Lecture 2: The Classical Linear Regression Model

Introduction

In lecture 1 we

- Introduced concept of an economic relation
- Noted that empirical economic relations do not resemble textbook relations
- Introduced a method to find the best fitting linear relation

No reference to mathematical statistics in all of this. Initially econometrics did not use the tools of mathematical statistics.

Mathematical statistics develops methods for the analysis of data generated by a random experiment in order to learn about that random experiment.
Is this relevant in economics?

Consider

- Wage equation: relation between wage and education, work experience, gender, …
- Macro consumption function: relation between (national) consumption and (national) income

What is the random experiment?

To make progress we start with the assumption that all economic relations are essentially deterministic, if we include all variables

\[ y = f(x_1, \ldots, x_w) \]
Hence, if we have data \( y_i, x_{i1}, \ldots, x_{iW}, i = 1, \ldots, n \) then

\[
y_i = f(x_{i1}, \ldots, x_{iW}) \quad , i = 1, \ldots, n
\]

Let \( \bar{x}_1, \ldots, \bar{x}_W \) be the sample averages of the variables and assume that \( f \) is sufficiently many times differentiable to have a Taylor series expansion around \( \bar{x}_1, \ldots, \bar{x}_W \), i.e. a polynomial approximation

\[
y_i = \beta_0' + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_W(x_{iW} - \bar{x}_W) + \cdots
\]

\[
\cdots + \gamma_1(x_{i1} - \bar{x}_1)^2 + \cdots + (x_{iW} - \bar{x}_W)^2 + \cdots
\]

\[
\cdots + \delta_1(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \cdots
\]
We divide $x_1, \ldots, x_W$ into three groups

1. Variables that do not vary in the sample (take this to be last $W - V$ variables), i.e. for $i = 1, \ldots, n$, $x_{i, V+1} = \overline{x}_{V+1}, \ldots, x_{i, W} = \overline{x}_W$. Example: gender if we consider a sample of women.
2. Variables in the relation that are omitted or cannot be included because they are unobservable. Let this be the next $V - K + 1$ variables.
3. Variables included in the relation, i.e. $x_1, \ldots, x_{K+1}$. 
To keep the relation simple we concentrate on the linear part. Hence the observations satisfy

\[ y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{iK-1} + \varepsilon_i \]

with

\[ \beta_0 = \beta_0' - \sum_{j=1}^{K=1} \beta_j \bar{x}_j \]

The remainder term contains all the omitted terms

\[ \varepsilon_i = \beta_K (x_{iK} - \bar{x}_K) + \cdots + \beta_V (x_{iV} - \bar{x}_V) + \]

\[ + \gamma_1 (x_{i1} - \bar{x}_1)^2 + \cdots + \gamma_V (x_{iV} - \bar{x}_V)^2 + \]

\[ + \delta_1 (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \cdots \]

We call \( \varepsilon_i \) the disturbance (of the exact linear relation).
Note that

\[ \beta_j = \frac{\partial f}{\partial x_j}(\bar{x}_1, \ldots, \bar{x}_W), \quad j = 1, \ldots, K - 1 \]

\[ \beta_0 = f(\bar{x}_1, \ldots, \bar{x}_W) - \sum_{j=1}^{K-1} \beta_j \bar{x}_j \]

Conclusions:

1. The slope coefficient \( \beta_j \) is the partial effect of \( x_j \) on \( y \).
2. The slope coefficients and the intercept depend on the variables that are constant in a sample, e.g. in a sample of women the coefficients in a wage relation may be different from those in a sample of only men.
3. Only in very special cases will the relation have a 0 intercept.
Consider the following experiment: Prediction of $y_i$ on the basis of $x_{i1}, \ldots, x_{i,K-1}$.

Assume that we have observed $x_{i1}, \ldots, x_{i,K-1}$. This does not tell us anything about the disturbance $\epsilon_i$ that depends on (many) variables beside $x_1, \ldots, x_{K-1}$. Hence, even if we know the coefficients $\beta_0, \ldots, \beta_{K-1}$, we cannot predict with certainty what $y_i$ is.

A variable with a value that is unknown before the experiment is performed is a random variable in probability theory.

As in classical random experiments (flipping coin, rolling die) randomness due to lack of knowledge.
In prediction experiment \( y_i \) is a random variable and hence so is

\[
\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{K-1} x_{i,K-1}
\]

If we have a sample of size \( n \), we have \( n \) replications of the random experiment

\[
y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_{K-1} x_{1,K-1} + \varepsilon_1
\]

\[
\vdots
\]

\[
y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{K-1} x_{n,K-1} + \varepsilon_n
\]

The distribution of the disturbance \( \varepsilon \) reflects the fact that we do not know the value of the disturbance if we know \( x_1, \ldots, x_{K-1} \).

In fact we make a more extreme assumption: the value of \( \varepsilon \) is completely unpredictable from \( x_1, \ldots, x_{K-1} \).
Assumption 1: \( \text{E}(\varepsilon_i \mid x_{i1}, \ldots, x_{i,K-1}) = 0 \)

In words: the disturbance \( \varepsilon \) is mean-independent of \( x_1, \ldots, x_{K-1} \).

Why do we need this assumption?

Consider

(1) \[ \varepsilon_i = \gamma(x_{i1} - \bar{x}_1) + \eta_i, \quad i = 1, \ldots, n \]

Hence, \( \varepsilon \) is (partially) predictable using \( x_1 \) (but not completely so).

Substitution gives

\[ y_i = \beta_0 - \gamma\bar{x}_1 + (\beta_1 + \gamma)x_{i1} + \cdots + \beta_{K-1}x_{i,K-1} + \eta_i \]
Remember that

\[ \beta_1 = \frac{\partial f}{\partial x_1} (\bar{x}_1, \ldots, \bar{x}_W) \]

Conclusion: If (1) holds then the coefficient of \( x_1 \) is not equal to the partial effect of \( x_1 \) on \( y \).

Failure of Assumption 1 is a failure of the ceteris paribus condition in the sample: a change in \( x_1 \) has two effects on \( y \), a direct effect \( \beta_1 \) and an indirect effect \( \gamma \). The latter effect is due to the fact that in a sample we cannot hold other relevant variables fixed/constant. Hence we only measure the partial effect if the omitted variables are not related with \( x_1 \).

Measuring partial effects is the goal of most empirical research in economics (and other social sciences). The biggest challenge in empirical research is to ensure the Assumption 1 holds.
Whether we care depends on our objectives. Compare homeowner who is interested in relation between house price and square footage of house.

- If he/she wants to predict the sales price of house as is no reason to be worried about interpretation of regression coefficient as partial effect.
- If he/she wants to evaluate the investment in an addition to the house the estimation of the partial effect is essential.

Two strategies to estimate partial effect

- Include all variables that are correlated with $x_1$ in the relation.
- Assign $x_1$ randomly, i.e. using a random experiment that is independent of anything, e.g. by flipping a coin if $x_1$ is dichotomous.
The CLR model

Re-label the variables $x_1, \ldots, x_{K-1}$ as $x_2, \ldots, x_K$ and introduce the variable

$$x_1 \equiv 1$$

i.e. $x_{i1} = 1, i = 1, \ldots, n$. Also re-label $\beta_0, \ldots, \beta_{K-1}$ to $\beta_1, \ldots, \beta_K$.

In the new notation for $i = 1, \ldots, n$

$$(2) \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \epsilon_i$$

This is the multiple linear regression model. The relation $(2)$ is linear in $x_1, \ldots, x_K$ and also linear in $\beta_1, \ldots, \beta_K$. The latter is essential. By re-interpreting $x_1, \ldots, x_K$ we can deal with relations that are non-linear in these variables.
Examples:

- **Polynomial in** $x$
  \[ y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \varepsilon \]

  **Define:** $x_1 = 1$, $x_2 = x$, $x_3 = x^2$, $x_4 = x^3$
• Log-linear relation

\[ \ln y = \beta_1 + \beta_2 \ln x + \epsilon \]

**Define:** \( x_1 \equiv 1, \quad x_2 = \ln x \)

**Note**

\[ \beta_2 = \frac{\partial \ln y}{\partial \ln x} \]

i.e. \( \beta_2 \) is the elasticity of \( y \) w.r.t. \( x \).
Economics 513

- Semi-log relation

\[ \ln y = \beta_1 + \beta_2 x + \varepsilon \]

Note: no re-definition is needed.

Also

\[ \beta_2 = \frac{\partial \ln y}{\partial x} = \frac{\partial y}{\partial x} \]

This is a semi-elasticity.
Assumption 1 in the new notation:

\[(3) \quad E(\varepsilon_i \mid x_{i1}, \ldots, x_{iK}) = 0, \quad i = 1, \ldots, n\]

Note that this implies that \(E(\varepsilon_i) = 0\). This is without loss of generality because a non-zero mean can be absorbed into the intercept.
In matrix notation (2) and (3) are

\[ y = X\beta + \varepsilon \]

and

\[ E(\varepsilon_i \mid x_i) = 0 \quad , i = 1, \ldots, n \]

with

\[ X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \]

(4)

Note that \( x_i \) is a \( K \)-vector.
The Classical (Multiple) Linear Regression (CLR) model is a set of assumptions mainly on the conditional distribution of $\varepsilon$ given $x_1, \ldots, x_K$.

Some assumptions are essential, some are convenient initial assumptions and can/will be relaxed.

The CLR model is appropriate in many practical situations and is the starting point for the use of mathematical statistical inference to measure economic relations.
CLR model

\[ y = X\beta + \varepsilon \]

**Assumption 1: Fundamental assumption**

\[ E(\varepsilon | X) = 0 \]

**Assumption 2: Spherical disturbances**

\[ E(\varepsilon \varepsilon' | X) = \sigma^2 I \]

**Assumption 3: Full rank**

\[ \text{rank}(X) = K \]
Discussion of the assumptions

Assumption 1 is shorthand for

\[ E(\varepsilon_i \mid X) = 0 \quad , i = 1, \ldots, n \]

Hence this is equivalent to

\[ E(\varepsilon_i \mid x_1, \ldots, x_n) = 0 \quad , i = 1, \ldots, n \]

Compare this with

\[ E(\varepsilon_i \mid x_i) = 0 \quad , i = 1, \ldots, n \]

By the law of iterated expectations, the current assumption implies the latter.
The current assumption states that not only $x_i$ but also $x_j, j \neq i$ is not related to $\varepsilon_i$. This is not stronger than the previous assumption if $x_1, \ldots, x_n$ are independent as in a random sample from a population. If these are not independent, as e.g. in time-series data, then this additional assumption may be too strong.

Assumption 1 is satisfied if $X$ is chosen independently of $\varepsilon_i$. In that case we can treat $X$ as a matrix of known constants. Therefore instead of Assumption 1 one sometimes sees

Assumption 1’: $X$ is a matrix of known constants determined independently of $\varepsilon$.

Note: Chosing/controlling $X$ is not enough.
Next, we consider assumption 2

Note

\[ \mathbf{\varepsilon \varepsilon'} = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1 \varepsilon_2 & \cdots & \varepsilon_1 \varepsilon_n \\ \varepsilon_2 \varepsilon_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ \varepsilon_n \varepsilon_1 & \cdots & \cdots & \varepsilon_n^2 \end{bmatrix} \]

Hence Assumption 2 implies that

\[ \mathbb{E}(\varepsilon_i^2 \mid X) = \sigma_i^2, \quad i = 1, \ldots, n \]

This is called homoskedasticity.
Assumption 2 also implies that

$$E(\varepsilon_i \varepsilon_j \mid X) = 0$$

Hence, given $X$ the disturbances are uncorrelated.
Example of failure of homoskedasticity: random coefficients

\[
y = \beta_0 + (\beta_1 + u)x + \varepsilon
\]

\[\uparrow\]

population variation
in coefficient

\[
= \beta_0 + \beta_1 x + \varepsilon + ux
\]

For the composite disturbance, if \(E(u \mid x) = 0\)

\[
E(\varepsilon + ux \mid x) = 0
\]

but

\[
\text{Var}(\varepsilon + ux \mid x) = \sigma^2 + \sigma_{\varepsilon u} x + \sigma_u^2 x^2
\]

with \(\sigma_{\varepsilon u} = E(\varepsilon u), \sigma_u^2 = E(u^2)\). Hence the composite error is heteroskedastic.
Failure of uncorrelated disturbances: serial correlation

Assume serial correlation of order 1 in disturbances

\[ \varepsilon_i = \rho \varepsilon_{i-1} + u_i \]

Applies in time-series.
Finally, consider Assumption 3.

If \( \text{rank}(X) = K \), then \( Xa = 0 \) if and only if \( a = 0 \) with \( a \) a \( K \)-vector, i.e. there is no linear relation between the \( K \) variables.
Failure of Assumption 3: wage equation

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \quad , i = 1, \ldots, n \]

with

- \( x_2 = \text{schooling (years)} \)
- \( x_3 = \text{age} \)
- \( x_4 = \text{potential experience} \)  
  (age-years in school-6)
Hence

\[ x_4 = x_3 - x_2 - 6 \]
Define for any constant $c$,

$$\tilde{\beta}_1 = \beta_1 - 6c, \quad \tilde{\beta}_2 = \beta_2 - c, \quad \tilde{\beta}_3 = \beta_3 + c, \quad \tilde{\beta}_4 = \beta_4 - c$$

Then also

$$y_i = \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \tilde{\beta}_3 x_{i3} + \tilde{\beta}_4 x_{i4} + \varepsilon_i, \quad i = 1, \ldots, n$$

Conclusion: We cannot distinguish between $\beta_1, \ldots, \beta_4$ and $\tilde{\beta}_1, \ldots, \tilde{\beta}_4$. For all $c$ these parameters are observationally equivalent.

Problem is also clear if we substitute for $x_4$

$$y_i = (\beta_1 - 6 \beta_4) x_{i1} + (\beta_2 - \beta_4) x_{i2} + (\beta_3 + \beta_4) x_{i3} + \varepsilon_i$$
By assumptions 1 and 2, $E(\varepsilon \mid X) = 0$, $\text{Var}(\varepsilon \mid X) = \sigma^2 I$. Sometimes it is assumed that

$$\varepsilon \mid X \sim N(0, \sigma^2 I).$$

Why is the normal distribution a natural choice?

In sequel we sometimes assume:

Assumption 4

$$\varepsilon \mid X \sim N(0, \sigma^2 I)$$
Linear regression as projection

Alternative interpretation of linear regression is as conditional mean function.

In previous derivation regression coefficient $\beta_j$ in multiple linear regression is the partial effect on $y$

$$\beta_j = \frac{\partial f}{\partial x_j}(\bar{x}_1, \ldots, \bar{x}_W)$$

For this result Assumption 1 was necessary.

Alternative: Consider $\beta_j$ as the coefficient of $x_j$ in a linear conditional mean function.

To keep things simple we consider the case of 1 explanatory variable (and an intercept).
The dependent variable \( y \) and the independent variable \( x \) have a joint population distribution with joint frequency distribution \( f(x, y) \).

Example: \( y \) = savings rate, \( x \) = income and \( f(x, y) \) is frequency distribution over all US households.

By a sample survey we obtain \( y_i, x_i, i = 1, \ldots, n \). This survey can be used to obtain an estimate of the population \( f(x, y) \) denoted by \( \hat{f}(x, y) \) (see table 1.1).

Note savings rate and income have been discretized.
### Table 1.1 Joint frequency distribution of $X = \text{income}$ and $Y = \text{savings rate}$.

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
<th>0.5</th>
<th>1.5</th>
<th>2.5</th>
<th>3.5</th>
<th>4.5</th>
<th>5.5</th>
<th>6.7</th>
<th>8.8</th>
<th>12.5</th>
<th>17.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.001</td>
<td>0.011</td>
<td>0.007</td>
<td>0.006</td>
<td>0.005</td>
<td>0.005</td>
<td>0.008</td>
<td>0.009</td>
<td>0.014</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>0.001</td>
<td>0.002</td>
<td>0.006</td>
<td>0.007</td>
<td>0.010</td>
<td>0.007</td>
<td>0.008</td>
<td>0.009</td>
<td>0.008</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.002</td>
<td>0.006</td>
<td>0.004</td>
<td>0.007</td>
<td>0.010</td>
<td>0.011</td>
<td>0.020</td>
<td>0.019</td>
<td>0.013</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.002</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.016</td>
<td>0.020</td>
<td>0.042</td>
<td>0.054</td>
<td>0.024</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.010</td>
<td>0.023</td>
<td>0.033</td>
<td>0.031</td>
<td>0.041</td>
<td>0.029</td>
<td>0.047</td>
<td>0.039</td>
<td>0.042</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.013</td>
<td>0.013</td>
<td>0.000</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$-0.05$</td>
<td>0.001</td>
<td>0.012</td>
<td>0.011</td>
<td>0.005</td>
<td>0.012</td>
<td>0.016</td>
<td>0.017</td>
<td>0.014</td>
<td>0.004</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>$-0.18$</td>
<td>0.002</td>
<td>0.008</td>
<td>0.013</td>
<td>0.006</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
<td>0.006</td>
<td>0.006</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>$-0.25$</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.006</td>
<td>0.009</td>
<td>0.007</td>
<td>0.005</td>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>$p(x)$</td>
<td>0.041</td>
<td>0.093</td>
<td>0.093</td>
<td>0.082</td>
<td>0.113</td>
<td>0.103</td>
<td>0.155</td>
<td>0.155</td>
<td>0.113</td>
<td>0.052</td>
<td></td>
</tr>
</tbody>
</table>

**Source:** Adapted from R. Kosobud and J. N. Morgan, eds., *Consumer Behavior of Individual Families over Two and Three Years* (Ann Arbor: Institute for Social Research, The University of Michigan, 1964), Table 5-5.
From the joint frequency in the sample we obtain the sample conditional frequency distribution (see table 1.2)

\[
\hat{f}(y \mid x) = \frac{\hat{f}(y, x)}{\hat{f}(x)}
\]

**Table 1.2** Conditional frequency distributions of \(Y = \) savings rate for given values of \(X = \) income.

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>1.5</th>
<th>2.5</th>
<th>3.5</th>
<th>4.5</th>
<th>5.5</th>
<th>6.7</th>
<th>8.8</th>
<th>12.5</th>
<th>17.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>.024</td>
<td>.118</td>
<td>.075</td>
<td>.073</td>
<td>.044</td>
<td>.049</td>
<td>.052</td>
<td>.058</td>
<td>.124</td>
<td>.077</td>
</tr>
<tr>
<td>.40</td>
<td>.024</td>
<td>.022</td>
<td>.064</td>
<td>.086</td>
<td>.088</td>
<td>.068</td>
<td>.052</td>
<td>.058</td>
<td>.071</td>
<td>.135</td>
</tr>
<tr>
<td>.25</td>
<td>.049</td>
<td>.064</td>
<td>.043</td>
<td>.086</td>
<td>.088</td>
<td>.107</td>
<td>.129</td>
<td>.123</td>
<td>.115</td>
<td>.115</td>
</tr>
<tr>
<td>.15</td>
<td>.049</td>
<td>.097</td>
<td>.097</td>
<td>.146</td>
<td>.142</td>
<td>.194</td>
<td>.271</td>
<td>.348</td>
<td>.212</td>
<td>.384</td>
</tr>
<tr>
<td>.05</td>
<td>.244</td>
<td>.247</td>
<td>.355</td>
<td>.378</td>
<td>.363</td>
<td>.281</td>
<td>.303</td>
<td>.252</td>
<td>.372</td>
<td>.135</td>
</tr>
<tr>
<td>.0</td>
<td>.317</td>
<td>.140</td>
<td>.000</td>
<td>.024</td>
<td>.009</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>-.05</td>
<td>.024</td>
<td>.129</td>
<td>.118</td>
<td>.061</td>
<td>.106</td>
<td>.155</td>
<td>.109</td>
<td>.090</td>
<td>.035</td>
<td>.058</td>
</tr>
<tr>
<td>-.18</td>
<td>.049</td>
<td>.086</td>
<td>.140</td>
<td>.073</td>
<td>.080</td>
<td>.078</td>
<td>.052</td>
<td>.052</td>
<td>.053</td>
<td>.038</td>
</tr>
<tr>
<td>-.25</td>
<td>.220</td>
<td>.097</td>
<td>.108</td>
<td>.073</td>
<td>.080</td>
<td>.068</td>
<td>.032</td>
<td>.019</td>
<td>.018</td>
<td>.058</td>
</tr>
<tr>
<td>Total</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

\(m_{y \mid x} \)
If there is an exact relation between $y$ and $x$, then every column should have one 1 and rest 0’s.

If not, we can consider the average of $y$ for every value of $x$. This is the conditional mean function.

In population

$$m(x) = E(y \mid x) = \sum_y y f(y \mid x)$$

with estimate

$$\hat{m}(x) = \sum_y y \hat{f}(y \mid x)$$
Note that $\hat{m}(x)$ may be rough: sampling variation around smooth population $m(x)$. 

*Figure 1.1* Conditional mean function: savings rate on income.
Why is \( m(x) = E(y \mid x) \) interesting?

One reason is optimal prediction. Assume joint population distribution \( f(x, y) \) is known and that you have a random draw from this distribution. Only \( x \) is revealed and you must predict \( y \). What is the best predictor \( h(x) \)?

Criterion: minimize expected squared prediction error

\[
E\left( (y - h(x))^2 \right) = E\left( ((y - m(x)) + (m(x) - h(x)))^2 \right) = \\
E\left( (y - m(x))^2 \right) + 2E\left( (y - m(x))(m(x) - h(x)) \right) + \\
E\left( (m(x) - h(x))^2 \right) \geq E\left( (y - m(x))^2 \right)
\]

and this lower bound is achieved if \( h(x) = m(x) \).

Conclusion: Optimal prediction is \( h(x) = E(y \mid x) \).
Now let us restrict to linear prediction

\[ h(x) = a + bx \]

**Best linear predictor**

\[
\min_{a,b} E \left( (y - a - bx)^2 \right)
\]

**First-order conditions with** \( u = y - a - bx \)

\[-2 E(u) = 0 \quad \Rightarrow \quad E(y) = a + b E(x)\]

\[-2 E(ux) = 0 \quad \Rightarrow \quad E(xy) = a + b E(x^2)\]
Solution (compare with OLS solution)

\[
b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}
\]

\[
a = \mathbb{E}(y) - b \mathbb{E}(x)
\]

If we replace population moments with sample moments we obtain

\[
\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
\hat{a} = \bar{y} - \hat{b}\bar{x}
\]

This is the OLS solution!
There is an important difference between the case that the conditional mean is linear and the case that the conditional mean is not linear.

If \( E(y \mid x) = a + bx \), we have for \( u = y - a - bx \)

\[
E(u \mid x) = E(y - a - bx \mid x) = E(y - E(y \mid x) \mid x) = 0
\]

This is Assumption 1 in the CLR model.

However, \( b \) has not a ‘structural’ interpretation, as a partial effect!
If conditional mean is not linear, we have from the first-order condition \( E(ux) = 0 \).

This is weaker:

\[
E(u \mid x) = 0 \quad \Rightarrow \quad E(ux) = 0
\]

but

\[
E(ux) = 0 \quad \text{not} \quad \Rightarrow \quad E(u \mid x) = 0
\]

Hence in that case Assumption 1 of the CLR model is not satisfied but a weaker uncorrelatedness assumption.
The Ordinary Least Squares (OLS) estimator

We have $n$ observations $y_i, x_{i1}, \ldots, x_{iK}, i = 1, \ldots, n$. We organize the data in the $n \times 1$ vector $y$ and the $n \times K$ matrix $X$ with

$$X = \begin{bmatrix}
1 & x_{12} & \cdots & x_{1K} \\
1 & x_{22} & \cdots & x_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n2} & \cdots & x_{nK}
\end{bmatrix}$$

The observations are a sample from a population and we assume that the joint distribution of $y, X$ is such that the CLR model is the appropriate statistical model, i.e. $y, X$ satisfy

$$y = X\beta + \varepsilon$$
for some $K \times 1$ vector $\beta$ of regression coefficients and some $n \times 1$ vector of random errors $\epsilon$ with a distribution that satisfies

$$E(\epsilon \mid X) = 0$$
$$E(\epsilon \epsilon' \mid X) = \sigma^2 I$$

This specifies the random experiment of which $y, X$ is the outcome (the CLR model).

We can now discuss statistical inference: estimation of population parameters and tests of hypotheses concerning the population parameters and other aspects of the population distribution.
Setup applies to both cross-section and time-series data. In first case \( y_i, x_{i2}, \ldots, x_{iK}, i = 1, \ldots, n \) is a random sample and the CLR assumptions on the population distribution can made on

\[
y_1 = x_1' \beta + \epsilon_1
\]

Because the observations are independent we can obtain the joint distribution of \( y, X \) from the marginal distributions. For time-series data the observations are not independent and the CLR model applies directly to the joint distribution of \( y, X \).
Estimation of $\beta$ and $\sigma^2$

The solution to minimization of sum of squared deviations/residuals

$$b = (X' X)^{-1} X' y$$

Note that for this $\text{rank}(X) = K$. This is an estimator of $\beta$ (only depends on the data): the Ordinary Least Squares (OLS) estimator of $\beta$.

Is the OLS estimator a good estimator?

In mathematical statistics estimators are evaluated by considering their sampling distribution, i.e. their distribution in repeated samples $(y_s, X_s), s = 1, \ldots, S$. 
All samples are realizations of CLR random experiment

\[ y_s = X_s \beta + \varepsilon_s, \quad s = 1, \ldots, S \]

The sampling distribution of the OLS estimator \( \hat{b} \) is the frequency distribution of \( b_s, s = 1, \ldots, S \) for \( S \) large. We can obtain this distribution by computer simulation (as in assignment 2).
Alternative is to use the CLR assumptions and rules of probability theory to derive (features) of the sampling distribution of $b$.

Consider

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

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

From this we can, using the CLR assumptions, find the conditional average of $b$ given $X$ (in the sampling distribution)

$$E(b \mid X) = \beta + (X'X)^{-1} X' E(\varepsilon \mid X) = \beta$$

Hence the unconditional average of $b$ is (law of iterated expectations)

$$E(b) = E_X (E(b \mid X)) = \beta$$

In words: under CLR assumptions the OLS estimator is unbiased for $\beta$. 
Beside mean consider the variance of $b$:

$$\text{Var}(b) = E[(b - E(b))(b - E(b))'] = E[(b - \beta)(b - \beta)']$$

We have

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

Upon substitution

$$\text{Var}(b) = E\left((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right)$$
Economics 513

We have

\[
E\left( (X'X)^{-1} X'\varepsilon\varepsilon' X (X'X)^{-1} | X \right) = \\
= (X'X)^{-1} X' E(\varepsilon\varepsilon' | X) X (X'X)^{-1} = \sigma^2 (X'X)^{-1}
\]

and hence

\[
\text{Var}(b) = E_X (\sigma^2 (X'X)^{-1}) = \sigma^2 E((X'X)^{-1})
\]

Note that \( \sigma^2 (X'X)^{-1} \) is an unbiased estimator of this variance.
In special case of constant and one regressor we have the unbiased variance estimator

\[
\text{Var}(b_2) = \frac{\sigma^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

Note that this decreases with \( \sigma^2 \) and with the variation in \( x \).
Optimality of OLS estimator in CLR model

Consider class of estimators for $\beta$ that are linear in $y$, i.e.

$$ b_L = Cy $$

For the OLS estimator $C = (X'X)^{-1}X'$, i.e. $C$ in general depends on $X$.

Gauss-Markov Theorem: In CLR model the OLS estimator is the Best Linear Unbiased (BLU) of $\beta$, i.e. it has the smallest variance of linear unbiased estimators.