Econometrics - lecture notes

Andrea Carriero

December 5, 2023

Contents

Part I
The linear regression model

1 Introduction

1.1 Univariate model

We will consider models in which we have 1 dependent variable y. We will try to explain this variable using some independent variables x.

Consider the model:

$$y_t = a + bx_t;\ t = 1, ..., T$$

this is a set of T equations:

$$\begin{aligned} y_1 &= a + bx_1 \\ y_2 &= a + bx_2 \\ &\vdots \\ y_T &= a + bx_T \end{aligned}$$

This system is overdetermined: there are T equations and 2 unknowns, a and b. There is not a solution, there are no a and b so that all these equations are satisfied (unless some equations are simply the same as others or linear combinations). We could drop T-k equations, then we would have a unique solution for a and b but that would mean loosing information. Instead, consider adding an error to the model:

$$y_t = a + bx_t + \varepsilon_t;\ t = 1, ..., T \tag{1}$$

this is a set of T equations:

$$\begin{aligned} y_1 &= a + bx_1 + \varepsilon_1 \\ y_2 &= a + bx_2 + \varepsilon_2 \\ &\vdots \\ y_T &= a + bx_T + \varepsilon_T \end{aligned}$$

We have now 2+T unknowns: a,b, and $\varepsilon_t;\ t = 1, ..., T$. The system now has more unknowns than equations and there are many solutions. We choose the solution that

minimizes the sum of the squared residuals

$$RSS = \varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_T^2$$

by minimizing the SSR we get the ordinary least squares estimates $\hat{a}$ and $\hat{b}$ (of a and b). The $\hat{y}_t$ is the regression line:

$$\hat{y}_t = \hat{a} + \hat{b}x_t$$

and the residual is:

$$y_t - \hat{y}_t = \hat{\varepsilon}_t.$$

This gives us:

$$\underset{data,\ actual}{y_t} = \underset{\text{model},fitted}{\hat{y}_t} + \underset{residuals}{\hat{\varepsilon}_t}$$

This is a key decomposition, which we will explore further later on.

The goal is now to extend the model to the case of more than one regressor. First, realize that the intercept is a regressor

$$y_t = a \times 1 + b \times x_t + \varepsilon_t;\ t = 1, ..., T \tag{2}$$

but a special one, as it is equal to 1 in all time periods:

$$\begin{aligned} y_1 &= a \times 1 + b \times x_1 + \varepsilon_1; \\ y_2 &= a \times 1 + b \times x_2 + \varepsilon_2; \\ &\vdots \\ y_T &= a \times 1 + b \times x_T + \varepsilon_T. \end{aligned}$$

The linear regression model with k regressors:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + ... + \beta_k x_{k,t} + \varepsilon_t$$

special case: $k = 2$, $x_{1,t} = 1$ for all $t = 1, ..., T$:

$$y_t = \beta_1 + \beta_2 x_{2,t} + \varepsilon_t,$$

which is the same model as (1) (with $a = \beta_1,\ b = \beta_2\ x_{2t} = x_t$).

1.2 Multivariate model

This course is about the general model

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + ... + \beta_k x_{k,t} + \varepsilon_t,\ t = 1, ..., T$$

where there are k regressors. This a time series model.

$$y_j = \beta_1 x_{1,j} + \beta_2 x_{2,j} + \beta_3 x_{3,j} + ... + \beta_k x_{k,j} + \varepsilon_j,\ j = 1, ..., N$$

This is a cross-section model. Panel data include both the time series and the cross section dimension.

Define the matrices:

$$\underset{T\times 1}{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix};\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix};$$

$$x_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,T} \end{bmatrix};\ x_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,T} \end{bmatrix};\ ...;\ x_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,T} \end{bmatrix};$$

$$\underset{T\times k}{X} = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & & \vdots \\ x_{1,T} & x_{2,T} & \cdots & x_{k,T} \end{bmatrix};\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix}$$

we have:

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon} \tag{3}$$

This compact notation is equivalent to

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + ... + \beta_k x_{k,t} + \varepsilon_t,\ t = 1, ..., T \tag{4}$$

1.3 Ordinary Least Squares

Derive the Ordinary Least Squares estimator for this model.

$$\begin{aligned} RSS &= \varepsilon_1^2 + \varepsilon_2^2 ... + \varepsilon_T^2 = \varepsilon'\varepsilon \\ &= \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \vdots & \varepsilon_T \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix} \\ &= \underbrace{(y - X\beta)'}_{\varepsilon'} \underbrace{(y - X\beta)}_{\varepsilon} \end{aligned}$$

The OLS problem:

$$\hat{\beta}_{OLS} = \arg\min RSS(\beta)$$

the solution to this is:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y.$$

How to interpret this? Divide both $X'X$ and $X'Y$ by T:

$$\hat{\beta}_{OLS} = \left(\frac{X'X}{T}\right)^{-1} \frac{X'Y}{T}$$

The quantities

$$\frac{X'X}{T} \to E[x_t x_t'], \ \frac{X'Y}{T} \to E[x_t y_t].$$

where $x_t = [x_{1t}, x_{2t}, ..., x_{kt}]$. So the estimator is given by the ratio of the (uncentered) sample covariance between X and Y and the variance of X. This is reminiscent of the formula for the model with 1 regressor:

$$\hat{\beta}_{OLS} = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)}.$$

Once we have the estimator we can compute the fitted values

$$\hat{Y} = X\hat{\beta}_{OLS}$$

and the decomposition:

$$Y = \hat{Y} + \hat{\varepsilon}$$

this is the actual, fitted, residual decomposition. The decomposition $Y = \hat{Y} + \hat{\varepsilon}$ is the basis to compute the R^2. Note that:

$$Y = X\hat{\beta}_{OLS} + \hat{\varepsilon} = X\underset{?}{\beta} + \underset{?}{\varepsilon}.$$

The $\hat{\varepsilon}$ is such that $\hat{\varepsilon}'\hat{\varepsilon}$ is the smallest possible.

2 Algebraic properties of the OLS estimator.

2.1 Orthogonality and its implications

We derived the OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

Recall the First Order Condition (FOC):

$$-(X'X)\hat{\beta}_{OLS} + X'Y = 0$$

that is:

$$X'(Y - X\hat{\beta}_{OLS}) = 0$$

or:

$$\underset{k\times T}{X'} \underset{T\times 1}{\hat{\varepsilon}} = 0$$

Under OLS the residuals $\hat{\varepsilon}$ are **orthogonal** to the regressors X. This implies the following properties:

1. If the model has an intercept, then the residuals have sample average exactly equal to 0:

$$\frac{1}{T}\mathbf{1}'\hat{\varepsilon} = \frac{1}{T}\underset{1\times T}{[1\ 1\ 1\ 1\ .....\ 1]}\begin{bmatrix}\hat{\varepsilon}_1\\ \hat{\varepsilon}_2\\ \vdots\\ \hat{\varepsilon}_T\end{bmatrix} = \frac{1}{T}(\hat{\varepsilon}_1 + \hat{\varepsilon}_2 + \hat{\varepsilon}_3 + ... + \hat{\varepsilon}_T) = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_T = 0$$

2. The regression line passes trough the mean of the data ($\bar{Y} = \overline{\hat{Y}}$):

$$\begin{aligned} Y &= \hat{Y} + \hat{\varepsilon} \\ \frac{1}{T}\mathbf{1}'Y &= \frac{1}{T}\mathbf{1}'\hat{Y} + \frac{1}{T}\mathbf{1}'\hat{\varepsilon} \\ \frac{1}{T}\sum_{t=1}^{T}Y_t &= \frac{1}{T}\sum_{t=1}^{T}\hat{Y}_t + \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_T \\ &= \frac{1}{T}\sum_{t=1}^{T}\hat{Y}_t + 0 \\ \bar{Y} &= \overline{\hat{Y}} \end{aligned}$$

3. OLS makes the $\hat{Y}$ such that it orthogonal to $\hat{\varepsilon}$:

$$\hat{Y}'\hat{\varepsilon} = (X\hat{\beta})'\hat{\varepsilon} = \hat{\beta}'X'\hat{\varepsilon} = \hat{\beta}'0 = 0,$$

Specifically OLS implies a partition of Y into two subspaces orthogonal to each other. We have:

$$\hat{Y} = X\hat{\beta} = X\underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}} = \underbrace{X(X'X)^{-1}X'}_{P_X}Y = P_XY$$

Where $P_X = X(X'X)^{-1}X$ is the projection matrix for the space spanned by the X. The matrix P_X projects things on the space spanned by the X. The residuals are:

$$\hat{\varepsilon} = Y - \hat{Y} = Y - P_XY = (I_n - P_X)Y = M_XY$$

where $M_X = (I_n - P_X)$ is the "residual maker" matrix" or "annichilator" matrix. The matrix M_X projects things on the space orthogonal to the X.

The matrices M_X and P_X have some important properties. They are symmetric:

$$\begin{aligned} P_X' &= X(X'X)^{-1}X' = P_X \\ M_X' &= I' - P_X' = I - P_X = M_X \end{aligned}$$

and idempotent:

$$\begin{aligned} P_XP_X &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' = P_X \end{aligned}$$

$$\begin{aligned} M_XM_X &= (I - P_X)(I - P_X) \\ &= I - P_X - P_X + P_XP_X \\ &= I - P_X = M_X \end{aligned}$$

M_X and P_X are mutually orthogonal:

$$P_XM_X = P_X(I - P_X) = P_X - P_XP_X = 0$$

The projection of X on X is X itself:

$$P_XX = X(X'X)^{-1}X'X = X$$

The projection of X on the space of M_X is 0:

$$M_XX = (I - P_X)X = X - P_XX = 0$$

Therefore we have:

$$Y = \hat{Y} + \hat{\varepsilon} = P_XY + M_XY$$

2.2 Goodness of fit

Recall the actual-fitted-residual decomposition:

$$Y = \hat{Y} + \hat{\varepsilon}$$

consider squaring both sides of the equation:

$$\begin{aligned} Y'Y &= (\hat{Y} + \hat{\varepsilon})'(\hat{Y} + \hat{\varepsilon}) \\ &= \hat{Y}'\hat{Y} + \hat{Y}'\hat{\varepsilon} + \hat{\varepsilon}'\hat{Y} + \hat{\varepsilon}'\hat{\varepsilon} \\ &= \hat{Y}'\hat{Y} + \hat{\varepsilon}'\hat{\varepsilon} \end{aligned}$$

note that:

$$\sum_{t=1}^{T} Y_t^2 = Y'Y = TSS;\ \sum_{t=1}^{T} \hat{Y}_t^2 = \hat{Y}'\hat{Y} = ESS;\ \sum_{t=1}^{T} \hat{\varepsilon}_t = \hat{\varepsilon}'\hat{\varepsilon} = RSS$$

we can write:

$$TSS = ESS + RSS$$

which gives:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

this is the *uncentered* R^2. However, the uncentered R^2 has a scaling problem. We need the centered R_C^2:

$$R_C^2 = \frac{CESS}{CTSS} = \frac{\sum_{t=1}^{T}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

where

$$\sum_{t=1}^{T}(Y_t - \bar{Y})^2 = CTSS;\ \sum_{t=1}^{T}(\hat{Y}_t - \bar{Y})^2 = CESS.$$

Moreover there is the adjusted $\bar{R}^2$ which takes into account how many regressors are in the model:

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}(1 - R_C^2),$$

when k increases, $\bar{R}^2$ reduces.

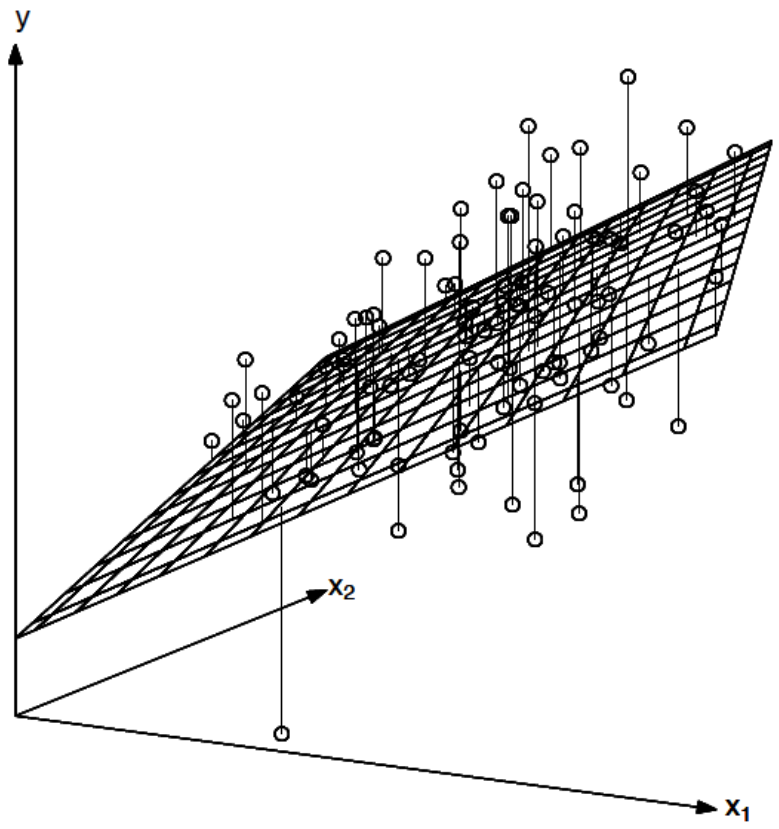


Figure 1: The regression plane of regression of y onto x_1, x_2. The plane has equation $\hat{y}_t = x_1\hat{\beta}_1 + x_2\hat{\beta}_2$.

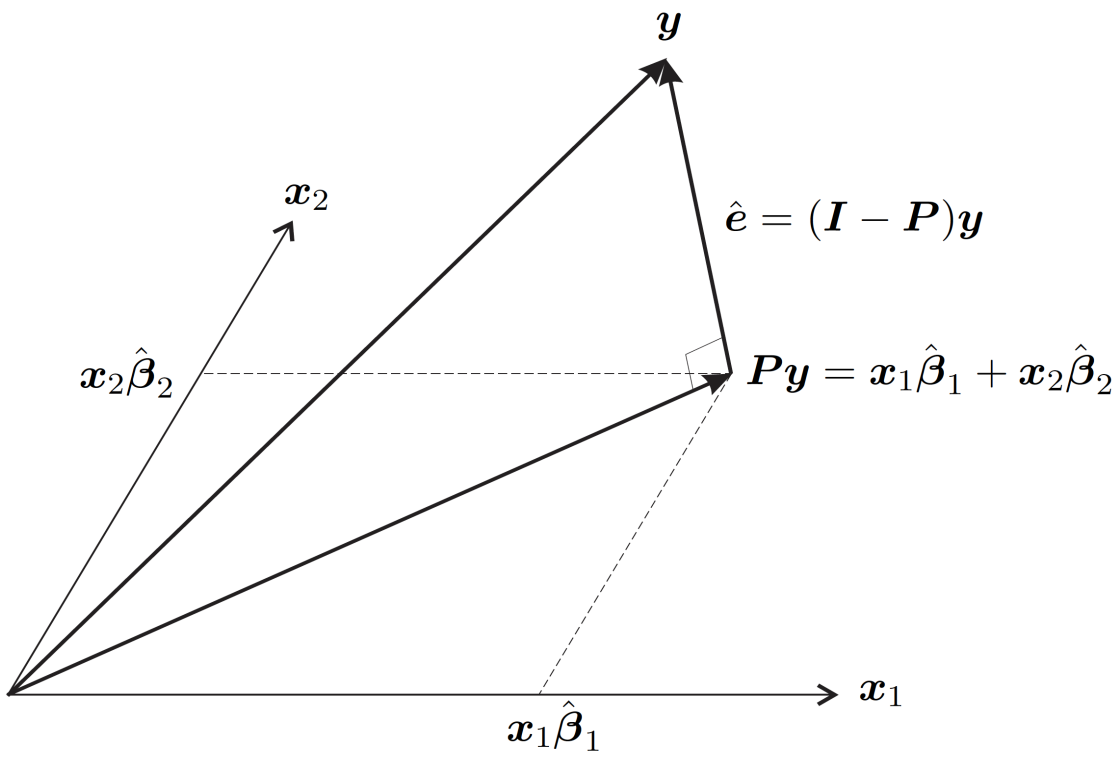


Figure 2: The orthogonal projection of y onto $\text{span}(x_1, x_2) = x_1\beta_1 + x_2\beta_2$

3 Small sample properties of the OLS estimator

So far, the vector

$$\underset{k\times k}{\hat{\beta}} = (X'X)^{-1}X'Y$$

is simply a vector of estimates. Note that we used one assumption to derive $\hat{\beta}$. We assumed that $(X'X)^{-1}$ exists. In order for this matrix to exist we need that $X'X$ is invertible. We require that $rank(X'X) = k$. If X has full rank k, then also $X'X$ will have full rank k.

$$(A1): \ X \text{ is full rank} = \text{no multicollinearity}$$

However, we have no theory to establish how good these estimates are. To do this, we need to add a probabilistic treatment to our model. We add:

$$(A2): X \text{ is deterministic, or fixed in repeated samples}$$

We postulate that ε is a random noise. We add the following assumptions:

$$(A3): E[\varepsilon] = 0$$

this implies $E[\varepsilon_t] = 0$, for all t. We add the assumption that:

$$(A4): E[\varepsilon\varepsilon'] = \sigma^2 I_T = \begin{bmatrix} \sigma^2 & 0 & & 0 \\ 0 & \sigma^2 & & \\ & & \ddots & \\ 0 & & 0 & \sigma^2 \end{bmatrix}$$

Note that

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}; \ \varepsilon' = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \vdots & \varepsilon_T \end{bmatrix}; \varepsilon\varepsilon' = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & & \varepsilon_1\varepsilon_T \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & & \\ & & \ddots & \\ \varepsilon_T\varepsilon_1 & & & \varepsilon_T^2 \end{bmatrix};$$

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1^2] & E[\varepsilon_1\varepsilon_2] & & E[\varepsilon_1\varepsilon_T] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2^2] & & \\ & & \ddots & \\ E[\varepsilon_T\varepsilon_1] & & & E[\varepsilon_T^2] \end{bmatrix}$$

so this assumption implies that

$$\begin{aligned} E[\varepsilon_t \varepsilon_s] &= 0, \text{ for all } t \neq s. \text{ no auto-correlation across periods/units} \\ E[\varepsilon_t^2] &= \sigma^2, \text{ for all } t. \text{ homoschedasticty (in conjunction with (A3))}^1 \end{aligned}$$

Finally, we add:

$$(A5) : \varepsilon \sim N(0, \sigma^2 I_N) : \text{ Gaussianity}$$

To summarize, we have the following set of assumptions.

$$\begin{aligned} (A1) &: X \text{ is full rank} = \text{no multicollinearity} \\ (A2) &: X \text{ is deterministic, or fixed in repeated samples} \\ (A3) &: E[\varepsilon] = 0 \\ (A4) &: E[\varepsilon\varepsilon'] = \sigma^2 I_T \text{ : sphericity (homoschedasticity, no autocorrelation)} \\ (A5) &: \varepsilon \sim N(\cdot, \cdot) : \text{ Gaussianity} \end{aligned}$$

So now we can use these to compute the properties of the OLS estimator.

3.1 Unbiasedness

lets say you want to estimate a coefficient θ, using an estimator $\hat{\theta}$. The estimator is unbiased iff:

$$E[\hat{\theta}] = \theta$$

First of all, recall the OLS estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

To study the properties, we need to put the estimator in the "sample error representation":

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \\ &= \underset{\text{actual value}}{\beta} + \underbrace{(X'X)^{-1}X'\varepsilon}_{\text{sampling error}} \end{aligned}$$

This is useful to compute the properties of the estimator. So consider taking the expectation:

$$E[\hat{\beta}] = E[\beta + (X'X)^{-1}X'\varepsilon]$$

now use the properties of the $E[\cdot]$ operator.

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[\underbrace{(X'X)^{-1}X'\varepsilon}_{\text{constant (A2)}}] \\ &= \beta + (X'X)^{-1}X' \underbrace{E[\varepsilon]}_{=0 \ (A3)} \\ &= \beta + (X'X)^{-1}X'0 \\ &= \beta \end{aligned}$$

3.2 Variance

Now, let us consider the variance of this estimator.

$$Var(\hat{\beta}) = Var(\beta + (X'X)^{-1}X'\varepsilon)$$

Here we can use the properties of the $Var[\cdot]$ operator.

$$\begin{aligned} Var(\hat{\beta}) &= Var(\beta + (X'X)^{-1}X'\varepsilon) \\ &= Var(\beta) + Var((X'X)^{-1}X'\varepsilon) + 2COV(...) \\ &= Var(\underbrace{(X'X)^{-1}X'\varepsilon}_{\text{constant, A2}}) \\ &= (X'X)^{-1}X'Var(\varepsilon)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_N X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

however, note that in our case β is not a random variable. It follows that $Var(\beta) = 0,\ 2COV(...) = 0$.

$$Var(Az) = AVar(z)A'$$

3.3 Distribution

Finally, the distribution of the OLS estimator.

$$\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$$

we know that

$$\begin{aligned}
\varepsilon &\sim N(0, \sigma^2 I_N) \\
(X'X)^{-1}X'\varepsilon &\sim N((X'X)^{-1}X'0, (X'X)^{-1}X'\sigma^2 I_N X(X'X)^{-1}) \\
&\sim N(0, \sigma^2(X'X)^{-1}) \\
\beta + (X'X)^{-1}X'\varepsilon &\sim N(0+\beta, \sigma^2(X'X)^{-1}) \\
\hat{\beta} &\sim N(\beta, \sigma^2(X'X)^{-1})
\end{aligned}$$

To sum up:

$A1-$ used to derive the estimator.
1) $\hat{\beta}$ is unbiased. We used A1-A3
2) $\hat{\beta}$ has variance $\sigma^2(X'X)^{-1}$. We used A1-A4
3) $\hat{\beta}$ has a normal distr. We used A1-A5

The expression

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

summarizes these results.

3.4 Gauss-Markov theorem

Next step, an optimality result. The OLS is the **Best Linear Unbiased Estimator** (BLUE). In the class of estimators that are 1) linear 2) unbiased, OLS is the one with the smallest variance. Gauss-Markov theorem.

Proof. consider another linear estimator.

$$b = \underset{k\times T}{C}\, Y$$

if $C = (X'X)^{-1}X'$ then $b = \hat{\beta}$. Require unbiasedness:

$$\begin{aligned}
E[b] &= \beta \\
E[b] &= E[CY] = E[C(X\beta + \varepsilon)] = E[CX\beta + C\varepsilon] = E[CX\beta] + E[C\varepsilon] \\
&= CX\beta + CE[\varepsilon] = CX\beta = \beta
\end{aligned}$$

this happens only if $CX = I$. So the class of linear unbiased estimators b is defined by the conditions:

$$\begin{aligned}
\text{any } b &= CY \ (linear) \\
\text{such that } CX &= I \ (unbiased)
\end{aligned}$$

Compute the variance:

$$\begin{aligned} Var(b) &= Var(CY) \\ &= Var(CX\beta + C\varepsilon) \\ &= Var(C\varepsilon) \\ &= CVar(\varepsilon)C' \\ &= C\sigma^2 IC' \\ &= \sigma^2 CC' \end{aligned}$$

Finally, define the matrix

$$D = C - (X'X)^{-1}X'$$

it follows that $C = D + (X'X)^{-1}X'$. Use this to compute:

$$\begin{aligned} Var(b) &= \sigma^2 CC' \\ &= \sigma^2\{D + (X'X)^{-1}X'\}\{D + (X'X)^{-1}X'\}' \\ &= \sigma^2\{D + (X'X)^{-1}X'\}\{D' + X(X'X)^{-1}\} \\ &= \sigma^2\{DD' + DX(X'X)^{-1} + (X'X)^{-1}X'D + (X'X)^{-1}X'X(X'X)^{-1}\} \\ &= \sigma^2\{DD' + 0 + 0 + (X'X)^{-1}\} \\ &= \sigma^2 DD' + \underbrace{\sigma^2(X'X)^{-1}}_{Var(OLS)}, \end{aligned}$$

where we used $DX = 0$, which is implied by the unbiasedness condition $CX = I$. Therefore:

$$Var(b) - Var(\hat{\beta}) = \sigma^2 DD',$$

where DD' is a positive definite matrix. this means that

$$Var(b) > Var(\hat{\beta})$$

therefore $\hat{\beta}$ has the smallest variance in the class of linear and unbiased estimators.

4 Inference on individual coefficients

We have considered the model

$$Y = X\underset{?}{\beta} + \varepsilon.$$

We have derived the estimator

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

We have seen that under A1-A5:

$$\hat{\beta} \sim N(\beta, \underbrace{\sigma^2(X'X)^{-1}}_{k\times k}).$$

In this case this is a k-variate random variable with multivariate Gaussian distribution. We want to make on the individual coefficient $\beta_j,\ j = 1, ..., k$. The distribution of the OLS estimator for the j-th element of the vector $\hat{\beta}$ is:

$$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1}]_{jj}).$$

The pivotal quantity:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}} \sim N(0,1)$$

can be used for inference. Specifically, for hypothesis testing and confidence intervals. Say we want to test $H_0 : \beta_j \overset{H_0}{=} \beta_0$. The pivotal quantity under the null is:

$$\hat{z} = \frac{\hat{\beta}_j - \beta_0}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}} \overset{H_0}{\sim} N(0,1)$$

Alternatively, we can compute confidence intervals:

$$\hat{\beta}_j \pm z_\alpha \times \sqrt{[\sigma^2(X'X)^{-1}]_{jj}}.$$

However, there is a problem:

$$Y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \underbrace{\sigma^2}\ I)$$

The value of σ^2 in unknown. We will solve this issue in two steps.

1) We will propose an estimator for σ^2

2) We will derive the distribution of $\hat{\beta}_j$ based on an unknown σ^2. This distribution will depend on the estimator of σ^2 and it will NOT be Gaussian.

4.1 Estimation of error variance

Step 1. We will propose an estimator. We want to estimate

$$\sigma^2 = E[\varepsilon_t^2]$$

from A4. We could use:

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2$$

This is unfeasible. A feasible estimator is:

$$\tilde{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$$

This is our proposed estimator. We want to check that is unbiased.

$$\tilde{\sigma}^2 = \frac{1}{T}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{T}\varepsilon' M' M\varepsilon = \frac{1}{T}\varepsilon' M\varepsilon$$

where we have used $\hat{\varepsilon} = MY = M\varepsilon$. The expectation is:

$$\begin{aligned}
E[\tilde{\sigma}^2] &= E[\frac{1}{T}\varepsilon' M\varepsilon] \\
&= \frac{1}{T}E[\underbrace{\varepsilon' M\varepsilon}_{1\times 1}] \\
&= \frac{1}{T}E[trace\{\varepsilon' M\varepsilon\}] \\
&= \frac{1}{T}E[trace\{\underbrace{M\varepsilon\varepsilon'}\}] \\
&= \frac{1}{T}trace\{E[M\varepsilon\varepsilon']\} \\
&= \frac{1}{T}trace\{M\,\underbrace{E[\varepsilon\varepsilon']}\} \\
&= \frac{1}{T}trace\{M\sigma^2 I\} \\
&= \frac{1}{T}\sigma^2 trace(M)
\end{aligned}$$

recall the matrix M

$$\begin{aligned} M &= I_T - X(X'X)^{-1}X' \\ trace\{M\} &= trace\{I_T\} - trace\{X(X'X)^{-1}X'\} \\ &= T - trace\{(X'X)^{-1}X'X\} \\ &= T - trace\{I_k\} \\ &= T - k \end{aligned}$$

so we have:

$$\begin{aligned} E[\tilde{\sigma}^2] &= \frac{1}{T}\sigma^2 trace(M) \\ &= \frac{1}{T}\sigma^2(T-k) \\ &= \frac{T-k}{T}\sigma^2 \end{aligned}$$

this is biased. We can adjust it:

$$E\left[\frac{T}{T-k}\tilde{\sigma}^2\right] = \frac{T}{T-k}E[\tilde{\sigma}^2] = \frac{T}{T-k}\frac{T-k}{T}\sigma^2 = \sigma^2$$

So the unbiased estimator is:

$$\hat{\sigma}^2 = \frac{T}{T-k}\tilde{\sigma}^2 = \frac{T}{T-k}\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 = \frac{1}{T-k}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$$

4.2 Distribution of OLS estimator when error variance is estimated

Now the second step is to use this estimator and derive the distribution of $\hat{\beta}$. We start from the result:

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma} \sim \chi^2_{T-k}$$

Now, note that $\frac{\varepsilon}{\sigma} \sim N(0, I_T)$. So this is a sum of squared normals, of which T-k are linearly independent. This is the sum of T-k independent standard Gaussians.

Remember the result we obtained:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Consider the following ratio:

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}} \sim N(0,1)}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}/(T-k)} \sim \sqrt{\chi^2_{T-k}/(T-k)}} \sim t_{T-k}$$

Both these results are based on M being a fixed matrix, which is the case under A2. Moreover, the degrees of freedom come precisely from the rank of M.

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}/(T-k)}} = \frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{(T-k)\sigma^2}}} = \frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{[(X'X)^{-1}]_{jj}}}}{\sqrt{\hat{\sigma}^2}} = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}}$$

we have a new pivotal quantity:

$$t = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}} \sim t_{T-k}$$

as opposed to:

$$z = \frac{\hat{\beta}_j-\beta_j}{\sqrt{[\sigma^2(X'X)^{-1}]_{jj}}} \sim N(0,1)$$

In practice we never know σ^2. For tests: $\mathrm{H}_0 : \beta_j = \beta_0$

$$\frac{\hat{\beta}_j-\beta_0}{\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}} \overset{H_0}{\sim} t_{T-k}$$

For C.I:

$$\hat{\beta}_j \pm t_\alpha \sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}$$

Softwares provide

$$\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}} \overset{H_0:\beta_j=0}{\sim} t_{T-k}$$

If T-k is large this approaches a Gaussian.

All this based on the distribution:

$$\hat{\beta}_j \sim t_{T-k}(\beta, [\hat{\sigma}^2(X'X)^{-1}]_{jj})$$

5 Partitioned Regressions and the Frish-Waugh Theorem

Recall the model:

$$\underset{T\times 1}{Y} = \underset{Txk}{X}\underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon}.$$

We are going to partition the regressors in 2 groups. The regressors are:

$$x_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,T} \end{bmatrix};\ x_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,T} \end{bmatrix};\ ...;\ x_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,T} \end{bmatrix};$$

and we grouped them in the matrix:

$$\underset{T\times k}{X} = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & & \vdots \\ x_{1,T} & x_{2,T} & \cdots & x_{k,T} \end{bmatrix}$$

Now we partition this matrix as follows:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \left[\underbrace{\begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{q,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{q,2} \\ \vdots & \vdots & & \vdots \\ x_{1,T} & x_{2,T} & \cdots & x_{q,T} \end{bmatrix}}_{X_1} \underbrace{\begin{bmatrix} x_{q+1,1} & \cdots & x_{k,1} \\ x_{q+1,2} & \cdots & x_{k,2} \\ \vdots & & \vdots \\ x_{q+T,2} & \cdots & x_{k,T} \end{bmatrix}}_{X_2} \right]$$

And accordingly:

$$\beta = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix},\ \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_q \end{bmatrix};\ \boldsymbol{\beta}_2 = \begin{bmatrix} \beta_{q+1} \\ \beta_{q+2} \\ \cdots \\ \beta_k \end{bmatrix}$$

Then we can rewrite our model:

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \varepsilon$$

and if you perform the product:

$$Y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \varepsilon$$

in X_1 there are q regressors, and in X_2 there are k-q regressors. We start from the normal equations.

$$X'X\hat{\beta} = X'Y$$

In the alternative notation:

$$\begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} Y$$

Do the multiplication:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix}$$

then,

$$\begin{aligned} X_1'X_1\hat{\boldsymbol{\beta}}_1 + X_1'X_2\hat{\boldsymbol{\beta}}_2 &= X_1'Y \\ X_2'X_1\hat{\boldsymbol{\beta}}_1 + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y \end{aligned}$$

Solve the first equation for $\hat{\boldsymbol{\beta}}_1$

$$\begin{aligned} (X_1'X_1)^{-1}X_1'X_1\hat{\boldsymbol{\beta}}_1 &= (X_1'X_1)^{-1}(X_1'Y - X_1'X_2\hat{\boldsymbol{\beta}}_2) \\ \hat{\boldsymbol{\beta}}_1 &= (X_1'X_1)^{-1}X_1'(Y - X_2\hat{\boldsymbol{\beta}}_2) \end{aligned}$$

Now we can plug this expression in the second equation and solve for $\hat{\boldsymbol{\beta}}_2$

$$\begin{aligned} X_2'X_1\hat{\boldsymbol{\beta}}_1 + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y \\ X_2'X_1\{(X_1'X_1)^{-1}X_1'(Y - X_2\hat{\boldsymbol{\beta}}_2)\} + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y \\ X_2'X_1(X_1'X_1)^{-1}X_1'Y - X_2'X_1\{(X_1'X_1)^{-1}X_1'X_2\hat{\boldsymbol{\beta}}_2 + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y \\ X_2'P_1Y - X_2'P_1X_2\hat{\boldsymbol{\beta}}_2 + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y \\ -X_2'P_1X_2\hat{\boldsymbol{\beta}}_2 + X_2'X_2\hat{\boldsymbol{\beta}}_2 &= X_2'Y - X_2'P_1Y \\ X_2'[I - P_1]X_2\hat{\boldsymbol{\beta}}_2 &= X_2'[I - P_1]Y \\ X_2'M_1X_2\hat{\boldsymbol{\beta}}_2 &= X_2'M_1Y \\ \hat{\boldsymbol{\beta}}_2 &= (X_2'M_1X_2)^{-1}X_2'M_1Y \end{aligned}$$

The same steps can be done to obtain a solution for $\hat{\boldsymbol{\beta}}_1$. We have:

$$\begin{aligned}
\underset{q\times 1}{\hat{\boldsymbol{\beta}}_1} &= (\underset{q\times T}{X_1'}\underset{T\times T}{M_2}\underset{T\times q}{X_1})^{-1}X_1'M_2Y \\
\underset{(k-q)\times 1}{\hat{\boldsymbol{\beta}}_2} &= (\underset{(k-q)\times T}{X_2'}\underset{T\times T}{M_1}\underset{T\times(k-q)}{X_2})^{-1}X_2'M_1Y
\end{aligned}$$

Consider the case in which X_1 and X_2 are orthogonal: $X_1'X_2 = 0$. This implies $X_1'M_2 = X_1'[I - X_2(X_2'X_2)X_2'] = X_1 - 0 = X_1$. Also we have that $X_2'M_1 = X_2'[I - X_1(X_1'X_1)^{-1}X_1'] = X_2'$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= (X_1'X_1)^{-1}X_1'Y \\
\hat{\boldsymbol{\beta}}_2 &= (X_2'X_2)^{-1}X_2'Y
\end{aligned}$$

Remark 1 *When the regressors are orthogonal, then the least square regression of Y on the two separate sets of variables gives the same OLS estimator as regressing Y on both groups at the same time.*

What else can we say?

$$\hat{\boldsymbol{\beta}}_2 = (X_2'\underbrace{M_1X_2}_{})^{-1}X_2'M_1Y$$

What is M_1X_2?

$$X_2 = X_1\gamma + \eta$$

The OLS estimator is:

$$\begin{aligned}
\hat{\gamma} &= (X_1'X_1)^{-1}X_1'X_2 \\
\hat{X}_2 &= X_1\hat{\gamma} = X_1(X_1'X_1)^{-1}X_1'X_2 = P_1X_2 \\
\hat{\eta} &= X_2 - \hat{X}_2 = X_2 - P_1X_2 = M_1X_2
\end{aligned}$$

So we note that:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= (X_2'M_1\underbrace{M_1X_2}_{\hat{\eta}})^{-1}X_2'M_1Y \\
&= (\hat{\eta}'\hat{\eta})^{-1}\hat{\eta}'Y,
\end{aligned}$$

i.e. the OLS estimator of the slope coefficient in the regression:

$$Y = \hat{\eta}\delta + v.$$

This is the Frish-Waugh Theorem

Proposition 2 *In the linear regression of Y on two sets of variables X_1 and X_2 the subvector $\hat{\boldsymbol{\beta}}_2$ is the set of coefficients obtained from regression of Y on the residuals $\hat{\eta}$, where $\hat{\eta}$ are the residuals of a regression of X_2 onto X_1.*

This extends to the case of individual regressors:

$$Y = x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k + \varepsilon$$

Remark 3 *The OLS estimator $\hat{\beta}_j$ in a multiple regression measures the effect of the regressor x_j on Y, after washing away the correlation of that regressor with the remaining regressors $x_1, x_2, ...x_{j-1,}x_{j+1,}...., x_k$.*

6 Inference on multiple coefficients and restricted least squares

6.1 Specifying multiple linear restrictions

Consider the model:

$$Y = X\beta + \varepsilon$$

full model:

$$Y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

restricted model:

$$Y = x_1\beta_1 + \varepsilon$$

The restriction is $\beta_2 = 0$.

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ 0 \end{bmatrix}$$

Generalize.

$$Y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \varepsilon$$

Imagine we want to impose the following restrictions:

$$\begin{aligned} a)\ \beta_1 &= 0 \Leftrightarrow \beta_1 = 0 \\ b)\ \beta_2 &= 1 \Leftrightarrow \beta_2 - 1 = 0 \\ c)\ \beta_2 + \beta_3 &= 5 \Leftrightarrow \beta_2 + \beta_3 - 5 = 0 \end{aligned}$$

We want to use the general representation:

$$\underset{j\times k}{R}\ \underset{k\times 1}{\beta} = \underset{j\times 1}{r}$$

In our example, we have j=3 and k=3

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix}$$

Now, we write them as follows:

$$m = R\beta - r = 0 \tag{5}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The vector m is the vector measuing the distance from the restrictions.

Let us give a few more examples of restrictions and put them in the form (5).

$$H_0 : \ \beta_i = 0$$

$$\begin{aligned} R &= [\ 0 \quad 0 \quad 0 \quad 1 \quad ... \quad 0\ ];\ r = 0 \\ m &= [\ 0 \quad 0 \quad 0 \quad 1 \quad ... \quad 0\ ] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \\ \beta_k \end{bmatrix} - 0 = 0 = \ \beta_i - 0 = 0 \end{aligned}$$

$$\begin{aligned} H_0 \quad &: \quad \beta_i = \beta_{i+1} \\ \beta_i - \beta_{i+1} \quad &= \quad 0 \end{aligned}$$

$$\begin{aligned} R &= [\ 0 \quad 0 \quad ... \quad 1 \quad -1 \quad ...\ ]\ ; r = 0 \\ m &= [\ 0 \quad 0 \quad ... \quad 1 \quad -1 \quad ...\ ] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \beta_{i+1} \\ \\ \beta_k \end{bmatrix} - 0 = 0 \end{aligned}$$

$$H_0 : \beta_2 + \beta_3 + \beta_4 = 0$$

$$R = \begin{bmatrix} 0 & 1 & 1 & 1 & ... & 0 \end{bmatrix}; r = 0$$

$$m = \begin{bmatrix} 0 & 1 & 1 & 1 & ... & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \beta_{i+1} \\ \\ \beta_k \end{bmatrix} - 0 = 0$$

Let us consider multiple restrictions

$$H_0 : \beta_1 = 0;\ \beta_2 = 0;\ \beta_3 = 0$$

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & 0 & ... & 0 \end{bmatrix};\ r = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$m = \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & 0 & ... & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \beta_{i+1} \\ \\ \beta_k \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Another one

$$H_0 : \beta_2 + \beta_3 = 1;\ \beta_4 + 3\beta_5 = 0;\ \beta_5 + \beta_6 = 0$$

$$R = \begin{bmatrix} 0 & 1 & 1 & 0 & ... & ... & .. & 0 \\ 0 & 0 & 0 & 1 & 3 & .. & .. & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & .. & .. \end{bmatrix}; r = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$m = \begin{bmatrix} 0 & 1 & 1 & 0 & ... & ... & .. & 0 \\ 0 & 0 & 0 & 1 & 3 & .. & .. & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & .. & .. \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \beta_{i+1} \\ \\ \beta_k \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

6.2 Wald statistic

The test starts from the discrepancy vector:

$$m = R\beta - r \stackrel{H_0}{=} 0$$

Imagine that we estimate the model

$$Y = X\beta + \varepsilon$$

getting $\hat{\beta}$. Then we can compute the estimate of the discrepancy vector

$$\hat{m} = R\hat{\beta} - r$$

The expected value of the estimated discrepancy is:

$$\begin{aligned} E[\hat{m}] &= E[R\hat{\beta} - r] \\ &= RE[\hat{\beta}] - r \\ &= R\beta - r = m \stackrel{H_0}{=} 0 \end{aligned}$$

Let us compute the variance:

$$\begin{aligned} Var[\hat{m}] &= Var[R\hat{\beta} - r] \\ &= Var[R\hat{\beta}] \\ &= RVar[\hat{\beta}]R' \\ &= R\sigma^2(X'X)^{-1}R' \end{aligned}$$

How can we devise a test for H_0? Imagine you have 3 restrictions

$$\begin{aligned} \beta_2 + \beta_3 - 1 &= 0; \\ \beta_4 + 3\beta_5 &= 0; \\ \beta_5 + \beta_6 &= 0 \end{aligned}$$

and after estimation, you find:

$$\begin{aligned} \hat{\beta}_2 + \hat{\beta}_3 - 1 &= 0.9 \\ \hat{\beta}_4 + 3\hat{\beta}_5 &= 0.0000001 \\ \hat{\beta}_5 + \hat{\beta}_6 &= 0.008 \end{aligned}$$

how can we decide how "well" these restrictions are satisfied? The answer is the Wald criterion. Divide the discrepancy by its standard deviation, to obtain a rescaled discrepancy:

$$\left(\sqrt{Var(\hat{m})}\right)^{-1}\hat{m} = \left(\sqrt{Var(\hat{m})}\right)^{-1}(R\hat{\beta} - r).$$

The rescaled discrepancy, is Gaussian and has variance I_j. The Wald criterion is:

$$\begin{aligned} W &= \hat{m}' Var(\hat{m})^{-1}\hat{m} \\ &= \hat{m}'[R\sigma^2(X'X)^{-1}]^{-1}R'\hat{m} \\ &= (R\hat{\beta} - r)'[R\sigma^2(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \\ &= \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \\ &= \underbrace{\frac{(R\hat{\beta} - r)'}{\sigma}}_{1\times j}\underset{j\times j}{[R(X'X)^{-1}R']^{-1}}\underbrace{\frac{(R\hat{\beta} - r)}{\sigma}}_{j\times 1} \sim \chi^2_j \end{aligned}$$

However, σ is unknown and needs to be estimated. The estimator is

$$\hat{\sigma} = \frac{1}{T-k}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 = \frac{1}{T-k}\hat{\varepsilon}'\hat{\varepsilon}.$$

Moreover, recall that

$$\frac{\hat{\sigma}^2}{\sigma^2}(T-k) = \frac{1}{\sigma^2}\hat{\varepsilon}'\hat{\varepsilon} = \frac{\varepsilon'}{\sigma}M\frac{\varepsilon}{\sigma} \sim \chi^2_{T-k}$$

Therefore:

$$\frac{\frac{(R\hat{\beta}-r)'}{\sigma}[R(X'X)^{-1}R']^{-1}\frac{(R\hat{\beta}-r)}{\sigma}/j \sim \chi^2_j}{\frac{\hat{\sigma}^2}{\sigma^2}(T-k)/(T-k) \sim \chi^2_{T-k}} \sim F_{j,T-k}$$

Simplifying:

$$\frac{(R\hat{\beta} - r)'[R\hat{\sigma}^2(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{j} \sim F_{j,T-k}$$

This allows to test for the restrictions.

6.3 Restricted least squares

Imagine we want to impose these restrictions on the model.

$$\begin{aligned} Y &= X\beta + \varepsilon \\ R\beta - r &= 0 \end{aligned}$$

this model can also be estimated via Least squares. The solution is:

$$\hat{\beta}_{RLS} = \hat{\beta}_{OLS} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{OLS} - r).$$

derive the RLS

$$\begin{aligned} \beta_R &= \arg\min S(\beta) \\ S(\beta) &= (Y - X\beta)'(Y - X\beta) \\ s.t.\ R\beta &= r \end{aligned}$$

$$L(\beta, \lambda) = S(\beta) + 2\lambda'\underbrace{(R\beta - r)}_{m}$$

lets solve this

$$\begin{aligned} \frac{\partial L(\beta,\lambda)}{\partial \beta} &= -2X'(Y - X\beta) + 2R'\lambda \\ &= -2X'(Y - X\hat{\beta}_R) + 2R'\lambda = 0 \\ &= -2X'(Y - X\hat{\beta}_R) + 2R'\lambda/\Theta = 0 \\ \\ &= -X'(Y - X\hat{\beta}_R) + R'\lambda = 0 \\ &= -X'Y + X'X\hat{\beta}_R + R'\lambda = 0 \\ &= X'X\hat{\beta}_R = X'Y - R'\lambda \\ &= \hat{\beta}_R = (X'X)^{-1}X'Y - (X'X)^{-1}R'\lambda \\ &= \hat{\beta}_R = \hat{\beta} - (X'X)^{-1}R'\lambda \end{aligned}$$

$$\begin{aligned} \frac{\partial L(\beta,\lambda)}{\partial \lambda} &= 2(R\beta - r) \\ &= 2(R\hat{\beta}_R - r) = 0 \\ &= R\hat{\beta}_R = r \end{aligned}$$

now lets substitute

$$\begin{aligned} R(\hat{\beta} - (X'X)^{-1}R'\lambda) &= r \\ R\hat{\beta} - R(X'X)^{-1}R'\lambda &= r \\ R\hat{\beta} - r &= R(X'X)^{-1}R'\lambda \\ \lambda &= [R(X'X)^{-1}R']^{-1}\underbrace{(R\hat{\beta} - r)} \end{aligned}$$

now we put this equation into the first one

$$\hat{\beta}_R = \hat{\beta} - \underbrace{(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}}_{C} \underbrace{(R\hat{\beta} - r)}_{\hat{m}}$$

and this is the RLS estimator. Check the SOC.

$$H_0 : R\beta - r = 0$$

3. Unbiasedness:

$$\begin{aligned} E[\hat{\beta}_R] &= E[\hat{\beta} - C(R\hat{\beta} - r)] \\ &= E[\hat{\beta}] - E[C(R\hat{\beta} - r)] \\ &= E[(X'X)^{-1}X'\varepsilon + \beta] - E[C(R\hat{\beta} - r)] \\ &= \beta + 0 + -E[C(R\hat{\beta} - r)] \\ &= \beta - C\underbrace{E[R\hat{\beta} - r]}\overset{H_0}{=} \beta - C0 = \beta \end{aligned}$$

instead, if H_0 is not satisfied

$$E[\hat{\beta}_R] = \beta - CE[R\beta - r] \neq \beta$$

Variance of $\hat{\beta}_R$:

$$\begin{aligned} \hat{\beta}_R &= \hat{\beta} - \underbrace{(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}}_{C} \underbrace{(R\hat{\beta} - r)}_{\hat{m}} \\ &= \hat{\beta} - C(R\hat{\beta} - r) \\ &= (I - CR)\hat{\beta} + Cr \end{aligned}$$

since $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ we have

$$\begin{aligned} Var(\hat{\beta}_R) &= Var[(I - CR)(\beta + (X'X)^{-1}X'\varepsilon) + Cr] \\ &= Var[(I - CR)((X'X)^{-1}X'\varepsilon)] \\ &= M_R Var((X'X)^{-1}X'\varepsilon)M_R' \\ &= M_R\sigma^2(X'X)^{-1}M_R'. \end{aligned}$$

where we defined $M_R = I - CR$. Finally, one can show that

$$M_R\sigma^2(X'X)^{-1}M_R' = M_R\underbrace{\sigma^2(X'X)^{-1}}$$

which gives $Var(\hat{\beta}_R) = M_R Var(\hat{\beta})$ and

$$\begin{aligned} V[\hat{\beta}_{RLS}] - Var(\hat{\beta}_{OLS}) &= (M_r - I)Var(\hat{\beta}_{OLS}) = CVar(\hat{\beta}_{OLS}) \\ &= -(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\sigma^2(X'X)^{-1}, \end{aligned}$$

a negative definite matrix.

To summarize, under H_0:

$$\begin{aligned} E[\hat{\beta}_{RLS}] &= E[\hat{\beta}_{OLS} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{OLS} - r)] \\ &= \beta - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}E[R\hat{\beta}_{OLS} - r] \end{aligned}$$

since under H_0 we have that $E[R\hat{\beta}_{OLS} - r] = 0$, $E[\hat{\beta}_{RLS}] \overset{H_0}{=} \beta$ and the $\hat{\beta}_{RLS}$ is unbiased and more efficient than $\hat{\beta}_{OLS}$. However, if the null H_0 is not satisfied, then the $\hat{\beta}_{RLS}$ has expectation:

$$\begin{aligned} E[\hat{\beta}_{RLS}] &= \beta - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}E[R\hat{\beta}_{OLS} - r] \neq 0 \\ \text{since } E[R\hat{\beta}_{OLS} - r] &\neq 0 \text{ under } H_0. \end{aligned}$$

so it is biased.

Remark 4 *Under H_0 the $\hat{\beta}_{RLS}$ estimator is unbiased and more efficient than $\hat{\beta}_{OLS}$. However, if H_0 is not satisfied, $\hat{\beta}_{RLS}$ is biased (while $\hat{\beta}_{OLS}$ is unbiased).*

7 The generalized linear regression model

recall the assumption CLRM:

- A1 - X is a matrix of full rank —> OLS estimator exists
- A2 - X is deterministic or fixed in repeated samples
- A3 - E[ε]=0
- A4 - E[$\varepsilon\varepsilon'$]=$\sigma^2 I_T$
- A5 - ε is Gaussian

Under these assumptions:

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(X'X)^{-1})$$

and:

$$\hat{\beta}_{OLS} \sim t(\beta, \hat{\sigma}^2(X'X)^{-1})$$

can be used for inference.

Now we start asking what happens as we remove some assumptions. Today's lecture focuses A4.

$$E[\varepsilon\varepsilon'] = \sigma^2 I_T$$

Sphericity.

$$Var(\varepsilon) = E[\varepsilon\varepsilon'] - E[\varepsilon]E[\varepsilon'] = \sigma^2 I_T$$

$$\sigma^2 I_T = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & & & \\ 0 & & & & \\ 0 & & & & \\ & & & & \sigma^2 \end{bmatrix}$$

$$\begin{aligned} Var(\varepsilon_t) &= \sigma^2 \\ Cov(\varepsilon_t, \varepsilon_s) &= 0 \end{aligned}$$

We are going to generalize this model.

$$\begin{aligned} A5bis.\ E[\varepsilon\varepsilon'] &= \sigma^2\Omega = \Sigma \\ &= \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & Cov(\varepsilon_1,\varepsilon_3) & ... & Cov(\varepsilon_1,\varepsilon_T) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & & & \\ Cov(\varepsilon_3,\varepsilon_1) & & & & \\ & & & & \\ Cov(\varepsilon_T,\varepsilon_1) & & & & Var(\varepsilon_T) \end{bmatrix} \end{aligned}$$

More generally:

$$E[\varepsilon\varepsilon'] = \sigma^2\Omega$$

The model with assumption $E[\varepsilon\varepsilon'] = \sigma^2\Omega$ is called the "generalized" linear regression model. Setting $\Omega = I$ gets us back to the CLRM.

1. Ask how the OLS estimator performs under this alternative set of assumptions.
2. Propose a more efficient estimator under these assumptions.

7.1 Properties of OLS estimator and HAC standard errors

First of all, lets consider bias.

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'\varepsilon + \beta \\
E[\hat{\beta}_{OLS}] &= E[(X'X)^{-1}X'\varepsilon + \beta] \\
&= (X'X)^{-1}X\underbrace{E[\varepsilon]}_{=0} + \beta \\
&= 0 + \beta
\end{aligned}$$

Then, consider the variance of the OLS estimator

$$\begin{aligned}
Var[\hat{\beta}_{OLS}] &= Var[(X'X)^{-1}X'\varepsilon + \beta] \\
&= Var[(X'X)^{-1}X'\varepsilon] \\
&= (X'X)^{-1}X'\underbrace{Var[\varepsilon]}_{\sigma^2\Omega}X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \neq Var[\hat{\beta}_{OLS}|CLRM] = \sigma^2(X'X)^{-1}
\end{aligned}$$

Therefore we have

$$Var[\hat{\beta}_{OLS}|GLRM] \neq Var[\hat{\beta}_{OLS}|CLRM]$$

Therefore, in presence of nonspherical disturbances the OLS standard errors are *incorrect*. Therefore, in this case we use

$$\begin{aligned}
Var[\hat{\beta}_{OLS}] &= \underbrace{\sigma^2}(X'X)^{-1}X'\underbrace{\Omega}X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1}
\end{aligned}$$

Consider

$$\Sigma = \sigma^2\Omega = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & Cov(\varepsilon_1,\varepsilon_3) & ... & Cov(\varepsilon_1,\varepsilon_T) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & & & \\ Cov(\varepsilon_3,\varepsilon_1) & & & & \\ & & & & \\ Cov(\varepsilon_T,\varepsilon_1) & & & & Var(\varepsilon_T) \end{bmatrix}$$

$$\frac{(T^2-T)}{2} + T = \frac{T^2-T+2T}{2} = \frac{T(T+1)}{2}$$

We try to estimate

$$X'\sigma^2\Omega X$$

which has $\frac{K(K+1)}{2}$.

White estimator (1980):

$$X'\widehat{\sigma^2\Omega}X = X'\begin{bmatrix} \hat{\varepsilon}_1^2 & 0 \\ & \\ 0 & \hat{\varepsilon}_T^2 \end{bmatrix} X = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \mathbf{x}_t\mathbf{x}_t'$$

Newey-West (1982):

$$\begin{aligned} X'\widehat{\sigma^2\Omega}X &= \sum_{p=-p^*}^{p^*}\sum_{t=p+1}^{T}\left(1-\frac{|p|}{1+p^*}\right)\hat{\varepsilon}_t\hat{\varepsilon}_{t-p}\mathbf{x}_t\mathbf{x}_{t-p}' \\ &= \sum_{t=1}^{T}\hat{\varepsilon}_t^2\mathbf{x}_t\mathbf{x}_t' + \sum_{p=1}^{p^*}\sum_{t=p+1}^{T}\left(1-\frac{p}{1+p^*}\right)\hat{\varepsilon}_t\hat{\varepsilon}_{t-p}(\mathbf{x}_t\mathbf{x}_{t-p}' + \mathbf{x}_{t-p}\mathbf{x}_t') \end{aligned}$$

with $p^* \approx T^{1/4}$. These are called Heteroschedasticity and autocorrelation consistent (HAC) robust estimators of the variance.

$$\widehat{Var[\hat{\beta}_{OLS}]} = (X'X)^{-1}X'\widehat{\sigma^2\Omega}X(X'X)^{-1}$$

then we can use

$$\hat{\beta}_{OLS} \tilde{} t(\beta, \widehat{Var[\hat{\beta}_{OLS}]})$$

for inference.

7.2 Efficient estimation via GLS

There is an alternative to the OLS estimator, which is actually more efficient.

$$\Omega^{-1} = P'P = \Omega^{-1/2'}\Omega^{-1/2}$$

The matrix P can be used to "rescale" the model.

$$Y = X\beta + \varepsilon$$

imagine pre-multiplying this by P:

$$\begin{aligned} PY &= PX\beta + P\varepsilon \\ Y^* &= X^*\beta + \varepsilon^* \end{aligned}$$

Let's compute the variance of the errors of the transformed model

$$\begin{aligned} Var(\varepsilon^*) &= Var(P\varepsilon) \\ &= PVar(\varepsilon)P' \\ &= P\sigma^2\Omega P' \\ &= \sigma^2 P\Omega P' \\ &= \sigma^2 P(P^{-1}P'^{-1})P' \\ &= \sigma^2 I \end{aligned}$$

where we used $\Omega = (\Omega^{-1})^{-1} = (P'P)^{-1} = P^{-1}P'^{-1}$. It follows that the transformed model

$$Y^* = X^*\beta + \varepsilon^*$$

satisfies the CLRM assumptions. And the OLS estimator of the transformed model:

$$\hat{\beta}^*_{OLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$

since this is the OLS estimator of a model that satisfies the CLRM assumptions, this is BLUE for β. We label this estimator the GLS estimator

$$\hat{\beta}_{GLS} = \hat{\beta}^*_{OLS}$$

In particular

$$\begin{aligned} \hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \\ &= (X'P'PX)^{-1}(PX)'PY \\ &= (X'P'PX)^{-1}X'P'PY \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y. \end{aligned}$$

$$\begin{aligned} E[\hat{\beta}_{GLS}] &= \beta \\ Var[\hat{\beta}_{GLS}] &= \sigma^2(X'\Omega^{-1}X)^{-1} \end{aligned}$$

and it is BLUE for β, so it is better than any other linear unbiased estimator, including the OLS estimator:

$$Var[\hat{\beta}_{GLS}|GLRM] = \sigma^2(X'\Omega^{-1}X)^{-1} < \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} = Var[\hat{\beta}_{OLS}|GLRM]$$

7.3 Feasible GLS

As it happen for the OLS estimator, we need to estimate $\sigma^2$and Ω^{-1}. We assume a functional form for Ω

$$\Omega = \Omega(\theta)$$

and estimate the subset of coefficients θ

$$\hat{\Omega} = \widehat{\Omega(\theta)}$$

generally, one assumes a functional form for the error terms and uses Maximum Likelihood.

$$\hat{\beta}_{FGLS} = (X'\widehat{\Omega^{-1}(\theta)}X)^{-1}X'\widehat{\Omega^{-1}(\theta)}Y$$

where F stands for "feasible". The FGLS is not BLUE. It is actually both biased and nonlinear. However

$$\hat{\beta}_{FGLS} \longrightarrow \hat{\beta}_{GLS}$$

So in a large sample, we converge to the unfeasible, optimal GLS estimator.

7.3.1 Heteroskedasticity and weighted least squares

For example let us assume that :

$$y_t = \alpha + \beta x_t + \varepsilon_t;\ \varepsilon_t \sim N(0, \sigma_t^2)$$

with

$$\sigma_t^2 = \lambda x_t^2$$

$$\sigma^2\Omega = \begin{bmatrix} E[\varepsilon_1^2] = \sigma_1^2 = \lambda x_1^2 & & 0 \\ 0 & E[\varepsilon_2^2] = \sigma_2^2 = \lambda x_2^2 & \\ & & \\ 0 & 0 & E[\varepsilon_T^2] = \sigma_T^2 = \lambda x_T^2 \end{bmatrix} = \lambda \begin{bmatrix} x_1^2 & & 0 \\ 0 & x_2^2 & \\ & & \ddots \\ 0 & & x_T^2 \end{bmatrix}$$

transform the model:

$$\begin{aligned}
\underbrace{\frac{1}{\sqrt{x_t^2}}y_t}_{y_t^*} &= \underbrace{\frac{1}{\sqrt{x_t^2}}\alpha}_{x_t^*} + \frac{1}{\sqrt{x_t^2}}\beta x_t + \underbrace{\frac{1}{\sqrt{x_t^2}}\varepsilon_t}_{\varepsilon_t^*} \\
y_t^* &= \alpha x_t^* + \beta + \varepsilon_t^*; \\
Var(\varepsilon_t^*) &= Var(\frac{1}{\sqrt{x_t^2}}\varepsilon_t) = \frac{1}{x_t^2}\lambda x_t^2 = \lambda
\end{aligned}$$

This is FGLS with $\theta = \{\lambda\}$ and $P = \begin{bmatrix} 1/x_1 & & 0 \\ 0 & 1/x_2 & \\ & & \\ 0 & 0 & 1/x_T \end{bmatrix}$. This exampe with heteroskedasticity is a.k.a Weighted Least Squares.

7.3.2 Autocorrelation and Cochrane-Orcutt procedure

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

$$\varepsilon_t = \varrho\varepsilon_{t-1} + u_t$$

$$Var[\varepsilon_t] = \frac{\sigma_u^2}{1-\varrho^2} = \sigma^2$$

$$\sigma^2\Omega = \frac{\sigma_u^2}{1-\varrho^2}\begin{bmatrix} 1 & \varrho & \varrho^2 & \varrho^{T-1} \\ \varrho & 1 & & \\ \varrho^2 & \varrho & & \\ & & & \\ \varrho^{T-1} & 0 & & 1 \end{bmatrix} = \sigma^2\Omega(\theta)$$

This is FGLS with $\theta = \{\varrho, \sigma_u^2\}$. Cochrane-Orcutt procedure:

$$y_t = \alpha + \beta x_t + \varepsilon_t \rightarrow \hat{\beta} \rightarrow \hat{\varepsilon}_t$$

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t \rightarrow \hat{\rho}$$

and iterate:

$$\begin{aligned}
y_t &= \alpha + \beta x_t + \varepsilon_t \\
\hat{\rho}y_{t-1} &= \hat{\rho}\alpha + \hat{\rho}\beta x_{t-1} + \hat{\rho}\varepsilon_{t-1}
\end{aligned}$$

$$y_t - \hat{\rho}y_{t-1} = (1-\hat{\rho})\alpha + \beta(x_t - \hat{\rho}x_{t-1}) + \underbrace{\varepsilon_t - \hat{\rho}\varepsilon_{t-1}}_{u_t}$$

7.3.3 Conditional heteroskedasticity and models of time varying volatilities

$$y_t = \alpha + \beta x_t + \varepsilon_t$$
$$\varepsilon_t \sim N(0, \sigma_t^2)$$

model ε_t^2 as follows:

$$\begin{aligned} \varepsilon_t^2 &= h_t^2 v_t^2, \\ v_t &\sim iidN(0,1) \\ h_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \text{ (ARCH 1)} \end{aligned}$$

this implies:

$$E[\varepsilon_t^2 | I_{t-1}] = E[h_t^2 v_t^2 | I_{t-1}] = h_t^2 E[v_t^2 | I_{t-1}] = h_t^2,$$

which shows that h_t^2 is the conditional heteroskedasticity. Note that the forecast error is:

$$\begin{aligned} u_t &= \varepsilon_t^2 - E[\varepsilon_t^2 | I_{t-1}] \\ &= \varepsilon_t^2 - h_t^2 \\ &\Longrightarrow \\ \varepsilon_t^2 &= h_t^2 + u_t \\ &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + u_t \end{aligned}$$

that is, the squared errors follow an AR(1) process. This model could be estimated with ML or with a recursive approach similar to Cochrane-Orcutt.

Stochastic volatility:

$$h_t^2 = \alpha_0 + \alpha_1 h_{t-1}^2 + \eta_t$$

note that differently from GARCH h_t^2 has its own disturbance. Estimated with ML, or Bayesian methods.

Part II
Random regressors and endogeneity

8 Independent regressors

So far we assumed:

$$Y = X \underbrace{\beta}_{unkown} + \underbrace{\varepsilon}_{\text{rand}\,om}$$

However, in economics X is random. There is the need to specify the relation between the two random objects in the model. The easiest case is one in which X and ε are independent.

- A1 - X is a matrix of full rank —> OLS estimator exists
- A2bis - X is random
- A3bis - E[$\varepsilon|X$]=0
- A4bis - E[$\varepsilon\varepsilon'|X$]=$\sigma^2 I_T$
- A5bis - $\varepsilon|X$ is Gaussian

  This set of assumptions ensures that X is random (A2) but independent from ε (because of A3-A5). Under these assumptions, it turns out that the OLS estimator is still BLUE.

8.1 The law of iterated expectations.

E[points vs team A]=3*2/10+1*3/10+0*5/10=6/10+3/10=9/10
E[points vs team B]=3*4/10+1*4/10+0*2/10=12/10+4/10+0=16/10

$$\begin{aligned} E[Points|X &= A] = 0.9 \\ E[points|X &= B] = 1.6 \end{aligned}$$

E[points]=E[points vs team A]*prob(playing with A)+E[points vs team B]*prob(playing with B)

$$E[points] = E[Points|A] * p(A) + E[Points|A] * p(B) +$$

We computed the expected values of the expected values.

$$\begin{aligned} E[points] &= E_X[E[Points|X]] \\ E[Z] &= E_X[E[Z|X]] \end{aligned}$$

We are going to change some assumptions

8.2 Bias and variance

- First, let us consider unbiasedness.

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'\varepsilon + \beta$$

First, we show that the estimator is *conditionally* unbiased.

$$\begin{aligned} E[\hat{\beta}|X] &= E[(X'X)^{-1}X'\varepsilon + \beta|X] \\ &= E[(X'X)^{-1}X'\varepsilon|X] + E[\beta|X] \\ &= (X'X)^{-1}X'\underbrace{E[\varepsilon|X]}_{=0} + \beta \\ &= \beta \end{aligned}$$

$$\begin{aligned} E[\varepsilon|X] &= 0 \Longrightarrow E[\varepsilon] = 0 \\ E[\varepsilon] &= E_X[\underbrace{E[\varepsilon|X]}] = E_X[0] = 0 \end{aligned}$$

How about unconditionally?

$$E[\beta] = E_X[E[\hat{\beta}|X]] = E_X[\beta] = \beta$$

therefore, it is unbiased.

Then we have also the variance.

$$\begin{aligned} Var(\hat{\beta}|X) &= Var((X'X)^{-1}X'\varepsilon + \beta|X) \\ &= Var((X'X)^{-1}X'\varepsilon|X) \\ &= (X'X)^{-1}X'\underbrace{Var(\varepsilon|X)}X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_T X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

which is the usual. Since

- The LIE is: $E[Z] = E_X[E[Z|X]]$
- The definition of variance is $Var(Z) = E[Z^2] - E[Z]^2$
  Imply $Var(Z) = E_X Var(Z|X) + Var_x E[Z|X]$, we have that:

$$\begin{aligned} Var(\hat{\beta}) &= E_X[Var(\hat{\beta}|X)] + Var_X \underbrace{(E[\hat{\beta}|X])} \\ &= E_X[Var(\hat{\beta}|X)] = E_X[\sigma^2(X'X)^{-1}] \end{aligned}$$

this is the variance.

8.3 Distribution of the OLS estimator and Gauss-Markov theorem

Therefore, the OLS estimator is NOT a Gaussian.

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$$

the average of the Gaussians above is NOT a Gaussian.

$$\hat{\beta} \sim ?$$

Consider the residuals:

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}|X = \frac{\varepsilon' M \varepsilon}{\sigma^2}|X \sim \chi^2_{T-k}$$

Moreover

$$\frac{\frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}}|X \sim N(0,1)}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}|X /(T-k)} \sim \chi^2_{T-k}} \sim t_{T-k}$$

and

$$\frac{(R\hat{\beta} - r)'[R\hat{\sigma}^2(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{j}|X \sim F_{j,T-k}$$

have densities that do not depend on X, hence they hold for every X which means once we integrate the X out they remain the same.

This means that we can make inference as usual, using the pivotal quantities above.

Finally, the same reasoning applies to the Gauss-Markov theorem.

$$Var(\hat{\beta}_{OLS}|X) \ < Var(\tilde{\beta}|X)$$

$$Var(\tilde{\beta}|X) = \sigma^2 DD' + \underbrace{\sigma^2 (X'X)^{-1}}_{Var(\hat{\beta}_{OLS}|X)},$$

Now, apply the LIE:

$Var(\tilde{\beta}) = E_X[Var(\tilde{\beta}|X)] = \sigma^2 E_X[DD'] + E_X[Var(\hat{\beta}_{OLS}|X)] = \sigma^2 E_X[DD'] + Var(\hat{\beta}_{OLS})$

hence $Var(\tilde{\beta}) - Var(\hat{\beta}_{OLS}) = \sigma^2 E_X[DD']$ is positive definite.

9 Asymptotic theory and large sample properties of the OLS estimator

So far X was random, but independent from ε

- A1 - X is a matrix of full rank —> OLS estimator exists
- A2bis - X is random
- A3bis - E[$\varepsilon|X$]=0
- A4bis - E[$\varepsilon\varepsilon'|X$]=$\sigma^2 I_T$
- A5bis - $\varepsilon|X$ is Gaussian

$$\varepsilon|X \sim N(0, \sigma^2 I_T)$$

What if X is not independent from ε? Small sample results break down. For example:

$$\begin{aligned} E[\hat{\beta}] &= \beta + E[(X'X)^{-1}X'\varepsilon] \\ &= \beta + E_X[E[(X'X)^{-1}X'\varepsilon|X]] \\ &= \beta + E_X[(X'X)^{-1}X'E[\varepsilon|X]] \end{aligned}$$

We must resort to asymptotics. As a silver lining, the Gaussianity assumption will no longer be necessary. Actually, one could invoke asymptotics even with independence between X and ε, just on the grounds of removing A5, which can be viewed as strong. In that case, removing A5 while keeping A3bis and A4bis means that unbiasedness remains valid, but normality happens only asymptotically.

- A1 - X is full rank
- A2ter - X is iid + not contemporaneously correlated with ε_t + regularity conditions (which ensure Laws of Large Numbers [LLN], Central Limit theorems [CLT] apply).
- A3 - E[ε]=0
- A4 - E[$\varepsilon\varepsilon'$]=$\sigma^2 I_T$.

  Note that A3 and A4 imply ε_t is iid.
- A5 - is no longer necessary

9.1 Asymptotic theory

What happens to random variables (estimators) as the sample size increases. $T \longrightarrow \infty,\ N \longrightarrow \infty$

$$\hat{\beta} \longrightarrow \beta$$

Definition 5 ***Convergence in probability.*** *The random variable x_n converges in probability to a constant c if $\lim_{n\to\infty} \Pr(|x_n - c| > \varepsilon) = 0$, for any positive ε. We write $x_n \xrightarrow{p} x$ or $p\lim x_n = c$. We say that c is probability limit (plim) of x_n.*

Theorem 6 ***Law of Large Numbers.*** *If $x_i,\ i = 1, ..., n$ is a random **i.i.d.** sample from a distribution with finite mean $E[x_i] = \mu$, then $p\lim(\bar{x}_n) = \mu$, where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean).*

Definition 7 ***Consistency.*** *If an estimator $\hat{\theta}$ is such that plim $\hat{\theta} = \theta$, then we call that estimator "consistent".*

An implication of the LLN and the definition of consistency is that the sample mean is a consistent estimator of μ. (It is an estimator, which converges in probability to μ).

Moving the interest to distributions:

Definition 8 ***Convergence in distribution.*** *The random variable x_n converges in distribution to a random variable x with cdf $F(x)$ if $\lim_{n\to\infty} |F(x_n) - F(x)| = 0$, for all continuity points in $F(x)$. This is written $x_n \xrightarrow{d} x$ and x is called the limiting distribution.*

Theorem 9 ***Central Limit Theorem.*** *If $x_i,\ i = 1, ..., n$ is a random **i.i.d.** sample from a distribution with finite mean $E[x_i] = \mu$ and finite variance $Var(x_i) = \sigma^2$ then $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2)$.*

The LLN will be used to establish consistency of the OLS estimator

The CLT will be used to derive the asymptotic distribution of the OLS estimator.

Continuous functions are limit-preserving:

Theorem 10 ***Continuous mapping theorem.*** *If $x_n \xrightarrow{p} c$ and $g(\cdot)$ is a continuous function, then $g(x_n) \xrightarrow{p} g(c)$. If $x_n \xrightarrow{d} x$ and $g(\cdot)$ is a continuous function, then $g(x_n) \xrightarrow{d} g(x)$.*

Theorem 11 ***The delta method.*** *Let a random variable be such that*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \xi$$

and $g(\cdot)$ *continuously differentiable in a neighborood of* μ*. Then*

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} G\xi$$

with $G = \frac{\partial g(u)}{\partial u}$*. For example if* $\xi \sim N(0, V)$ *then*

$$g(\hat{\mu}) \xrightarrow{d} N(g(\mu), GVG').$$

9.2 Large sample properties of the OLS estimator

- A1 - X is full rank
- A2ter - X is i.i.d. with $Cov(x_t, \varepsilon_t) = 0$, that is no *contemporaneous* correlation between regressors and error term. We also require finite second moment $E[x_t x_t'] = Q_{XX} < \infty$ and positive definite. Furthermore, for asymptotic normality we will require finite fourth moments: $Q_{x\varepsilon} = Var[x_t \varepsilon_t] < \infty$.
- A3+A4 - ε_t is i.i.d. with $E[\varepsilon_t] = 0, Var(\varepsilon_t) = \sigma^2$.

  Under this set of assumptions, we can show consistency and asymptotic normality.

9.2.1 Consistency

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= \beta + (X'X)^{-1}X'\varepsilon \\
&= \beta + \left(\frac{X'X}{T}\right)^{-1}\frac{X'\varepsilon}{T}
\end{aligned}$$

What are these quantities?

$$\frac{X'X}{T} = \frac{1}{T}\sum_{t=1}^{T} x_t x_t' \xrightarrow{p} E[x_t x_t'] = Q_{xx};$$

because $x_t x_t'$ is iid and therefore a LLN applies. Compactly, $plim\frac{X'X}{T} = Q_{xx}$. Also

$$\frac{X'\varepsilon}{T} = \frac{1}{T}\sum_{t=1}^{T} x_t \varepsilon_t \xrightarrow{p} E[x_t \varepsilon_t] = 0$$

because $x_t\varepsilon_t$ is iid and thereforea LLN applies, and because $E[x_t\varepsilon_t] = 0$ due to the fact that $E[\varepsilon_t] = 0$ and $Cov(x_t\varepsilon_t) = 0$. Compactly, $plim\frac{X'\varepsilon}{T} = 0$. Now using the continuous mapping theorem:

$$\begin{aligned} p\lim\hat{\beta} &= \beta + plim\left(\frac{X'X}{T}\right)^{-1} plim\frac{X'\varepsilon}{T} \\ &= \beta \end{aligned}$$

Hence the OLS is consistent.

9.2.2 Asymptotic normality

For the asymptotic distribution, we will need a CLT. Let us start focusing on the distribution of $\sqrt{T}(\hat{\beta} - \beta)$:

$$\begin{aligned} \hat{\beta} &= \beta + \left(\frac{X'X}{T}\right)^{-1}\frac{X'\varepsilon}{T} \\ \sqrt{T}(\hat{\beta} - \beta) &= \left(\frac{X'X}{T}\right)^{-1}\sqrt{T}\frac{X'\varepsilon}{T} \end{aligned}$$

Let us focus on the term

$$\sqrt{T}\frac{X'\varepsilon}{T} = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t,$$

which is $\sqrt{T}$ times the sample mean of $x_t\varepsilon_t$. Since $x_t\varepsilon_t$ is i.i.d. with mean $E[x_t\varepsilon_t] = 0$ and variance

$$\begin{aligned} Q_{x\varepsilon} &= Var[x_t\varepsilon_t] \\ &= E[x_t\varepsilon_t(x_t\varepsilon_t)'] - E[x_t\varepsilon_t]E[(x_t\varepsilon_t)'] \\ &= E[x_t\varepsilon_t\varepsilon_t'x_t'] \\ &= E[x_tx_t'\varepsilon_t^2], \end{aligned}$$

the CLT implies

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t - \underset{E[x_t\varepsilon_t]}{0}\right) \xrightarrow{d} N(0, \underset{Var[x_t\varepsilon_t]}{Q_{x\varepsilon}}).$$

Then we have, by applying the continuous mapping theorem and the delta method:

$$\begin{aligned}\sqrt{T}(\hat{\beta}-\beta) &= \underbrace{\left(\frac{X'X}{T}\right)^{-1}}_{\xrightarrow{p} Q_{xx}} \underbrace{\frac{X'\varepsilon}{\sqrt{T}}}_{\xrightarrow{d} N(0,Q_{x\varepsilon})} \\ &\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} Q_{xx}^{-1} N(0, Q_{x\varepsilon}) \\ &\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N(0, Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}), \qquad (6)\end{aligned}$$

which is the asymptotic distribution of $\sqrt{T}(\hat{\beta}-\beta)$. Rearranging gives the asymptotic distribution of $\hat{\beta}$:

$$\hat{\beta} \xrightarrow{d} N(\beta, \frac{1}{T} Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}) \qquad (7)$$

The variance $\frac{1}{T} Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}$ goes to 0 as T goes to infinity, which is not surprising since we know that $\hat{\beta} \xrightarrow{p} \beta$. The representation (7) is useful when constructing test statistics and standard errors. However, for theoretical purposes the representation (6) is more useful as it retains non-degenerate limits as the sample sizes diverge.

9.2.3 Summary of asymptotic results

So far we have seen that under these conditions:

- $\text{plim} \frac{X'X}{T} = plim \frac{1}{T} \sum x_t x_t' = Q_{xx} = E[x_t x_t']$ (LLN holds for $x_t x_t'$)
- $\text{plim } \frac{X'\varepsilon}{T} = plim \frac{1}{T} \sum x_t \varepsilon_t = E[x_t \varepsilon_t]$ (LLN holds for $x_t \varepsilon_t$)
- $E[x_t \varepsilon_t] = 0$ (no contemporaneous correlation)

We can show that OLS is consistent: $\hat{\beta} \xrightarrow{p} \beta$
Also assuming:

- $\sqrt{T}\left(\frac{X'\varepsilon}{T} - E[x_t \varepsilon_t]\right) \xrightarrow{d} N(0, Var[x_t \varepsilon_t])$ (CLT holds for $x_t \varepsilon_t$)
- $Var[x_t \varepsilon_t] = Q_{x\varepsilon}$ p.d.

We can show that OLS is asymptotically Gaussian: $\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N(0, Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1})$ and $\hat{\beta} \xrightarrow{d} N(\beta, \frac{1}{T} Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1})$.

9.3 Asymptotic variance estimation

Can we estimate the asymptotic variance?

$$AsyVar(\hat{\beta}) = Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}$$

We are going to consider two cases, which depend on the assumption about the conditional variance of the errors. Consistent estimator of Q_{xx}^{-1} is $\left(\frac{X'X}{T}\right)^{-1}$. We need to consistently estimate the matrix $Q_{x\varepsilon}$. In general:

$$Q_{x\varepsilon} = Var[x_t \varepsilon_t] = E[x_t x_t' \varepsilon_t^2] - E[x_t \varepsilon_t] E[x_t \varepsilon_t]' = E[x_t x_t' \varepsilon_t^2].$$

Under the assumption that

$$E[\varepsilon_t^2 | x_t] = \sigma^2,$$

that is, conditional homoskedasticity.[2]

$$Q_{x\varepsilon} = E[x_t x_t' \varepsilon_t^2] = E[E[x_t x_t' \varepsilon_t^2 | x_t]] = E[x_t x_t' E[\varepsilon_t^2 | x_t]] = E[x_t x_t' \sigma^2] = \sigma^2 E[x_t x_t'] = \sigma^2 Q_{xx}.$$

So we have that

$$AsyVar(\hat{\beta}) = Q_{xx}^{-1} \sigma^2 Q_{xx} Q_{xx}^{-1} = \sigma^2 Q_{xx}^{-1}$$

and

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q_{xx}^{-1}) \tag{8}$$

and Q_{xx}^{-1} can be estimated consistently via:

$$\hat{\sigma}^2 \left(\frac{X'X}{T}\right)^{-1} \xrightarrow{p} \sigma^2 Q_{xx}^{-1}$$

That is

$$\widehat{AsyVar}(\hat{\beta}) = \hat{\sigma}^2 \left(\frac{X'X}{T}\right)^{-1}$$

Going back to (8) and plugging in this consistent estimator of the asymptotic variance we have:

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \hat{\sigma}^2 \left(\frac{X'X}{T}\right)^{-1}\right),$$

rearranging:

$$\hat{\beta} \xrightarrow{d} N\left(\beta, \hat{\sigma}^2 (X'X)^{-1}\right),$$

[2] Actually, the slighlty weaker assumption $Cov(x_t x_t', \varepsilon_t^2) = 0$ would suffice: $E[x_t x_t' \varepsilon_t^2] = E[x_t x_t'] E[\varepsilon_t^2] = \sigma^2 Q_{xx}$.

which is the usual! Note that

$$Var(\hat{\beta}|X) = \hat{\sigma}^2(X'X)^{-1} = \frac{1}{T}\widehat{AsyVar}(\hat{\beta})$$

and goes to 0 as $T \longrightarrow \infty$:

$$\underset{\longrightarrow 0}{\frac{1}{T}} \underset{\rightarrow \sigma^2 Q_{xx}^{-1}}{\widehat{AsyVar}}(\hat{\beta}).$$

Let us now consider the more general case with conditional heteroskedasticity. In this case ε_t is still independent across time, but is not identically distributed, so $E[\varepsilon_t^2|x_t] = \sigma^2$ does not apply. We have

$$Q_{x\varepsilon} = E[x_t x_t' \varepsilon_t^2]$$

a consistent estimator is

$$\hat{Q}_{x\varepsilon} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \hat{\varepsilon}_t^2 = \frac{1}{T} X' \widehat{\sigma^2 \Omega} X,$$

that is, the White (1980) estimator. Indeed note that:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \hat{\varepsilon}_t^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \varepsilon_t^2 - \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' (\hat{\varepsilon}_t^2 - \varepsilon_t^2)$$

where $\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \varepsilon_t^2 \xrightarrow{p} Q_{x\varepsilon}$ and $\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' (\hat{\varepsilon}_t^2 - \varepsilon_t^2) \xrightarrow{p} 0$. Therefore we have:

$$\widehat{AsyVar}(\hat{\beta}) = \left(\frac{X'X}{T}\right)^{-1} \frac{1}{T} X' \widehat{\sigma^2 \Omega} X, \left(\frac{X'X}{T}\right)^{-1} \longrightarrow AsyVar(\hat{\beta}) = Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}.$$

Plugging this estimator into (6) gives:

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \left(\frac{X'X}{T}\right)^{-1} \frac{1}{T} X' \widehat{\sigma^2 \Omega} X \left(\frac{X'X}{T}\right)^{-1}\right)$$

and rearranging gives:

$$\hat{\beta} \xrightarrow{d} N\left(\beta, (X'X)^{-1} X' \widehat{\sigma^2 \Omega} X (X'X)^{-1}\right), \tag{9}$$

which is the usual OLS expression with White heteroskedasticity consistent estimator. Note that again.

$$Var(\hat{\beta}|X) = (X'X)^{-1} X' \widehat{\sigma^2 \Omega} X (X'X)^{-1} = \frac{1}{T}\widehat{AsyVar}(\hat{\beta})$$

which of course goes to 0 as $T \longrightarrow \infty$ since $\widehat{AsyVar}(\hat{\beta}) \rightarrow \sigma^2 Q_{xx}^{-1} Q_{x\varepsilon} Q_{xx}^{-1}$

A similar reasoning could be applied in a case with ε_t autocorrelated. However, this would require invoking different CLT and LLN to account for dependence. The resulting asymptotic distribution would be as in (9) with $X'\widehat{\sigma^2\Omega}X$ based on a HAC estimator, e.g. the Newey-West estimator.

Remark 12 *The asymptotic distributions derived in this Section are the same we derived in the small sample using normality of the errors and fixed regressors, and this also in the case of autocorrelated and heteroschedastic disturbances. In other words, if we have a large sample we can resort to these asymptotic arguments to avoid assuming that the errors are Gaussian, and also to allow for random regressors. The assumptions we need to be able to resort to the asyptotic argument is that the random regressors are iid and not contemporaneously correlated with the errors of the model, plus some regularity conditions.*

9.4 A digression on time series models

What we discussed used CLT and LLN to derive consistency and asymptotic distributions. In order for the LLN and CLT to hold we have made requirements on the heterogeneity and dependence of the regressors X, in particular assuming they were independent (so no dependence) and identically distributed (so no heterogeneity). In a time series model, such assumptions are not tenable, however they can be "substituted" by stationarity (which limits the heterogeneity) and ergodicity (which limits the dependence). Indeed it is possible to derive LLN and CLT that apply to dependent and/or heterogeneous data, as long as the regressors are stationary and ergodic. If that is the case, then the same steps can be used to derive the consistency and asymptotic distrbution of a time series model.

This also means that in a dynamic model (a time series model in which the regressor is the lag of the dependent variable) such as:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

the coefficient ϕ can be consistently estimated with OLS, and the OLS estimator has an asymptotically Gaussian distribution. However, it is key that the errors are white noise for this result to hold. This can be achieved by specifying a rich enough lag specification.

Finally, as we shall see later on, even in presence of nonstationarity the OLS estimator is still consistent (actually, it is super-consistent) even though it converges to a non-standard distribution.

10 Endogeneity and instrumental variables

We have seen that under the assumption $Cov(x_t, \varepsilon_t) = 0$ "standard" inference is still valid. There are many applications in which such assumption is untenable. Sometimes it happens that:

$$Cov(x_t, \varepsilon_t) \neq 0 \rightarrow Endogeneity,$$

and x_t is called an endogenous variable.

10.1 Examples of endogeneity

10.1.1 Omitted variables

We have seen that, if we omit a variable we have a bias (omitted variable bias). We also have inconsistency:

$$DGP : y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{10}$$

with X_1 and X_2 mutually correlated and uncorrelated with ϵ.

$$MODEL : y = X_1\delta_1 + v$$

where we have

$$v = X_2\beta_2 + \epsilon,$$

which gives:

$$\begin{aligned} \hat{\delta}_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'\{X_1\delta_1 + v\} \\ &= \beta_1 + \left(\frac{1}{T}X_1'X_1\right)^{-1} \underbrace{\frac{1}{T}X_1'v}_{\longrightarrow E[x_{1t}v_t]} \end{aligned}$$

and

$$E[x_{1t}v_t] = E[x_{1t}(x_{2t}\beta_2 + \varepsilon_t)] = E[x_{1t}x_{2t}]\beta_2 \neq 0$$

10.1.2 Measurement error

Similar to omitted variable. Imagine you want to estimated earnings as a funcion of education:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

where $x_i = Education$, but we only observe $\tilde{x}_i = Schooling$, which is an imperfect proxy for education:

$$\tilde{x}_i = x_i + u_i.$$

So the model we can estimate is:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + \varepsilon_i \\ &= \beta_1 + \beta_2(\tilde{x}_i - u_i) + \varepsilon_i \\ &= \beta_1 + \beta_2 \tilde{x}_i + w_i \end{aligned}$$

with $w_i = \varepsilon_i - \beta_2 u_i$. In this model $E[\tilde{x}_i w_i] = E[(x_i + u_i)(\varepsilon_i - \beta_2 u_i)] = E[x_i(\varepsilon_i - \beta_2 u_i)] + E[u_i(\varepsilon_i - \beta_2 u_i)] = -\beta_2 E[u_i^2]$. It follows that

$$\frac{1}{T}\sum \tilde{x}_i w_i \to -\beta_2 E[u_i^2] \neq 0.$$

Note the negative value: attenuation bias (we are assuming β_2 positive as it should be expected).

10.1.3 Selection bias

This happens when we are trying to estimate a causal relationship and the regressors X are not selected in a randomized way. Here the problem of endogeneity is a feature of the model/experimental design, which does not match appropriately the causal relation we want to represent. The textbook example is the study of the effects of a drug on some patients. Let y_i denote a patient health status and let x_i denote wether he took the drug ($x_i = 1$) or not ($x_i = 0$). The linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

which implies

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

And allows to answer questions such as: what is the expected effect of this drug? One answer is:

$$E[y_i|x_i = 1] - E[y_i|x_i = 0] = \beta_0 + \beta_1 \cdot 1 - (\beta_0 + \beta_1 \cdot 0) = \beta_1. \tag{11}$$

Note however that this answer is not necessarily satisfactory. We cannot limit ourselves to computing the difference between the expected health of those who took the drug $E[y_i|x_i = 1]$ and those who did not $E[y_i|x_i = 0]$, because there might be

other carachteristics of the individuals playing some other explanatory role. What we really are after is

$$E[y_i^1|x_i = 1] - E[y_i^0|x_i = 1] \tag{12}$$

where

$$\begin{cases} y_i^0 = \text{health status of individual } i \text{ when he does not take the drug} \\ y_i^1 = \text{health of individual } i \text{ when he takes the drug} \end{cases},$$

irrespectively of wether they actually took the drug or not. This is Rubin's concept of potential output. Expression (12) then denotes the difference between the expected health of those who took the drug $E[y_i|x_i = 1] = E[y_i^1|x_i = 1]$ and the *same* people (therefore people for whom $x_i = 1$), under the hypothetical scenario that they did not take it, $E[y_i^0|x_i = 1]$. Since we do not observe $y_i^0|x_i = 1$ (this is called a conterfactual) this creates a complication. We have:

$$\begin{aligned} E[y_i|x_i &= 1] - E[y_i|x_i = 0] \\ &= E[y_i^1|x_i = 1] - E[y_i^0|x_i = 0] \ (observables) \\ &= E[y_i^1|x_i = 1] + (\underbrace{-E[y_i^0|x_i = 1]}_{count \text{ erf } actual} + E[y_i^0|x_i = 1]) - E[y_i^0|x_i = 0] \\ &= E[y_i^1|x_i = 1] - E[y_i^0|x_i = 1] + \underbrace{E[y_i^0|x_i = 1] - E[y_i^0|x_i = 0]}_{Selection \ Bias}. \end{aligned}$$

The steps above show that the desired quantity $E[y_i^1|x_i = 1] - E[y_i^0|x_i = 1]$ coincides to the observable quantity $E[y_i|x_i = 1] - E[y_i|x_i = 0]$ ony if the selection bias term is 0. This term represents the effect of how the patients taking the drug are selected. Note that, if they are selected randomly, that is, independently from the potential outcome, we have that $E[y_i^0|x_i = 1] = E[y_i^0|x_i = 0] = E[y_i^0]$ and the selection bias term is 0. But if they are not selected randomly, for example the choice to take the drug is on a voluntary basis, or is given only among people that are regularly visiting their doctor, those patients with better health and better possible outcomes are less likely to be taking the drug and $E[y_i^0|x_i = 1] < E[y_i^0|x_i = 0]$.

Moving on to the regression, the observed outcome can be written as a function of the potential outcomes:

$$\begin{aligned} y_i &= y_i^0 + (y_i^1 - y_i^0)x_i \\ &= E[y_i^0] + (y_i^1 - y_i^0)x_i + (y_i^0 - E[y_i^0]) \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i. \end{aligned}$$

Evaluating the expectations under treatment and no treatment gives:

$$\begin{aligned} E[y_i|x_i &= 1] = \beta_0 + \beta_1 + E(\varepsilon_i|x_i = 1) \\ E[y_i|x_i &= 0] = \beta_0 + E(\varepsilon_i|x_i = 0) \\ E[y_i|x_i &= 1] - E[y_i|x_i = 0] = \beta_1 + E(\varepsilon_i|x_i = 1) - E(\varepsilon_i|x_i = 0) \end{aligned}$$

which shows that if there is any kind of dependence between the regressor x_i and the error term ε_i, then β_1 is not identified from the data. If randomization is not possible, then there are two solutions. Note that we can interpret this also this example as an instance of omitted variable problem/measument error, and adding a control variable to the conditional mean of the model, for example age or overall health can remove the selection bias. The other solution is to use a z_i instrument correlated with x_i but uncorrelated with the potential outcomes (i.e. uncorrelated with the errors). Why? because then we can project x_i on the space of z_i and obtain a regression

$$\begin{aligned} y_i &= \beta_0 + \beta_1(P_z x_i + M_z x_i) + \varepsilon_i. \\ &\quad \beta_0 + \beta_1 \hat{x}_i \ + \tilde{\varepsilon}_i \end{aligned}$$

where $\hat{x}_i \ = P_z x_i$ and $\tilde{\varepsilon}_i = \beta_1 M_z x_i + \varepsilon_i$ are by construction orthogonal, hence the selection bias of this regression is 0.

10.1.4 Simultaneous equation models

A basic demand-supply model Simultaneous equation models represent a set of causal relationships. A basic demand - supply model is:

$$\begin{cases} q_t^d = \beta p_t + \varepsilon_t^d \\ q_t^s = \gamma p_t + \varepsilon_t^s \end{cases}$$

In equilibrium $q_t^d = q_t^s = q_t$ which implies the model can be written as:

$$\begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t^d \\ \varepsilon_t^s \end{bmatrix}.$$

Solving the model gives:

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \frac{\gamma}{\gamma-\beta} & -\frac{\beta}{\gamma-\beta} \\ \frac{1}{\gamma-\beta} & -\frac{1}{\gamma-\beta} \end{bmatrix} \begin{bmatrix} \varepsilon_t^d \\ \varepsilon_t^s \end{bmatrix} = \begin{bmatrix} \frac{\gamma\varepsilon_t^d - \beta\varepsilon_t^s}{\gamma-\beta} \\ \frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma-\beta} \end{bmatrix}. \tag{13}$$

Clearly both q_t and p_t depend on both shocks. What happens if we try to estimate the demand equation via OLS?[3] We would have:

$$\begin{aligned}
\hat{\beta} &= \frac{\frac{1}{T}\Sigma p_t q_t}{\frac{1}{T}\Sigma p_t^2} \\
&= \frac{\frac{1}{T}\Sigma \frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma-\beta}\frac{\gamma\varepsilon_t^d - \beta\varepsilon_t^s}{\gamma-\beta}}{\frac{1}{T}\Sigma\left(\frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma-\beta}\right)^2} \\
&= \frac{\frac{1}{T}\Sigma(\varepsilon_t^d - \varepsilon_t^s)(\gamma\varepsilon_t^d - \beta\varepsilon_t^s)}{\frac{1}{T}\Sigma\left(\varepsilon_t^d - \varepsilon_t^s\right)^2} \\
&= \frac{\frac{1}{T}\Sigma\{\gamma(\varepsilon_t^d)^2 - \beta\varepsilon_t^d\varepsilon_t^s - \gamma\varepsilon_t^d\varepsilon_t^s + \beta(\varepsilon_t^s)^2\}}{\frac{1}{T}\Sigma\{(\varepsilon_t^d)^2 + (\varepsilon_t^s)^2 - 2\varepsilon_t^d\varepsilon_t^s\}} \\
&= \frac{\longrightarrow \gamma E[(\varepsilon_t^d)^2] + \beta E[(\varepsilon_t^s)^2]}{\longrightarrow E[(\varepsilon_t^d)^2] + E[(\varepsilon_t^s)^2]} \\
\longrightarrow\ & \frac{\gamma\sigma_d^2 + \beta\sigma_s^2}{\sigma_d^2 + \sigma_s^2}.
\end{aligned} \tag{14}$$

This is not the elasticity we are looking for (the demand elasticity β) but rather a mix of demand and supply elasticities (the demand elasticity β and the supply elasticity γ).[4]

In this situation the coefficient $\frac{\gamma\sigma_d^2+\beta\sigma_s^2}{\sigma_d^2+\sigma_s^2}$ is identified but γ and β are not. This is not surprising, it is a direct consequence of the simultaneity of this model: it is

[3] The same would happen if we tried to estimate the supply equation.

[4] Only in the special case in which the shocks to one of the two curves are so small to be negligible we could consistently estimate β and γ. This is close to a situation in which only one of the two curves is being shocked, and its movements after the shocks allow us to trace out the other curve.

possible to show that in model (13) we have[5]

$$E[q_t|p_t] = \frac{\gamma\sigma_d^2 + \beta\sigma_s^2}{\sigma_d^2 + \sigma_s^2} p_t, \tag{15}$$

which is indeed the probability limit to which the OLS estimator is converging (see (14)). This is an observable property of the joint distribution, and indeed the linear regression of q_t on p_t yields a consistent estimator of this quantity.

Clearly in this (and any other) simultaneous equation model the object βp_t is **not** the conditional mean, but a conceptual causal object capturing what would happen to q_t if we moved p_t only via shocks ε_t^s while keeping fixed ε_t^d. This amounts to a situation in which only one of the two curves is being shocked, and its movements after the shocks allow us to trace out the other curve. A notation for this object was given by Pearl (2010): $\beta p_t = E[q_t|do(\varepsilon_t^d = \bar{\varepsilon}_t^d)]$. This latter expectation and the conditional expectation $E[q_t|p_t]$ in (15) will coincide only if $E[\varepsilon_t^d|p_t] = 0$, that is if there is independence between ε_t^d and p_t, a condition that by construction does not hold in a simultaneous model.

Problem: it is important to emphasize that the causal object βp_t is **not** a characteristic of the joint distribution and therefore **cannot** be inferred from the data without additional assumptions.

Solution: find a variable that shifts one curve but not the other. This would introduce in the system a channel trough which we can reproduce the ideal counterfactual situation of keeping one curve fixed while moving the other. If the instrumental variable only moves one of the two curves, then it is possible to trace the curve that does not respond to the instrument. For example, define w_t as the number of days with temperatures below 0 Celsius, which would cause a reduction in the yield of the agricultural sector. We can think of w_t as a component of the supply shocks:

$$\varepsilon_t^s = hw_t + u_t^s$$

[5]From (13) we have:

$$\begin{aligned}
E\begin{bmatrix} q_t \\ p_t \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
Var\begin{bmatrix} q_t \\ p_t \end{bmatrix} &= (\gamma-\beta)^{-2}\begin{bmatrix} \gamma & -\beta \\ 1 & -1 \end{bmatrix}\begin{bmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix}\begin{bmatrix} \gamma & 1 \\ -\beta & -1 \end{bmatrix} \\
&= (\gamma-\beta)^{-2}\begin{bmatrix} \sigma_s^2\beta^2 + \sigma_d^2\gamma^2 & \sigma_s^2\beta + \sigma_d^2\gamma \\ \sigma_s^2\beta + \sigma_d^2\gamma & \sigma_d^2 + \sigma_s^2 \end{bmatrix} \\
E[q_t|p_t] &= (\sigma_s^2\beta + \sigma_d^2\gamma)(\sigma_d^2 + \sigma_s^2)^{-1} p_t,
\end{aligned}$$

where the last line follows from the properties of the multivariate Gaussian.

and we can write the model as

$$\left\{ \begin{array}{c} q_t^d = \beta p_t + \varepsilon_t^d \\ q_t^s = \gamma p_t + h w_t + u_t^s \end{array} \right. .$$

Note that w_t only impacts the supply and is uncorrelated with both the other shocks. Putting in matrix notation and solving the model gives:

$$\begin{aligned} \left[\begin{array}{c} q_t \\ p_t \end{array} \right] &= (\gamma - \beta)^{-1} \left[\begin{array}{cc} \gamma & -\beta \\ 1 & -1 \end{array} \right] \left[\begin{array}{c} 0 \\ h w_t \end{array} \right] + (\gamma - \beta)^{-1} \left[\begin{array}{cc} \gamma & -\beta \\ 1 & -1 \end{array} \right] \left[\begin{array}{c} \varepsilon_t^d \\ u_t^s \end{array} \right] \\ &= \left[\begin{array}{c} -\frac{h}{\gamma-\beta}\beta \\ -\frac{h}{\gamma-\beta} \end{array} \right] w_t + \left[\begin{array}{c} \frac{\gamma\varepsilon_t^d - \beta u_t^s}{\gamma-\beta} \\ \frac{\varepsilon_t^d - u_t^s}{\gamma-\beta} \end{array} \right] \end{aligned} \tag{16}$$

Use the second equation to find the conditional expectation $E[p_t|w_t]$:

$$p_t = \underbrace{-\frac{h}{\gamma-\beta} w_t}_{p_t^* = E[p_t|w_t]} + \frac{\varepsilon_t^d - u_t^s}{\gamma - \beta}. \tag{17}$$

Key step: using the conditional expectation p_t^* we can re-write the model (16) as:

$$\left[\begin{array}{c} q_t \\ p_t \end{array} \right] = \left[\begin{array}{c} \beta \\ 1 \end{array} \right] p_t^* + \left[\begin{array}{c} \frac{\gamma\varepsilon_t^d - \beta u_t^s}{\gamma-\beta} \\ \frac{\varepsilon_t^d - u_t^s}{\gamma-\beta} \end{array} \right]. \tag{18}$$

It is clear that in this model $E[q_t|p_t^*] = \beta p_t^*$, i.e. the conditional expectation $E[q_t|p_t^*]$ coincides with the causal object of interest $\beta p_t = E[q_t|do(\varepsilon_t^d = \bar{\varepsilon}_t^d)]$. The OLS estimator in the regression of q_t on p_t^* would converge to this value:

$$\begin{aligned} \hat{\beta} &= \frac{\frac{1}{T}\Sigma p_t^* q_t}{\frac{1}{T}\Sigma (p_t^*)^2} \\ &= \frac{\frac{1}{T}\Sigma p_t^* (\beta p_t^* + v_t)}{\frac{1}{T}\Sigma (p_t^*)^2} \\ &= \frac{\frac{1}{T}\Sigma (\beta (p_t^*)^2 + p_t^* v_t)}{\frac{1}{T}\Sigma (p_t^*)^2} \\ &= \beta + \frac{\frac{1}{T}\Sigma p_t^* v_t}{\frac{1}{T}\Sigma (p_t^*)^2} \longrightarrow \beta, \end{aligned} \tag{19}$$

where the last step follows from $E[p_t^* v_t] = 0$.

The regression of q_t on p_t^* appearing in the first line of (18) can be interpreted as a version of the demand equation in which p_t has been swapped with p_t^* (with a remainder term ending up in the disturbances). The second equation of (18) instead is the "first stage" regression, projecting p_t on a space orthogonal to the demand shocks, to obtain a variable p_t^* that is uncorrelated to the demand shock. In practice p_t^* needs to be estimated, which can be done using the linear regression:

$$p_t = \delta w_t + \eta_t$$

with $\hat{p}_t = \hat{\delta} w_t$ a consistent estimate of $p_t^* = E[p_t|w_t] = -\frac{h}{\gamma-\beta} w_t$. The operational version of (19) is therefore:

$$\begin{aligned} \hat{\beta} &= \frac{\frac{1}{T}\Sigma \hat{p}_t(\beta \hat{p}_t + v_t)}{\frac{1}{T}\Sigma(\hat{p}_t)^2} \\ &= \beta + \frac{\frac{1}{T}\Sigma \hat{p}_t v_t}{\frac{1}{T}\Sigma(\hat{p}_t)^2} \longrightarrow \beta, \end{aligned}$$

since $\hat{p}_t \longrightarrow p_t^*$ and $E[p_t^* v_t] = 0$.[6]

Of course, it is key that only one of the two curves is moved by the instrument. Imagine this is not the case the model becomes:

$$\begin{aligned} \begin{bmatrix} q_t \\ p_t \end{bmatrix} &= (\gamma-\beta)^{-1} \begin{bmatrix} \gamma & -\beta \\ 1 & -1 \end{bmatrix} \begin{bmatrix} kw_t \\ hw_t \end{bmatrix} + (\gamma-\beta)^{-1} \begin{bmatrix} \gamma & -\beta \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \varepsilon_t^d \\ u_t^s \end{bmatrix} \\ &= (\gamma-\beta)^{-1} \begin{bmatrix} k\gamma - h\beta \\ k-h \end{bmatrix} w_t + \begin{bmatrix} \frac{\gamma\varepsilon_t^d - \beta u_t^s}{\gamma-\beta} \\ \frac{\varepsilon_t^d - u_t^s}{\gamma-\beta} \end{bmatrix}. \end{aligned}$$

Using the second equation we can obtain the conditional expectation

$$p_t^* = E[p_t|w_t] = -\frac{k-h}{\gamma-\beta} w_t,$$

but this is no longer proportional to the expectation $E[q_t|p_t^*]$ appearing in the first equation and therefore cannot be used to pin-down β by re-writing the system in the form (18). The form (18) would ensure that $E[q_t|p_t^*] = \beta p_t^*$ but it requires the exclusion restriction $k = 0$.

[6] Another way to see how the introduction of w_t identifies β is to note that using representation (16) one can compute $E[q_t|w_t] = -\frac{h}{\gamma-\beta}\beta w_t$ and $E[p_t|w_t] = -\frac{h}{\gamma-\beta} w_t$, and then note that $\frac{E[q_t|w_t]}{E[p_t|w_t]} = \beta$. In the sample, one could regress q_t on w_t to obtain $(\Sigma w_t^2)^{-1}\Sigma w_t q_t$ and p_t on w_t to obtain $(\Sigma w_t^2)^{-1}\Sigma w_t p_t$, and then take the ratio $\frac{(\Sigma w_t^2)^{-1}\Sigma w_t q_t}{(\Sigma w_t^2)^{-1}\Sigma w_t p_t} = (\Sigma w_t p_t)^{-1}\Sigma w_t q_t$ which is the IV estimator. This approach is less efficient that the TSLS approach, because it uses w_t instead of the projection of p_t on w_t.

Full information system estimation approach While the system written as in (18) obviously lends itself to a TSLS estimation procedure, an alternative procedure involves estimating the system in (16) and then solve for the parameters of interest, that is estimate the form:

$$\left[\begin{array}{c} q_t \\ p_t \end{array} \right] = \left[\begin{array}{c} a_1 \\ a_2 \end{array} \right] w_t + \left[\begin{array}{c} v_t^d \\ v_t^s \end{array} \right]$$

with

$$\left[\begin{array}{c} a_1 \\ a_2 \end{array} \right] = \left[\begin{array}{c} -\frac{h}{\gamma-\beta}\beta \\ -\frac{h}{\gamma-\beta} \end{array} \right]; \tag{20}$$

$$Var\left(\left[\begin{array}{c} v_t^d \\ v_t^s \end{array} \right] \right) = \left[\begin{array}{cc} \sigma_{v^d}^2 & \sigma_{v^d v^s} \\ \sigma_{v^d v^s} & \sigma_{v^s}^2 \end{array} \right] = \left[\begin{array}{cc} \frac{\sigma_{\varepsilon^d}^2\gamma^2+\sigma_{u^s}^2\beta^2}{(\gamma-\beta)^2} & \frac{\sigma_{\varepsilon^d}^2\gamma+\sigma_{u^s}^2\beta}{(\gamma-\beta)^2} \\ \frac{\sigma_{\varepsilon^d}^2\gamma+\sigma_{u^s}^2\beta}{(\gamma-\beta)^2} & \frac{\sigma_{\varepsilon^d}^2+\sigma_{u^s}^2}{(\gamma-\beta)^2} \end{array} \right]. \tag{21}$$

There are five structural coefficients $h, \gamma, \beta, \sigma_{\varepsilon^d}^2, \sigma_{u^s}^2$ and five reduced form coefficients $a_1, a_2, \sigma_{v^d}^2, \sigma_{v^d v^s}, \sigma_{v^s}^2$. The latter can be consistently estimated with OLS and then one can solve for the structural parameters. In this example (20) identifies β trough the ratio a_1/a_2, and then assuming knowkledge of γ one can find $h = -a_2(\gamma - \beta)$. To find γ one can use the ratios of the variance conditions in (21)

$$\frac{\sigma_{v^d v^s}}{\sigma_{v^s}^2} = \frac{\sigma_{\varepsilon^d}^2\gamma + \sigma_{u^s}^2\beta}{\sigma_{\varepsilon^d}^2 + \sigma_{u^s}^2}; \frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} = \frac{\sigma_{\varepsilon^d}^2\gamma^2 + \sigma_{u^s}^2\beta^2}{\sigma_{\varepsilon^d}^2 + \sigma_{u^s}^2},$$

writing them as:

$$\gamma = \frac{\sigma_{v^d v^s}}{\sigma_{v^s}^2} + \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\frac{\sigma_{v^d v^s}}{\sigma_{v^s}^2} - \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\beta, \tag{22}$$

$$\gamma^2 = \frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} + \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} - \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\beta^2, \tag{23}$$

and solving for $\frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}$ and γ. To do so, first equalize the square of (22) with (23) to obtain $\frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}$:

$$\begin{aligned} \left(\frac{\sigma_{v^d v^s}}{\sigma_{v^s}^2} + \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\left(\frac{\sigma_{v^d v^s}}{\sigma_{v^s}^2} - \beta \right) \right)^2 &= \frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} + \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2}\left(\frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} - \beta^2 \right) \\ \rightarrow \quad \frac{\sigma_{u^s}^2}{\sigma_{\varepsilon^d}^2} &= \frac{\left(\frac{\sigma_{v^s}^2}{\sigma_{v^d v^s}} \right)^2 \frac{\sigma_{v^d}^2}{\sigma_{v^s}^2} - 1}{\left(\frac{\sigma_{v^s}^2}{\sigma_{v^d v^s}}\beta - 1 \right)^2}. \end{aligned}$$

Now use this ratio in (23) to solve for γ^2:

$$\gamma^2 = \frac{\sigma^2_{v^d}}{\sigma^2_{v^s}} + \frac{\left(\frac{\sigma^2_{v^s}}{\sigma_{v^d v^s}}\right)^2 \frac{\sigma^2_{v^d}}{\sigma^2_{v^s}} - 1}{\left(\frac{\sigma^2_{v^s}}{\sigma_{v^d v^s}}\beta - 1\right)^2}\left(\frac{\sigma^2_{v^d}}{\sigma^2_{v^s}} - \beta^2\right) = \frac{\left(\beta - \frac{\sigma^2_{v^d}}{\sigma_{v^d v^s}}\right)^2}{\left(\frac{\sigma^2_{v^s}}{\sigma_{v^d v^s}}\beta - 1\right)^2}$$

Taking the square root and rearranging yields $\gamma = \frac{\beta\sigma_{v^d v^s} - \sigma^2_{v^d}}{\beta\sigma^2_{v^s} - \sigma_{v^d v^s}}$, hence also γ is identified. Finally we can use $\frac{\sigma^2_{\varepsilon^d} + \sigma^2_{u^s}}{(\gamma-\beta)^2} = \sigma^2_{v^s}$ and $\frac{\sigma^2_{\varepsilon^d}\gamma + \sigma^2_{u^s}\beta}{(\gamma-\beta)^2} = \sigma_{v^d v^s}$ in (21) to straightforwardly solve for $\sigma^2_{\varepsilon^d}$ and $\sigma^2_{u^s}$.

A 3 variables macroeconomic model Here is a minimial example of a modern macro model:

$$\begin{aligned} IS &: \tilde{y}_t = a_1(i_t - \tilde{\pi}_t) + a_2\tilde{y}_{t-1} + \varepsilon^y_t \\ PC &: \tilde{\pi}_t = b_1 E_t[\tilde{\pi}_{t+1}] + b_2(\tilde{y}_t - y^*_t) + \varepsilon^\pi_t \\ TR &: i_t = c_1(\tilde{y}_t - y^*_t) + c_2(\tilde{\pi}_t - \tilde{\pi}^*_t) + c_3 i_{t-1} + \varepsilon^i_t \end{aligned}$$

where $\tilde{y}_t = \ln Y_t - \ln Y_{t-1}$ is the growth rate of output, $\tilde{\pi}_t = \ln P_t - \ln P_{t-1}$ the growth rate of prices, and i_t the (nominal) interest rate. The IS (Investment - Savings) curve is relating output to interest rates. The higher interest rates are, the lower output is. The Phillips Curve in which inflation depends on inflation expectations $E_t[\tilde{\pi}_{t+1}]$ and the output gap. Note that $\tilde{\pi}_t$ is related to $\tilde{y}_t$ trough the IS curve: higher inflation reduces the cost of borrowing, which increases investments and output. However higher output pushes inflation up trough the Phillips curve (more firms are looking for workers so salaries go up, costs go up, eventually prices go up). The PC effect counterbalances the IS effect until we reach a stable equilibrium. Finally, the Taylor Rule depicts the central bank actions in pursuing its goals of inflation and output deviations from $\tilde{\pi}^*_t$ and y^*_t are adjusted via increases/decreases in the interest rate.

We can simplify this model by setting $E_t[\tilde{\pi}_{t+1}] = \tilde{\pi}_{t-1}$ (adaptive expectations) and $\tilde{\pi}^*_t = y^*_t = 0$ (this is just to simplify notation).

$$\begin{aligned} IS &: \tilde{y}_t = a_1(i_t - \tilde{\pi}_t) + a_2\tilde{y}_{t-1} + \varepsilon^y_t \\ PC &: \tilde{\pi}_t = b_1\tilde{\pi}_{t-1} + b_2\tilde{y}_t + \varepsilon^\pi_t \\ TR &: i_t = c_1\tilde{y}_t + c_2\tilde{\pi}_t + c_3 i_{t-1} + \varepsilon^i_t \end{aligned}$$

Since every variable interacts contemporaneously with all the others, one cannot estimate any of these three equations via OLS. Note you cannot even perform a

forecast, as you would need the forecast from one equation to forecast another. The model above is a system of differential equations.

$$\begin{bmatrix} 1 & a_1 & -a_1 \\ -b_2 & 1 & 0 \\ -c_1 & -c_2 & 1 \end{bmatrix} \begin{bmatrix} \tilde{y}_t \\ \tilde{\pi}_t \\ i_t \end{bmatrix} = \begin{bmatrix} a_2 & 0 & 0 \\ 0 & b_1 & 0 \\ 0 & 0 & c_3 \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\pi}_{t-1} \\ i_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \\ \varepsilon_t^i \end{bmatrix}$$

Consider solving it:

$$\begin{aligned} \begin{bmatrix} \tilde{y}_t \\ \tilde{\pi}_t \\ i_t \end{bmatrix} &= \begin{bmatrix} 1 & a_1 & -a_1 \\ -b_2 & 1 & 0 \\ -c_1 & -c_2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} a_2 & 0 & 0 \\ 0 & b_1 & 0 \\ 0 & 0 & c_3 \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\pi}_{t-1} \\ i_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & a_1 & -a_1 \\ -b_2 & 1 & 0 \\ -c_1 & -c_2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \\ \varepsilon_t^i \end{bmatrix} \\ &= \frac{1}{D} \begin{bmatrix} a_2 & -b_1(a_1 - a_1c_2) & c_3a_1 \\ a_2b_2 & -b_1(a_1c_1 - 1) & c_3a_1b_2 \\ a_2(c_1 + b_2c_2) & b_1(c_2 - a_1c_1) & c_3(a_1b_2 + 1) \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\pi}_{t-1} \\ i_{t-1} \end{bmatrix} \\ &\quad + \frac{1}{D} \begin{bmatrix} 1 & -(a_1 - a_1c_2) & a_1 \\ b_2 & -(a_1c_1 - 1) & a_1b_2 \\ c_1 + b_2c_2 & c_2 - a_1c_1 & a_1b_2 + 1 \end{bmatrix} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \\ \varepsilon_t^i \end{bmatrix} \end{aligned}$$

with $D = a_1b_2 - a_1c_1 - a_1b_2c_2 + 1$. Note that the autoregressive matrix above can now be estimated with OLS! However, what we get is not the structural parameters we wanted! They are the parameters of the reduced form, the only form identified in the data. There is no guarantee that we can back the structural parameters from the reduces form parameters.

10.2 Instruments

The problem is solved trough the use of some alternative variables, called instruments. Z

- Z to be uncorrelated with the errors $Cov(z_t, \varepsilon_t) = 0 \Leftrightarrow E[z_t\varepsilon_t] = 0$. (Validity or exogeneity).
- Z to be correlated (as much as possible) with the X. (Relevance).
- $E[z_tz_t'] = Q_{zz} < \infty$, $E[z_tx_t'] = Q_{zx} < \infty$. (Regularity conditions).

Start with as many instruments as variables. The starting point is the moment condition:

$$E[Z'\varepsilon] = 0$$

suggests the moment estimator

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

that is:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$$

which of course turns out to be consistent (as we know that the moment condition holds).

$$\begin{aligned}
\hat{\beta}_{IV} &= (Z'X)^{-1}Z'Y \\
&= (Z'X)^{-1}Z'(X\beta + \varepsilon) \\
&= (Z'X)^{-1}Z'X\beta + (X'Z)^{-1}Z'\varepsilon \\
&= \beta + \underbrace{\left(\frac{Z'X}{T}\right)^{-1}}_{\to Q_{zx}} \underbrace{\frac{Z'\varepsilon}{T}}_{\to E[Z'\varepsilon]=0} \\
&\to \beta
\end{aligned}$$

Asymptotic distribution:

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) = \underbrace{\left(\frac{Z'X}{T}\right)^{-1}}_{\to Q_{zx}} \sqrt{T}\frac{Z'\varepsilon}{T}$$

and

$$\sqrt{T}\frac{Z'\varepsilon}{T} \to N(0, Var[z_t\varepsilon_t])$$

with $Var[z_t\varepsilon_t] = E[z_t z_t'\varepsilon_t^2] = E_Z[z_t z_t' E[\varepsilon_t^2|z_t]] = Q_{zz}\sigma^2$. Hence

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) \to N(0, \sigma^2 Q_{zx}^{-1} Q_{zz} Q_{zx}^{-1\prime})$$

a consistent estimator of the asymptotic variance is

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{T}\left(\frac{Z'X}{T}\right)^{-1}\left(\frac{Z'Z}{T}\right)\left(\frac{X'Z}{T}\right)^{-1} = \hat{\sigma}^2\left(\frac{Z'X}{T}\right)^{-1}\left(\frac{Z'Z}{T}\right)\left(\frac{X'Z}{T}\right)^{-1}$$

and we can write:

$$\hat{\beta}_{IV} \to N(\beta, \hat{\sigma}^2 (Z'X)^{-1} Z'Z(X'Z)^{-1}).$$

10.3 Two stage least squares

Suppose that you found a valid and relevant set of instruments Z. They can possibly be more than the Xs. Do the following.

$$X = Z\gamma + u$$

$$\begin{aligned} \hat{\gamma} &= (Z'Z)^{-1}Z'X \\ \hat{X} &= Z\hat{\gamma} \end{aligned}$$

these are the fitted values.

$$\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$$

It follows $\hat{X}$

$$\begin{aligned} \hat{\beta}_{2SLS} &= (\hat{X}'X)^{-1}\hat{X}'Y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (X'P_zX)^{-1}X'P_zY \\ &= (X'P_zP_zX)^{-1}X'P_zY \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \end{aligned}$$

which is the two stage least squares. The second stage is

$$Y = \hat{X}\beta + \varepsilon$$

The first stage was

$$X = Z\gamma + u$$

It is possible to show that TSLS is the most efficient choice among all possible transformations of Z. The asymptotic distribution is

$$\hat{\beta}_{2SLS} \to N(\beta, \hat{\sigma}^2(\hat{X}'X)^{-1}\hat{X}'\hat{X}(X'\hat{X})^{-1}).$$

where $(\hat{X}'X)^{-1}\hat{X}'\hat{X}(X'\hat{X})^{-1} = (\hat{X}'\hat{X})^{-1}$ and $\hat{\sigma}^2 = (Y - X\ \hat{\beta})'(Y - X\ \hat{\beta})/(T - k)$. Important: note $\hat{\sigma}^2$ uses the endogenous regressors and **not** the instrumented ones $(Y - \hat{X}\ \hat{\beta})'(Y - \hat{X}\ \hat{\beta})/(T - k)$.

10.4 Hausman test

Consider the null hypothesis of exogeneity. The OLS estimator $\hat{\beta}_{OLS}$ will be consistent. Then take a 2SLS estimator $\hat{\beta}_{2SLS}$ derived from Z instruments. Under the Null hypothesis of exogeneity, this estimator is also consistent.

$$p\lim(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) = 0. \tag{24}$$

Under the alternative only $\hat{\beta}_{2SLS}$ is consistent. Hence we can devise a test for the null of exogeneity using (24) as the null. In order to re-scale the discrepancy vector, we need to standardize it by its variance:

$$(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})' asyVar(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}).$$

where

$$\begin{aligned} asyVar((\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})) &= asyVar(\hat{\beta}_{2SLS}) + asyVar(\hat{\beta}_{OLS}) \\ &\quad -2asyCov(\hat{\beta}_{2SLS}, \hat{\beta}_{OLS}). \end{aligned}$$

Now note that under the null $\hat{\beta}_{OLS}$ will be BLUE and $\hat{\beta}_{2SLS}$ will be inefficient.[7] This implies[8] that $asyVar(\hat{\beta}_{2SLS}) = asyCov(\hat{\beta}_{2SLS}, \hat{\beta}_{OLS})$ which simplifies things as we do not need to estimate the covariance:

$$asyVar((\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})) = asyVar(\hat{\beta}_{OLS}) - asyVar(\hat{\beta}_{2SLS}).$$

[7] To see this more clearly, consider the difference in the estimated asymptotic variances of the two estimators under the null of exogeneity:

$$\sigma^2(\hat{X}'\hat{X})^{-1} - \sigma^2(X'X)^{-1}$$

consider the object:

$$\begin{aligned} & (X'X) - (X'Z)(Z'Z)^{-1}(Z'X) \\ = \; & X'X - XP_zX \\ = \; & X'M_zX \end{aligned}$$

positive definite. Since $A - B > 0 \Leftrightarrow A^{-1} < B^{-1}$ we have that the OLS has always smaller variance under the null of exogeneity.

[8] This follows from Hausmann's result:

$$\begin{aligned} & Cov(\hat{\beta}_{efficient}, \hat{\beta}_{efficient} - \hat{\beta}_{inefficient}) \\ = \; & Var(\hat{\beta}_{efficient}) - Cov(\hat{\beta}_{efficient}, \hat{\beta}_{inefficient}) = 0 \end{aligned}$$

which implies that $Var(\hat{\beta}_{efficient}) = Cov(\hat{\beta}_{efficient}, \hat{\beta}_{inefficient})$.

Therefore the Wald statistic can be written as

$$W = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'(asyVar(\hat{\beta}_{2SLS}) - asyVar(\hat{\beta}_{OLS}))^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$$

and estimated using

$$\widehat{W} = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'(\hat{\sigma}^2[(\hat{X}'\hat{X})^{-1} - (X'X)^{-1}])^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}).$$

Since there will be some Xs that are exogenous (e.g. the intercept) this quadratic form will generally be of reduced rank $K^* = K - K_0$ where K_0 is the number of exogenous Xs and K^* the number of endogenous Xs (call these X^*). Hence the inversion of the matrix in squared brakets needs to be performed using a generalized inverse procedure. In practice though, an equivalent variable addition test due to Wu (1973) can be performed easily by running the augmented regression:

$$y = X\beta + \hat{X}^*\gamma + \varepsilon^*$$

where X are all the variables (both endogenous and exogenous) and $\hat{X}^*$ are the projection of the endogenous variables X^* on the instruments. The null of exogeneity corresponds to $\gamma = 0$. If there is endogeneity, it will be picked up by γ. The Wald statistic for $\gamma = 0$ has distribution $\mathrm{F}_{K^*,n-K-K^*}$ and this procedure is equivalent to the Hausmann test described above.

Part III
Likelihood based methods

11 Maximum Likelihood estimation

Fisher, 1925. Maximum Likelihood estimation is more widely applicable (not only linear regression models). For iid data:

$$L(y;\theta) = \prod_{t=1}^{T} f(y_t;\theta)$$

more generally for dependent (but id) variables:

$$L(y;\theta) = \prod_{t=1}^{T} f(y_t|y_{t-1},\ldots,y_1;\theta)$$

Maximum Likelihood Estimator (MLE):

$$\widehat{\theta}_{ML} = \arg\max_{\theta} L(y;\theta)$$

Usually better to use the log-likelihood function:

$$l(y;\theta) = \ln L(y;\theta) = \sum_{t=1}^{T} \ln f(y_t \mid y_{t-1},\ldots,y_1;\theta)$$

First order condition $s(y;\theta) = \frac{\partial l}{\partial \theta} = 0$. [$s(y;\theta)$ is called the score vector]. Nice asymptotic properties:

- Consistency: $\text{plim}_{T\to\infty}\widehat{\theta}_{ML} = \theta$
- Asymptotic normality: $\sqrt{T}(\widehat{\theta}_{ML}-\theta) \longrightarrow N(0,[\frac{1}{T}I(\theta)]^{-1})$ and $\widehat{\theta}_{ML} \longrightarrow N(\theta,[I(\theta)]^{-1})$.
- Here $I(\theta)$ is the Fisher information matrix. Defined as the variance of the score vector:

$$I(\theta) = E\left[\frac{\partial l}{\partial \theta}\frac{\partial l}{\partial \theta}'\right]$$

- It also equals (minus) the expected value of the Hessian (information matrix equality):

$$I(\theta) = E\left[\frac{\partial l}{\partial \theta}\frac{\partial l}{\partial \theta}'\right] = -E\left[\frac{\partial^2 l}{\partial \theta \partial \theta'}\right]$$

- Asymptotic efficiency: $[I(\theta)]^{-1}$.The MLE has variance at least as small as the best unbiased estimate of θ. The MLE is generally not unbiased, but its bias is small making the comparison with unbiased estimates and the Cramer–Rao bound appropriate.

- Estimators of asymptotic variance matrix $[\frac{1}{T}I(\theta)]^{-1}$ (since $\sqrt{T}(\widehat{\theta}_{ML} - \theta) \longrightarrow N(0, [\frac{1}{T}I(\theta)]^{-1})$) are based on the information matrix computed at the ML estimates:

$$\left[\frac{1}{T}\sum_{t=1}^{T}(\frac{\partial l_t(y;\widehat{\theta}_{ML})}{\partial \theta})(\frac{\partial l_t(y;\widehat{\theta}_{ML})}{\partial \theta})'\right]^{-1} \rightarrow [\frac{1}{T}I(\theta)]^{-1} \text{ or } \left[-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_t(y;\widehat{\theta}_{ML})}{\partial \theta \partial \theta'}\right]^{-1} \rightarrow [\frac{1}{T}I(\theta)]^{-1}.$$

- Since $\widehat{\theta}_{ML} \longrightarrow N(\theta, [I(\theta)]^{-1})$ this gives

$$\left[\sum_{t=1}^{T}(\frac{\partial l_t(y;\widehat{\theta}_{ML})}{\partial \theta})(\frac{\partial l_t(y;\widehat{\theta}_{ML})}{\partial \theta})'\right]^{-1} \rightarrow [I(\theta)]^{-1} \text{ or } \left[-\sum_{t=1}^{T}\frac{\partial^2 l_t(y;\widehat{\theta}_{ML})}{\partial \theta \partial \theta'}\right]^{-1} \rightarrow [I(\theta)]^{-1}.$$

- Invariance. If $g(\theta)$ is a continuous function of θ, the ML estimator of $g(\theta)$ is $g(\widehat{\theta}_{ML})$

11.1 ML estimation of CLRM

Linear Model

$$y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 I)$$

Gaussianity:

$$f(\varepsilon) = (2\pi)^{-T/2}\left|\sigma^2 I\right|^{-1/2}\exp\{-0.5(\varepsilon'(\sigma^2 I)^{-1}\varepsilon)\}$$

Transformation:

$$f_y(y|X) = \left|\frac{\partial r^{-1}(y)}{\partial y}\right| f_\varepsilon(r^{-1}(y))$$

where $\varepsilon = r^{-1}(y) = y - X\beta$ and $\left|\frac{\partial r^{-1}(y)}{\partial y}\right| = 1$.

$$f(y) = (2\pi)^{-T/2}\left|\sigma^2 I\right|^{-1/2}\exp\{-0.5((y - X\beta)'(\sigma^2 I)^{-1}(y - X\beta))\}$$

in logs:

$$\ln L = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Let $\theta = (\beta' \ \sigma^2)'$. First order conditions: $s(\theta; y) = \frac{\partial l}{\partial \theta} = 0$ gives

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= -\frac{1}{\sigma^2}(-X'y + X'X\beta) = 0 \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) = 0
\end{aligned}$$

$$\begin{aligned}
\widehat{\beta}_{ML} &= (X'X)^{-1}X'y = \widehat{\beta}_{OLS} \\
\widehat{\sigma}^2_{ML} &= \frac{1}{T}(y - X\widehat{\beta}_{ML})'(y - X\widehat{\beta}_{ML}) = \frac{1}{T}\hat{\varepsilon}'\hat{\varepsilon}
\end{aligned}$$

Second order conditions

$$\begin{bmatrix} \frac{\partial^2 l}{\partial\beta\partial\beta'} & \frac{\partial^2 l}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 l}{\partial\sigma^2\partial\beta} & \frac{\partial^2 l}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{X'X}{\sigma^2} & -\frac{X'\varepsilon}{\sigma^4} \\ -\frac{(X'\varepsilon)'}{\sigma^4} & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\varepsilon'\varepsilon \end{bmatrix}$$

Information matrix:

$$[I(\theta)] = -E[\frac{\partial l_t}{\partial\theta\partial\theta'}] = -\begin{bmatrix} E[-\frac{X'X}{\sigma^2}] = -\frac{X'X}{\sigma^2} & E[-\frac{X'\varepsilon}{\sigma^4}] = 0 \\ E[-\frac{X'\varepsilon}{\sigma^4}] = 0 & E\left[\frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\varepsilon'\varepsilon\right] = -\frac{T}{2\sigma^4} \end{bmatrix}$$

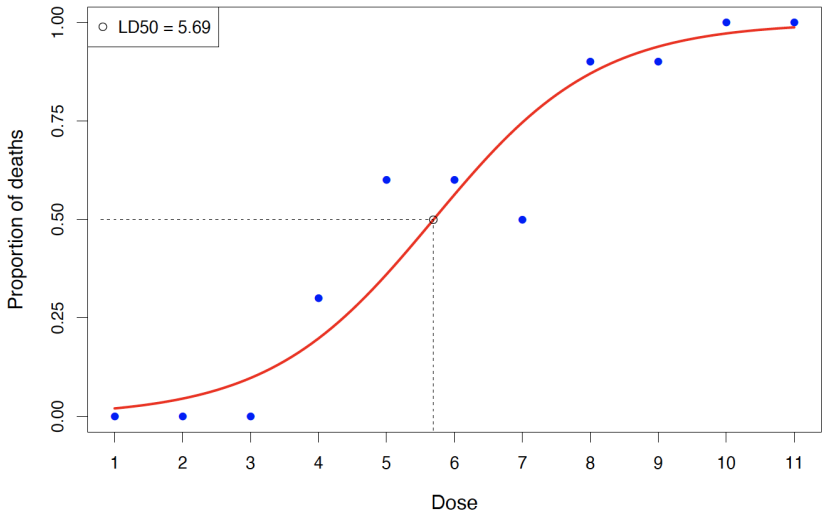
where we used $E[\frac{\varepsilon'\varepsilon}{\sigma^6}] = E[\frac{\sigma^2 T}{\sigma^6}] = \frac{T}{\sigma^4}$. We have:

$$[I(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/T \end{bmatrix}$$

11.2 Logistic regression

[Example from Efron, "Computer age statistical inference"] An experimental new anti-cancer drug called Xilathon is under development. Before human testing can begin, animal studies are needed to determinesafe dosages. To this end, a bioassay or dose–response experiment was carried out: $N = 11$ groups of $n = 10$ mice each were injected with increasing amounts of Xilathon. Let

$$y_i = \text{number of mice dying in group } i$$



The counts are modelled with a binomial:

$$y_i \overset{iid}{\sim} Bi(n_i = 10, \pi_i), \;\; i = 1, ..., N = 11$$

where π_i is the probability of death in group i, which is $prob(death|x_i)$. The observed sample proportion is:

$$p_i = y_i/n_i = y_i/10$$

which is the direct evidence for π_i ($\pi_i = E[p_i|X]$). By plotting the p_i we can see that a linear regression such as

$$p_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is not ideal. Indeed we have that $E[p_i|x_i] = \beta_0 + \beta_1 x_i$ which is not necessarily a probabiliy. Moreover, the figure above suggests a nonlinear relationship. Instead, for the conditioanl mean use a non-linear funcion with desirable properties. In particular

$$p_i = \Lambda(X_i'\beta) + \varepsilon_i$$

where

$$\Lambda(X_i'\beta) = \frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e^{(\beta_0+\beta_1 x_i)}}$$

is a good candidate as it is between 0 and 1 and allows nonlinear behaviours. In this case we have

$$\pi_i = E[p_i|X] = \frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e^{(\beta_0+\beta_1 x_i)}}$$

and

$$1 - \pi_i = 1 - E[p_i|X] = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_i)}}.$$

Taking logs of the ratio:

$$\lambda_i = \ln \frac{\pi_i}{1 - \pi_i} = \ln e^{(\beta_0 + \beta_1 x_i)} = \beta_0 + \beta_1 x_i$$

shows that this is a model in which the log of the ratio between the probabilities π_i and $1 - \pi_i$ is a linear function of x_i. The coefficient λ_i is the logit parameter.

The likelihood is:

$$\Pr ob(y_1, y_2, ..., y_N|X) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \Pi_{y_i=0}(1 - \pi_i)\Pi_{y_i=1}\pi_i = \Pi_{i=1}^{N}(1 - \pi_i)^{n_i - y_i}\pi_i^{y_i},$$

and if we recall that

$$\pi_i = \frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}} = \Lambda(X_i'\beta)$$

we can write, in logs:

$$\ln L \propto \sum_{i=1}^{N} (n_i - y_i) \ln(1 - \Lambda(X_i'\beta)) + y_i \ln \Lambda(X_i'\beta)$$

The first order condition is

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} \left[(n_i - y_i) \frac{-\Lambda'(X_i'\beta)}{1 - \Lambda(X_i'\beta)} + y_i \frac{\Lambda'(X_i'\beta)}{\Lambda(X_i'\beta)} \right] X_i = 0.$$

Note that

$$\Lambda'(X_i'\beta) = \frac{e^{(\beta_0 + \beta_1 x_i)}}{(1 + e^{(\beta_0 + \beta_1 x_i)})^2} = \Lambda(X_i'\beta)(1 - \Lambda(X_i'\beta))$$

so we have

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= \sum_{i=1}^{N} \left[-(n_i - y_i)\Lambda(X_i'\beta) + y_i(1 - \Lambda(X_i'\beta)) \right] X_i \\ &= \sum_{i=1}^{N} \left[-n_i\Lambda(X_i'\beta) + y_i\Lambda(X_i'\beta) + y_i - y_i\Lambda(X_i'\beta)) \right] X_i \\ &= \sum_{i=1}^{N} \left[y_i - n_i\Lambda(X_i'\beta) \right] X_i = 0 \end{aligned}$$

which is a set of orthogonality conditions based on the logistic function $\Lambda(X_i\beta)$ as opposed as a linear specification $X_i\beta$. The second derivative is:

$$H = \frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} n_i \Lambda'(X_i'\beta) X_i X_i' = -\sum_{i=1}^{N} n_i \Lambda(X_i'\beta)(1-\Lambda(X_i'\beta)) X_i X_i'$$

which is always negative definite. The simpler case in which $n_i = 1$ is the "discrete choice model".

The solution is also the one that minimizes the deviance:

$$D(\pi_i, p_i) = 2n_i \left[p_i \ln \frac{p_i}{\pi_i} + (1-p_i) \ln \frac{(1-p_i)}{(1-\pi_i)} \right],$$

where the term in the square brackets is the expectation taken using the probabilities p of the logarithmic difference between the probabilities p and π. Deviance is **analogous to squared error** in ordinary regression theory, and can be seen as its generalization. It is twice the "Kullback–Leibler divergence" the preferred name in the information-theory literature.[9] When $p_i = \pi_i$ the deviance is 0.

11.3 ML estimation of GLRM

We no longer necessarily have iid data. Linear Model

$$y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2\Omega)$$

$$\begin{aligned} \ln L &= -\frac{T}{2} \ln \left|2\pi\sigma^2\Omega\right| - \frac{1}{2\sigma^2}(y-X\beta)'\Omega^{-1}(y-X\beta) \\ &= -\frac{T}{2}\log 2\pi - \frac{T}{2}\ln\left|\sigma^2\right| - \frac{1}{2}\log|\Omega| - \frac{1}{2\sigma^2}(y-X\beta)'\Omega^{-1}(y-X\beta), \end{aligned}$$

here we assume Ω is known, hence we can just focus on:

$$\ln L \propto -\frac{T}{2}\log\left|\sigma^2\right| - \frac{1}{2\sigma^2}(y-X\beta)'\Omega^{-1}(y-X\beta)$$

First order conditions: $s(\theta; y) = \frac{\partial l}{\partial\theta} = 0$ gives

$$\begin{aligned} \frac{\partial l}{\partial\beta} &= -\frac{1}{\sigma^2}(-X'\Omega^{-1}y + X'\Omega^{-1}X\beta) = 0 \\ \frac{\partial l}{\partial\sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(y-X\beta)'\Omega^{-1}(y-X\beta) = 0 \end{aligned}$$

[9] The KL divergence is the expected value of a log-likelihood ratio if the data are actually drawn from the distribution at the numerator of the ratio, i.e. under the null hypothesis for which that ratio is computed.

$$\begin{aligned}\widehat{\beta}_{ML} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \widehat{\beta}_{GLS} \\ \widehat{\sigma}^2_{ML} &= \frac{1}{T}(y - X\widehat{\beta}_{ML})'\Omega^{-1}(y - X\widehat{\beta}_{ML}) = \frac{1}{T}\hat{\varepsilon}'\Omega^{-1}\hat{\varepsilon}\end{aligned}$$

This assumed Ω known. If Unknown, we need to consider $\Omega(\theta)$ and maximize w.r.t. θ as well.

11.3.1 ARCH models

Consider a Gaussian ARCH(m) process:

$$\begin{aligned}y_t &= \beta' x_t + \eta_t; \\ \eta_t &= h_t\varepsilon_t;\ \varepsilon_t \sim N(0,1) \\ h_t^2 &= \alpha_0 + \alpha_1\eta_{t-1}^2 + ... + \alpha_m\eta_{t-m}^2,\end{aligned}$$

note h_t^2 is a deterministic function of η_{t-1}^2. Conditions on α_0 and α_1 are needed ($\alpha_0 > 0; \alpha_1 \geq 0$ and $\alpha_1 < 1$ at the least. More for existence of higher moments). The conditional expectation is:

$$E[\eta_t|I_{t-1}] = E[h_t\varepsilon_t|I_{t-1}] = h_t E[\varepsilon_t|I_{t-1}] = 0$$

The conditional variance is:

$$V[\eta_t|I_{t-1}] = E[\eta_t^2|I_{t-1}] - E[\eta_t|I_{t-1}]^2 = E[h_t^2\varepsilon_t^2|I_{t-1}] = h_t^2 E[\varepsilon_t^2|I_{t-1}] = h_t^2$$

It follows that:

$$\eta_t|I_{t-1} \sim N(0, h_t^2).$$

Note that the unconditional distribution of η_t is not normal. It is a mixture of normal distributions with different (and random) variances. It can be shown that the resulting unconditional distribution has fatter tails than a normal distribution:

$$\kappa(\eta_t) = \frac{E[\eta_t^4]}{E[\eta_t^2]^2} = \frac{E[h_t^4\varepsilon_t^4]}{E[h_t^2\varepsilon_t^2]^2} = \frac{E[h_t^4]E[\varepsilon_t^4]}{E[h_t^2]^2 E[\varepsilon_t^2]^2} = \frac{E[h_t^4]}{E[h_t^2]^2}3 \geq 3$$

where the last inequality follows from Jensen's inequality $E[f(x)] \geq f(E[x])$.Therefore, the ARCH is consistent with the fact thar returns are not normal. It is also consistent with the volatility clustering phenomenon. To see this, consider the forecast error:

$$\eta_t^2 - E[\eta_t^2|I_{t-1}] = \eta_t^2 - h_t^2 = \nu_t$$

which implies

$$\eta_t^2 = h_t^2 + \nu_t = \alpha_0 + \alpha_1 \eta_{t-1}^2 + \nu_t$$

which means η_t^2 follows an AR(1).

As we said, the conditional distribution of η_t is $\eta_t | I_{t-1} \sim N(0, h_t^2)$, with pdf:

$$f(\eta_t | I_{t-1}) \propto \frac{1}{\sqrt{h_t^2}} \exp\left(-\frac{\eta_t^2}{2h_t^2}\right).$$

The conditional density of y_t is therefore:

$$f(y_t | I_{t-1}; \boldsymbol{\alpha}, \beta) \propto \frac{\exp\left(-\frac{1}{2}\frac{(y_t - \beta' x_t)^2}{\alpha_0 + \alpha_1 (y_{t-1} - \beta' x_t)^2 + ... + \alpha_m (y_{t-m} - \beta' x_t)^2}\right)}{\sqrt{(\alpha_0 + \alpha_1 (y_{t-1} - \beta' x_t)^2 + ... + \alpha_m (y_{t-m} - \beta' x_t)^2)}}$$

(Jacobian is 1). Here $\boldsymbol{\alpha}$=$\alpha_0, \alpha_1, ..., \alpha_m$ is the vector of coefficients. Now consider a sample of T observations. The Likelihood (conditional on the first m observations) is:

$$\begin{aligned} L &\propto f(y_T, ..., y_{m+1} | \boldsymbol{\alpha}, \beta, y_m, ..., y_1) \\ &= f(y_T | I_{T-1}) \times f(y_{T-1} | I_{T-2}) \times ... \times f(y_{m+1} | I_m) \\ &= \Pi_{t=m+1}^{T} \frac{\exp\left(-\frac{1}{2}\frac{(y_t - \beta' x_t)^2}{\alpha_0 + \alpha_1 (y_{t-1} - \beta' x_t)^2 + ... + \alpha_m (y_{t-m} - \beta' x_t)^2}\right)}{\sqrt{(\alpha_0 + \alpha_1 (y_{t-1} - \beta' x_t)^2 + ... + \alpha_m (y_{t-m} - \beta' x_t)^2)}} \end{aligned}$$

The function $\log L$ can be minimized wrt $\boldsymbol{\alpha}$ and β. Remember there are restrictions on $\boldsymbol{\alpha}$, so it is a constrained estimation. Engle (1982) shows that the information matrix is such that the ML estimators of $\boldsymbol{\alpha}$ and β are independent. Parameters $\boldsymbol{\alpha}, \beta$ can be estimated with full efficiency based only on a consistent estimate of the other.

11.3.2 ARMA models

Consider the ARMA(2,1):

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + u_t + \vartheta u_{t-1}, \quad u_t \sim \text{i.i.d. } N\left(0, \sigma_u^2\right).$$

One could write down the likelihood for the entire sample and maximize, as we did in the GARCH example. There is a quicker way to evaluate the likelihood using:

1. The prediction error decomposition, and
2. A recursive algorithm known as the Kalman filter

Prediction error decomposition To understand what the PED is, consider the simpler model:

$$Y_t = \phi_1 Y_{t-1} + u_t, \quad u_t \sim \text{i.i.d. } N\left(0, \sigma_u^2\right)$$

and the forecast:

$$\begin{aligned} Y_{t|t-1} &\equiv E[Y_t|I_{t-1}] = \phi_1 Y_{t-1}, \\ \Sigma_{t|t-1} &\equiv VAR[Y_t|I_{t-1}] = \sigma_u^2. \end{aligned}$$

This forecast has forecast error:

$$v_{t|t-1} \equiv Y_t - E[Y_t|I_{t-1}] = u_t,$$

which has variance $Var[v_{t|t-1}] = \sigma_u^2 = \Sigma_{t|t-1}$, i.e. the same as $VAR[Y_t|I_{t-1}]$. The likelihood for observation t can be written as:

$$l(Y_t|Y_{1:t-1};\theta) \propto -\ln|\Sigma_{t|t-1}(\theta)| - v'_{t|t-1}(\theta)\,\Sigma^{-1}_{t|t-1}(\theta)\,v_{t|t-1}(\theta).$$

For more complex models, e.g. the ARMA model, computing the PED moments $\Sigma_{t|t-1}$ and $v_{t|t-1}$ can be easily done using a prediction-update algorithm known as the Kalman Filter.

11.4 State space models (linear and Gaussian)

The linear Gaussian SSM consists of two equations

$$\begin{aligned} \text{Space (measurement, observation):} \quad & Y_t = \Phi s_t + \varepsilon_t, \\ \text{State (transition):} \quad & s_t = F s_{t-1} + \eta_t. \end{aligned}$$

$$\begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \sim iidN\left(0, \begin{bmatrix} \Omega_\varepsilon & 0 \\ 0 & \Omega_\eta \end{bmatrix}\right).$$

- Y_t vector of N observed variables, while s_t vector of k unobserved states
- Φ and F are $N \times k$ and $k \times k$ coefficient matrices
- Intercepts or additional exogenous regressors in both equations are omitted but can be introduced easily.
- Similarly, time variation in the coefficient matrices Φ and F can be allowed

The Kalman Filter - Learning about states from data

- An algorithm that produces and updates linear projections of the latent variable s_t given observations of Y_t.
- Useful in its own right, and is also employed in ML estimation.
- Define:

$$s_{t|s} := E\left[s_t|Y_{1:s}\right], \quad P_{t|s} := \mathrm{Var}\left[s_t|Y_{1:s}\right].$$

- We then wish to do:

$$\begin{array}{rl} \text{Filtering:} & s_{t|t} \text{ and } P_{t|t}, \quad t = 1, ..., T. \\ \text{Smoothing:} & s_{t|T} \text{ and } P_{t|T}, \quad t = 1, ..., T. \end{array}$$

Derivation of Kalman Filter

- Assume you start by knowing $s_{t-1} \sim N(s_{t-1|t-1}, P_{t-1|t-1})$. The filter is a rule to update to $s_t \sim N(s_{t|t}, P_{t|t})$ once we observe data Y_t.
- The first step is to find the joint distribution of states and data in t, conditional on past observations $(1, ..., t-1)$:

$$\left[\begin{array}{c} s_t \\ Y_t \end{array}\right] \Bigg| I_{t-1} \sim N\left(\left[\begin{array}{c} s_{t|t-1} \\ Y_{t|t-1} \end{array}\right], \left[\begin{array}{cc} P_{t|t-1} & C'_{t|t-1} \\ C_{t|t-1} & \Sigma_{t|t-1} \end{array}\right]\right) \tag{25}$$

- The moments above can be calculated easily using the equations of the system, $Y_t = \Phi s_t + \varepsilon_t, \quad s_t = F s_{t-1} + \eta_t \Rightarrow$

$$\begin{array}{rcl} s_{t|t-1} & = & E\left[s_t|Y_{t-1}\right] = FE\left[s_{t-1}|Y_{t-1}\right] + E\left[\eta_t|Y_{t-1}\right] = Fs_{t-1|t-1}, \\ Y_{t|t-1} & = & E\left[Y_t|Y_{t-1}\right] = \Phi E\left[s_t|Y_{t-1}\right] + E\left[\varepsilon_t|Y_{t-1}\right] = \Phi s_{t|t-1}. \\ P_{t|t-1} & = & \mathrm{Var}\left[s_t|Y_{t-1}\right] = FP_{t-1|t-1}F' + \Omega_\eta, \\ \Sigma_{t|t-1} & = & \mathrm{Var}\left[Y_t|Y_{t-1}\right] = \Phi P_{t|t-1}\Phi' + \Omega_\varepsilon, \\ C_{t|t-1} & = & \mathrm{Cov}\left(Y_t, s_t|Y_{t-1}\right) = \Phi P_{t|t-1}. \end{array} \tag{26}$$

Conditional and joint normals

- We have now specified the distribution $(s_t, Y_t|Y_{t-1})$ we now look for the distribution $(s_t|Y_t, Y_{t-1}) = (s_t|Y_t)$.

- This is easy to do using basic results regarding Normal distributions: Let (a, b) be normally distributed,

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Omega_{aa} & \Omega_{ab} \\ \Omega_{ba} & \Omega_{bb} \end{bmatrix} \right). \tag{27}$$

  Then the conditional distribution of a conditional on b is given as

$$a|b \sim N\left(\mu_{a|b}, \Omega_{a|b}\right), \tag{28}$$

  where

$$\mu_{a|b} = \mu_a + \Omega_{ab}\Omega_{bb}^{-1}\left(b - \mu_b\right), \quad \Omega_{a|b} = \Omega_{aa} - \Omega_{ab}\Omega_{bb}^{-1}\Omega_{ba}.$$

Updating a linear projection

- Set (27) to be the joint distribution in (25)

$$\begin{bmatrix} a = s_t \\ b = Y_t \end{bmatrix} \Bigg| Y_{t-1} \sim N\left(\begin{bmatrix} \mu_a = s_{t|t-1} \\ \mu_b = Y_{t|t-1} \end{bmatrix}, \begin{bmatrix} \Omega_{aa} = P_{t|t-1} & \Omega_{ab} = C'_{t|t-1} \\ \Omega_{ba} = C_{t|t-1} & \Omega_{bb} = \Sigma_{t|t-1} \end{bmatrix} \right)$$

- Applying the result (28):

$$s_t | \, Y_{t-1}, Y_t \sim N\left(\mu_{a|b} = s_{t|t}, \Omega_{a|b} = P_{t|t}\right) \tag{29}$$

  with

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + C'_{t|t-1}\Sigma_{t|t-1}^{-1}\left(Y_t - Y_{t|t-1}\right) && (30) \\ P_{t|t} &= P_{t|t-1} - C'_{t|t-1}\Sigma_{t|t-1}^{-1}C_{t|t-1} && (31) \end{aligned}$$

- So we have moved from $(s_{t-1}|Y_{t-1})$ to $(s_t, Y_t|Y_{t-1})$ (prediction) and then from $(s_t, Y_t|Y_{t-1})$ to $(s_t|Y_t)$ (update).
- We can now repeat and use $(s_t|Y_t)$ to move forward to $(s_{t+1}|Y_{t+1})$
- Using $C_{t|t-1} = \Phi P_{t|t-1}$ (see (26)) the updating equations can be re-written as:

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + K_{t|t-1}v_{t|t-1} && (32) \\ P_{t|t} &= P_{t|t-1} - K_{t|t-1}\Phi P_{t|t-1} && (33) \end{aligned}$$

  with

$$K_{t|t-1} = P_{t|t-1}\Phi'\Sigma_{t|t-1}^{-1} \tag{34}$$

  denoting the Kalman Gain and $v_{t|t-1} = Y_t - Y_{t|t-1}$ denoting the 1-step ahead prediction error.

The Kalman Filter recursions The algorithm works as follows:

0) Start with an initial condition $(s_{t-1}|Y_{t-1}) \sim N(s_{t-1|t-1}, P_{t-1|t-1})$

1) Use the prediction equations to find the moments of $(s_t, Y_t|Y_{t-1})$:

$$\begin{aligned} s_{t|t-1} &= F s_{t-1|t-1} \text{ (state prediction)} \\ P_{t|t-1} &= F P_{t-1|t-1} F' + \Omega_\eta \text{ (variance of state prediction)} \\ Y_{t|t-1} &= \Phi s_{t|t-1} \text{ (measurement prediction)} \\ \Sigma_{t|t-1} &= \Phi P_{t|t-1} \Phi' + \Omega_\varepsilon \text{ (variance of measurement prediction)} \end{aligned}$$

1.5 Compute the Kalman gain:

$$K_{t|t-1} = P_{t|t-1} \Phi' \Sigma_{t|t-1}^{-1}$$

2 Observe the prediction error $v_{t|t-1} = Y_t - Y_{t|t-1}$ and update to find $(s_t|Y_t) \sim N(s_{t|t}, P_{t|t})$:

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + K_{t|t-1} v_{t|t-1} \\ P_{t|t} &= P_{t|t-1} - K_{t|t-1} \Phi P_{t|t-1} \end{aligned}$$

Likelihood via prediction error decomposition

- As a by product, the algorithm will provide the rime series of $v_{t|t-1}$ and $\Sigma_{t|t-1}$ for $t = 1, ...T$.
- So at each $t = 1, ...T$ we can compute and store:

$$l(Y_t|Y_{1:t-1}; \theta) \propto -\ln |\Sigma_{t|t-1}(\theta)| - v'_{t|t-1}(\theta) \Sigma_{t|t-1}^{-1}(\theta) v_{t|t-1}(\theta)$$

  where

$$\theta = f^{-1}(\Phi,\ F,\ \Omega_\varepsilon,\ \Omega_\eta)$$

- The sum of the likelihoods of the forecast errors $\sum l_t(\theta)$ provides the likelihood of the whole system
- Therefore the KF offers a fast way to evaluate the likelihood of a SS model.

The ARMA example

$$
\begin{aligned}
Y_t &= \begin{bmatrix} 1 & \vartheta \end{bmatrix} \begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} = s_{1t} + \vartheta s_{2t} \\
\begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} &= \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} s_{1t-1} \\ s_{2t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \end{bmatrix},
\end{aligned}
$$

with $\Omega_\varepsilon = 0$, $\Omega_\eta = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{bmatrix}$. Indeed:

$$
\begin{aligned}
Y_t &= \phi_1 s_{1t-1} + \phi_2 s_{2t-1} + u_t + \vartheta(+\phi_1 s_{1t-2} + \phi_2 s_{2t-2} + u_{t-1}) \\
&= \phi_1(s_{1t-1} + \vartheta s_{1t-2}) + \phi_2(s_{2t-1} + \vartheta s_{2t-2}) + u_t + \vartheta u_{t-1} \\
&= \phi_1(s_{1t-1} + \vartheta s_{2t-1}) + \phi_2(s_{1t-2} + \vartheta s_{2t-2}) + u_t(1 + \vartheta)
\end{aligned}
$$

12 James - Stein Estimators

12.1 James-Stein result

Let $\{z_t\}_{t=1}^T$ be a sequence of iid k-variate Gaussians with mean θ and variance I_p:

$$z_t|\theta \sim N(\theta, I_k).$$

As an estimator of the population mean, the sample mean

$$\hat{\theta}^{ML} = \bar{z}_T = \frac{\Sigma_{t=1}^T z_t}{T}$$

is the ML estimator and as such it is Minimum Variance Unbiased.

However, Stein (1956) showed that this is inadmissible with respect to the MSE loss whenever $k \geq 3$. (Inadmissible is decision theory jargon to say that is strictly dominated). Moreover, James and Stein (1961) proposed an alternative estimator which always improves over $\hat{\theta}^{ML}$ in terms of MSE (when $k \geq 3$):

$$\hat{\theta}^{JS} = \left(1 - \frac{(k-2)}{\|\bar{z}_T\|^2}\right) \bar{z}_T,$$

where $\|\cdot\|$ is the Euclidian norm, hence

$$\|\bar{z}_T\|^2 = \left(\sqrt{\bar{z}_1^2 + \bar{z}_2^2 + ... + \bar{z}_k^2}\right)^2 = \bar{z}_T' \bar{z}_T$$

and we can re-write

$$\hat{\theta}^{JS} = \left(1 - \frac{(k-2)}{\bar{z}_T' \bar{z}_T}\right) \bar{z}_T.$$

Since $\frac{(k-2)}{\bar{z}_T' \bar{z}_T}$ is a positive quantity, $\hat{\theta}^{JS}$ is closer to 0 than $\hat{\theta}^{ML}$: shrinkage.

12.2 Application to CLRM

This result applies to the CLRM as well. In that case we have that the k-variate Gaussian is:

$$\hat{\beta}|\beta \sim N(\beta, \sigma^2(X'X)^{-1})$$

Standardizing gives

$$\underbrace{\sigma^{-1}(X'X)^{1/2}\hat{\beta}}_{z_t}|\beta \sim N(\underbrace{\sigma^{-1}(X'X)^{1/2}\beta}_{\theta}, I_k).$$

Here we only have 1 realization of $\hat{\beta}$ so $T = 1$ and:

$$\begin{aligned} \hat{\theta}^{ML} &= \bar{z}_T = \frac{\Sigma_{t=1}^{1} z_t}{1} = z_1 = \frac{1}{\sigma}(X'X)^{1/2}\hat{\beta} \\ \bar{z}_T'\bar{z}_T &= \frac{1}{\sigma^2}\hat{\beta}'X'X\hat{\beta} \end{aligned}$$

and the JS of θ is:

$$\hat{\theta}^{JS} = \left(1 - \frac{\sigma^2(k-2)}{\hat{\beta}'X'X\hat{\beta}}\right)\frac{1}{\sigma}(X'X)^{1/2}\hat{\beta} \to \underbrace{\sigma^{-1}(X'X)^{1/2}\beta}_{\theta}$$

To go back to the representation in terms of β just undo the standardization:

$$\left(\sigma^{-1}(X'X)^{1/2}\right)^{-1}\hat{\theta}^{JS} = \left(1 - \frac{\sigma^2(k-2)}{\hat{\beta}'X'X\hat{\beta}}\right)\hat{\beta} \to \left(\sigma^{-1}(X'X)^{1/2}\right)^{-1}\theta = \beta$$

that is

$$\hat{\beta}^{JS} = \left(1 - \frac{\sigma^2(k-2)}{\hat{\beta}'X'X\hat{\beta}}\right)\hat{\beta} \to \beta.$$

12.3 Bayesian interpretation

This result was introduced in frequentist terms. However, the JS estimator has a straightworward Bayesian interpretation. Suppose we wish to estimate the coefficient vetor:

$$\mu_1, \mu_2, ..., \mu_k$$

on which we have prior

$$\mu_i \sim N(M_0, V_0),$$

by using a single observation vector x:

$$x_1, x_2, ..., x_k$$

with likelihood

$$x_i|\mu_i \sim N(\mu_i, 1).$$

The MLE is

$$\hat{\mu}_i^{ML} = x_i$$

The Bayesian posterior mean is

$$\begin{aligned}
\hat{\mu}_i^{Bayes} &= (V_0^{-1}+1)^{-1}(V_0^{-1}M_0 + x_i) \\
&= \left(\frac{1+V_0}{V_0}\right)^{-1}(V_0^{-1}M_0 + x_i) \\
&= (V_0+1)^{-1}(M_0 + V_0 x_i) \\
&= M_0 + \frac{V_0}{V_0+1}(x_i - M_0) \qquad (35)
\end{aligned}$$

It turns out that the JS estimator is the plug-in version of (35)

$$\hat{\mu}_i^{JS} = \hat{M}_0 + \frac{\widehat{V_0}}{V_0+1}(x_i - \hat{M}_0)$$

with

$$\begin{aligned}
\hat{M}_0 &= \bar{x} \\
\frac{\widehat{V_0}}{V_0+1} &= 1 - \frac{k-3}{S};\ S = \Sigma_{i=1}^k (x_i - \bar{x})^2.
\end{aligned}$$

Substituting gives:

$$\hat{\mu}_i^{JS} = \bar{x} + \left(1 - \frac{k-3}{S}\right)(x - \bar{x}). \qquad (36)$$

This is Lindlay (1962) version of JS estimator, which shrinks towards $\bar{x}$ rather than towards 0. The standard case is imediately obtained by setting $M_0 = \hat{M}_0 = 0$. In this case, the MSE benefit comes from the fact that we are using the other observations to form a point to which to shrink the estimates. Note that we are using the data to compute an estimate of the prior mean $\hat{M}_0 = \bar{x}$! This is called "Empirical Bayes".

Table 7.1 *Eighteen baseball players;* `MLE` *is batting average in first 90 at bats;* `TRUTH` *is average in remainder of 1970 season; James–Stein estimator* `JS` *is based on arcsin transformation of MLEs. Sum of squared errors for predicting* `TRUTH`*:* `MLE` *.0425,* `JS` *.0218.*

Player	MLE	JS	TRUTH	x
1	.345	.283	.298	11.96
2	.333	.279	.346	11.74
3	.322	.276	.222	11.51
4	.311	.272	.276	11.29
5	.289	.265	.263	10.83
6	.289	.264	.273	10.83
7	.278	.261	.303	10.60
8	.255	.253	.270	10.13
9	.244	.249	.230	9.88
10	.233	.245	.264	9.64
11	.233	.245	.264	9.64
12	.222	.242	.210	9.40
13	.222	.241	.256	9.39
14	.222	.241	.269	9.39
15	.211	.238	.316	9.14
16	.211	.238	.226	9.14
17	.200	.234	.285	8.88
18	.145	.212	.200	7.50

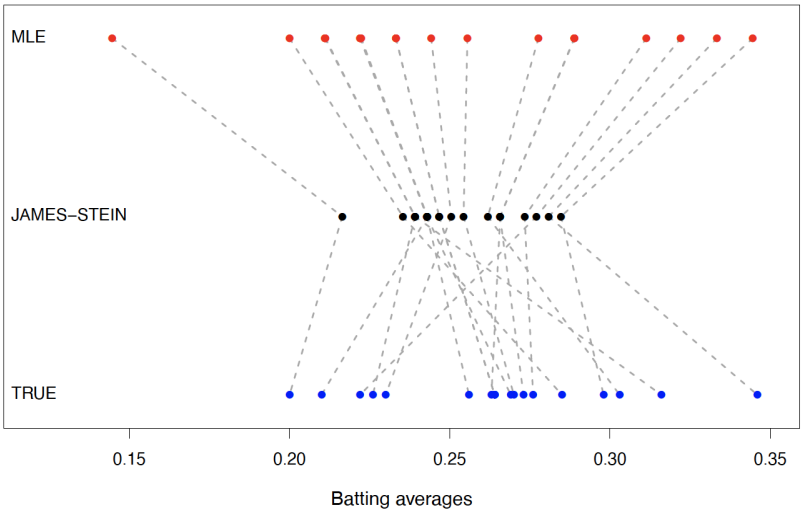


Figure 7.1 Eighteen baseball players; top line MLE, middle James–Stein, bottom true values. Only 13 points are visible, since there are ties.

13 Bayesian treatement of the linear regression model

13.1 Introduction

- Start with:

$$y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 I_T)$$

and get

$$\hat{\beta} = (X'X)^{-1} X'Y$$

- Add data:

$$\begin{bmatrix} Y \\ Y_1 \end{bmatrix} = \begin{bmatrix} X \\ X_1 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ \varepsilon_1 \end{bmatrix};\ \varepsilon \sim N(0, \sigma^2 I_{T+T_1})$$

and get

$$\begin{aligned} \hat{\beta} &= \left(\begin{bmatrix} X' & X_1' \end{bmatrix} \begin{bmatrix} X \\ X_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X' & X_1' \end{bmatrix} \begin{bmatrix} Y \\ Y_1 \end{bmatrix} \\ &= \left(X'X + X_1'X_1 \right)^{-1} (X'Y + X_1'Y_1) \end{aligned}$$

Bayesian approach

1. The researcher starts with a prior belief about the coefficient β. The prior belief is in the form of a distribution $p(\beta)$

$$\beta \sim N(\beta_0, \Sigma_0)$$

2. Collect data and write down the likelihood function as before $p(Y|\beta)$.

3. Update your prior belief on the basis of the information in the data. Combine the prior distribution $p(\beta)$ and the likelihood function $p(Y|\beta)$ to obtain the posterior distribution $p(\beta|Y)$

Key identities

- These three steps come from Bayes Theorem:

$$p(\beta|Y) = \frac{p(Y|\beta) \times p(\beta)}{p(Y)} \propto p(Y|\beta) \times p(\beta)$$

- Useful identities:

$$\underset{joint}{p(Y,\beta)} = \underset{data\ density}{p(Y)} \times \underset{posterior}{p(\beta|Y)} = \underset{likelihood}{p(Y|\beta)} \times \underset{prior}{p(\beta)}$$

- $p(Y)$ is the data density (also known as marginal likelihood).It is the constant of integration of the posterior:

$$\int p(\beta|Y)d\beta = \frac{1}{p(Y)} \int \underbrace{p(Y|\beta) \times p(\beta)}_{posterior\ kernel} d\beta = 1,$$

therefore it is not needed if we are only interested in the posterior kernel. The posterior kernel is sufficient to compute e.g. mean and variance of $p(\beta|Y)$.

$$p(\beta|Y) \propto p(Y|\beta) \times p(\beta)$$

Derivation of posterior We assume that σ^2 is known, recall that k is the number of regressors. Set prior distribution for $\beta \sim N(\beta_0, \Sigma_0)$

$$\begin{aligned} p(\beta) &= (2\pi)^{-\frac{k}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left[-0.5\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)\right] \\ &\propto \exp\left[-0.5\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)\right] \end{aligned}$$

Obtain data and form the likelihood function:

$$\begin{aligned} p(Y|\beta) &= (2\pi)^{-\frac{T}{2}} \left|\sigma^2 I_T\right|^{-\frac{1}{2}} \exp[-0.5\left(Y-X\beta\right)'\left(\sigma^2 I_T\right)^{-1}\left(Y-X\beta\right)] \\ &\propto \exp[-0.5\left(Y-X\beta\right)'\left(Y-X\beta\right)/\sigma^2] \end{aligned}$$

Obtain the posterior kernel

$$\begin{aligned} p(\beta|Y) &\propto p(Y|\beta) \times p(\beta) \\ &\propto \exp[-0.5\left(Y-X\beta\right)'\left(Y-X\beta\right)/\sigma^2] \exp[-\frac{1}{2}\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)] \\ &\propto \exp[-0.5\{\left(Y-X\beta\right)'\left(Y-X\beta\right)/\sigma^2 + \left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)] \end{aligned} \quad (37)$$

One can show that

$$p(\beta|Y) \propto \exp[-0.5\left(\beta-\beta_1\right)'\Sigma_1^{-1}\left(\beta-\beta_1\right)]$$

which is the kernel of a normal distribution with moments

$$\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1} \tag{38}$$

$$\beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'Y\right), \tag{39}$$

therefore we conclude:

$$\beta|Y \sim N(\beta_1, \Sigma_1)$$

Details Completing the squares of (37) gives:

$$\begin{aligned}\ln p(\beta|Y) \propto\ & \beta'\Sigma_0^{-1}\beta - \beta'\Sigma_0^{-1}\beta_0 - \beta_0'\Sigma_0^{-1}\beta + \beta_0'\Sigma_0^{-1}\beta_0 + \\ & \frac{1}{\sigma^2}Y'Y - \frac{1}{\sigma^2}Y'X\beta - \frac{1}{\sigma^2}\beta'X'Y + \frac{1}{\sigma^2}\beta'X'X\beta\end{aligned}$$

regrouping gives:

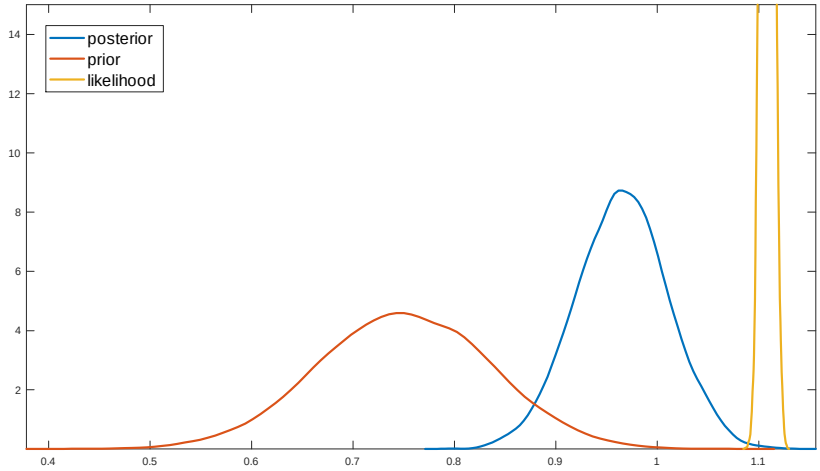
$$\begin{aligned}\ln p(\beta|Y) \propto\ & \beta'\underbrace{(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X)}_{\Sigma_1^{-1}}\beta - \beta'\underbrace{(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'Y)}_{\Sigma_1^{-1}\beta_1} + \\ & - \underbrace{(\beta_0'\Sigma_0^{-1} + \frac{1}{\sigma^2}Y'X)}_{\beta_1'\Sigma_1^{-1}}\beta + \beta_0'\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}Y'Y\end{aligned} \tag{40}$$

where the elements in braces are those in definitions (38) and (39). Rewrite the first term as:

$$\begin{aligned}\beta'\Sigma_1^{-1}\beta &= (\underbrace{\beta - \beta_1}_{} + \underbrace{\beta_1}_{})'\Sigma_1^{-1}(\beta - \beta_1 + \beta_1) \\ &= \underbrace{(\beta - \beta_1)'}\Sigma_1^{-1}\underbrace{(\beta - \beta_1)} + (\underbrace{\beta_1}_{})'\Sigma_1^{-1}(\underbrace{\beta_1}_{}) \\ &\quad + \underbrace{(\beta - \beta_1)'}\Sigma_1^{-1}(\underbrace{\beta_1}_{}) + (\underbrace{\beta_1}_{})'\Sigma_1^{-1}\underbrace{(\beta - \beta_1)}.\end{aligned}$$

The last three terms in the expression contain 5 terms: of these 3 terms $\beta_1'\Sigma_1^{-1}\beta_1$ with signs +,-,- simplify to just $-\beta_1'\Sigma_1^{-1}\beta_1$ and the other two terms are $\beta'\Sigma_1^{-1}\beta_1$ and ${\beta_1}'\Sigma_1^{-1}\beta$, which simplify with the 2nd and 3rd terms in (40). Therefore, we have:

$$\begin{aligned}\ln p(\beta|Y) &\propto (\beta - \beta_1)'\Sigma_1^{-1}(\beta - \beta_1) - \beta_1'\Sigma_1^{-1}\beta_1 + \beta_0'\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}Y'Y \\ &\propto (\beta - \beta_1)'\Sigma_1^{-1}(\beta - \beta_1)\end{aligned}$$



Comparison with OLS (or ML)

- Note that, as $X'X\hat{\beta} = X'Y$, we have:

$$\beta_1 = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'X\hat{\beta}\right)$$

- Σ_0^{-1} and $\Sigma_0^{-1}\beta_0$ are the prior moments, can be interpreted as dummy observations/pre-sample observations.
- Without the priors, these moments are simply the OLS estimates. Without the data, these moments are simply the priors
- The mean is a weighted average of the prior and OLS. The weights are inversely proportional to the variance of prior and data information
- Setting $\Sigma_0^{-1} = \frac{\lambda}{\sigma^2}I_k$ and $\beta_0 = 0$ gives the Ridge regression:

$$\beta_1 = (\lambda I_k + X'X)^{-1}X'Y.$$

The Likelihood Principle

- Consider tossing a drawing pin (Lindley and Phillips 1976). Luigi says he tossed it 12 times and obtained :

$$\{U, U, U, D, U, D, U, U, U, U, U, D\}$$

- You -as a statistician- are asked to give a 5% rejection region for the null that U and D are equally likely
- Obtaining 9 U's out of 12 suggests that the chance of its falling uppermost (U) exceeds 50%.The results that would even more strongly support this conclusion are:

  $$(10, 2),\ (11, 1),\ \textit{and}\ (12, 0),$$

  so that, under the null hypothesis $\theta = 1/2$, the chance of the observed result, or more extreme, is:

  $$\left\{\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right\}\left(\theta = \frac{1}{2}\right)^{12} = \mathbf{7.5}\% > 5\%$$

- Hence, you do NOT reject the null that U and D are equally likely (50%).
- However, now Luigi tells you: "but I didn't set to throw the pin 12 times. My plan was to throw the pin until 3 Ds appeared"
- Does this change your inference? Yes it does
- Under the new scenario, the more exreme events would be:

  $$(10, 3),\ (11, 3),\ (12, 3),\ ...,$$

  while the events $(10, 2),\ (11, 1)$, and $(12, 0)$ actually can NOT take place under this design.
- So the chance of the observed result under the null hypothesis becomes:

  $$\left\{1 - \binom{10}{2}\left(\frac{1}{2}\right)^{11} - \binom{9}{2}\left(\frac{1}{2}\right)^{10} - ... - \binom{2}{2}\left(\frac{1}{2}\right)^{3}\right\} = \mathbf{3.25}\% < 5\%$$

- Why is this happening? Because the two setups imply a different stopping rule (stop at 12 draws, or stop at 3 D draws). This, more generally, alters the sample space.
- Things are even more problematic. Think if Luigi says "I just kept drawing the pin until lunch was served". How would you tackle this?
- Confidence intervals similarly demand consideration of the sample space. Indeed, so does every statistical technique, with the exception of maximum likelihood.

> Lindley and Phillips (1976): Many people's intuition says this specification is irrelevant. Their argument might more formally be expressed by saying that the evidence is of 12 honestly reported tosses, 9 of which were U; 3, D. Furthermore, these were in a particular order, that reported above. Of what relevance are things that might have happened [e.g. no lunch], but did not?

Indeed, this helps us understand **The LIKELIHOOD PRINCIPLE**:

1. All the information about θ obtainable from an experiment is contained in the likelihood function for θ given the data. There is no informatin beyond that contained in the likelihood and the data.

2. If two likelihood functions for θ (from the same or different experiments) are proportional to one another, then they contain the same information about θ.

By using only the likelihood, and nothing else from the experiment, the answer to the problem is the same regardless of the stopping rule. Indeed, let $x_1 = \#U$ in experiment 1 $x_2 = \#U$ in experiment 2. In experiment 1 (E1) we have a binomial density:

$$f_\theta^1(x_1) = \binom{12}{x_1}\theta^{x_1}(1-\theta)^{12-x_1} \Longrightarrow \ell_\theta^1(9) = \binom{12}{9}\theta^9(1-\theta)^3$$

In experiment 2 (E2) we have a negative binomial density:

$$f_\theta^2(x_2) = \binom{x_2+3-1}{x_2}\theta^{x_2}(1-\theta)^3 \Longrightarrow \ell_\theta^2(9) = \binom{11}{9}\theta^9(1-\theta)^3$$

In this situation, the Likelihood Principle says that:

- 1. For experiment E1 alone the information about θ is contained solely in $\ell_\theta^1(9)$;
  2. For experiment E2 alone the information about θ is contained solely in $\ell_\theta^2(9)$;
  3. Since $\ell_\theta^1(9)$ and $\ell_\theta^2(9)$ are proportional as functions of θ, the information about θ in the two experiments is identical.

Error variance

- The distribution of $\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2$ is a chi-square:

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 = \frac{\varepsilon'}{\sigma}(I - P_X)\frac{\varepsilon}{\sigma} \sim \chi^2_{T-k}$$

- Write down $\hat{\varepsilon}'\hat{\varepsilon}$ in terms of the bias adjusted estimator $\tilde{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(T-k)$

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 = (T-k)\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{T-k}$$

- In a chi-square we have $E[\chi^2_v] = v$ and $Var[\chi^2_v] = 2v$. From the moments of $\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2$ we can infer the moments of $\tilde{\sigma}^2$:

$$\begin{aligned} E[(T-k)\tilde{\sigma}^2/\sigma^2] &= T-k \Longrightarrow E[\tilde{\sigma}^2] = \sigma^2 \\ Var[(T-k)\tilde{\sigma}^2/\sigma^2] &= 2(T-k) \Longrightarrow Var[\tilde{\sigma}^2] = 2\sigma^4/(T-k) \end{aligned}$$

- Note that a rescaled chi-square is a Gamma:

$$s^2\tilde{\sigma}^2 \sim \chi^2_v \Longleftrightarrow \tilde{\sigma}^2 \sim \Gamma\left(\frac{v}{2}, \frac{s^2}{2}\right)$$

- The distribution of $\tilde{\sigma}^2$ is Gamma with $v = (T-k)$ and $s^2 = (T-k)/\sigma^2$. Specifically:

$$(T-k)\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{T-k} \Longleftrightarrow \tilde{\sigma}^2 \sim \Gamma\left(\frac{T-k}{2}, \frac{(T-k)/\sigma^2}{2}\right). \tag{41}$$

- It has mean

$$\frac{T-k}{2} \Big/ \frac{(T-k)/\sigma^2}{2} = \sigma^2$$

and variance

$$\frac{T-k}{2} \Big/ \left(\frac{(T-k)/\sigma^2}{2}\right)^2 = 2\sigma^4/(T-k)$$

- Now imagine that σ^2 was a random variable (not a coefficient) and consider (41) as a function of σ^2. The distribution of $\frac{1}{\sigma^2}$ is a Gamma:

$$(T-k)\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{T-k} \Longleftrightarrow \frac{1}{\sigma^2} \sim \Gamma\left(\frac{T-k}{2}, \frac{(T-k)\tilde{\sigma}^2}{2}\right)$$

and that of σ^2 is an inverse Gamma:

$$(T-k)\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{T-k} \Longleftrightarrow \sigma^2 \sim \Gamma^{-1}\left(\frac{T-k}{2}, \frac{(T-k)\tilde{\sigma}^2}{2}\right)$$

where $v = (T-k)$ and $s^2 = (T-k)\tilde{\sigma}^2$.

- Note that $(T-k)\tilde{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}$, so we can write

$$\sigma^2 \sim \Gamma^{-1}\left(\frac{T-k}{2}, \frac{\hat{\varepsilon}'\hat{\varepsilon}}{2}\right)$$

- All this suggests to use an inverse Gamma as a prior for σ^2.

Conditional Posterior of error variance 1. Set prior distribution $\sigma^2 \sim \Gamma^{-1}(v_0/2, s_0^2/2)$

$$p(\sigma^2) = [\Gamma(\nu_0/2)]^{-1}(s_0^2/2)^{\frac{\nu_0}{2}}(\sigma^2)^{-\frac{\nu_0+2}{2}}\exp\left(-s_0^2/2\sigma^2\right)$$

2. Obtain data and form the likelihood function

$$p(Y|\beta,\sigma^2) = (2\pi)^{-\frac{T}{2}}\left|\sigma^2 I_T\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(Y-X\beta)'\left(\sigma^2 I_T\right)^{-1}(Y-X\beta)\right]$$

3. Obtain the conditional posterior kernel

$$p\left(\sigma^2|Y,\beta\right) \propto (\sigma^2)^{-\frac{T+\nu_0+2}{2}}\exp\left[-\{s_0^2 + (Y-X\beta)'(Y-X\beta)\}/2\sigma^2\right]$$

which is the kernel of an inverse gamma

$$\Gamma^{-1}(v_1/2, s_1^2/2)$$

with

$$v_1 = T + v_0,\ s_1^2 = s_0^2 + (Y-X\beta)'(Y-X\beta).$$

13.2 The CLRM with independent N-IG prior

- In summary, when the prior for β and σ^2 is:

$$\beta \sim N(\beta_0, \Sigma_0);\ \sigma^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$$

we derived the conditional posteriors:

$$\beta|\sigma^2, Y \sim N\left(\beta_1, \Sigma_1\right);\ \sigma^2|\beta, Y \sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right)$$

with moments:

$$\begin{aligned}
\Sigma_1 &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1} \\
\beta_1 &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'X\hat{\beta}\right) \\
\nu_1 &= \nu_0 + T \\
s_1^2 &= s_0^2 + (y - X\beta)'(y - X\beta)
\end{aligned}$$

- Simple Monte Carlo simulation is not an option to obtain the joint $\beta, \sigma^2|Y$. Other methods are needed$\longrightarrow$ Gibbs Sampling (MCMC)
- Integrating out σ^2 analytically to obtain $\beta|Y$ is not an option. Computing the marginal likelihood in closed form is not feasible.
- However, the model is more flexible than the one with conjugate prior for it does not require proportionality between the prior on β and the error variance.

13.3 Gibbs sampling

Set starting values $\theta_1^{j=0}, ..., \theta_k^{j=0}$.

1. Sample $\theta_i^{j=1}$ from
$$f\left(\theta_i^1|\theta_1^1, \theta_2^1, ..., \theta_{i-1}^1, \theta_{i+1}^0, ..., \theta_k^0\right)$$
2. Set $j = j + 1$ and repeat (1) until $j = J$:
$$f\left(\theta_i^j|\theta_1^j, \theta_2^j, ..., \theta_{i-1}^j, \theta_{i+1}^{j-1}, ..., \theta_k^{j-1}\right).$$

- Gibbs sampling is a special case of more general MCMC sampling
- As $J \to \infty$ the joint and marginal distributions of simulated $\{\theta_1^j, ...\theta_K^j\}_{j=1}^m$ converge at an exponential rate to the joint and marginal distributions of $\theta_1....\theta_k$
- For simple models (e.g. linear regressions, also multivariate), this happens *really fast*

- By construction the Gibbs sampler produces draws than are autocorrelated
  - Some burn-in required
  - What is the efficiency/mixing?
- **Convergence**
  - Time series plots
  - Tests of equality across independent chains e.g. Geweke's (1992) convergence diagnostic test for equal means
  - Potential Scale Reduction Factors (PSRF), Gelman and Rubin (1992)
- **Mixing**
  - Time series plots
  - Autocorrelograms
  - Inefficiency factors (IF) give an idea of how far we are from i.i.d. sampling
- Check out the R / MATLAB packages COnvergence DiAgnostics
- Gelman and Rubin (1992) and Brooks, S.P. and Gelman, A. (1998)
- Let $\{\theta_{mj}\}_{j=1}^{J}$ be the $m-$th simulated chain, $m = 1, \ldots, M$. Let $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the $m-$th chain, and let the overall sample posterior mean be $\hat{\theta} = \Sigma_{m+1}^{M} \hat{\theta}_m / M$.
- There are two ways to estimate the variance of the stationary distribution σ^2:
  - The mean of the empirical variance within each chain:

$$W = \frac{1}{M} \Sigma_{m+1}^{M} \hat{\sigma}_m^2$$

  - The empirical variance from all chains combined:

$$V = \frac{J-1}{J} W + \frac{M+1}{MJ} B,$$

    where $B = \frac{J}{M-1} \Sigma_{m+1}^{M} (\hat{\theta}_m - \hat{\theta})^2$ is the empirical between-chain variance

- If the chains have converged, then both W and V are unbiased. Otherwise the first method will underestimate the variance, since the individual chains have not had time to range all over the stationary distribution, and the second method will overestimate the variance, since the starting points were chosen to be overdispersed.

- The convergence diagnostic is:

$$PSRF = \sqrt{\frac{V}{W}}$$

- Brooks and Gelman (1997) have suggested, if $PSRF < 1.2$ for all model parameters, one can be fairly confident that convergence has been reached.

- More reassuring (and common) is to apply the more stringent condition $PSRF <$ 1.1ors

- The inefficiency factor (IF) $1 + 2\sum_{k=1}^{\infty} \rho_k$, where ρ_k is the k-th order autocorrelation. This is the inverse of the relative numerical efficiency measure of Geweke (1992). Usually estimated as the spectral density at frequency zero with Newey-West kernel (with a 4% bandwidth).

- i.i.d. sampling (e.g. MC sampling) features IF=1, here IF < 20 are considered good.

- Note that mixing can *always* be improved artificially by a practice called thinning. Thinning (or skip sampling) is only advisable if you have space constraints, since it always implies loss of information

13.4 Generalized linear regression model

- Gibbs sampling is powerful. For example, we can easily extend the model we are considering:

$$y_t = \beta x_t + \varepsilon_t \tag{42}$$
$$\varepsilon_t = \phi\varepsilon_{t-1} + u_t,\ u_t \sim iidN(0, \sigma_u^2) \tag{43}$$

  In this model there are 3 (groups of) parameters: β, ϕ, σ_u^2. Also, we have $\sigma_\varepsilon^2 = Var(\varepsilon_t) = \sigma_u^2/(1-\phi^2)$

- Consider the Cochrane-Orcutt transformation:

$$P = \begin{bmatrix} -\phi & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & & -\phi \quad 1 \end{bmatrix}$$

$$Py = PX\beta + P\varepsilon \tag{44}$$

where $[P\varepsilon]_t = -\phi\varepsilon_{t-1} + \varepsilon_t$.

- The model in (44) is a Generalized LRM, with error variance $Var(P\varepsilon) = PVar(\varepsilon)P' = \sigma_\varepsilon^2 PP' = \frac{\sigma_u^2}{1-\phi^2}PP' = \Omega(\phi, \sigma_u^2)$. We have:

$$P(\phi)y = P(\phi)X\beta + P(\phi)\varepsilon \tag{45}$$

which has likelihood:

$$p(Y|\beta, \sigma^2, \phi) = (2\pi)^{-\frac{T}{2}} |\Omega|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y - X\beta)'(\Omega)^{-1}(Y - X\beta)\right]$$

where recall $\Omega = \Omega(\phi, \sigma_u^2)$

- One can specify:

$$\beta \sim N(\beta_0, \Sigma_0);\ \phi \sim N(\phi_0, \Sigma_{\phi_0});\ \sigma_u^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$$

- Under knowledge of σ_u^2 and ϕ, this gives the following posterior for β :

$$\beta|\phi, \sigma^2, y \sim N(\beta_1, \Sigma_1)$$

$$\Sigma_1 = \left(\Sigma_0^{-1} + X'\Omega(\phi, \sigma_u^2)^{-1}X\right)^{-1},\ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + X'\Omega(\phi, \sigma_u^2)^{-1}Y\right)$$

which is simply the average of a GLS estimator and the prior.

- Then, under knowledge of β, it is easy to use the model in (42) to derive $\varepsilon|\beta, y$ and this can be used as an observable in (43). This gives

$$\varepsilon_1 = \varepsilon_0\phi + u,\ u_t \sim N(0, \sigma_u^2 I_{T-1}) \tag{46}$$

which is a standard linear regression model with AR coefficient ϕ and error variance σ_u^2.

- Given the prior specified in (43), the posteriors will be:

$$\begin{aligned}
\phi|\beta,\sigma_u^2,y &\sim N(\phi_1,\Sigma_{\phi_1});\\
\Sigma_{\phi_1} &= \left(\Sigma_{\phi_0}^{-1}+\frac{1}{\sigma_u^2}\varepsilon_0'\varepsilon_0\right)^{-1},\ \beta_1=\Sigma_{\phi_1}\left(\Sigma_{\phi_0}^{-1}\phi_0+\frac{1}{\sigma_u^2}\varepsilon_0'\varepsilon_1\right)\\
\sigma_u^2|\beta,\phi,y &\sim \Gamma^{-1}\left(\frac{\nu_0+T}{2},\frac{s_0^2+(\varepsilon_1-\varepsilon_0\phi)'(\varepsilon_1-\varepsilon_0\phi)}{2}\right)
\end{aligned}$$

14 Appendix A: Moments of the error variance in the LRM

Start with:

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 = (T-k)\hat{\sigma}^2/\sigma^2 \sim \chi^2$$

where

$$\begin{aligned} E[(T-k)\hat{\sigma}^2/\sigma^2] &= (T-k) \\ Var[(T-k)\hat{\sigma}^2/\sigma^2] &= 2(T-k) \end{aligned}$$

$$\begin{aligned} E[(T-k)\hat{\sigma}^2/\sigma^2] &= (T-k) \Longrightarrow (T-k)E[\hat{\sigma}^2]/\sigma^2 = (T-k) \Longrightarrow E[\hat{\sigma}^2] = \sigma^2 \\ Var[(T-k)\hat{\sigma}^2/\sigma^2] &= 2(T-k) \Longrightarrow (T-k)^2 Var[\hat{\sigma}^2]/\sigma^4 = 2(T-k) \Longrightarrow Var[\hat{\sigma}^2] = 2\sigma^4/(T-k). \end{aligned}$$

Alternative derivation which also shows the distribution:

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 = (T-k)\hat{\sigma}^2/\sigma^2 = s_0\hat{\sigma}^2 \sim \chi^2_{T-k}$$

where $s_0^2 = (T-k)/\sigma^2$. Note that if

$$s_0\hat{\sigma}^2 \sim \chi^2_{T-k}$$

then $\hat{\sigma}^2$ is a gamma:

$$\hat{\sigma}^2 \sim \Gamma\left(\frac{T-k}{2}, \frac{s_0^2}{2}\right)$$

with mean

$$\frac{T-k}{2}/\frac{s_0^2}{2} = (T-k)/s_0^2 = (T-k)/((T-k)/\sigma^2) = \sigma^2$$

and variance

$$\frac{T-k}{2}/\left(\frac{s_0^2}{2}\right)^2 = 2(T-k)/s_0^4 = 2(T-k)/((T-k)^2/\sigma^4) = 2\sigma^4/(T-k)$$

Towards a conjugate prior Finally note that

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 = (T-k)\hat{\sigma}^2/\sigma^2 = s_0\hat{\sigma}^2 \sim \chi^2_{T-k}$$

implies that

$$\frac{1}{\sigma^2} \sim \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}}\chi^2_{T-k}$$

that is, σ^2 is the inverse of a rescaled chi-square distribution, i.e. it is an inverse gamma. This offers a choice for a conjugate prior for σ