Lecture 5. Ordinary Least Squares: Estimation, Inference and Predicting Outcomes

Let us review the basics of the linear model. We have \( n \) units (individuals, firms, or other economic agents) drawn randomly from a large population. On each unit we observe an outcome \( y_i \) for unit \( i \), and a \( K \)-dimensional vector of explanatory variables \( x_i = (x_{i1}, x_{i2}, \ldots, x_{iK})' \) (where typically the first covariate is a constant, \( x_{i1} = 1 \) for all \( i = 1, \ldots, n \)). We are interested in explaining the distribution of \( y_i \) in terms of the explanatory variables \( x_i \) using a linear model:

\[
y_i = \beta' x_i + \varepsilon_i.
\] (1)

In matrix notation,

\[
Y = X\beta + \varepsilon,
\]

or, avoiding vector and matrix notation completely,

\[
y_i = \beta_1 \cdot x_{i1} + \ldots + \beta_K \cdot x_{iK} + \varepsilon_i = \sum_{k=1}^{K} \beta_k \cdot x_{ik} + \varepsilon_i.
\]

We consider a sequence of increasingly weaker assumptions on the relation between \( \varepsilon_i \) and \( x_i \). First, we assume that the residuals \( \varepsilon_i \) are independent of the covariates or regressors, and normally distributed with mean zero and variance \( \sigma^2 \):

**Assumption 1** \( \varepsilon_i | x_i \sim \mathcal{N}(0, \sigma^2) \).

We can weaken this considerably. First, we could relax normality and only assume independence:

**Assumption 2** \( \varepsilon_i \perp x_i \),

combined with the normalization that \( \mathbb{E}[\varepsilon_i] = 0 \). We can even weaken this assumption further by requiring only mean-independence.
Assumption 3 \( \mathbb{E}[\varepsilon_i|x_i] = 0, \)

or even further, requiring only zero correlation:

Assumption 4 \( \mathbb{E}[\varepsilon_i \cdot x_i] = 0. \)

We will also assume that the observations are drawn randomly from some population. We can also do most of the analysis by assuming that the covariates are fixed, but this complicates matters for some results, and it does not help very much.

Assumption 5 The pairs \((x_i, y_i)\) are independent draws from some distribution, with the first two moments of \(x_i\) finite.

The (ordinary) least squares estimator for \(\beta\) solves

\[
\min_\beta \sum_{i=1}^n (y_i - \beta' x_i)^2.
\]

This leads to

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

Note that we adopt an alternative notation for the OLS estimator \(\hat{\beta}\) instead of \(b\). In the sequel I will also use \(\hat{\varepsilon}\) for \(e\), and \(\hat{\sigma}^2\) for \(s^2\). The (exact) distribution of the OLS estimator is

\[
\hat{\beta} \sim \mathcal{N} \left( \beta, \sigma^2 \cdot (X'X)^{-1} \right).
\]

Without the normality of the \(\varepsilon\) it is difficult to derive the exact distribution of \(\hat{\beta}\). However, we have

\[
\hat{\beta} = \beta + \left( \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i.
\]
Hence, even under assumption 5, $E(x_i \varepsilon_i) = 0$, so that by the Law of Large Numbers

$$\frac{1}{n} \sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} 0$$

and

$$\frac{1}{n} \sum_{i=1}^{n} x_i x'_i \xrightarrow{p} E(xx')$$

so that

$$\hat{\beta} \xrightarrow{p} \beta$$

Also, under assumption 5 (and all stronger assumptions 1-4) and a second moment condition on $\varepsilon$ (variance finite and equal to $\sigma^2$), we can establish asymptotic normality by the Central Limit Theorem.

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x'_i\right)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{d} N\left(0, \sigma^2 \cdot E[xx']^{-1}\right).$$

Typically we do not know $\sigma^2$. We can estimate it as

$$\hat{\sigma}^2 = \frac{1}{n - K - 1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2.$$  

Dividing by $n - K - 1$ rather than $n$ corrects for the fact that $K + 1$ parameters are estimated before calculating the residuals $\hat{\varepsilon}_i = y_i - \hat{\beta}' x_i$. This correction does not matter in large samples, and in fact the maximum likelihood estimator

$$\hat{\sigma}^2_{ml} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2,$$

is a perfectly reasonable alternative. If assumption 1 holds then

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$$
So in practice, whether we have asymptotic normality or not, we will use the following
distribution for \( \hat{\beta} \):

\[
\hat{\beta} \approx \mathcal{N}(\beta, \hat{V}),
\]

where

\[
V = \frac{\sigma^2}{n} \cdot (\mathbb{E}[xx'])^{-1} = \sigma^2 \cdot (\mathbb{E}[XX])^{-1},
\]

and

\[
\hat{V} = \hat{\sigma}^2 \cdot \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1}.
\]

Often we are interested in one particular coefficient. Suppose for example we are inter-
ested in \( \beta_k \). In that case we have

\[
\hat{\beta}_k \approx \mathcal{N}(\beta_1, \hat{V}_{kk}),
\]

where \( V_{ij} \) is the \((i, j)\) element of the matrix \( V \). We can use this for constructing confidence
intervals for a particular coefficient. For example, a 95\% confidence interval for \( \beta_1 \) would be

\[
\left( \hat{\beta}_k - 1.96 \cdot \sqrt{\hat{V}_{kk}}, \hat{\beta}_k + 1.96 \cdot \sqrt{\hat{V}_{kk}} \right).
\]

We can also use this to test whether a particular coefficient is equal to some preset number.
For example, if we want to test whether \( \beta_k \) is equal to 0.1, we construct the t-statistic

\[
t = \frac{\hat{\beta}_k - 0.1}{\sqrt{\hat{V}_{kk}}},
\]

and compare it to a normal distribution.

Let us look at some real data. The following regressions are estimated on data from
the National Longitudinal Survey of Youth (NLSY). The data set used here consists of 935
observations on usual weekly earnings, years of education, and experience (calculated as age minus education minus six). Table 1 presents some summary statistics for these 935 observations. The particular data set consists of men between 28 and 38 years of age at the time the wages were measured.

We will use these data to look at the returns to education. Mincer developed a model that leads to the following relation between log earnings, education and experience for individual $i$:

$$\log(earnings)_i = \beta_1 + \beta_2 \cdot educ_i + \beta_3 \cdot exper_i + \beta_4 \cdot exper_i^2 + \epsilon_i.$$  

Estimating this on the NLSY data leads to

$$\hat{\log(earnings)}_i = 4.016 + 0.092 \cdot educ_i + 0.079 \cdot exper_i - 0.002 \cdot exper_i^2$$

(0.222) (0.008) (0.025) (0.001)

The estimated standard deviation of the residuals is $\hat{\sigma} = 0.41$. In brackets are the standard errors for the parameters estimates, based on the square roots of the diagonal elements of the variance estimate (3).

Using the estimates and the standard errors we can construct the confidence intervals. For example, a 95% confidence interval for the returns to education, measured by the parameter
\( \beta_2, \) is

\[
CI_{0.95} = (0.0923 - 1.96 \cdot 0.008, 0.0923 + 1.96 \cdot 0.008) = (0.0775, 0.1071).
\]

The t-statistic for testing \( \beta_2 = 0.1 \) is

\[
t = \frac{0.0923 - 0.1}{0.008} = 1.0172,
\]

so at the 90% level we do not reject the hypothesis that \( \beta_2 \) is equal to 0.1.

Now suppose we wish to use these estimates for predicting a more complex change. For example, suppose we want to see what the estimated effect is on the log of weekly earnings of increasing a person’s education by one year. Because changing an individual’s education also changes their experience (in this case it automatically reduces it by one year), this effect depends not just on \( \beta_1 \). To make this specific, let us focus on an individual with twelve years of education (high school), and ten years of experience (so that exper^2 is equal to 100). The expected value of this person’s log earnings is

\[
\hat{\log(earnings)} = 4.016 + 0.092 \cdot 12 + 0.079 \cdot 10 - 0.002 \cdot 100 = 5.7191
\]

Now change this person’s education to 13. Their experience will go down to 9 and exper^2 will go down to 81. Hence the expected log earnings is

\[
\hat{\log(earnings)} = 4.016 + 0.092 \cdot 13 + 0.079 \cdot 9 - 0.002 \cdot 81 = 5.7696
\]

The difference is \( \hat{\beta}_2 - \hat{\beta}_3 - 19 \cdot \hat{\beta}_4 = 0.0505 \). Hence the expected gain of an additional year of education, taking into account the effect on experience and experience squared is the difference between these two predictions, which is equal to 0.051. Now the question is what the standard error for this prediction is. The general way to answer this question is as follows. The vector of estimated coefficients \( \hat{\beta} \) is approximately normal with mean \( \beta \) and variance \( V \).
We are interested in a linear combination of the $\beta$'s, namely $\beta_2 - \beta_3 - 19 \cdot \beta_4 = \lambda \beta$, where $\lambda = (0, 1, -1, -19)'$. Therefore

$$\lambda' \hat{\beta} \sim N(\lambda' \beta, \lambda' V \lambda),$$

where $V$ is the variance in equation (2). In the above example, we have the following values for the covariance matrix $V$:

$$V = \begin{pmatrix}
0.0494 & -0.0011 & -0.0047 & 0.0001 \\
0.0001 & 0.000 & 0.0000 \\
0.0006 & 0.0000 & 0.0000 \\
0.0006 & 0.0000 & 0.0000 \\
\end{pmatrix}.$$  

Hence the standard error of $\lambda' \hat{\beta}$ is 0.0096, and the 95% confidence interval for $\lambda' \beta$ is $(0.0317, 0.0693)$.

The second method for getting an estimate and standard error for $\lambda' \beta$ is very easy in the linear case. We are interested in an estimator for $\lambda' \beta$. To analyze this we reparametrize from

$$\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\end{pmatrix} \quad \text{to} \quad \begin{pmatrix}
\gamma = \beta_2 - \beta_3 - 19 \cdot \beta_4 \\
\beta_3 \\
\beta_4 \\
\end{pmatrix},$$

The inverse of the transformation is

$$\begin{pmatrix}
\beta_1 \\
\beta_2 = \gamma + \beta_3 + 19 \cdot \beta_4 \\
\beta_3 \\
\beta_4 \\
\end{pmatrix}.$$

Hence we can write the regression function as

$$\log(\text{earnings})_i = \beta_1 + (\gamma + \beta_3 + 19 \cdot \beta_4) \cdot \text{educ}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}^2_i + \varepsilon_i$$

$$= \beta_1 + \gamma \cdot \text{educ}_i + \beta_2 \cdot (\text{exper}_i + \text{educ}_i) + \beta_4 \cdot (\text{exper}^2_i + 19\text{educ}_i) + \varepsilon_i.$$
Hence to get an estimate for $\gamma$ we can regress log earnings on a constant, education, experience minus education and experience-squared minus 19 times education. This leads to the estimated regression function

$$\hat{\text{log(earnings)}}_i = 4.016 + 0.051 \cdot \text{educ}_i + 0.079 \cdot (\text{exper}_i + \text{educ}_i) - 0.002 \cdot (\text{exper}_i^2 + 19 \cdot \text{educ}_i).$$

(0.222) (0.010) (0.025) (0.001)

Now we obtain the estimate and standard error directly from the regression output.

Let us also look at a nonlinear version of this. Suppose we are interested in the regression of log earnings on education. The estimated regression function is

$$\hat{\text{log(earnings)}}_i = 5.0455 + 0.0667 \cdot \text{educ}_i.$$

(0.0849) (0.0062)

The estimate for $\sigma^2$ is $\hat{\sigma}^2 = 0.1744$. Now suppose we are interested in the average effect of increasing education by one year for an individual with currently eight years of education, not on the log of earnings, but on the level of earnings. At $x$ years of education the expected level of earnings is

$$\mathbb{E}[\text{earnings}|\text{educ} = x] = \exp(\beta_1 + \beta_2 \cdot x + \sigma^2/2),$$

using the fact that if $Z \sim N(\mu, \sigma^2)$, then $\mathbb{E}[\exp(Z)] = \exp(\mu + \sigma^2/2)$.

Hence the parameter of interest is

$$\theta = \exp(\beta_1 + \beta_2 \cdot 9 + \sigma^2/2) - \exp(\beta_1 + \beta_2 \cdot 8 + \sigma^2/2).$$

Getting an estimate for $\theta$ is easy. Just plug in the estimates for $\beta$ and $\sigma^2$ to get:

$$\hat{\theta} = \exp(\hat{\beta}_1 + \hat{\beta}_2 \cdot 9 + \hat{\sigma}^2/2) - \exp(\hat{\beta}_1 + \hat{\beta}_2 \cdot 8 + \hat{\sigma}^2/2) = 19.9484.$$
However, we also need a standard error for this estimate. Let us write this more generally as \( \theta = g(\gamma) \), where \( \gamma = (\beta', \sigma^2)' \). We have an approximate distribution for \( \gamma \):

\[
\sqrt{n}(\hat{\gamma} - \gamma) \approx N(0, \Omega).
\]

Then by the Delta method,

\[
\sqrt{n}(g(\hat{\gamma}) - g(\gamma)) \approx N\left(0, \frac{\partial g}{\partial \gamma}' \Omega \frac{\partial g}{\partial \gamma}\right).
\]

In this case,

\[
\frac{\partial g}{\partial \gamma} = \begin{pmatrix}
\exp(\beta_0 + \beta_1 \cdot 9 + \sigma^2/2) - \exp(\beta_0 + \beta_1 \cdot 8 + \sigma^2/2) \\
9 \cdot \exp(\beta_0 + \beta_1 \cdot 9 + \sigma^2/2) - 8 \cdot \exp(\beta_0 + \beta_1 \cdot 8 + \sigma^2/2) \\
\frac{1}{2} \exp(\beta_0 + \beta_1 \cdot 9 + \sigma^2/2) - \frac{1}{2} \exp(\beta_0 + \beta_1 \cdot 8 + \sigma^2/2)
\end{pmatrix}.
\]

We estimate this by substituting estimated values for the parameters, so we get

\[
\frac{\hat{\partial g}}{\partial \gamma} = \begin{pmatrix}
\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 9 + \hat{\sigma}^2/2) - \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 8 + \hat{\sigma}^2/2) \\
9 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 9 + \hat{\sigma}^2/2) - 8 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 8 + \hat{\sigma}^2/2) \\
\frac{1}{2} \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 9 + \hat{\sigma}^2/2) - \frac{1}{2} \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 8 + \hat{\sigma}^2/2)
\end{pmatrix} = \begin{pmatrix}
19.9484 \\
468.5779 \\
9.9742
\end{pmatrix}.
\]

To get the variance for \( \hat{\theta} = g(\hat{\gamma}) \), we also need the full covariance matrix, including for the parameter \( \sigma^2 \). Using the fact that because of the normal distribution the estimator for \( \sigma^2 \) is independent of the estimators of the other parameters, and that it has asymptotic variance equal to \( 2\sigma^4 \), we have

\[
\hat{\Omega} = \begin{pmatrix}
6.7382 & -0.4873 & 0.0000 \\
0.0362 & 0.0000 & 0.0608
\end{pmatrix}.
\]

Hence the variance for the parameter of interest is

\[
\frac{1}{n} \cdot \frac{\partial g}{\partial \gamma}(\hat{\gamma})' \hat{\Omega} \frac{\partial g}{\partial \gamma}(\hat{\gamma}) = \frac{1}{n} \cdot \begin{pmatrix}
19.9484 \\
468.5779 \\
9.9742
\end{pmatrix}' \begin{pmatrix}
6.7382 & -0.4873 & 0.0000 \\
0.0362 & 0.0000 & 0.0608
\end{pmatrix} \begin{pmatrix}
19.9484 \\
468.5779 \\
9.9742
\end{pmatrix} = 1.2756^2.
\]
Finally, suppose we are interested in the average effect on the level of earnings of increasing education levels by one year. That is, for each individual we estimate the effect of increasing their education level by one year, from whatever level it was, followed by averaging over all individuals. In terms of the parameters of the linear regression model, the parameter of interest is now a much messier function:

$$\theta = g(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \left( \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) - \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) \right).$$

Substituting estimated values for the parameters leads to

$$\hat{\theta} = g(\hat{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \left( \exp(\hat{\beta}_1 + \hat{\beta}_2 \cdot \text{educ}_i + \hat{\sigma}^2/2) - \exp(\hat{\beta}_1 + \hat{\beta}_2 \cdot \text{educ}_i + \hat{\sigma}^2/2) \right) = 29.0527.$$

Even though this is a much messier function, the principle is the same. We already have the covariance matrix $\Omega$ for $\beta$ and $\sigma^2$, we just need the derivatives of the new transformation $g(\gamma)$:

$$\frac{\partial g}{\partial \gamma} = \frac{1}{n} \sum_{i=1}^{n} \left( \begin{array}{c} \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) - \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) \\
(\text{educ}_i + 1) \exp(\beta_1 + \beta_2 (\text{educ}_i + 1) + \sigma^2/2) - \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) \\
\frac{1}{2} \exp(\beta_1 + \beta_2 \cdot (\text{educ}_i + 1) + \sigma^2/2) - \frac{1}{2} \exp(\beta_1 + \beta_2 \cdot \text{educ}_i + \sigma^2/2) \end{array} \right).$$

The standard error in this case is 2.8895.