1. OLS in finite samples

January 13, 2022

Best linear unbiased estimator is found by deriving the linear unbiased estimator with the lowest variance.

1 Introduction

We use the OLS to estimate the parameters in the multiple linear regression model - it is written as follows:

$$\boldsymbol{Y} = \underset{n\times k}{\boldsymbol{X}}' \times \underset{k\times 1}{\beta} + \epsilon$$

We find an estimator for β by minimizing the sum of squared residuals (SSR). We denote the residuals as $\boldsymbol{e} = \boldsymbol{Y} - \mathbf{X}\boldsymbol{b}$ where $\boldsymbol{b}$ is a candidate for β:

$$\begin{aligned} S(\boldsymbol{b}) &= (\boldsymbol{Y} - \mathbf{X}\boldsymbol{b})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{b}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\mathbf{X}\boldsymbol{b} - \boldsymbol{b}'\mathbf{X}'\boldsymbol{Y} + \boldsymbol{b}'\mathbf{X}'\mathbf{X}\boldsymbol{b} \end{aligned}$$

We want to minimize with respect to $\boldsymbol{b}$, so we use the first order conditions:

$$\frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{b}} = -2\mathbf{X}'\boldsymbol{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{b} = \mathbf{0}$$

We can then rewrite the first order conditions into the normal equations:

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

Where $\hat{\beta}$ is the vector that solves them. We can then solve for $\hat{\beta}$ and obtain the Ordinary Least Squares (OLS) estimator:

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

2 Assumptions

Considering the classical Gauss-Markov (GM) assumptions, where $(A2 - GM)$ says that $\{\epsilon_1, ..., \epsilon_n\}$ and $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ are fully statistically independent. This is a very strong assumption, which is unlikely to hold outside of controlled experiments. Therefore we replace the classical GM assumptions with our Least Squares Assumptions:

$(A1)$: rank $(\mathbf{X}) = k \rightarrow$ no multicollinearity implying k≤n and **X'X** is positive definite and thereby invertible

$(A2)$: $E[\epsilon|\mathbf{X}] = 0 \rightarrow$ Using LIE this implies mean independece.

$(A3)$: $Var[\epsilon_i|\mathbf{X}] = \sigma^2,$ for all $i = 1, ..., n \rightarrow$ Implying homoskedasticity

$(A4)$: $Cov[\epsilon_i, \epsilon_j|\mathbf{X}] = 0,$ for all $i = 1, ..., n,\ i \neq j \rightarrow$ This assumption rules out autocorrelation

We make additional assumptions:

$$
\begin{aligned}
&(A5) : \epsilon|\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \text{ implying that } \hat{\beta}|\mathbf{X} \sim N\left(\beta, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\\
&(A6) : \text{plim}\left(\frac{1}{2}\mathbf{X}'\mathbf{X}\right) = E\left[\mathbf{X}_i\mathbf{X}_i'\right] \equiv \mathbf{Q}\\
&(A7) : \text{The sequence } \left\{\boldsymbol{X}_i\epsilon_i\right\} \text{ for } i = 1, ..., n \text{ is } i.i.d \text{ with } E\left[\boldsymbol{X}_i\epsilon_i\right] = \mathbf{0}
\end{aligned}
$$

Where assumption $A7$ replaces assumption $A2$.

3 Properties

Under assumptions $A1 - A5$ the OLS estimator is BLUE, which means it is the best linear unbiased estimator. This means it is the estimator with the lowest mean squared error and the smallest variance.

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

$$\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underbrace{\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}]}_{=0, A2} = \boldsymbol{\beta}$$

Under our assumptions, the OLS estimator will be BLUE: Best Linear Unbiased Estimator
The covariance matrix of $\hat{\beta}$ is

$$Var[\hat{\beta}|\mathbf{X}] = E[\left(\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|x\right)$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \underbrace{E[\varepsilon\varepsilon'|\mathbf{X}]}_{=V[\varepsilon|\mathbf{X}]=\sigma^2, A3} \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$= \sigma^2\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'}\right)^{-1}$$

An unbiased estimator for the variance $\sigma^2 = Var\left(\varepsilon_i\right)$ is: $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon}$

4 Hypothesis testing

Testing the null hypothesis that $\beta_j = \beta_j^0 : t = \frac{\hat{\beta}_j - \beta_j^0}{SE\left(\hat{\beta}_j\right)} \sim t\left(n-k\right)$

Testing g restrictions: $F = \frac{\left(\hat{\varepsilon'}_R\hat{\varepsilon}_R - \hat{\varepsilon}'\hat{\varepsilon}\right)/g}{\hat{\varepsilon}'\hat{\varepsilon}/(n-k)} \sim F\left(g, n-k\right)$
Further, $\hat{\varepsilon'}_R\hat{\varepsilon}_R$ is the SSR from the restricted model, and $\hat{\varepsilon'}\hat{\varepsilon}$ is the SSR under H_1 (unrestricted model)

4.1 Linear restrictions

We look at the linear model:

$$\underset{n\times1}{Y} = \underset{n\times k}{\mathbf{X}}\underset{k\times1}{\beta} + \underset{n\times1}{\varepsilon}$$

We want to test g linear restrictions ($g < k$):

$$H_0 : \mathbf{R}\beta = \boldsymbol{q}$$
$$H_1 : \mathbf{R}\beta \neq \boldsymbol{q}$$

where $\mathbf{q}$ is a $[g \times 1]$-vector and $\mathbf{R}$ is a $[g \times k]$-matrix, and with $\operatorname{rank}(\mathbf{R}) = g$
Under H_0 is called the restricted model, since we impose restrictions and the model under H_1 is the unrestricted model.
There are different ways to test this, including $F-$ test, Wald test and the $t-$ test . The underlying idea of $t-$, $F-$ and Wald test is to estimate the unrestricted model and check whether the quantity $\mathbf{R}\hat{\beta} - \boldsymbol{q}$ is close to zero.

4.2 Testing OLS assumptions

4.2.1 Tests for heteroskedasticity

We can use the White test to test for heteroskedasticity. Heteroskedasticity is when variance of ϵ_i depends on i or the value of x_i. Assumption $A3$ and $A4$ says $Var\,(\epsilon|\mathbf{X}) = \sigma^2\mathbf{I}_n$ which implies homoskedasticity. Therefore we want to test whether these assumptions are violated.
Steps of the White test for heteroskedasticity:

1. Estimate the model under H_0: Homoskedasticity
2. Regress the squared residuals from step 1 on all the explanatory variables, along with their squares and cross-products.
3. Calculate an $F-$ or $LM-$test for joint significance of the explanatory variables of step 2, using the R^2 from step 2.

If the null of homoskedasticity is rejected and we know the form of the heteroskedasticity, then we can use the Generalized Least Squares (GLS) - however, this is typically not the case. When we do not know the form of the heteroskedasticity, we can use the White Standard Errors. The White Standard Errors are given as follows:

$$\widehat{Var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{n} \hat{\sigma}_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'\right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

Where $\hat{\sigma}_i^2$ is an estimate of the unknown σ_i^2.
When estimating the White Standard Errors there are several choices for the estimate of $\hat{\sigma}_i^2$:

1. We can set $\hat{\sigma}_i^2 = \hat{\varepsilon}_i^2$. However this tends to underestimate the standard errors.
2. We can also set $\hat{\sigma}_i^2 = (n/(n-k))\,\hat{\varepsilon}_i^2$, which is a better alternative than above.
3. Another alternative is to set $\hat{\sigma}_i^2 = \hat{\varepsilon}_i^2/(1-h_i)$, where h_i is the i'th diagonal element of $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.
4. The last option is to se $\hat{\sigma}_i^2 = \hat{\varepsilon}_i^2/(1-h_i)^2$, which is a reliable choice and preferred in practice.

The Breuch-Pagan test is similar to the White test but without squares and cross-products in Step 2.

4.2.2 Tests for autocorrelation

If we want to test for autocorrelation we can use the Durbin-Watson test for first-order autocorrelation, and the Breush-Godfrey test for autocorrelation up to lag p. Autocorrelation means that $Cov\,(u_i, u_j) \neq 0$ for $i \neq j$, which is a violation of our assumption $A4$. If there is autocorrelation we can use the Newey-West standard errors, which are both robust against heteroskedasticity and autocorrelation (HAC)

4.2.3 Test for normality

We can also make a test on assumption $A5$. Here we can use the Jarque-Bera test for normality. We find the skewness and kurtosis for our test and sees if they are statistically close enough to 0 and 3.

2. OLS in large samples

January 13, 2022

In this subject we turn to asymptotic properties which is an approximation to the finite samples

1 Introduction

We use the OLS to estimate the parameters in the multiple linear regression model - it is written as follows:

$$Y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i, \quad i = 1, 2, ..., n$$

This can be written in matrix notation:

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}} \times \underset{k\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

We find an estimator for β by minimizing the sum of squared residuals (SSR). We denote the residuals as $\boldsymbol{e} = \boldsymbol{Y} - \mathbf{X}\boldsymbol{b}$ where $\boldsymbol{b}$ is a candidate for β:

$$\begin{aligned} S(\boldsymbol{b}) &= (\boldsymbol{Y} - \mathbf{X}\boldsymbol{b})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{b}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\mathbf{X}\boldsymbol{b} - \boldsymbol{b}'\mathbf{X}'\boldsymbol{Y} + \boldsymbol{b}'\mathbf{X}'\mathbf{X}\boldsymbol{b} \end{aligned}$$

We want to minimize with respect to $\boldsymbol{b}$, so we take the first order conditions:

$$\frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{b}} = -2\mathbf{X}'\boldsymbol{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{b} = \mathbf{0}$$

We can then rewrite the first order conditions into the normal equations:

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

Where $\hat{\beta}$ is the vector that solves them. We can then solve for $\hat{\beta}$ and obtain the Ordinary Least Squares (OLS) estimator:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{Y}$$

2 Assumptions

We have our Least Squares Assumptions that hold for finite samples:

$(A1)$: rank $(\mathbf{X}) = k \rightarrow$ no multicollinearity implying k≤n and **X'X** is positive definite and thereby invertible

$(A2)$: $\mathbf{E}\left[\epsilon|\mathbf{X}\right] = \mathbf{0} \rightarrow$ Using LIE this implies A1-GM

$(A3)$: $Var\left[\epsilon_i|\mathbf{X}\right] = \sigma^2,$ for all $i = 1, ..., n \rightarrow$ Implying homoskedasticity →using LIE this implies A3-GM

$(A4)$: $Cov\left[\epsilon_i, \epsilon_j|\mathbf{X}\right] = 0,$ for all $i = 1, ..., n,\ i \neq j$ (rules out autocorrelation)

We make additional assumptions:

$$\begin{aligned} &(A5) : \epsilon|\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \text{ implying that } \hat{\beta}|\mathbf{X} \sim N\left(\beta, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \\ &(A6) : \text{plim}\left(\frac{1}{2}\mathbf{X}'\mathbf{X}\right) = E\left[\mathbf{X}_i\mathbf{X}_i'\right] \equiv \mathbf{Q} \\ &(A7) : \text{The sequence } \left\{\boldsymbol{X}_i\epsilon_i\right\} \text{ for } i = 1, ..., n \text{ is } i.i.d \text{ with } E\left[\boldsymbol{X}_i\epsilon_i\right] = \mathbf{0} \end{aligned}$$

We make additional assumptions, where $A7$ replace $A2$. Because models often become too complex to derive their exact statical properties like in finite samples, we use the asymptotic properties that are only an approximation to the finite sample, but it becomes better as the sample size grow.

3 Properties

3.1 Unbiasedness

Under assumptions $A1 - A5$ the OLS estimator is BLUE, which means it is the best linear unbiased estimator. This means it is the estimator with the lowest mean squared error and the smallest variance.

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

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

It is worth noting that unbiasedness is not very informative since we normally only have a single sample at hand so consistency is more informative.

3.2 Consistency

Consistency means that the estimated value approaches the actual, true parameter value as the sample size increases. We can rewrite the OLS estimator as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i \\ &= \boldsymbol{\beta} + \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'}_{\boldsymbol{Q}}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i\end{aligned}$$

By using the law of large numbers along with assumption ***A6*** and **A7**, we can get:

$$plim(\hat{\beta}) = \beta + \mathbf{Q}^{-1}\underbrace{\mathbb{E}\left[\boldsymbol{X}_i\varepsilon_i\right]}_{=0\ A7} = \beta$$

This establishes that the OLS estimator is consistent.

3.3 A7

A7 states the error term in observation i is uncorrelated with all of the explanatory variables in observation i, for all i.
Using LIE, we can show that

$$E[\mathbf{X_i}\varepsilon_i] = E[E[\mathbf{X_i}\varepsilon_i|\mathbf{X}]] = E[\mathbf{X_i}\underbrace{E[\varepsilon_i|\mathbf{X}]}_{=0\ A2}] = 0$$

When A7 is violated, the OLS estimator is neither unbiased nor consistent, which means there is endogeneity. This can be solved by IV.

3.4 A5, the distribution

In the finite sample the approximate distribution is: $\hat{\beta} \sim N\left(\beta, \sigma^2 \left(\sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i'}\right)^{-1}\right)$.
Using the expression for $\hat{\beta}$, we can rewrite the equation as

$$\hat{\beta} - \beta = \left(\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i\right)$$

and scale it by $\sqrt{n}$ to circumvent the issue of degeneracy.

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i\right)$$

The last part is mean zero by A7 and LIE gives us the variance

$$E[\mathbf{X_i}\varepsilon_\mathbf{i}^\mathbf{2}\mathbf{X_i'}] = E[E[\mathbf{X_i}\varepsilon_i^2\mathbf{X_i'}|\mathbf{X}]] = \underbrace{\sigma^2}_{A3:\ E[\varepsilon^2|\mathbf{X}]=\sigma^2} E[\mathbf{X_i X_i'}]$$

Because the sequences of $X_i\epsilon_i$ are an iid sequence and the first two moments are finite, we can say, that the final term is normally distributed with mean 0 and the variance derived

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X_i}\varepsilon_i \xrightarrow{d} N\left(0, \sigma^2\boldsymbol{Q}\right)$$

We know that we can multiply the distributions mean by Q^{-1} and we can pre and post multiply by Q^{-1} to get the variance of the entire distribution. This implies the following asymptotic distribution of the OLS estimator:

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, \sigma^2\mathbf{Q^{-1}QQ^{-1}}\right) = N\left(0, \sigma^2\mathbf{Q^{-1}}\right)$$

In a finite sample this suggest the approximate distribution $\hat{\beta} \overset{a}{\sim} N\left(\beta, \sigma^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i X_i'}\right)^{-1}\right)$ where the unknown matrix $\mathbf{Q}$ is approximated by $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i X_i'}$.

4 Maximum Likelihood

4.1 Introduction

Some models will not be appropriate to estimate by OLS, e.g. moving-average models. So we need a more general estimation framework which is where maximum likelihood comes into play.
The idea of ML is:

1. The dataset was generated by some known model
2. The form and distribution of the model is known, but the parameter values are not and must be estimated.

The ML principle is to select the estimate of the parameters of the model as the value that render the actual observed data most likely to occur.
The likelihood function is the density, written as a function of the unknown parameters:

$$\begin{aligned} L(\theta; Y) &= L\left(\theta; Y_1, \ldots, Y_n\right) \\ &\equiv p\left(Y_1, \ldots, Y_n; \theta\right) \\ &= \prod_{i=2}^{n} p\left(Y_i \mid Y_{i-1}, \ldots, Y_1; \theta\right) p\left(Y_1; \theta\right) \end{aligned}$$

If the observations are independent (and thus random sample), it becomes:

$$L(\boldsymbol{\theta}; \boldsymbol{Y}) = \prod_{i=1}^{n} p\left(Y_i; \boldsymbol{\theta}\right)$$

It is simpler to work with the log-likelihood function:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{Y}) &= \log L(\boldsymbol{\theta}; \mathbf{Y}) \\ &= \sum_{i=1}^{n} \log p\left(Y_i; \boldsymbol{\theta}\right) \end{aligned}$$

Where the last equality follows only under independence.
We can do this since the logarithm is a monotonically increasing function, the argument that maximizes the log-likelihood function is the same as the one that maximizes the likelihood function.
The maximum likelihood estimator $\hat{\theta}_{ML}$ is the value that maximizes the likelihood function and it is the value of the parameters that makes the observed sample most likely.
When a unique MLE $\hat{\theta}_{ML}$ exist it can be found by solving the likelihood equations:

$$\left.\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}}\right|_{\hat{\theta}_{ML}} = \mathbf{0}$$

It is consistent, asymptotic normality, asymptotic efficient and achieves Cramér-Rao lower bound for consistent estimators. Optimal asymptotically but not in the finite sample.
Looking at the Gaussian linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, $\varepsilon|\mathbf{X} \sim N\left(0, \sigma^2 \mathbf{I}_n\right)$ and it follows that $\mathbf{Y}|\mathbf{X} \sim N\left(\mathbf{X}\beta, \sigma^2 \mathbf{I_n}\right)$ which is the multivariate normal distribution. The log-likelihood function is then:

$$l\left(\beta, \sigma^2; \mathbf{Y}|\mathbf{X}\right) = \ln p\left(\mathbf{Y}|\mathbf{X}; \beta, \sigma^{\mathbf{2}}\right) = -\frac{n}{2}\ln\left(2\pi\right) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\beta\right)'\left(\mathbf{Y} - \mathbf{X}\beta\right)$$

The last term on the right is the SSR ($S\left(\mathbf{b}\right)$), hence its derivative is $\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{-2X'Y + 2X'Xb}$
FOC:

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}\left(\mathbf{X'Y} - \mathbf{X'X}\beta\right) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\mathbf{Y} - \mathbf{X}\beta\right)'\left(\mathbf{Y} - \mathbf{X}\beta\right) = 0$$

MLE:

$$\hat{\beta}_{ML} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'Y}$$

$$\hat{\sigma^2}_{ML} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}_{ML}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\beta}_{ML}\right)$$

- MLE of β is the same as in OLS
- MLE of σ^2 is equal to $\frac{n-k}{n}$ times the unbiased OLS estimator. So $\hat{\sigma^2}_{ML}$ is biased but consistent.

4.2 Properties of ML

The maximum likelihood estimator has following asymptotic properties
Consistency:

$$\text{plim}\left(\hat{\theta}_{ML}\right) = \theta$$

Asymptotic Normality:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$$

Where $\mathbf{V}$ is the asymptotically covariance matrix.
$\hat{\theta}_{ML}$ is asymptotically efficient and achieves the Cramér-Rao lower bound for consistent estimators. This means that it has the lowest possible variance for any unbiased estimator.
The MLE is optimal asymptotically, but it is does not have such properties for finite samples.

4.3 Hypothesis testing

Likelihood Ratio examine wether a reduced model provides the fit as a full model. So the idea behind the Likelihood Ratio test is to test if the restrictions are valid. If they are valid, then imposing them should not lead to a large reduction in the log-likelihood function.
The test statistic is:

$$LR = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\hat{\boldsymbol{\theta}}_R\right)\right) \xrightarrow{d} \chi^2(g)$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood based on the unrestricted estimates that are valid under H_1 and $\ell\left(\hat{\theta}_R\right)$ is the log-likelihood based on the restricted estimates that are valid under the null.

3. The endogeneity problem

January 13, 2022

1 Introduction

When we consider the linear model

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}} \times \underset{k\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

We use the asymptotic properties of OLS estimator when the sample size goes toward infinity, since economic theory sometimes becomes to complex to derive the finite sample properties.
OLS is consistent if $A7$ holds, which is $E\left(\boldsymbol{x}_i\varepsilon_i\right) = 0$. This means that for each i, the error term has zero mean and is uncorrelated with any explanatory variables.
A7 is often violated in practice. This is often due to measurement errors, omitted variable biases and/or simultaneous equations. When A7 is violated, OLS is neither unbiased, nor consistent, which is **the endogeneity problem.**
When $A7$ **is violated**, we can estimate β with instrumental variables estimator and still get a consistent estimate. **However we only want to use this estimator when A7 is violated since the IV estimator is less efficient than OLS.**
To use this estimator we need instruments, which are variables not included in $\mathbf{X}_i$ that are uncorrelated with the error term but correlated with the explanatory variables that are responsible for the violation of A7.
The solution to the problem is **IV** (Instrumental variables).

- IV conditions:
  1. $m \geq k$ it is the **Order condition.** Exogenous variables serve as their own instruments (e.g. the constant)
  2. $plim\left(\frac{1}{n}\mathbf{Z}'\varepsilon\right) = E[\mathbf{Z_i}\varepsilon_i] = 0$. Instruments must be uncorrelated with the error term. **Validity**
  3. $plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) = E[\mathbf{Z_i X_i'}] \equiv \mathbf{Q_{ZX}}$, $rank\left(\mathbf{Q_{ZX}}\right) = k$. This is the rank condition. The instruments must be correlated with the explanatory variables. **Relevance.**
  4. $plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right) = E[\mathbf{Z_i Z_i'}] \equiv \mathbf{Q_{ZZ}}$, $rank(\mathbf{Q_{ZZ}}) = m$, no perfect collinearity between instruments → Q_{ZZ} is invertible. (Equivalent to A6)

  - Also: $(A3')\, Var[\varepsilon_i|\mathbf{X},\mathbf{Z}] = \sigma^2$ and $(A4')\, Cov\left(\varepsilon_i, \varepsilon_j|\mathbf{X},\mathbf{Z}\right) = 0$

- Summing up: When selecting instruments we need to make sure:
  1. Uncorrelation with the error term: $E\left(\mathbf{Z_i}\varepsilon_i\right) = 0 \rightarrow$ Valid instrument
  2. Correlation with the explanatory variable that cause the violation of $(A7)$: $Cov\left(Z_i X_i\right) \neq 0 \rightarrow$ Relevant instrument

- Consistent estimates of our model yields: $\hat{\beta}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{Y}$ in the **exactly identified** case, $k = m$

- In the **over-identified** case, $m > k$, we use GMM and obtain $\hat{\beta}_{IV} = \left(\mathbf{X}'\mathbf{P_Z}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{Y}$ using a weighting matrix. Since there are more equations than unknowns leading to not all sample moments being equal to zero.

2 Examples

2.1 Measurement error

If we want to estimate the following regression model

$$Y_i = \alpha + \beta X_i^* + \varepsilon_i, \quad i = 1, \dots, n$$

under the classical assumptions and that $\varepsilon_i \sim \text{i.i.d} \cdot \left(0, \sigma_\varepsilon^2\right)$.
We cannot see X_i^* directly. We observe X_i instead of X_i^* and that

$$X_i = X_i^* + u_i$$

Where $u_i \sim \text{i.i.d.} \left(0, \sigma_u^2\right)$ is the measurement error with $\mathbb{E}\left[X_i \varepsilon_i\right] = 0$.
So we need to estimate β from:

$$Y_i = \alpha + \beta X_i + \nu_i, \quad \nu_i = \varepsilon_i - \beta u_i$$

Then we get a different OLS estimator

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$$

$$= \beta + \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(\left(\varepsilon_i - \overline{\varepsilon}\right) - \beta\left(u_i - \overline{u}\right)\right)}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$$

Taking the plim yields:

$$plim\left(\hat{\beta}_{\text{OLS}}\right) = \beta + \frac{Cov\left[X_i, \left(\varepsilon_i - \beta u_i\right)\right]}{Var\left[X_i\right]}$$

$$= \beta - \beta\frac{Cov\left[X_i, u_i\right]}{Var\left[X_i\right]}$$

We can now make one of two mutually exclusive assumptions regarding the measurement error:

$$\mathbb{E}\left[X_i u_i\right] = 0, \left(X_i^* = X_i - u_i\right) \quad (1)$$
$$\mathbb{E}\left[X_i^* u_i\right] = 0, \left(X_i = X_i^* + u_i\right) \quad (2)$$

1. If the **first assumption holds** the OLS estimator is **still consistent but have a larger error variance.**
   (a) It will be consistent because $Cov\left(X_i, u_i\right) = 0$
2. But if the s**econd hold** then the OLS estimator $\hat{\beta}$ **will generally be inconsistent**, which leads to $\hat{\alpha}$ **to also be inconsistent.** We can see this from the probability limit of the OLS

$$\begin{aligned} plim\left(\hat{\beta}_{\text{OLS}}\right) &= \beta - \beta\frac{Cov\left[X_i, u_i\right]}{Var\left[X_i\right]} \\ &= \beta - \beta\frac{\sigma_u^2}{Var\left[X_i\right]} = \beta\left(1 - \frac{\sigma_u^2}{Var\left[X_i\right]}\right) \end{aligned}$$

   It is only consistent if X_i^* is measured without error, meaning that $\sigma_u^2 = 0$.

Solution to the measurement error (type 2)

- Can be solved if there exists an instrument which fulfills the following: $Cov\left(z_i, x_i\right) = \sigma_{z,x} \neq 0$ **(relevant)** and $E\left(z_i u_i\right) = 0 = E\left(z_i \varepsilon_i\right)$ **(valid).**
- The IV-estimator is:

$$\hat{\beta}_{IV} = \frac{\frac{1}{n}\sum_{i=1}^n \left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^n \left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}$$

- It is consistent since

$$plim\left(\hat{\beta}_{IV}\right) = \beta + \frac{Cov\left(Z_i, (\varepsilon_i - \beta u_i)\right)}{Cov\left(Z_i, X_i\right)}$$

$$= \beta$$

2.2 Omitted variables

The observations on Y_i are generated from the model:

$$Y_i = \alpha + \beta X_{i1} + \gamma X_{i2} + \varepsilon_i, \quad i = 1, \dots, n$$

Satisfying the usual assumptions, and in particular

$$\mathbb{E}\left[X_{i1}\varepsilon_i\right] = \mathbb{E}\left[X_{i2}\varepsilon_i\right] = 0$$

We are interested only in β so we estimate the following model with OLS by omitting X_{i2}:

$$Y_i = \alpha + \beta X_{i1} + \nu_i, \quad \nu_i = \gamma_i X_{i2} + \varepsilon_i$$

We omit X_{i2} if we do not know the true model or X_{i2} is not directly measurable.
The OLS estimator for β:

$$\hat{\beta}_{\text{OLS}} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i1} - \bar{X}_1\right)\left(\left(\varepsilon_i - \bar{\varepsilon}\right) + \gamma\left(X_{i2} - \bar{X}_2\right)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i1} - \bar{X}_1\right)^2}$$

Taking the probability limits on the OLS estimator for β:

$$plim\left(\hat{\beta}_{\text{OL}}\right) = \beta + \gamma\frac{Cov\left[X_{i1}, X_{i2}\right]}{Var\left[X_{i1}\right]}$$

Therefore OLS is only consistent if X_{i1} and X_{i2} are uncorrelated or if $\gamma = 0$. If $\hat{\beta}_{OLS}$ is inconsistent, it carries over to $\hat{\alpha}_{OLS}$.
So we can leave out variables that belong in the regression as long as they are uncorrelated with the remaining variables.
A solution for omitted variable is to find an instrumental variable Z_i that is correlated with X_{i1}, but uncorrelated with X_{i2} and ε_i, then we need to use IV estimator to get an consistent estimate of β.

$$Cov\left[Z_i, X_{i1}\right] \neq 0$$
$$\mathbb{E}\left[Z_i X_{i2}\right] = \mathbb{E}\left[Z_i\varepsilon_i\right] = 0$$

Taking the probability limit to the IV estimator:

$$plim\left(\hat{\beta}_{IV}\right) = \beta + \frac{Cov\left[Z_i, (\varepsilon_i + \gamma X_{i2})\right]}{Cov\left[Z_i, X_{i1}\right]} = \beta$$

We see that it is consistent.

2.3 Simultaneous equation models

We look at the structural form of a two-equation model:

$$Y_1 = \alpha_1 Y_2 + \beta_1 Z_1 + \varepsilon_1$$
$$Y_2 = \alpha_2 Y_1 + \beta_2 Z_2 + \varepsilon_2$$

Where Z_1 and Z_2 are exogenous variables.
We can combine the two equation to one equation for Y_2, but only if $\alpha_1\alpha_2 \neq 1$.

$$\begin{aligned} Y_2 &= \alpha_2\alpha_1 Y_2 + \alpha_2\beta_1 Z_1 + \alpha_2\varepsilon_1 + \beta_2 Z_2 + \varepsilon_2 \\ Y_2(1 - \alpha_1\alpha_2) &= \alpha_2\beta_1 Z_1 + \beta_2 Z_2 + \alpha_2\varepsilon_1 + \varepsilon_2 \\ Y_2 &= \underbrace{\frac{\alpha_2\beta_1}{1 - \alpha_1\alpha_2}}_{=\pi_{21}} Z_1 + \underbrace{\frac{\beta_2}{1 - \alpha_1\alpha_2}}_{=\pi_{22}} Z_2 + \underbrace{\frac{\alpha_2\varepsilon_1 + \varepsilon_2}{1 - \alpha_1\alpha_2}}_{=\upsilon_2} \end{aligned}$$

From this we can see that Y_2 and ε_1 are generally correlated and then OLS is inconsistent when estimating Y_1. For the estimator to be consistent, we would require $\alpha_2 = 0$ and $Corr\,(\varepsilon_1, \varepsilon_2) = 0$. Similarly when $\alpha_1 = 0$ and ε_1 and ε_2 are uncorrelated, Y_1 would be uncorrelated with ε_2 and the OLS would be consistent.
The solution if Y_2 and ε_1 is correlated:
Use Z_2 as instrument for Y_2 when estimating Y_1.
If Z_2 had also been present in the equation for Y_1, we would not be able to use it as an instrument for Y_2 and would have to find something else.

2.4 Comparing OLS and the IV estimator

If there is no endogeneity of the initial explanatory variables, both OLS and IV will have consistent estimates, but OLS will be more efficient:

$$Var[\hat{\beta}_{IV}] = \frac{Var[\hat{\beta}_{OLS}]}{R^2_{X,Z}} \geq Var[\hat{\beta}_{OLS}]$$

4. Instrumental Variables

January 13, 2022

1 Introduction

When we want to estimate the following regression model:

$$Y_i = \boldsymbol{X}_i'\beta + \epsilon_i, \quad i = 1, 2, ..., n$$

we need assumption $A7$, $E\left(\boldsymbol{x}_i\varepsilon_i\right) = 0$, in order for the OLS estimator to be consistent. This means that for each i, the error term has zero mean and is uncorrelated with any explanatory variables. However, $A7$ is often violated in practice, making the OLS biased and inconsistent, which is the endogeneity problem.

When $A7$ is violated, we can estimate β with instrumental variables estimator and still get a consistent estimate. However we only want to use this estimator when $A7$ is violated since the IV estimator is less efficient than OLS. To use this estimator we need instruments, which are variables not included in $\mathbf{X}_i$ that are uncorrelated with the error term but correlated with the explanatory variables that are responsible for the violation of A7.

2 Assumptions

We look at the linear model:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim i.i.d.\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$$

Where $\mathbf{X}$ is an $n \times k$ matrix and $\mathbb{E}\left[\mathbf{X}'\varepsilon\right] \neq 0$.

We can identify m instruments arranged into the $n \times m$ matrix $\mathbf{Z}$ that satisfy the following conditions:

- IV conditions:
  1. $m \geq k$ is the **Order condition.** Exogenous variables serve as their own instruments (e.g. the constant)
  2. $plim\left(\frac{1}{n}\mathbf{Z}'\varepsilon\right) = E[\mathbf{Z_i}\varepsilon_i] = \mathbf{0}$. Instruments must be uncorrelated with the error term. **Validity**
  3. $plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) = E[\mathbf{Z_i}\mathbf{X_i'}] \equiv \mathbf{Q_{ZX}}$, $rank\left(\mathbf{Q_{ZX}}\right) = k$. This is the rank condition. The instruments must be correlated with the explanatory variables. **Relevance.**
  4. $plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right) = E[\mathbf{Z_i}\mathbf{Z_i'}] \equiv \mathbf{Q_{ZZ}}$, $rank(\mathbf{Q_{ZZ}}) = m$, no perfect collinearity between instruments → Q_{ZZ} is invertible. Equivalent to A6 where $\mathbf{Q}_{ZZ}$ is finite and invertible.
- Equation (1) is referred to as the Order Condition, which says that we must have at least as many instruments as explanatory variables.
- Assumption (2) requires the instruments to be uncorrelated with the error term, exogeneity.
- Equation (3) is the rank condition: the instruments must be sufficiently linearly related to the explanatory variables (relevance).
- Equation (4) is saying that there cannot be perfect collinearity between the instruments.

We assume assumption $A1$, $A2$ and our new assumptions for the IV estimator are:

$$(A3'): Var\left[\epsilon_i|\mathbf{X}, \mathbf{Z}\right] = \sigma^2, \text{ for all } i = 1, ..., n$$
$$(A4'): Cov\left[\epsilon_i, \epsilon_j|\mathbf{X}, \mathbf{Z}\right] = 0 \text{ for all } i = 1, ..., n$$

3 The IV estimator

There are two cases of estimation: the exactly identification where the number of instruments equals the number of parameters, $m = k$, and the over-identification where $m > k$, meaning there are more instruments than parameters.

3.1 Exact identification

Condition (2) is the moment condition, which we use to derive the Instrumental Variables estimator. This is done by replacing the population mean with the sample mean.
In the exactly identified case, where $m = k$, the IV estimator is given by the solution of the following m equations:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \left(Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$$

which yields:

$$\hat{\boldsymbol{\beta}}_{IV} = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i Y_i = \left(\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{Z}'\boldsymbol{Y}$$

3.2 Over identification

In the over-identified case, we have $m > k$:
As there are more equations than unknowns, we cannot in general set all the sample moments equal to zero.
Instead we need to use the Generalized Method of Moments and set the equations close to zero by minimizing the objective:

$$\begin{aligned} S(\hat{\beta}) &= (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &\equiv (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{P}_z(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

Where

$$\mathbf{P}_Z \equiv \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$$

The first order condition then becomes:

$$\left.\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}}\right|_{\hat{\beta}_{IV}} = -2\mathbf{X}'\mathbf{P}_Z\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{IV}\right) = \mathbf{0}$$

So the IV estimator becomes:

$$\hat{\beta}_{IV} = \left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_Z\boldsymbol{Y}$$

Another way to obtain the IV estimator is by a Two-Stage Least Squares procedure. This involves running two OLS regressions in a sequence.

1. Run a multivariate regression of $\mathbf{X}$ on $\mathbf{Z}$ to obtain a matrix of fitted values $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P}_Z\mathbf{X}$$

2. Regress $\mathbf{Y}$ on $\hat{\mathbf{X}}$ to obtain the IV estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IV} &= \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\boldsymbol{Y} \\ &= \left(\mathbf{X}'\mathbf{P}_Z\boldsymbol{X}\right)^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{Y} \end{aligned}$$

3.3 Properties of the IV estimator

- Consistency:

$$\hat{\beta}_{IV} = \left(\mathbf{X'Z}\left(\mathbf{Z'Z}\right)^{\mathbf{-1}}\mathbf{Z'X}\right)^{-1}\mathbf{X'Z}\left(\mathbf{Z'Z}\right)^{\mathbf{-1}}\mathbf{Z'}\left(\mathbf{X}\beta+\varepsilon\right)$$

$$= \beta + \left(\frac{1}{n}\mathbf{X'Z}\left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{Z'Z}\right)^{\mathbf{-1}}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{Z'X}\right)^{-1}\frac{1}{n}\mathbf{X'Z}\left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{Z'Z}\right)^{\mathbf{-1}}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{Z'}\varepsilon$$

Using IV conditions and taking the probability limit (plim), yields:

$$plim(\hat{\beta}_{IV}) = \beta + \left(\mathbf{Q'_{ZX}Q^{-1}_{ZZ}Q_{ZX}}\right)^{-1}\mathbf{Q'_{ZX}Q^{-1}_{ZZ}} * \mathbf{0}$$

$$= \beta$$

- The IV estimator is approximately normally distributed:

$$\hat{\beta}_{IV} \overset{a}{\sim} N\left(\beta, \sigma^2\left(\mathbf{X'P_ZX}\right)^{-1}\right)$$

- The error variance can be consistently estimated as:

$$\sigma^2_{IV} = \frac{1}{n-k}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}_{\mathbf{IV}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}_{IV}\right)$$

- Both IV and OLS are consistent in case of $E\left[\mathbf{X'}\varepsilon\right]=0$, but OLS will be more efficient

$$Var\left[\hat{\beta}_{IV}\right] = \frac{Var\left[\hat{\beta}_{OLS}\right]}{R^2_{X,Z}} \geq Var\left[\hat{\beta}_{OLS}\right]$$

where $R^2_{X,Z}$ is from the first step of 2SLS.

4 Hypothesis testing

4.1 Exogeneity Tests

If all the regressors are exogenous, then OLS is consistent and more efficient than IV. So we only want to use IV if some of the regressors are endogenous.
We want to test the following hypotheses:

$$\begin{aligned} H_0 :& \mathbb{E}\left[\mathbf{X'}\varepsilon\right] = \mathbf{0} \quad \textit{Exogeneity} \\ H_1 :& \mathbb{E}\left[\mathbf{X'}\varepsilon\right] \neq \mathbf{0} \quad \textit{Endogeneity} \end{aligned}$$

Under the null, both OLS and IV are consistent. We can:

1. Estimate by OLS
2. Estimate by IV using the instruments for the k^+ potentially endogenous regressors $\mathbf{X}^+$.
3. Compare the estimates: If they differ significantly, we conclude at least one regressor is endogenous.

A way to test for exogeneity is to run the Hausman Test. In this test we calculate the squared difference of the estimators, normalized by the covariance of the difference.

- **Hausman test for exogeneity:** Define $\mathbf{d} = \left(\hat{\beta}_{\mathbf{IV}} - \hat{\beta}_{\mathbf{OLS}}\right)$ and the test statistic is (squared difference):

$$H = \mathbf{d}' \hat{Var}[\mathbf{d}]^{-1}\mathbf{d}$$

$$= \mathbf{d}' \left[\hat{\sigma}^2 \left(\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\right]^{-1} \mathbf{d} \sim \chi^2\left(k^+\right)$$

  This is hard since the middle part has to be invertible (full rank) and it is invalid in case of weak instruments.

- **An alternative test for exogeneity**
  - Consider the general model, where $\mathbf{X_i^*}$ is a $(k - k^+) * 1$ vector of exogenous variables:

$$Y_i = \beta_0 + \left(\mathbf{X_i^+}\right)' \beta_1 + \left(\mathbf{X_i^*}\right)' \beta_2 + u_i$$

  - Suppose we have a $m \geq k$ IV's in $\mathbf{Z_i}$, we can test by:
    - $*$ 1) Obtaining the residuals $\hat{\upsilon}_{ij}$ from the k^+ auxiliary regressions:

$$X_{ij}^+ = \alpha_0 + Z_i'\alpha_1 + \upsilon_{ij}, \quad j = 1, ..., k^+$$

      and stack them into a $k^+ * 1$ vector $\hat{\upsilon}_i$.
    - $*$ 2) Then estimate:

$$Y_i = \beta_0 + \left(\mathbf{X_i^+}\right)' \beta_1 + \left(\mathbf{X_i^*}\right)' \beta_2 + \hat{\upsilon}_i\gamma + e_i$$

      abd test if $\gamma = 0$ using a Wald test with k^+ degrees of freedom, or an F-test with $(k^+, n - k - k^+)$ degrees of freedom.

4.2 Relevance Tests

We want the instruments to be as correlated with the regressors as possible for various reasons:

- The rank condition requires it
- The lower the correlation, the higher the variance of the IV estimator
- The lower the correlation, the higher chance of potential inconsistency due to endogenous instruments

We can perform a simple test to check if the instruments are relevant to the endogenous regressors. It is done in the following steps:
(Testing if $Cov\left(\mathbf{X}, \mathbf{Z}\right) \neq 0$)

1. Start by taking each of the endogenous regressors and regress them on all the instruments including the exogenous variables
2. Perform a $t-$ or $F-$test for the significance of the instruments.

4.3 Validity Tests

For consistency of the IV estimator we require $\mathbb{E}\left[\mathbf{Z}'\varepsilon\right] = \mathbf{0}$. To determine if the instruments are exogenous, we want to test the following hypotheses:

$$\begin{aligned} H_0 :& \mathbb{E}\left[\mathbf{Z}'\varepsilon\right] = \mathbf{0}, \quad \text{(Valid Instruments)} \\ H_1 :& \mathbb{E}\left[\mathbf{Z}'\varepsilon\right] \neq \mathbf{0}. \quad \text{(Invalid Instruments)} \end{aligned}$$

Note: When a instrument is exogenous, we say its valid.
However we can only test this assumption if we have more instruments than regressors $m > k$ (if we are in the over-identified case) IMPORTANT!

- First we start of by defining the $[n * m^*]$ matrix $\mathbf{Z}^*$ so that its columns are a subset of those of **Z,** and let $\mathbf{Z}^+$ be the remaining m^+ columns.
  - Note: Z^* are the m^* instrument that we **are sure are exogenous:** IV using the subset $\mathbf{Z}^*$ is **consistent** under **both** H_0 **and** H_1**.**
  - Note: $m = m^* + m^+$
- Under the null, all the instruments are exogenous, so the IV estimator using **Z** is consistent
- We can then
  - Estimate using **Z** to obtain $\hat{\beta}_{IV}$
  - Estimate using $\mathbf{Z}^*$ to obtain $\hat{\beta}^*_{IV}$
  - If the results differ significantly, conclude that one or more of the instruments in $\mathbf{Z}^+$ violate exogeneity
  - (This means that $\mathbf{Z}^+$ and m^+ are the instruments we are unsure of.
- We can use the **Hausman Test** from before to compare the two estimators:
  - Under the null, both estimators are consistent
  - Under the alternative, only $\hat{\beta}^*_{IV}$ is consistent.
  - (Notice it is very similar to the exogeneity test setup)
  - The Hausman test is conducted like before, and with m^+ degrees of freedom.

$$d = \left(\hat{\beta}_{IV} - \hat{\beta}^*_{IV}\right)$$

Additional tests to be aware of:

- We define the **J-statistic** as the objective function from $\mathbf{S}\left(\hat{\beta}\right) = \mathbf{G}\left(\hat{\beta}\right)' \hat{\mathbf{W}}\mathbf{G}\left(\hat{\beta}\right)$ (from the our over-identified case (where we had specified the G's))

$$J\left(\hat{\beta}\right) = \mathbf{G}\left(\hat{\beta}\right)' \hat{\mathbf{W}}\mathbf{G}\left(\hat{\beta}\right)$$

  We not let

$$J = \mathbf{G}\left(\hat{\beta}\right)' \hat{\mathbf{W}}\mathbf{G}\left(\hat{\beta}\right)$$

  where we use all the elements.
  And we let

$$J^* = \mathbf{G}\left(\hat{\beta^*}\right)' \hat{\mathbf{W}}^*\mathbf{G}\left(\hat{\beta^*}\right)$$

  where we only use the "safe" exogenous instruments.
- An easy way to compare the estimators is to check if the difference between J and J* differs significantly from zero.
- The test statistic is then:

$$J - J^* = G\left(\hat{\beta}\right)' \hat{W}G\left(\hat{\beta}\right) - G\left(\hat{\beta^*}\right)' \hat{W^*}G\left(\hat{\beta^*}\right) \overset{a}{\sim} \chi^2\left(m - m^*\right)$$

  or in others words, $\overset{a}{\sim} \chi^2\left(m^+\right)$.
  - If the difference hereof is too large, we conclude that $E\left[\left(\mathbf{Z}^+\right)' \varepsilon\right] \neq 0$ and as a consequence thereof, $E\left[\mathbf{Z}'\varepsilon\right] = 0$, maintaining the assumption that $E\left[\left(\mathbf{Z}^*\right)' \varepsilon\right] = 0$

- Another test is the **Sargan-Hansen Test,** which is a special case of the Hausman Test we get when $m^* = k$.
  - Note: When $m^* = k$, $J^* = 0$
  - The test statistic then becomes

    $$J = G\left(\hat{\beta}\right)' \hat{W} G\left(\hat{\beta}\right) \overset{a}{\sim} \chi^2\left(m-k\right)$$

    If the null is rejected, we conclude that some of the moments conditions and thus instruments are invalid
  - :
  - The Sargan-Hansen Test can be conducted with an **LM approach**
    - * 1) Estimate

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

      by IV to obtain the residuals $\hat{\varepsilon}_{IV} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}$
    - * 2) Perform the auxiliary regression:

      $$\hat{\varepsilon}_{IV} = \mathbf{Z}\gamma + \mathbf{u}$$

      by OLS to obtain the R^2
    - * 3) Calculate

      $$LM = nR^2 \overset{a}{\sim} \chi^2\left(m-k\right)$$

5. Generalized Method of Moments

January 13, 2022

1 Introduction

The general idea of Method of Moments estimation:

1. Specify one or more moment conditions
2. Replace the population conditions by their sample analogues
3. Solve for the estimator in question

The main benefits of using GMM

- Robust to distributional assumptions
- Handles non-linear models
- Handles models that are formulated as moment conditions

1.1 Method of moments

If we look at $Y_1, \ldots, Y_n$ which is an i.i.d. sample from a population with unknown mean μ, so that:

$$\mathbb{E}\left[Y_i - \mu\right] = 0, \quad i = 1, \ldots, n$$

Replacing the population mean by its sample analogue we get:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}\right) = 0 \Longleftrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

We can use a similar strategy to obtain estimates of higher-order moments.
For example the second raw moment:

$$\mathbb{E}\left[Y_i^2 - \mu_2\right] = 0, \quad i = 1, \ldots, n$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - \hat{\mu}_2\right) = 0 \Longleftrightarrow \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

We can estimate the first raw and second centered moments simultaneously, using the following moment conditions:

$$\mathbb{E}\left[\begin{pmatrix} Y_i - \mu \\ \left(Y_i - \mu\right)^2 - \mu_2^* \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

2 Generalized Method of Moments

In the general case with k unknown parameters and m distinct moment conditions:

$$\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\theta}_0; \boldsymbol{w}_i, \boldsymbol{Z}_i\right)\right] \equiv \mathbb{E}\left[\boldsymbol{g}_i\left(\boldsymbol{\theta}_0\right)\right] = \mathbf{0}, \quad i = 1, \ldots, n$$

where $\mathbf{g}_i : \mathbb{R}^k \rightarrow \mathbb{R}^n$ are known functions of the observed data $\mathbf{w}_i$, the instruments $\mathbf{Z}_i$ and the true underlying parameters θ_0

The GMM estimator is obtained by replacing the population moments by their sample analogues.

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

If

- $m = k$: the model is exactly identified
- $m > k$: the model is over-identified (Is solved by the numerical methods)
- $m < k$: the model is unidentified (Is not possible to solve, there are ∞ many solutions)

2.1 GMM Estimator

We define the $m \times 1$ moment vector:

$$\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\hat{\boldsymbol{\theta}})$$

In an over-identified model, there generally exist no exact solutions to the system of equations, since there are more equation than unknown parameters.
Instead we minimize $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})$ using a weighted sum of squares:

$$\hat{\boldsymbol{\theta}}_{GMM} = \underset{\hat{\theta}}{argmin} \boldsymbol{G}_n'(\hat{\boldsymbol{\theta}})\mathbf{W}_n\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})$$

Where $\mathbf{W}_n$ is called the weighting matrix.
$\hat{\boldsymbol{\theta}}_{GMM}$ often cannot be found analytically, so we have to rely on numerical methods.

3 Weighting Matrix

The weighting matrix, $\boldsymbol{W}_n$, is any positive definite matrix with the dimension $m \times m$ that may depend on data but not on the parameters, $\boldsymbol{\theta}$.
The purpose of this matrix is to determine the relative importance of violations of the individual moment conditions.
We can look at an example of a Weighting matrix:
Suppose $m = 2$ and let $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}) = (G_1, G_2)'$ and

$$\mathbf{W} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

Then $\boldsymbol{G}_n'(\hat{\boldsymbol{\theta}})\mathbf{W}_n\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}) = aG_1^2 + bG_2^2$. If we are more sure of the first moment conditions than the second, we can set $a \gg b$
But generally moments are not independent, so some correlation is expected, hence it may not be desirable to set the off-diagonal elements to zero.
Since the GMM estimator is already inefficient compared to the ML estimator, we might want to choose the W matrix such that the GMM estimator is efficient within the class of GMM estimators.
So we choose the weighting matrix to minimize the variance of the estimator, so the estimator is asymptotically efficient.
The optimal weighting matrix then becomes:

$$\begin{aligned} \boldsymbol{W}_{opt} &= R_0^{-1} \\ &= \left(E\left[\boldsymbol{g}_i\left(\boldsymbol{\theta}_0\right)\boldsymbol{g}_i'\left(\boldsymbol{\theta}_0\right)\right]\right)^{-1} \end{aligned}$$

Where R_0 is the asymptotic covariance matrix of the sample moments.
Since θ_0 is unknown, it is in practice not possible to use R_0^{-1} as weighting matrix
The solution to this is to use 2-step GMM

1. First step is to use $\boldsymbol{W_0} = \boldsymbol{I}$ to obtain $\hat{\boldsymbol{\theta}}_1$ which is consistent, but not efficient
2. Next calculate $\hat{\boldsymbol{W}}_1 = \left(\hat{\boldsymbol{R}}_1\left(\hat{\boldsymbol{\theta}}_1\right)\right)^{-1} = \frac{1}{n}\left(E\left[\boldsymbol{g}_i\left(\hat{\boldsymbol{\theta}}_1\right)\boldsymbol{g}_i'\left(\hat{\boldsymbol{\theta}}_1\right)\right]\right)^{-1}$ and use this to obtain $\hat{\boldsymbol{\theta}}_2$

If we iterate, we add the following steps:

1. Next calculate $\hat{\boldsymbol{W}}_2 = \left(\hat{\boldsymbol{R}}_2\left(\hat{\boldsymbol{\theta}}_2\right)\right)^{-1}$ and use this to obtain $\hat{\boldsymbol{\theta}}_3$
2. Continue until $\left|\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-1}\right| \approx 0$ since we converge to choose $\boldsymbol{W_n} = R_0^{-1}$.

4 Properties

Choosing the matrix using this procedure then the GMM estimator get the following properties:
The estimator is asymptotically consistent, hence

$$plim(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0$$

The estimator is asymptotically efficient within the class of GMM estimators
In finite sample, the estimator is approximately normal distributed

$$\hat{\theta}_{GMM} \overset{a}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{n}V\right)$$

5 Special cases

5.1 OLS and GMM

- Consider $\mathbf{Y_i} = \mathbf{X_i'}\beta_0 + \varepsilon_i$
- Use A7 to obtain:

  $$E\left[\mathbf{X_i}\varepsilon\right] = E\left[\mathbf{X_i}\left(\mathbf{Y_i} - \mathbf{X_i'}\beta_\mathbf{0}\right)\right] = 0$$

  (By inserting for the residuals)
  This is a vector of m moment conditions so $m = k \rightarrow$ exactly identified case

- Replacing the population mean by sample analogue, as we do it GMM:

  $$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i}\left(\mathbf{Y_i} - \mathbf{X_i'}\hat{\beta}_\mathbf{GMM}\right) = 0$$

  Solving for beta yields

  $$\hat{\beta}_{GMM} = \left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{Y_i}\right) = \left(\mathbf{X'X}\right)^{-1}\left(\mathbf{X'Y}\right)$$

- OLS is a special case of GMM. We don't need a weighting matrix because $m = k$.
- One can extent this with White or Newey West standard errors
- All the tests from OLS still apply

5.2 2SLS

- Suppose that $E\left[\mathbf{X_i}\varepsilon_\mathbf{i}\right] \neq 0$, which means A7 doesn't hold, but that we have access to a set of instruments $\mathbf{Z_i}$ that satisfy the moment condition

  $$E\left[\mathbf{Z_i}\varepsilon_\mathbf{i}\right] = E\left[\mathbf{Z_i}\left(\mathbf{Y_i} - \mathbf{X_i'}\beta_\mathbf{0}\right)\right] = 0$$

- If the model is exactly-identified, we can solve directly for the estimator using the method of moments

  $$\hat{\beta}_{GMM} = \left(\sum_{i=1}^{n}\mathbf{Z_i}\mathbf{X_i'}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z_i}\mathbf{Y_i}\right) = \left(\mathbf{Z'X}\right)^{-1}\left(\mathbf{Z'Y}\right)$$

  But if we are in the over-identified case, we first have to determine the weighting matrix (assuming A3 and A4 holds)

  $$\hat{\mathbf{W}}_\mathbf{n} = \left(\hat{\mathbf{R}}_\mathbf{n}\right)^{-1} = \left(\frac{1}{n}\sigma^2\mathbf{Z'Z}\right)^{-1}$$

- We then get the objective function:

$$S(\beta) = \underbrace{(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{Z}}_{G_n(\beta)'}(\mathbf{Z}'\mathbf{Z})^{\mathbf{-1}}\underbrace{Z'(Y-X\beta)}_{\mathbf{G_n}(\beta)}$$

  optimizing, we obtain the following GMM estimator

$$\hat{\beta}_{GMM} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{P_Z}\mathbf{Y}$$

  where $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{\mathbf{-1}}\,\mathbf{Z}^{'}$

- In the over-identified IV case, we have a weighting matrix we have to account for. (The one above)

$$\hat{\beta}_{GMM} = \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}_\mathbf{n}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}_\mathbf{n}\mathbf{Z}'\mathbf{Y}$$

  where we know $\hat{\mathbf{W}}_\mathbf{n} = \left(\frac{\mathbf{1}}{\mathbf{n}}\sigma^{\mathbf{2}}\mathbf{Z}'\mathbf{Z}\right)^{\mathbf{-1}}$

- The estimated covariance matrix is:

$$\hat{Var}\left[\hat{\beta}_{GMM}\right] = \left(\mathbf{X}'\mathbf{Z}\mathbf{W}_\mathbf{n}\hat{\mathbf{Z}}'\mathbf{X}\right)^{-1}$$

5.3 NLS and GMM

- Consider the non-linear model:

$$\mathbf{Y_i} = f(\mathbf{X_i};\beta_\mathbf{0}) + \varepsilon_i$$

- Now A7 is not sufficient so we use A2 instead: $E[\varepsilon_i|\mathbf{X_i}] = 0 \Rightarrow E[h(\mathbf{X_i})\,\varepsilon] = 0$
- The moment condition is:

$$E\left[\frac{\partial f(\mathbf{X_i};\beta)}{\partial\beta}(Y_i - f(\mathbf{X_i};\beta))\right] = 0$$

- This can be solved using numerical methods

6 Hypothesis testing

We want to test whether the imposed moment conditions are true. So, if the expected value of our moment is zero.

$$\begin{aligned} H_0 : E[g_i(\theta_0)] &= 0 \sim \text{Valid moment condition, hence consistent GMM} \\ H_1 : E[g_i(\theta_0)] &\neq 0 \sim \text{Invalid moment condition, hence inconsistent GMM} \end{aligned}$$

This only works when $m > k$. Then we use the J test (Sargan test)

$$J = G\left(\hat{\theta}_{GMM}\right)'\hat{R}_n^{-1}G\left(\hat{\theta}_{GMM}\right) \overset{a}{\sim} \chi^2(m-k)$$

1. Estimate the model using all the moment conditions.
2. Estimate using only a subset of "trusted" moment conditions.
3. We can also use the Hausman test to compare the results.

If we reject, we conclude that at least one of the "suspect" moment conditions is invalid.

6. Univariate Stationary Time series

January 13, 2022

1 Introduction

The aim of time series analysis is based on a sample of observations to identify the underlying data-generating process (DGP), which we then can use to do forecasting. We use forecasting in many economic situations, and therefore forecasting is very important. For example we use forecasting in macroeconomics to forecast inflation, GDP and unemployment rates. We also use it in the financial markets, to forecast interest rates, stock prices, and exchange rates. In order to do optimal forecasting we need to know the statistical properties of various time series models such that we can correctly identify the data generating process. Simple time series models often outperform more complicated structural models. The simplest process is the so-called white noise process, which satisfies the following assumptions:

$$E\left(\epsilon_t\right) = 0 \quad \forall t$$
$$E\left(\epsilon_t^2\right) = \sigma^2 \quad \forall t$$
$$E\left(\epsilon_t \epsilon_s\right) = 0 \quad \forall t \neq s$$

2 ARMA models (model selection, estimation, diagnostics and forecasting)

The $ARMA\left(p,q\right)$ (AutoRegressive Moving Average) model is given by:

$$y_t = c + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q}$$

where ϵ_t is the white noise process. The $ARMA\left(p,q\right)$ model is a combination of the $AR(p)$ (AutoRegressive) model and the $MA(q)$ (Moving Average) model.

2.1 Stationarity

The main concern regarding a time series $\boldsymbol{y} = \ldots, y_1, y_2, \ldots, y_t, \ldots, y_T, \ldots$ and the sample we observe $\boldsymbol{y}_T = \{y_1, y_2, \ldots, y_t, \ldots, y_T\}$ is whether it is stationary or non-stationary. We have that y is weakly stationary (also called covariance stationary) if

$$(i): \quad E\left(y_t\right) = \mu_y = \text{ constant}$$
$$(ii): \quad var\left(y_t\right) = \sigma_y^2 = \text{ constant}$$
$$(iii): \quad cov\left(y_t, y_{t-s}\right) = \gamma_s = f(s)$$

where γ_s is called the auto-covariance function - it is a measure of persistence.

2.2 Estimation

When estimating the models we will use maximum likelihood under the assumption of $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ - meaning that the errors are assumed to be Gaussian. For iid data with marginal pdf $f\left(y_t; \boldsymbol{\theta}\right)$, the joint density of the sample $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}'$ is the product of the marginal densities:

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

This means that the log-likelihood function can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log\left(f\left(y_t; \boldsymbol{\theta}\right)\right)$$

However this does not hold for a sample from a covariance stationary time series, since the random variables in the sample are not iid. A solution to this is to try to determine the joint density directly - however this requires, among other things, the $T \times T$ variance covariance matrix $var\left(\boldsymbol{y}\right)$. Another way to do it uses factorization of the joint density into a series of conditional densities and the density of a set of initial values.

2.2.1 The likelihood function for a Gaussian $ARMA\left(p, q\right)$ process

We can find the likelihood function for an $ARMA\left(p, q\right)$ process by conditioning on the initial values of the y's and the ϵ's, where we can then calculate the sequence of errors using:

$$\epsilon_t = y_t - c - \alpha_1 y_{t-1} - \ldots - \alpha_p y_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

The standard is to set the initial values of the ϵ equal to zero and the initial values of y equal to their actual values. Based on these values we do a iteration started at $t = p + 1$, which gives us the following log-likelihood function:

$$\ell(\boldsymbol{\theta}) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log\left(\sigma^2\right) - \sum_{t=p+1}^{T}\left[\frac{\epsilon_t^2}{2\sigma^2}\right]$$

where $\boldsymbol{\theta} = \left(c, \alpha_1, \ldots, \alpha_p, \theta_1, \ldots, \theta_q, \sigma^2\right)$.

2.3 Model selection

When selecting the model, we want to determine the lag orders. The aim is the most parsimonious model (the one with the fewest lags) with error being WN.
We can choose between our models based on the expected theoretical correlation patterns:

Process	Autocorrelation function	Partial autocorrelation function
$AR\left(p\right)$	Decays exponentially	Cuts off after lag p
$MA(q)$	Cuts off after lag q	Decays exponentially
$ARMA\left(p, q\right)$	Decays exponentially after q	Decays exponentially after p

2.3.1 ACF & PACF

The autocorrelation function (ACF) describes the correlation between Y_t and its lag Y_{t-s} as a function of s:

$$\rho_s = \frac{cov\left(Y_t, Y_{t-s}\right)}{var\left(Y_t\right)} = \frac{\gamma_s}{\gamma_0}$$

The partial autocorrelation function (PACF) gives the partial correlation of a stationary times series with its own lagged values all else is constant.

2.3.2 Diagnostics

After selecting a model, we check whether the residuals are approximately white noise, i.e. that they:

- Are normally distributed
- Have constant variance
- Are serially uncorrelated
  - To test if the model is serially uncorrelated, we can use an ARCH LM test
    - 1) Estimate the model under H_0 : homoscedasticity up to lag g
    - 2) Regress squared residuals from step 1 on a constant and lagged squared residuals up to order g

$$e_t^2 = \beta_0 + \sum_{s=1}^{g} \beta_s e_{t-s}^2 + u_t$$

    - 3) Calculate the LM-test (using R^2 from step 2) for joint significance of the explanatory variables in step 2. The test statistic is asymptotically $\chi^2(g)$ distributed

3 Motivation for time-varying volatility and ARCH models

With high-frequency data the assumption that ε_t all have the same variance σ^2 is often violated. When modeling a time-varying variance without violating the stationarity assumption, we need to use the the Autoregressive conditional heteroskedasticity (ARCH) model. Heteroskedasticity suggest volatility in the error term. So in the ARCH model the volatility is not constant over the time series but conditional on the time. However volatility is not persistent, it can jump up in one period and then jump back. If we need persistent volatility we need to look at GARCH since this model include the volatility of the previous period in the volatility of the current period.

3.1 ARCH model

The most simple model to use is the so-called $ARHC(1)$ model. This model is given by:

$$\sigma_t^2 = \varpi + \alpha\varepsilon_{t-1},$$

Here we need to impose the assumptions $\varpi \geq 0$ and $\alpha \geq 0$, to ensure that $\sigma_t^2 \geq 0$. The unconditional variance of ε_t is then given by:

$$\sigma^2 = E\left(\varepsilon_t^2\right) = \varpi + \alpha E\left(\varepsilon_{t-1}^2\right),$$

which has a stationary solution:

$$\sigma^2 = \frac{\varpi}{1-\alpha}$$

provided $|\alpha| < 1$. The $ARCH(1)$ model can be generalized to an $ARCH(p)$ process.

4 Forecasting

- Forecasting is of great importance in economics and our best estimate for h periods ahead depends on the information available at time t:

$$\hat{y}_{t+h|t} = E_t\left(y_{t+h}\right)$$

- There are two types of forecasting:
  - 1) **In-sample forecasting:** This is mainly used to measure the model fit.
    You estimate the model based on the observations $\{1, 2, ..., T\}$ and construct forecasts for the observations $\{2, ..., T\}$

- 2) **Out-of-sample forecasting:** Estimate the model based on the observations $\{1, 2, ..., t\}$ and construct forecasts for the observations $\{t+1, ..., T\}$. Distinguish two different approaches:
  - **Dynamic forecasting:** With the data from time $\{1, 2, ..., t\}$, you forecast and estimate for all periods after, $\{t+1, ..., T\}$, $\hat{y}_{t+1|t}, \hat{y}_{t+2|t}, ..., \hat{y}_{t+h|t}$
  - **Static forecasting:** Is based on the 'rolling window' of data. For every period, we re-estimate the parameters for each forecast: $\hat{y}_{t+1|t}, \hat{y}_{t+2|t+1}, ..., \hat{y}_{t+h|t+h-1}$

- To evaluate the usefulness of the forecasts we use loss functions, e.g. RMSE (Root mean square error: $\sqrt{\frac{1}{H}\sum_{h=1}^{H}\hat{\varepsilon}_{t+h}^{2}}$) and MAE (Mean absolute error: $\frac{1}{H}\sum_{h=1}^{H}|\hat{\varepsilon}_{t+h}^{2}|$). (The MAE is more robust to outliers)

- We can compare two forecast models using **Diebold-Mariano test:**

$$DM = \frac{\hat{d}}{\sqrt{Var\left(\hat{d}\right)}} \overset{a}{\sim} N\left(0,1\right)$$

where

$$\hat{d} = \frac{1}{H}\sum_{h=1}^{H}\left[L\left(\hat{\varepsilon}_{1,t+h}\right) - L\left(\hat{\varepsilon}_{2,t+h}\right)\right]$$

with $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ denoting the forecast errors from model 1 and 2, respectively, and L(ε) denoting the loss function, which is assumed to be $L\left(\varepsilon\right) = \varepsilon^2$.
If the test is significantly different from 0, it is an indication that one of the forecast is better than the other.

  - In choosing the best model one should note that the model that has the best in-the-sample fit, not necessarily has the best out-of-sample fit. This is due to parameter and model uncertainty, while also the data generating process can suffer from structural breaks and hence vary over time.

7. Univariate Non-Stationary Time Series

January 13, 2022

1 Introduction

The aim of time series analysis is based on a sample of observations to identify the underlying data-generating process (DGP), which we then can use to do forecasting. We use forecasting in many economic situations, and therefore forecasting is very important. The simplest time series process is the so-called white noise process, which satisfies the following assumptions:

$$\begin{aligned} E\left(\epsilon_t\right) &= 0 \quad \forall t \\ E\left(\epsilon_t^2\right) &= \sigma^2 \quad \forall t \\ E\left(\epsilon_t \epsilon_s\right) &= 0 \quad \forall t \neq s \end{aligned}$$

The main concern regarding a time series $\boldsymbol{y} = \ldots, y_1, y_2, \ldots, y_t, \ldots, y_T, \ldots$ and the sample we observe $\boldsymbol{y}_T = \{y_1, y_2, \ldots, y_t, \ldots, y_T\}$ is whether it is stationary or non-stationary. We have that y is weakly stationary (also called covariance stationary) if

$$\begin{aligned} (i): &\quad E\left(y_t\right) = \mu_y = \text{ constant} \\ (ii): &\quad var\left(y_t\right) = \sigma_y^2 = \text{ constant} \\ (iii): &\quad cov\left(y_t, y_{t-s}\right) = \gamma_s = f(s) \end{aligned}$$

where γ_s is called the auto-covariance function - it is a measure of persistence. A time series is non-stationary when some of the stationarity conditions are violated.

In this topic, we look at unit roots as a particular time series model. Knowing the right distribution of the underlying process is important for forecasting, policies (to distinguish between permanent and transitory effects),and to avoid spurious regressions (which often stem from regressing two independent unit root processes on each other)

2 Deterministic and stochastic trends

2.1 Deterministic Trend model

An example of a non-stationary model is the deterministic trend model. The deterministic trend model is given by:

$$y_t = c + \lambda t + \epsilon_t$$

which has $E\left(y_t\right) = c + \lambda t$ for $\lambda \neq 0$. This type of model is said to be trend stationary, and shocks to this type of model will only have transitory effects.
If we consider a stationary $AR(1)$ model with a deterministic linear trend term:

$$y_t = c + \alpha y_{t-1} + \lambda t + \epsilon_t$$

where ϵ_t is white noise with $var\left(\epsilon_t\right) = \sigma^2$. Using the lag operator we can rewrite the model in the following way:

$$\begin{aligned} y_t &= \frac{c}{1-\alpha} + \frac{\lambda}{1-\alpha L}t + \frac{1}{1-\alpha L}\epsilon_t \\ &= \mu_0 + \mu_1 t + \frac{1}{1-\alpha L}\epsilon_t \end{aligned}$$

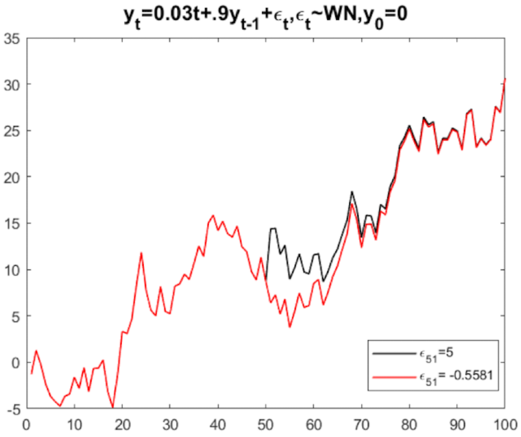
where $\mu_0 = \frac{c}{1-\alpha}$ and $\mu_1 = \frac{\lambda}{1-\alpha L}$.
The mean of y_t is then

$$E\left(y_t\right) = \mu_0 + \mu_1 t$$

which contains a linear trend, while the variance is constant, $var\left(y_t\right) = \frac{\sigma^2}{1-\alpha^2}$. This means that the original process is not stationary, but when we take the deviation from the mean we get a stationary process:

$$y_t - E\left(y_t\right) = y_t - \mu_0 - \mu_1 t = \frac{1}{1-\alpha L}\epsilon_t$$

This means that y_t is trend stationary, and since ϵ_t is white noise shocks only have transitory effects.



2.2 Stochastic Trend Model

2.2.1 Random walk process

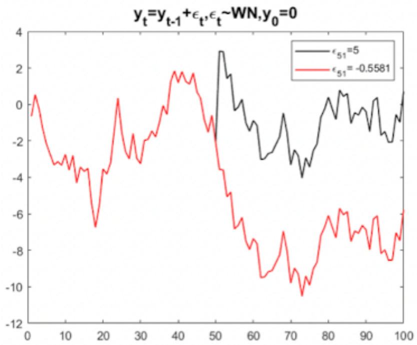
Another example of a model that is non-stationary, and thereby violate the stationarity conditions, is the random walk model. The random walk model is given by:

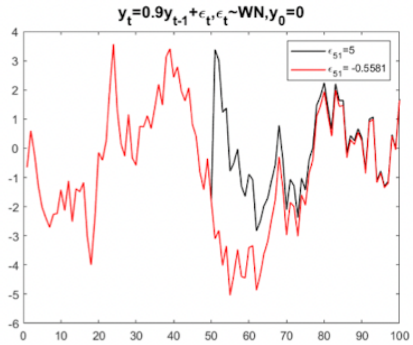
$$y_t = y_{t-1} + \epsilon_t$$

where ϵ_t is white noise with $var\left(\epsilon_t\right) = \sigma^2$. The random walk model has $E\left(y_t\right) = y_0$ and $var\left(y_t\right) = t\sigma^2$. This is an example of a unit root process, also called a stochastic trend model. This type of model is called difference stationary, and shocks to this type of model will have permanent effects.
By recursive substitution we can write the random walk model as follows:

$$y_t = y_0 + \sum_{i=0}^{t-1} \epsilon_{t-i}$$

Again, since ϵ_t is a white noise, we see that $E\left(y_t\right) = y_0$ and $var\left(y_t\right) = t\sigma^2$. The last term of the equation $\sum_{i=0}^{t-1} \epsilon_{t-i}$ is called the stochastic trend. Shocks to the process are accumulated over time and thus have permanent effects.





2.2.2 Random walk with a drift

We consider the random walk with a drift:

$$y_t = c + y_{t-1} + \epsilon_t$$

By recursive substitution we can rewrite the model as follows:

$$y_t = y_0 + ct + \sum_{i=0}^{t-1} \epsilon_{t-i}$$

Where we can see that $E\left(y_t\right) = y_0 + ct$ and $var\left(y_t\right) = t\sigma^2$. This random walk with a drift model has both a deterministic trend ct and a stochastic trend. These types of models are also called unit root processes. When rewriting these two models we get:

$$\begin{array}{rc} \text{Random walk:} & y_t - y_{t-1} = \epsilon_t \\ \text{Random walk with drift:} & y_t - y_{t-1} = c + \epsilon_t \end{array}$$

This implies that the first difference of the models is stationary, meaning y_t is difference stationary.

A more general class of unit root processes is generated as a RW (with drift) where innovations are allowed to be general stationary process i.e. an $ARMA(p, q)$.
An $ARIMA\left(p, d, q\right)$ process is a process which after having been differenced d times is an $ARMA\left(p, q\right)$ process

3 Unit root tests

A series that needs to be differenced once to get stationarity is called an $I(1)$ process, and a stationary series is called an $I(0)$ process. We can a feeling of this by looking at the Autocorrelation Function. For a stationary series, the autocorrelations decline rapidly as the lag increases, while for a non-stationary series, the autocorrelation declines very slowly.
We can test the presence of unit roots more formally by using a DF, KPSS, or PP test.

- Extra: As the null hypothesis is often that there is a unit root process, we run into the problem of a non-standard and non-normal asymptotic distribution to which there are no convenient closed form expressions. Consequently, critical values must be calculated using simulation methods.
    - Consider the following DGP

      $$y_t = \alpha y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \; white \; noise$$

      where an (intuitive) hypothesis for testing for a unit root is

      $$H_0 : \alpha = 1 \quad \text{The process has a unit root, i.e. is I(1)}$$

      $$H_1 : |\alpha| < 1 \quad \text{The process is stationary, i.e. is I(0)}$$

- Note that for the OLS estimator of α

  $$\sqrt{T}\left(\hat{\alpha}-\alpha\right) \overset{d}{\to} N\left(0, 1-\alpha^2\right)$$

  which, under the null hypothesis becomes

  $$\sqrt{T}\left(\hat{\alpha}-1\right) \overset{d}{\to} N\left(0, 0\right)!$$

  (This is no bueno)

3.1 Dickey Fuller Test

The null hypothesis is that the process has a unit root, i.e. is $I(1)$ and the alternative is that the process is stationary i.e. is $I(0)$.
There are different DF test procedures depending on the underlying data generating process.
As an example, I look at an $AR(1)-$process without constant and deterministic trend:

$$y_t = \alpha y_{t-1} + \epsilon_t$$

$$\begin{aligned} H_0 :& \alpha = 1 \\ H_1 :& |\alpha| < 1 \end{aligned}$$

With the usual asymptotic properties we would have

$$\sqrt{T}(\widehat{\alpha}-1) \xrightarrow{d} \mathcal{N}(0,0)$$

under the null hypothesis. This means we cannot use the usual asymptotic if y_t is an unit root process. Therefore we re-parameterize the model in the following way:

$$\begin{gathered} y_t - y_{t-1} = (\alpha-1)y_{t-1} + \epsilon_t \\ \Updownarrow \\ \Delta y_t = \gamma y_{t-1} + \epsilon_t \end{gathered}$$

This implies the following hypothesis test:

$$\begin{aligned} H_0 &: \gamma = 0 \text{ The process has a unit root} \\ H_1 &: \gamma < 0 \text{ The process is stationary} \end{aligned}$$

In principle, we have to difference the series until we are able to reject the null:

$$Y_t = \begin{cases} \text{Reject null} & \Rightarrow Y_t \text{ is stationary} \\ \text{Do not reject null} & \Rightarrow \Delta Y_t \begin{cases} \text{Reject null} & \Rightarrow \Delta Y_t \text{ is stationary} \\ \text{Do not reject null} & \Rightarrow \text{We might have unit root} >1 \end{cases} \end{cases}$$

The Dicker-Fuller test of null of a unit root is then:

$$t_{DF} = \frac{\widehat{\gamma}}{\text{s.e. }(\widehat{\gamma})} \sim tDF_1 - \text{distribution}$$

where the 5% critical value in tDF_1 is -1.94 for large sample sizes. 1
With the following auxiliary regressions

1. If the DGP is given as $y_t = c + \alpha y_{t-1} + \varepsilon_t$ where ε_t is white noise, the auxiliary regression becomes

   $$\Delta y_t = c + \gamma y_{t-1} + \varepsilon_t$$

2. If the DGP is given as $y_t = c + \lambda t + \alpha y_{t-1} + \varepsilon_t$ where ε_t is white noise the auxiliary regression becomes

   $$\Delta y_t = c + \lambda t + \gamma y_{t-1} + \varepsilon_t$$

3.2 Augmented Dickey-Fuller test

The Dickey-Fuller distribution requires the innovations in the auxiliary regression to be serially uncorrelated. This motivates the ADF:

$$\Delta y_t = c + \lambda t + \gamma y_{t-1} + \sum_{i=1}^{k} \beta_i \Delta y_{t-i} + \varepsilon_t$$

where k is chosen such that the residuals are white noise. The Augmented Dickey-Fuller test statistic t_{ADF} has the same asymptotic distribution as the t_{DF}.
Practical challenges in the ADF test:

1. Should we include a constant, a constant and a linear trend or none of the two? This is important because the asymptotical distribution of the t-statistic will change under the null hypothesis dependent on the case we are in (De tre cases fra note i Notability, tror jeg)
   (a) If we neglect relevant terms, we have a tendency to maintain the null too often.
2. How many lags should we include?
   (a) If we include too few lags, the test will be positively biased, and we will reject the null too often.
   (b) If we include too many lags, the test will have too low power and we will maintain the null too often.

3.3 PP-test

This is an alternative to the ADF -test. Where the ADF-test uses additional lags of the first-difference variable in order to account for serial correlation in the errors, the Phillips-Perron test applies a non-parametric correction. Note that the problem with the serially correlated error terms is that y_{t-1} becomes endogenous.
Here, we also have that a null hypothesis of a unit root and an alternative hypothesis of a stationary process.

3.4 KPSS-test

Important: Here the hypothesis is opposite of PP and ADF!

$$H_0 : \text{The process is stationary, i.e. I(0)}$$

$$H_1 : \text{The process has a unit root, i.e. I(1)}$$

- Calculate the test-statistic as follows:
  - 1) Perform an auxiliary regression of y_t on a constant (and possible a deterministic trend term) and save the residuals $\{e_t\}_{t=1}^{T}$
  - 2) Compute the cumulative residuals function

    $$S_t = \sum_{s=1}^{t} e_s, \ \wedge \forall t$$

  - 3) Compute the test statistic

    $$KPSS = T^{-2} \sum_{t=1}^{T} \frac{S_t^2}{\hat{\sigma}^2}$$

    where $\hat{\sigma}^2$ is an estimator for the 'long-run variance'.

4 Extra (only for questions)

4.1 Higher order integration

Economic time series are some times quite smooth and trending which could indicate that these series are $I(2)$. If $y_t \sim I(2)$ then

$$\Delta^2 y_t = (1-L)^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \sim I(0)$$

4.2 Forecasting non-stationary models

4.2.1 Forecasting linear deterministic trend models

Consider the simple linear deterministic trend model

$$\begin{aligned} y_{t+1} &= c + \lambda (t+1) + \varepsilon_{t+1} \\ &\vdots \\ y_{t+h} &= c + \lambda (t+h) + \varepsilon_{t+h} \end{aligned}$$

where $E(\varepsilon_{t+1}|I_t) = 0$ and $var(\varepsilon_t) = \sigma^2$. The h-step ahead forecast for this model is

$$\hat{y}_{t+h|t} = c + \lambda (t+h)$$

with forecast error variance $\hat{\sigma}^2$

4.2.2 Forecasting random walk (RW)

Consider the random walk with drift

$$y_{t+1} = c + y_t + \varepsilon_{t+1}$$

$$y_{t+2} = c + y_{t+1} + \varepsilon_{t+2} = 2c + y_t + \varepsilon_{t+1} + \varepsilon_{t+2}$$

$$y_{t+h} = c + y_{t+h-1} + \varepsilon_{t+h} = hc + y_t + \varepsilon_{t+1} + ... + \varepsilon_{t+h}$$

where $E(\varepsilon_{t+1}|I_t) = 0$ and $var(\varepsilon_t) = \sigma^2$. The h-step ahead forecast for this model is then

$$\hat{y}_{t+h|t} = hc + y_t$$

with forecast error variance of $h\sigma^2$
If there is no drift, we have that: $c = 0$, which means that $\hat{y}_{t+h|t} = hc + y_t = y_t$

5 Motivation for time-varying volatility and ARCH models

With high-frequency data the assumption that ε_t all have the same variance σ^2 is often violated. When modeling a time-varying variance without violating the stationarity assumption, we need to use the the Autoregressive conditional heteroskedasticity (ARCH) model. Heteroskedasticity suggest volatility in the error term. So in the ARCH model the volatility is not constant over the time series but conditional on the time. However volatility is not persistent, it can jump up in one period and then jump back. If we need persistent volatility we need to look at GARCH since this model include the volatility of the previous period in the volatility of the current period.

5.1 ARCH model

The most simple model to use is the so-called $ARCH(1)$ model. This model is given by:

$$\sigma_t^2 = \varpi + \alpha \varepsilon_{t-1},$$

Here we need to impose the assumptions $\varpi \geq 0$ and $\alpha \geq 0$, to ensure that $\sigma_t^2 \geq 0$. The unconditional variance of ε_t is then given by:

$$\sigma^2 = E\left(\varepsilon_t^2\right) = \varpi + \alpha E\left(\varepsilon_{t-1}^2\right),$$

which has a stationary solution:

$$\sigma^2 = \frac{\varpi}{1-\alpha}$$

provided $|\alpha| < 1$. The $ARCH(1)$ model can be generalized to an $ARCH\,(p)$ process.

5.2 GARCH model

If we need persistent volatility, we use the GARCH model, that also depends on the volatility of the previous period.

- GARCH(1,1) is

  $$\sigma_t^2 = \overline{\omega} + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

  By recursive substitution, the GARCH$(1,1)$ model can be rewritten as an $ARCH\,(\infty)$ model.

- Threshold $GARCH\,(1,1)$

  $$\sigma_t^2 = \overline{\omega} + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma I_{t-1} \varepsilon_{t-1}^2$$

  where $I_{t-1} = 1$ if $\varepsilon_{t-1} < 0$ and zero otherwise.
  If $\gamma > 0$ negative shocks have a larger impact on future volatility than do positive shocks of the same magnitude.

- Exponential $GARCH\,(1,1)$

  $$log\left(\sigma_t^2\right) = \overline{\omega} + \beta log\left(\sigma_{t-1}^2\right) + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha \frac{|\varepsilon_{t-1}|}{\sigma_{t-1}}$$

  where $\gamma \neq 0$ implies asymmetry and $\gamma < 0$ implies that negative shocks have a larger impact on future volatility than positive shocks of the same magnitude do.

- $GARCH - M$: In financial theory certain sources of risk are priced by the market. if σ_t^2 is an appropriate measure of risk, the conditional variance may enter the conditional mean function of y_t

  $$y_t = \mathbf{x_t'}\theta + \delta \sigma_t^2 + \varepsilon_t \qquad \varepsilon_t | I_{t-1} \sim N\left(0, \sigma_t^2\right)$$

  where the process for σ_t^2 could be any of those presented above. Other measures of risk in the mean equation can be included.

The models can be estimated using maximum likelihood and one can make volatility forecast too.

8 - Cointegration and Spurious Regression

January 13, 2022

1 Introduction and ADL

Often we want to model the dynamic relation between two or more variables. To do so, we can use the Autoregressive Distributed Lag model, $ADL\left(p,r\right)$. The $ADL(1,1)$ model is given by:

$$y_t = \alpha + \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t$$

The primary interest in these types of models is the analysis of short-run and long-run effects of a change in the explanatory variable. In the $ADL\left(1,1\right)$ model, β_0 is the immediate impact on y_t, and $\frac{\beta_0+\beta_1}{1-\phi}$ is the long-run multiplier, and can be interpreted as the long-run impact on y of a permanent change in $x$of one unit.
In the $ADL\left(1,1\right)$ model, both x and y are stationary.
If x_t changes the immediate impact on y is called the impact multiplier $\left(SR\right)$.

$$\begin{aligned} \text{Immediate impact SR: } \frac{\partial y_t}{\partial x_t} &= \beta_0 \\ \text{One period:} \frac{\partial y_{t+1}}{\partial x_t} &= \phi\beta_0 + \beta_1 \\ \text{Two periods:} \frac{\partial y_{t+2}}{\partial x_t} &= \phi\left(\phi\beta_0 + \beta_1\right) \end{aligned}$$

Continuing and accumulating the effects we get the LR-multiplier:

$$\begin{aligned} \text{LR-multiplier: } &= \beta_0 + \left(\phi\beta_0 + \beta_1\right) + \phi\left(\phi\beta_0 + \beta_1\right) + \phi^2\left(\phi\beta_0 + \beta_1\right) ... \\ &= \beta_0 + \left(1 + \phi + \phi^2 + ...\right)\left(\phi\beta_0 + \beta_{_1}\right) \\ &= \frac{\beta_0 - \beta_0\phi}{1-\phi} + \frac{\phi\beta_0 + \beta_1}{1-\phi} \\ &= \frac{\beta_0 + \beta_1}{1-\phi} \end{aligned}$$

The last one is the LR impact on y of a permanent change in x of one unit. Note that this derivation requires $|\phi| < 1$, which is ensured because y is stationary, i.e. $I\left(0\right)$.
The general $ADL\left(p,r\right)$ model with one explanatory variable is

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=0}^{r} x_{t-j}\beta_j + \varepsilon_t$$

And from this, we can see that the long-run equilibrium is

$$E\left(y_t\right) = \frac{\alpha}{1-\sum_{i=1}^{p}\phi_i} + \frac{\sum_{j=0}^{r}\beta_j}{1-\sum_{i=1}^{p}\phi_i} E\left(x_t\right)$$

- Provided usual assumptions, estimate ADL using OLS
- The choice of lag length in dynamic models are based on:
  - No serial correlation in errors
  - Significant coefficients
  - Minimization of information criteria

2 Error correction models

The error correction model is where we estimate the short-run model, after we have restricted the long-run properties of the model. An example of this is considering the $ADL\,(1,1)$ model:

$$y_t = \alpha + \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t$$

We then make manipulations:

$$\begin{aligned} y_t - y_{t-1} &= \alpha + \phi y_{t-1} - y_{t-1} + \beta_0 x_t - \beta_0 x_{t-1} + \beta_0 x_{t-1} + \beta_1 x_{t-1} + \varepsilon_t \\ &= \alpha - (1-\phi)\, y_{t-1} + \beta_0\,(x_t - x_{t-1}) + (\beta_0 + \beta_1)\, x_{t-1} + \varepsilon_t \\ \Delta y_t &= \beta_0 \Delta x_t - (1-\phi)\left[y_{t-1} - \frac{\alpha}{(1-\phi)} - \frac{(\beta_0+\beta_1)}{(1-\phi)} x_{t-1}\right] + \varepsilon_t \\ &= \beta_0 \Delta x_t - (1-\phi)\left[y_{t-1} - \alpha^* - \beta^* x_{t-1}\right] + \varepsilon_t \sim ECM \end{aligned}$$

A long-run relation we could impose is, for example, $y = x$ which implies that $\frac{(\beta_0+\beta_1)}{(1-\phi)} = 1$ and $\frac{\alpha}{1-\phi} = 0$ and inserted in (1):

$$\Delta y_t = \beta_0 \Delta x_t - (1-\phi)\{y_{t-1} - x_{t-1}\} + \epsilon_t$$

There are two systematic effects on the change in the dependent variable Δy_t

1. The instantaneous multiplier effect $\beta_0 \Delta x_t$, due to changes in the explanatory variable
2. The partial correction for the deviation from the long-run equilibrium $(\phi - 1)\{y_{t-1} - x_{t-1}\}$

If $0 < \phi < 1$ and $y_{t-1} > x_{t-1}$ then Δy_t will decrease, i.e. y will move in the direction of equilibrium (and vice versa if $y_{t-1} < x_{t-1}$)

3 Definition of cointegration

If y_t and x_t are two $I(1)$ processes, then, in general, $y_t - \beta x_t$ is an $I(1)$ process for any number β. However, if a $\beta \neq 0$ exists, such that $y_t - \beta x_t$ is an $I(0)$ process, then we say that y and x are cointegrated. This means that they share a common trend. We call β the cointegration parameter and $[1, \beta]'$ the cointegration vector.

The economic intuition behind this phenomenon is based on the belief that a certain pair of economic variables should not diverge from each other by too great an extend in the long run. This means that such variables may drift apart in the short run, but in the long run economic forces will begin to bring them together again. Examples of such variables are:

- Interest rates of different maturities.
- Consumption and income.
- Prices and wages.

4 Spurious regression

A spurious regression is a regression that provides misleading statistical evidence of a linear relationship between independent non-stationary variables. It is very important to be aware of the stationarity of the individual time series. In the case of non-stationarity, the variance of the time series is not fluctuating around a constant mean, and the OLS will become misleading. For instance, if both Y_t and X_t are generated by a random walk process, a regression of Y_t on X_t could yield to a suprious regression, i.e. fairly high R^2 even when there are no actual correlation.
We can tests for spurious regressions by regressing first differences of the series on each other. Only truly related series would yield high R^2 in this case.

$$y_t - y_{t-1} = (\beta_0 + \beta_1 x_t + u_t) - (\beta_0 + \beta_1 x_{t-1} + u_{t-1})$$

$$\Delta y_t = \beta_1 \Delta x_t + \Delta u_t$$

5 Engle-Granger test

If we have a hypothesized value of β we can test for cointegration in the following way:

- Define a new variable, $d_t = y_t - \beta x_t$, and apply the Augmented Dickey-Fuller test to d_t. The ADF has the following null- and alternative hypothesis:

$$H_0 : \gamma = 0 \quad \text{The process has a unit root, i.e. is } I(1)\text{(Are not cointegrated)}$$
$$H_1 : \gamma < 0 \quad \text{The process is stationary, i.e. is } I(0)\text{(Are cointegrated)}$$

- If we reject the null hypothesis, that the process has a unit root, we conclude that y_t and x_t are cointegrated. This means that the null hypothesis is that y_t and x_t are not cointegrated.

When the potential cointegration parameter β is unknown, we have to estimate it and thereafter apply the ADF test to the residuals:

$$\widehat{d_t} = y_t - \widehat{\alpha} - \widehat{\beta} x_t$$

This is The Engle-Granger test. The problem with this test is that under the null hypothesis, the two series are not cointegrated. This means that we are running a spurious regression. The critical values in this test are therefore different from the usual critical values associated with the ADF test, because they have to take into account the estimation of β.
The Engle-Granger test follows the following procedure:

1. First we examine whether y_t and x_t are unit root processes. It only makes sense to test for cointegration if this is the case.

2. The second step is to run the regression $y_t = \alpha + \beta x_t + d_t$, and thereby obtain the residuals $\hat{d}_t$.

3. We then run the regression $\Delta \hat{d}_t$ on $\hat{d}_{t-1}$, and possibly lags of $\Delta \hat{d}_t$ in order to account for serial correlation.

$$\Delta \hat{d}_t = c + \gamma \hat{d}_{t-1} + \sum_{i=1}^{k} \alpha_i \Delta \hat{d}_{t-i} + u_i$$

4. At last we compare the t-statistic on $\Delta \hat{d}_{t-1}$ to the desired critical value. If the t-statistic is above the critical value, we have evidence that $y_t - \beta x_t$ is $I(0)$ fro some β. This means that y_t and x_t are cointegrated.

- If one account for serial correlation by robust standard errors instead of using lags in the ADF test → Phillips-Ouliaris cointegration-test

- If y_t and x_t contain drift terms, $E(y_t)$ and $E(x_t)$ are linear functions of time, but the strict definition of cointegration requires $y_t - \beta x_t$ to be $I(0)$ without a trend → we have to take the trend into account: We run the second step of EG-test with $y_t = \alpha + \lambda t + \beta x_t + d_t$ instead (other critical values)
- EG-test can also be used for cointegration of three or more variables however the critical values must be adjusted
- Caveat; EG cannot test for more than one cointegration relationship (if k variables there are $k-1$ potential cointegration relationships) and it is sensitive to which variable is the regressand (→ Johansen's VAR approach)

5.1 Error-correction models and cointegration

If y and x are $I(1)$ and cointegrated, an error correction model can be made as:

$$\Delta y_t = \theta + \gamma \Delta x_t + \delta \hat{d}_{t-1} + \varepsilon_t$$

all therms in this are $I(0)$.

- If we know α and β we can estimate the error correction model by regressing Δy_t on Δx_t and $d_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$. If not, we have to estimate α and β to obtain $\hat{d}_{t-1} = y_{t-1} - \hat{\alpha} - \hat{\beta} x_{t-1}$
- When testing the other parameters in the error correction model, we can ignore the preliminary estimation of α and β (asymptotically).

9. Vector Autoregressive Models

January 13, 2022

1 Introduction

We look at the multivariate case of the autoregressive model, which allows us to account for the data generating process for all included variables and model the dynamic interrelation between variables which include causality and impulse-response functions.

2 Vector Autoregressive Models

The general vector autoregressive model is given by:

$$Y_t = \underbrace{\delta}_{(k\times 1)} + \underbrace{\Theta_1}_{(k\times k)}\underbrace{Y_{t-1}}_{(k\times 1)} + \cdots + \underbrace{\Theta_p}_{(k\times k)}\underbrace{Y_{t-p}}_{(k\times 1)} + \underbrace{\epsilon_t}_{(k\times 1)} \quad t = 1, 2, \ldots, T \quad (1)$$

where $\epsilon \sim iid\,(0, \Sigma)$

$$\sum_{(k\times k)} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix}$$

where $E\,(\varepsilon_t) = 0$, $Var\,(\varepsilon_t) = \underbrace{\sum}_{k\times k}$, $cov\,(\varepsilon_t, \varepsilon_{t-1}) = 0 \quad \forall\, l > 0$, $\varepsilon_t \sim iid\,(0, \sum)$.

The total number of parameters to be estimated is $k + pk^2 + \frac{1}{2}k\,(k+1)$

2.1 Properties of the VAR(1)

The $VAR\,(1)$ is:

$$\mathbf{Y_t} = \delta + \mathbf{\Theta Y_{t-1}} + \varepsilon_\mathbf{t}$$

By recursive substitution, we have that:

$$\mathbf{Y_t} = \mathbf{\Theta^t Y_0} + \sum_{\mathbf{i=0}}^{\mathbf{t-1}} \mathbf{\Theta^i}\delta + \sum_{\mathbf{i=0}}^{\mathbf{t-1}} \mathbf{\Theta^i}\varepsilon_{\mathbf{t-i}}$$

By assuming that $Y_0 = 0$ and the process is stationary iff:

$$\mathbf{E}\left(\mathbf{Y_t}\right)= \sum_{\mathbf{i=0}}^{\infty} \mathbf{\Theta^i} \delta = \left(\mathbf{I} - \mathbf{\Theta}\right)^{\mathbf{-1}} \delta$$

$$\mathbf{Var}\left(\mathbf{Y_t}\right)= \sum_{\mathbf{i=0}}^{\infty} \mathbf{\Theta^i} \sum \left(\mathbf{\Theta'}\right)^{\mathbf{i}}$$

$$\mathbf{Cov}\left(\mathbf{Y_t}, \mathbf{Y_{t-m}}\right)= \sum_{\mathbf{i=0}}^{\infty} \mathbf{\Theta^{m+i}} \sum \left(\mathbf{\Theta'}\right)^{\mathbf{i}}$$

For the $VAR(1)$ process to be stationary, it must hold that $\Theta^j \to 0$ as $j \to \infty$. From the eigenvalue problem, we have that $\Theta^i c_j = \lambda_j^i c_j$ which means that we can condition stationarity on the eigenvalues lying inside the complex unit circle (eigenvalues are the inverse of the roots). If we have stationarity, we get the properties above.
Slides: We can check for stationarity by calculating the eigenvalues for Θ since $\Theta^i c_j = \lambda_j^i c_j$, where λ_j and c_j are the j'th egeinvalue and eigenvector, respectively, of Θ. Hence, $|\lambda_j| < 1, \quad \forall j$ for the process to be stationary.

2.2 Estimation of Stationary VAR Models

When estimating a stationary VAR, we assume $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$, which means that ϵ_t follows a k-dimensional multivariate normal distribution.
The $VAR(p)$ model conditioned on the first p observations have the following log-likelihood function:

$$\begin{aligned}\ell = &-\frac{(T-p)k}{2}\log(2\pi) - \frac{T-p}{2}\log(det(\Sigma)) \\ &-\frac{1}{2}\sum_{t=p+1}^{T}\left(Y_t - \delta - \sum_{i=1}^{p}\Theta_i Y_{t-i}\right)' \Sigma^{-1}\left(Y_t - \delta - \sum_{i=1}^{p}\Theta_i Y_{t-i}\right)\end{aligned}$$

When estimating the $VAR(p)$ model, we use maximum likelihood based on the log-likelihood function or apply OLS to each of the k equations in (1) and then estimate Σ as

$$\hat{\Sigma} = \frac{1}{T-p}\sum_{t=p+1}^{T} \hat{\epsilon}_t \hat{\epsilon}_t'$$

2.3 Model selection and diagnostic checks

The lag order can be determined by checking :

- Significant coefficients
- Minimizing information criteria
- No serial correlation in the innovations

When specifying the model, it is important to be aware of the curse of dimensionality: the total number of parameters in a $VAR(p)$ model increase by k^2 for each additional lag.
With the assumption $\epsilon_t \sim idd\,(o, \Sigma)$, we need to check that there is no serial correlation and constant variance. If we also assume normally distributed errors, then this also has to be checked.

3 Granger causality tests

The Granger causality test is a test for determining whether one time series is useful in forecasting another. We say that a variable y_t that evolves over time Granger-causes another evolving variable y_t if predictions of the value of y_t based on its own past values and on the past values of x_t are better than predictions of y_t based only on y_t's own past values.
If we look at a bi-variate $VAR(2)$ system:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \theta_{11}^1 & \theta_{12}^1 \\ \theta_{21}^1 & \theta_{22}^1 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \theta_{11}^2 & \theta_{12}^2 \\ \theta_{21}^2 & \theta_{22}^2 \end{pmatrix} \begin{pmatrix} x_{t-2} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{x,t} \\ \epsilon_{y,t} \end{pmatrix}$$

in this case we can say that

- y_t does not Granger-cause x_t when $\theta_{12}^1 = \theta_{12}^2 = 0$
- x_t does not Granger-cause y_t when $\theta_{21}^1 = \theta_{21}^2 = 0$

We can use a Wald test to test for Granger causality.

4 Impulse-response functions

Their main purpose is to describe the evolution of a model's variables in reaction to a shock in one or more variables. Granger-causality may not tell us the complete story about the interactions between the variables of a system.
If we would like to look at the impulse response relationship between two variables we turn to the impulse-response functions. Here, we trace out the effect of an exogenous shock in one of the variables on some or all of the other variables.
If we first look at the univariate $AR(1)$ model which can be rewritten as an $MA(\infty)$ model:

$$\begin{aligned} y_t &= \theta y_{t-1} + \epsilon_t \\ &= \sum_{i=0}^{\infty} \theta^i \epsilon_{t-i} \end{aligned}$$

This implies:

$$\frac{\partial y_t}{\partial \epsilon_{t-i}} = \theta^i$$

We can generalize this to the multivariate case:

$$Y_t = \epsilon_t + \sum_{i=1}^{\infty} \Psi_i \epsilon_{t-i}$$

where Y_t and ϵ_t are $k \times 1$ vectors and Ψ_i is a $k \times k$\$ matrix which has the following interpretation

$$\Psi_i = \frac{\partial Y_{t+i}}{\partial \epsilon_t'} = \begin{bmatrix} \frac{\partial y_{1,t+i}}{\partial \epsilon_{1t}} & \cdots & \frac{\partial y_{1,t+i}}{\partial \epsilon_{kt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{k,t+i}}{\partial \epsilon_{1t}} & \cdots & \frac{\partial y_{k,t+i}}{\partial \epsilon_{kt}} \end{bmatrix}$$

Ψ_i as a function of i is known as the impulse-response function.

- Note:
  - As we assume covariance stationarity (that the eigenvalues are numerically smaller than 1), the marginal impact of the shock will vanish away as we look further into the future.
  - Ψ_i as a function of i is known as the impulse response function
  - On the diagonal is the self-effect.

5 Forecasting in the VAR model

The minimum MSE (Mean squared error) predictor for forecast horizon h at forecast origin t is the conditional expected value: $\hat{Y}_{t+h|t} = E_t\left(Y_{t+h}\right)$. To illustrate, we can look at $VAR\left(1\right)$ process:

$$\mathbf{Y_t} = \delta + \mathbf{\Theta} \mathbf{Y_{t-1}} + \varepsilon_\mathbf{t}$$

For instance, the 2-period forecast will be

$$\begin{aligned}\mathbf{\hat{Y}_{t+2|t}} &= \mathbf{E}\left(\mathbf{Y_{t+2}}\right) \\ &= \mathbf{E}\left(\delta + \mathbf{\Theta} \mathbf{Y_{t+1}} + \varepsilon_\mathbf{t+1}\right) \\ &= \delta + \mathbf{\Theta} \mathbf{E}\left(\mathbf{Y_{t+1}}\right) \\ &= \delta + \mathbf{\Theta}\left(\delta + \mathbf{\Theta} \mathbf{Y_t}\right) \\ &= \left(\mathbf{I_n} + \mathbf{\Theta}\right)\delta + \mathbf{\Theta^2} \mathbf{Y_t}\end{aligned}$$

Hence, our h period afhead forecast will be

$$\hat{Y}_{t+h|t} = \left(I_n + 1 + ... + \Theta^{h-1}\right)\delta + \Theta^h Y_t$$

Taking the limit and using the stationarity condition: $\Theta^j \to 0$ as $j \to \infty$

$$\lim_{h\to\infty} \hat{Y}_{t+h|t} = \left(I_n - \Theta\right)^{-1}\delta$$

That is, when we look at the limit, our forecast equals the unconditional expected value: $E_t\left(Y_{t+h}\right) \to \mu$ for $h \to \infty$

10 - Panel Data Models

January 13, 2022

1 Introduction

There are many advantages of combining the time series and the cross section dimension. Some of the advantages are:

- An increase in the sample size
- Analyze the effects of policy changes
- Analyze the effect of time

Panel data is a time series for the same cross-sectional units in the dataset. Panel data allows us to control for unobserved characteristics, that are constant over time.The basic panel data model for cross sectional unit i $(i = 1, 2, \dots, N)$ is

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, 2, \dots, T$$

where x_{it} is a $K \times 1$ vector of observable explanatory variables, u_{it} is the idiosyncratic error with mean 0, and c_i is time-invariant, independent of u_{it}, and has mean 0 (without loss of generality) and variance σ_c^2.
Strict exogeneity implies that the explanatory variables in each time period are uncorrelated with the error in each time period:

$$Corr\left(x_{is}, u_{it}\right) = 0, \quad s, t = 1, \dots, T$$

If $Corr\left(\boldsymbol{x}_{it}, c_i\right) = 0$ we have a random effects model, while if $Corr\left(x_{it}, c_i\right) \neq 0$ we have a fixed effects model.

2 Fixed effects model

The general idea in fixed effects estimation is to eliminate the unobserved component c_i trough a transformation of the equation.

2.1 Estimation

2.1.1 Fixed effect estimator

We start by removing the "within" mean

$$\bar{y}_{i.} = \bar{x}'_i \cdot \beta + c_i + \bar{u}_i$$

Where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^T y_{it}$, $\overline{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^T \mathbf{x_{it}}$, $\overline{u}_i = \frac{1}{T}\sum_{t=1}^T u_{it}$
Subtracting from the main equation of interest:

$$y_{it} - \bar{y}_{i.} = \left(x_{it} - \bar{x}_{i.}\right)' \beta + \left(u_{it} - \bar{u}_{i.}\right)$$

$$\Leftrightarrow$$

$$\ddot{y}_{it} = \ddot{x}'_{it}\beta + \ddot{u}_{it} \tag{1}$$

The fixed effects estimator, $\widehat{\boldsymbol{\beta}}_{FE}$, is obtained by using pooled OLS on equation (1). We minimize the sum of squared residuals $min_{\beta}\left(\ddot{u}_{it}'\ddot{u}_{it}\right)$

$$\hat{\beta}_{FE}=\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{x}_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{y}_{it}$$

If we have, that:

$$E\left(\ddot{x}_{it}\ddot{u}_{it}\right)=0$$

and that the rank condition is satisfied, then the fixed effects estimator will be consistent. If we also assume homoskedasticity and no serial correlation, the fixed effects estimator is also efficient.

2.1.2 First difference estimator

Another way to estimate the fixed effects model is to use the first difference estimator. This is done by lagging

$$y_{it}=x_{it}'\beta+c_i+u_{it},\quad t=1,2,\ldots,T$$

one period and subtracting, which gives

$$\Delta y_{it}=\Delta x_{it}'\beta+\Delta u_{it},\quad t=2,3,\ldots,T \tag{2}$$

We can then use pooled OLS on equation (2) to obtain $\widehat{\boldsymbol{\beta}}_{FD}$. Similar to before, $\hat{\beta}_{FD}$ is estimated like so:

$$\hat{\beta}_{FE}=\left(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta x_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta y_{it}$$

This estimator is consistent if

$$E\left(\Delta x_{it}\Delta u_{it}\right)=0$$

and if the rank condition is satisfied.

2.1.3 Fixed effect estimator vs first difference estimator

Under the assumption of strict exogeneity and satisfied rank conditions both estimators are consistent, therefore the choice between the two estimators depends on efficiency. For the first difference estimator to be efficient, it implies that u_{it} is a random walk, whereas for the fixed effects estimator to be efficient it implies no serial correlation.

3 Random effects model

In the random effects model we assume homoskedasticity and no serial correlation of the idiosyncratic error. ($Cov\left(x_{it},c_i\right)=0$ and $Cov\left(x_{it},u_{it}\right)=0$
The model can be written as:

$$y_{it}=x_{it}'\beta+v_{it}$$

where $v_{it}=c_i+u_{it}$. Again we use the pooled OLS to estimate $\widehat{\boldsymbol{\beta}}_{OLS}$.

$$\hat{\beta}_{OLS}=\left(\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}x_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}y_{it}$$

The composite error term v_{it} has the following properties:

$$\begin{aligned}E\left(v_{it}\right)&=0\\E\left(v_{it}^2\right)&=\sigma_c^2+\sigma_u^2\\E\left(v_{it}v_{is}\right)&=E\left(c_i^2+c_iu_{it}+c_iu_{it}+u_{it}u_{it}\right)=\sigma_c^2\end{aligned}$$

Which leads to the following variance-covariance matrix:

$$\underset{(T\times T)}{\Omega} = E\left(v_i v_i'\right) = \begin{pmatrix} \sigma_c^2+\sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2+\sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_u^2 \end{pmatrix}$$

$$= \sigma_u^2 I_T + \sigma_c^2 i_T i_T'$$

Note: i_t denotes a $T \times 1$ vector of ones.
This means that the OLS does not provide efficient estimates since our error term is autocorrelated, $E\left[v_{it}v_{is}\right] = \sigma_c^2$. However, since we know the structure of the covariance matrix, Ω, we can consistently and efficiently estimate $\hat{\beta}_{RE}$ by generalized least squares, (GLS). The strategy behind the GLS is as follows:

1. Transform the model by premultiplying with an unknown $T \times T$ matrix $A = \Omega^{-1/2}$

$$\begin{aligned} \underset{(T\times 1)}{y_i} &= \underset{(T\times k)}{x_i^{'}} \underset{(k\times 1)}{\beta} + \underset{(T\times 1)}{v_i} \\ Ay_i &= Ax_i^{'}\beta + Av_i \end{aligned}$$

2. From this new model we require that:

$$\begin{aligned} E(Av_i) &= 0 \\ var(Av_i) &= A\mathit{\Omega} A' = I_T \end{aligned}$$

   which implies that

$$\mathit{\Omega} = (A'A)^{-1}$$

3. The model can now be rewritten as

$$\tilde{y}_i = \tilde{x}_i^{'}\beta + \tilde{v}_i$$

   Where we use the transformed values

$$\tilde{y}_i = Ay_i,\ \tilde{x}_i = x_i A^{'}, \quad \tilde{v}_i = Av_i$$

4. We then estimate the transformed model by OLS, which gives us the GLS estimator, $\boldsymbol{\beta}_{GLS}$.

$$\beta_{\mathbf{GLS}} = \left(\sum_{\mathbf{i=1}}^{\mathbf{N}} \tilde{\mathbf{x}}_{\mathbf{i}}\tilde{\mathbf{x}}_{\mathbf{i}}'\right)^{\mathbf{-1}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \tilde{\mathbf{x}}_{\mathbf{i}}\mathbf{y}_{\mathbf{i}}$$

$$= \left(\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x}_{\mathbf{i}}\mathbf{A}'\mathbf{A}\mathbf{x}_{\mathbf{i}}'\right)^{\mathbf{-1}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x}_{\mathbf{i}}\mathbf{A}'\mathbf{A}\mathbf{y}_{\mathbf{i}}$$

$$= \left(\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x}_{\mathbf{i}}\boldsymbol{\Omega}^{\mathbf{-1}}\mathbf{x}_{\mathbf{i}}'\right)^{\mathbf{-1}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x}_{\mathbf{i}}\boldsymbol{\Omega}^{\mathbf{-1}}\mathbf{y}_{\mathbf{i}}$$

This estimator will be efficient.

3.1 Fixed effect or random effect?

- Use fixed effect estimation if:
  - $Corr\left(x_{it}, c_i\right) \neq 0$ (random effects estimation is inconsistent)
  - Estimates of c_i are of special interest
- Use random effects estimation if
  - $Corr\left(x_{it}, c_i\right) = 0$ (fixed effects estimation is inefficient)
  - Key explanatory variables are constant over time (time-invariant variables cannot be included in fixed effects estimation)
  - Key explanatory variables display very little variation over time (fixed effects estimation can lead to imprecise estimates)

4 Hausman test

We can conduct a Hausman test for random effects. We know that, if $corr\left(x_{it}, c_i\right) = 0$ both random effects and fixed effects estimation are consistent, while if $corr\left(x_{it}, c_i\right) \neq 0$ only fixed effects is consistent. The idea behind the Hausman test is therefore to compare the two set of estimates, where a statistically significant difference is interpreted as evidence against the random effects assumption, $corr\left(x_{it}, c_i\right) = 0$.
Test statistic:

$$H = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)' \left[\hat{V}_{FE} - \hat{V}_{RE}\right]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \sim \chi^2\left(M\right)$$

where $\hat{V}$denotes the estimated covariance matric and $M \leq K$ denotes the number of time-varying explanatory variables.