Econometrics II

Chapter 2
Ordinary Least Squares and Linear Regression Model

Marie Paul

University of Duisburg-Essen
Ruhr Graduate School in Economics

Summer Semester 2020

Outline of lecture

2. Ordinary Least Squares and Linear Regression Model
   - Conditional Expectation Function and Best Linear Predictor
   - The OLS estimator
   - The Linear Regression Model
   - Recap: Interpreting OLS Results

Self-study plan for Lecture II

- This lecture shall repeat OLS from a microeconometric, practical point of view.
- For our second virtual meeting, please work through all slides of Lecture II.
- Please read in Cameron and Trivedi (2005) in Chapter 4 (Linear Models): 4.1. Introduction, 4.2. Regressions and Loss Functions, 4.3. Example: Returns to Schooling, 4.4. Ordinary Least Squares (as far as the text relates to the slides).
- Please read Section 2.5, 2.15, 2.18, 2.19 in Hansen (2019) attached to my email (as far as the text relates to the slides).
- On the slides, you sometimes find questions under the heading "your own research". These questions relate to your own research or your research ideas. Write down your answers and discuss them with at least one fellow PhD student.
- Collect remaining questions and topics to discuss them in our virtual meeting.

Outline

Loss function

In general terms, **regression** refers to a **set of procedures** to study the relationship between an **outcome** variable y and a set of **regressors x**.

Think of the purpose of regression to be **conditional prediction** of y given **x**, denoted $\widehat{y}$. Let $e \equiv y - \widehat{y}$ denote the **prediction error** and

$$\mathsf{L}(e) = \mathsf{L}(y - \widehat{y}), \tag{2.1}$$

the **loss** associated with the error e for some **decision maker** ($\partial \mathsf{L}/\partial e > 0$).

With $(y, \widehat{y})$ random, the decision maker want to minimize the **expected value** of the loss function, $\mathsf{E}[\mathsf{L}(e)]$. If the prediction depends on **x**, the expected loss is

$$\mathsf{E}[\mathsf{L}(e|\mathbf{x})]. \tag{2.2}$$

The question is: **Which** loss function $\mathsf{L}(\cdot)$ should be chosen?

Loss function

The **general** answer is that it depends on the particular decision maker's own **preferences** (or the **disutility** s/he derived from the prediction error).

The **practical** answer is the **convention** that $\mathrm{L}(e) = e^2$, i.e., **squared loss**:

$$\mathrm{E}[(y - \widehat{y})^2 | \mathbf{x}]. \tag{2.3}$$

Hence, the (risk-averse) decision maker becomes **exponentially** more unhappy as the prediction error increases.

The choice of loss function is not innocuous since it will determine the **optimal predictor**.

The optimal predictor of y under squared loss is the **conditional expectation function (CEF)**:

$$\arg\min_{\widehat{y}} \mathrm{E}[(y - \widehat{y})^2 | \mathbf{x}] = \mathrm{E}[y | \mathbf{x}]. \tag{2.4}$$

The CEF is the best unrestricted predictor of the dependent variable.

Optimal predictor

Example 2.1 (Optimal predictor under squared loss)

We want to minimize the expected prediction error $e = (y - \widehat{y})$ given $\mathbf{x}$ under squared loss, $\mathsf{L}(e) = e^2$, i.e.:

$$\min_{\widehat{y}} \mathsf{E}[\mathsf{L}(e|\mathbf{x})] = \min_{\widehat{y}} \mathsf{E}[(y - \widehat{y})^2|\mathbf{x}] \tag{2.5}$$

Define $\mathsf{E}[y|\mathbf{x}] = \mu_x$. Basic algebra yields

$$\begin{aligned}
\mathsf{E}[(y - \widehat{y})^2|\mathbf{x}] &= \mathsf{E}[((y - \mu_x) + (\mu_x - \widehat{y}))^2|\mathbf{x}] \\
&= \mathsf{E}[(y - \mu_x)^2 + (\mu_x - \widehat{y})^2 + 2(y - \mu_x)(\mu_x - \widehat{y})|\mathbf{x}] \\
&= \mathsf{Var}[y|\mathbf{x}] + (\mu_x - \widehat{y})^2 + 2(\mu_x - \mu_x)(\mu_x - \widehat{y}) \\
&= \mathsf{Var}[y|\mathbf{x}] + (\mu_x - \widehat{y})^2,
\end{aligned} \tag{2.6}$$

implying that we should set $\widehat{y} = \mu_x$ to minimize expected loss.

Optimal predictor

In the most general case **no structure** is placed on $\mathsf{E}[y|\mathbf{x}]$ and estimation is by nonparametric regression. But usually a parametric **model** is specified for $\mathsf{E}[y|\mathbf{x}]$:

$$\mathsf{E}[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta}), \tag{2.7}$$

where $g(\cdot)$ is a known function and $\boldsymbol{\beta}$ is a vector of **parameters** to be estimated.

The optimal prediction is then equal to $\widehat{y} = g(\mathbf{x}, \widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}}$ is chosen according to

$$\sum_{i=1}^{N} \mathsf{L}(e_i) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2. \tag{2.8}$$

to minimize the sum of squared prediction errors in a sample of N observations.

Linear Regression Model

If the conditional expectation function (CEF) $\mathrm{E}[y|\mathbf{x}]$ is linear in $\mathbf{x}$, so that $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ we have the linear regression model:

$$y = \mathbf{x}'\boldsymbol{\beta} + u, \tag{2.9}$$

$$\mathrm{E}[u|\mathbf{x}] = 0 \tag{2.10}$$

Usually $\mathbf{x}$ includes a one so that the model includes an intercept.
The linear prediction here coincides with the CEF. The parameter $\boldsymbol{\beta}$ has **structural or causal interpretation** and consistent estimation of $\boldsymbol{\beta}$ by OLS implies consistent estimation of $\mathrm{E}[y|\mathbf{x}]$.

The linear projection model

Very generally, the **best linear predictor coefficient** of y given **x** under squared error loss can be found by minimizing the loss equation and this gives

$$\beta = (\mathrm{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathrm{E}[\mathbf{x}y]. \tag{2.11}$$

which is the sample analogue of the OLS estimator. **x** again includes an intercept. The error is $u = y - \mathbf{x}'\beta$ and rewritting gives the decomposition of y into linear predictor and error:

$$y = \mathbf{x}'\beta + u. \tag{2.12}$$

This equation is called the linear projection model. We call $\mathbf{x}'\beta$ is the **best linear predictor** of y given **x** or the **linear projection** of y onto **x**.

The linear projection model needs only mild regularity conditions. It can be shown that $\mathrm{E}[u] = 0$ and $\mathrm{Cov}[\mathbf{x}, u] = 0$ are implied by the first order conditions. Thus, the **linear regression is a special case of the projection model**, but not vice versa because the linear projection model does not necessarily satisfy $\mathrm{E}[u|\mathbf{x}] = 0$.

Linear prediction and regression

If we have a **linear CEF** and estimate it by OLS we can thus always give it the **nonstructural reduced form interpretation** of the best linear prediction.

But a **structural interpretation** of OLS requires that the conditional mean of the error term, given regressors, equals zero: $\mathrm{E}[u|\mathbf{x}] = 0$. This is not generally true.

The linear projection model may also be used as an approximation for a **non-linear CEF**: $\mathrm{E}[y|\mathbf{x}] \neq \mathbf{x}'\boldsymbol{\beta}$ being the best linear predictor.

But note that the coefficient definition and the linear equation in the linear projection model comes from **the definition of choosing the best linear predictor**. In an economic model $\boldsymbol{\beta}$ might be defined differently and then interpreting the coefficient from the linear prediction model might not be useful.

Linear Projection as an Approximation

Example 2.2 (taken from Hansen 2019)

$$\mathrm{E}[lwage|experience] = m(experience) \tag{2.13}$$

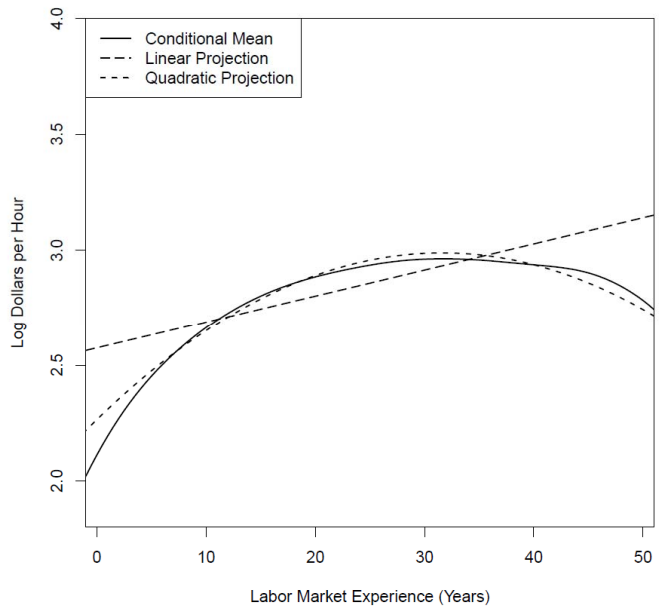

Figure 2.9: Linear and Quadratic Projections of log($wage$) onto Experience

Linear Projection as an Approximation

Example 2.3 (Returns to Schooling)

$$lnwage_i = \alpha s_i + \mathbf{x}_{2i}'\boldsymbol{\beta} + u_i. \tag{2.14}$$

$\widehat{\alpha} = 0.10$.

This regression can be used in a descriptive way: a one-year increase in schooling is associated with 10% higher earnings, controlling for all the factors included in $\mathbf{x}_2$.

Policy interest is in the impact of an exogenous change in schooling on earnings, i.e. the causal effect of an additional year of schooling. But because schooling is an endogenous variable, the regression is not causally meaningful.

Outline

Model specification

The simplest example of regression is the OLS estimator in the **linear regression model**.

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \tag{2.15}$$

where the subscript i is often dropped.
Or in matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.16}$$

The OLS estimator is

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.17}$$

Identification

We focus on the ability of the OLS estimator to permit **identification of the CEF** $\mathrm{E}[y|\mathbf{x}]$. For the linear model the parameter $\boldsymbol{\beta}$ is identified if

$$\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} \tag{2.18}$$

and

$$\mathbf{X}\boldsymbol{\beta}^{(1)} = \mathbf{X}\boldsymbol{\beta}^{(2)} \text{if and only if} \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}. \tag{2.19}$$

The first condition says that the conditional mean is correctly specified and the second implies that $\mathbf{X}'\mathbf{X}$ is nonsingular which is needed to compute the OLS estimate.

Distribution of the OLS estimator

Recap of Econometrics I:
Assume the data generating process (dgp) is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ so that the model is correctly specified plus some additional assumptions. Then the OLS estimator is consistent plim $\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}$ under the condition $E[\mathbf{x}_i, u_i] = 0$.

The asymptotic distribution of the OLS estimator (limits and expectations dropped) is

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{OLS}) \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}), \tag{2.20}$$

Heteroskedastic-consistent estimate of the variance matrix:

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}_{OLS}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{2.21}$$

Outline

Assumptions for Cross-Section Regression

Set of **assumptions appropriate for many applied settings** (made by White, 1980).

More general than in an **introductory treatment of OLS** which presents often **too restrictive assumptions** for applications with micro survey data.

Discuss the most important assumptions. See **Cameron and Trivedi (2005)** for the complete list of assumptions.

Stratified Random Sampling

Assumption 1: The data $(y_i, \mathbf{x}_i)$ are independent and not identically distributed (inid) over i.

Permits distribution to differ over i, e.g. stratified random sampling.

Rules out clustering (consistency still holds, but need different variance matrix).

On Sampling schemes

The iid assumption is unlikely in many practical applications. A few examples we will return to later:

1. **Pooled cross sections**: If data is sampled over several time periods and the sample distribution changes over time (**inid**).
   - But violated in time-series/panel data due to **autocorrelation**.
2. **Stratified** sampling: Individuals are sampled **within** groups (strata) with different distributions/dependency structures.
   - Stratified **random** sampling: Usually no problem for consistency.
   - **Cluster** sampling: Strata are randomly sampled but individuals **within** strata are correlated. Need to adapt standard errors.
   - **Spatial** dependence: Dependence across nearby geographical units. Need to adapt standard errors.
3. Sample **selection**: Non-random sampling where some **selection mechanism** governs sampling (**truncation** or survey non-response).
4. Attrition: Non-random **censoring** of individuals in longitudinal data.

Correctly Specified Model

Assumption 2: The model is correctly specified so that $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$.

Seems obvious, but refers also to the correspondence of the model and the data. Apart from stating that the model is **linear** in **x** and $\boldsymbol{\beta}$s are the same across individuals, it says there are **no omitted variables** in the regression, and **no measurement error** in the regressors.
If assumption 2 fails then OLS can only be interpreted as an optimal linear predictor.

Weakly Exogenous Regressors

Assumption 4: The errors have zero mean conditional on regressors $E[u_i|\mathbf{x}_i] = 0$.

The **zero conditional means assumption** combined with Assumption 2 implies that $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ so that **the conditional means is indeed** $\mathbf{x}'\boldsymbol{\beta}$.

All factors in the error term are uncorrelated with **x** (and any function of **x**) and have on average zero impact on y.

Implies $Cov[\mathbf{x}, u] = 0$ which is sufficient for consistency. To estimate **partial effects** of each x_j on $E[y|\mathbf{x}]$ over a broad range of values for **x**, we want $E[u_i|\mathbf{x}_i] = 0$.

If the assumption $E[u|\mathbf{x}] = 0$ fails, we have endogenous regressors and OLS is inconsistent.

On Omitted variables

Suppose the dgp is again linear but has the following form

$$y = \mathbf{x}'\boldsymbol{\beta} + z\alpha + \nu, \tag{2.22}$$

where z is a scalar regressor and x and z are uncorrelated with ν. Suppose that y is regressed on **x only** with z **omitted**. This means that

$$y = \mathbf{x}'\boldsymbol{\beta} + \underbrace{(z\alpha + \nu)}_{\mu}, \tag{2.23}$$

so that the error term is now instead μ.

If z is correlated with $\mathbf{x}$, then $\mathrm{E}[\mu\mathbf{x}] \neq 0$ and OLS will be inconsistent for $\boldsymbol{\beta}$ since

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \underbrace{(N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{z}}_{\delta}\,\alpha + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\boldsymbol{\nu}), \tag{2.24}$$

which yields

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \delta\alpha, \tag{2.25}$$

where the sign of δ depend on α and the correlation between $\mathbf{z}$ and $\mathbf{X}$.

Endogeneity

Endogeneity is a collection for inconsistency issues arising from a correlation between a **regressor** of interest and the **error** term.

This includes the problems of **omitted variables**, **sample selection**, **measurement error** as well as **reverse causality**.

Different approaches to deal with endogeneity issues are **instrumental variables**, **fixed effects**, **difference-in-differences** and **regression discontinuity** designs.

Additional assumptions

Assumption 3 states that under iid need to assume that second moments exists. Under inid more involved. Assumption 3 is formulated such that stochastic regressors are permitted. This is usually the case in survey data.

Assumption 5 states independent regression errors are assumed. Different variances (heteroscedasticity) are allowed.

Assumption 6 is a technical assumption on the variance matrix.

Identification strategies

Example 2.4 (Your own research)

Discuss the most important assumptions for a linear regression you have estimated e.g. in your Master thesis.

Outline

Interpreting OLS Results

Example 2.5 (Wage Regression 1 based on cross-section from SOEP)

Source	SS	df	MS	Number of obs	=	1,958
				F(1, 1956)	=	73.14
Model	6659.39515	1	6659.39515	Prob > F	=	0.0000
Residual	178099.278	1,956	91.0528006	R-squared	=	0.0360
				Adj R-squared	=	0.0356
Total	184758.673	1,957	94.4091329	Root MSE	=	9.5422

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-3.690785	.4315668	-8.55	0.000	-4.537164	-2.844406
_cons	18.12718	.2996592	60.49	0.000	17.53949	18.71486

What is the average wage of males? Of females? Three ways to see whether coefficient is significant?

Interpreting OLS Results

Example 2.6 (Wage Regression 2 based on cross-section from SOEP)

Source	SS	df	MS		
Model	217.248936	7	31.0355622	Prob > F	0.0000
Residual	350.84896	1,950	.179922543	R-squared	0.3824
				Adj R-squared	0.3802
Total	568.097895	1,957	.290290187	Root MSE	.42417

Number of obs	=	1,958				
F(7, 1950)	=	172.49				
Prob > F	=	0.0000				
R-squared	=	0.3824				
Adj R-squared	=	0.3802				
Root MSE	=	.42417				

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
novoc	-.2476475	.0283773	-8.73	0.000	-.3033006	-.1919944
univ	.4639993	.0244413	18.98	0.000	.4160655	.5119331
experience	.0417119	.0030667	13.60	0.000	.0356975	.0477263
experiencesq	-.0007487	.0000755	-9.92	0.000	-.0008967	-.0006006
east	-.3369118	.0224919	-14.98	0.000	-.3810225	-.2928011
immig	-.1177419	.0314804	-3.74	0.000	-.1794807	-.0560032
female	-.1638751	.019595	-8.36	0.000	-.2023045	-.1254457
_cons	2.359362	.0312044	75.61	0.000	2.298165	2.42056

Interpret the coefficient of high skilled (univ). Why is the coefficient on female lower than before?

Interpreting OLS Results

Example 2.7 (Wage Regression 3 based on cross-section from SOEP)

Source	SS	df	MS	Number of obs	=	1,958
				F(8, 1950)	=	1073.38
Model	577003.445	8	72125.4307	Prob > F	=	0.0000
Residual	131029.093	1,950	67.1944067	R-squared	=	0.8149
				Adj R-squared	=	0.8142
Total	708032.539	1,958	361.610081	Root MSE	=	8.1972

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
novoc	10.12283	.6478173	15.63	0.000	8.85234	11.39331
apprent	12.93475	.6030307	21.45	0.000	11.7521	14.1174
univ	21.67021	.7114998	30.46	0.000	20.27483	23.06559
experience	.4643214	.0592648	7.83	0.000	.3480924	.5805504
experiencesq	-.0072086	.0014587	-4.94	0.000	-.0100694	-.0043477
east	-5.117456	.4346602	-11.77	0.000	-5.969904	-4.265009
immig	-1.962585	.6083646	-3.23	0.001	-3.155698	-.7694718
female	-3.132496	.3786776	-8.27	0.000	-3.875152	-2.389841

What happened here? Two changes....

Interpreting OLS Results

Example 2.8 (Rent Regression)

```
.
. generate toparea_luxbath = toparea * luxbath

. regress rent squarem toparea luxbath toparea_luxbath kitchen

      Source |       SS       df       MS              Number of obs =    2053
-------------+------------------------------           F(  5,  2047) =  493.93
       Model |  67587422.5     5  13517484.5            Prob > F      =  0.0000
    Residual |  56021152.9  2047  27367.4416           R-squared     =  0.5468
-------------+------------------------------           Adj R-squared =  0.5457
       Total |   123608575  2052  60238.0972           Root MSE      =  165.43

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     squarem |    6.479404   .1515729    42.75   0.000     6.182151    6.776657
     toparea |     133.468    27.5207     4.85   0.000     79.49653    187.4395
     luxbath |    70.44169   13.37241     5.27   0.000     44.21674    96.66664
toparea_lux~h |    122.8264   66.13567     1.86   0.063    -6.873787    252.5266
     kitchen |    147.6964   14.23297    10.38   0.000     119.7838     175.609
        cons |     98.4085   10.87756     9.05   0.000     77.07626    119.7407
------------------------------------------------------------------------------
```

Interpret the interaction effect.