Modern Basic Econometrics ⊖ Lectures

Introduction to Matrix Approach to Econometrics

Esben Høg

Department of Mathematical Sciences
Aalborg University

18. februar 2016

List of Slides

1. Introduction

- In the following I will treat a number of subjects closely related to the standard multiple regression models in Econometrics.
- These subjects are more or less thoroughly examined in other courses. The major difference is that I almost exclusively use the notation of matrix algebra here. However the first half of the present slide set is perhaps more or less well known to most of you.
- There are several reasons to make yourself well acquainted with this:
  - Linear models are easier to work with in matrix notation.
  - A lot of scientific papers use this notation.
  - Several programming languages are matrix-oriented, which ease the implementation of calculations.

2. The Multiple Regression Model

- The general linear multiple regression model with k **explanatory** variables and n observations is fundamentally a system of n equations:

$$
\begin{aligned}
y_1 &= \beta_1 + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + u_1, \\
y_2 &= \beta_1 + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + u_2, \\
y_3 &= \beta_1 + \beta_2 x_{32} + \ldots + \beta_k x_{3k} + u_3, \\
&\vdots \\
y_n &= \beta_1 + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + u_n.
\end{aligned}
$$

- with $u_1, \ldots, u_n \sim \text{iid} - \mathcal{N}(0, \sigma^2)$.

- If we construct the following vectors and matrices

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

- and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},$$

- the regression model can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\Leftrightarrow \quad \underset{n\times 1}{\boldsymbol{y}} = \underset{n\times k}{\boldsymbol{X}} \underset{k\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{u}}, \quad \boldsymbol{u} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

- This is perhaps known from the theory of linear normal models.

3. The Multivariate Normal Distribution

We introduce some concepts from multivariate normal (Gaussian) distributions.

$\boldsymbol{y}$ is Multivariate Normal Distributed with mean vector $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$.

$\Downarrow$

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

$\Downarrow$

Multivariate density for $\boldsymbol{y}$ is

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right).$$

Mean vector:

Covariance matrix:

4. Conditioning a multivariate normal distribution

Partitioned vectors and covariance matrix

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the distribution of $\boldsymbol{y}_2$ conditional on $\boldsymbol{y}_1$ is multivariate normal with

$$\boldsymbol{y}_2 \,|\, \boldsymbol{y}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1)\,,\, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

In particular this is important if $\boldsymbol{y}$ is a time series, and for example $\boldsymbol{y}_2$ corresponds to a future period and $\boldsymbol{y}_1$ corresponds to the present. Then under normality we have a conditional normal distribution for the future observations given the present observations.

5. The Bivariate Normal Distribution

Special case: The bivariate case, that is $\boldsymbol{y}_1 = y_1$ and $\boldsymbol{y}_2 = y_2$ are both scalars,

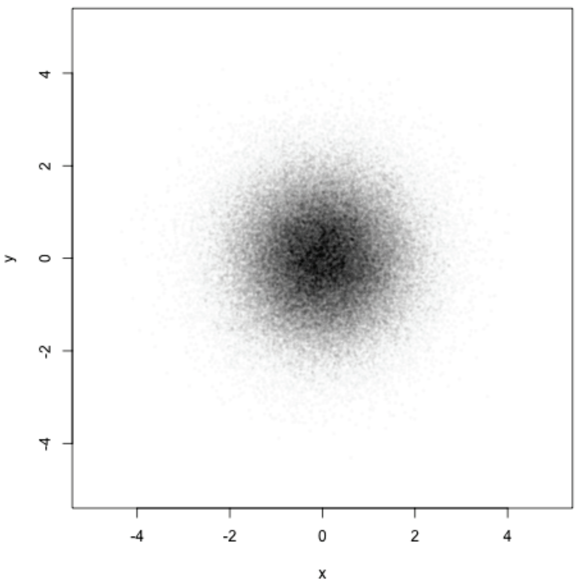
$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \; \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \; \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Under normality y_2 is conditionally normal given y_1:

$$y_2 \mid y_1 \sim \mathcal{N}\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(y_1 - \mu_1)\,,\, (1-\rho^2)\sigma_2^2\right),$$

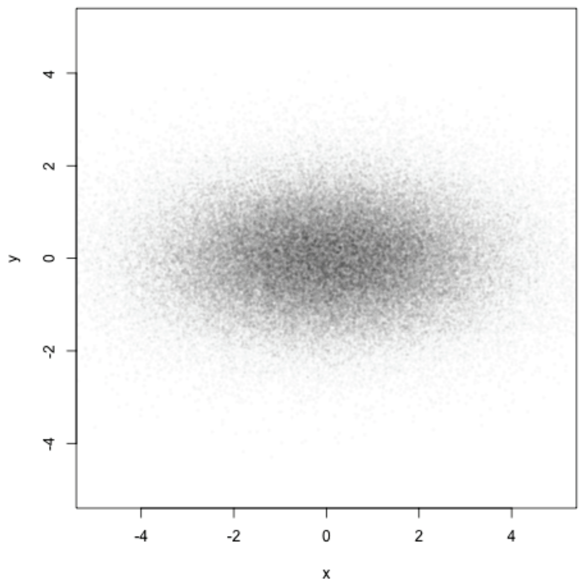
where ρ is the correlation coefficient between y_1 and y_2.

➤ Examples for (x, y) bivariate normal distributed (no. 1).



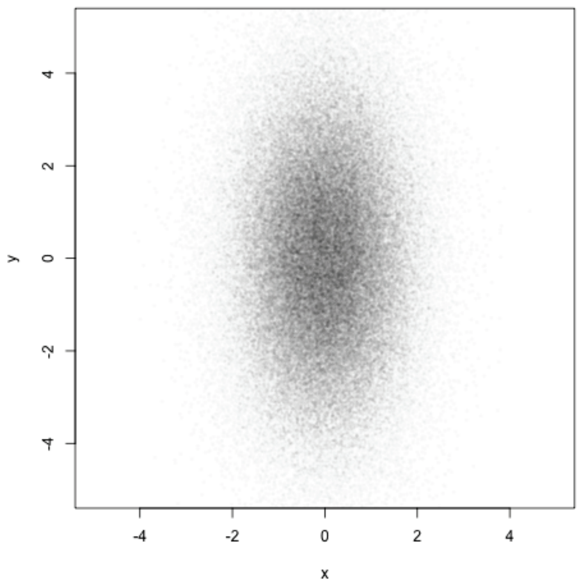
$X \sim \mathcal{N}(0,1),\ \ Y \sim \mathcal{N}(0,1)$

$\rho = 0$



$X \sim \mathcal{N}(0,2),\ \ Y \sim \mathcal{N}(0,1)$

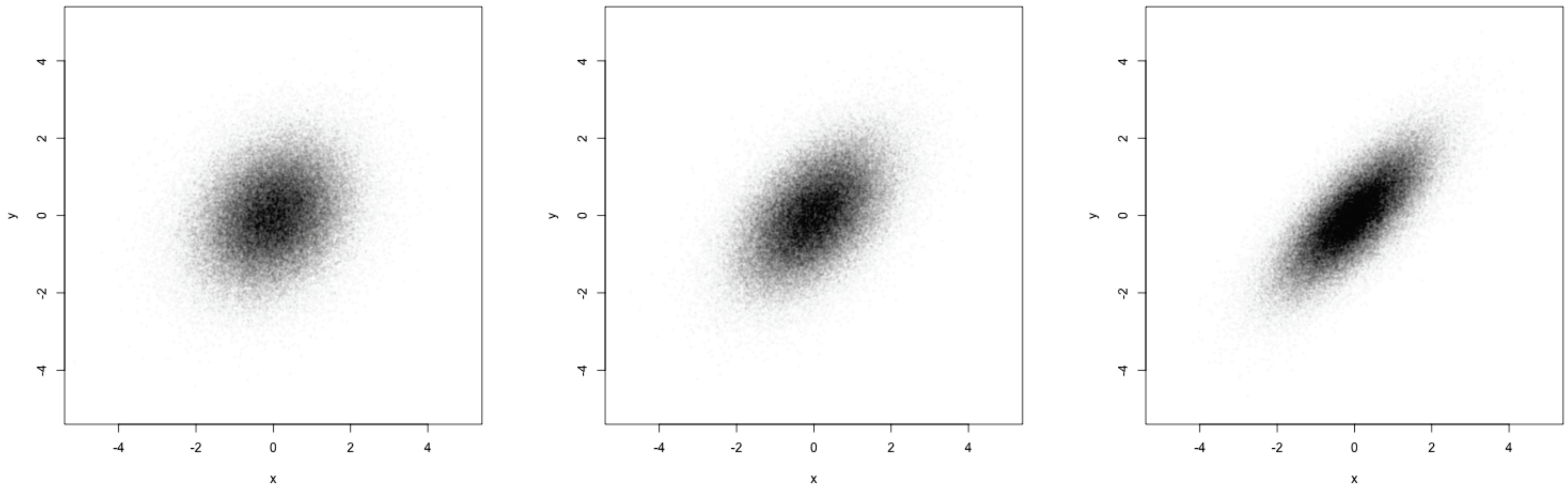
$\rho = 0$



$X \sim \mathcal{N}(0,1),\ \ Y \sim \mathcal{N}(0,2)$

$\rho = 0$

➤ Examples for (x, y) bivariate normal distributed (no. 2).

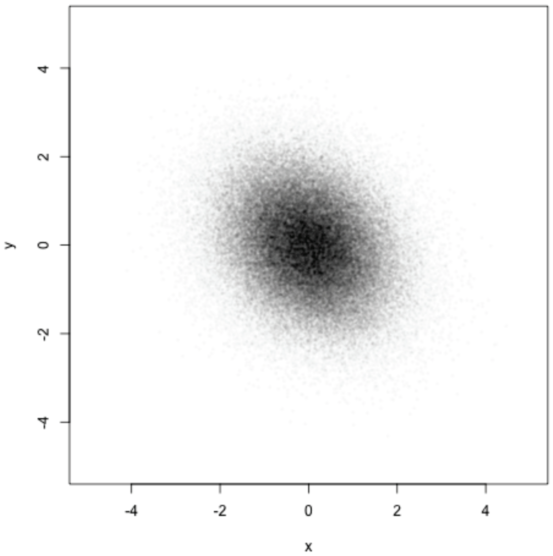


$X \sim \mathcal{N}(0,1), \; Y \sim \mathcal{N}(0,1)$ $\rho = 0.25$

$X \sim \mathcal{N}(0,1), \; Y \sim \mathcal{N}(0,1)$ $\rho = 0.5$

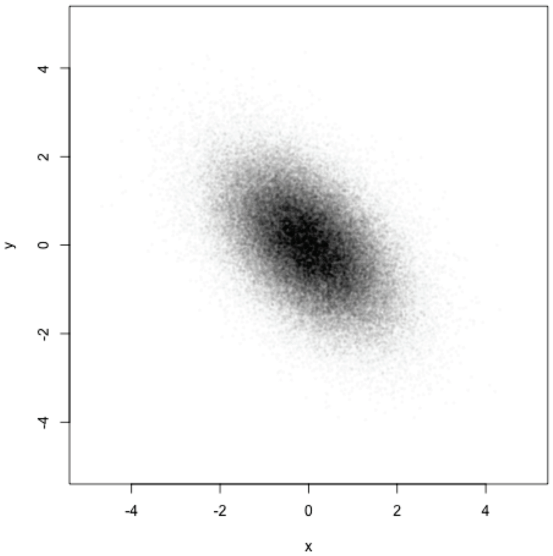
$X \sim \mathcal{N}(0,1), \; Y \sim \mathcal{N}(0,1)$ $\rho = 0.75$

- Examples for (x, y) bivariate normal distributed (no. 3).



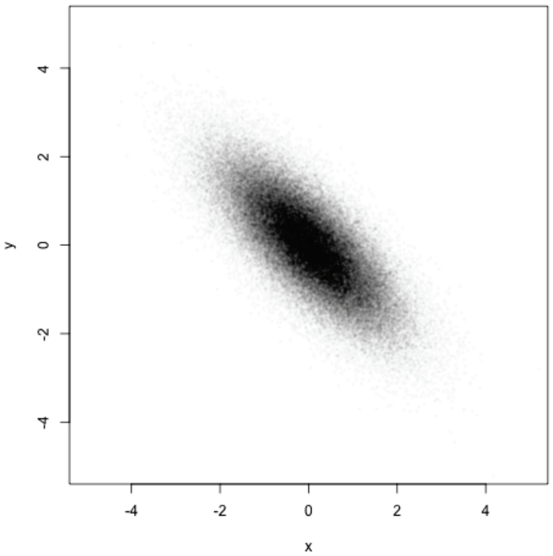
$X \sim \mathcal{N}(0,1),\;\; Y \sim \mathcal{N}(0,1)$

$\rho = -0.25$



$X \sim \mathcal{N}(0,1),\;\; Y \sim \mathcal{N}(0,1)$

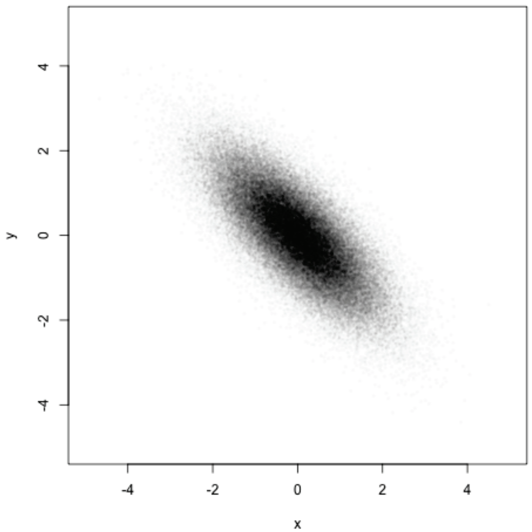
$\rho = -0.5$



$X \sim \mathcal{N}(0,1),\;\; Y \sim \mathcal{N}(0,1)$

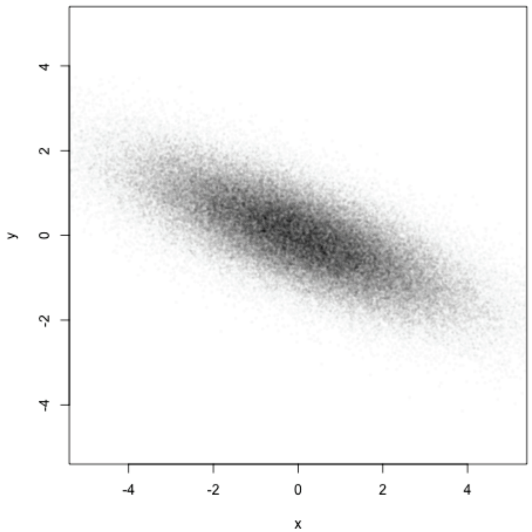
$\rho = -0.75$

- Examples for (x, y) bivariate normal distributed (no. 4).



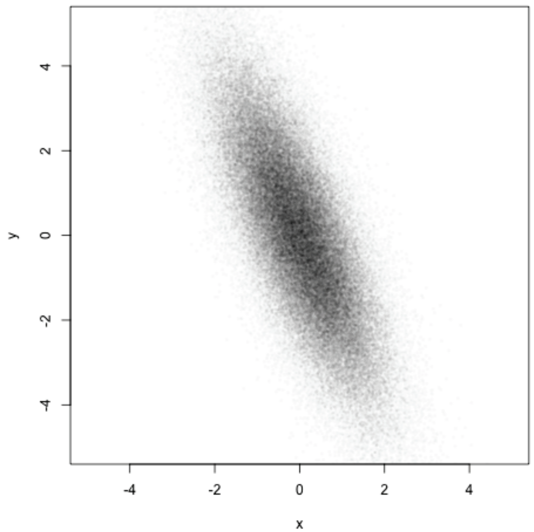
$X \sim \mathcal{N}(0,1),\ Y \sim \mathcal{N}(0,1)$

$\rho = -0.75$



$X \sim \mathcal{N}(0,2),\ Y \sim \mathcal{N}(0,1)$

$\rho = -0.75$



$X \sim \mathcal{N}(0,1),\ Y \sim \mathcal{N}(0,2)$

$\rho = -0.75$

Classical Assumptions in Econometrics:

A1. zero conditional mean: $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$

A2. No perfect collinearity: $\text{rank}(\boldsymbol{X}) = k \Leftrightarrow |\boldsymbol{X}^\top \boldsymbol{X}| \neq 0$

A3 Homoskedasticity and no serial correlation $\text{var}(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$

A4. Normality of errors: $\boldsymbol{u} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$

where:

$$\sigma^2 \boldsymbol{I}_n = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

The case of a General Covariance Matrix

- The general case where we may have **heteroscedasticity** and/or auto-correlation corresponds to cases where the covariance matrix of $\boldsymbol{u}$ is no longer $\sigma^2 \boldsymbol{I}_n$. Instead the covariance matrix of $\boldsymbol{u}$ has the following general form:

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^\top) \quad &= \begin{bmatrix} \mathrm{var}(u_1) & \mathrm{cov}(u_1 u_2) & \dots & \mathrm{cov}(u_1 u_n) \\ \mathrm{cov}(u_2 u_1) & \mathrm{var}(u_2) & \dots & \mathrm{cov}(u_2 u_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(u_n u_1) & \mathrm{cov}(u_n u_2) & \dots & \mathrm{var}(u_n) \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} = \boldsymbol{\Sigma}.
\end{aligned}
$$

Example: Heteroscedasticity

- For so-called heteroscedasticity in the form:

$$\mathrm{var}(u_i) = \sigma_i^2 = \sigma^2 z_i^2 \quad \text{for } i = 1, \ldots, n,$$

and where $z_1, \ldots, z_n$ are some known values,

- the "General Covariance Matrix" then has the form

$$\mathrm{var}(\boldsymbol{u}) = \boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} z_1 & 0 & \ldots & 0 \\ 0 & z_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & z_n \end{bmatrix}.$$

Example: Autocorrelation

- For 1st order autocorrelation in the form:

$$u_t = \rho u_{t-1} + \epsilon_t, \quad \epsilon_t \sim nid(0, \sigma_\epsilon^2) \quad \text{for } t = 1, \ldots, T,$$

- it turns out that the "General Covariance Matrix" has the form

$$\mathsf{var}(\boldsymbol{u}) = \boldsymbol{\Sigma} = \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \ldots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \ldots & \rho^{T-3} & \rho^{t-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \ldots & \rho & 1 \end{bmatrix}.$$

- Exercise: Show that $\boldsymbol{\Sigma}$ indeed has this form.

6. Least Squares and differentiation w.r.t. a vector

- The first task is to estimate the regression parameters: $\beta_1, \ldots, \beta_k$ by OLS.
- OLS minimizes the sum of squared residuals:

$$\min_{\beta_1,\beta_2,\ldots,\beta_k} \sum_{i=1}^{n} u_i^2 = \min_{\beta_1,\beta_2,\ldots,\beta_k} Q(\beta_1, \beta_2, \ldots, \beta_k) = \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}).$$

- We interpret the sum of squared errors as a function of the unknown parameter vector $\boldsymbol{\beta}$. By minimizing $Q(\boldsymbol{\beta})$ wrt. $\boldsymbol{\beta}$ we find the OLS estimate $\hat{\boldsymbol{\beta}}$.

- To do that we may differentiate $Q(\boldsymbol{\beta})$ wrt. $\boldsymbol{\beta}$, put the 1st order derivatives equal to zero and solve the resulting k equations which we call the **normal equations**.

- The second order differentiation w.r.t. $\boldsymbol{\beta}$ produces a square matrix (a Hessian) which should be positive definite in order for the normal equations to define a minimum.

- We use the notation: $\boldsymbol{u}^\top = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$, where $\underset{n\times 1}{\boldsymbol{u}}$ is the column vector of residuals.
- Obviously by using the definition of inner product

$$\begin{aligned} \underset{1\times n}{\boldsymbol{u}^\top} \underset{n\times 1}{\boldsymbol{u}} &= u_1u_1 + u_2u_2 + \cdots u_nu_n \\ &= \textstyle\sum_{i=1}^n u_i^2 = Q(\boldsymbol{\beta}). \end{aligned}$$

- which motivates the need to calculate a derivative like

$$\frac{\partial \boldsymbol{u}^\top \boldsymbol{u}}{\partial \boldsymbol{\beta}}.$$

- By definition: $\underset{n\times 1}{\boldsymbol{u}} = \underset{n\times 1}{\boldsymbol{y}} - \underset{n\times k}{\boldsymbol{X}}\underset{k\times 1}{\boldsymbol{\beta}}$. If we insert this we obtain

$$\begin{aligned}
Q(\boldsymbol{\beta}) = \boldsymbol{u}^\top \boldsymbol{u} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta},
\end{aligned}$$

- and since $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{y}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} - \frac{\partial 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$,
- we obviously have to know how to calculate these three scalar functions of $\boldsymbol{\beta}$:

$$\frac{\partial \boldsymbol{y}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}}, \qquad \frac{\partial 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} \quad \text{and} \quad \frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}}.$$

6.1. Some formulas of Matrix Differentiation

- Consider the case where we want to differentiate a scalar function of a vector, $f(\boldsymbol{\delta})$, with respect to that very same vector $\underset{k\times 1}{\boldsymbol{\delta}}$.
- Then the rule is:

$$\frac{\partial f(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{\delta})}{\partial \delta_1} \\ \frac{\partial f(\boldsymbol{\delta})}{\partial \delta_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{\delta})}{\partial \delta_k} \end{bmatrix}.$$

- From the previous we have three "relevant" forms of scalar functions of $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y}, \qquad f(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} \quad \text{and} \quad f(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

- I.e. a constant, a linear function, and a quadratic form.

1. The constant function is

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y}.$$

Therefore we have

$$\frac{\partial \boldsymbol{y}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial\left(\sum_{i=1}^{n} y_i^2\right)}{\partial \beta_1} \\ \frac{\partial\left(\sum_{i=1}^{n} y_i^2\right)}{\partial \beta_2} \\ \vdots \\ \frac{\partial\left(\sum_{i=1}^{n} y_i^2\right)}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \underset{k \times 1}{\mathbf{0}}.$$

2. The linear function is
$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{y} = \boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta}.$$

This is because $\underset{k\times n}{\mathbf{X}^\top}\underset{n\times 1}{\boldsymbol{y}}$ is a $k \times 1$ vector (call it $\boldsymbol{a}$ for example), we have with $\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}^\top$

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{a} = \beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_k a_k = \boldsymbol{a}^\top \boldsymbol{\beta}.$$

So

$$\frac{\partial(\boldsymbol{\beta}^\top \boldsymbol{a})}{\partial \boldsymbol{\beta}} = \frac{\partial(\boldsymbol{a}^\top \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial(\beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_k a_k)}{\partial \beta_1} \\ \frac{\partial(\beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_k a_k)}{\partial \beta_2} \\ \vdots \\ \frac{\partial(\beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_k a_k)}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \boldsymbol{a} = \mathbf{X}^\top \boldsymbol{y}.$$

3. The quadratic form is

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

$\underset{k\times n}{\boldsymbol{X}^\top}\underset{n\times k}{\boldsymbol{X}}$ is a symmetric $k \times k$ matrix. Call it $\boldsymbol{A}$. Then we have

$$\begin{aligned} f(\boldsymbol{\beta}) &= \underset{1\times k}{\boldsymbol{\beta}^\top}\underset{k\times k}{\boldsymbol{A}}\underset{k\times 1}{\boldsymbol{\beta}} \\ &= \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \\ a_{1k} & a_{2k} & \dots & a_{kk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}. \end{aligned}$$

This derivative in this case is best understood via a simple illustration.

Therefore, let us calculate

$$f(\boldsymbol{\beta}) = \underset{1\times 3}{\boldsymbol{\beta}^\top}\underset{3\times 3}{\boldsymbol{A}}\underset{3\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}\beta_1 + a_{12}\beta_2 + a_{13}\beta_3 & a_{12}\beta_1 + a_{22}\beta_2 + a_{23}\beta_3 & a_{13}\beta_1 + a_{23}\beta_2 + a_{33}\beta_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$= a_{11}\beta_1^2 + a_{12}\beta_1\beta_2 + a_{13}\beta_1\beta_3 + a_{12}\beta_2\beta_1 + a_{22}\beta_2^2 + a_{23}\beta_2\beta_3 + a_{13}\beta_3\beta_1 + a_{23}\beta_3\beta_2 + a_{33}\beta_3^2$$

$$= a_{11}\beta_1^2 + a_{22}\beta_2^2 + a_{33}\beta_3^2 + 2a_{12}\beta_1\beta_2 + 2a_{13}\beta_1\beta_3 + 2a_{23}\beta_3\beta_2.$$

So,

$$\frac{\partial \boldsymbol{\beta}^\top \boldsymbol{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_2} \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_3} \end{bmatrix} = \begin{bmatrix} 2(a_{11}\beta_1 + a_{12}\beta_2 + a_{13}\beta_3) \\ 2(a_{22}\beta_2 + a_{12}\beta_1 + a_{23}\beta_3) \\ 2(a_{33}\beta_3 + a_{13}\beta_1 + a_{23}\beta_2) \end{bmatrix}$$

$$= \begin{bmatrix} 2(a_{11}\beta_1 + a_{12}\beta_2 + a_{13}\beta_3) \\ 2(a_{12}\beta_1 + a_{22}\beta_2 + a_{23}\beta_3) \\ 2(a_{13}\beta_1 + a_{23}\beta_2 + a_{33}\beta_3) \end{bmatrix}$$

$$= 2\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 2\boldsymbol{A}\boldsymbol{\beta}.$$

Hence $\frac{\partial(\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}$.

- So much tedious algebra to convince ourselves that

$$\frac{\partial \boldsymbol{y}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} = \mathbf{0}, \qquad \frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y}, \qquad \text{and} \qquad \frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

- Hence we must have:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{y}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} - 2\frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

- And therefore the normal equations are

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \Leftrightarrow \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}.$$

- There exist three **equivalent** necessary and sufficient conditions that ensure that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible:

  1. The columns in $\boldsymbol{X}$ are linearly independent.
  2. $\boldsymbol{X}^\top\boldsymbol{X}$ has full rank, that is $\rho\left(\boldsymbol{X}^\top\boldsymbol{X}\right) = k$.
  3. $\left|\boldsymbol{X}^\top\boldsymbol{X}\right| \neq 0$.

- If these conditions are not met we are faced with so–called perfect collinearity.

- If however the three conditions are fulfilled $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$ is well defined and the OLS estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

6.2. The distribution of $\hat{\boldsymbol{\beta}}$

- Note that $\hat{\boldsymbol{\beta}}$ is also the MLE of $\boldsymbol{\beta}$.
- Given $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$. From the theory of linear normal models we then know that $\hat{\boldsymbol{\beta}}$ is normally distributed.
- It is easily shown that

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta}, \text{ therefore } \hat{\boldsymbol{\beta}} \text{ is an unbiased estimator of } \boldsymbol{\beta}.$$

- Furthermore it is easily shown that the variance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\operatorname{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

- Hence $\hat{\boldsymbol{\beta}}$ is multivariate normal distributed according to

$$\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}).$$

- This fact is used to test linear hypotheses on $\boldsymbol{\beta}$.
- We estimate σ^2 by the unbiased estimator s^2:

$$s^2 = \frac{(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k} = \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{n-k}.$$

- Then

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{x^{jj}}} \sim t(n-k),$$

where x^{jj} is the jth diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

6.3. The Gauss-Markov theorem

- This theorem states that the OLS estimate $\hat{\boldsymbol{\beta}}$ is BLUE, *(Best Linear Unbiased Estimator)*.

BLUE:

1. Unbiased: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.
2. Efficient (min. variance): $\text{var}(\hat{\boldsymbol{\beta}})$ smallest possible.
3. Linear function of the y's.

2. Says that the estimate has the smallest[a] variance among all possible unbiased linear estimators.

[a] A square matrix $\boldsymbol{A}$ is here defined to be smaller than another square matrix $\boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive definite.

7. The decomposition of sums of squares

- We can construct the usual decomposition of the variation of $\boldsymbol{y}$, since

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}},$$

where $\hat{\boldsymbol{u}}$ is the residual vector. Then

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{y} &= (\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}})^\top (\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}) \\ &= \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \hat{\boldsymbol{u}} + \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}. \end{aligned}$$

- Since
$$\begin{aligned}\hat{\beta}^\top X^\top \hat{u} &= \hat{\beta}^\top X^\top (y - X\hat{\beta}) \\ &= ((X^\top X)^{-1} X^\top y)^\top X^\top (y - X(X^\top X)^{-1} X^\top y) \\ &= y^\top X(X^\top X)^{-1} X^\top (y - X(X^\top X)^{-1} X^\top y) \\ &= y^\top X(X^\top X)^{-1} X^\top y - y^\top X(X^\top X)^{-1} X^\top X(X^\top X)^{-1} X^\top y \\ &= y^\top X(X^\top X)^{-1} X^\top y - y^\top X(X^\top X)^{-1} X^\top y = 0,\end{aligned}$$

- we have
$$y^\top y = \hat{\beta}^\top X^\top X\hat{\beta} + \hat{u}^\top \hat{u}.$$

- This only provides us with the sum of squares, but what we want is the squared deviations from averages. This is obtained by subtracting $n\bar{y}^2 = \frac{1}{n}(\boldsymbol{\iota}^\top \boldsymbol{y})^2$ on both sides:

$$\boldsymbol{y}^\top \boldsymbol{y} - \frac{1}{n}(\boldsymbol{\iota}^\top \boldsymbol{y})^2 = \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} - \frac{1}{n}(\boldsymbol{\iota}^\top \boldsymbol{y})^2 + \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$$

Total variation	=	Explained variation	+	Unexplained variation
SS Total		SS Explained		SS Residuals

Here:

SSE: The part of the variation in y, which is explained by $\boldsymbol{X}$. *SSR*: The part of the variation in y, which is not explained by the model.

- The coefficient of determination is now defined as:

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

- and the adjusted coefficient of determination

$$\boxed{\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-k}(1 - \mathrm{R}^2)}$$

- Note the intuition in the definition of $\mathrm{R}^2_{\mathrm{adj}}$: Mean sums of squares are used instead of sums of squares.

8. Geometric interpretation of OLS

- Given the OLS estimators $\hat{\boldsymbol{\beta}}$, the *predicted* value of $\boldsymbol{y}$ is defined by

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\boldsymbol{P}}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}.$$

- The residuals are defined as

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} = \underbrace{[\boldsymbol{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\boldsymbol{M}}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{u},$$

- Where

$$\boldsymbol{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \quad \text{and} \quad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$$

are the so-called fundamental OLS matrices.

- They are **idempotent** (and symmetric and hence by definition projection matrices) and satisfy the equations $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$, and

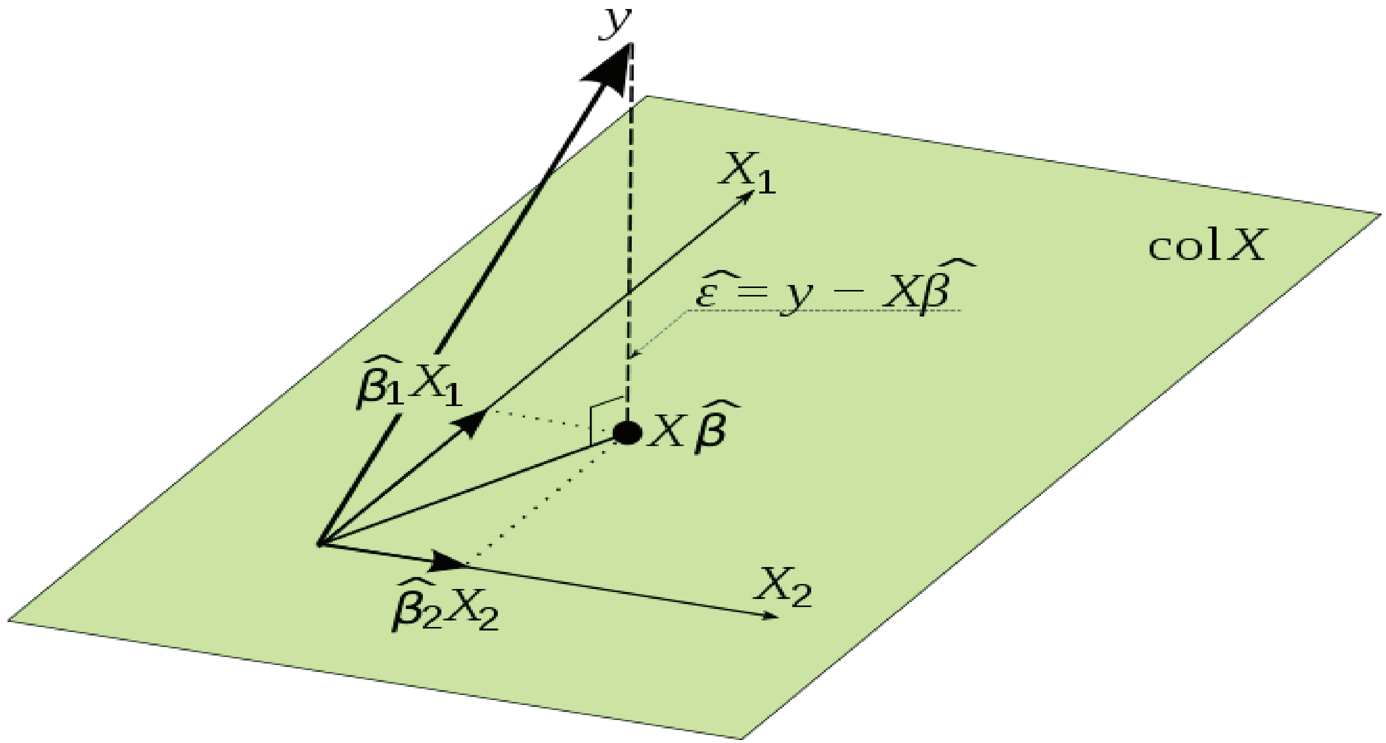
$$\begin{aligned} \hat{\boldsymbol{y}}^\top \hat{\boldsymbol{u}} = (\boldsymbol{P}\boldsymbol{y})^\top [\boldsymbol{M}\boldsymbol{y}] &= \boldsymbol{y}^\top \boldsymbol{P}^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\ &= \boldsymbol{y}^\top (\boldsymbol{P}^\top - \boldsymbol{P}^2)\boldsymbol{y} \\ &= \boldsymbol{y}^\top (\boldsymbol{P} - \boldsymbol{P})\boldsymbol{y} \\ &= 0, \end{aligned}$$

  which shows that the residuals and the predicted responses are orthogonal.

- Further it can be seen that

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}} \\ &= \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y}. \end{aligned}$$

- The projection related to OLS:



- The previous and the fact that both $\boldsymbol{P}$ and $\boldsymbol{M}$ are projection matrices with $\boldsymbol{P}+\boldsymbol{M} = \boldsymbol{I}$, show that OLS geometrically can be interpreted as a decomposition of $\boldsymbol{y}$ in two orthogonal components.
  - $\boldsymbol{P}$ projects $\boldsymbol{y}$ on the (*hyper*-)plane , which is spanned by the columns in $\boldsymbol{X}$, and
  - $\boldsymbol{M}$ projects $\boldsymbol{y}$ on the (*hyper*-)plane, which is orthogonal to the first (*hyper*-)plane.
- Thus, $\boldsymbol{y}$ may be seen as a sum of two orthogonal vectors, that constitute the sides in a right-angled triangle enclosing the right angle.

9. Linear Restrictions

- We want to apply some non linear restrictions to the model:

$$
\begin{aligned}
r_{11}\beta_1 + r_{12}\beta_2 + \ldots + r_{1k}\beta_k &= r_1 \\
r_{21}\beta_1 + r_{22}\beta_2 + \ldots + r_{2k}\beta_k &= r_2 \\
&\vdots \\
r_{q1}\beta_1 + r_{q2}\beta_2 + \ldots + r_{qk}\beta_k &= r_q.
\end{aligned}
$$

- In matrix notation this is written $\underset{q\times k}{\boldsymbol{R}}\ \underset{k\times 1}{\boldsymbol{\beta}} = \underset{q\times 1}{\boldsymbol{r}}$, where the elements of $\boldsymbol{R}$ gives the restrictions.

9.1. Estimation

- Vi want to estimate the model under such restrictions.
- Each row in $\boldsymbol{R}$ is a linear restriction of the coefficient vector.
- Typically $\boldsymbol{R}$ will have few rows and many zeros in each row.
- **Examples**:
  - One of the coefficients is 0, $\beta_j = 0$:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{r} = 0.$$

  - Two of the coefficients are equal, $\beta_k = \beta_j$:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & \dots & -1 & \dots & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{r} = 0.$$

➤ More examples:

- Some of the coefficients add up to 1, $\beta_2 + \beta_3 + \beta_4 = 1$:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & \dots & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{r} = 1.$$

- Some of the coefficients are 0, $\beta_1 = \beta_2 = \beta_4 = 0$:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

- To estimate the model under the restrictions we set out the Lagrange function to minimizing:

$$L(\boldsymbol{\lambda},\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X\beta})^{\top}(\boldsymbol{y}-\boldsymbol{X\beta}) + \boldsymbol{\lambda}^{\top}(\boldsymbol{R\beta}-\boldsymbol{r})$$

$$= \boldsymbol{y}^{\top}\boldsymbol{y} - 2\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{y} + \boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X\beta} + \boldsymbol{\lambda}^{\top}(\boldsymbol{R\beta}-\boldsymbol{r}).$$

- The first order conditions are

$$\frac{\partial L(\boldsymbol{\lambda},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^{\top}\boldsymbol{y} + 2\boldsymbol{X}^{\top}\boldsymbol{X\beta} + \boldsymbol{R}^{\top}\boldsymbol{\lambda} = 0,$$

$$\frac{\partial L(\boldsymbol{\lambda},\boldsymbol{\beta})}{\partial \boldsymbol{\lambda}} = \boldsymbol{R\beta} - \boldsymbol{r} = 0.$$

- Let $\hat{\hat{\boldsymbol{\beta}}}$ denote the estimate of $\boldsymbol{\beta}$ under the restrictions.

- From the first set of restrictions we get

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{2}\boldsymbol{R}^\top \boldsymbol{\lambda}).$$

- Inserting in the last set of restrictions reveals

$$\boldsymbol{r} = \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{2}\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top \boldsymbol{\lambda},$$

- which again may be solved for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = 2\left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{r}\right].$$

- This can be used to find $\hat{\hat{\boldsymbol{\beta}}}$:

$$
\begin{aligned}
\hat{\hat{\boldsymbol{\beta}}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{y} - \frac{1}{2}\boldsymbol{R}^\top \overbrace{2\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{r}\right]}^{\boldsymbol{\lambda}}) \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{r}\right] \\
&= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\left[\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right].
\end{aligned}
$$

- It is worth noting that if $\hat{\boldsymbol{\beta}}$ fulfills the restrictions then

$$
\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r} = 0 \qquad \text{which implies} \qquad \hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}.
$$

9.2. Test linear restrictions

- We will see how to test if the restrictions are valid.
- Thus we want to test the null hypothesis: $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$.
- This is done under assumption A4 via the following F-statistic

$$F = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top [\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}/(n-k)} \sim F(q, n-k)$$

- If A4 is not fulfilled, then we must use the so–called Wald statistic:

$$W = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top [\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}/(n-k)} \sim \chi^2(q) \text{ (approximately}$$

- Note that the F-statistic is exactly equal to:

$$F = \frac{(\hat{\boldsymbol{u}}_*^\top \hat{\boldsymbol{u}}_* - \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}})/q}{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}/(n-k)} \sim F(q, n-k),$$

- where

  $\hat{\boldsymbol{u}}_*^\top \hat{\boldsymbol{u}}_*$ is the sum of the squared residuals from the restricted model, and

  $\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$ is the sum of squared residuals from the unrestricted model.

- A nice feature of using the matrix version on the previous slide is that you only need the estimates from one regression, namely the unrestricted one.

9.3. The Chow Forecast test

- The idea is that, if the parameter vector is constant, then we have a specific confidence that *out-of-sample* predictions will fall within specified bounds.
  - these bounds are the well-known prediction intervals computed from sample data.
- Hence, large prediction errors will be critical for the constancy hypothesis.

- To test this hypothesis we split the sample in two.
  - Use n_1 observations for estimation, and
  - use $n_2 = n - n_1$ for testing.

  1. For time series we would usually take the first n_1 for estimation.
  2. In cross-sections we could e.g. split the sample according to a size variable.

- It will be appropriate to reserve 5-15 percent of the observations for testing.

- In splitting the sample, we use the following notation:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \begin{array}{l} \leftarrow \quad n_1 \times 1 \\ \leftarrow \quad n_2 \times 1 \end{array} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \begin{array}{l} \leftarrow \quad n_1 \times k \\ \leftarrow \quad n_2 \times k \end{array}$$

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} \begin{array}{l} \leftarrow \quad n_1 \times 1 \\ \leftarrow \quad n_2 \times 1 \end{array}$$

- The complete model can then be written as

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{u}_1, \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{\alpha} + \boldsymbol{u}_2. \end{aligned}$$

- The null hypothesis is

$$\mathrm{H}_0 : \boldsymbol{\alpha} = \boldsymbol{\beta}.$$

- To use a dummy variable approach write the second equation as

$$\begin{aligned} \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{\alpha} + \boldsymbol{u}_2 = \boldsymbol{X}_2\boldsymbol{\alpha} + \boldsymbol{u}_2 + \boldsymbol{X}_2\boldsymbol{\beta} - \boldsymbol{X}_2\boldsymbol{\beta} \\ &= \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{X}_2(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \boldsymbol{u}_2 \\ &= \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{u}_2. \end{aligned}$$

- Hence if $\boldsymbol{\gamma} = \mathbf{0}$ then $\boldsymbol{\alpha} = \boldsymbol{\beta}$.
- Note that $\boldsymbol{X}_2(\boldsymbol{\alpha} - \boldsymbol{\beta})$ is a $n_2 \times 1$ "parameter-vector".
- Now the model may be written compactly as

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} \\ \boldsymbol{X}_2 & \boldsymbol{I}_{n_2} \end{bmatrix}}_{\boldsymbol{Z}} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}.$$

- Now it is simple to estimate $\begin{bmatrix} \boldsymbol{\beta} & \boldsymbol{\gamma} \end{bmatrix}^\top$:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \boldsymbol{y}$$

$$= \left(\begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \boldsymbol{I}_{n_2} \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \boldsymbol{I}_{n_2} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \boldsymbol{I}_{n_2} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \boldsymbol{y}_1 \\ \boldsymbol{y}_2 - \mathbf{X}_2(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \boldsymbol{y}_1 \end{bmatrix}.$$

- Hence $\hat{\boldsymbol{\beta}}$ estimate $\boldsymbol{\beta}$ using n_1 data points.

- Noting that

$$\boldsymbol{X}_2(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{y}_1 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}_2,$$

we discover that $\hat{\boldsymbol{\gamma}} = \boldsymbol{y}_2 - \hat{\boldsymbol{y}}_2$ is the prediction errors, when we are trying to predict $\boldsymbol{y}_2$ using an estimate of $\boldsymbol{\beta}$ based on n_1 observations.

- Hence to test for parameter constancy we test

$$\mathsf{H}_0 : \boldsymbol{\gamma} = \boldsymbol{0},$$

using our test for linear restrictions to the model with the dummy variable $\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_{n_2} \end{bmatrix}^\top$.

- The restriction matrices are:

$$\boldsymbol{R} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n_2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{r} = \begin{bmatrix} \mathbf{0} \end{bmatrix}.$$

- Hence, the test is:

$$F = \frac{\left(\boldsymbol{R}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix}\right)^\top [(\boldsymbol{R}\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{R}^\top]^{-1}\left(\boldsymbol{R}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix}\right)/n_2}{\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}}/\underbrace{(n_1-k)}_{=n_1+n_2-(k+n_2)}} \sim F(n_2, n_1-k).$$

- If A4 does not apply, then use:

$$W = \frac{(\boldsymbol{R}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix})^\top [(\boldsymbol{R}\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix})}{\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}}/(n_1 - k)} \sim \chi^2(n_2).$$

9.4. Chow's Breakpoint Test

- If the forecast subset is large enough it might be better to estimate two regression functions, one for each sub-sample, and then test for common parameters.
- Then the unrestricted model may be written:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix},$$

  where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the k-vectors of the two sub-samples.
- The null hypothesis of no structural break is then

$$\mathrm{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2.$$

- Again, use the test for linear restrictions to this model. We just have to set up the restriction matrices:

$$\boldsymbol{R} = \left[\begin{array}{cc} \boldsymbol{I}_k & -\boldsymbol{I}_k \end{array} \right] \quad \text{and} \quad \boldsymbol{r} = \left[\begin{array}{c} \boldsymbol{0} \end{array} \right].$$

- Hence, the test is:

$$F = \frac{(\boldsymbol{R} \left[\begin{array}{c} \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \right])^\top [(\boldsymbol{R}\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{R}^\top]^{-1} (\boldsymbol{R} \left[\begin{array}{c} \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \right])/k}{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}/(n-2k)} \sim F(k, n-2k).$$

- If A4 does not apply, then use the chi-squared alternative (the Wald test).

- It is not hard to modify this test to relax the restriction on the intercept term, or on the slope.

10. Heteroskedasticity Defined

- A general *Heteroskedastic* model violates A3.
- It can be presented as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}).$$

where

$$\begin{aligned}\sigma^2\boldsymbol{\Sigma} &= var(\boldsymbol{u}) = \\ &= E(\boldsymbol{u}\boldsymbol{u}^\top) = \end{aligned}\begin{bmatrix} var(u_1) & cov(u_1, u_2) & \dots & cov(u_1, u_n) \\ cov(u_2, u_1) & var(u_2) & \dots & cov(u_2, u_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_n, u_1) & cov(u_n, u_2) & \dots & var(u_n) \end{bmatrix}.$$

11. OLS with Heteroskedastic Errors

- For a general covariance matrix the following results hold:

1. The OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is unbiased and consistent.

2. The OLS estimator is inefficient.

3. The message is the following:
   - We can get a fair estimate of the parameter vector using OLS, even when A3 is not satisfied, but
   - We can not perform hypothesis testing on the parameter vector, because the standard errors estimated using OLS are WRONG (i.e. biased and inconsistent).

- The correct variance matrix of the OLS estimator is

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top|\boldsymbol{X}\right] \\
&= E\left[((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u})((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u})^\top|\boldsymbol{X}\right] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top E\left[\boldsymbol{u}\boldsymbol{u}^\top|\boldsymbol{X}\right]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma^2\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}
\end{aligned}$$

  - Hence, tests based on $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ are invalid.

To sum up:

- The usual OLS t-statistics are **not** t-distributed under hetero-skedasticity, and the problem can not be resolved by using large sample sizes. This is also the case for the F-test.

- One will usually underestimate $\mathsf{SE}(\hat{\beta}_{j\text{OLS}})$, which implies that confidence bands get to short and t-statistics get to large.
  - Thus, hypotheses which should be retained may be rejected.

12. Heteroskedasticity-Robust Inference

- In the **simple** regression model with heteroskedasticity:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad \mathrm{var}(u_i) = \sigma_i^2$$

we have an easy correction.

- If we don't know σ_i^2, then we may use:

$$\widehat{\mathrm{var}(\hat{\beta}_{1\mathrm{OLS}})}^{\mathrm{HRSE}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}.$$

- This estimator of the standrad error is consistent.

- In the **multiple** regression model one should use

$$\widehat{\text{var}(\hat{\beta}_{jOLS})}^{\text{HRSE}} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{\left(\sum_{i=1}^{n} \hat{r}_{ij}^2\right)^2},$$

- where $\hat{r}_{ij}$ are the residuals from the following regression:

$$x_{ji} = \delta_0 + \delta_1 x_{1i} + \ldots + \delta_{j-1} x_{j-1,i} + \delta_{j+1} x_{j+1,i} + \ldots + \delta_k x_{ki} + r_{ij},$$

- E.g. the residuals from the regression where x_j are regressed on the remaining explanatory variables.

- $\sqrt{\widehat{\text{var}(\hat{\beta}_{jOLS})}^{\text{HRSE}}}$ is known as the **heteroskedasticity-robust standard error** for β_j.

13. GLS Estimation

- We have something like: $\sigma_i^2 = \sigma^2 z_i$ with z_i known.
- The trick is:

$$\frac{y_i}{\sqrt{z_i}} = \alpha \frac{1}{\sqrt{z_i}} + \beta \frac{x_i}{\sqrt{z_i}} + \frac{u_i}{\sqrt{z_i}}, \tag{1}$$

$$\Rightarrow \quad \mathsf{var}\left(\frac{u_i}{\sqrt{z_i}}\right) = \sigma^2 \quad \text{for all } i$$

- and now OLS is applicable on (1), $\hat{\beta} = \hat{\beta}_{\text{GLS}}$ with:

$$\widehat{\mathsf{var}(\hat{\beta}_{\text{GLS}})} = \frac{s^2}{(\sum^n x_i^2/z_i) - (\sum^n x_i/z_i)^2 (\sum^n z_i^{-1})^{-1}}, \quad s^2 = \frac{SSE_{\text{GLS}}}{n-2}.$$

- This method is also called **Weighted Least Squares** WLS.

- Hence in general: Least Squares estimation, when $\boldsymbol{\Sigma}$ is known, can be done by transforming the model such that the classical assumptions are fulfilled.

- As $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix we can write it as

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^\top.$$

- By the calculus rules for eigen-values and -vectors we know

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &= \boldsymbol{C}\boldsymbol{\Lambda}^{-1}\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}(\boldsymbol{C}\boldsymbol{\Lambda}^{-1/2})^\top \\ &= \boldsymbol{P}^\top\boldsymbol{P},\end{aligned}$$

and

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{-1})^{-1} = (\boldsymbol{P}^\top\boldsymbol{P})^{-1} = \boldsymbol{P}^{-1}(\boldsymbol{P}^\top)^{-1}.$$

Important general Rule

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{\Sigma} \text{ is invertible}$$
$$\Updownarrow$$
$$\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \boldsymbol{I}).$$

Here $\boldsymbol{\Sigma}^{-1/2}$ is the matrix defined such that $\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Sigma}^{-1/2})^\top \boldsymbol{\Sigma}^{-1/2}$. Let $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1/2}$. Then $\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}^\top \boldsymbol{P}$. Exactly as defined on the previous slide.

- Hence in general; Least Squares estimation, when $\boldsymbol{\Sigma}$ is known, can be done by transforming the model such that the classical assumptions (classical design criteria) on the covariance matrix are fulfilled. See next slide.

- Now use $\boldsymbol{P}$ to define:

$$\boldsymbol{y}_* = \boldsymbol{P}\boldsymbol{y}, \qquad \boldsymbol{X}_* = \boldsymbol{P}\boldsymbol{X}, \qquad \boldsymbol{u}_* = \boldsymbol{P}\boldsymbol{u},$$

- and note that this implies that

$$\begin{aligned}
\operatorname{var}(\boldsymbol{u}_*|\boldsymbol{X}) &= E[\boldsymbol{u}_*\boldsymbol{u}_*^\top|\boldsymbol{X}] = E[\boldsymbol{P}\boldsymbol{u}(\boldsymbol{P}\boldsymbol{u})^\top|\boldsymbol{X}] \\
&= E[\boldsymbol{P}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{P}^\top|\boldsymbol{X}] = \boldsymbol{P}E[\boldsymbol{u}\boldsymbol{u}^\top|\boldsymbol{X}]\boldsymbol{P}^\top \\
&= \boldsymbol{P}\sigma^2\boldsymbol{\Sigma}\boldsymbol{P}^\top \\
&= \sigma^2\boldsymbol{P}\boldsymbol{P}^{-1}(\boldsymbol{P}^\top)^{-1}\boldsymbol{P}^\top \\
&= \sigma^2\boldsymbol{I}.
\end{aligned}$$

- Hence, the transformed errors satisfy the design criteria, and it is valid to apply OLS to the model

$$\boldsymbol{y}_* = \boldsymbol{X}_* \boldsymbol{\beta} + \boldsymbol{u}_*, \qquad \boldsymbol{u}_* \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

- Hence, the estimates of $\boldsymbol{\beta}$ and σ^2 are the usual ones:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{GLS}} &= (\boldsymbol{X}_*^\top \boldsymbol{X}_*)^{-1} \boldsymbol{X}_*^\top \boldsymbol{y}_* \\
&= ((\boldsymbol{P}\boldsymbol{X})^\top \boldsymbol{P}\boldsymbol{X})^{-1} (\boldsymbol{P}\boldsymbol{X})^\top \boldsymbol{P}\boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}.
\end{aligned}$$

$$s^2_{\text{GLS}} = \frac{(\boldsymbol{y}_* - \boldsymbol{X}_*\hat{\boldsymbol{\beta}}_{\text{GLS}})^\top(\boldsymbol{y}_* - \boldsymbol{X}_*\hat{\boldsymbol{\beta}}_{\text{GLS}})}{n-k}$$
$$= \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{GLS}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{GLS}})}{n-k}$$

- ➥ These are called GLS estimators.

- Variance of $\hat{\boldsymbol{\beta}}_{\text{GLS}}$:
$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = \sigma^2(\boldsymbol{X}_*^\top\boldsymbol{X}_*)^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}.$$

- and the estimated variance:
$$\widehat{\text{var}(\hat{\boldsymbol{\beta}}_{\text{GLS}})} = s^2_{\text{GLS}}(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}.$$

- In conclusion this shows that if $\boldsymbol{\Sigma}$ is known, then it is easy to take care of a general covariance matrix for the disturbances.

14. FGLS estimation

- FGLS is short for Feasible Generalized Least Squares, and it is the method to use when $\boldsymbol{\Sigma}$ is unknown.
- In the general cases the situation is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is the unknown covariance matrix of the disturbances.

Note that the scaling factor σ^2 is included in $\boldsymbol{V}$.

- Now if we can come up with a consistent estimate of $\boldsymbol{V}$, call it $\hat{\boldsymbol{V}}$, then we are home free:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{FGLS}} &= (\boldsymbol{X}^\top \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{V}}^{-1} \boldsymbol{y}, \\ \text{var}(\hat{\boldsymbol{\beta}}_{\text{FGLS}}) &= (\boldsymbol{X}^\top \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1}. \end{aligned}$$

14.1. Whites HCSE

- Halbert White suggested some simple heteroskedasticity consistent standard errors for the OLS estimator in the case where $\boldsymbol{V}$ has the following structure:

$$\boldsymbol{V} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix},$$

  where $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ are unknown.

- This estimator is exactly equal to the robust one we encountered previously.

- Remember that the correct variance matrix for the OLS estimator is

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

- Hence, to perform FGLS we need an estimate of

$$\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \cdots & \boldsymbol{x}_1^\top & \cdots \\ \cdots & \boldsymbol{x}_2^\top & \cdots \\ & \vdots & \\ \cdots & \boldsymbol{x}_n^\top & \cdots \end{bmatrix}$$

$$= \sum_{t=1}^{n} \sigma_t^2 \boldsymbol{x}_t \boldsymbol{x}_t^\top,$$

where $\boldsymbol{x}_t = (1, x_{2t}, \cdots, x_{kt})$ is the t'th row of $\boldsymbol{X}$.

- The White estimator of $\text{var}(\hat{\boldsymbol{\beta}})$, then, is to replace σ_t^2 with the squared tth residual $\hat{u}_t^2$, such that

$$\widehat{\text{var}(\hat{\boldsymbol{\beta}})}^{\text{HCSE}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{V}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \begin{bmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_n^2 \end{bmatrix} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

- In general, one should always report these standard errors with the regression output.

Problem 1

- Verify the formulas on slide 9 from the notation on slides 6 and 8.
- Show that $\boldsymbol{\Sigma}$ on slide 17 has the form as postulated.
- Show that s^2 from slide 31 is unbiased
- Show that the OLS estimator $\hat{\boldsymbol{\beta}}$ on slide 32 has the smallest variance among all possible unbiased linear estimators.

Problem 2

- Consider the equation
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$
Suppose that the rows of $\boldsymbol{X}$ are a random sample from some distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} > 0$, and suppose that $\mathsf{var}(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$, and $\mathsf{E}(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\theta}$, where $\boldsymbol{\theta} \neq \mathbf{0}$ does not depend on n.
- Show that $\mathsf{E}(\boldsymbol{u}) = \mathbf{0}$ and $\mathsf{var}(\boldsymbol{u}) = (\sigma^2 + \boldsymbol{\theta}^\top \boldsymbol{\Sigma}\boldsymbol{\theta})\boldsymbol{I}_n$, so that the random disturbances have mean $\mathbf{0}$, constant variance, and are uncorrelated with one another, but are correlated with $\boldsymbol{X}$.
- Show that the OLS estimator of $\boldsymbol{\beta}$ satisfies $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{u}$.
- Show that $\mathsf{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} + \boldsymbol{\theta}$ and $\mathsf{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$.

Problem 3

- Let $\hat{\boldsymbol{\beta}}$ be the $(k \times 1)$ vector of OLS estimates.
  1. Show that for any $(k+1) \times 1$ vector $\mathbf{b}$, we can write the sum of squared residuals as
  $$SSR(\boldsymbol{b}) = \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{b})^\top \mathbf{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}).$$
  (Hint: Write $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = [\hat{\boldsymbol{u}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})]^\top[\hat{\boldsymbol{u}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})]$ and use the fact that $\boldsymbol{X}^\top \hat{\boldsymbol{u}} = \mathbf{0}$.
  2. Explain how the expression for $SSR(\boldsymbol{b})$ in part 1. above proves that $\hat{\boldsymbol{\beta}}$ uniquely minimizes $SSR(\boldsymbol{b})$ over all possible values of $\boldsymbol{b}$, assuming $\boldsymbol{X}$ has rank $k \times 1$.

Problem 4

Let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Let $\boldsymbol{A}$ be a $(k+1) \times (k+1)$ nonsingular matrix and define $z_t \equiv \boldsymbol{x}_t \boldsymbol{A}$, $t = 1, \ldots, n$, where $\boldsymbol{x}_t$ is the tth row in $\boldsymbol{X}$. Therefore, z_t is $1 \times (k+1)$ and is a nonsingular linear combination of $\boldsymbol{x}_t$. Let $\boldsymbol{Z}$ be the $n \times (k+1)$ matrix with rows z_t. Let $\tilde{\boldsymbol{\beta}}$ denote the OLS estimate from a regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$.

- Show that $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}$.
- Let $\hat{y}_t$ be the fitted values from the original regression and let $\tilde{y}_t$ be the fitted values from regressing $\boldsymbol{y}$ on $\boldsymbol{Z}$. Show that $\hat{y}_t = \tilde{y}_t$, for all $t = 1, \ldots, n$. How do the residuals from the two regressions compare?
- Show that the estimated variance matrix for $\tilde{\boldsymbol{\beta}}$ is $s^2 \boldsymbol{A}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{A}^{-1\top}$, where s^2 is the usual variance estimate from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$.

Problem 5

Consider the usual linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{y}$ is $(n \times 1)$, $\boldsymbol{X}$ is $(n \times 5)$, $\boldsymbol{\beta}$ is (5×1), and $\boldsymbol{u}$ is $(n \times 1)$, $\mathsf{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathsf{var}(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$. The sample size n is equal to 500, and $\boldsymbol{X}$ is non-stochastic and has full rank. OLS is used to estimate the following coefficients and variance matrix of the coefficients

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 1.1 \\ -3 \\ 0.5 \\ 0.4 \end{pmatrix}, \quad \widehat{\mathsf{var}(\hat{\boldsymbol{\beta}})} = \begin{bmatrix} 0.5 & & & & \\ 0.2 & 0.2 & & & \\ 0.1 & 0.3 & 0.1 & & \\ -0.2 & -0.1 & 0.1 & 1 & \\ -0.5 & -0.7 & -0.2 & -0.5 & 0.1 \end{bmatrix}.$$

Problem 5 (continued)

Test the following hypotheses. In each case set up the test for the hypothesis, write down the $\boldsymbol{R}$ and $\boldsymbol{r}$ matrices and do the F and Wald tests, cf. slide 47.

1. $\mathrm{H}_0 : \beta_1 = 1$
2. $\mathrm{H}_0 : -\beta_2 + \beta_4 + \beta_5 = 0$
3. $\mathrm{H}_0 : \beta_1 = 1, \quad -\beta_2 + \beta_4 + \beta_5 = 0.$

HINTS to problems

- Problem 1: The Gauss-Markov Theorem, slide page 24. The idea of the proof is to choose another arbitrary linear estimator for $\boldsymbol{\beta}$. Call this estimator $\tilde{\boldsymbol{\beta}}$. It has the form $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$. Define a matrix $\boldsymbol{D} = \boldsymbol{A} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and write the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ as something that depends on $\boldsymbol{D}$ and $\boldsymbol{X}$, and recognize that $\mathsf{var}(\tilde{\boldsymbol{\beta}}) - \mathsf{var}(\hat{\boldsymbol{\beta}})$ must be positive semi definite.

- Problem 2, second bullet on slide 76: Here you should use the **vector versions** of two general and important rules: (1) The law of iterated expectations: $\mathsf{E}[Y] = \mathsf{E}[\mathsf{E}[Y|Z]]$ for any pair of random variables. (2) The law of unconditional variance: $\mathsf{var}[Y] = \mathsf{E}[\mathsf{var}[Y|Z]] + \mathsf{var}[\mathsf{E}[Y|Z]]$ for any pair of random variables.

- Problem 3: Should be straightforward.

- Problem 4: Should be straightforward when you realize that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}$.

- Problem 5: Rewrite the F and Wald statistics on slide 47 using the fact that

$$s^2 = \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{n-k} \quad \text{and} \quad \widehat{\mathsf{var}(\hat{\boldsymbol{\beta}})} = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$