Simple Regression Model (Part I)

Pedram Jahangiry

JON M. HUNTSMAN
SCHOOL OF BUSINESS
UtahStateUniversity

The Simple Regression Model

Definition of the simple linear regression model:

we are interested in "explaining y in terms of x" or

"studying how y varies with changes in x"

Intercept

Slope parameter

$$y = \beta_0 + \beta_1 x + u$$

Dependent variable

Independent variable

Error term,
Disturbance,
Unobservables,…

TABLE 2.1 Terminology for Simple Regression

Y	*X*
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Interpretation of the simple linear regression model

$$y = \beta_0 + \beta_1 x + u$$

"Studies how y varies with changes in x "

$$\frac{\Delta y}{\Delta x} = \beta_1 \quad \text{as long as} \quad \frac{\Delta u}{\Delta x} = 0$$

By how much does the dependent variable change if the independent variable is increased by one unit?

Interpretation only correct if all other things remain equal when the independent variable is increased by one unit

β_1 is **slope parameter**, holding other factros in u fixed.

β_0 is **intercept parameter**, also called the **constant term**.

The simple linear regression model is rarely applicable in practice (why?) but its discussion is useful!

Example: Soybean yield and fertilizer

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Rainfall,
land quality, ...

Measures the effect of fertilizer on yield, holding all other factors fixed

Example: A simple wage equation

$$wage = \beta_0 + \beta_1 educ + u$$

Labor force experience,
work ethic, intelligence, ...

Measures the change in hourly wage given another year of education, holding all other factors fixed

When is there a causal interpretation? $y = \beta_0 + \beta_1 x + u$

- $E(u) = 0$, As long as β_0 is included in the equation. Without loss of generality. (why?)

- **mean independence** assumption + $E(u) = 0$: Zero conditional mean assumption

$$E(u|x) = 0$$

The explanatory variable must not contain information about the mean of the unobserved factors

Example: wage equation

$$wage = \beta_0 + \beta_1 educ + u$$

e.g. intelligence …

The conditional mean independence assumption is **unlikely** to hold because individuals with more education will also be more intelligent on average.

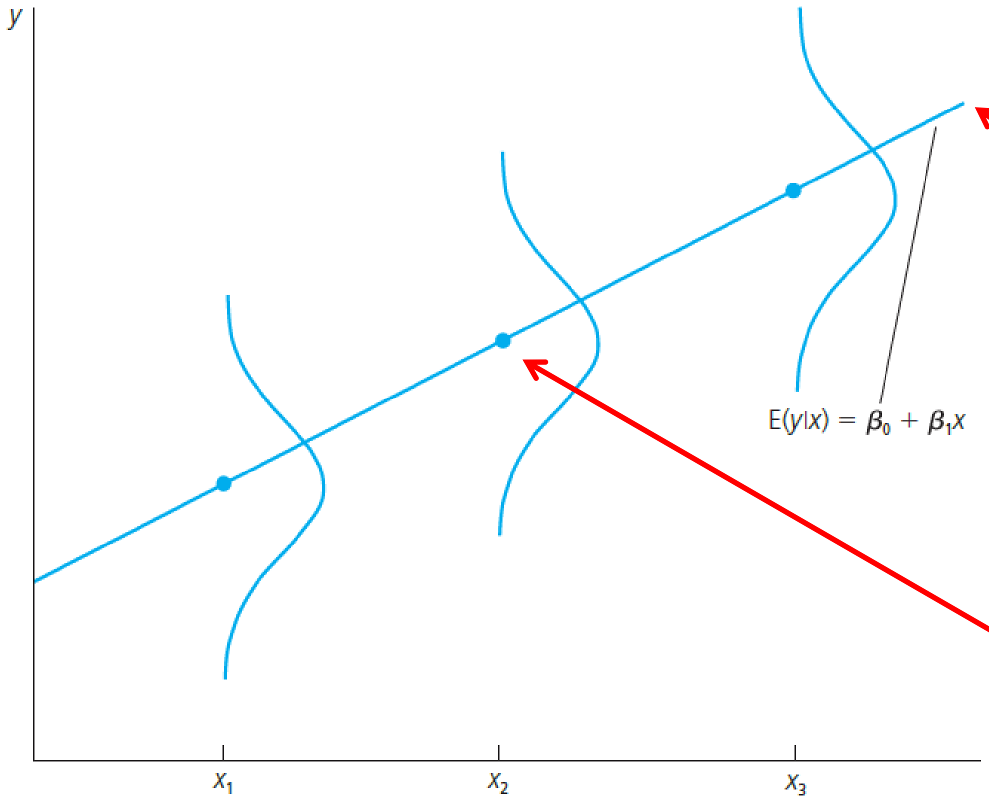
Population regression function (PRF)

The conditional mean independence assumption implies that

$$
\begin{aligned}
E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\
&= \beta_0 + \beta_1 x + E(u|x) \\
&= \beta_0 + \beta_1 x
\end{aligned}
$$

This means that the **conditional average** value of the dependent variable can be expressed as a linear function of the explanatory variable

Population Regression Function (PRF)



Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x_2$

Ordinary Least Squares estimates (OLS)

The purpose of regression analysis is to take a theoretical equation like:

$$y = \beta_0 + \beta_1 x + u$$

And use data to create an estimated equation:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x$$

- ✓ **Ordinary Least Squares (OLS)** is most widely used method to obtain estimates.
- ✓ OLS has become the standard point of reference.

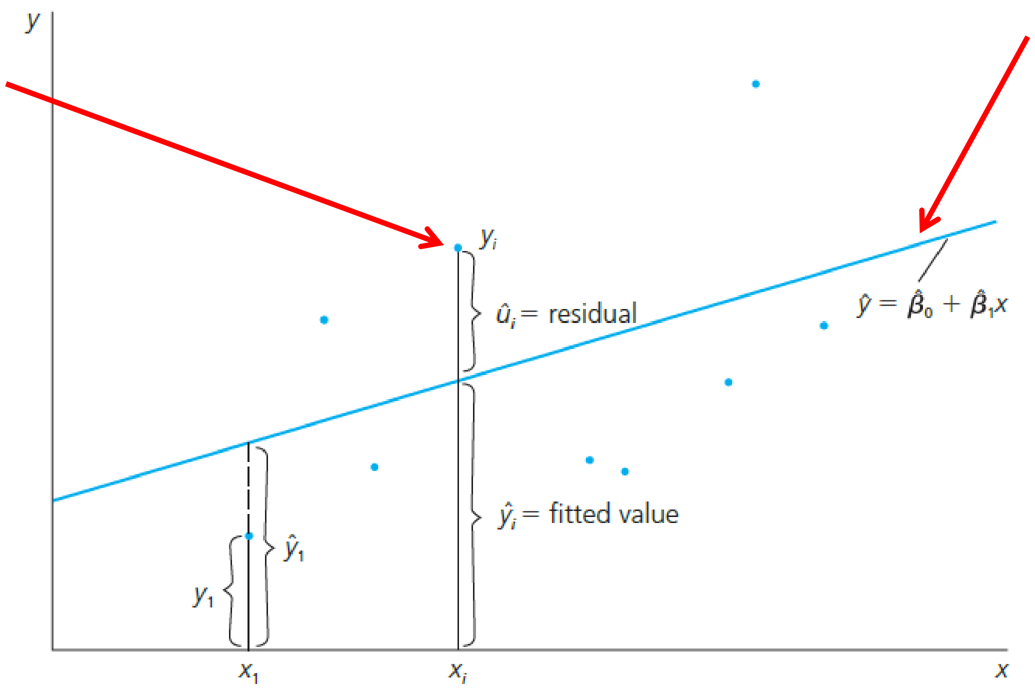
OLS regression line or Sample regression function (SRF)

Fit a regression line (as good as possible) through the data points:

For example, the i-th data point

$$(x_i, y_i)$$

Fitted regression line



Deriving the ordinary least squares estimates (cont'd)

$$y_i = \beta_0 + \beta_1 \, x_i + u_i$$

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \, x_i$$

Minimize sum of squared regression residuals:

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Ordinary Least Squares (OLS) estimates

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

Example: Wage and Education

$$wage = \beta_0 + \beta_1 educ + u$$

Hourly wage in dollars

Years of education

Fitted regression:

$$\widehat{wage} = -0.90 + 0.54\ educ$$

Intercept

In the sample, one more year of education was associated with an increase in hourly wage by $0.54

Intercept interpretation?

Example: CEO Salary and Return on Equity

$$salary = \beta_0 + \beta_1 roe + u$$

Salary in thousands of dollars

Average return on equity of the CEO's firm

Fitted regression

$$\widehat{salary} = 963.191 + 18.501\ roe$$

Intercept

If the return on equity increases by 1 percent, then salary is predicted to change by $18,501

Intercept interpretation?

What is the predicted salary when roe=30?

Terminology: Regressing Salary on ROE

$$\widehat{salary} = 963.191 + 18.501\ roe$$

x_i

y_i

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{u}_i = y_i - \hat{y}_i$

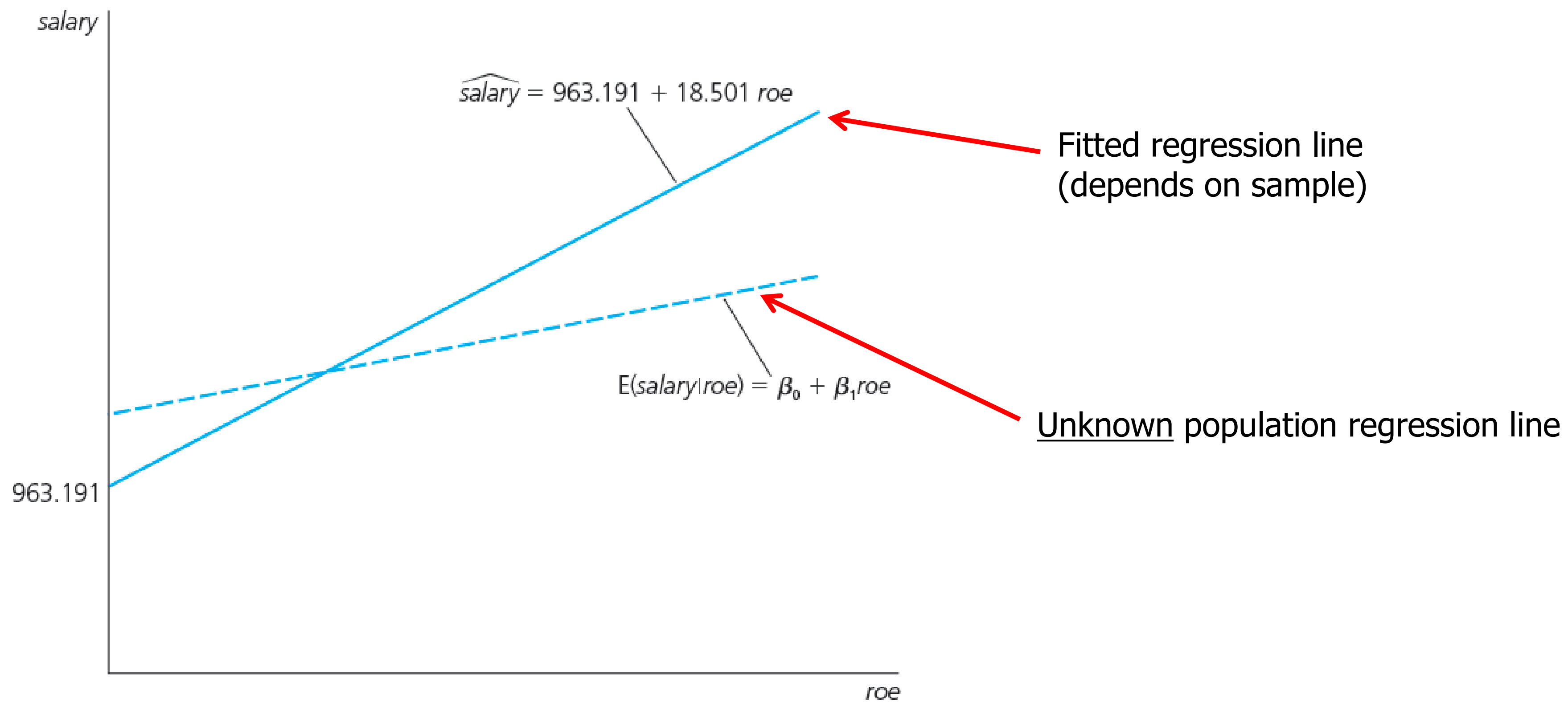
TABLE 2.1 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	–129.0581
2	10.9	1001	1164.854	–163.8542
3	23.5	1122	1397.969	–275.9692
4	5.9	578	1072.348	–494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	–188.2151
7	16.4	1078	1266.611	–188.6108
8	16.3	1094	1264.761	–170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	–616.7726
11	25.9	567	1442.372	–875.3721
12	26.8	933	1459.023	–526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	–438.7678
15	56.3	2011	2004.808	6.191895

For example , CEO number 12's salary was $526,023 lower than predicted **(Over Prediction)**

- ❑ If $\hat{u}_i > 0$, SRF underpredicts y_i
- ❑ If $\hat{u}_i < 0$, SRF overpredicts y_i

CEO Salary and Return on Equity: PRF vs SRF



Properties of OLS on any sample of data

Fitted values and **residuals**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad \widehat{u}_i = y_i - \widehat{y}_i$$

Fitted or predicted values

Deviations from regression line (= residuals)

Algebraic properties of OLS regression:

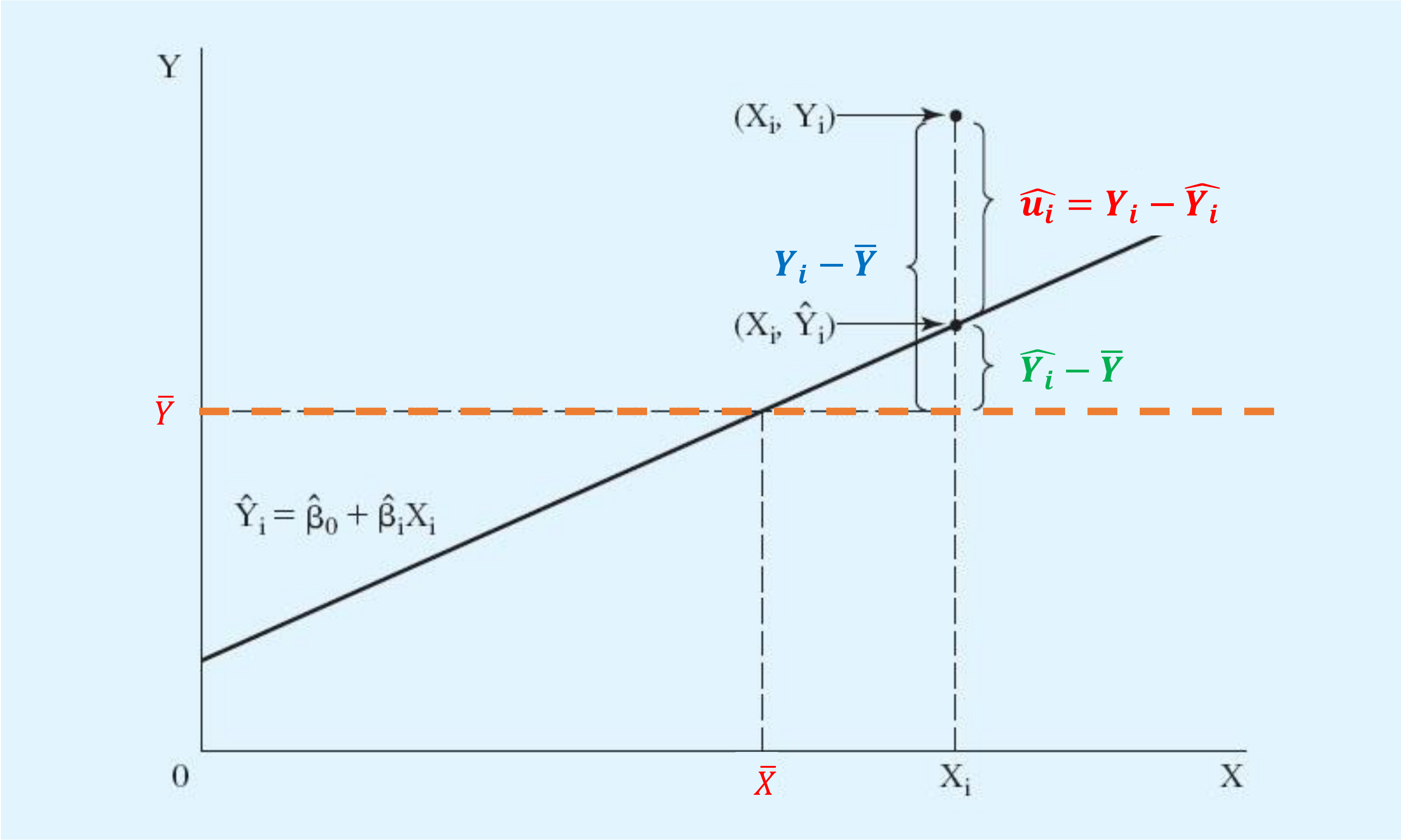
$$\sum_{i=1}^{n} \widehat{u}_i = 0 \qquad \sum_{i=1}^{n} x_i \widehat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$$

Deviations from regression line sum up to zero

Covariance between deviations and regressors is zero

Sample averages of y and x lie on regression line

Decomposition of the variance in y



Measures of Variation

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Total sum of squares,
Represents total sample variation in y

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Explained sum of squares,
Represents variation Explained by regression

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i{}^2$$

Residual sum of squares,
Represents variation **not** Explained by regression

Decomposition of Total Variation

$$SST = SSE + SSR$$

Total variation Explained part Unexplained part

Goodness-of-fit (R^2 or coefficient of determination)

How well does the explanatory variable explain the dependent variable?

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

R-squared measures the fraction of the total variation that is explained by the regression

CEO Salary and return on equity : R^2

$$\widehat{salary} = 963.191 + 18.501\ roe$$

$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3% of the total variation in salaries

Caution: A high R-squared does not necessarily mean that the regression has a causal interpretation!

What happens to coefficients and R^2 if $y \rightarrow \alpha y$? α is a constant

What happens to coefficients and R^2 if $x \rightarrow \alpha x$? α is a constant