Lecture 4

Chapters 1.1-1.4

1.1 The Simple Linear Model

In the stats review, we learned about certain statistics that tell us something about the relationship between two variables:

1. Covariance
2. Correlation

In this chapter will focus on how one variable affects the other through regression analysis.

The simple linear regression model – we hypothesize that X determines Y through some model:

$$\boxed{Y_i} = \beta_1 + \beta_2 \boxed{X_i} + u_i$$

Y- dependent variable

X- explanatory variable, independent variable, or regressor

observable

β_1- intercept

β_2 - slope

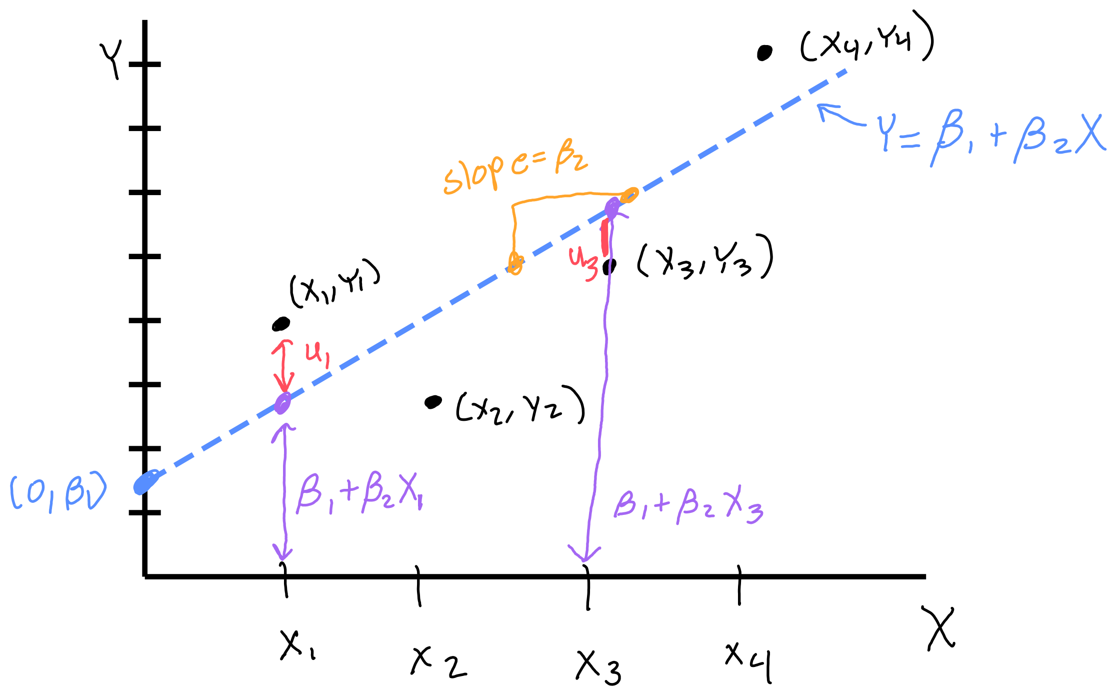
model parameters

unobserved

u – disturbance term

i – observation

Suppose we have the four points graphed below. We want to find the line that best fits the points. Because of the disturbance term, the points will not fall exactly on the line.



1.2 Least Squares Regression

The line that goes through the data is called the fitted regression line:

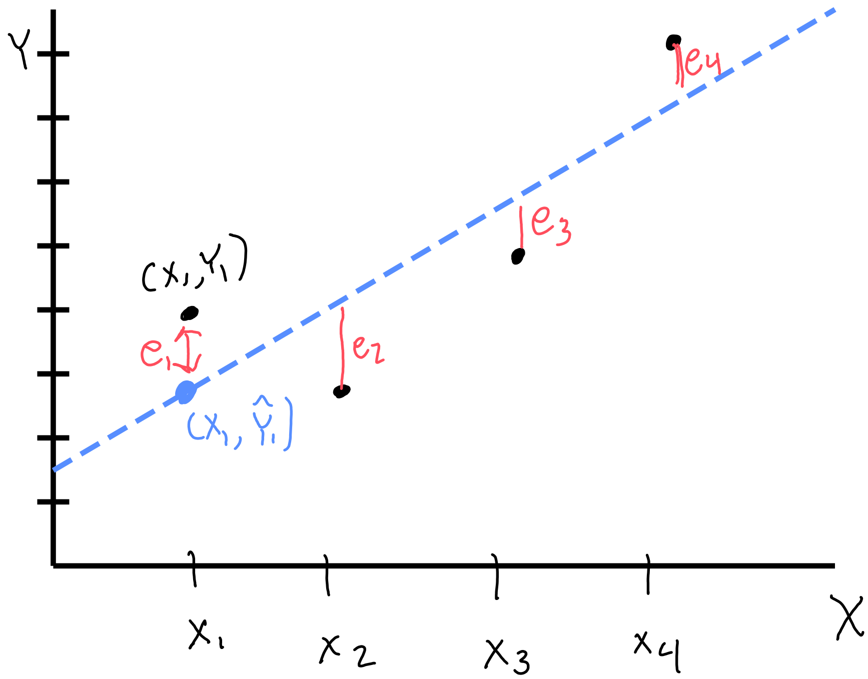
$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2}X_i$$

$\widehat{Y_i}$ – fitted value of Y

$\widehat{\beta_1}$ – estimate of β_1

$\widehat{\beta_2}$ – estimate of β_2

Residual – The difference between the actual value of Y and the fitted value of Y.



Exercise 1: From the definition of residuals, write an equation where residual is denoted as e_i. Simplify the equation so that it does not contain $\widehat{Y_i}$.

$$e_i = Y_i - \hat{Y}_i$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$e_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

$$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

Formula – residual sum of squares, RSS

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

Least Squares Criterion – estimating β_1 and β_2 by minimizing the RSS.

Ordinary least squares estimator – OLS, the estimator that is based on the least squares criterion.

1.3 Derivation of the regression coefficients

Exercise 2: Using the points below and the definitions of residuals and least squares criterion, find the estimates for $\widehat{\beta_1}$ and $\widehat{\beta_2}$.

X	Y	$\hat{Y} = b_1 + b_2 X_i$	$e = Y - \hat{Y}$
1	4	$b_1 + b_2$	$4 - b_1 - b_2$
2	3	$b_1 + 2b_2$	$3 - b_1 - 2b_2$
3	5	$b_1 + 3b_2$	$5 - b_1 - 3b_2$
4	8	$b_1 + 4b_2$	$8 - b_1 - 4b_2$

$$RSS = \sum e_i^2 = (4-b_1-b_2)^2 + (3-b_1-2b_2)^2 + (5-b_1-3b_2)^2 + (8-b_1-4b_2)^2$$

$$= 16 + b_1^2 + b_2^2 - 8b_1 - 8b_2 + 2b_1b_2 + 9 + b_1^2 + 4b_2^2 - 6b_1 - 12b_2 + 4b_1b_2$$

$$+ 25 + b_1^2 + 9b_2^2 - 10b_1 - 30b_2 + 6b_1b_2 + 64 + b_1^2 + 16b_2^2 - 16b_1 - 64b_2 + 8b_1b_2$$

$$RSS = 114 + 4b_1^2 + 30b_2^2 - 40b_1 - 114b_2 + 20b_1b_2$$

$$\frac{\partial RSS}{\partial b_1} = 0 \rightarrow 8b_1 - 40 + 20b_2 = 0$$

$$8b_1 = 40 - 20b_2$$

$$b_1 = 5 - 2.5b_2$$

$$\frac{\partial RSS}{\partial b_2} \rightarrow 60b_2 - 114 + 20b_1 = 0$$

$$60b_2 + 20(5 - 2.5b_2) = 114$$

$$60b_2 + 100 - 50b_2 = 114$$

$$10b_2 = 14$$

$$\boxed{b_2 = 1.4}$$

$$b_1 = 5 - 2.5(1.4)$$

$$b_1 = 5 - 3.5$$

$$\boxed{b_1 = 1.5}$$

Go to desmos example & plot points & regression

Exercise 3: Using the general form of Y, find the normal equations for solving for β_1 and β_2.

$$e_i = Y_i - b_1 - b_2 X_i$$

$$\rightarrow RSS = (Y_1 - b_1 - b_2 X_1)^2 + (Y_2 - b_1 - b_2 X_2)^2 + \dots + (Y_n - b_1 - b_2 X_n)^2$$

$$RSS = (Y_1^2 + b_1^2 + b_2^2 X_1^2 - 2b_1 Y_1 - 2b_2 X_1 Y_1 + 2b_1 b_2 X_1)$$
$$+ \dots + (Y_n^2 + b_1^2 + b_2^2 X_n^2 - 2b_1 Y_n - 2b X_n Y_n + 2b_1 b_2 X_n)$$

$$RSS = \sum Y_i^2 + n b_1^2 + b_2 \sum X_i^2 - 2b_1 \sum Y_i - 2b_2 \sum X_i Y_i + 2b_1 b_2 \sum X_i$$

$$\frac{\partial RSS}{\partial b_1}: \quad 2nb_1 - 2\sum Y_i + 2b_2 \sum X_i = 0$$

$$2nb_1 = 2\sum Y_i - 2b_2 \sum X_i$$

$$\cancel{2}\cancel{n} b_1 = \cancel{2}\cancel{n}\bar{Y} - \cancel{2} b_2 \cancel{n} \bar{X}$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

$$\boxed{\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}}$$

$$\bar{X} = \frac{1}{n}\sum X_i$$
$$\rightarrow n\bar{X} = \sum X_i$$
$$\bar{Y} = \frac{1}{n}\sum Y_i$$
$$\rightarrow n\bar{Y} = \sum Y_i$$

$$\frac{\partial RSS}{\partial b_2}: \quad \cancel{2} b_2 \sum X_i^2 - \cancel{2}\sum X_i Y_i + \cancel{2} b_1 \sum X_i = 0$$

$$b_2 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - b_2 \bar{X}) n\bar{X} = 0$$

$$\boxed{b_2 \sum X_i^2} \boxed{- \sum X_i Y_i + n\bar{X}\bar{Y}} \boxed{- n b_2 \bar{X}^2} = 0$$

$$b_2 \left(\sum X_i^2 - n\bar{X}^2\right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$b_2 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

normal equations

also $\beta_2 = \frac{S_{XY}}{S_X^2}$

$$\boxed{\beta_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Exercise 4: Use the normal equations we just found and re-solve for the points from exercise 2.

$\bar{X} = \frac{1}{4}(1+2+3+4) = \frac{1}{4}(10) = \boxed{2.5}$ $\quad \bar{Y} = \frac{1}{4}(4+3+5+8) = \frac{20}{4} = \boxed{5}$

$\Sigma(x_i - \bar{x})(Y_i - \bar{Y}) = (1-2.5)(4-5) + (2-2.5)(3-5) + (3-2.5)(5-5) + (4-2.5)(8-5)$

$= (-1.5)(-1) + (-0.5)(-2) + (0.5)(0) + (1.5)(3) = \boxed{7}$

$\Sigma(x_i - \bar{x})^2 = (1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2$

$= -1.5^2 + (-0.5)^2 + 0.5^2 + 1.5^2 = \boxed{5}$

$\boxed{\hat{\beta}_2 = \frac{7}{5} = 1.4}$

$\boxed{\hat{\beta}_1 = 5 - 1.4(2.5) = 1.5}$

Exercise 5: Solve for the regression coefficients for the following points.

X	Y
1	7
2	8
3	5
4	4

$\bar{x} = \frac{1}{4}(1+2+3+4) = \frac{1}{4}(10) = \boxed{2.5}$

$\bar{Y} = \frac{1}{4}(7+8+5+4) = \frac{1}{4}(24) = \boxed{6}$

$\Sigma(x_i - \bar{x})(Y_i - \bar{Y}) = (1-2.5)(7-6) + (2-2.5)(8-6) + (3-2.5)(5-6)$
$+ (4-2.5)(4-6)$

$= (-1.5)(1) + (-0.5)(2) + (0.5)(-1) + (1.5)(-2)$

$= -1.5 - 1 - 0.5 - 3 = \boxed{-6}$

$\Sigma(x_i - \bar{x})^2 = \boxed{5}$

$\boxed{\hat{\beta}_2 = \frac{-6}{5} = -1.2}$

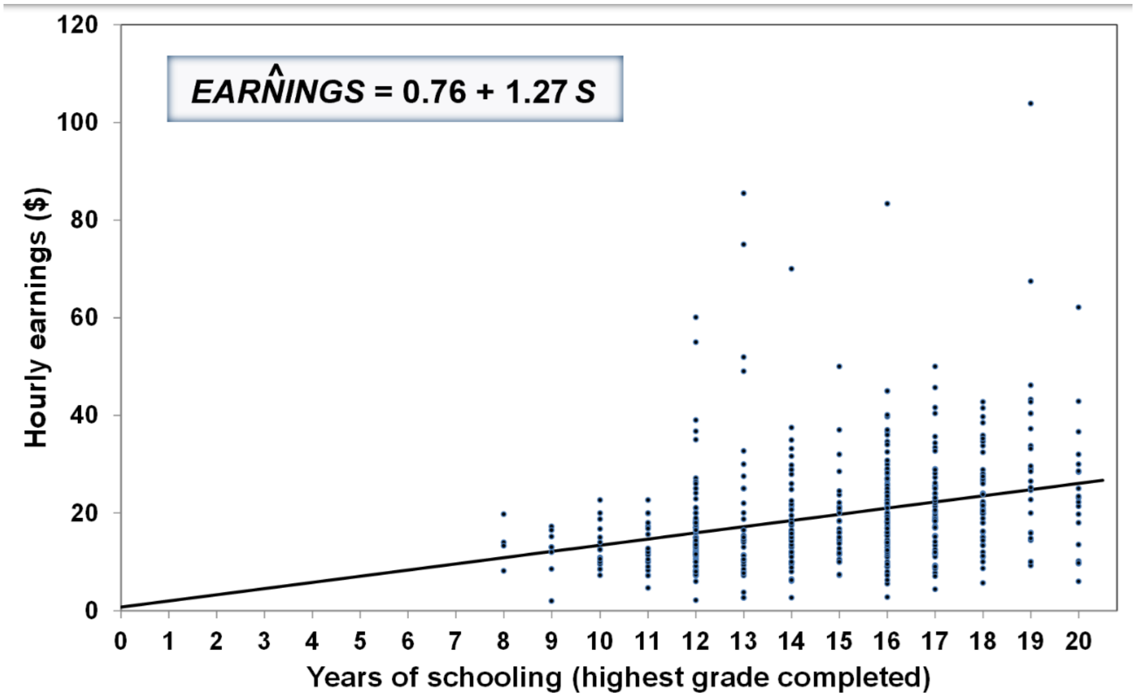
$\hat{\beta}_1 = 6 - (-1.2)(2.5)$

$\hat{\beta}_1 = 6 - (-3)$

$\boxed{\hat{\beta}_1 = 9}$

1.4 Interpretation of Regression Coefficients

Suppose we have data on schooling and earnings. We run the regression $Earnings_i = \beta_1 + \beta_2 School_i + u_i$. The output we got is shown below.



Exercise 6: Interpret the β_2 coefficient.

1 additional Year of school will increase hourly earnings by $1.27 on average.

Exercise 7: Interpret the β_1 coefficient. Does this make sense?

A person with 0 years of school will earn $0.76 per hour.

No, because this is below minimum wage. Since the data only goes to 8 years of school, when we go to far away (to zero) the meaning of the coefficients start to lose meaning.

Exercise 8: How much would you expect someone to earn if they went to school for 10 years?

$0.76 + 1.27(10)$

$0.76 + 12.7$

$\boxed{13.46}$

Exercise 9: How much would you expect someone to earn if they went to school for 15 years?

$0.76 + 1.27(15)$

$0.76 + 19.05$

$\boxed{\$19.81}$

Exercise 10: How much would you expect someone's earnings to change by if they went back to school for 2 years?

$\Delta E = \Delta\beta_1 + \beta_2 \Delta S$

$\Delta E = (0.76 - 0.76) + 1.27(2)$

$\Delta E = 2.5\cancel{4}4$

They should expect an increase of \$2.54

Exercise 11: Suppose you ran a regression of SAT scores on college GPA and found the fitted model:

$$GPA_i = 2.0 + 0.01SAT_i$$

Interpret the β_2 coefficient.

Earning 1 additional point on the SAT increases college GPA by 0.01