a Bayesian approach to sign restrictions developed by Uhlig (2005). We return to this below. Here is a recent example of sign restrictions. Cieslak and Pang (2021) consider a VAR in daily data on stock returns and daily short-, medium- and long-term bond yield changes. They specify sign restrictions like growth surprises raise stock prices and yields but the effect on short yields is bigger than that on long yields. They similarly identify monetary policy and risk premium surprises.

5. External instruments. Suppose that there is some variable Z_t that is correlated with a particular structural shock, but not with others. For example, Z_t could be the change in interest rates in a small window around an FOMC announcement, which we might think is correlated with a structural monetary policy shock, but not other structural shocks (Gertler and Karadi (2013)). Let this structural shock be the first shock, without loss of generality. Then $E(Z_t u_t) = R_1 E(Z_t \varepsilon_t)$. If we regress u_t on Z_t we identify R_1 up to scale.

External instruments have become very widely used. There are two other ways of implementing the idea:

(i) One alternative implementation is proposed by Paul (2020). It involves augmenting the VAR with the instrument. So the equation is

$$Y_t = A_1 Y_{t-1} + ... + A_n Y_{t-n} + \gamma Z_t + \varepsilon_t$$

The contemporaneous effects of the shock are estimated by γ (subject to rescaling the first element to 1). Then the remaining impulse responses are implied by the VAR. For example, the impulse response at lag 1 is $A_1\gamma$. This gives numerically the same impact impulse response as the regression of u_t on Z_t. The estimated impulse responses at longer horizons may be slightly different.

(ii) Another is to do a VAR in the variable $(Z_t, Y_t')'$ with a recursive ordering (so the instrument is ordered first). This is sometimes called an internal instrument. See Plagborg-Moller and Wolf (2021) for more details. The idea is that we get an augmented VAR (called a proxy SVAR) of the form:

$$\begin{pmatrix} 1 & 0 \\ 0 & A_0 \end{pmatrix} \begin{pmatrix} Z_t \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} Z_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \beta\eta_{1t} \\ \eta_t \end{pmatrix}$$

This can be identified by a Cholesky ordering.

6. Identification from heteroskedasticity. Suppose that there are two regimes in which the structural errors have variance $\Sigma_{\varepsilon 1}$ and $\Sigma_{\varepsilon 2}$. Normalize R to have 1s on the diagonal. Then the

variance-covariance matrix of reduced form errors is $R\Sigma_{\varepsilon,1}R'$ in the first regime and $R\Sigma_{\varepsilon,2}R'$ in the second regime. The total number of unknowns is

R	$p^2 - p$
$\Sigma_{\varepsilon,1}$	p
$\Sigma_{\varepsilon,2}$	p
Total	$p^2 + p$

The number of equations is $p(p+1)/2$ in each period, of $p^2 + p$ in total. So the number of equations equals the number of unknowns, giving a just-identified problem.

Two problems:
(1) What if R changes across regimes?
(2) What if the variance-covariance matrix of the structural errors does not change (much) across regimes?

Rigobon and Sack (2003) estimate the effects of monetary policy shocks on asset prices in this way in a VAR with stock prices interest rates and short-term interest rates. The two regimes are (i) days of FOMC announcements and monetary policy testimonies and (ii) all other days. It seems reasonable to suppose that the variance of structural monetary policy shocks is greater in the first regime than the second. They estimate the effects of a surprise tightening in monetary policy on stock price; a 25 basis point monetary policy surprise lowers stock prices by about 2 percentage points.

7. Every possible identification implies a different forecast error variance decomposition. We could select the identification based on this. For example, Francis et al. (2014) consider a VAR in labor productivity, hours, the consumption-output ratio and the investment-output ratio. They want to identify a technology shock. They do so by maximizing the fraction of the forecast-error variance in labor productivity at the 40 quarter horizon that is explained by the technology shock.

8. Narrative identification. This is an idea pioneered by Antolin-Diaz and Rubio-Ramirez (2018). The idea is that we have strong beliefs about the sign of a shock in a given period. For example, we would asset that in the oil crisis, the oil supply shock was negative. Each possible identification implies a complete set of structural shocks. If an identification implies a shock that violates the restriction, then it gets ruled out. This is very similar to sign restrictions, except that it applies to the shock in specific periods, not the impulse response.

Notes
1. Often researchers like to write a structural VAR as

$$A_0 Y_t = A_1 Y_{t-1} .... + A_n Y_{t-n} + \varepsilon_t$$

where ε_t are the structural shocks. This is another way of writing the VAR, with $A_0 = R^{-1}$. Pre-multiplying through by R gives the reduced form VAR.
2. We can also write $\varepsilon_t = Su_t$ where $S = R^{-1}$. If R_i is the ith column of R, and S_i is the ith row of S, then $S_i = R_i'\Sigma^{-1}$.

3. If you are just interested in the effect of one shock (very common) and are using the Cholesky identification, then the full ordering of the variables does not actually matter. All that matters is which variables are above and below the shock of interest. So if the monetary policy shock is ordered last (or first) the ordering of the other variables does not make any difference.

Estimation with restrictions

It is well known the multivariate maximum likelihood estimation of a system of equations is equivalent to OLS equation-by-equation if the coefficients are unrestricted and
(i) All regressions have the same right-hand-side variables, or
(ii) The errors are uncorrelated.

Accordingly a VAR with the same number of lags in each equation, or with unrestricted coefficient estimates, can be estimated by OLS and this is numerically the same as the system MLE. But otherwise it is not. The MLE maximizes the log-likelihood function

$$-\frac{1}{2}\Sigma_{t=1}^{T}\varepsilon_t'\Sigma^{-1}\varepsilon_t - \frac{1}{2}\log|\Sigma|$$

where $\varepsilon_t = A(L)Y_t$ with respect to Σ and the parameters in $A(L)$

Inference about VAR coefficients and impulse responses

Let α be the vector $vec([A_1 \;\; A_2 .... A_n]')$ which is of order $p^2 n$x1 and let $\hat{\alpha}$ and $\hat{\Sigma}$ denote the OLS estimates of α and Σ. Also define

$$Q = \begin{pmatrix} \Gamma(0) & \Gamma(1) & \dots & \Gamma(n-1) \\ \Gamma(1) & \Gamma(0) & & \\ \vdots & & \ddots & \vdots \\ \Gamma(n-1) & & \cdots & \Gamma(0) \end{pmatrix}$$

where $\Gamma(j) = E(Y_t Y_{t-j}')$. Then

$$T^{1/2}(\hat{\alpha} - \alpha) \rightarrow_d N(0, \Sigma \otimes Q^{-1})$$

and

$$T^{1/2}(vech(\hat{\Sigma}) - vech(\Sigma)) \rightarrow_d N(0, J)$$

Moreover these distributions are independent of each other. If we suppose that the errors are normal, then

$$J = 2D_p(\Sigma \otimes \Sigma)D_p'$$

where $vech(.) = D_p vec(.)$. For example, $D_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

This allows us to conduct tests for individual coefficients in the VAR and to do tests for Granger causality. Also, it gives standard errors for impulse responses, because an impulse response is a nonlinear function of α and Σ, which can be called $f(\alpha, \Sigma)$ and so, by the delta method (Runkle (1987))

$$T^{1/2}\{f(\hat{\alpha}, \hat{\Sigma}) - f(\alpha, \Sigma)\} \rightarrow_d N(0, f_\alpha(\Sigma \otimes Q^{-1}) f_\alpha' + f_\Sigma J f_\Sigma')$$

An alternative is to use a bootstrap. The bootstrap involves the following steps:
1. Estimate the VAR.
2. Resample from the residuals with replacement.
3. Construct new artificial datasets.
4. In each, estimate the coefficient of interest (e.g. impulse response). Repeat (2)-(4) (say) 1,000 times.
5. A confidence interval can be formed from the percentiles of the distribution in (4). Note that this uses the "other percentile" principle.

Another alternative is to use the bias-adjusted bootstrap (Kilian (1998)). This involves the following steps:
1. Estimate the VAR.
2. Resample from the residuals with replacement.
3. Construct new artificial datasets.
4. In each, estimate the coefficient vector α. Repeat (2)-(4) 1,000 times.
5. The OLS estimate of α can now be bias-adjusted. Resample from the bias-adjusted VAR estimate residuals with replacement.
6. Construct new artificial datasets.
7. In each, add the bias adjustment to the estimate of α and then compute the coefficient of impulse. Repeat (5)-(7) 1,000 times.
8. Form a confidence interval from the percentiles of the distribution in (7).

The bootstrap-based bias-adjustment of Kilian (1998) is useful, whether the objective is inference for impulse responses or forecasting. An alternative is that Nicholls and Pope (1989) gives an analytical expression for the bias in a VAR (1) with an intercept. This expression is

$$E(\hat{\mathrm{A}}) - A = \Sigma\{(I_n - A)^{-1} + A(I_n - A^2)^{-1} + \Sigma_{i=1}^{n}\lambda_i(I_n - \lambda_i \mathrm{A})^{-1}\}\Gamma(0) + O(T^{-3/2})$$

where $\lambda_1, \lambda_2 .... \lambda_n$ denote the eigenvalues of A and $\Gamma(0)$ is the variance-covariance matrix of Y_t, which solves the equation $\Gamma(0) = A\Gamma(0)A' + \Sigma$ which implies that

$$vec(\Gamma(0)) = (I_{n^2} - (A \otimes A))^{-1} vec(\Sigma)$$

The advantage of the bootstrap is its general applicability.

Yet another alternative is to form a Bayesian interval (Sims and Zha (1999)), perhaps interpreted as a frequentist interval. Of course, the interval depends on the choice of prior.

Bayesian Treatment of VARs: Non-informative prior
Write the model in the form $Y = XA + u$. The "noninformative" prior for α and Σ is proportional to $|\Sigma|^{-(n+1)/2}$. With this prior:

- The posterior for α given Σ is $N(\hat{\alpha}, \Sigma \otimes (X'X)^{-1})$.
- The posterior for Σ is inverse-Wishart with parameters $e'e$ and $T - np$, where e denotes the matrix of VAR residuals.

Sidenote: Wishart and inverse-Wishart distributions.
Suppose that $X_1, X_2, ...X_n$ are iid with mean 0 and variance covariance matrix Ω. Then $\Sigma_{i=1}^{n} X_i X_i'$ is $W(\Omega, n)$. It is a multivariate analog of a χ^2.
We say that A is inverse-Wishart(Ω, n) if A^{-1} is $W(\Omega^{-1}, n)$.
Note that if A is $IW(\Omega, n)$ and Ω is pxp then $E(A) = \frac{1}{n-p-1}\Omega$.

So the recipe is one first takes draws of Σ, then draws of α and then one can build up a posterior distribution for any object of interest (e.g. impulse response). The percentiles of this distribution give a Bayesian confidence interval.

It turns out that the delta method often has confidence intervals for impulse responses that are biased and too short. The ordinary bootstrap does not do much better. The Bayesian methods and bias-adjusted bootstrap do not fully control coverage, but get effective coverage that is in many cases reasonably close to the nominal level.

Bayesian Treatment of VARs: Normal-Inverse Wishart Prior
In the last section, we discussed a non-informative prior. But informative priors may be useful in forecasting. In forecasting with VARs, a challenge is that the number of parameters can be quite large, and informative priors may be helpful. Consider the VAR

$$y_t = k + A_1 y_{t-1} + A_2 y_{t-2} ... + A_n y_{t-n} + u_t \tag{4}$$

where y_t is an px1 vector and u_t is i.i.d. $N(0, \Sigma)$. Write the model in the form $Y = XA + u$ and consider the prior:

$$\begin{aligned} vec(A) \,|\, \Sigma &\sim N(vec(A_0), \Sigma \otimes N_0^{-1}) \\ \Sigma &\sim IW(S_0, v_0) \end{aligned}$$

where A_0 , N_0, v_0 and S_0 are all prior hyperparameters. Then the posterior is given by:

$$\begin{aligned} vec(A) \,|\, \Sigma &\sim N(vec(A_P), \Sigma \otimes N_P^{-1}) \\ \Sigma &\sim IW(S_P, v_P) \end{aligned} \tag{5}$$

where

$$\begin{aligned} v_P &= T + v_0 \\ N_P &= N_0 + X'X \\ A_P &= N_P^{-1} N_0 A_0 + N_P^{-1} X'Y \\ S_P &= S_0 + (Y - X\hat{A})'(Y - X\hat{A}) + \hat{A}X'X\hat{A} + A_0' N_0 A_0 - A' N_P A \end{aligned}$$

Simulating from this posterior involves Gibbs sampling.

Uhlig (2005) and others consider a Bayesian approach to imposing sign restrictions. Any matrix R that satisfies $\Sigma = RR'$ can be written as $R = PU$ where the Cholesky factorization of Σ is $\Sigma = PP'$ and U is an orthonormal matrix. The standard prior for U is uniform on the space of orthonormal matrices (and independent of the other priors). Rubio-Ramirez, Waggoner and Zha

(2010) propose a beautifully simple algorithm for this. Let X be a matrix of independent standard normal random numbers, and let $X = QR$ be its QR decomposition (this is a decomposition into an orthonormal matrix Q and an upper triangular matrix R: qr function in Matlab). Normalize the diagonal elements of R to be positive. Now Q is uniformly distributed on the space of orthonormal matrices.
So the algorithm for imposing sign restrictions goes as follows:

1. Take a draw from the posterior, iterating between these two conditional densities, in (5) discarding the first 1,000 or so burn-in draws. This is a simple case of Gibbs sampling.

2. Work out the Cholesky factor of Σ and take a random orthonormal matrix. Compute R and see if the sign restrictions are satisfied.

3. If the sign restrictions are satisfied, combine the chosen R with the reduced form VAR parameters in (1) and work out impulse responses etc. If not, repeat step 2.

4. Repeat 1-3 to get draws from the posterior.

Baumeister and Hamilton (2015) point out that this procedure brings in implicit prior information over and above the sign restrictions.

Minnesota Prior
The Minnesota prior of Doan, Litterman and Sims (1984) is an informative prior that has been found by many to be helpful for forecasting purposes. Consider the VAR

$$y_t = k + A_1 y_{t-1} + A_2 y_{t-2} ... + A_n y_{t-n} + u_t \tag{6}$$

where y_t is an px1 vector and u_t is i.i.d. $N(0, \Sigma)$.
In the Minnesota prior, the priors for k, $A = [A_1 \; A_2 .... A_p]'$ and Σ are mutually independent and

$$p(k) \sim N(0, \kappa I_p) \tag{7}$$

$$p(vec(A)) \sim N(0, \Omega_A) \tag{8}$$

and

$$p(\Sigma) \propto |\Sigma|^{-(p+1)/2} \tag{9}$$

where κ is a large number, Ω_A is a diagonal matrix, the prior variance for the ij th element of A_k is $\frac{\lambda^2}{k^2}\frac{\sigma_i^2}{\sigma_j^2}$ where λ is a hyperparameter that measures the overall tightness of the prior and σ_i^2 is the residual variance from fitting an AR(1) to y_{it}. Note that the prior proposed by Doan, Litterman and Sims differed slightly from this in that they set the prior mean of the diagonal elements of A_1 to 1. In other words, they were shrinking towards a random walk prior: the variant here instead follows Banbura, Giannone and Reichlin (2009) in instead shrinking towards a white noise prior.

The Gibbs sampler can then be used to take draws from the posterior of the parameters. Specifically, the posterior of $vec(A)$ conditional on Σ is

$$N((\Omega_A^{-1} + \Sigma^{-1} \otimes X'X)^{-1} vec(X'Y\Sigma^{-1}), (\Omega_A^{-1} + \Sigma^{-1} \otimes X'X)^{-1})$$

while the posterior for Σ conditional on A is

$$IW((Y - XA)'(Y - XA), T)$$

where $IW(.,.)$ denotes the inverse-Wishart distribution.

Priors directly on the structural impulse responses

Generally researchers impose priors on the VAR parameters, but economists have more ideas about structural impulse responses. Plagborg-Moller (2016) specifies priors for the $D(L)$ parameters and then works out the posterior. With a given $D(L)$, the likelihood is that of a vector moving average and the posterior is then the prior times the likelihood, though efficient numerical implementation is complicated. The choice of R is not explicitly required in this procedure. The likelihood will have multiple peaks, but the posterior should be single-peaked.

Local Projections

Local projections is an idea proposed by Jorda (2005). Instead of estimating a VAR, estimate the univariate regression:

$$y_{t+s} = \beta_s s_t + \gamma_1 y_{t-1} ... + \gamma_n y_{t-n} + u_t$$

where s_t is a shock for different values of s. Note that the errors are bound to be serially correlated. The impulse response is different for each s and this is the sense in which it is local. It can be applied in many contexts. Like panel data:

$$y_{i,t+s} = \beta_s s_{i,t} + \alpha_i + \lambda_t + \gamma_1 y_{i,t-1} ... + \gamma_n y_{i,t-n} + u_{i,t}$$

Or adding interaction effects:

$$y_{t+s} = \beta_s s_t + \omega_x x_t + \lambda_s s_t x_t + \gamma_1 y_{t-1} ... + \gamma_n y_{t-n} + u_t$$

And Barnichon and Brownlees (2018) have a proposal for smoothing the impulse responses. In population, local projections and a VAR are equivalent (Plagborg-Moller and Wolf, 2021), but of course in the finite samples they are not.

A Cholesky identified VAR can be estimated by local projections. Suppose that we have a Cholesky ordering and we are interested in the effect of structural shock i and we let Y_t^* denote the vector of all elements of Y_t ordered above i. Then we consider the regression

$$Y_{t+s} = \beta_s Y_{t,i} + \gamma_0 Y_t^* + \gamma_1 Y_{t-1} ... + \gamma_n Y_{t-n} + u_t$$

If shock i is ordered first then Y_t^* is empty and so that term is simply omitted, but otherwise it is present.

Econ 607 Jonathan Wright

Fall 2023

Handout on Panel VARs

Suppose that Y_{it} is an mx1 vector of variables for country i in time period t. There are n countries and T time periods. The cross-sectional units do not have to be countries, but they most often are. A panel VAR has the potential to give big efficiency gains. Canova and Ciccarelli (2013) give a nice survey of panel VAR methods, mainly from the Bayesian perspective.

We can consider a panel VAR of the form:

$$Y_{it} = \alpha_i + A_{1i}Y_{it-1} + A_{2i}Y_{it-2} ... + A_{pi}Y_{it-p} + u_{it}$$

where u_{it} are reduced form shocks. The coefficients are allowed to vary across the cross-sectional units. Without any further restrictions, this amounts to something like a separate panel for each country. It's not quite because it would actually be a seemingly unrelated regression system. You can estimate it by OLS equation by equation, but feasible GLS is asymptotically more efficient if the errors are correlated across countries. How this works is that we think of each equation as being a regression of the form $Y_i = X_i\beta_i + u_i$ and we then stack the equations to yield:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

We then estimate this twice: first by OLS, and then by feasible GLS using the estimated residuals to estimate a covariance matrix of the errors of the form $\Sigma \otimes I_T$.

One approach is to assume that the slope coefficients are the same for all countries (homogenous). This can be done by a pooled approach with fixed effects, stacking nT observations. This is easy. One problem is that it is inconsistent if T is small. Henceforth let's simplify everything to one lag.

Anderson and Hsiao (1982) propose an IV approach. This takes the equation differencing out the fixed effects:

$$Y_{it} - Y_{it-1} = A_1(Y_{it-1} - Y_{it-2}) + (u_{it} - u_{it-1})$$

and then estimating this by IV using Y_{it-2} as an instrument. Arellano and Bond (1991) propose a GMM approach. This is based on the moment conditions:

$$E([(Y_{it} - Y_{it-1}) - A_1(Y_{it-1} - Y_{it-2})]Y_{it-j}) = 0,\ t = 3,...T;\ j = 1,...t-2$$

Notice that this is not a standard IV estimator any more because there is an instrument for each time period and the number of instruments differs by time period. For t=3, there is 1 instrument, for t=2 there are 2

instruments and so on. Where we have just one lag, the minimum number of time periods T is 3, and this gives just identification. More lags would give overidentifying restrictions. With two lags, the minimum number of time periods T would be 5.

A related alternative, proposed by Arellano and Bover (1995) uses a forward orthogonal transformation instead. Define

$$Y_{it}^{*} = \sqrt{\frac{T-t}{T-t+1}}(Y_{it} - \frac{Y_{i,t+1} + Y_{i,t+2}....Y_{i,T}}{T-t}), t = 2...T-1$$

Then this transformation lets us wash out the fixed effects as:

$$Y_{it}^{*} = A_1 Y_{it-1}^{*} + u_{it}^{*}$$

And then we can do GMM based on the moment conditions:

$$E([Y_{it}^{*} - A_1 Y_{it}^{*}]Y_{it-j}) = 0,\ t = 2,...T;\ j = 1,...t-1$$

Now any variable dated $t-1$ is available as an instrument, so you get more instruments. In either of these approaches, you don't necessarily want to use all of the instruments that are available (Roodman 2009) in finite samples.

Yet another approach is to stick with pooled least squares estimation with a dummy variable. Kiviet (1995) has a formula for the bias in that case, and you can correct for the bias. In some simulations, Judson and Owen (1999) find that the best results are obtained with this bias correction, with the Anderson-Hsiao IV estimator coming in second.

Uribe and Yue (2006) is an example of a panel VAR with common slope coefficients and fixed effects. The variables are output, investment, trade balance and real interest rates in emerging markets and the US real interest rate. The US real interest rate is specified as a univariate autoegression. The main objective is to study the effects of US and local real interest rate shocks. These are identified by a Cholesky identification. Each emerging market is considered on its own.

Another example of a panel VAR with common slope coefficients and fixed effects is Love and Zicchino (2006). This considers a panel VAR with about 8000 firms from 36 countries. They consider firm level variables like sales-to-capital ratios, investment-to-capital ratio and cash flow and consider a panel VAR of the form:

$$Y_{it} = \alpha_i + d_{ct} + A_1 Y_{it-1} + \varepsilon_{it}$$

where the i subscript indexes firms. So there are constant parameters, but we add in country-time fixed effects. Structural impulse response anaylsis is conducted using a Cholesky identification.

Often researchers might want some structure, but not to assume that the slope coefficients are the same for all countries. Pesaran and Smith (1995) assume that:

$$A_{1i} = A + \eta_i$$

We can then write the VAR in the form:

$$Y_{it} = \alpha_i + AY_{it-1} + \eta_i Y_{it-1} + u_{it}$$

which can then be estimated by pooled GLS, but giving only estimates of the fixed effects and the matrix A. Another alternative (Pesaran and Smith (1995)) is the mean group estimator. This estimates the VAR for each country separately, or by the seemingly unrelated regression estimator, and then simply averages the estimators. Or, if you are interested in the impulse responses, you could estimate the impulse response for each country and then average these. Either way, let $\hat{\theta}_i$ be the estimated object of interest for country i. The mean group estimator is $\hat{\theta}_{MG} = n^{-1}\Sigma\hat{\theta}_i$ and the generally used estimate of the variance is $\frac{1}{n(n-1)}\Sigma_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_{MG})(\hat{\theta}_i - \hat{\theta}_{MG})'$.

An example is Gambacorta, Hoffmann and Peersman (2014). They consider a VAR for 8 countries in four variables: monthly output, prices, the VIX and the size of central bank balance sheets. They estimate this VAR by SURE for each country separately. For the identification of the SVAR, they combine sign identification with zero restrictions. Concretely, it is assumed that for each country P_i is the Cholesky factor of the reduced form variance-covariance matrix and then that the reduced form errors are P_iQ times the structural errors. The matrix Q is specified as:

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}$$

where θ is in $[0, 2\pi]$. If the impact impulse response of the monetary policy shock meets the sign criteria (the VIX does not go up and the balance sheet does go up) for all 8 countries, then that draw is accepted, otherwise it is rejected. Then once they have the impulse responses, they average these across the 8 countries. Finally they take the percentiles of these impulse responses. In this exercise, the researcher considers a monetary policy shock in each country without taking a stand on how the monetary policy shocks are correlated across countries.

The setup so far rules out the idea of dynamic interdependencies (lags of one country affecting another). Going to the other extreme, we could define $Y_t = (Y_{1t}', Y_{2t}', ...Y_{nt}')'$ and then write a huge VAR of the form:

$$Y_t = AY_{t-1} + u_t$$

This has an enormous number of parameters. If there are p lags, there would be pm^2n^2 parameters. So something has to be done to reduce the dimension of the parameters. One can impose a lot of zeros, which may get us back to something like we had without interdependencies. Or, if we let α denote the vector of parameters, we might impose a factor structure on the parameters (Canova and Ciccarelli (2004, 2009)).

Pesaran et al. (2004) propose a solution to the dimensionality problem when modeling country interdependencies that he terms a GVAR (Global VAR). The setup is that there are N+1 countries, with country 0 being the US. For each country, the vector of variables is Y_{it} and from the perspective of country i the average for the rest of the world is Y_{it}^*. We then specify:

$$Y_{it} = A_1Y_{it-1} + A_2Y_{it}^* + A_3Y_{it-1}^* + u_{it}$$

The rest of the world variable is defined as $Y_{it}^{*} = \Sigma_{j=0}^{n} \omega_{ij} Y_{it}$ and the weights could be trade weights, but the point is that they are not being estimated. The weights must sum to one and $\omega_{ii} = 0$. This model can be estimated by OLS for each country separately, or it could be estimated as a system, but it is generally estimated by OLS for each country separately. The shocks are assumed to be iid. Stacking the country variables as $Y_t = (Y_{0t}', Y_{1t}', ...Y_{nt}')'$ we can again write this VAR in the form:

$$Y_t = AY_{t-1} + u_t$$

But this time there are far fewer parameters to estimate. The estimation is done purely at the country level; stacking into a single equation is for the purposes of looking at effects of shocks.

There will be the question though of identification restrictions for a structural VAR. Once you allow dynamic interdependencies, having a Cholesky ordering becomes very difficult. Dees et al. (2007) have a global VAR where they order the US variables first and then have an ordering within the US variables, and then in this way get a Cholesky identification of a structural monetary policy shock. But most of the time when researchers are using a GVAR for impulse response analysis, they give up on trying to identify structural shocks and instead use generalized impulse response function.

As to standard errors, there are a number of ways of proceeding. Writing the VAR in the form $Y_t = AY_{t-1} + u_t$ immediately allows us to construct a bootstrap, drawing from u_t by resampling with replacement.

Handout on Chow-Lin Interpolation

A very common practical problem in empirical macroeconomics is that we have data at low frequency, but would like to estimate that time series at a higher frequency. For example, GDP is produced only once a quarter, but we would like to have monthly estimates of GDP. A reasonable way of solving this problem of interpolating high-frequency data entails looking at series that are observed at higher frequency and using the relationship between these series and the series that we observe at low frequency. For example, there are monthly data on consumption, industrial production, retail sales, employment etc that are strongly positively associated with GDP. We might use these to infer what monthly GDP would have been. The intuition is that in a quarter where these monthly indicators were falling over the quarter, we would want to assign a disproportionate share of GDP to the first month of the quarter.

A classic paper by Chow and Lin (1971) proposed a method for doing this. In fact, under the assumptions that they make, it gives the best linear unbiased estimate of the high-frequency data. Without loss of generality, I will assume that the problem is to interpolate a quarterly time series to the monthly frequency. But the same approach could be used to interpolate an annual time series to the quarterly frequency and so on.

Suppose that $y_t^{(Q)}$ is the quarterly time series, so that $y_t^{(Q)} = y_{t,1} + y_{t,2} + y_{t,3}$ where $y_{t,1}$ denotes the series in the first month of the quarter and so on. In month i of quarter t we observe other variables and assume that

$$y_{t,i} = \beta_1 x_{1,t,i} + \beta_2 x_{2,t,i} ... + \beta_p x_{p,t,i} + u_{t,i} \tag{1}$$

and that

$$u_{t,i} = aLu_{t,i} + \varepsilon_{t,i}$$

where L denotes the monthly lag operator, which could be in the previous quarter and $\varepsilon_{t,i}$ is iid with mean zero and variance σ^2 Accordingly, the 3Tx3T variance-covariance matrix of monthly errors in (1) is

$$V = \begin{pmatrix} 1 & a & a^2 & .. & a^{3T-1} \\ a & 1 & a^2 & .. & a^{3T-2} \\ & & & & \\ a^{3T-1} & a^{3T-2} & a^{3T-3} & .. & 1 \end{pmatrix} \tag{2}$$

If $y^{(Q)}$ and $y^{(M)}$ denote the vectors of quarterly and monthly observations respectively, of lengths T and $3T$, then $y^{(Q)} = Cy^{(M)}$ where

$$C = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & .. & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & .. & 0 & 0 & 0 \\ & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & .. & 1 & 1 & 1 \end{pmatrix}$$

Let $X^{(M)}$ be the 3Txp matrix of the monthly variables and define $X^{(Q)} = CX^{(M)}$. We can write equation (1) in the form $y^{(M)} = X^{(M)}\beta + u$ and then premultiply this by C to get

$$y^{(Q)} = X^{(Q)}\beta + u^{(Q)} \tag{3}$$

where $u^{(Q)} = Cu$. In this regression the variance-covariance matrix of the errors is CVC' and multiplying this out, the first autocorrelation of the errors can be calculated as

$$\frac{a^5 + 2a^4 + 3a^3 + 2a^2 + a}{3 + 2a^2 + 4a} \tag{4}$$

With all this, here are the steps in the Chow-Lin interpolation procedure:

1. Construct $X^{(Q)} = CX^{(M)}$.
2. Run the regression in (3) by OLS, giving the estimator
$$\hat{\beta}_{OLS} = (X^{(Q)\prime} X^{(Q)})^{-1} X^{(Q)\prime} y^{(Q)}$$
3. Calculate the first order autocorrelation of the residuals from this OLS regression.
4. Find the value of a that sets the value of (4) to this autocorrelation (use the Matlab function fsolve). Call this $\hat{a}$. Plug this into (2) to get an estimate of V, $\hat{V}$.
5. Now run the regression in (3) by feasible GLS, giving the estimator
$$\hat{\beta}_{FGLS} = (X^{(Q)\prime}(C\hat{V}C')^{-1} X^{(Q)})^{-1} X^{(Q)}(C\hat{V}C')^{-1} y^{(Q)}$$
6. Obtain the corresponding residuals, $\hat{u}^{(Q)}_{FGLS}$.
7. Chow and Lin show that the best linear unbiased estimator of the monthly data $y^{(M)}$ is
$$\hat{y}^{(M)} = X^{(M)}\hat{\beta} + \hat{V}C'(C\hat{V}C')^{-1}\hat{u}^{(Q)}_{FGLS}$$

The Kalman filter provides an alternative (rather similar) way of interpolating monthly data.

Econ 607 Jonathan Wright
Fall 2023

Handout on Filtering

The filtering problem is that there is unobserved variable (a "state" variable) that evolves by some law of motion and there are observed variables that are related to the unobserved state. It might sound an arcane problem, but as we'll see in a bit, it has an enormous number of important macroeconomic applications.

The basic filtering problem is a linear model in what is known as state space form. This is the model where we observe y_t while α_t is an unobserved state and

$y_t = Z_t\alpha_t + \varepsilon_t$ (the *measurement* equation)
$\alpha_t = T\alpha_{t-1} + \eta_t$ (the *transition* equation)
where $\varepsilon_t \sim iidN(0,H)$ and $\eta_t \sim iidN(0,Q)$

The Kalman Filter
The Kalman Filter allows inference to be done in the basic state space model. In this model $\alpha_t \mid Y_s$ is normal; let $\alpha_{t|s}$ and $P_{t|s}$ denote its mean and variance. We then have

Updating Equations
$\alpha_{t|t} = \alpha_{t|t-1} + P_{t|t-1}Z_t'F_t^{-1}(y_t - Z_t\alpha_{t|t-1})$
$P_{t|t} = P_{t|t-1} - P_{t|t-1}Z_t'F_t^{-1}Z_tP_{t|t-1}$
where
$F_t = Z_tP_{t|t-1}Z_t' + H$

Prediction Equations
$\alpha_{t+1|t} = T\alpha_{t|t}$
$P_{t+1|t} = TP_{t|t}T' + Q$

If α_t is stationary, we can initialize from the unconditional mean and variance-covariance matrix of α_t. The mean is just zero. The variance is $P_{t|t}$ which is the solution to the equation:
$P_{0|0} = TP_{0|0}T' + Q$
the solution to which is
$vec(P_{0|0}) = (I - T \otimes T)^{-1}vec(Q)$

So, in Matlab, $P_{0|0}$ is simply
reshape(inv(eye(n^2)-kron(T,T))*reshape(Q,n^2,1),n,n)
where n is the number of elements in the state vector α_t.

Or we can treat the mean and variance of α_0 as unknown parameters. Or, especially if the state vector is nonstationary, we can set $\alpha_{0|0}$ to zero and $P_{0|0}$ to κI where κ is a large number, constituting a "diffuse prior."

Whatever the initialization, we then iterate through the updating and predictive equations to get $\alpha_{t|t-1}$ and $\alpha_{t|t}$.The Kalman filter has two potential purposes: (i) estimation and (ii) inference about the state vector. For the first of these, we have the log-likelihood:

$$l = \Sigma_{t=1}^{T} \log(f(y_t \mid Y_{t-1}))$$
$$y_t \mid Y_{t-1} \sim N(Z_t\alpha_{t|t-1}, Z_t P_{t|t-1} Z_t' + H) = N(Z_t\alpha_{t|t-1}, F_t)$$
$$\therefore l = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\Sigma_{t=1}^{T}\log|F_t| - \frac{1}{2}\Sigma_{t=1}^{T} v_t' F_t^{-1} v_t$$

where $v_t = y_t - Z_t\alpha_{t|t-1}$

The updating equation for the mean of the state can be written as

$$\alpha_{t|t} = \alpha_{t|t-1} + K_t(y_t - y_{t|t-1})$$

where $K_t = P_{t|t-1}Z_t'F_t^{-1}$ is the Kalman gain.

For inference about the state vector, we already have $\alpha_{t|t}$, the "filtered" estimates. But we might want $\alpha_{t|T}$, the "smoothed" estimates. These are obtained with one more set of recursions known as the Kalman smoother:

$$\alpha_{t|T} = \alpha_{t|t} + J_t(\alpha_{t+1|T} - \alpha_{t+1|t})$$
$$P_{t|T} = P_{t|t} + J_t(P_{t+1|T} - P_{t+1|t})J_t'$$

where

$$J_t = P_{t|t}T'P_{t+1|t}^{-1}$$

The Kalman filter/smoother needs values of the parameters. We can use numerical methods to find the parameter values that maximize the likelihood, and then plug these in to get filtered and smoothed estimates of the states, which are also of interest.

In some cases, it will turn out that matrices are singular. The formulas will still go through, but replacing the inverse by the Moore-Penrose generalized inverse.

In Bayesian work, we often want to take a draw from the distribution of the state vector conditional on the full dataset and parameters. This is accomplished by the device of Carter and Kohn (1994) that works as follows:

(a) Running the Kalman filter to get $\theta_{t|t}$, $P_{t|t}$ and $P_{t+1|t}$ for all t.

(b) Take a draw for θ_t from the $N(\theta_{T|T}, P_{T|T})$ distribution.

(c) Cycle back taking draws

$$\theta_t \sim N(\theta_{t|t} + P_{t|t}T'P_{t+1|t}^{-1}(\theta_{t+1} - T\theta_{t|t}), P_{t|t} - P_{t|t}T'P_{t+1|t}^{-1}TP_{t|t})$$

which amounts to taking draws from the distribution given by the Kalman smoother.

The EM Algorithm

This idea of using the Kalman filter to obtain the log-likelihood and then maximize it numerically is easier said than done—at least if the number of parameters is large. The EM algorithm is a device for maximizing the log-likelihood function obtained from the Kalman filter. This makes it easy to handle models with a relatively large number of parameters, and it is indeed an essential tool for many modern Kalman filter applications.

The idea of the EM algorithm is to obtain the smoothed estimates of the state vector and then to treat these as observed data and maximize the likelihood function. There are closed form expressions for these parameter estimates. Essentially it boils down to OLS regressions with the state vector as observable, except with an adjustment for the fact that there is measurement error in the state variable. With these parameter estimates, the Kalman smoother is run again, and the algorithm goes back and forth. At each iteration, the log-likelihood function should increase. The Kalman smoother is the "E" step (E for Expectation); the maximization is the "M" step. The algorithm converges once the change is small.

Given the smoothed estimates, here are the expressions for the parameter estimates in the Kalman filter in the "M" step.

$$\hat{T} = \Sigma_{t=1}^{T}\{\alpha_{t|T}\alpha_{t-1|T}' + P_{t,t-1|T}\}[\Sigma_{t=1}^{T}\{\alpha_{t-1|T}\alpha_{t-1|T}' + P_{t-1|T}\}]^{-1}$$

$$\hat{H} = \frac{1}{T}\Sigma_{t=1}^{T}(y_t - Z_t\alpha_{t|T})'(y_t - Z_t\alpha_{t|T}) + \frac{1}{T}\Sigma_{t=1}^{T}Z_tP_{t|T}Z_t'$$

$$\hat{Q} = \frac{1}{T}\Sigma_{t=1}^{T}\eta_t\eta_t' + \frac{1}{T}\Sigma_{t=1}^{T}[I \ -T]\begin{pmatrix} P_{t|T} & P_{t,t-1|T} \\ P_{t,t-1|T}' & P_{t-1|T} \end{pmatrix}[I \ -T]'$$

where $\eta_t = \alpha_{t|T} - T\alpha_{t-1|T}$ and $P_{t,t-1|T} = E[(\alpha_t - \alpha_{t|T})(\alpha_{t-1} - \alpha_{t-1|T})']$. Computing $P_{t,t-1|T}$ is the only tricky bit. This involves yet another set of recursions, sometimes known as the *lag-one covariance smoother*. Like the ordinary smoother, it starts at the end with

$$P_{T,T-1|T} = (I - P_{T-1|T}Z_T'F_T^{-1})TP_{T-1|T-1}$$

and then cycles backward using

$$P_{t,t-1|T} = P_{t|t}J_{t-1}' + J_t(P_{t+1,t|T} - TP_{t|t})J_{t-1}'$$

Another thing to note is that these formulas require $\alpha_{0|T}$ and $P_{1,0|T}$ but these can be got by continuing the smoothing recursions back from 1 more period.

In many applications, Z_t is a parameter, G. The EM algorithm also gives the MLE of this. It is

$$\hat{G} = \Sigma_{t=1}^{T}y_t\alpha_{t|T}'[\Sigma_{t=1}^{T}\{\alpha_{t|T}\alpha_{t|T}' + P_{t|T}\}]^{-1}$$

The initial condition is a tricky thing in the EM algorithm. A standard device is to the mean of the α_0 to $\alpha_{0|T}$ (which can be obtained by continuing the smoother for one more iteration) while holding the variance of α_0 fixed at some reasonable value.

Applications of the Kalman Filter

1. Time-varying parameter models

$y_t = \beta_t' x_t + \varepsilon_t$

$\beta_t = \beta_{t-1} + \eta_t$

where $(\varepsilon_t, \eta_t')'$ is iid $N(0, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \Sigma_\eta \end{pmatrix}$ and the x_t s are strictly exogenous. This is immediately in state space form.

2. Random-walk-plus noise model.

$y_t = \mu_t + \varepsilon_t$

$\mu_t = \mu_{t-1} + \eta_t$

where $(\varepsilon_t, \eta_t')'$ is iid $N(0, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \Sigma_\eta \end{pmatrix}$. This is useful for a permanent-transitory decomposition. Again, it is immediately a model in state space form.

3. Mixed frequency interpolation. Earlier, we discussed Chow-Lin interpolation. The Kalman filter gives an alternative. Consider the problem where there is a variable that is observed (say) monthly and another that is observed only quarterly that is an aggregate of an unobserved monthly series. Let the observed monthly series be q_t and let the observed quarterly series be $w_t^{(Q)} = w_t + w_{t-1} + w_{t-2}$. Suppose that the state vector is

$$\alpha_t = (q_t, w_t, w_{t-1}, w_{t-2})'$$

We assume that q_t and w_t follow a VAR (1), giving the transition equation. In the last month of each quarter, the measurement equation is

$$\begin{pmatrix} q_t \\ w_t^{(Q)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \alpha_t$$

while in all other months, the measurement equation is

$$q_t = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \alpha_t$$

4. A "factor" model. Suppose that c_t is the unobservable "state of the economy." But we do observe an nx1 vector of indicators y_t and assume that

$y_t = \gamma f_t + u_t$

$\phi(L) f_t = \eta_t$

$D(L) u_t = \varepsilon_t$

and η_t and the elements of ε_t are mutually uncorrelated iid and normal with mean zero. Without loss of generality, let's suppose $\phi(L)$ is an AR(2) and $D(L)$ is a VAR(2). Then the state vector is

$\alpha_t = (f_t, f_{t-1}, u_t', u_{t-1}')'$.

The transition equation is

$$\alpha_t = \begin{pmatrix} \phi_1 & \phi_2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & D_1 & D_2 \\ 0 & 0 & I & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix}$$

and the measurement equation is

$$y_t = \begin{pmatrix} \gamma & 0 & I & 0 \end{pmatrix} \alpha_t$$

This is one way of getting the common component from a bunch of series. An alternative is the *principal components* estimator. Say that we have T observations and y_t is nx1. Then we can solve the problem:

$$(\hat{f}_t, \hat{\lambda}) = \arg\min_{f_t, \lambda_i} \Sigma_{i=1}^{n} \Sigma_{t=1}^{T} (y_{it} - \lambda_i' f_t)^2$$

and we say that f_t are the principal components and λ_i is the loading of the ith variable. Let F be the Txr matrix of principal components: $(f_1 \ f_2 .... f_T)'$, let Λ be the nxr matrix of loadings: $(\lambda_1 \ \lambda_2 .... \lambda_n)'$ and let Y be the Txn data matrix $(y_1 \ y_2 .... y_T)'$. We can solve this problem subject to a normalization that $\frac{1}{T} F'F = I$ and $\Lambda'\Lambda$ is diagonal. The solution is that $\hat{F}$ is $\sqrt{T}$ times the matrix of eigenvectors associated with the r largest eigenvalues of $\frac{YY'}{nT}$. And $\hat{\Lambda} = \frac{Y'\hat{F}}{T}$.

5. Aruoba, Diebold and Scotti (2009) propose a measure of the state of the economy based on mixed frequency data. The idea is that there is a latent daily variable that is the state of the economy that follows (for simplicity) an AR(1):

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} ... + \phi_p x_{t-p} + u_t$$

and then there are observations of weekly, monthly and quarterly data (including initial claims, industrial production and real GDP growth). Define the "cumulator" variables

$$C_{Weekly}^t = \zeta_t C_{Weekly}^{t-1} + x_t$$

where $\zeta_t = 1$ on the first day of a week and zero otherwise and let $C_{monthly}^t$ and $C_{quarterly}^t$ be defined similarly. Let the state vector be $\alpha_t = (x_t, x_{t-1}, ... x_{t-p+1}, C_{weekly}^t, C_{monthly}^t, C_{quarterly}^t)'$. The measurement equation relates each of the observed series to the state vector. For example, initial claims will be $\beta_{0,claims} + \beta_{1,claims} C_{weekly}^t + \text{error}$.

Drifting Coefficients VAR

This is an example of a Bayesian model that uses filtering methods and that is useful in macroeconomics (Cogley and Sargent (2001)).

The setup is a VAR which can be written in the form

$$y_t = (I \otimes x_t')\theta_t + \varepsilon_t$$

where $\theta_{t+1} = \theta_t + v_t$ and (ε_t', v_t') is iid $N(0,V)$.

Start out by specifying that the prior for θ_0 is $N(\bar{\theta}, \bar{P})$ and the prior for V is $IW(\bar{V}^{-1}, \bar{T})$.

We can work out the posterior of $\theta = \{\theta_t\}_{t=1}^T$ and V by Gibbs sampling.

1. Take a draw from the distribution of θ conditional on V and Y (the data). This is accomplished by the device of Carter and Kohn (1994) that works as follows in the current context:

(a) Running the Kalman filter to get $\theta_{t|t}$, $P_{t|t}$ and $P_{t+1|t}$ for all t.

(b) Take a draw for θ_t from the $N(\theta_{T|T}, P_{T|T})$ distribution.

(c) Cycle back taking draws

$$\theta_t \sim N(\theta_{t|t} + P_{t|t}P_{t+1|t}^{-1}(\theta_{t+1} - \theta_{t|t}), P_{t|t} - P_{t|t}P_{t+1|t}^{-1}P_{t|t})$$

2. Take a draw from the distribution of V conditional on θ and Y. Let $\hat{V}$ denote the residual variance-covariance matrix. This is given by a draw from a $IW([\bar{V} + T\hat{V}]^{-1}, \bar{T} + T)$ distribution.

3. Repeat 1 and 2 many times to build up the posterior distribution, tossing away the first set of draws as "burn-in", as is usual with the Gibbs sampler.

In this application there is one further wrinkle. This algorithm will generate draws of θ_t that are explosive. Cogley and Sargent add in a further prior that the roots of θ_t are stationary. This is easily imposed. For each draw in step 1, check that the entire series $\{\theta_t\}_{t=1}^T$ is stationary. If it is not, simply repeat 1(b) and 1(c) until it is.

Extended Kalman Filter

This is one approach to approximately estimating a nonlinear filtering model. The basic setup is as follows:

$$y_t = f(\alpha_t) + v_t$$
$$\alpha_t = h(\alpha_{t-1}) + w_t$$

where v_t is iid N(0,H) and w_t is iid N(0,Q). The state vector α_t is Lx1. This uses the following recursions:

$$\alpha_{t|t} = \alpha_{t|t-1} + P_{t|t-1}\frac{df(\alpha_{t|t-1})'}{d\alpha_t}F_t^{-1}(y_t - f(\alpha_{t|t-1}))$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}\frac{df(\alpha_{t|t-1})'}{d\alpha_t}F_t^{-1}\frac{df(\alpha_{t|t-1})}{d\alpha_t}P_{t|t-1}$$

where

$$F_t = \frac{df(\alpha_{t|t-1})}{d\alpha_t}P_{t|t-1}\frac{df(\alpha_{t|t-1})'}{d\alpha_t} + H$$

Prediction Equations

$$\alpha_{t+1|t} = h(\alpha_{t|t})$$

$$P_{t+1|t} = \frac{dh(\alpha_{t|t})}{d\alpha_t}P_{t|t}\frac{dh(\alpha_{t|t})'}{d\alpha_t} + Q$$

which effectively use Taylor series expansions for the variance terms. The derivatives could be computed analytically (better) or numerically. The log-likelihood is then approximated by:

$$-\frac{T}{2}\log(2\pi) - \frac{1}{2}\Sigma_{t=1}^{T}\log|F_t| - \frac{1}{2}\Sigma_{t=1}^{T}v_t'F_t^{-1}v_t$$

where $v_t = y_t - f(\alpha_{t|t-1})$. The extended smoother is worked out by backward recursions:

$$\alpha_{t|T} = \alpha_{t|t} + J_t(\alpha_{t+1|T} - \alpha_{t+1|t})$$

$$P_{t|T} = P_{t|t} + J_t(P_{t+1|T} - P_{t+1|t})J_t'$$

where $J_t = P_{t|t}\frac{dh(\alpha_{t|t})}{d\alpha_t}'P_{t+1|t}^{-1}$.

Unscented Kalman Filter

This is a powerful approach and more modern approach to estimating a nonlinear filtering model. The basic setup is the same as for the extended Kalman filter, but it doesn't rely on taking derivatives which makes it better for very nonlinear (or non-differentiable) functions. The unscented Kalman filter is a set of recursions starting from $\alpha_{0|0}$ and $P_{0|0}$ the unconditional mean and variance of the states.

Given $\alpha_{t|t}$ and $P_{t|t}$ pick "sigma points" which are meant to approximate the distribution of α_t at time t:

$$\chi_{0,t} = \alpha_{t|t}$$

$$\chi_{i,t} = \alpha_{t|t} + (\sqrt{(L+\lambda)P_{t|t}})_i, i = 1,...L$$

$$\chi_{i,t} = \alpha_{t|t} - (\sqrt{(L+\lambda)P_{t|t}})_i, i = L+1,...2L$$

with weights

$$W_0^m = \frac{\lambda}{L+\lambda},\ W_0^c = \frac{\lambda}{L+\lambda} + 1 - \alpha^2 + \beta$$

$$W_i^m = W_i^c = \frac{1}{2(L+\lambda)}, i \neq 0$$

where $(\sqrt{A})_i$ is the ith column of the lower triangular Cholesky factor of A, $\lambda = (\alpha^2 - 1)L$ and α and β are constants that might be set to 10^{-3} and 2 respectively.

We then use the recursions:

$$\alpha_{t+1|t} = \Sigma_{i=0}^{2L} W_i^m h(\chi_{i,t})$$

$$P_{t+1|t} = \Sigma_{i=0}^{2L} W_i^c (h(\chi_{i,t}) - \alpha_{t+1|t})(h(\chi_{i,t}) - \alpha_{t+1|t})' + Q$$

We then get a new set of sigma points to approximate the distribution of α_{t+1} at time t:

$$\overline{\chi}_{0,t} = \alpha_{t+1|t}$$

$$\overline{\chi}_{i,t} = \alpha_{t+1|t} + (\sqrt{(L+\lambda)P_{t+1|t}})_i, i = 1,...L$$

$$\overline{\chi}_{i,t} = \alpha_{t+1|t} - (\sqrt{(L+\lambda)P_{t+1|t}})_i, i = L+1,...2L$$

We then use the recursions:

$$\hat{y}_{t+1|t} = \Sigma_{i=0}^{2L} W_i^m f(\overline{\chi}_{i,t}) \text{ and } F_{t+1} = \Sigma_{i=0}^{2L} W_i^c [f(\overline{\chi}_{i,t}) - \hat{y}_{t+1|t}][f(\overline{\chi}_{i,t}) - \hat{y}_{t+1|t}]' + H$$

$$P_{t+1}^{\alpha y} = \Sigma_{i=0}^{2L} W_i^c [\overline{\chi}_{i,t} - \alpha_{t+1|t}][f(\overline{\chi}_{i,t}) - \hat{y}_{t+1|t}]'$$

Given that we assume normality, we then have the updating equations:

$$\alpha_{t+1|t+1} = \alpha_{t+1|t} + P_{t+1}^{\alpha y} F_{t+1}^{-1} (y_{t+1} - \hat{y}_{t+1|t})$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1}^{\alpha y} F_{t+1}^{-1} P_{t+1}^{\alpha y'}$$

And the log-likelihood is simply:

$$-\frac{T}{2}\log(2\pi) - \frac{1}{2}\Sigma_{t=1}^{T} \log|F_t| - \frac{1}{2}\Sigma_{t=1}^{T} (y_t - \hat{y}_{t|t-1})' F_t^{-1} (y_t - \hat{y}_{t|t-1}).$$

In some cases, the matrix of which you are taking the Cholesky factor to compute the sigma points will be singular or near singular. A trick for fixing this is to use the LDU decomposition. If you have a matrix A, and you take [L,D]=ldl(A) in Matlab, then this will give matrices L and D such that $A = LDL'$. Now set L*sqrt(max(D,0)). This will be a more numerically stable version of the lower Cholesky factor (a bit like the Moore-Penrose trick described above).

A neat thing that you can verify is that if you apply the unscented Kalman filter to a linear Gaussian model, then it will numerically give you exactly the same results, even though it looks quite different.

An unscented Kalman smoother is also available. When running the first set of recursions (for $\alpha_{t+1|t}$ and $P_{t+1|t}$), add in one more equation:

$$C_{t+1} = \Sigma_{i=0}^{2L} W_i^c (h(\chi_{i,t}) - \alpha_{t+1|t})(\chi_{i,t} - \alpha_{t|t})'$$

Then run the smoother backwards starting from $\alpha_{T|T}$ and $P_{T|T}$ as follows:

$$\alpha_{t|T} = \alpha_{t|t} + C_{t+1} P_{t+1|t}^{-1} (\alpha_{t+1|T} - \alpha_{t+1|t})$$

$$P_{t|T} = P_{t|t} + C_{t+1} P_{t+1|t}^{-1} (P_{t+1|T} - P_{t+1|t}) P_{t+1|t}^{-1} C_{t+1}'$$

Simulation Based Filtering

The Kalman filter assumes normality, although it gives the best linear estimate of the state even without normality. Simulation based methods are one way to handle non-normality (and indeed nonlinearity). Consider the model:

$$y_t = Z_t \alpha_t + \varepsilon_t$$
$$\alpha_t = T \alpha_{t-1} + \eta_t$$

where the errors have arbitrary densities. Consider the following algorithm:

- Take a set of n draws from the distribution of α_0.
- Using the model, get implied draws of α_1. Call these $\{\tilde{\alpha}_{1i}\}_{i=1}^n$.
- Now compute $q_{1i} = p(y_1 \mid \tilde{\alpha}_{1i})$ for each draw.
- Resample from $\{\tilde{\alpha}_{1i}\}_{i=1}^n$ with replacement picking each draw with probability $w_{1i} = q_{1i} / \Sigma_{j=1}^n q_{1j}$. This is a now a set of draws from the distribution of $\alpha_1 \mid y_1$.
- Continue in the same was cycling through the whole sample.
- The density of y_t conditional on Y_{t-1} can be approximated as $\Sigma_{i=1}^n p(y_t \mid \tilde{\alpha}_{ti})$ and the log likelihood is $\Sigma_{t=1}^T \log p(y_t \mid Y_{t-1})$.

There is also a smoother. As before, you first run the filter and save the outputs. The outputs are $\{\tilde{\alpha}_{ti}\}_{i=1t=1}^{n\ T}$ and $\{w_{ti}\}_{i=1t=1}^{n\ T}$. Now we construct alternatives weights (probabilities) for the smoother. These start from $\overline{w}_{Ti} = w_{Ti}$ and then cycle backwards setting

$$\overline{w}_{ti} = \Sigma_{j=1}^n \overline{w}_{t+1j} \frac{w_{ti}\, p(\tilde{\alpha}_{t+1j} \mid \tilde{\alpha}_{ti})}{\Sigma_{k=1}^n w_{tk}\, p(\tilde{\alpha}_{t+1j} \mid \tilde{\alpha}_{tk})}$$

Note that $\Sigma_{i=1}^n \overline{w}_{ti} = 1$. The density of α_t conditional on Y_T can now be approximated by resampling from $\{\tilde{\alpha}_{ti}\}_{i=1}^n$ picking each draw with probability $\overline{w}_{ti}$.

Hamilton Switching Model

This is an important nonlinear filtering model.

$y_t = \alpha + \beta S_t + \varepsilon_t$

where S_t is a Markov switching process.

$P(S_t = 1 \mid S_{t-1} = 1) = p$

$P(S_t = 0 \mid S_{t-1} = 0) = q$

The error ε_t is $N(0, \sigma^2)$

<u>Backgorund and Steady state of a Markov Chain</u>

Define $p_t = (P(S_t = 0), P(S_t = 1))'$. Then $p_t = \Pi p_{t-1}$

where Π is the transition matrix and in this case

$$\Pi = \begin{pmatrix} q & 1-p \\ 1-q & p \end{pmatrix}$$

Under regularity conditions, $p_t \to p^*$, the "steady state". If so

$p^* = \pi p^*$

and in this case

$$p^* = \begin{pmatrix} q & 1-p \\ 1-q & p \end{pmatrix} p^*$$

Write $p^* = (h, 1-h)'$. Then

$$h = qh + (1-p)(1-h) \Rightarrow h = \frac{1-p}{2-p-q}$$

So the steady state of the Markov chain is $p^* = (\frac{1-p}{2-p-q}, \frac{1-q}{2-p-q})'$

Now back to the model. The log likelihood is

$\Sigma_{t=1}^{T} \log(f(y_t \mid Y_{t-1}))$

and

$$f(y_t \mid Y_{t-1}) = N(\alpha, \sigma^2) P(S_t = 0 \mid Y_{t-1}) + N(\alpha + \beta, \sigma^2) P(S_t = 1 \mid Y_{t-1})$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(y_t - \alpha)^2}{2\sigma^2}) P(S_t = 0 \mid Y_{t-1}) + \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(y_t - \alpha - \beta)^2}{2\sigma^2}) P(S_t = 1 \mid Y_{t-1})$$

a mixture of normals.

<u>Prediction equations</u>

$$P(S_t = 1 \mid Y_{t-1}) = pP(S_{t-1} = 1 \mid Y_{t-1}) + (1-q) P(S_{t-1} = 0 \mid Y_{t-1})$$

$$P(S_t = 0 \mid Y_{t-1}) = (1-p) P(S_{t-1} = 1 \mid Y_{t-1}) + qP(S_{t-1} = 0 \mid Y_{t-1})$$

<u>Updating equations (from Bayes Theorem)</u>

$$P(S_t = 0 \mid Y_t) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(y_t - \alpha)^2}{2\sigma^2}) P(S_t = 0 \mid Y_{t-1}) / f(y_t \mid Y_{t-1})$$

$$P(S_t = 1 \mid Y_t) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(y_t - \alpha - \beta)^2}{2\sigma^2}) P(S_t = 1 \mid Y_{t-1}) / f(y_t \mid Y_{t-1})$$

<u>Starting iterations</u>

Only need $P(S_1 = 1 \mid Y_0) = P(S_1 = 1)$, the unconditional probability. We know that

$$P(S_1 = 1) = \frac{1-p}{2-p-q}$$

and this allows us to start the recursions.

The model is useful for fitting business cycles (Hamilton (1989)). A few last comments:

(i) In practice, we would replace ε_t by an AR process.

(ii) Also it is possible to let p and q depend on variables at time $t-1$.

(iii) There is also a smoother that can be run backwards to get state probabilities conditional on the whole sample. This uses the recursions

$$P(S_t=1|Y_T)=\frac{P(S_{t+1}=0|Y_T)P(S_t=1|Y_t)P(S_{t+1}=0|S_t=1)}{P(S_{t+1}=0|Y_t)}+\frac{P(S_{t+1}=1|Y_T)P(S_t=1|Y_t)P(S_{t+1}=1|S_t=1)}{P(S_{t+1}=1|Y_t)}$$

for $t=T-1,T-2,...$ starting from $P(S_T=1|Y_T)$.

Bayesian Approach to the Hamilton Model

A Bayesian approach to the Hamilton switching model is also available. Sometimes maximum-likelihood estimation of the model creates too much switching and the Bayesian approach allows us to keep the system in one state for longer periods of time, which has more economic appeal.

Suppose for simplicity that σ^2 is known. Let the prior for $\theta=(\alpha,\beta)'$ be $N(\theta_0,\Theta_0)$. Let p and q have beta priors that are $\beta(u_{11},u_{10})$ and $\beta(u_{00},u_{01})$, respectively. A Gibbs sampler is available for drawing from the posterior. The idea is that there are three steps:

1. Draw from the states conditional on α, β, p and q. To do this involves the following sub-steps:

(a) Apply the Hamilton filter to work out $\{P(S_t=1|Y_t)\}_{t=1}^T$ and $\{P(S_{t+1}=1|Y_t)\}_{t=1}^T$

(b) Take a draw of S_T which will be 1 with probability $P(S_T=1|Y_T)$ and zero otherwise.

(c) Because the regime is a Markovian process, the probability mass function of the vector of states conditional on the vector of data can be written as

$$P(S_t=1|\mathrm{S}_{t+1},Y_T)=P(S_t=1|\mathrm{S}_{t+1},Y_t)=\frac{P(S_{t+1}|S_t=1,Y_t)P(S_t=1|Y_t)}{P(S_{t+1}|Y_t)}=\frac{P(S_{t+1}|S_t=1)P(S_t=1|Y_t)}{P(S_{t+1}|Y_t)}$$

for $t=T-1$, where $P(S_{t+1}|S_t=1)$ means the probability of the observed draw for S_{t+1} given that $S_t=1$, obtained from the transition matrix. For example, if $S_{t+1}=1$, then $P(S_{t+1}|S_t=1)=p$. Take a draw of S_{T-1} from this distribution.

(d) Repeat step (c) backwards for $t=T-2,T-3,...1$ to generate the entire series of states.

2. Conditional on the states, take a draw from the distribution of p and q, If n_{00} is the number of transitions from state 0 to state 0 (from the draws of the states in part 1) and n_{01}, n_{11} and n_{10} are defined similarly, then the posterior for p is $\beta(u_{11}+n_{11},u_{10}+n_{10})$ and the posterior for q is $\beta(u_{00}+n_{00},u_{01}+n_{01})$.

3. Take a draw from the posterior distribution of α and β which is $N(\theta_1,\Theta_1)$ where

$$\Theta_1=(\Theta_0^{-1}+\frac{1}{\sigma^2}X'X)^{-1}$$

$$\theta_1=\Theta_1(\Theta_0^{-1}\theta_0+\frac{1}{\sigma^2}X'Y)$$

and X is the Tx2 matrix, the tth row of which is $(1,S_t)$ with the draw of the state from part 1 and Y is the Tx1 data vector.

News Announcements

In the U.S. (and many other countries) information about the economy is released at precisely scheduled times. Most economic data in the U.S. come out at 8:30am, although announcement by the Fed about the target federal funds rate are instead at 2:15pm. Studying the effects of these announcements on asset prices is about as close as we can get in macroeconomics to a natural experiment. Edderington and Lee (1993) is a pioneering paper.

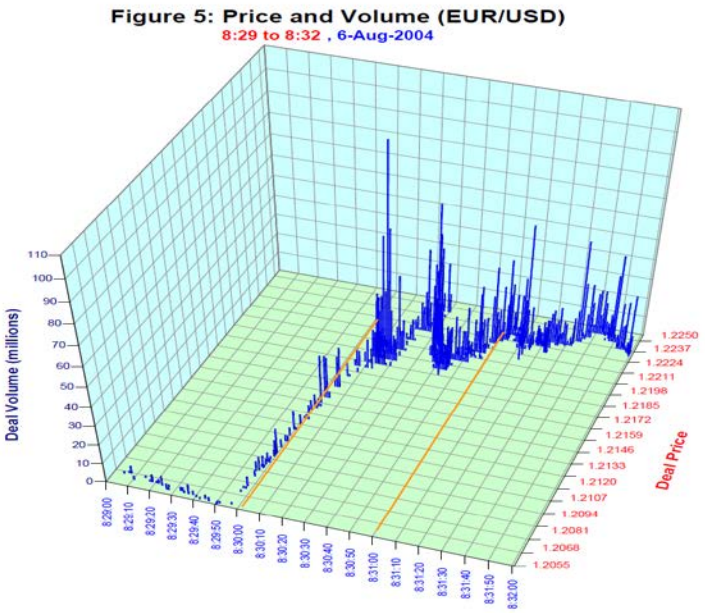
Ahead of each announcement, there are survey expectations of what the announced value will be (from Money Market Services). The surprise component of the announcement is then given as

$$\text{Surprise}=\text{Actual} - \text{Expected (from survey)}$$

and it is common to scale these surprises by their standard deviation. With high-frequency asset price data, we can then look at the effects of these announcements on asset prices. Consider (say) an announcement at 8:30am. The idea is to run a regression of the form

$$\text{Price at 8:45} - \text{Price at 8:25} = \alpha + \beta * \text{Surprise} + \text{Error}$$

A question that comes up is a how small a window you want to use to measure the price impact (in this equation it is a 20 minute window). From an efficient-markets perspective, agents should update their asset prices very quickly following the news announcement. And it generally looks like they do. This chart gives an example



It shows the euro-dollar exchange rate around a nonfarm payrolls announcement on August 6, 2004. The announcement came out at 8:30 am and indicated that the U.S. labor market was

weaker than had been expected. The bars in the chart show the amount that was traded at each price in each second. Price is on the right side in red, time is alone the front, and the heights of the bars are the volumes. What you see is that the price jumped within seconds of the announcement without much trading volume. Then the volume picked up and there was high volatility around the new price level.

So from this perspective, you want some very short window to measure asset price changes. Longer windows just add noise to the left-hand-side variable, which reduces the precision of the slope coefficient estimate. Of course, sometimes one wonders how long the effect really sticks for. That's a hard question to answer, at least via the event-study methodology. With sufficiently large windows, coefficients will be so imprecisely estimated that we will rarely find anything to be significant.

For one announcement, it is customary not to use the survey expectations. This is the announcement by the FOMC of the target for the federal funds rate. That's because there are futures markets which are bets on the average level of the federal funds rate for each calendar month. These are effectively bets on the average level of the federal funds rate for that month. The bet is set at some level, F, which is the futures price. When the actual realized funds rate, R, is known, one party pays the other party \$50*(R-F). where F and R are both measured in basis points. Under risk-neutrality, the federal funds futures rate should be the expectation of the realized funds rate. The definition of the realized rate is the average expected daily federal funds rate over all days in the month.

The FOMC sets the target for the federal funds rate at meetings 8 times a year. Assume that the actual and target funds rate are the same, that there will be no intermeeting rate changes, and that investors are risk-neutral. If there is no meeting *next* month, then the expected decision from the FOMC meeting can be read off as the next-month futures rate just before the FOMC announcement. If there is a meeting *next* month, then we have to use the current month contract instead. Suppose that the FOMC meeting is on day D(F) and that the number of days in the month is D. Then the current-month futures rate before the FOMC announcement is

$$\text{Futures} = \frac{D(F)}{D} * O + (\frac{D - D(F)}{D}) * E$$

where O and E denote the old funds rate, and the expectation for the funds rate, respectively. We can immediately use this to solve for the expected funds rate after the meeting as

$$E = \frac{D * \text{Futures-}D(F) * O}{D - D(F)}$$

The futures-based expectation of the federal funds rate can be used as the expectation instead of the survey quote. This approach was first used by Kuttner (2001). An advantage of this is that it is more timely. The survey is taken the Friday before the FOMC announcement; whereas the futures quote can be from the day before, or even five minutes before the announcement time.

Announcements contain more than just one headline number. For macroeconomic news, there is a wealth of disaggregated data. For the FOMC announcement, there is a statement that market participants use to infer expectations about the future path of policy. Often there is more news in the statement than in the actual decision concerning the target for the federal funds rate. Two approaches have been taken to quantifying the information in the text of the statement:

(i) Gürkaynak, Sack and Swanson (2005) use the change in futures quotes for subsequent month in a short window around the FOMC announcement to measure the statement surprise. More precisely, they take the piece that is orthogonal to the target funds rate surprise. Federal funds futures are liquid only about 6 months out. To get interest rate futures quotes at longer horizons, one needs to use Eurodollar futures, which are bets on the level of the three-month interest rate.
(ii) Lucca and Trebbi (2009) use linguistic scoring methods.

Here a number of the findings that researchers have obtained on regressing changes in asset prices in short windows around announcements on the unexpected components of those announcements:

(i) There are systematic relations between surprises and changes in asset prices. In general, stronger-than-expected data drives interest rates up, bond prices down, and the foreign exchange value of the dollar up. In lower frequency data, it is very hard to find a relationship between exchange rates and macroeconomic fundamentals. For all practical purposes, in low-frequency data, the exchange rate is a random walk that is unrelated to fundamentals. Nevertheless, stronger-than-expected data in the U.S. causes the dollar to appreciate (Andersen, Bollerslev, Diebold and Vega (2003). So there is some link between fundamentals and the exchange rate, at least in small windows around news announcements.

(ii) Announcements cause jumps in the conditional mean of asset prices and longer-lasting increases in both volume and volatility.

(iii) Interest rates are quite sensitive to macroeconomic news, including long-term interest rates. More on this later.

(iv) Announcements of tighter-than-expected monetary policy by the Fed cause stock prices to fall and affect interest rates at all maturities, although the effect on long-term bond yields is modest. A 25 basis point surprise tightening lowers stock prices by about 1 percent.

(v) Researchers have tried hard to find time-variation in the sensitivity of asset prices to news. They have had some success, but not much. One finding of particular interest is that rising unemployment is associated with increasing stock prices during economic expansions ,but with falling stock prices during recessions (Boyd, Jagannathan and Hu (2005), Andersen, Bollerslev, Diebold and Vega (2007)). The Gordon growth model provides a natural framework for thinking about this. There are two offsetting effects: bad news about the economy lowers dividend expectations but also lower discount rates. The balance between the two appears to be different in expansions and recessions.

The event study methodology can also consider effects at lower frequency. For example, Ottonello and Winberry (2020) look at the effects of monetary policy surprises on firm investment at the quarterly frequency. In these cases, the surprises are aggregated within the quarter. Also, the effects of the surprises can be allowed to vary using interaction effects. For example, Ottonello and Winberry consider panel data regressions of the form:

$$I_{it} = \alpha_i + \gamma_t + \beta_1 MPS_t + \beta_2 L_{it} MPS_t + \beta_3 L_{it} + \varepsilon_{it}$$

where I_{it} and L_{it} are investment and leverage of firm i and MPS_t is the monetary policy surprise.

Reverse Engineering Probabilities from Asset Prices: Ps and Qs

Let us suppose that there are K possible states of the world and that there are K assets. Suppose that the ith asset has a payoff of $P(i, j)$ in the jth state of the world. Let $p(i)$ be the price of the ith asset and let $\pi(j)$ be the probability of the jth state of the world occurring. Finally, let $p = (p(1), p(2), ...p(K))'$, $\pi = (\pi(1), \pi(2), ...\pi(K))'$ and P denote the matrix the ijth element of which is P. The last asset has a payoff of 1 in all states.

Finally, also assume that investors are risk-neutral, so that the price that they pay for any asset is equal to its expected payoff. Then we can write

$$p = P\pi$$

Provided that the matrix P is invertible, we can reverse-engineer the probabilities of the different states as

$$\pi = P^{-1}p$$

The inclusion of the asset which has a payoff of 1 in all states is a device to force the probabilities to add to 1.

All this is under the assumption of risk-neutrality. But we generally think that agents are not risk neutral. The prices that they pay will in any case, under minimal assumptions, satisfy the relation

$$p(i) = E(M(j)P(i, j)) = \Sigma_{j=1}^{K}\pi(j)M(j)P(i, j)$$

where $M(j)$ is the pricing kernel. Now define

$$\tilde{\pi}(j) = \frac{\pi(j)M(j)}{\Sigma_{i=k}^{K}\pi(k)M(k)} = \frac{\pi(j)M(j)}{E(M)}$$

Since a pricing kernel is non-negative, it is easy to check that $\tilde{\pi}(j)$ satisfies the requirements to be a probability (non-negative, less than one, and sums to 1 over all values of j). And we can write

$$p(i) = E(M)\Sigma_{j=1}^{K}\tilde{\pi}(j)P(i, j) = \frac{1}{R_f}\Sigma_{j=1}^{K}\tilde{\pi}(j)P(i, j)$$

where R_f is the gross riskfree rate. Over a short period, we might think that this is close to 1. We would then have

$$p(i) = E^{Q}(P(i, j))$$

where this expectation is taken with respect to "fake" probabilities. These are the true probabilities if agents are risk-neutral, but otherwise they are fake probabilities that overweight bad states of the world and underweight good states. Investors are pricing these assets as though they are risk neutral but with these distorted probabilities. They are therefore called risk-neutral probabilities, as opposed to the actual probabilities which are called physical probabilities. A

common terminology is that the physical probabilities are the P-measure and the risk-neutral probabilities are the Q-measure.

The probabilities that we might reverse-engineer from the prices of assets are under the Q-measure. Only if we assume that agents are risk-neutral do they happily also turn out to be probabilities under the P-measure.

A very standard place in which we want to back out probabilities of different states of the world is working out the probability density for the price of an asset that are implied by options prices. A call option with a strike price of K has a payoff at maturity of $\max(0, P-K)$ where P is the price of the underlying asset at maturity. Meanwhile, the put option with the same strike price has a payoff of $\max(0, K-P)$. If we neglect any discounting, the price that I will pay today for the call option is $E^Q(\max(0, P-K))$ and the price for the put option is $E^Q(\max(0, K-P))$.

If the number of options prices is equal to the number of possible values for P, then we can use the observed options prices to solve for the probability density function for the price of the asset. As an illustration, suppose that we have prices of call options on oil with the following strikes

Strike Price	Call Option Price
70	10.50
75	6.00
80	3.00
85	1.00

and suppose that the possible prices for oil at maturity are 70, 75, 80, 85 and 90. The payoffs of the options under the different scenarios are as follows

Option	P=70	P=75	P=80	P=85	P=90
70	0	5	10	15	20
75	0	0	5	10	15
80	0	0	0	5	10
85	0	0	0	0	5
	1	1	1	1	1

with the last row representing the asset with a payoff of one in all states of the world. So if I let this be the matrix P and take my vector of prices as $p = (10.5, 6, 3, 1, 1)'$, I can easily solve for the risk-neutral probabilities which are $P^{-1}p$.

Outcome	Risk-Neutral Probability
70	10%
75	30%
80	20%
85	20%
90	20%

One thing to be careful of in these calculations is put-call parity. Put-call parity is an arbitrage relationship that exists between put and call options at the same strike price. If you include put and call options in the set of asset prices used to reverse engineer the probability density, then some of these will be redundant. It will manifest itself in a singular matrix P. You avoid the problem simply by deleting the redundant assets.

There are recent clever uses of options. One example is from Martin (2017) to get a lower bound for the equity risk premium, and a bound that is not so low as to be uninformative. Here is the logic. The equity risk premium is:

$$E(R_{t+1}) - RF_t = E(R_{t+1}^2 M_{t+1}) - RF_t - [E(R_{t+1}^2 M_{t+1}) - E(R_{t+1})]$$

Denoting moments under risk neutral probabilities with an asterisk:

$$Var^*(R_{t+1}) = E^*(R_{t+1}^2) - E^*(R_{t+1})^2 = RF_t E(M_{t+1} R_{t+1}^2) - RF_t E(M_{t+1} R_{t+1})^2 = RF_t E(M_{t+1} R_{t+1}^2) - RF_t$$

Plugging this into the first equation gives

$$E(R_{t+1}) - RF_t = \frac{1}{RF_t} Var^*(R_{t+1}) - Cov(M_{t+1} R_{t+1}, R_{t+1})$$

The paper argues that the second term is negative in lots of models, which means that $\frac{1}{RF_t} Var^*(R_{t+1})$ is a lower bound on the equity risk premium. This can in turn be measured from options prices as:

$$\frac{1}{RF_t} \text{var}^*(R_{t+1}) = \frac{2}{S_t} \{ \int_0^F P_t(K) dK + \int_F^\infty C_t(K) dK \}$$

where S_t is the stock price today, F is the forward stock price and P_t and C_t denote prices of put and call options.

Kremers and Martin (2018) have a somewhat related approach to explaining, or at least modeling, failures of uncovered interest parity (UIP). After some algebra, you can write:

$$E_t(\frac{S_{t+1}}{S_t}) = \frac{RF_t}{RF_t^*} + \frac{1}{RF_t^*} Cov_t^*(\frac{S_{t+1}}{S_t}, R_{t+1}) - Cov_t(M_{t+1} R_{t+1}, \frac{S_{t+1}}{S_t})$$

where S_t is the exchange rate. Under UIP, only the first term exists. The second term is something that can be measured from quanto index contacts. These are futures based on (say) the S&P500 index but in a non-dollar currency. It turns out that the second term explains a lot of the failure of UIP.

Bond Markets and the Term Structure of Interest Rates

The most basic building block of bond analysis is a zero coupon bond. This gives the holder the right to \$1 in *n* years time. Let $P_{ZC}(n)$ denote the price of this bond.

The continuously compounded yield on this bond is

$$y_{ZC}^{cc}(n) = [\log(1) - \log(P_{ZC}(n))] / n = -\frac{1}{n}\log(P_{ZC}(n))$$

and we can write

$$P_{ZC}(n) = \exp(-ny_{ZC}^{cc}(n))$$

The yield with annual compounding is $\exp(y_{ZC}^{cc}) - 1$. The yield with semiannual compounding is $2[\exp(\frac{y_{ZC}^{cc}(n)}{2}) - 1]$. In academic finance, we work with zero coupon bonds and continuous compounding.

Forward Rates

Zero-coupon bonds of different maturities can be combined to guarantee an interest rate on an investment to be made in the future. Let $P_{ZC}(n)$ be the price of an *n*-period zero-coupon bond
Suppose that I

- Buy $\frac{P_{ZC}(n+1)}{P_{ZC}(n)}$ *n*-period bonds. I pay $P_{ZC}(n)\frac{P_{ZC}(n+1)}{P_{ZC}(n)} = P_{ZC}(n+1)$

and

- Sell one *(n+1)*-period bond for which I receive $P_{ZC}(n+1)$

Working through what happens:
1. The two cash flows cancel out exactly today.
2. In year *n*, I receive $P_{ZC}(n+1) / P_{zc}(n)$
3. In year *n+1*, I pay \$1.

I have in this way synthesized borrowing from *m* to *m+1*, at a rate locked in today.

The continuously-compounded return that I have locked in is

$$f_{n,n+1} = \log(1) - \log(\frac{P_{ZC}(n+1)}{P_{ZC}(n)}) = -\log(\frac{P_{ZC}(n+1)}{P_{ZC}(n)}) = -\log(\frac{\exp(-(n+1)y_{ZC}^{cc}(n+1))}{\exp(-ny_{ZC}^{cc}(n))}$$

$$\therefore f_{n,n+1} = (n+1)y_{ZC}^{cc}(n+1) - ny_{ZC}^{cc}(n)$$

and this is known as the continuously-compounded one-year zero-coupon forward rate from *n* to *n+1* years hence.

Recall that the continuously-compounded return on an *m*-year zero-coupon bond is

$$y_{ZC}^{cc}(n) = -\frac{1}{n}\log P_{ZC}(n)$$

That can be decomposed as

$$y_{ZC}^{cc}(n) = -[\log(\frac{P_{ZC}(1)}{P_{ZC}(0)}) + \log(\frac{P_{ZC}(2)}{P_{ZC}(1)}) + \log(\frac{P_{ZC}(3)}{P_{ZC}(2)}) ... + \log(\frac{P_{ZC}(n)}{P_{ZC}(n-1)})] / n$$

$$y_{ZC}^{cc}(n) = [f_{0,1} + f_{1,2} + f_{2,3} .... + f_{n-1,n}] / n$$

The zero-coupon yield can thus be decomposed into the average of a string of one-year forward rates.

● The ***m*-period forward rate beginning *n* years' hence** is the implied rate at which the investor would contract to borrow for *m* years beginning in *n* years' time. This is

$$f_{n,n+m} = \frac{1}{m}[(n+m)y_{ZC}^{cc}(n+m) - ny_{ZC}^{cc}(n)] = \frac{n}{m}[y_{ZC}^{cc}(n+m) - y_{ZC}^{cc}(m)] + y_{ZC}^{cc}(n+m)$$

● The **instantaneous forward rate** is the implied rate at which the investor would contract today to borrow for an arbitrarily short period in *n* years' time. This is

$$f^{INST}(n) = \lim_{m\to 0} f_{n,n+m} = \lim_{m\to 0} \frac{n}{m}[y_{ZC}^{cc}(n+m) - y_{ZC}^{cc}(n)] + y_{ZC}^{cc}(n+m)$$

$$\therefore f^{INST}(m) = \frac{dny_{ZC}^{cc}(n)}{dn} = -\frac{d\log P_{ZC}(n)}{dn}$$

So we can decompose a zero-coupon yield into an average of instantaneous forward rates:

$$y_{ZC}^{cc}(n) = \frac{1}{n}\int_0^n f^{INST}(t)dt$$

The point of forward rates is that they allow us to isolate long-term effects on bond yields that are separate from short-term levels of interest rates.

An influential paper using forward rates was Gürkaynak, Sack and Swanson (2005). This paper looked at the effects of news announcements on forward interest rates. It found that ten-year-ahead forward rates were very sensitive to macroeconomic surprises. A distant-horizon forward rate is (algebraically) the sum of long-term inflation expectations, the long-term expectations of real short-term interest rates, and the risk premium. Gürkaynak, Sack and Swanson argued that the sensitivity of long-term forward rates to news represents unanchored long-term inflation expectations. There is some discussion of the appropriate interpretation, but it is in any event very noteworthy that long-term forward rates jump around so much on one particular data release.

Coupon-Bearing Securities

Bonds in the U.S. pay coupons twice a year, but we will pretend that bonds pay annual coupons and will work with annual compounding for expositional simplicity.

Consider a coupon-bond that pays C each year and $\$1+C$ at the maturity of the bond in n years time. Let $P_{C,n}$ be the price of this bond.

The bond can be thought of as a bundle of zero coupon bonds and so the price must satisfy

$$P_{C,n} = CP_{ZC}(1) + CP_{ZC}(2).... + (1+C)P_{ZC}(n)$$

The yield with annual compounding is defined as the value of $y_{C,m}$ that satisfies the equation

$$P_{C,n} = \frac{C}{(1+y_{C,n})} + \frac{C}{(1+y_{C,n})^2} + \frac{C}{(1+y_{C,n})^3}.... + \frac{1+C}{(1+y_{C,n})^n} = C\Sigma_{j=1}^{n}(\frac{1}{1+y_{C,n}})^j + (\frac{1}{1+y_{C,n}})^n$$

Whereas academic work looks at zero coupon yields with continuous compounding, the market convention is to look at yields on coupon bearing securities wiith compounding at the coupon frequency (bond-equivalent, coupon-equivalent)

Three Special Cases

1. If $y_{C,n}=C$, then

$$P_{C,n} = y_{C,n}\Sigma_{j=1}^{n}(\frac{1}{1+y_{C,n}})^j + \frac{1}{(1+y_{C,n})^n} = 1$$

and the bond is said to be trading "at par" and

$$y_n^{PAR} = y_{C,n} = C$$

In this case:

$$1 = P_{c,n} = CP_{ZC}(1) + CP_{ZC}(2).... + (1+C)P_{ZC}(n) = C\Sigma_{j=1}^{n}P_{ZC}(j) + P_{ZC}(n)$$

$$C\Sigma_{j=1}^{n}P_{ZC}(j) = 1 - P_{ZC}(n)$$

$$\therefore C = y_n^{PAR} = \frac{1-P_{ZC}(n)}{\Sigma_{j=1}^{n}P_{ZC}(n)}$$

2. When the maturity n is infinite (perpetuity),

$$P_{C,n} = C\Sigma_{j=1}^{\infty}(\frac{1}{1+y_{C,n}})^j = C[\frac{(\frac{1}{1+y_{C,n}})}{1-(\frac{1}{1+y_{C,n}})}] = C\frac{1}{1+y_{C,n}-1} = \frac{C}{y_{C,n}}$$

In this case:

$$\frac{C}{y_{C,n}} = P_{C,n} = CP_{ZC}(1) + CP_{ZC}(2)... = C\Sigma_{j=1}^{\infty}P_{ZC}(j)$$

$$\frac{C}{y_{C,n}} = C\Sigma_{j=1}^{\infty}P_{ZC}(j)$$

$$\therefore y_{C,n} = \frac{1}{\Sigma_{j=1}^{\infty}P_{ZC}(j)}$$

3. If the coupon is zero, then $P_{0,n} = P_{ZC}(n) = \frac{1}{(1+y_{0,n})^n}$.

Relationship Between Price and Yield

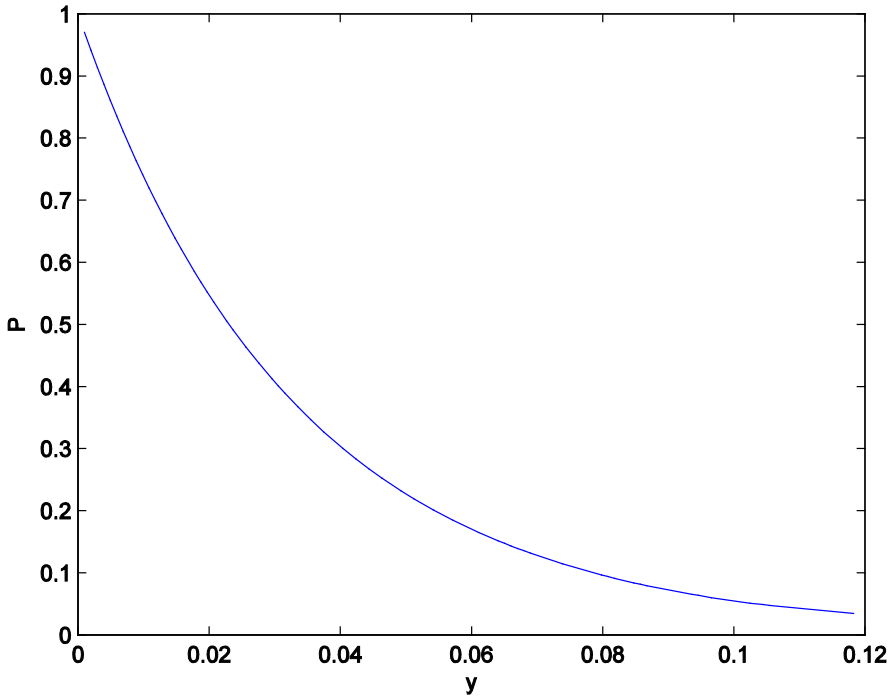
For a zero-coupon *n*-period bond, recall that

$P_{ZC}(n) = \exp(-ny_{ZC}^{cc}(n))$

and hence

$P_{ZC}(n) = \exp(-n\log[y_{ZC}(n)+1])$

Here is a plot of $P_{ZC}(n)$ against $y_{ZC}(n)$ for n=30.



The relationship is downward sloping but is also convex.

Duration

Duration is a central concept in bond market maths.

The (MacAulay) **duration** of a bond is

$$D = -\frac{d\log(P)}{dy}(1+y)$$

The **modified duration** of a bond is simply

$$D_{\text{mod}} = -\frac{d\log(P)}{dy} = \frac{D}{1+y}$$

What does duration mean? It means three things at once:

1. If yield rises one percentage point, then price falls by about D_{mod} percent. In a first order Taylor series expansion about price P_0 and yield y_0,

$$P \approx P_0 + \frac{dP}{dy}(y - y_0)$$

$$\therefore \frac{P - P_0}{P_0} \approx \frac{1}{P_0}\frac{dP}{dy}(y - y_0) = -D_{\text{mod}}(y - y_0)$$

In other words duration is the slope of the relationship between log price and yield.

2. A measure of the impact of **small parallel** shifts in the yield curve on the value of a fixed income portfolio. If I have a portfolio of B_1 units of a bond with modified duration D_1, B_2 units of a bond with duration D_2,....B_q units of a bond with duration D_q, then a one percent parallel shift in the yield curve will lead to a percentage change in the portfolio value of $-\Sigma_{j=1}^{q} B_q D_q$

3. The weighted average of how long an investor has to wait to get money (hence "duration")

$$D = \frac{1}{P}\{\frac{C}{(1+y)} + \frac{2C}{(1+y)^2} + \frac{3C}{(1+y)^3} .... + \frac{m(1+C)}{(1+y)^m}\}$$

The Nelson-Siegel curve

Bonds only trade with certain coupons and certain maturities. It is important to have a smoothed implied zero-coupon curve. This can be obtained in a number of ways. A popular method is to use the Nelson-Siegel functional form. In this , the instantaneous forward rates are of the form:

$$f_t = \beta_0 + \beta_1 \exp(-n/\tau) + \beta_2 (n/\tau)\exp(-n/\tau)$$

and the yields, from integration, are:

$$y_t(n) = \beta_0 + \beta_1 \frac{1 - \exp(-n/\tau)}{n/\tau} + \beta_2 [\frac{1 - \exp(-n/\tau)}{n/\tau} - \exp(-n/\tau)]$$

The four parameters: β_0, β_1, β_2 and τ can all be obtained by minimizing the difference between observed yields on coupon-bearing securities and the implied yields from this curve. For more details, see Gürkaynak, Sack and Wright (2007).

The Expectations Hypothesis and Affine Term Structure Models

The expectations hypothesis says that bonds are priced as though agents are risk neutral. This means that the price of a zero-coupon bond is

$$P_{ZC}(n) = E(\int_{s=0}^{n} e^{-r(s)} ds)$$

where $r(s)$ is the instantaneous risk-free rate at time s. This almost says that yields will be the expected average short-term rate over the life of the bond.[1] Except that there is a Jensen's

[1] Or you could instead define the expectations hypothesis as saying that yields are the expected average short-term rate over the life of the bond, in which case the Jensen's inequality term means that agents are not exactly risk neutral.

inequality term which means that actually the yields will be slightly higher than the expectation of average short rates. But, unless the maturity is very long, this is a small effect.

Here a number of tests of the expectations hypothesis:

1. Suppose I buy an *n*-period zero coupon bond today. And I sell it in one year as an (*n-1*)-period zero coupon bond. My holding period return from this exercise is:

$$\log(P_{t+1}(n-1)) - \log(P_t(n))$$

and the excess bond returns, or excess holding period return, over the one period interest rate is:

$$exrt = \log(P_{t+1}(n-1)) - \log(P_t(n)) - y_t(1)$$

Now the expectations hypothesis says that this ought not to be forecastable *ex-ante*. But when regressions of the form

$$exrt_{t,t+1} = \alpha + \beta' x_t + \varepsilon_{t,t+1}$$

are run using term structure variables at time t as the predictors, it turns out that they are significant. For example, the steeper is the slope of the yield curve at time t, the higher excess returns subsequently tend to be. Cochrane and Piazzesi (2005) instead use the term structure of forward rates as predictors, and get extremely strong forecasting power.

2. Another approach, was adopted by Campbell and Shiller (1991). The expectations hypothesis implies that the expectation of the future interest rate from m to n periods hence is the forward rate over that period. So

$$E_t(y_{t+m}(n-m)) = \frac{n}{n-m} y_t(n) - \frac{m}{n-m} y_t(m)$$

$$\therefore E_t(y_{t+m}(n-m) - y_t(n)) = \frac{m}{n-m}(y_t(n) - y_t(m))$$

and so if we regress $y_{t+m}(n-m) - y_t(n)$ onto $\frac{m}{n-m}(y_t(n) - y_t(m))$, one ought to get a slope coefficient of 1. But in the data, if m is short (say 3 months) and n is long (say 10 years), the estimated slope coefficient is negative. When the yield curve is steep, according to the expectations hypothesis, long-term interest rates should be *rising*, but in fact they are *falling*.[2]

3. Campbell and Shiller also considered a second test of the expectations hypothesis. The EH implies that the n-period yield is the expected average *m*-period interest rate over the next n periods:

$$y_t^{(n)} = \frac{1}{k} E_t(\Sigma_{i=0}^{k-1} y_{t+im}^{(m)})$$

where $k = n/m$. This means that

[2] If the expectations hypothesis were right up to a constant risk premium, then this would just be absorbed into the constant of the Campbell and Shiller regression.

$$y_t^{(n)} - y_t^{(m)} = \frac{1}{k} E_t(\Sigma_{i=0}^{k-1} y_{t+im}^{(m)}) - y_t^{(m)}$$

$$\therefore y_t^{(n)} - y_t^{(m)} = \Sigma_{i=1}^{k-1}(1-\frac{i}{k}) E_t(y_{t+im}^{(m)} - y_{t+(i-1)m}^{(m)})$$

and so if we consider the regression

$$\Sigma_{i=1}^{k-1}(1-\frac{i}{k})(y_{t+im}^{(m)} - y_{t+(i-1)m}^{(m)}) = \alpha + \beta(y_t(n) - y_t(m)) + \varepsilon_t$$

which is a regression of a weighted-average of future short-term yield changes onto the slope of the term structure, then one ought to get a slope coefficient β that is equal to one. The dependent variable in this equation can be thought of as the *perfect-foresight* term spread, as it is the term spread that would prevail at time t if the path of m period interest rates over the next n periods were correctly anticipated. Whereas tests (1) and (2) give consistent evidence against the expectations hypothesis, the results of test (3) are more mixed. The slope coefficient is positive and in some cases not significantly different from 1.

Affine Models

Affine term structure models provide a way of modeling the term structure of interest rates. From the basic asset pricing relation, the price of the bond must be

$$P_{ZC}(n) = E_t(\Pi_{j=1}^{n} M_{t+j})$$

which can be seen from recursive substitution. Assume that the pricing kernel is conditionally lognormal

$$M_{t+1} = \exp(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\varepsilon_{t+1})$$

where $\lambda_t = \lambda_0 + \lambda_1 X_t$ is an affine function of an mx1 vector of state variables, X_t, ε_{t+1} is iid $N(0, I)$, and $r_t = \delta_0 + \delta_1' X_t$ is the one-period interest rate. Assume further that the vector of state variables follows a vector autoregression (VAR)

$$X_{t+1} = \mu + \Phi X_t + \Sigma\varepsilon_{t+1} \tag{1}$$

It then follows that

$$P_{ZC}(n) = \exp(A_n + B_n' X_t)$$

where A_n is a scalar and B_n is an m x1 vector that satisfy the recursions

$$A_{n+1} = -\delta_0 + A_n + B_n'(\mu - \Sigma\lambda_0) + \frac{1}{2} B_n'\Sigma\Sigma' B_n$$

$$B_{n+1} = (\Phi - \Sigma\lambda_1)' B_n - \delta_1$$

starting from $A_1 = -\delta_0$ and $B_1 = -\delta_1$. The bond prices given in this way are the same as though agents were risk-neutral ($\lambda_0 = \lambda_1 = 0$), but the state vector followed an alternative law of motion:

$$X_{t+1} = \mu^* + \Phi^* X_t + \Sigma\varepsilon_{t+1} \quad (2)$$

where $\mu^* = \mu - \Sigma\lambda_0$ and $\Phi^* = \Phi - \Sigma\lambda_1$. Equations (1) and (2) are the physical and risk neutral representations of the law of motion for the state vector, or the P and Q measures, respectively. Note that the variance-covariance matrix of the shocks is the same in both equations (1) and (2). The yields are in turn given by

$$y_t(n) = -\frac{1}{n}\log(P_{ZC}(n)) = -\frac{A_n}{n} - \frac{B_n'}{n} X_t \quad (3)$$

A fairly simple way of estimating the parameters of this model is available when the elements of the state vector are observed. in this case, we can estimate equation (1) by least squares and can estimate δ_0 and δ_1 by an OLS regression of short-term interest rates onto X_t. The remaining parameters are λ_0 and λ_1. These can be estimated by minimizing the distance between actual yields and the yields that are implied by the model (equation (3)). More involved approaches to estimation are of course available, and these can apply when the state vector is latent.

A different approach to estimating the risk premium on longer term bonds is to look at the difference between long-term interest rates and expectations of average short-term interest rates over the life of the bond obtained from surveys. At a minimum, it is useful as a robustness check.

Arbitrage-Free Nelson-Siegel

A number of papers have been written by Jens Christensen, Frank Diebold and Glenn Rudebusch that propose a powerful but simple approach to term structure modeling in a number of papers.

First start out with the Nelson-Siegel curve

$$y_t(n) = \beta_{0t} + \beta_{1t}\frac{1-\exp(-n/\tau)}{n/\tau} + \beta_{2t}[\frac{1-\exp(-n/\tau)}{n/\tau} - \exp(-n/\tau)]$$

If we fix the parameter τ, the remaining parameters (β_{0t}, β_{1t} and β_{2t}) can be estimated each time period and then a VAR can be used to forecast future values of these parameters and hence future interest rates. Notice that I have added time subscripts to the parameters that are allowed to change over time. This approach was proposed by Diebold and Li (2006). It appears to give good interest rate forecasts.

The problem with this is that it will typically imply that there are arbitrage opportunities. But the dynamic Nelson-Siegel approach can be integrated into an affine term structure model. Consider an affine term structure model with three factors and suppose that under the Q-measure (equation (2)), $\mu^* = 0$ and $\Phi^* = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\theta & \theta \\ 0 & 0 & -\theta \end{pmatrix}$. Then something amazing happens. When we go to work out the bond prices and yields, the expression for yields is approximately

$$y_t(n) \approx X_{1t} + (\frac{1-e^{-\theta n}}{\theta n})X_{2t} + (\frac{1-e^{-\theta n}}{\theta n} - e^{-\theta n})X_{3t}$$

But this is now of the same functional form as the dynamic Nelson-Siegel curve, with $\tau = 1/\theta$, $X_{1t} = \beta_{0t}$, $X_{2t} = \beta_{1t}$ and $X_{3t} = \beta_{2t}$. The parameters β_{0t}, β_{1t} and β_{2t} follow a VAR(1) under the P-measure, given by equation (1).

Affine Models with Macroeconomic Factors

It seems natural to include macroeconomic variables in the state vector X_t for an affine term structure model. If one takes a VAR including yields and macroeconomic variables (growth, inflation), lags of the macroeconomic variables Granger-cause future yields. In equation (1), the block of Φ that gives the predictive power of today's macro variables for tomorrow yields is nonzero. So far so good for including macroeconomic variables in the state vector.

But there is a problem. From equation (3), the term structure of yields is a function of the state vector. Unless there is some singularity, equation (3) applied over different maturities can be inverted to recover the state vector. However, if one regresses inflation or GDP growth on yields, the R-squared values are modest. The only way to get this singularity is if the block of Φ^* in equation (2) that gives the predictive power of today's macro variables for tomorrow yields *is* equal to zero. The macroeconomic variables are then said to be *hidden* or *unspanned* factors. They do not show up in today's yields, but are important for forecasting. Joslin, Priesbsch and Singleton (2009) is one of a number of recent papers advocating treating macroeconomic variables as hidden factors.

The Bond Premium Puzzle

The methods for understanding the term structure of interest rates discussed up till now are statistical methods, not models based on optimizing behavior of agents. We would probably prefer a more structural model. Some authors put more structure in the law-of-motion of the factors (not just saying that it is a VAR(1)). Some authors use a pricing kernel derived from theory. Some do both. One generic problem is that most papers in which the pricing kernel is derived from theory cannot readily explain why bond risk premia are on average large and positive. In other words, they cannot explain why yield curves on average slope up. After all, in a recession, marginal utility of consumption is high and bond prices rise. This should make bonds the perfect hedge, commanding a negative risk premium!

The upward sloping yield curve can be thought of as the bond premium puzzle, that is analogous to the equity premium puzzle. Some recent work has made progress on explaining the bond premium puzzle from a utility-founded model (e.g. Piazzesi and Schneider (2006) and Rudebusch and Swanson (2008, 2009)). The flavor of the argument is that consumption growth and inflation both have persistent and transitory components, and the correlation between the persistent component of inflation and consumption growth is negative. This means that inflation erodes the value of a nominal bond precisely when consumption growth is low and so marginal utility is high. That makes nominal bonds risky and causes the yield curve to slope up.[3]

[3] To get this effect to be quantitatively important, either Epstein-Zin preferences or habit formation are apparently also needed.

Index-Linked Bonds

Most governments now issue index-linked bonds. In the U.S., they are called TIPS. With these bonds, the coupons and principal payments are tied to inflation. The yields on the bonds are calculated just as before, but these are real yields. A technical point that is of some importance and usefulness is that, at least in the U.S., the indexation to the principal can never be negative (although the indexation to the coupons can be). In the event of deflation, the TIPS holder gets their money back.

One of the useful spinoffs of index-linked bonds is that they provide measures of inflation expectations. Let $y_N(n)$ and $y_R(n)$ denote the nominal and real (zero-coupon continuously compounded) yields with a maturity of n, respectively. Then $\pi_{BE}(n) = y_N(n) - y_R(n)$ will be the breakeven inflation rate, or inflation compensation. It is the level of inflation which would, *ex-post*, make an investor indifferent between holding a nominal and a real security. If investors were risk neutral, it would be inflation expectations. But there are clearly other forces driving these spreads. See Campbell, Shiller and Viceria (2009) and Gürkaynak, Sack and Wright (2010) for more discussion.

Index-linked bonds have been around for about thirty years in the UK and over ten years in the U.S., so there are now enough data to do good empirical work on these bonds.

Handout on Forecasting

1. Predictive Regressions in Finance

The time series evidence for time-varying risk premia comes from equations for forecasting excess returns. The form of the equation is

$$r_{t,t+h} = \alpha + \beta' x_t + \varepsilon_{t,t+h}$$

where $r_{t,t+h}$ denotes a return from time t to $t+h$ and x_t is some predictor. Then the ex-ante expected excess return, or risk premium is $\alpha + \beta' x_t$, which will be time-varying as long as $\beta \neq 0$. For stock returns, the predictors usually used are the short-term interest rate, the dividend yield or the consumption-wealth ratio. For excess bond returns, the predictors are the slope of the yield curve or the term structure of forward rates.

When $h > 1$, the forecast errors have an overlapping structure and so the standard errors have to allow for serial correlation (Newey-West or Hansen-Hodrick standard errors). Still the relationship has something of a spurious regression, at least for large h, because the left-hand-side and right-hand-side variables are both persistent.

These regressions also have a potential for bias. To see this consider the case where $h = 1$ where the regressor is an AR(1)

$$r_{t+1} = \alpha + \beta' x_t + \varepsilon_{t+1}$$
$$x_{t+1} = \mu + \phi x_t + u_{t+1}$$

Now suppose the correlation between ε_{t+1} and u_{t+1} is δ. Then

$$\varepsilon_{t+1} = \delta \frac{\sigma_\varepsilon}{\sigma_u} u_{t+1} + \sqrt{1-\delta^2}\sigma_\varepsilon \eta_t$$

where η_t is iid with mean zero and variance 1, uncorrelated with u_{t+1}. Now, neglecting the intercept, we have

$$\hat{\beta} - \beta = \frac{\Sigma x_t \varepsilon_{t+1}}{\Sigma x_t} = \delta \frac{\sigma_\varepsilon}{\sigma_u} \frac{\Sigma x_t u_{t+1}}{\Sigma x_t} + \sqrt{1-\delta^2}\sigma_\varepsilon \frac{\Sigma x_t \eta_{t+1}}{\Sigma x_t}$$

The second term is asymptotically normal. But the first has an AR(1) type distribution which will be non-normal and biased, if ϕ is 1 (or "local to unity"). That gives an intuition for a bias. And, the bias will be a downward/upward bias if δ is positive/negative. Moreover, it is natural to expect the correlation to be big. For example, if excess stock returns come in above

expectations, the dividend yield should fall—so in this predictive regression, the correlation should be negative.

So there are strategies for predictive regressions:

1. Bias-adjustment. Estimate $\delta\sigma_\varepsilon / \sigma_u$ and the AR bias and hence adjust β by the estimated bias. Concretely, Stambaugh's bias-adjusted estimate for the case $h=1$ is

$$\hat{\beta}_{BA} = \hat{\beta} + \frac{Cov(\varepsilon_{t+1}, u_{t+1})}{Var(u_{t+1})}[\frac{1+3\hat{\phi}}{T}]$$

Or the bootstrap could be used for bias-adjustment.

2. The standard regression will have severe size distortions if ϕ is big and h is large. Standard errors that are robust to autocorrelation won't do the job. Hodrick (1992) came up with two tricks for dealing with this

(a) Reorganizing the long-horizon regression. For illustration, let's drop all the intercepts and assume that the regression is

$$r_{t,t+h} = \beta x_t + \varepsilon_{t,t+h}$$

Now under stationarity, $\beta = \frac{Cov(x_t, r_{t,t+h})}{Var(x_t)} = \frac{Cov(x_t^{(h)}, r_{t,t+1})}{Var(x_t)}$

where $x_t^{(h)} = x_t + x_{t-1} ... + x_{t-h+1}$. So if I run the regression

$$r_{t,t+1} = \gamma x_t + u_{t,t+1}$$

then $\beta = 0 \Leftrightarrow \gamma = 0$. We can test the hypothesis that $\beta = 0$ by testing the necessary and sufficient condition that $\gamma = 0$. But that has much less in the way of size distortions because it's not regressing one very persistent process on another (and that's always what's bad news).

(b) Alternative standard errors for the long-horizon regression. We can stick with the original regression and the estimate of the slope coefficient. The OLS estimate of β satisfies $T^{1/2}(\hat{\beta} - \beta) \to_d N(0, V)$. Instead of estimating V as

$$(T^{-1}\Sigma x_t x_t')^{-1} \Omega (T^{-1}\Sigma x_t x_t')^{-1}$$

where Ω is the zero-frequency spectral density of $x_t \varepsilon_{t,t+h}$, we can use the estimate

$$(T^{-1}\Sigma x_t x_t')^{-1} T^{-1}\Sigma w_{t+1} w_{t+1}' (T^{-1}\Sigma x_t x_t')^{-1}$$

where $w_{t+1} = r_{t+1} x_t^{(h)}$ which is asymptotically the same thing if $\beta = 0$ (homework problem). But it seems to work much better in small sample sizes.

Both of these methods rely on stationarity, which makes them suspect. But it is also true that they work remarkably well in relevant sample sizes. It's a little surprising that they aren't used more often. Perhaps it is because of the limitation that they only test the hypothesis that $\beta = 0$.

2. Forecast Efficiency.

Another forecasting problem is that we have some forecast (from a financial market, a survey, a a central bank etc.) and we want to test that it is the conditional expectation given the information set at that time (an "efficient" forecast). The device is to run a forecast efficiency regression. Suppose that the forecast is made at time t for time $t+h$. The realized value is y_{t+h}. and the forecast is $\hat{y}_{t+h|t}$. The regression is

$$y_{t+h} - \hat{y}_{t+h|t} = \beta' x_t + \varepsilon_{t+h}$$

where x_t is a vector of variables in the information set at time t. We could consider $x_t = 1$, which is just a check for bias. We could consider $x_t = (1, \hat{y}_{t+h|t})$, which is the "Mincer-Zarnowitz" regression. Or any other variables in the information set at the time that the forecast is made. In all cases, for an efficient forecast, we want $\beta = 0$. Newey-West standard errors can be used.

Even if β is nonzero, it doesn't necessarily mean that the forecast is not a conditional expectation because there can be structural breaks that agents are learning about slowly. For example, all forecasts of inflation underpredicted inflation in the 1970s and overpredicted in the 1980s. If you look back and apply these tests, inflation forecasts will appear to be biased down in the 70s and up in the 80s. But a more realistic interpretation is that there were changes in the economy that the forecasters were learning about slowly.

A variant on this is that forecast revisions for a fixed period forecast should be serially uncorrelated. So $\hat{y}_{t+h|t+1} - \hat{y}_{t+h|t}$ should be uncorrelated with $\hat{y}_{t+h|t} - \hat{y}_{t+h|t-1}$.

A recent paper by Patton and Timmerman (2010) considers jointly testing forecast efficiency across multiple horizons. Consider the regression

$$y_{t+h} = \alpha + \beta_0 \hat{y}_{t+h|t} + \Sigma_{j=1}^{h-1} \beta_j [\hat{y}_{t+h|t+j} - \hat{y}_{t+h|t+j-1}] + \varepsilon_t$$

So this is a regression of the outcome on the h-quarter forecast and a sequence of forecast revisions. Under the null that the forecasts are efficient at all horizons, $\alpha = 0$ and $\beta_0 = \beta_1 = \beta_2 ... = \beta_{h-1} = 1$ as this is the only way that $E_{t+h-j}(y_{t+h}) = \hat{y}_{t+h-j} \forall j = 1,...h$. This test is more efficient than testing at each horizon separately; the power gains seem in practice to be quite big.

3. Direct and Iterative Forecasts

Suppose that we want to forecast some variable (inflation or growth) h periods in the future. We can consider two approaches. One is to write down a VAR in the variable to be forecast and other variables and to iterate this forward to get the required forecast. The other is to run the "direct" forecasting regression

$$y_{t+h} = \alpha + \beta y_{t-1} + \gamma x_{t-1} + \varepsilon_{t+h}$$

perhaps including more lags. The iterative forecast has to be most efficient as long as the VAR is correctly specified. The direct forecast however may be more robust to model misspecification. In simulations, it generally seems that the iterative forecast works best.

4. Diebold Mariano tests

Out-of-sample root mean square prediction error (RMSPE) is a natural metric for the quality of point forecasts. Given two competing forecasts, we can work out their out-of-sample RMSPEs in recursive or rolling schemes. Let $\hat{u}_{1t}$ and $\hat{u}_{2t}$ denote the two prediction errors. The idea of the Diebold-Mariano test is to apply a t test to the series $z_t = u_{1t}^2 - u_{2t}^2$ and see if the mean is zero or not. Concretely, take the test statistic

$$\frac{T^{-1}\Sigma_{t=1}^{T}(\hat{u}_{1t}^2 - \hat{u}_{2t}^2)}{\hat{\sigma}/\sqrt{T}}$$

where T is the number of time periods for the out-of-sample forecast comparison and $\hat{\sigma}^2$ is the sample variance of $\hat{u}_{1t}^2 - \hat{u}_{2t}^2$. This is simply a t-statistic testing the hypothesis that

$$E(u_{1t}^2) = E(u_{2t}^2)$$

and it has a standard normal null limiting distribution. This all works well for "non-nested" forecast comparisons, that is where the neither model is nested in the other. Unfortunately, in many cases, the forecasts are nested. For example, if the two models are based on models

$$y_t = \beta_1' x_{1t} + u_{1t}$$
$$y_t = \beta_1' x_{1t} + \beta_2' x_{2t} + u_{2t}$$

then the only way that the two models will have the same prediction error variance is if $\beta_2 = 0$ in which case $u_{1t} = u_{2t}$. The fact that the errors are the same means that the normal distribution for the Diebold-Mariano statistic does not apply. Fortunately, it has been derived by Clark and McCracken (2002).

5. Forecasting with large datasets

Prediction with large datasets is an important recent development in econometrics. Suppose that we want to forecast inflation or growth h periods in the future and we have a set of predictors $\{x_{it}\}_{i=1}^{n}$. Intuitively, the idea is to try to combine the information in all of these predictors while avoiding estimating too many free parameters. There are many ways of setting up this idea.

1. The factor-augmented autoregression. We can take the first k principal components of $\{x_{it}\}_{i=1}^{n}$ that are assumed to be stationary, call these $f_{1t}, \ldots . f_{kt}$ and can then consider the regression

$$y_{t+h} = \alpha + \beta y_{t-1} + \Sigma_{i=1}^{k} \gamma_k f_{k,t-1} + \varepsilon_{t+h}$$

which can clearly be used for prediction. To form the principal components we first rescale the data to have mean zero and variance 1 (this means that units are arbitrary) and then perform the eigenvalue decomposition on its covariance matrix. Concretely, the Matlab code

```
for j=1:size(x,2); x(:,j)=(x(:,j)-mean(x(:,j)))/std(x(:,j)); end;
[coeff, score] = princomp(x);
```

will put the ordered principal components of the matrix x in the columns of score.

2. The factor-augmented VAR. This simply fits a VAR to $(y_t, f_{1t}, \ldots f_{kt})'$ and iterates this forward to provide forecasts of y_{t+h}.

3. Bayesian VARs. A Bayesian VAR will impose some prior on the parameters. This is often a deliberately informative prior, to combat overfitting and to shrink the model to some simple specification. Even if you don't adopt the Bayesian philosophy, this can still be a useful forecasting device. The Minnesota prior (covered earlier) is a good example. Banbura, Giannone and Reichlin (2009) find that it gives good forecasts.

4. Equal-weighted averaging. Suppose that we take models each of which is (say) of the form

$$y_{t+h} = \alpha + \beta y_{t-1} + \gamma x_{it-1} + \varepsilon_{t+h}$$

There are in total n such models, all of which are quite simplistic, and each of which gives a forecast. Now suppose that we simply average the forecasts from all of these models. The idea that simply *averaging* forecasts (with equal weights) works better than trying to estimate optimal weights is part of the folklore of forecasting (Bates and Granger (1969)) and is surprising, but true.

5. Bayesian model averaging. Suppose that there are n models, each of the form

$$y_i = Z\gamma + X_i\beta_i + \varepsilon_i$$

and pretend that the regressors are strictly exogenous and the errors are iid normal with mean zero and variance σ^2. So the parameter vector for model i is $\theta_i = (\gamma', \beta_i', \sigma^2)'$. Z is a matrix of regressors that are common to all the models. Without loss of generality, assume that Z and X_i are orthogonal (it is without loss of generality because one can always replace X_i by regression residuals).

Assume that the prior for β_i conditional on σ_i^2 is $N(0, \phi\sigma^2(X_i'X_i)^{-1})$ and that the prior for $(\gamma', \sigma^2)'$ is proportional to $1/\sigma^2$. The OLS estimate of β_i is

$$\hat{\beta}_i = (X_i'X_i)^{-1} X_i' y$$

and the OLS estimate of γ is $(Z'Z)^{-1}Z'y$.
Zellner (1971) shows that:

1. The posterior means are $\tilde{\beta}_i = \frac{\hat{\beta}_i \phi}{1+\phi}$ and $\hat{\gamma}$. The forecast from model i is $\tilde{y}_i = Z\hat{\gamma} + X_i\tilde{\beta}_i$.

2. The likelihood for model i, M_i, is

$$P(D \mid M_i) = \int P(D \mid \theta_i, M_i) P(\theta_i \mid M_i) d\theta_i \propto (\frac{1}{1+\phi})^{p_k/2} [\frac{SSR}{1+\phi} + \frac{\phi SSU}{1+\phi}]^{-(T-p)/2}$$

where SSR is the sum of squared residuals from the restricted regression of y on Z and SSU is the sum of squared residuals from the unrestricted regression of y on Z and X_i.

Now suppose that all the models are equally likely *a priori*, so that the model prior is $P(M_i) = 1/n$. Then the posterior probability that each model is true is given by

$$P(M_i \mid D) = \frac{P(D \mid M_i)P(M_i)}{\Sigma_{j=1}^{n} P(D \mid M_j)P(M_j)}$$

and the BMA forecast will be

$$\Sigma_{i=1}^{n} P(M_i \mid D)\tilde{y}_i$$

Now all the assumptions of strict exogeneity, iid errors etc. are unreasonable in a forecasting context, but this can be seen as just a pragmatic shrinkage device, which seems to work well in a number of applications.

Sometimes one takes just one variable at a time in each model. Or, one can use all possible permutations of variables. In this latter case, it may be appealing not to have a prior that all models are equally likely. One could instead take a set of K candidate regressors and consider as models all possible permutations of these variables (of which there are 2^K) but the prior that the ith model is correct is

$$P(M_i) = \rho^{p_i}(1-\rho)^{K-p_i}$$

where p_i is again the number of regressors in the ith model. The parameter ρ controls the model size. The expected number of regressors under the prior is ρK. With this prior BMA can easily be implemented, at least as long as K is not too big.

6. The dynamic factor model.

$$X_t = a_0 f_t + a_1 f_{t-1} \ldots + a_q f_{t-q} + \varepsilon_t$$

where f_t follows a VAR. This can be written as a static factor model where the factors are $F_t = (f_t', f_{t-1}' \ldots f_{t-q}')'$. We call f_t the dynamic factors and F_t the static factors. The model has the implication that $E_t(X_{t+h})$ is a linear function of F_t and the data. If this model is correct, estimating the static principal components and using those for forecasting will not be the most efficient available method. Intuitively, in this case, the factor augmented autoregression will be like OLS and alternative methods will be more efficient, like GLS. Forni, Hallin, Lippi and Reichlin (2005) give more specifics.

7. LASSO. This takes all the predictors but solves the problem

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} (Y - X\beta)'(Y - X\beta)$$

subject to the constraint

$$\Sigma_{j=1}^{p} \mid \beta_j \mid \leq c$$

Solving this problem will push some parameters towards zero and some all the way to zero. Hence it is the Least Absolute Shrinkage and Selection Operator. This can equivalently be written as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta}(Y - X\beta)'(Y - X\beta) + \lambda\Sigma_{j=1}^{p} | \beta_j |$$

It leaves the question of selecting the *tuning parameter*, c. There are a number of ways of doing this. A common one is cross-validation. This involves defining $\hat{\beta}_{LASSO(c,-i)}$ as the LASSO estimate leaving out observation i and using the tuning parameter c. We would then select the tuning parameter as:

$$\hat{c} = \arg\min_{c} \Sigma_{i=1}^{n}(y_i - x_i'\hat{\beta}_{LASSO(c,-i)})^2$$

In LASSO we nearly always standardize all the variables to have variance 1 and mean 0 and then run the regression without an intercept. Otherwise the penalty term will depend on the units in which regressors are measured.

If there is a single regressor, it turns out that LASSO has simple form:

$$\hat{\beta}_{LASSO} = \hat{\beta} - \frac{\lambda}{2} \text{ if } \hat{\beta} > \frac{\lambda}{2}$$

$$\hat{\beta} + \frac{\lambda}{2} \text{ if } \hat{\beta} < -\frac{\lambda}{2}$$

$$0 \text{ otherwise}$$

where $\hat{\beta}$ is the OLS estimator. This will hold for many regressors if the regressors are orthonormal.

It is tricky to do inference for LASSO. It is often done by a bootstrap holding the tuning parameter fixed.

The penalty term in LASSO is an L1-penalty. You could have an L2-penalty. This is called ridge regression. It shrinks parameters to zero, but never all the way, so there is no selection in ridge. The ridge regression has a simple closed form:

$$\hat{\beta}_{RIDGE} = (X'X + \lambda I)^{-1} X'Y$$

The elastic net combines both L1 and L2 penalties. LASSO, ridge and the elastic net are part of Machine Learning which is a very active area of research in statistics/econometrics. Machine Learning aims to use statistical algorithms to give good predictions in situations with many possible predictors.

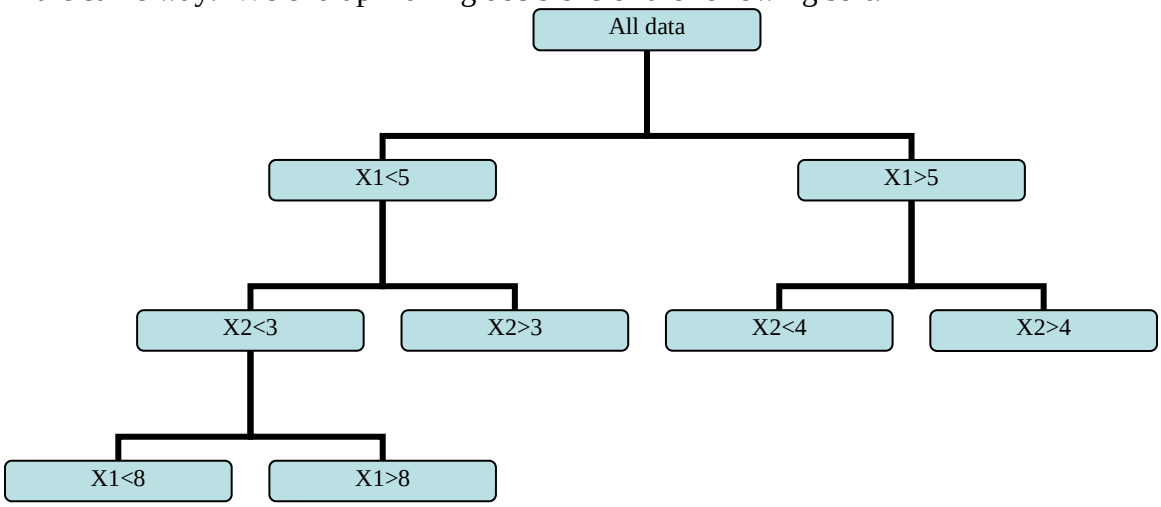
The basic idea at the root of all of these penalization estimators comes from Stein's paradox. This says that if θ is a vector of 3 or more parameters and X is an unbiased estimator of θ, then X is not the minimum mean square error estimator of θ. Imparting some bias (shrinkage towards zero) helps.

9. Regression Trees. This is a very different approach, also popular in the Machine Learning literature. There is no model or likelihood, unlike in any of the methods considered above. Rather it is a classification algorithm. It comes in many flavors, but here is the idea. Suppose that I have a dependent variable $\{y_i\}_{i=1}^{n}$ and two predictors $\{x_{1i}\}_{i=1}^{n}$ and $\{x_{2i}\}_{i=1}^{n}$.

I now search across all the variables and all possible cutoffs to solve the problem:

$$k^*, j^* = \arg\min_{j,k} \Sigma_{i=1}^{n}(y_i - \overline{y}_1 1(x_{ji} \le k) - \overline{y}_2 1(x_{ji} > k))^2$$

where $\bar{y}_1 = \frac{1}{n_1}\Sigma_{i=1}^{n} y_i 1(x_{ji} \le k)$ and $\bar{y}_2 = \frac{1}{n_2}\Sigma_{i=1}^{n} y_i 1(x_{ji} > k)$. That lets us split the data according to whether $x_{ji} \le k$ or $x_{ji} > k$. Then at each node, we consider splitting the data based into two in the same way. We end up making decisions of the following sort:



Each of the terminal nodes is then a "leaf" of the tree. The value of the dependent variable is then averaged for each leaf. When making a prediction, based on the explanatory variables you determine which leaf the observation belongs to, and the forecast is just the average of the dependent variable for that leaf. We can continue until each leaf has reached a minimum size.

This will tend to produce too many leaves. Machine learning people would call this algorithm "greedy". You want to have a way of stopping, and there are many options:
1. You could use cross validation to estimate the minimum leaf size and then use this as your chosen minimum leaf size.
2. You could stop when the p-value of the significance of the difference between two leaves is above 0.05.
3. You could split the sample into a training sample and a testing sample. Use the training sample to construct the tree. But then for each node, evaluate whether you would be better off merging them based on prediction in the testing sample. This process is called pruning.

6. What Forecasts Economic Activity and Inflation?

There is of course an enormous literature trying to find predictors that do a good job in forecasting real economic activity and/or inflation. These are mostly asset prices. In contrast to the "large dataset" methods that have been popular in recent econometric work, these methods try to find a single model that is hoped to be helpful for prediction.

Economic Activity.

An old idea is to use the slope of the term structure to forecast future economic activity. If recessions are times that the Fed tightens monetary policy to create economic slack in order to disinflate, then an inverted yield curve should presage a recession. At least up the mid 1980s that seemed to be true, in the sense that the more the yield curve sloped down, the lower growth would subsequently be.

A more recent idea is to use corporate spreads to forecast growth. A recent paper on this is Gilchrist, Yankov and Zakrajsek (2008). They find that expected default rates implicit in corporate bond spreads have considerable predictive power for future real economic activity (they use nonfarm payrolls and industrial production).

Inflation

The oldest approach is to use a Phillips curve relationship for predicting inflation

$$\pi_{t+1} = \alpha + a(L)\pi_t + b(L)X_t + \varepsilon_{t+1}$$

where X_t represents some measure(s) of real activity, traditionally the unemployment rate. Or Gali and Gertler (1999) prefer to use marginal cost. An alternative approach is to use the slope of the term structure. However, Ang, Bekaert and Wei (2007) argue convincingly that surveys do a better job of forecasting inflation than either. It seems that for inflation there are occasional breaks in the level of inflation that standard regressions have a hard time capturing, but that a judgmental forecast can pick up better.

6. Density Forecasts

Density forecasts are also worth a mention. Any point forecast coupled with an estimate of the variance (and a distributional assumption) naturally implies a density forecast. If the variance is constant, this density forecast is a rather silly one, in that the percentiles of the density forecast are moving in lockstep with the point forecast. But the volatility could be specified to follow a GARCH process, or a stochastic volatility model, giving a genuine density forecast.

Quantile regression offers an alternative. Quantile regression has increased in popularity in cross-sectional econometrics, but is also being used a little in forecasting. A conventional

regression estimates the *mean* of Y, conditional on X. A quantile regression instead estimates the τ th *quantile* of Y conditional on X. It is still linear, so the model is of the form

$$y_t(\tau) = \beta_0 + \beta_1 x_{1t} + ...\beta_p x_{pt} = \beta' x_t$$

where $y_t(\tau)$ denotes the τ th quantile of Y. It can be estimated by

$$\hat{\beta} = \arg\min_\beta \Sigma_{t=1}^T \rho_\tau (y_t - \beta' x_t)$$

where $\rho_\tau(z) = z(\tau - 1(z<0))$. In the special case $\tau = 1/2$

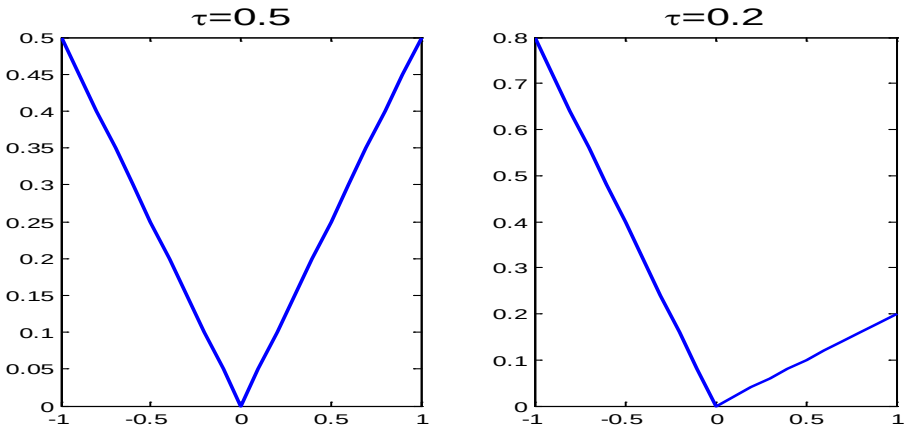
$$\hat{\beta} = \arg\min_\beta \Sigma_{t=1}^T \{y_t - \beta' x_t\}\{\frac{1}{2} - 1(y_t - \beta' x_t < 0)\}$$

$$\therefore \hat{\beta} = \arg\min_\beta \frac{1}{2}\Sigma_{t=1}^T \{y_t - \beta' x_t\}1(y_t - \beta' x_t > 0) - \frac{1}{2}\Sigma_{t=1}^T \{y_t - \beta' x_t\}1(y_t - \beta' x_t < 0)$$

$$\therefore \hat{\beta} = \arg\min_\beta \Sigma_{t=1}^T |y_t - \beta' x_t|$$

which is a least absolute deviations (LAD) estimator. This estimates the *median* of Y conditional on X. But quantile regression allows us to estimate all the percentiles of Y conditional on X, and hence to obtain an estimate of the conditional density.[1]

In general the function $\rho_\tau(.)$ is piecewise linear and is known as the "check" function. We can easily draw the function $\rho_\tau(.)$ for different choices of the quantile.



Some of the density forecasts that get attention in macroeconomics are judgmental forecasts. The SPF has asked respondents to place probabilities on GDP growth and inflation falling in certain bins four time a year for a long time. There are other similar survey density forecasts.

[1] If we assume that the data are generated by a linear homoskedastic regression model in which the error density is iid with density $f(.)$, then $T^{1/2}(\hat{\beta} - \beta) \to_d N(0, \frac{\tau(1-\tau)}{f(0)^2} M^{-1})$ where $M = \text{plim} T^{-1}\Sigma_{t=1}^T x_t x_t'$. The iid assumption is unreasonable in most contexts—but more general standard errors are available. Note that the GMM distribution theory does *not* apply directly because the objective function is not differentiable.

It is natural to want to evaluate a density forecast. whether obtained from a model or judgment. The technology for doing this was provided in Diebold, Gunter and Tay (1998) and other papers. The idea is simple, using the probability integral transform result that you learned in statistics last year. If I have a sequence of density forecasts, and observe $\{Z_t\}_{t=1}^T$–the time series of the *percentiles* of where the realizations were observed in the *ex-ante* forecast density, then $\{Z_t\}_{t=1}^T$ should be uniform on the unit interval and should be iid (if the forecasts are non-overlapping). So for example, this means that the outcome should be less than the forecast median half the time and more than the forecast median half the time, and which it is should be purely random over time. And moreover, Z_t should be orthogonal to *everything* that was known at the time that the forecast was made.

Here are two specific tests:

● Let $\hat{F}(r) = \frac{1}{r}\Sigma_{t=1}^T 1(Z_t \le r)$ denote the empirical cdf of $\{Z_t\}_{t=1}^T$. Let $F(r)$ be the standard uniform cdf ($F(r) = r$). Under the null hypothesis that the density forecast is correctly specified, $\hat{F}(r)$ should be "close" to $F(r)$. This can be assessed by a Kolmogorov-Smirnov test. The test statistic is

$$KS = \sup_{0 \le r \le 1} |\hat{F}(r) - F(r)|$$

Under the null hypothesis

$$T^{1/2} KS \rightarrow_d \sup_{0 \le r \le 1} |B(r) - rB(1)|$$

where $B(r)$ is a standard Brownian motion (and so $B(r) - rB(1)$ is a Brownian bridge).

● We can test the autocorrelation of $\{Z_t\}_{t=1}^T$.

An idea that combines these two is to use the Berkowitz LR test (Berkowitz (2000)). If the density is correctly specified then $\{Z_t\}_{t=1}^T$ is iid uniform. If Φ denotes the standard normal cdf, and $\eta_t = \Phi^{-1}(Z_t)$ then $\{\eta_t\}_{t=1}^T$ must be iid standard normal. Now consider the AR(1)

$$\eta_t = \phi_0 + \phi_1 \eta_{t-1} + v_t$$

Under the null, $\phi_0 = 0$, $\phi_1 = 1$ and the variance of v_t is 1. We can do a joint test of the three hypotheses by a LR test. If the log-likelihood function is $l(\phi_0, \phi_1, \sigma_v^2)$ then the test statistic is

$$2[l(\hat{\phi}_0, \hat{\phi}_1, \hat{\sigma}_v^2) - l(0,0,1)]$$

For point forecasts, we considered root-mean-square prediction error as a metric of forecast quality. For density forecasts, a natural analog is the predictive likelihood—where in the forecast density the actual realization occurred. Obviously one would like this to be as big as possible. If $f_t(.)$ is the forecast density at time t for y_{t+1} and T^* denotes the number of out-of-sample periods, the predictive likelihood is

$$\Sigma_{t=1}^{T^*} \log f_t(y_{t+1})$$

If we have two competing density forecasts $f_{1,t}(.)$ and $f_{2,t}(.)$, we can compare their predictive likelihoods by defining $\xi_{1,t} = \log f_{1,t}(y_{t+1})$ and $\xi_{2,t} = \log f_{2,t}(y_{t+1})$ and then considering the test statistic

$$\frac{T^{-1}\Sigma_{t=1}^{T^*}(\xi_{1,t+1} - \xi_{2,t+1})}{\sqrt{T^{-1}\Sigma_{t=1}^{T^*}(\xi_{1,t+1} - \xi_{2,t+1})^2} / \sqrt{T}}$$

This is just like the Diebold-Mariano statistic, but replacing squared errors from a point forecast with realizations of the predictive density. See Amisano and Giacomini (2007).

Econ 607 Jonathan Wright
Fall 2023

Handout on Unit Roots, Spurious Regressions and Cointegration

The simplest time series model is an AR(1)

$$y_t = \alpha y_{t-1} + u_t$$

In the case $|\alpha| < 1$, we have

$$T^{1/2}(\hat{\alpha} - \alpha) \to_d N(0, 1-\alpha^2)$$

But this breaks down in the case $\alpha = 1$, which is a random walk. The "knife-edge" case where α is exactly equal to one arguably isn't that interesting *per se*. More importantly, this result doesn't work well unless the sample size is enormous if α is close to, but less than, 1.

Definitions for Nonstationary time series

A time series is a **random walk** if $y_t = y_{t-1} + u_t$ where u_t is iid.

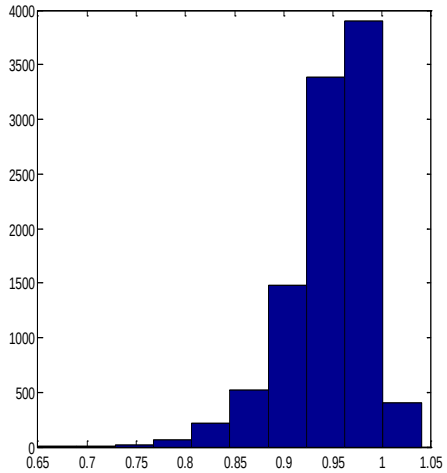
A time series is a **martingale** if $E_{t-1}(y_t) = y_{t-1}$.

A time series is a **martingale difference sequence** if $E_{t-1}(y_t) = 0$.

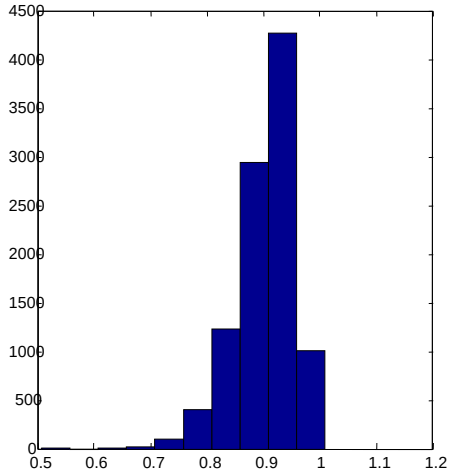
A time series is (weakly) **stationary** if it's first two moments exist and do not change over time.
A time series is **invertible** if it can be written as an autoregression.
A time series is **I(0)** if it is both stationary and invertible.
A time series is **I(d)** if its' dth differences are I(0). If a time series is I(1), it is said to have a unit root.
An **ARIMA(p,d,q)** model is a time series the dth differences of which form a stationary and invertible ARMA(p,q) model.

Simulated distribution of OLS estimator of α if $T = 100$

α=1



α=0.95

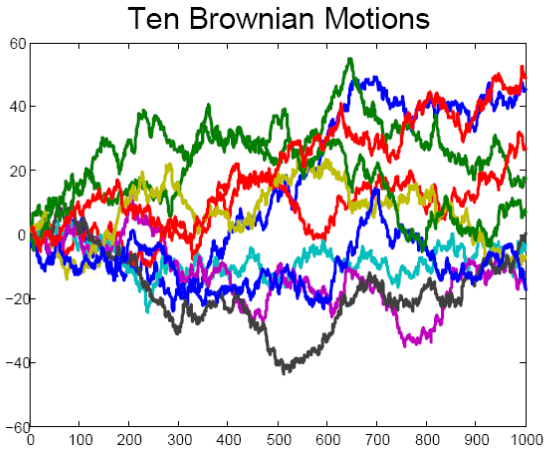


Both are skewed to the left.

Normality doesn't work and for this and many non-standard problems in econometrics, we need to introduce new tools; Brownian motion and a functional central limit theorem.

The stochastic process $B(t)$ is a **Brownian motion** if

1. $B(0) = 0$
2. $B(t) - B(s) \sim N(0, \sigma^2(t-s))$ for any $t > s$
3. If $t_1 < t_2 \le t_3 < t_4$ then
   $B(t_2) - B(t_1)$ is independent of $B(t_4) - B(t_3)$



Brownian motion properties:

- A Brownian motion is a martingale
- $Cov(B(s), B(t)) = \min(s,t)$, for a "standard" Brownian motion ($\sigma^2 = 1$)

Basic functional central limit theorem.

Suppose that $\varepsilon_1, \varepsilon_2, ...\varepsilon_T$ are iid with mean 0 and variance σ^2 and $2+\delta$ finite moments. Let $S_t = \Sigma_{s=1}^{t} \varepsilon_s$. Define the function $S_T(r) = \frac{1}{T^{1/2}\sigma} S_{[Tr]}$, where [.] denotes the integer part of the argument.

The functional central limit theorem says that

$S_T(r) \Rightarrow B(r)$

where "$\Rightarrow$" means convergence in distribution uniformly in r and $B(r)$ is a standard Brownian motion on the unit interval.

Functional central limit theorem with non-iid errors.

Suppose that $\varepsilon_1, \varepsilon_2, ...\varepsilon_T$ are stationary with mean 0 and (average) zero-frequency spectral density ω^2 satisfying suitable conditions. Let $S_t = \Sigma_{s=1}^{t}\varepsilon_s$. Define the function $S_T(r) = \frac{1}{T^{1/2}\omega}S_{[Tr]}$. Then $S_T(r) \Rightarrow B(r)$.

Continuous mapping theorem Suppose that $X_T \rightarrow_d X$ (uniformly in r, if applicable). Then $f(X_T) \rightarrow_d f(X)$ where $f(.)$ is any continuous function.

Derivation of the limiting distribution of OLS for the random walk model

Suppose that $y_t = \alpha y_{t-1} + u_t$ with $\alpha = 1$ and $y_0 = 0$

$y_{t-1} = TS_{t-1}\int_{(t-1)/T}^{t/T} dr = T\int_{(t-1)/T}^{t/T} S_{[Tr]}dr = T^{3/2}\sigma\int_{(t-1)/T}^{t/T} S_T(r)dr$

$y_{t-1}^2 = TS_{t-1}^2\int_{(t-1)/T}^{t/T} dr = T\int_{(t-1)/T}^{t/T} S_{[Tr]}^2dr = T^2\sigma^2\int_{(t-1)/T}^{t/T} S_T(r)^2 dr$

$\therefore \Sigma_{t=1}^{T}y_{t-1} = T^{3/2}\sigma\int_0^1 S_T(r)dr$ and $\Sigma_{t=1}^{T}y_{t-1}^2 = T^2\sigma^2\int S_T^2(r)dr$

$\therefore T^{-3/2}\Sigma_{t=1}^{T}y_{t-1} = \sigma\int_0^1 S_T(r)dr \Rightarrow \sigma\int_0^1 B(r)dr$ and $T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2 = \sigma^2\int_0^1 S_T(r)^2 dr \Rightarrow \sigma^2\int_0^1 B(r)^2 dr$

But $\hat{\alpha} - 1 = \frac{\Sigma_{t=1}^{T}y_{t-1}u_t}{\Sigma_{t=1}^{T}y_{t-1}^2}$

$y_{t-1}u_t = \frac{1}{2}\{y_{t-1}(y_t - y_{t-1}) + (y_t - u_t)u_t\} = \frac{1}{2}\{y_{t-1}y_t - y_{t-1}^2 + y_tu_t - u_t^2\} = \frac{1}{2}\{y_t^2 - y_{t-1}^2 - u_t^2\}$

$\therefore \hat{\alpha} - 1 = \frac{\frac{1}{2}\{y_T^2 - \Sigma_{t=1}^{T}u_t^2\}}{\Sigma_{t=1}^{T}y_{t-1}^2}$

$\therefore T(\hat{\alpha} - 1) = \frac{\frac{1}{2}\{T^{-1}y_T^2 - T^{-1}\Sigma_{t=1}^{T}u_t^2\}}{T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2} = \frac{\frac{1}{2}\{[T^{-1/2}y_T]^2 - T^{-1}\Sigma_{t=1}^{T}u_t^2\}}{T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2}$

$T^{-1/2}y_T \Rightarrow \sigma B(1)$, $T^{-1}\Sigma_{t=1}^{T}u_t^2 \rightarrow_p \sigma^2$ and $T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2 \Rightarrow \sigma^2\int_0^1 B(r)^2 dr$

Combining these pieces, $T(\hat{\alpha} - 1) \rightarrow_d \frac{\frac{1}{2}\{B(1)^2 - 1\}}{\int_0^1 B(r)^2 dr}$

Notice that

- The scaling is by T not $T^{1/2}$. So the estimator is "superconsistent"
- The distribution is not normal
- We could also write the distribution of the numerator as $T^{-1}\Sigma_{t=1}^{T}y_t\Delta y_t \rightarrow_d \sigma^2\int_0^1 B(r)dB(r)$.

Consistency of the two results arises from $\int_0^1 B(r)dB(r) = \frac{1}{2}B(1)^2 - \frac{1}{2}$ (special case of Ito's lemma).

The t-statistic testing the hypothesis that $\alpha = 1$ is

$$\frac{(\hat{\alpha} - 1)}{s}\{\Sigma_{t=1}^{T} y_{t-1}^2\}^{-1/2}$$

where $s^2 = T^{-1}\Sigma(y_t - \hat{\alpha} y_{t-1})^2$. The t-statistic converges in distribution to $\frac{\frac{1}{2}\{B(1)^2 - 1\}}{\sqrt{\int_0^1 B(r)^2 dr}}$ under the null. This is the Dickey-Fuller distribution. It is common to re-write this as a t-test in the equation

$$\Delta y_t = (\alpha - 1) y_{t-1} + u_t$$

The critical values for a 5 percent one-sided test rejecting if $\alpha < 1$ is -1.95. If an intercept is included; the critical values is -2.89.

Often, we want to test the hypothesis that $\alpha = 1$ where the error term is not iid under the null, and instead follows an AR(p). We can then

$$\Delta y_t = (\alpha - 1) y_{t-1} + \Sigma_{j=1}^{p} \phi_j \Delta y_{t-j} + u_t$$

and again test the hypothesis that the first coefficient is equal to zero. This is called an Augmented Dickey-Fuller test. The distribution is again the Dickey-Fuller distribution.

Nelson and Plosser (1982) was a seminal paper testing for unit roots in macroeconomic time series , finding that the hypothesis of a unit root could not be rejected for 13 out of 14 time series considered.

Initially there were two reasons why the profession was excited about whether there were unit roots or not. First, it was thought to tell us what the main sources of business cycle fluctuations were. If there is a unit root in real output, a shock today lasts infinitely far into the future, which may suggest that it is a technology shock and if there is little persistence, the shocks must come from monetary policy and fiscal shocks. West (1988) and Christiano and Eichenbaum argue compellingly against this view.

Second, it was thought that the way we should do inference about relations among time series depends on whether they have unit roots. That part still seems right, but the same issues crop up if time series are very persistent but do not have exact unit roots. And it is impossible to tell these two apart in the sample sizes that are actually available.

Beveridge-Nelson Decomposition

Suppose that x_t is a random walk, u_t is stationary and $y_t = kx_t + u_t$. The process y_t is I(1), but the "degree" of nonstationarity depends on k. The Beveridge-Nelson decomposition is a way to decompose any I(1) series into random walk and stationary pieces.

B-N decomposition: If y_t is any series such that $\Delta y_t = C(L)\varepsilon_t$ then

$$y_t = C(1)\Sigma_{s=1}^{t}\varepsilon_s + C^*(L)\varepsilon_t$$

where $C^*(L) = \Sigma_{j=0}^{\infty} c_j^* L^j$ and $c_j^* = -\Sigma_{i=j+1}^{\infty} c_i$. The first piece is a random walk; the second piece is stationary.

Proof: $y_t = \Sigma_{s=1}^{t}\Delta y_s = \Sigma_{s=1}^{t} C(L)\varepsilon_s$

$\therefore y_t = (c_0\varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2}..) + (c_0\varepsilon_{t-1} + c_1\varepsilon_{t-2}...)..$

$= (\Sigma_{j=0}^{\infty} c_j - \Sigma_{j=1}^{\infty} c_j)\varepsilon_t + (\Sigma_{j=0}^{\infty} c_j - \Sigma_{j=2}^{\infty} c_j)\varepsilon_{t-1}...$

$= [C(1) + c_0^*]\varepsilon_t + [C(1) + c_1^*]\varepsilon_{t-1}...$

$= C(1)\Sigma_{s=1}^{t}\varepsilon_s + C^*(L)\varepsilon_t$

The spectral density of Δy_t at frequency zero is $\frac{\sigma^2}{2\pi}C(1)^2$. So the "size" of the random walk piece is proportional to the zero-frequency spectral density of the first differences. A key point is that a time series may be I(1) yet have only a very small random walk component.

Spurious Regressions

Suppose that x_t and y_t are two *unrelated* random walks. In a regression of one on the other, the coefficient is likely to be significant and the R-squared is likely to be high. But there is in fact no relation between the series! It's called a spurious regression.

Yule (1927) and Granger and Newbold (1974) found this in simulations. Phillips (1986) gave some analytical results. Suppose that

$x_t = x_{t-1} + u_{t,x}$, $y_t = y_{t-1} + u_{t,y}$

and the shocks are iid and mutually independent with variances σ_x^2 and σ_y^2, respectively. Let

$$X_T(r) = \frac{1}{T^{1/2}\sigma_x} x_{[Tr]} \text{ and } Y_T(r) = \frac{1}{T^{1/2}\sigma_y} y_{[Tr]}.$$

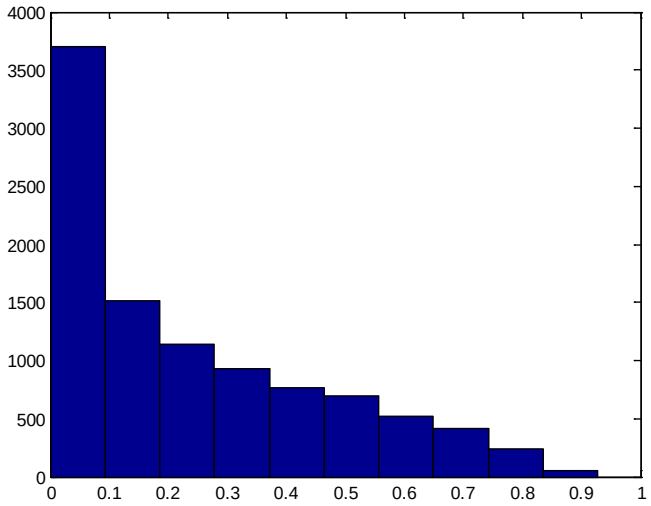
Then $X_T(r) \Rightarrow U(r)$ and $Y_T(r) \Rightarrow V(r)$ where these are two independent Brownian motions.

We have $\hat{\beta} = \frac{\Sigma_{t=1}^{T} x_t y_t}{\Sigma_{t=1}^{T} x_t^2} = \frac{T^{-2}\Sigma_{t=1}^{T} x_t y_t}{T^{-2}\Sigma_{t=1}^{T} x_t^2} \Rightarrow \frac{\sigma_y \int U(r)V(r)dr}{\sigma_x \int U(r)^2 dr}$

$$R^2 = 1 - \frac{\Sigma_{t=1}^{T}(y_t - \hat{\beta}x_t)^2}{\Sigma_{t=1}^{T} y_t^2} = 1 - \frac{\Sigma_{t=1}^{T} y_t^2 - 2\hat{\beta}\Sigma_{t=1}^{T} x_t y_t + \hat{\beta}^2\Sigma_{t=1}^{T} x_t^2}{\Sigma_{t=1}^{T} y_t^2} = \frac{2\hat{\beta}\Sigma_{t=1}^{T} x_t y_t - \hat{\beta}^2\Sigma_{t=1}^{T} x_t^2}{\Sigma_{t=1}^{T} y_t^2}$$

$$\therefore R^2 \to_d \frac{2\frac{\sigma_y \int U(r)V(r)dr}{\sigma_x \int U(r)^2 dr}\sigma_x\sigma_y \int U(r)V(r)dr - \{\frac{\sigma_y \int U(r)V(r)dr}{\sigma_x \int U(r)^2 dr}\}^2 \sigma_x^2 \int U(r)^2 dr}{\sigma_y^2 \int V(r)^2 dr} = \frac{[\int U(r)V(r)dr]^2}{\int U(r)^2 dr \int V(r)^2 dr}$$

This is a simulation of this limiting distribution. The mean is a bit above 0.2.



In a similar way, it can be shown that the Durbin-Watson statistic converges in probability to zero. A problem is that spurious regressions can arise if the time series do not have unit roots, but are just very persistent. A context where this comes up is a long-horizon predictive regression in finance where an h-period return is regressed on some predictor, like a dividend-yield.

Cointegration

The fact that two time series have unit roots, does not mean that a relationship between them is a spurious regression. It is also possible that they are cointegrated.

Cointegration: Approximate Definition. Two drunks are walking in the park, but are tied together with a piece of string. Each is a random walk. The two drunks are cointegrated.

More formal definition. Two nonstationary time series x_t and y_t are said to be cointegrated if they are both I(1) but if there is exists some linear combination $u_t = y_t - \beta x_t$, for $0 < k < \infty$, that is I(0).

We can rewrite the definition of cointegration as $y_t = \beta x_t + u_t$ where the regressor is I(1) and the error term is I(0). This model has intriguing statistical properties

- OLS is superconsistent (meaning $T(\hat{\beta} - \beta)$ converges to a distribution, that is a function of Brownian motions).
- If x_t is strictly exogenous (independent of the error at all leads and lags), then t- and F-statistics associated with OLS have their usual normal and χ^2 limiting distributions.
- If x_t is not strictly exogenous, there are estimators other than OLS such that t- and F-statistics have normal and χ^2 limiting distributions. A popular choice is dynamic OLS which estimates the relationship

$$y_t = \beta x_t + d(L)\Delta x_t + u_t$$

where $d(L)$ is a two-sided polynomial (Stock and Watson (1993)). Another choice is the maximum likelihood estimator proposed by Soren Johansen that is implemented in Eviews.

More econometric detail on cointegration.

A cointegrating system can be written in "triangular form"

$y_t = \beta x_t + \varepsilon_{1t}$

$\Delta x_t = \varepsilon_{2t}$

where ε_{1t} and ε_{2t} are I(0). Suppose that $S_{1t} = \Sigma_{s=1}^{t}\varepsilon_{1s}$ and $S_{2t} = \Sigma_{s=1}^{t}\varepsilon_{2s}$. Let $S_{1T}(r) = \frac{1}{T^{1/2}\sigma_1} S_{1[Tr]}$ and $S_{2T}(r) = \frac{1}{T^{1/2}\sigma_1} S_{2[Tr]}$. Suppose further that $S_{1T}(r) \Rightarrow W_1(r)$ and $S_{2T}(r) \Rightarrow W_2(r)$ where $W_1(r)$ and $W_2(r)$ are two standard Brownian motions.

The OLS coefficient in a cointegrating regression has the distribution

$$T(\hat{\beta} - \beta) \rightarrow \frac{\sigma_1 \int_0^1 W_2(r) dW_1(r)}{\sigma_2 \int_0^1 W_2(r)^2 dr}$$

because $T^{-1}\Sigma_{t=1}^{T} x_t \varepsilon_{1t} \rightarrow_d \sigma_1\sigma_2 \int_0^1 W_2(r) dW_1(r)$ and $T^{-2}\Sigma_{t=1}^{T} x_t^2 \rightarrow_d \sigma_2^2 \int_0^1 W_2(r)^2 dr$. The null distribution of the t-statistic is

$$\frac{\int_0^1 W_2(r) dW_1(r)}{\sqrt{\int_0^1 W_2(r)^2 dr}}$$

Under strict exogeneity, the two Brownian motions are independent, and so this distribution reduces to normal. Otherwise it is a nonstandard distribution. Likewise, under strict exogeneity, the distribution of an F-statistic is χ^2. Otherwise it is a function of Brownian motions.

Now consider three cases

Case (a) ε_{1t} and ε_{2t} are iid and mutually uncorrelated. In this case x_t is strictly exogenous. So OLS allows standard inference.

Case (b) ε_{1t} and ε_{2t} are iid but mutually correlated. Let $\tilde{\varepsilon}_{1t} = \varepsilon_{1t} - d\varepsilon_{2t}$ where $d = Cov(\varepsilon_{1t}, \varepsilon_{2t}) / Var(\varepsilon_{2t})$. Then

$y_t = \beta x_t + d\Delta x_t + \tilde{\varepsilon}_{1t}$

$\Delta x_t = \varepsilon_{2t}$

and now x_t is strictly exogenous. Dynamic OLS will work; standard OLS will not give standard inference (though it will still be superconsistent).

Case (c) ε_{1t} and ε_{2t} are just I(0). Let $\tilde{\varepsilon}_{1t} = \varepsilon_{1t} - d(L)\varepsilon_{2t}$ where $d(L)$ is such that $\tilde{\varepsilon}_{1t}$ and ε_{2t} are orthogonal at all leads and lags. Then

$y_t = \beta x_t + d(L)\Delta x_t + \tilde{\varepsilon}_{1t}$

$\Delta x_t = \varepsilon_{2t}$

and again x_t is strictly exogenous.

The Bottom Line

If two I(1) variables are cointegrated, then we would want to estimate their relationship by OLS or dynamic OLS.

On the other hand, if they are *not* cointegrated, we would be concerned about a spurious regression and would want to estimate the relationship in first differences, or transforming it in some other suitable way.

In principle, the way to tell if the I(1) variables y_t and x_t are cointegrated or not is to apply a unit root test to the residuals from the regression of y_t on x_t. The null hypothesis is of a unit root; and so of *no* cointegration. But the critical values are different from the ordinary unit root test (tabulated by Engle and Yoo (1987)).

If two series are I(1) and are cointegrated, then they have an error correction representation:

$$a(L)\Delta y_t = c + b(L)\Delta x_t + \alpha(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + \varepsilon_t$$

This is intuitive as the model corrects some of the cointegrating error. One way of estimating this model is to first estimate the cointegrating regression (i.e. β_0 and β_1) and then estimate the error correction model by regressing Δy_t on lags of Δy_t, Δx_t and lags of Δx_t and $y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 x_{t-1}$. A system with an error correction model for Δy_t and an analogous equation for Δx_t is a vector error correction model (VECM). A VECM is a restricted VAR.

Near Unit Roots

However, the problems associated with unit roots do not arise just in the knife-edge case of an exact unit root. The downward bias in an AR model was well known by 1950, long before unit roots. Yet, it is essentially the same problem.

A device that gives a better approximation to the small sample behavior of estimators and test statistics with near unit roots is to specify that in the AR(1) $y_t = \alpha_T y_{t-1} + u_t$, $\alpha_T = 1 + c/T$. The parameter depends on the sample size, which is a fiction, but it is a useful fiction if it provides a good approximation to the distribution of estimators and test statistics.

In this case, if $Y_T(r) = \frac{1}{T^{1/2}\sigma} y_{[Tr]}$, then $Y_T(r) \Rightarrow J_c(r)$ where

$$dJ_c(r) = cJ_c(r)dr + dB(r)$$

is an Ornstein-Uhlenbeck process. It is the continuous-time analog of an AR(1). If in fact $\alpha_T = 1 + c/T$, but we test for an exact unit root, the limiting distribution of the t-statistic testing the hypothesis that $\alpha = 1$ is

$$t \to_d \frac{\frac{1}{2}(J_c(1)^2 - 1)}{\{\int_0^1 J_c(r)^2 dr\}^{1/2}}$$

Proof: $\hat{\alpha}-1=\frac{\Sigma_{t=1}^{T}y_{t-1}\Delta y_t}{\Sigma_{t=1}^{T}y_{t-1}^2}$

$$y_{t-1}\Delta y_t=\frac{1}{2}\{y_{t-1}(y_t-y_{t-1})+(y_t-\Delta y_t)\Delta y_t\}=\frac{1}{2}\{y_{t-1}y_t-y_{t-1}^2+y_t\Delta y_t-\Delta y_t^2\}=\frac{1}{2}\{y_t^2-y_{t-1}^2-\Delta y_t^2\}$$

$$\therefore \hat{\alpha}-1=\frac{\frac{1}{2}\{y_T^2-\Sigma_{t=1}^{T}\Delta y_t^2\}}{\Sigma_{t=1}^{T}y_{t-1}^2}$$

$$\therefore T(\hat{\alpha}-1)=\frac{\frac{1}{2}\{T^{-1}y_T^2-T^{-1}\Sigma_{t=1}^{T}\Delta y_t^2\}}{T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2}=\frac{\frac{1}{2}\{[T^{-1/2}y_T]^2-T^{-1}\Sigma_{t=1}^{T}\Delta y_t^2\}}{T^{-2}\Sigma_{t=1}^{T}y_{t-1}^2}$$

$$T^{-1}\Sigma_{t=1}^{T}\Delta y_t^2=T^{-1}\Sigma_{t=1}^{T}(y_t-y_{t-1})^2=T^{-1}\Sigma_{t=1}^{T}(u_t+\frac{c}{T}y_{t-1})^2=T^{-1}\Sigma_{t=1}^{T}u_t^2+c\frac{1}{T^3}\Sigma_{t=1}^{T}y_{t-1}^2+2\frac{c}{T^2}\Sigma_{t=1}^{T}y_{t-1}u_t$$

$$\therefore T^{-1}\Sigma\Delta y_t^2\rightarrow_p \sigma^2$$

$$\therefore T(\hat{\alpha}-1)\rightarrow_d \frac{\frac{1}{2}(J_c(1)^2-1)}{\int_0^1 J_c(r)^2\,dr}$$

$$\therefore t\rightarrow_d \frac{\frac{1}{2}(J_c(1)^2-1)}{\{\int_0^1 J_c(r)^2\,dr)^{1/2}}$$

So we can form a confidence interval for c by inverting the acceptance region of this test (Stock (1991)). The idea is that we form a grid of values of c. Test each, by comparing the observed test statistic with the critical values from this distribution. The set of values of c for which this test accepts (the *acceptance region of the test*) is the confidence set for c. Then divide by the sample size and add 1 to get the corresponding confidence set for α.

Econ 607 Jonathan Wright
Fall 2023

Handout on Conditional Heteroskedasticity

It is very common for macroeconomic and especially finance time series to exhibit bursts of volatility. Modeling this seems important for forecasting and other purposes. The original model was autoregressive conditional heteroskedasticity (ARCH) which specifies that

$$r_t = \mu + \sigma_t \varepsilon_t$$
$$\sigma_t^2 = \omega + \alpha(r_{t-1} - \mu)^2$$
$$\sigma_1^2 = \omega / (1-\alpha)$$

where ε_t is iid standard normal. We can compute the kurtosis of this process as follows:

$$E(\sigma_t^4) = \omega^2 + 3\alpha^2 E(\sigma_t^4) + 2\alpha\omega E(\sigma_t^2) = \omega^2 + 3\alpha^2 E(\sigma_t^4) + 2\alpha\omega^2 / (1-\alpha) = 3\alpha^2 E(\sigma_t^4) + \omega^2[\frac{1+\alpha}{1-\alpha}]$$

$$\frac{E(\sigma_t^4 \varepsilon_t^4)}{E(\sigma_t^2 \varepsilon_t^2)^2} = \frac{3E(\sigma_t^4)}{\omega^2 / (1-\alpha)^2} = \frac{3\omega^2 \frac{1+\alpha}{1-\alpha}\frac{1}{1-3\alpha^2}}{\omega^2 / (1-\alpha)^2} = \frac{3\omega^2(1+\alpha)\frac{1}{1-3\alpha^2}}{\omega^2/(1-\alpha)} = \frac{3\omega^2(1+\alpha)(1-\alpha)}{\omega^2(1-3\alpha^2)} = \frac{3(1-\alpha^2)}{(1-3\alpha^2)} > 3$$

As a result we see that not only does this model allow for bursts of volatility, but it also accounts for fat tails.

Estimation is fairly easy by maximum likelihood as the log-likelihood function is (apart from a constant):

$$-\frac{1}{2}\Sigma_{t=1}^T \log(\sigma_t^2) - \frac{1}{2}\Sigma_{t=1}^T (\frac{r_t - \mu}{\sigma_t^2})$$

and can be numerically maximized with respect to the parameters α, μ and ω.

A test for the null hypothesis that $\alpha = 0$ is obtained from the R^2 in a regression of $(r_t - \bar{r})^2$ on $(r_{t-1} - \bar{r})^2$. Under the null, $T * R^2$ has a $\chi^2(1)$ limiting distribution. This test has a Lagrange multiplier interpretation. It has the useful feature that the parameters of the model do not need to be estimated.

The model has been extended in a great many ways. Three in particular are:
(i) Generalized ARCH (GARCH). A GARCH(p,q) model is

$$r_t = \mu + \sigma_t \varepsilon_t$$
$$\sigma_t^2 = \omega + \Sigma_{i=1}^q \alpha_i (r_{t-i} - \mu)^2 + \Sigma_{i=1}^p \beta_i \sigma_{t-i}^2$$

and this can also be estimated by maximum likelihood.

(ii) GARCH in mean

$$r_t = \mu + \lambda\sigma_t + \sigma_t \varepsilon_t$$

$$\sigma_t^2 = \omega + \Sigma_{i=1}^{q} \alpha_i (r_{t-i} - \mu)^2 + \Sigma_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

(iii) Exponential GARCH

$$r_t = \mu + \sigma_t \varepsilon_t$$

$$\log \sigma_t^2 = \omega + \alpha \log(\sigma_{t-1}^2) + \beta[\theta \varepsilon_t + (|\varepsilon_t| - E(|\varepsilon_t|))]$$

Since ε_t is standard normal, $E(|\varepsilon_t|) = \sqrt{2/\pi}$. This model is particularly useful for representing stock returns, because they not only show burst of volatility, but volatility tends to rise when returns are low. This model can capture a skewness effect of this sort.

Any number of extensions to this framework have been proposed. One can add in explanatory variables, or have nonlinear or multivariate specifications. A general challenge is ensuring that the variances remain positive. Of course, any parameterization which gives negative variances will in turn have a likelihood of minus infinity.

Stochastic volatility.

Stochastic volatility is a different type of model, which arguably fits the data better, and is in some ways more appealing. In it, volatility is random. The simplest model is

$$r_t = \sigma_t \varepsilon_t$$

$$\log(\sigma_t^2) = \omega + \phi \log(\sigma_{t-1}^2) + \sigma_u u_t$$

where the errors ε_t and u_t are iid standard normal, and mutually independent. The model has therefore three parameters: ω, ϕ and σ_u.

One can write the model as

$$\log(r_t^2) = \log(\sigma_t^2) + \log(\varepsilon_t^2)$$

$$\log(\sigma_t^2) = \omega + \phi \log(\sigma_{t-1}^2) + \sigma_u u_t$$

which is a model in state space form with $\log(\sigma_t^2)$ as the state, except that the error term in the measurement equation is nonnormal. The Kalman filter can still be used as if the measurement error were normal (the Kalman filter is still the best linear estimate but it will not be optimal). Doing the filter right is harder. The particle filter is one way. Kim, Shephard and Chib (1998) propose another, that has the useful spinoff of being a good general algorithm for a linear filtering problem with non-Gaussian errors.

To see how this algorithm works, write the model as

$$\alpha_t = \log(\sigma_t^2)$$

$$\alpha_t = \omega + \phi \alpha_{t-1} + \sigma_u u_t$$

$$\log(r_t^2) = \alpha_t + \Sigma_{i=1}^{n} q_{it} v_{it}$$

where exactly one of $\{q_{1t}, q_{2t}, ...q_{nt}\}$ is equal to 1 and the others are all equal to zero, $P(q_{it} = 1) = p_i$ and $v_{it} \sim iidN(\mu_i, \sigma_i^2)$. In other words, the log-$\chi^2(1)$ random error is being approximated by a mixture of normals. The parameters of this mixing distribution are *not* to be estimated…rather they are given by matching moments and, for n=7, are as follows:

	p_i	μ_i	σ_i^2
i=1	0.00730	-11.40039	5.79596
i=2	0.10556	-5.24321	2.61369
i=3	0.00002	-9.83726	5.17950
i=4	0.04395	1.50746	0.16735
i=5	0.34001	-0.65098	0.64009
i=6	0.24566	0.52478	0.34023
i=7	0.25750	-2.35859	1.26261

Source: Kim, Shephard and Chib (1998).

We start with a prior for the parameters. For example, we could have an inverse-Wishart prior for σ_u^2, a rescaled beta prior for ϕ and a normal prior for ω. We can then use Gibbs sampling to work out the distribution of α_t and the parameters, conditional on Y_T. Here is the algorithm:

1. Hold the $\{q_{it}\}$s and the parameters fixed. Now this is a linear state space model. We use the simulation smoother to take a draw of the $\{\alpha_t\}s$.
2. Hold the α_t s and the parameters fixed. Now we can work out $\log(\varepsilon_t^2) = \log(y_t^2) - \alpha_t = \Sigma_{i=1}^n q_{it} v_{it}$. Now

$$P(q_{it} = 1 | \{\alpha_t\}_{t=1}^T, Y_T) = \frac{N(\mu_i, \sigma_i^2) p_i}{\Sigma_{j=1}^n N(\mu_j, \sigma_j^2) p_j}$$

and this gives us $q_{it} | \alpha_t, Y_T$.
3. Finally take draws of the parameters conditional on $\{q_{it}\}$s and α_t. Expressions for the posteriors are in Kim, Shephard and Chib (1998).

Then repeat in the usual way.

Realized volatility
In the last ten years or so, high-frequency data have become available. Taking the sum of squares of five-minute returns gives a natural estimator of daily volatility. In fact, in the limit as the observation window size goes to zero, the sum of squared returns is an arbitrarily good estimator of variance under some assumptions (that rule out market microstructure noise, infrequent trading and so on). This is called realized volatility. It is has the desirable feature of being model-free.

One might wish to evaluate a forecast of volatility. People at one stage used to run regressions of the form

$$r_t^2 = \alpha + \beta \hat{\sigma}_t^2 + \varepsilon_t$$

to evaluate a forecast of volatility on day t, where $\hat{\sigma}_t^2$ is the forecast obtained from a GARCH model and r_t is the return on that day. They would get very low R-squared values from this

regression, and conclude that the volatility forecasts were poor. Instead, Andersen and Bollerslev (1998) considered the regression

$$RVOL_t = \alpha + \beta \hat{\sigma}_t^2 + \varepsilon_t$$

where $RVOL_t$ denotes the realized volatility on day t. Suddenly, the R-squared value became something in the range 40-50 percent. It seems that the daily squared return is too noisy a measure of volatility. The problem was not that GARCH models gave poor forecasts, but that daily squared return was too noisy a measure to evaluate those forecasts.

We can fit time series models directly to realized volatility to obtain forecasts. This also raises an intriguing possibility for measuring risk-aversion, considered by Bollerslev, Tauchen and Zhou (2009). We can get a forecast of the realized volatility from fitting an autoregression. We can get options implied volatility for the same period. The difference between those two is a risk premium that compensates investors for uncertainty about volatility. Under some assumptions, that is proportional to risk aversion. So the spread between a forecast of realized volatility and options-implied volatility represents a market-based measure of risk aversion.

Another concept is the realized semivariance. The downside realized semivariance is $\Sigma r_t^2 1(r_t < 0)$ where r_t is the tth intraday return on day t. The upside realized semivariance is $\Sigma r_t^2 1(r_t > 0)$ and the two together sum to the realized variance. We can similarly estimate downside and upside betas. Ang et al. (2006) show that downside and upside risk are priced differently.

Handout on Structural Stability

Let's consider a regression model of the form

$$y_t = x_t'\beta_t + \varepsilon_t$$

where ε_t is an mds and $T^{-1}\Sigma_{t=1}^{[T\lambda]} x_t x_t' \to_p \lambda\Sigma_x$, uniformly in λ. Also

$$T^{-1/2}\Sigma_{t=1}^{[T\lambda]} x_t \varepsilon_t \Rightarrow \sigma\Sigma_x^{-1/2} B(\lambda)$$

where σ^2 is the variance of ε_t and $B(r)$ is a k-dimensional standard Brownian motion. The null hypothesis is that

$$\beta_t = \beta \ \ \forall t$$

where β_t is a k-dimensional parameter vector. The alternative could be

$$\beta_t = \beta, \ t \le r$$

$$\beta_t = \beta + \gamma, \ t > r$$

where r, the break date is either known a priori or unknown. Or the alternative could be

$$\beta_t = \beta_{t-1} + \eta_t$$

where η_t is $iid(0, G)$.

The Chow test

The standard Chow test gives a test of the null of stability against the first kind of alternative, when the break date is known. The test statistic is of the form

$$F(r) = \frac{SSR - SSU}{SSU/(T-k)}$$

where SSR and SSU are restricted an unrestricted sums of squares respectively.

But the break date is rarely known ahead of time. Even when we think that the potential break date is obvious, it's really because we already took a peak at the data. So it is usually more appropriate to treat the break date as unknown. We can try every possible break date and work out the statistic

$$\sup F = \sup_{\pi T \le \lambda \le (1-\pi)T} F(\lambda)$$

Andrews (1993) derived the distribution of this test statistic. Here is the derivation. In the restricted regression, let the OLS estimator be $\hat{\beta}$. As $\hat{\beta} - \beta = (\Sigma_{t=1}^T x_t x_t')^{-1}\Sigma_{t=1}^T x_t \varepsilon_t$, it follows that

$$T^{1/2}(\hat{\beta} - \beta) \to \sigma\Sigma_x^{-1/2} B(1)$$

In the unrestricted regression, let the break date be $T\lambda$ and let the OLS estimators before and after the break be $\hat{\beta}_1$ and $\hat{\beta}_2$. We similarly have

$$T^{1/2}(\hat{\beta}_1 - \beta_1) \to \sigma\Sigma_x^{1/2}\lambda^{-1} B(\lambda)$$

and

$$T^{1/2}(\hat{\beta}_2 - \beta_2) \to \sigma\Sigma_x^{1/2}(1-\lambda)^{-1}B(1-\lambda)$$

The restricted sum of squared residuals is:

$$\Sigma_{t=1}^{T}(y_t - \hat{\beta}'x_t)^2 = \Sigma_{t=1}^{T}(\varepsilon_t - (\hat{\beta}-\beta)'x_t)^2 = \Sigma_{t=1}^{T}\varepsilon_t^2 + (\hat{\beta}-\beta)'\Sigma_{t=1}^{T}x_t x_t'(\hat{\beta}-\beta) - 2(\hat{\beta}-\beta)'\Sigma_{t=1}^{T}x_t\varepsilon_t$$

and the unrestricted sum of squared residuals is

$$\Sigma_{t=1}^{T}\varepsilon_t^2 + (\hat{\beta}_1-\beta)'\Sigma_{t=1}^{T\lambda}x_t x_t'(\hat{\beta}_1-\beta) - 2(\hat{\beta}_1-\beta)'\Sigma_{t=1}^{T\lambda}x_t\varepsilon_t + (\hat{\beta}_2-\beta)'\Sigma_{t=T\lambda+1}^{T}x_t x_t'(\hat{\beta}_2-\beta) - 2(\hat{\beta}_2-\beta)'\Sigma_{t=T\lambda+1}^{T}x_t\varepsilon_t$$

The numerator of the Chow statistic is therefore

$$(\hat{\beta}-\beta)'\Sigma_{t=1}^{T}x_t x_t'(\hat{\beta}-\beta) - 2(\hat{\beta}-\beta)'\Sigma_{t=1}^{T}x_t\varepsilon_t - (\hat{\beta}_1-\beta)'\Sigma_{t=1}^{T\lambda}x_t x_t'(\hat{\beta}_1-\beta) + 2(\hat{\beta}_1-\beta)'\Sigma_{t=1}^{T\lambda}x_t\varepsilon_t$$

$$-(\hat{\beta}_2-\beta)'\Sigma_{t=T\lambda+1}^{T}x_t x_t'(\hat{\beta}_2-\beta) + 2(\hat{\beta}_2-\beta)'\Sigma_{t=T\lambda+1}^{T}x_t\varepsilon_t$$

Taking the limits of each element gives:

$$SSR - SSU \to_d \sigma^2 B(1)'B(1) - 2\sigma^2 B(1)'B(1) - \lambda^{-2}B(\lambda)'\lambda B(\lambda) + 2\sigma^2\lambda^{-1}B(\lambda)'B(\lambda)$$

$$-(1-\lambda)^{-2}B(1-\lambda)'(1-\lambda)B(1-\lambda) + 2\sigma^2(1-\lambda)^{-1}B(1-\lambda)'B(1-\lambda)$$

$$\therefore SSR - SSU \to_d \sigma^2\{\frac{B(\lambda)'B(\lambda)}{\lambda} + \frac{B(1-\lambda)'B(1-\lambda)}{1-\lambda} - B(1)'B(1)\}$$

$$= \sigma^2\{\frac{B(\lambda)'B(\lambda)}{\lambda} + \frac{[(B(1)-B(\lambda))'(B(1)-B(\lambda))]}{1-\lambda} - B(1)'B(1)\}$$

$$= \frac{\sigma^2}{\lambda(1-\lambda)}\{(1-\lambda)B(\lambda)'B(\lambda) + \lambda[(B(1)-B(\lambda))'(B(1)-B(\lambda))] - \lambda(1-\lambda)B(1)'B(1)\}$$

$$= \frac{\sigma^2}{\lambda(1-\lambda)}\{(1-\lambda)B(\lambda)'B(\lambda) + \lambda B(1)'B(1) - 2\lambda B(\lambda)'B(1) + \lambda B(\lambda)'B(\lambda) - \lambda(1-\lambda)B(1)'B(1)\}$$

$$= \frac{\sigma^2}{\lambda(1-\lambda)}\{B(\lambda)'B(\lambda) + \lambda^2 B(1)'B(1) - 2\lambda B(\lambda)'B(1)\} = \frac{\sigma^2}{\lambda(1-\lambda)}[B(\lambda)-\lambda B(1)]'[B(\lambda)-\lambda B(1)]$$

The denominator of the Chow statistic is $SSU/(T-k) = \Sigma_{t=1}^{T}\varepsilon_t^2/(T-k) + o_p(1) \to_p \sigma^2$ and so for any break point $T\lambda$, the limiting distribution of the Chow statistic is:

$$F(\lambda) \Rightarrow \frac{(B(\lambda)-\lambda B(1))(B(\lambda)-\lambda B(1))'}{\lambda(1-\lambda)}$$

For any fixed λ, this is a χ^2 distribution on k degrees of freedom. But this distribution holds jointly in all λ and so the sup-F statistic has the distribution:

$$\sup F \to_d \sup_{\pi \le \lambda \le 1-\pi} \frac{(B(\lambda)-\lambda B(1))(B(\lambda)-\lambda B(1))'}{\lambda(1-\lambda)}$$

This is the distribution that was derived by Andrews (1993). It is nonstandard, but it can be simulated and is tabulated in Andrews (1993).

The Chow test compares parameters before and after a potential break date. But it will not work well if the alleged break is at the very end (or very start) of the sample. That is because one sample will have too few observations to obtain a precise estimate. A different kind of break test is appropriate in this case. It estimates the parameters on the first (large) part of the data and then assesses the size of the residuals on the second (small) part of the data.

Suppose that the parameters are estimated on a sample of size T but that we wish to check for a structural break over a subsequent sample of size T^* where $T^*/T = \theta$.

Let $\hat{\beta}$ denote the OLS estimate from the original sample and let $e_t = y_t - \hat{\beta}' x_t$ denote the forecasting error for an observation in the subsequent sample, $t = T+1, T+2, ..T+T^*$. Let V denote an estimator of 2π times the spectral density of $x_t\varepsilon_t$. Then we can form the test statistic

$$[T^{*-1/2}\Sigma_{t=T+1}^{T+T^*} x_t e_t]'(V(1+\theta))^{-1}[T^{*-1/2}\Sigma_{t=T+1}^{T+T^*} x_t e_t]$$

which will have a χ^2 pointwise limiting distribution under the null of no structural break on k degrees of freedom. One can also search over possible break-dates and this will have the same distribution as the sup-F test.

Estimating the Break Date

A separate question is estimation of the break date, given that there is a one-time structural break. The value of λ that maximizes the sup-F statistic can be thought of as an estimator of the break date. Bai (1997) considers estimation of the break date. Consider the model

$$y_t = X_t'\beta + X_t'\delta 1(t > r) + \varepsilon_t$$

and so there is a break at time r. Note that r is the break *date*, not the fraction of the way through the sample that the break occurs. Let's estimate the parameters, including the break date, my minimizing the sum of squared residuals:

$$(\hat{\alpha}, \hat{\beta}, \hat{r}, \hat{\delta}\} = \arg\min_{\{\alpha,\beta,r,\delta\}} \Sigma(y_t - X_t'\beta - X_t'\delta 1(t > r))^2$$

which has to be done numerically. Suppose that $M = E(x_t x_t')$ and the shocks are conditionally homoskedastic. Then Bai shows that

$$\frac{\delta' M \delta}{\sigma^2}(\hat{r} - r) \rightarrow_d \arg\max_s \{W(s) - \frac{|s|}{2}\}$$

where $W(s)$ is a two-sided Brownian motion (defined over $-\infty < s < \infty$). A two-sided Brownian motion is

$W_1(s)$ if $s > 0$

0 if $s = 0$

$W_2(-s)$ if $s < 0$

where W_1 and W_2 are two independent standard Brownian motions defined over $0 < s < \infty$.

This means that if we define

$$\hat{m} = \frac{\hat{\delta}'\hat{M}\hat{\delta}}{\hat{\sigma}^2}$$

from the inverse of the acceptance region of a test, a 95 percent confidence interval for r will be given by $[\hat{r} - P_{0.975}m, \hat{r} - P_{0.025}m]$ where P_α denotes the α percentile of the distribution of $\arg\max_s \{W(s) - \frac{|s|}{2}\}$.

The Nyblom Test

Turning to the random walk parameter alternative, a Lagrange multiplier test does not require the alternative to be estimated (convenient because that would involve running the Kalman filter). Nyblom (1989) shows that the LM statistic is

$$L = \frac{1}{T^2}\Sigma_{r=1}^{T}(\Sigma_{s=r}^{T}\hat{z}_s)'\hat{\Sigma}_x^{-1}(\Sigma_{s=r}^{T}\hat{z}_t)/\hat{\sigma}^2$$

where $\hat{z}_t = x_t e_t$ and $e_t = \varepsilon_t - x_t'(\hat{\beta} - \beta)$. Now

$$\Sigma_{s=r}^{T}\hat{z}_s = \Sigma_{s=1}^{T}\hat{z}_s - \Sigma_{s=1}^{r-1}\hat{z}_s = -\Sigma_{s=1}^{r-1}\hat{z}_s = -\Sigma_{s=1}^{r-1}\{x_s\varepsilon_s - x_s x_s'(\hat{\beta} - \beta)\}$$

$$\therefore T^{-1/2}\Sigma_{s=r}^{T}\hat{z}_s = -T^{-1/2}\Sigma_{s=1}^{r-1}x_s\varepsilon_s + T^{-1}\Sigma_{s=1}^{r-1}x_s x_s'(T^{-1}\Sigma_{t=1}^{T}x_t x_t')^{-1}T^{-1/2}\Sigma_{t=1}^{T}x_t\varepsilon_t$$

$$\therefore T^{-1/2}\Sigma_{s=r}^{T}\hat{z}_s \to -\sigma\Sigma_x^{1/2}B(\lambda) + \lambda\sigma\Sigma_x^{1/2}B(1) = -\sigma\Sigma_x^{1/2}(B(\lambda) - \lambda B(1))$$

$$\therefore L \to \int_0^1 (B(\lambda) - \lambda B(1))'(B(\lambda) - \lambda B(1))d\lambda$$

Again this is a nonstandard distribution that has to be tabulated.

Handout on Threshold Models

The threshold model can be popular in a number of contexts. For example, the Phillips curve may be steep at very high or low level of unemployment, but flat in an intermediate range. A simple threshold model is:

$$y_i = \beta_1' x_i + \varepsilon_i,\ q_i < \theta$$
$$y_i = \beta_2' x_i + \varepsilon_i,\ q_i \geq \theta$$

and θ is the threshold parameter. The model may be written as:

$$y_i = \beta_1' x_i + (\beta_2 - \beta_1)' q_i 1(q_i \geq \theta) + \varepsilon_i$$

Of course, q_i could be equal to x_i which would make for a kink in the relationship between y_i and x_i, but the model is more general than that. In the case of the kinked model, we might well want to impose the restriction that $\beta_1'\theta = \beta_2'\theta$ so that there is no jump at the threshold parameter.

Least squares estimation of the parameters is the natural way of proceeding. This can be done by concentrated optimization. Suppose that I fix the parameter θ. Then all the other parameters of the model can be estimated in closed form by simple OLS (or restricted least squares if there is a parameter restriction of the sort described in the last paragraph). Let the sum of squared residuals be $S(\theta)$. The estimate of θ is then $\arg\min_\theta S(\theta)$.

Inference in this model is difficult. Hansen (2000) proposes a way of getting a useable asymptotic distribution. He adopts the device of saying that $\beta_2 - \beta_1 = cn^{-\alpha}$ with $c \neq 0$ and $0 < \alpha < 1/2$. Under homoscedasticity, and if $M = E(x_i x_i')$, $\sigma^2 = Var(\varepsilon_i)$ and f is the density of q_i evaluated at the true threshold parameter:

$$n^{1-2\alpha} \frac{f(c'Mc)}{\sigma^2}(\hat{\theta} - \theta) \rightarrow_d \arg\max_{-\infty < r < \infty}[W(r) - \frac{1}{2}|r|]$$

where $W(r)$ is a two-sided Brownian motion. (The similarity of this to the Bai result on estimating the date of a structural break is no accident.)

We might want to test the hypothesis that $\beta_1 = \beta_2$, meaning that there is no actual threshold effect. The likelihood ratio test of this hypothesis is:

$$F = \frac{SSR - SSU}{SSU / n}$$

where SSU depends on the estimated threshold, $\hat{\gamma}$. Unfortunately the asymptotic distribution of this test statistic depends on nuisance parameters and cannot be tabulated. The bootstrap strategy of Hansen (1996) is however available and does get the correct asymptotic distribution. The idea of this bootstrap is as follows:

1. Hold fixed the regressors and the threshold variable.

2. Obtain the residuals.
3. In each bootstrap draw, resample from the residuals with replacement and build up a new dataset assuming that there is no threshold effect and that $\beta_1 = \beta_2 = 0$.
4. Compute the F-statistic.
5. Repeat steps 3-4 many times to obtain the bootstrap distribution of the F statistic.
6. Reject the null of no threshold effect only if the actual F statistic is above the 95th percentile of the bootstrap distribution.

Other notes.
1. The transition function can be smoother than in the simple threshold model. We might specify that:

$$y_t = \beta_1' x_i + (\beta_2 - \beta_1)' \frac{1}{1 + \exp(-\gamma(q_i - a))} + \varepsilon_i$$

2. We could have a model in which the right hand side variable is a lagged dependent variable, along the lines:

$$y_t = \beta_1 y_{t-1} + \varepsilon_t, y_{t-1} < \theta$$
$$y_t = \beta_2 y_{t-1} + \varepsilon_t, y_{t-1} \geq \theta$$

This would be a threshold autoregression.
3. There is no restriction on the number of regimes to being just 2. For example people estimate models of interest rates where there is fast mean reversion for either very high or very low interest rates, but slow mean reversion in between.
4. We can also lag the threshold variable, and this can be called the *delay* parameter.

Econ 607 Jonathan Wright
Fall 2023

Handout on Solving Rational Expectations Linear Difference Models

Suppose that we have a model of the form

$$AE_t y_{t+1} = By_t + C\varepsilon_t \quad (1)$$

where $y_t = (k_t', d_t')'$, k_t is an n_k x1 vector of predetermined variables that are given at time t (they were determined earlier), d_t is an n_d x1 vector of forward-looking (jump) variables that can be controlled at time t, and ε_t is an iid shock.

Let us define the matrices $F = A^{-1}B$ and $G = A^{-1}C$. Suppose that all the eigenvalues of F are unique, and that we can write $F = V^{-1}JV$ where J is a diagonal matrix of eigenvalues and V is the inverse of the matrix of corresponding eigenvectors. This is called the Jordan form (there is an extension to the case of repeated eigenvalues).

Blanchard and Kahn (1980) showed that the condition for equation (1) to have a unique solution is that $A^{-1}B$ has n_k eigenvalues less than 1 in absolute value and n_d eigenvalues greater than one in absolute magnitude. If there are too many stable roots, then there will be multiple equilibria. If there are too few, them there will be no solution.

The Blanchard and Kahn solution method relies on the matrix A being invertible. Other techniques are available instead when it is not. Here is the Blanchard-Kahn method:

1. Obtain the Jordan form of $F = V^{-1}JV$. Order the Jordan blocks such that the first n_k diagonal elements are less than 1 in absolute value and the remainder are greater than 1 in absolute value.
2. Partition F, G, V^{-1} and J conformably with y_t as

$$F = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix},\ G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix},\ V^{-1} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \text{ and } J = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}$$

3. The solution is then

$$k_{t+1} = [F_{11} - F_{12}V_{22}^{-1}V_{21}]k_t + [G_1 - F_{12}V_{22}^{-1}J_2^{-1}(V_{21}G_1 + V_{22}G_2)]\varepsilon_t \quad (2)$$

$$d_t = -V_{22}^{-1}V_{21}k_t - V_{22}^{-1}J_2^{-1}[V_{21}G_1 + V_{22}G_2]\varepsilon_t \quad (3)$$

So if I wanted to simulate data from this process, the steps would be:
1. Start out with $k_0 = 0$.
2. Draw the exogenous process $\{\varepsilon_t\}$.
3. Compute $\{k_t\}$ from the VAR in (2).
4. Compute $\{d_t\}$ from (3).

Idea of the solution method

Where this is coming from is that I can rewrite equation (1) as

$$AE_t y_{t+1} = By_t + Cx_t$$

$$\therefore E_t y_{t+1} = A^{-1}By_t + A^{-1}Cx_t$$

$$\therefore E_t y_{t+1} = V^{-1}JVy_t + A^{-1}Cx_t$$

$$\therefore E_t Vy_{t+1} = JVy_t + VA^{-1}Cx_t$$

$$\therefore E_t y^*_{t+1} = Jy^*_t + C^*x_t$$

where $y^*_t = Vy_t$. But the matrix J is diagonal, so this allows me to decouple the system into two parts: one stable, and one unstable. I solve the transformed equations separately to get y^*_t and then rotate back to get the solution for $y_t = (k'_t, d'_t)'$.

Example 1

As an example, consider the benchmark forward-looking new-Keynesian model (in which the steady state level of inflation is zero):

$$\pi_t = \kappa g_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} \tag{4}$$

$$g_t = E_t g_{t+1} - \sigma(i_t - E_t \pi_{t+1} - \bar{i}) + \varepsilon_t \tag{5}$$

$$i_t = \bar{i} + \phi_1 \pi_t + \phi_2 g_t \tag{6}$$

Substituting (6) into (5) gives

$$g_t = E_t g_{t+1} - \sigma(\phi_1 \pi_t + \phi_2 g_t - E_t(\pi_{t+1})) + \varepsilon_t \tag{7}$$

The combination of (4) and (7) can be written as:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -\gamma_f & 0 \\ 0 & \sigma & 1 \end{pmatrix} \begin{pmatrix} \pi_t \\ E_t \pi_{t+1} \\ E_t g_{t+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ \gamma_b & -1 & \kappa \\ 0 & \sigma\phi_1 & 1+\sigma\phi_2 \end{pmatrix} \begin{pmatrix} \pi_{t-1} \\ \pi_t \\ g_t \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \varepsilon_t \tag{8}$$

which is clearly a model in the form of equation (1) and can be solved by the Blanchard-Kahn method. In this example, there is one predetermined variable, π_{t-1} and there are two jump variables π_t and g_t.

Example 2

Example 2 is a little simpler, but the same kind of setup:

$$\pi_t = \kappa g_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} \tag{9}$$

$$g_t = \phi_1 g_{t-1} + \phi_2 \pi_{t-1} + \varepsilon_t \tag{10}$$

This can be written as:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\kappa & 1 & -\gamma_f \end{pmatrix} \begin{pmatrix} g_t \\ \pi_t \\ E_t \pi_{t+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & 0 \\ 0 & 0 & 1 \\ 0 & \gamma_b & 0 \end{pmatrix} \begin{pmatrix} g_{t-1} \\ \pi_{t-1} \\ \pi_t \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \varepsilon_t \tag{11}$$

which is another model in the form of equation (1) and can be solved by the Blanchard-Kahn method. In this example, there are two predetermined variables, g_{t-1} and π_{t-1} and there is one jump variable π_t.