Seminar 1: OLS Regression

Simple and Multiple Linear Regression

Giulio Rossetti*

giuliorossetti94.github.io

January 23, 2026

* email: giulio.rossetti.1@wbs.ac.uk

Roadmap

Part 1: Theory

- Exercise 1: Wage and Education
- Exercise 2: Fertility and Education
- Exercise 3: College GPA Prediction

Part 2: Practice (MATLAB)

- Exercise 4: CEO Salaries (ceosal1)
- Exercise 5: CEO Salaries (ceosal2)
- Exercise 6: OLS Unbiasedness (Monte Carlo)

Summary

Roadmap

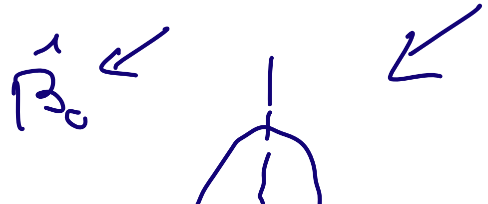
Part 1: Theory

- Exercise 1: Wage and Education
- Exercise 2: Fertility and Education
- Exercise 3: College GPA Prediction

Part 2: Practice (MATLAB)

- Exercise 4: CEO Salaries (ceosal1)
- Exercise 5: CEO Salaries (ceosal2)
- Exercise 6: OLS Unbiasedness (Monte Carlo)

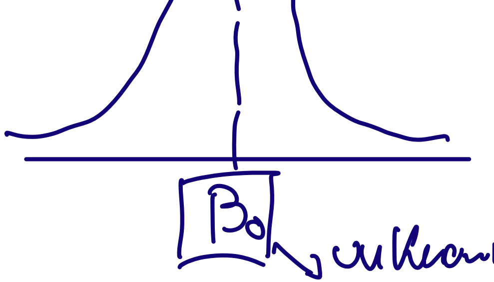
Summary



Exercise 1

Simple Linear Regression Model

Model: $Wage = \beta_0 + \beta_1 Education + u$



Data: UK workforce in 2013 (12 individuals)

Individual	1	2	3	4	5	6	7	8	9	10	11	12
Wage	3.1	3.2	3	6	5.3	8.8	11	5	3.6	18	6.3	8.1
Education	11	12	11	8	12	16	18	12	12	17	16	13

Goal: Estimate β_0 and β_1 using OLS.

$$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$$

Exercise 1

OLS Estimators

For the general SLR model $y = \beta_0 + \beta_1 x + u$:

OLS Formulas

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{cov(x, y)}{var(x)}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\hat{\beta}_0$

$\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Exercise 1

Computing the Estimates

Step 1: Compute sample means: $\bar{y} = 6.78$, $\bar{x} = 13.17$

Step 2: Compute the slope coefficient

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{99.43}{95.67} = 1.04$$

RV

Step 3: Compute the intercept

$$\hat{\beta}_0 = 6.78 - 1.04 \times 13.17 = -6.90$$

RV

Exercise 1

Interpretation

β_1

Estimated Model

$$\widehat{Wage} = -6.90 + 1.04 \times Education$$

Interpretation:

- $\hat{\beta}_1 = 1.04$: An additional year of education is associated with a £1.04 increase in hourly wage.
- $\hat{\beta}_0 = -6.90$: Predicted wage for zero years of education (extrapolation).

$R^2 = 1$

Exercise 1

Goodness-of-Fit: R^2

R-Squared Definition

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \in [0,1]$$

Components:

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 206.48$ (Total Sum of Squares)
- $SSR = \sum_{i=1}^{n}\hat{u}_i^2 = 103.13$ (Residual Sum of Squares)

Result:

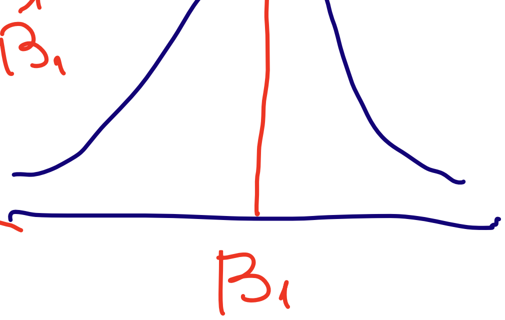
$$R^2 = 1 - \frac{103.13}{206.48} = 0.50$$

Education explains 50% of the variation in wages.

Exercise 1

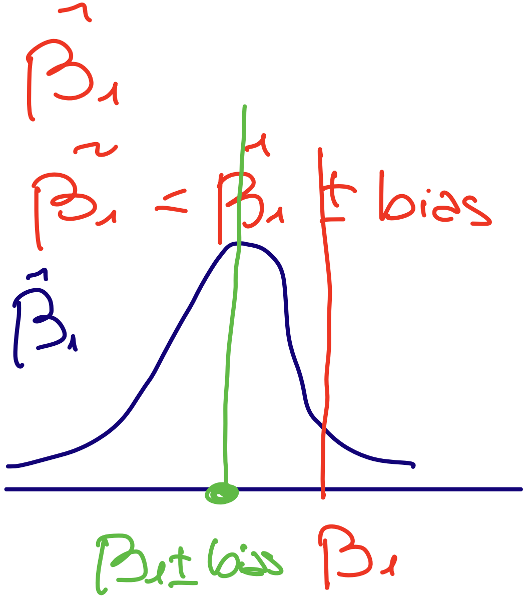
Extending to Multiple Regression

Extended Model: $Wage = \beta_0 + \beta_1 Education + \beta_2 Expertise + u$



Why include more variables?

- MLR investigates the marginal effect of multiple factors
- Holds fixed other factors otherwise hidden in u
- Reduces omitted variable bias



Interpretation of $\hat{\beta}_j$:

- Change in y due to a one-unit increase in x_j, *ceteris paribus*

Exercise 2

Omitted Variable Bias

Model: $Fertility = \beta_0 + \beta_1 Education + u$

Question: What factors are in u? Are they correlated with Education?

Potential factors in u:

- Income
- Intelligence
- Age
- Leisure time

Exercise 2

Why OVB is a Problem

Correlation concerns:

- **Income**: Higher income $\rightarrow$ easier to raise children; correlated with education
- **Intelligence**: Affects both education and fertility decisions

Key Insight

Given potential correlation between $Education$ and factors in u, the ceteris paribus effect is unlikely to be uncovered from this SLR model.

$\Rightarrow$ Omitted Variable Bias (OVB) may arise!

Exercise 3

The Model

Model: $colgpa = \beta_0 + \beta_1 \cdot hsperc + \beta_2 \cdot sat + u$

Variables: $colgpa$ = College GPA; $hsperc$ = HS percentile; sat = SAT score

Source	SS	df	MS
Model	490.606706	2	245.303353
Residual	1303.58897	4134	.315333567
Total	1794.19567	4136	.433799728

Number of obs	= 4137
F(2, 4134)	= 777.92
Prob > F	= 0.0000
Root MSE	= .56155

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsperc	-.0135192	.0005495	-24.60	0.000	-.0145965	-.012442
sat	.0014762	.0000653	22.60	0.000	.0013482	.0016043
_cons	1.391757	.0715424	19.45	0.000	1.251495	1.532018

Exercise 3

OLS Estimates

Estimated Model

$$\widehat{colgpa} = 1.392 - 0.013 \cdot hsperc + 0.0015 \cdot sat$$

Interpretation:

- $\hat{\beta}_2 = 0.0015$: A 1-point increase in SAT raises GPA by 0.0015

Exercise 3

Part (a): Predicted GPA

Question: What is the predicted GPA when $hsperc = 20$ and $sat = 1050$?

Solution:

$$\widehat{colgpa} = 1.392 - 0.013 \times 20 + 0.0015 \times 1050$$
$$= 1.392 - 0.26 + 1.575 = 2.707$$

Interpretation: A student in the top 20% with SAT = 1050 is expected to have a GPA of about 2.7.

Exercise 3

Part (b): GPA Difference Between Students

Question: Students A and B have the same $hsperc$, but A's SAT is 200 points lower. Predicted GPA difference?

Solution: Use the difference equation:

$$\Delta colgpa = \hat{\beta}_1 \cdot \Delta hsperc + \hat{\beta}_2 \cdot \Delta sat$$

With $\Delta hsperc = 0$ and $\Delta sat = -200$:

$$\Delta colgpa = -0.013 \times 0 + 0.0015 \times (-200) = -0.30$$

Interpretation: Student A's GPA is expected to be 0.30 lower.

Exercise 3

Part (c): SAT Difference for GPA Gap

Question: Holding $hsperc$ fixed, what SAT difference leads to a 0.50 GPA difference?

Solution: Set $\Delta colgpa = 0.50$ and $\Delta hsperc = 0$:

$$0.50 = 0.0015 \cdot \Delta sat \quad \Rightarrow \quad \Delta sat = \frac{0.50}{0.0015} = 333.33$$

Interpretation: A student needs a SAT score about 333 points higher to have a GPA 0.5 points above a peer from the same percentile.

Exercise 3

Part (d): Goodness-of-Fit

From STATA output: $SSR = 1303.59$, $SST = 1794.20$

Compute R^2:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{1303.59}{1794.20} = 0.27$$

Interpretation

Only 27% of variation in college GPA is explained by $hsperc$ and SAT.

$\Rightarrow$ Other relevant variables may be omitted from the model.

Roadmap

Part 1: Theory

- Exercise 1: Wage and Education
- Exercise 2: Fertility and Education
- Exercise 3: College GPA Prediction

Part 2: Practice (MATLAB)

- Exercise 4: CEO Salaries (ceosal1)
- Exercise 5: CEO Salaries (ceosal2)
- Exercise 6: OLS Unbiasedness (Monte Carlo)

Summary

Exercise 4

CEO Salaries and Firm Performance

Dataset: `ceosal1.txt` (209 observations, year 1990)

Variables:

- $salary$: CEO salary in thousands of dollars
- roe: Return on equity (average 1988-1990)
- $sales$: Firm sales in millions of dollars

Models to estimate:

1. Simple: $salary = \beta_0 + \beta_1 \cdot roe + u$
2. Multiple: $salary = \beta_0 + \beta_1 \cdot roe + \beta_2 \cdot sales + u$

Exercise 4

MATLAB Code (Part 1: Load Data)

```
clear all
load ceosal1.txt
salary = ceosal1(:,1);
sales = ceosal1(:,3);
roe = ceosal1(:,4);
n = 209;
y = salary;

% Histogram
histogram(salary)
```

Exercise 4

MATLAB Code (Part 2: OLS Estimation)

```
% Simple Linear Regression (SLR)
X1 = [ones(n,1) roe];
betahat1 = inv(X1'*X1)*X1'*y;            % OLS estimator
uhat1 = y - X1*betahat1;                 % Residuals
R2_1 = 1 - uhat1'*uhat1/(var(y)*(n-1)); % R-squared

% Multiple Linear Regression (MLR)
X2 = [ones(n,1) roe sales];
betahat2 = inv(X2'*X2)*X2'*y;            % OLS estimator
uhat2 = y - X2*betahat2;                 % Residuals
R2_2 = 1 - uhat2'*uhat2/(var(y)*(n-1)); % R-squared
```

Exercise 4

Key Formulas in Matrix Form

OLS Estimator

$$\hat{\beta} = (X'X)^{-1}X'y$$

Residuals

$$\hat{u} = y - X\hat{\beta}$$

R-Squared

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{(n-1)\cdot\mathrm{Var}(y)} = 1 - \frac{SSR}{SST}$$

Exercise 4

The Design Matrix

SLR: $X_1 = \begin{bmatrix} 1 & roe_1 \\ 1 & roe_2 \\ \vdots & \vdots \\ 1 & roe_n \end{bmatrix}$

MLR: $X_2 = \begin{bmatrix} 1 & roe_1 & sales_1 \\ 1 & roe_2 & sales_2 \\ \vdots & \vdots & \vdots \\ 1 & roe_n & sales_n \end{bmatrix}$

The column of ones captures the constant term β_0.

Exercise 5

CEO Salaries with Sales and Profits

Dataset: `ceosal2.txt` (177 observations, year 1990)

Variables:

- $salary$: CEO compensation in thousands of dollars
- $sales$: Firm sales in millions of dollars
- $profits$: Firm profits in millions of dollars

Model: $salary = \beta_0 + \beta_1 \cdot sales + \beta_2 \cdot profits + u$

Exercise 5

MATLAB Code

```
clear all
load ceosal2.txt
salary = ceosal2(:,1);
sales = ceosal2(:,7);
profits = ceosal2(:,8);
n = 177;

X = [ones(n,1) sales profits];
K = size(X,2);  % Number of regressors (including constant)
y = salary;

histogram(salary)  % Check distribution

% OLS Estimation
betahat = inv(X'*X)*X'*y;            % OLS estimator
uhat = salary - X*betahat;           % Residuals
R2 = 1 - uhat'*uhat/(var(y)*(n-1)); % R-squared
```

Exercise 5

Tasks

1. **Histograms**: Visualize distributions of salary, sales, profits
   - Check for skewness, outliers

2. **Beta estimates**: $(X'X)^{-1}X'y$
   - $\hat{\beta}_0$: Baseline salary
   - $\hat{\beta}_1$: Effect of sales on salary
   - $\hat{\beta}_2$: Effect of profits on salary

3. R^2: Goodness-of-fit

Exercise 6

Proving OLS is Unbiased via Simulation

Goal: Use Monte Carlo simulation to show $E[\hat{\beta}] = \beta$

Approach:

1. Use $\hat{\beta}$ from Exercise 5 as the “true” β
2. Generate many simulated datasets with known β
3. Estimate $\hat{\beta}$ for each simulation
4. Compare average $\hat{\beta}$ to the true β

Data Generating Process:

$$y_{\text{sim}} = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Exercise 6

MATLAB Code

```
% Monte Carlo Simulation
nobs = 10000;              % Number of simulations
betasim = zeros(nobs, K);
mu = 0;
sigma = 1;

for i = 1:nobs
    e = mu + sigma * randn(n, 1);      % Random errors
    ysim = X * betahat + e;            % Simulated y
    betasim(i,:) = inv(X'*X) * X' * ysim;  % OLS estimate
end

% Compare average estimates to true values
[mean(betasim)' betahat]
```

Exercise 6

Interpreting the Results

Output comparison:

Parameter	$\overline{\hat{\beta}}_{\text{sim}}$	β_{true}
β_0	$\approx \hat{\beta}_0$	$\hat{\beta}_0$
β_1	$\approx \hat{\beta}_1$	$\hat{\beta}_1$
β_2	$\approx \hat{\beta}_2$	$\hat{\beta}_2$

Conclusion

The average of OLS estimates across simulations converges to the true values. This confirms OLS is **unbiased**.

Roadmap

Part 1: Theory

- Exercise 1: Wage and Education
- Exercise 2: Fertility and Education
- Exercise 3: College GPA Prediction

Part 2: Practice (MATLAB)

- Exercise 4: CEO Salaries (ceosal1)
- Exercise 5: CEO Salaries (ceosal2)
- Exercise 6: OLS Unbiasedness (Monte Carlo)

Summary

Summary

Part 1 - Theory:

- **Ex 1:** OLS estimation, R^2, SLR vs MLR
- **Ex 2:** Omitted Variable Bias
- **Ex 3:** Prediction with MLR

Part 2 - Practice:

- **Ex 4:** CEO salaries with ROE and sales
- **Ex 5:** CEO salaries with sales and profits
- **Ex 6:** Monte Carlo simulation

Key: OLS: $\hat{\beta} = (X'X)^{-1}X'y$; R^2 measures goodness-of-fit

Key Formulas

OLS Estimator

$$\hat{\beta}_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)} \quad \text{or} \quad \hat{\beta} = (X'X)^{-1}X'y$$

Goodness-of-Fit

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$$

Prediction

$$\hat{y} = X\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots$$