Regression Analysis

- The multiple linear regression model with $k$ explanatory variables assumes that the $t$th observation of the dependent or endogenous variable $y_t$ is described by the linear relationship

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \epsilon_t,$$

$$t = 1, \ldots, T,$$

where the $x_{ti}$'s, $t = 1, \ldots, T$, $i = 1, \ldots, k$, are the independent, exogenous, or explanatory variables.

- If there is a constant in the regression, then, for example, $x_{t1} = 1$ for all $t$. 
To compactify the notation, define

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{bmatrix}, \]

\[ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}, \]

so that

\[ y = X\beta + \epsilon. \]

We assume that \( T > k \), and \( X \) has full column rank, i.e., \( \text{rank}(X) = k \). Then we know that also \( \text{rank}(X'X) = \text{rank}(X) = k \).

If there is a constant in the regression, then the first column of \( X \) is just a column of ones.
• For the disturbance or error process \( \{\epsilon_t\}_{t=1}^T \), we assume that

(i) \( \mathbb{E}(\epsilon|X) = 0 \)
(ii) The covariance matrix of the errors is

\[
\text{Cov}(\epsilon|X) = \sigma^2 I_T = \begin{bmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2
\end{bmatrix},
\]

i.e., the errors are homoskedastic (all have the same variance \( \sigma^2 \)) and uncorrelated, i.e.,

\[
\mathbb{E}(\epsilon_t \epsilon_\tau) = \begin{cases} 
\sigma^2 & t = \tau \\
0 & t \neq \tau.
\end{cases}
\]

• We also frequently assume

\[
\epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),
\]

i.e.,

\[
\epsilon \sim \mathcal{N}(0, \sigma^2 I_T), \quad (1)
\]

where \( I_T \) is the identity matrix of dimension \( T \).
The normality assumption may often be acceptable for monthly returns, but is less adequate for higher frequencies such as weekly or daily.

A simple test for normality is provided by the Jarque–Bera test (JB), which is

\[
JB = T \left( \frac{\hat{S}^2}{6} + \frac{(\hat{K} - 3)^2}{24} \right) \overset{asy}{\sim} \chi^2(2),
\]

where

\[
\hat{S} = \frac{T^{-1} \sum_t (R_{it} - \bar{R}_i)^3}{\hat{\sigma}_i^3}, \quad \hat{K} = \frac{T^{-1} \sum_t (R_{it} - \bar{R}_i)^4}{\hat{\sigma}_i^4}
\]

are the sample skewness and sample kurtosis, respectively.

Note that both measures are location and scale free, i.e., not affected by a linear transformation \( \tilde{R}_i = a + bR_i \).
• The *skewness* checks for asymmetries (recall that the normal distribution is symmetric and has skewness zero).

• The *kurtosis*, due to the fourth moment involved, gives very large values a higher weight and thus checks whether the tails of the empirical distribution are thicker than those of the normal, i.e., whether the probability of very large observations (of either sign) is more likely than under normality (which has kurtosis equal to three).

• A kurtosis significantly larger than three indicates *excess kurtosis*, i.e., larger probability of large losses and gains.

• Such a density typically has also more weight in the center and less weight in the “shoulders” of the distribution.

• Considerable excess kurtosis is typical for returns at daily or weekly frequencies, which is clearly of great importance for risk management.
- Normality is rejected at the 5% level if JB exceeds the right tail critical value of the $\chi^2(2)$ distribution, which is approximately given by 5.99.
- 19 DAX stocks, monthly data, January 1995–December 2004, index is the DAX

<table>
<thead>
<tr>
<th>i</th>
<th>Company</th>
<th>skewness</th>
<th>kurtosis</th>
<th>JB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Allianz</td>
<td>0.37*</td>
<td>5.98***</td>
<td>46.82***</td>
</tr>
<tr>
<td>2</td>
<td>BASF</td>
<td>0.09</td>
<td>3.16</td>
<td>0.30</td>
</tr>
<tr>
<td>3</td>
<td>BayHypo</td>
<td>1.08***</td>
<td>8.66***</td>
<td>182.23***</td>
</tr>
<tr>
<td>4</td>
<td>BMW</td>
<td>−0.06</td>
<td>3.33</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>Bayer</td>
<td>−0.23</td>
<td>3.96**</td>
<td>5.56*</td>
</tr>
<tr>
<td>6</td>
<td>CommBank</td>
<td>−0.01</td>
<td>6.04***</td>
<td>45.69***</td>
</tr>
<tr>
<td>7</td>
<td>Conti</td>
<td>−0.24</td>
<td>3.71</td>
<td>3.67</td>
</tr>
<tr>
<td>8</td>
<td>DtBank</td>
<td>−0.10</td>
<td>3.92**</td>
<td>4.34</td>
</tr>
<tr>
<td>9</td>
<td>Eon</td>
<td>−0.49**</td>
<td>3.17</td>
<td>4.93*</td>
</tr>
<tr>
<td>10</td>
<td>Henkel</td>
<td>−0.12</td>
<td>3.14</td>
<td>0.37</td>
</tr>
<tr>
<td>11</td>
<td>Linde</td>
<td>−0.13</td>
<td>3.00</td>
<td>0.33</td>
</tr>
<tr>
<td>12</td>
<td>Lufthansa</td>
<td>−0.48**</td>
<td>4.51***</td>
<td>15.79***</td>
</tr>
<tr>
<td>13</td>
<td>MAN</td>
<td>−0.03</td>
<td>4.38***</td>
<td>9.45***</td>
</tr>
<tr>
<td>14</td>
<td>RWE</td>
<td>−0.27</td>
<td>2.88</td>
<td>1.57</td>
</tr>
<tr>
<td>15</td>
<td>Schering</td>
<td>0.15</td>
<td>2.66</td>
<td>1.04</td>
</tr>
<tr>
<td>16</td>
<td>Siemens</td>
<td>0.19</td>
<td>3.80*</td>
<td>3.87</td>
</tr>
<tr>
<td>17</td>
<td>Thyssen</td>
<td>0.20</td>
<td>3.39</td>
<td>1.59</td>
</tr>
<tr>
<td>18</td>
<td>TUI</td>
<td>0.60***</td>
<td>6.99***</td>
<td>86.02***</td>
</tr>
<tr>
<td>19</td>
<td>VW</td>
<td>0.01</td>
<td>2.71</td>
<td>0.42</td>
</tr>
</tbody>
</table>
*, **, and *** indicates significance at the 10%, 5%, and 1% level, respectively.
Ordinary Least Squares (OLS) Estimation

- The OLS estimator is defined to be the vector $\hat{\beta}$ which minimizes the residual sum of squares,

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

$$= y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$= \hat{\epsilon}'\hat{\epsilon},$$

where

$$\hat{\epsilon} = y - X\hat{\beta}.$$

- To minimize the sum of squares, we calculate

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2Xy + 2X'X\hat{\beta} = 0,$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'y. \quad (4)$$
For example, consider the single index model (SIM),

\[ R_{it} = \alpha_i + \beta_i R_{Mt} + \epsilon_{it}, \quad t = 1, \ldots, T. \quad (5) \]

In this case,

\[
X = \begin{bmatrix}
1 & R_{M1} \\
1 & R_{M2} \\
\vdots & \vdots \\
1 & R_{MT}
\end{bmatrix}, \quad y = \begin{bmatrix}
R_{i1} \\
R_{i2} \\
\vdots \\
R_{iT}
\end{bmatrix}, \quad (6)
\]

\[
X'X = \begin{bmatrix}
T & \sum_{t=1}^{T} R_{Mt} \\
\sum_{t=1}^{T} R_{Mt} & \sum_{t=1}^{T} R_{Mt}^2
\end{bmatrix}, \quad (7)
\]

\[
X'y = \begin{bmatrix}
\sum_{t=1}^{T} R_{it} \\
\sum_{t=1}^{T} R_{Mt} R_{it}
\end{bmatrix}, \quad (8)
\]
\[
(X'X)^{-1} = \left[ \frac{T \sum R_{Mt} \sum R_{Mt}^2}{T \sum R_{Mt}^2 - (\sum R_{Mt})^2} \right]^{-1}
\]

\[
= \left[ \frac{\sum R_{Mt}^2 - \sum R_{Mt}^2}{T \sum R_{Mt}^2 - (\sum R_{Mt})^2} \right]^{-1}
\]

\[
= \frac{1}{T} \begin{bmatrix}
\bar{R}_M^2 & -\bar{R}_M \\
-\bar{R}_M & 1
\end{bmatrix}
\]

where \( \bar{R}_M = T^{-1} \sum R_{Mt}, \bar{R}_M^2 = T^{-1} \sum R_{Mt}^2, \) and
\( s_M^2 = \bar{R}_M^2 - \bar{R}_M^2 = T^{-1} \sum (R_{Mt} - \bar{R}_M)^2 \).

- Hence

\[
\begin{bmatrix}
\hat{\alpha}_i \\
\hat{\beta}_i
\end{bmatrix}
= \frac{1}{T} \begin{bmatrix}
\frac{\bar{R}_M^2}{s_M^2} & -\bar{R}_M \\
-\bar{R}_M & 1
\end{bmatrix}
\begin{bmatrix}
\sum R_{it} \\
\sum R_{Mt}R_{it}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\frac{\bar{R}_M^2}{s_M^2} & -\bar{R}_M \\
-\bar{R}_M & 1
\end{bmatrix}
\begin{bmatrix}
\bar{R}_i \\
\bar{R}_M \bar{R}_i
\end{bmatrix}.
\]

That is

\[
\hat{\beta}_i = \frac{R_M R_i - \bar{R}_i \bar{R}_M}{s_M^2} = \frac{T^{-1} \sum_t (R_{it} - \bar{R}_i)(R_{Mt} - \bar{R}_M)}{s_M^2} = \frac{s_{iM}}{s_M^2}.
\]

\[
\hat{\alpha}_i = \frac{R_M^2 R_i - \bar{R}_M \bar{R}_M R_i}{s_M^2} = \frac{R_i (R_M^2 - \bar{R}_M^2) + \bar{R}_M (R_M \bar{R}_i - \bar{R}_M R_i)}{s_M^2} = \frac{R_i - \frac{R_M \bar{R}_i - \bar{R}_M \bar{R}_i}{s_M^2} \bar{R}_M}{s_M^2} = \bar{R}_i - \hat{\beta}_i \bar{R}_M.
\]
Goodness of Fit

- The coefficient of determination is often used to informally judge how well the model explains the observed variations in the endogenous variable.

- This is based on

\[
\hat{e}'\hat{e} = (y - X\hat{\beta})(y - X\hat{\beta})' \\
= y'y' - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \\
= y'y' - \hat{y}'\hat{y}, \quad (9)
\]

where \( \hat{y} = X\hat{\beta} \).

\[
1\hat{\beta}'X'y = \hat{\beta}X'X(X'X)^{-1}X'y = \hat{\beta}'X'X\hat{\beta}.
\]

13
• If there is a constant term in the regression, the residuals sum to zero, and we have \( \bar{y} = \hat{y} \), so that

\[
s_y^2 = s_{\bar{y}}^2 + s_{\epsilon}^2, \tag{10}
\]

and the \( R^2 \) is

\[
R^2 = \frac{s_{\bar{y}}^2}{s_y^2} = 1 - \frac{s_{\epsilon}^2}{s_y^2}, \quad 0 \leq R^2 \leq 1.
\]

• In the context of the single index model (SIM), the coefficient of determination can be interpreted as an estimate of the fraction of variability of asset \( i \) that is due to movements of the market (proportion of nondiversifiable risk).

• The remaining part of the variability is then due to nonmarket factors and can be diversified away.

• If there is no constant term in the regression, we can still define a \( R^2 \), but this does not have the same interpretation, as (10) is no longer valid.
• That is, (9) still holds, but it cannot be interpreted in terms of a variance decomposition.
• 19 DAX stocks, monthly data, January 1995–December 2004, index is the DAX

<table>
<thead>
<tr>
<th>i</th>
<th>Company</th>
<th>$\hat{\beta}_i$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Allianz</td>
<td>1.26</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>BASF</td>
<td>0.77</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>BayHypo</td>
<td>1.38</td>
<td>0.47</td>
</tr>
<tr>
<td>4</td>
<td>BMW</td>
<td>0.85</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>Bayer</td>
<td>1.01</td>
<td>0.57</td>
</tr>
<tr>
<td>6</td>
<td>CommBank</td>
<td>1.21</td>
<td>0.53</td>
</tr>
<tr>
<td>7</td>
<td>Conti</td>
<td>0.83</td>
<td>0.43</td>
</tr>
<tr>
<td>8</td>
<td>DtBank</td>
<td>1.08</td>
<td>0.63</td>
</tr>
<tr>
<td>9</td>
<td>Eon</td>
<td>0.43</td>
<td>0.23</td>
</tr>
<tr>
<td>10</td>
<td>Henkel</td>
<td>0.48</td>
<td>0.21</td>
</tr>
<tr>
<td>11</td>
<td>Linde</td>
<td>0.61</td>
<td>0.32</td>
</tr>
<tr>
<td>12</td>
<td>Lufthansa</td>
<td>1.02</td>
<td>0.52</td>
</tr>
<tr>
<td>13</td>
<td>MAN</td>
<td>0.96</td>
<td>0.42</td>
</tr>
<tr>
<td>14</td>
<td>RWE</td>
<td>0.54</td>
<td>0.26</td>
</tr>
<tr>
<td>15</td>
<td>Schering</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>16</td>
<td>Siemens</td>
<td>1.35</td>
<td>0.64</td>
</tr>
<tr>
<td>17</td>
<td>Thyssen</td>
<td>1.01</td>
<td>0.50</td>
</tr>
<tr>
<td>18</td>
<td>TUI</td>
<td>1.05</td>
<td>0.47</td>
</tr>
<tr>
<td>19</td>
<td>VW</td>
<td>0.89</td>
<td>0.39</td>
</tr>
</tbody>
</table>
• Problem of $R^2$ is that it does not take into account the number of regressors.

• The *adjusted* $R^2$ is

$$R^2_{adj} = 1 - \frac{s^2 \epsilon}{s^2 y} \frac{T - 1}{T - k} = 1 - (1 - R^2) \frac{T - 1}{T - k},$$

which may decrease with the inclusion of a new regressor, and may even become negative (so the term *R squared* may appear misleading).

• However, we may want to compare different factor models for returns by applying criteria which are directly related to the prospective use of the model, i.e., out–of–sample forecasting of the correlation structure.

• For example, a simple approach is to calculate global minimum variance portfolios over the out–of–sample period and subsequently compare the *realized* portfolio variances implied by different models.
Distribution of $\hat{\beta}$

- We have

\[
\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon)
\]
\[
= \beta + (X'X)^{-1}X'\epsilon,
\]

where $\epsilon_t \sim N(0, \sigma^2 I_T)$.

- It follows that $\hat{\beta}$ is normally distributed with mean $\beta$ (unbiasedness) and covariance matrix

\[
(X'X)^{-1}X'Cov(\epsilon)X(X'X)^{-1} = \sigma^2(X'X)^{-1},
\]

i.e.,

\[
\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}).
\]
Estimation of $\sigma^2$

- An unbiased estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{T - k}. \quad (11)$$

- Further results (under normality of the errors) are:

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2} \sim \chi^2(T - k), \quad (12)$$

a $\chi^2$ distribution with $T - k$ degrees of freedom.

- $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.
Testing Hypotheses about $\hat{\beta}$

- Consider a hypothesis about a single element of $\beta$, i.e.,
  \[ H_0 : \beta_i = \beta_i^0. \]

- The variance of $\hat{\beta}_i$ is $\sigma^2_{\beta_i} = \sigma^2(X'X)^{-1}_{ii}$, and, under $H_0$,
  \[ \frac{\hat{\beta}_i - \beta_i^0}{\sqrt{\sigma^2_{\beta_i}}} \sim N(0, 1). \]  
  \[(13)\]

- As we don’t know $\sigma^2$, we can use $\hat{\sigma}^2$ instead, but this changes the distribution of (13), due to the sampling variance of $\hat{\sigma}^2$.

- We have seen that $(T - k)\hat{\sigma}^2/\sigma^2 \sim \chi^2(T - k)$. 
• Thus, dividing the $N(0, 1)$ variable in (13) by $\sqrt{\sigma^2/\sigma^2}$ produces a Student’s $t$ distribution with $T - k$ degrees of freedom, i.e., the $t$ statistic

$$
t = \frac{\hat{\beta}_i - \beta^0_i}{\sqrt{\hat{\sigma}^2_{\beta_i}}} \sim t(T - k),$$

where $\hat{\sigma}^2_{\beta_i} = \hat{\sigma}^2 (X'X)_{ii}^{-1}$.

• A symmetric $(1 - \alpha)$ confidence interval can be calculated via

$$
\hat{\beta}_i \pm t_{1-\alpha/2, T-k} \sqrt{\hat{\sigma}^2_{\beta_i}}, \quad (14)
$$

where $t_{1-\alpha/2, T-k}$ is the $1 - \alpha/2$–quantile of the $t$ distribution with $T - k$ degrees of freedom.

• The decision rule for a two–sided $t$–test is then to reject the null at level $\alpha$ if the (absolute value of the) $t$ statistic exceeds the $(1 - \alpha/2)$–quantile of the Student’s $t$ distribution with $T - k$ degrees of freedom.
- In the two–sided $t$–test above, both positive and negative deviations from the null can lead to a rejection of $H_0$.

- That is, we test $H_0 : \beta_i = \beta_i^0$ against the alternative $H_1 : \beta_i \neq \beta_i^0$.

- If we are only interested in deviations in one direction, for example $H_1 : \beta_i > \beta_i^0$, we refer to a one–sided test.

- The corresponding null hypothesis is $H_0 : \beta_i \leq \beta_i^0$ and we reject $H_0$ if

$$t > t_{1-\alpha,T-k}, \quad (15)$$

where $t_{1-\alpha,T-k}$ is the $(1 - \alpha)$–quantile of $t$ distribution with $T - k$ degrees of freedom.
• Next, consider a hypothesis of the form

\[ H_0 : \mathbf{R}\beta = \mathbf{r}, \]

where \( \mathbf{R} \) is a \( q \times k \) matrix of rank \( q \).

• For example, if, in the single index model, we want to test \( \alpha_i = 0 \) and \( \beta_i = 1 \) for asset \( i \),

\[
\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.
\]

• If \( H_0 \) is true, then

\[
\mathbf{R}\widehat{\beta} - \mathbf{r} \sim N(0, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}').
\]

• Therefore

\[
W = \frac{(\mathbf{R}\widehat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\beta} - \mathbf{r})}{\sigma^2} \sim \chi^2(q).
\]
• As before, we cannot use this directly, as $\sigma^2$ is unknown, but we can use:

• Let $Y_1 \sim \chi^2(\nu_1)$, and $Y_2 \sim \chi^2(\nu_2)$, and let $Y_1$ and $Y_2$ be independent. Then the ratio

$$F = \frac{Y_1/\nu_1}{Y_2/\nu_2} = \frac{Y_1 \nu_2}{Y_2 \nu_1} \sim F(\nu_1, \nu_2),$$

i.e., $F$ has an $F$–distribution with $\nu_1$ degrees of freedom in the numerator and $\nu_2$ degrees of freedom in the denominator.

• Use the fact that $(T - k)\hat{\sigma}^2/\sigma^2 \sim \chi^2(T - k)$, and independence of $\hat{\beta}$ and $\hat{\sigma}^2$.

The test statistic

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{q\hat{\sigma}^2} \sim F(q, T - k).$$
• Many alternative representations exist for the \( F \)-statistic.

• For example, we can consider the restricted least squares problem, i.e., the minimization of the RSS under the constraint that \( R\hat{\beta}_R = r \).

• Working through the first–order conditions associated with the Lagrangian

\[
L = (y - X\beta)'(y - X\beta) + \lambda'(R\beta - r), \quad (16)
\]

the restricted estimator is

\[
\hat{\beta}_R = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).
\]

• Calculations show that

\[
\hat{\epsilon}_R'\hat{\epsilon}_R = \hat{\epsilon}'\hat{\epsilon} + (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).
\]
Therefore, three alternative forms of $F$ with appealing interpretation are

$$F = \frac{\hat{\epsilon}' \hat{\epsilon} - \hat{\epsilon}' \hat{\epsilon}}{q \hat{\epsilon}^2} = \frac{(\hat{\epsilon}' \hat{\epsilon} - \hat{\epsilon}' \hat{\epsilon})/q}{\hat{\epsilon}' \hat{\epsilon}/(T - k)}.$$ 

$$= \frac{(R^2 - R_R^2)/q}{(1 - R^2)/(T - k)}$$ 

$$= \frac{(\hat{\beta}_R - \hat{\beta})(X'X)(\hat{\beta}_R - \hat{\beta})}{q \hat{\sigma}^2}.$$
Optimality Properties of $\hat{\beta}$

1) Gauss-Markov Theorem

The OLS estimator is the **Best Linear Unbiased Estimator (BLUE)**, i.e., for any other unbiased linear estimator $\tilde{\beta}$ of $\beta$, we have that

$$\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta})$$

is a positive semidefinite matrix.

To see this, let

$$\tilde{\beta} = Ay$$

be unbiased for $\beta$.

Define

$$B = A - (X'X)^{-1}X',$$

so that

$$\tilde{\beta} = [B + (X'X)^{-1}X']y$$

$$= [B + (X'X)^{-1}X'](X\beta + \epsilon)$$

$$= (BX + I_k)\beta + (B + (X'X)^{-1}X')\epsilon.$$
Unbiasedness requires $BX = 0$, so that

$$
\tilde{\beta} = \beta + (B + (X'X)^{-1}X')\epsilon,
$$

and the covariance matrix of $\tilde{\beta}$ is

$$
\text{Cov}(\tilde{\beta}) = \sigma^2(B + (X'X)^{-1}X')(B + (X'X)^{-1}X')'
$$

$$
= \sigma^2BB'
$$

$$
+ \sigma^2[BX(X'X)^{-1} + (X'X)^{-1}X'B']
$$

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

$$
= \sigma^2[BB' + (X'X)^{-1}]
$$

Therefore, the covariance matrix of the competing estimator exceeds that of the OLS estimator by $\sigma^2BB'$, which is a positive semidefinite matrix.
Now consider any linear combination $w'\beta$ of the elements of the parameter vector.

The sampling variance of $w'\tilde{\beta}$ is

$$\text{Var}(w'\tilde{\beta}) = w'[\sigma^2BB' + \sigma^2(X'X)^{-1}]w$$

$$\geq \sigma^2w'(X'X)^{-1}w.$$  

This shows that the sampling variance of $w'\tilde{\beta}$ is less than or equal to that of the corresponding combination of any other linear unbiased estimator of $\beta$.

By choosing $w$ to be the $i$th unit vector in $\mathbb{R}^k$, an implication of this is

$$\text{Var}(\hat{\beta}_i) \leq \text{Var}(\tilde{\beta}_i), \quad i = 1, \ldots, k.$$
2) Given our distributional assumption about the error term (iid normality), we can say even more, by relying on the Cramér–Rao lower bound (CRLB).

In fact, in this case, the OLS estimator has the smallest variance among all unbiased estimators (not only linear estimators), meaning that it is an **UMVUE** (**U**iformly **M**inimum–**V**ariance **U**nbias**E**stimator).
Prediction

• Suppose we want to predict

\[ y_{T+1} = x_{T+1}\beta + \epsilon_{T+1}, \]

using our OLS–fitted model.

• Our predictor is

\[ \hat{y}_{T+1} = x_{T+1}\hat{\beta} = x_{T+1}(X'X)^{-1}X'y, \]

which has prediction error

\[ x_{T+1}\hat{\beta} - y_{T+1} = x_{T+1}(\hat{\beta} - \beta) - \epsilon_{T+1}. \]

• This is unbiased in the sense that the prediction error has zero expectation, and the prediction error variance is

\[ \mathbb{E}[(\hat{y}_{T+1} - y_{T+1})^2] = \sigma^2[x_{T+1}(X'X)^{-1}x_{T+1}'] + 1]. \]
This is the best linear unbiased predictor, because a linear unbiased predictor of the form $\tilde{y}_{T+1} = ay$ can be written as $[x_{T+1}(X'X)^{-1}X' + b]y$, with $b = a - x_{T+1}(X'X)^{-1}X'$.

This has prediction error

$$\tilde{y}_{T+1} - y_{T+1} = [x_{T+1}(X'X)^{-1}X' + b](X\beta + \epsilon)$$

$$- x_{T+1}\beta - \epsilon_{T+1}$$

$$= bX\beta$$

$$+ [x_{T+1}(X'X)^{-1}X' + b]\epsilon - \epsilon_{T+1},$$

so that $bX = 0$ is required for unbiasedness.

The forecast error variance is then

$$\mathbb{E}[(\tilde{y}_{T+1} - y_{T+1})^2]$$

$$= \sigma^2(x_{T+1}(X'X)^{-1}x'_{T+1} + bb' + 1)$$

$$\geq \sigma^2(x_{T+1}(X'X)^{-1}x'_{T+1} + 1)$$

$$= \mathbb{E}[(\tilde{y}_{T+1} - y_{T+1})^2].$$
Generalized Least Squares

• If $\epsilon \sim \mathcal{N}(0, \Sigma)$, where $\Sigma$ is a covariance matrix, the OLS estimator need no longer be efficient according to the Gauss-Markov theorem.

• However, we can find nonsingular $C$ such that $CC' = \Sigma$, for example, by employing the Choleski decomposition, with $C$ nonsingular.

• Then, with $\tilde{\epsilon} = C^{-1}\epsilon$, we have $E(\tilde{\epsilon}\tilde{\epsilon}') = C^{-1}\Sigma(C')^{-1} = C^{-1}CC'(C')^{-1} = I$.

• Thus, we may consider the transformed equation

$$C^{-1}y = C^{-1}X + C^{-1}\epsilon,$$

with GLS (generalized least squares) estimator

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y.$$
• We will use this when we discuss the Black–Litterman approach.