CHAPTER 7: REGRESSION

STATISTICS (FERM)

Lecturer: Nguyen Minh Quan, PhD
quannm@hcmiu.edu.vn

CONTENTS

Introduction to Regression

- Want to study relationship between quantities.

- In many situations, there is a single response variable Y, also called the dependent variable, which depends on the value of a set of input, also called independent, variables $x_1, ..., x_r$.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

- In practice, it subjects to random error e:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$$

Or, can re-write

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$$

This equation is called a linear regression equation.

Introduction to Regression

- The coefficients β_j are called the regression coefficients.
- A regression equation containing a single independent variable - that is, one in which $r = 1$ - is called a simple regression equation:

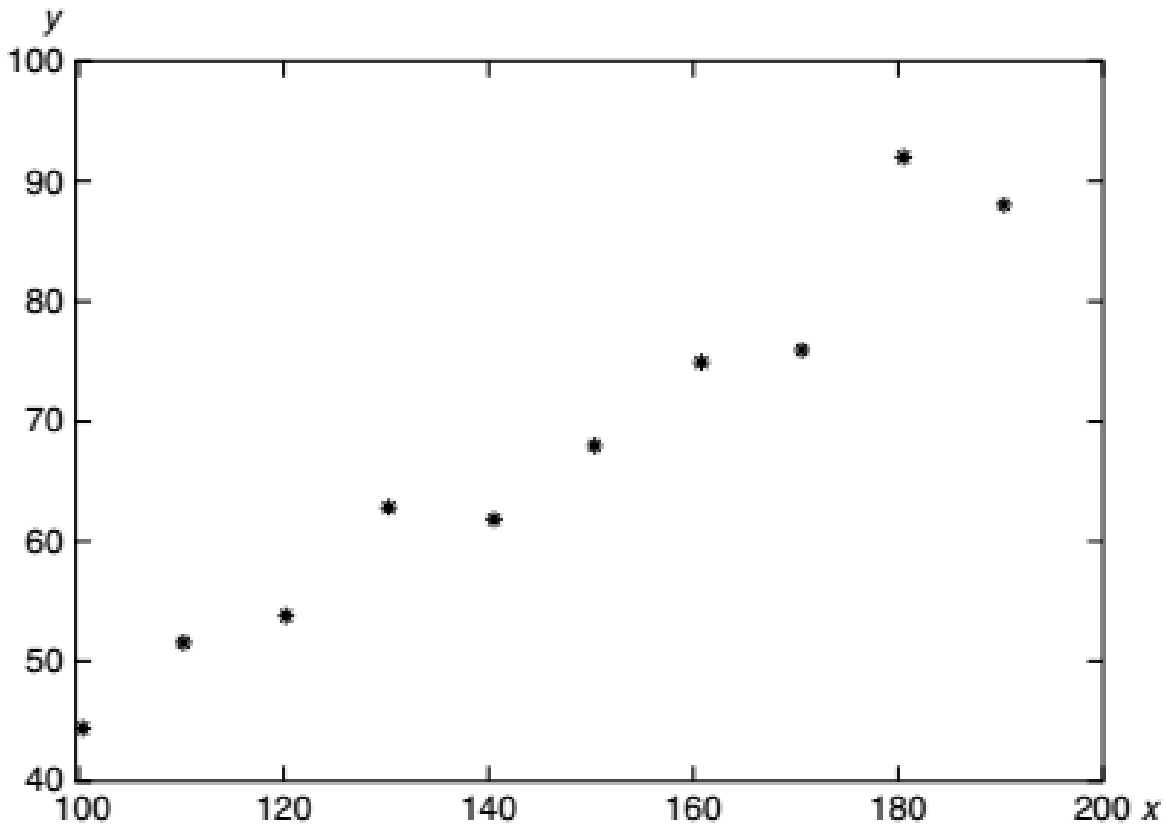
$$Y = \alpha + \beta x + e,$$

where e, representing the random error, is a random variable having mean 0.

- A regression equation contains many independent variables is called a multiple regression equation.

Introduction to Regression: Scatter diagram

As an example, considering 10 data pairs (x_i, y_i), $i = 1, ..., 10$. A plot of y_i versus x_i is called a scatter diagram.



Scatter plot.

Least Squares Estimators of the Regression Parameters

- Suppose that the responses Y_i corresponding to the input values x_i, $i = 1, \cdots, n$ are to be observed and used to estimate α and β in a simple linear regression model.
- If A and B are the estimators of α and β, then the sum of the squared differences between the estimated responses and the actual response values — call it SS— is given by

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- The method of least squares chooses as estimators of α and β the values of A and B that minimize SS.

Least Squares Estimators of the Regression Parameters

- We set

$$\frac{\partial SS}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0$$

$$\frac{\partial SS}{\partial B} = -2\sum_{i=1}^{n}x_i(Y_i - A - Bx_i) = 0$$

- This leads to the normal equations:

$$\sum_{i=1}^{n}Y_i = nA + B\sum_{i=1}^{n}x_i$$

$$\sum_{i=1}^{n}x_iY_i = A\sum_{i=1}^{n}x_i + B\sum_{i=1}^{n}x_i^2$$

Least Squares Estimators of the Regression Parameters

We let

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Theorem

The least squares estimators of β and α corresponding to the data set $x_i, Y_i,\ i = 1, ..., n$ are, respectively,

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$A = \overline{Y} - B\bar{x}$$

The straight line $A + Bx$ is called the estimated regression line.

Least Squares Estimators of the Regression Parameters

Example

A large midwestern bank is planning on introducing a new word processing system to its secretarial staff. To learn about the amount of training that is needed to effectively implement the new system, the bank chose eight employees of roughly equal skill. These workers were trained for different amounts of time and were then individually put to work on a given project. The following data indicate the training times and the resulting times (both in hours) that it took each worker to complete the project.

Worker	Training time (= x)	Time to complete project (= Y)
1	22	18.4
2	18	19.2
3	30	14.5
4	16	19.0
5	25	16.6
6	20	17.7
7	10	24.4
8	14	21.0

Least Squares Estimators of the Regression Parameters

Example

(a) What is the estimated regression line?
(b) Predict the amount of time it would take a worker who receives 28 hours of training to complete the project.
(c) Predict the amount of time it would take a worker who receives 50 hours of training to complete the project.

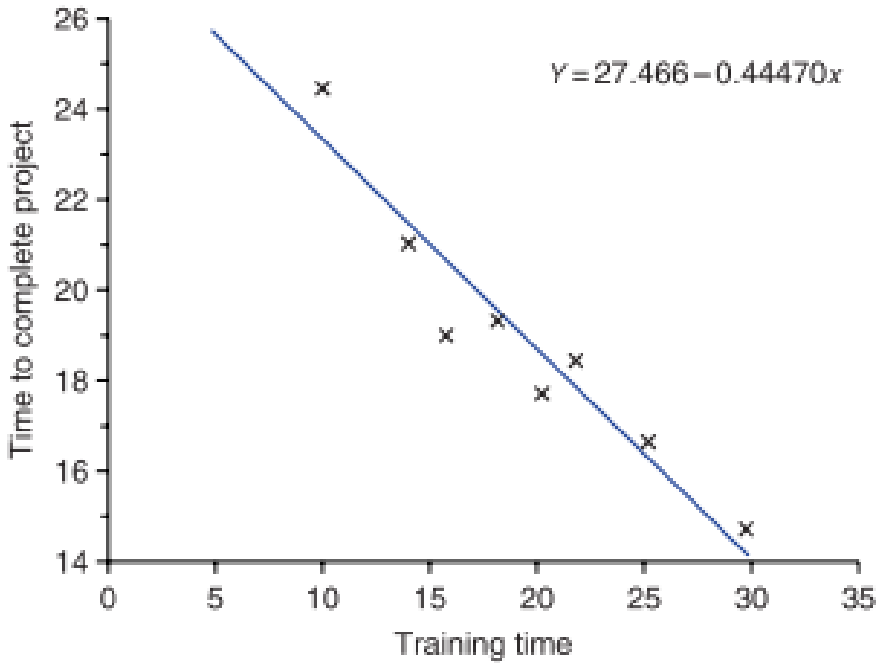
Solution

(a) We compute the least-squares estimators by the given formula in slide #8:

$$A = 27.46606, B = -0.4447002.$$

The estimated regression line is as follows: $Y = 27.46606 - 0.4447002x$.

Least Squares Estimators of the Regression Parameters



(b) $27.466 - 28(0.445) = 15.006$.

(c) Data is not available to answer this question.

Least Squares Estimators of the Regression Parameters

Example

This Table displays data on age and price for a sample of cars of a particular make and model. We refer to the car as the Orion, but the data, obtained from the Asian Import edition of the Auto Trader magazine, is for a real car. Ages are in years; prices are in hundreds of dollars, rounded to the nearest hundred dollars.

Age and price data for a sample of 11 Orions

Car	Age (yr) x	Price ($100) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

Least Squares Estimators of the Regression Parameters

Example

a. Determine the regression equation for the data.
b. Graph the regression equation and the data points.
c. Use the regression equation to predict the price of a 3-year-old Orion.

Solution

(a)

Table for computing the regression equation for the Orion data

Age (yr) x	Price ($100) y	xy	x^2
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326

$$B = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = -20.26, A = \overline{Y} - B\bar{x} = 195.47$$

(c) $195.47 - 20.26 \times 3 = 134.69$

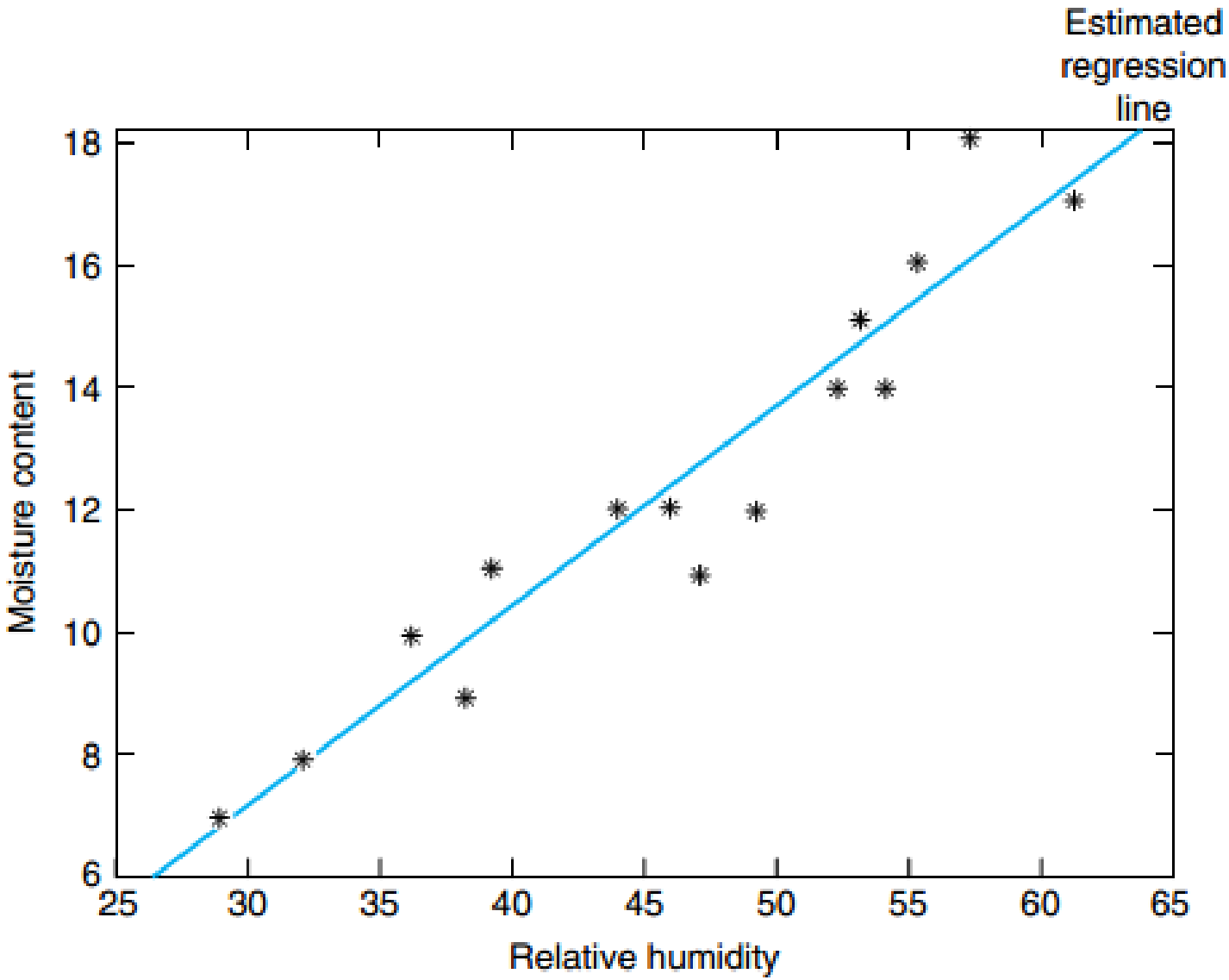
Least Squares Estimators of the Regression Parameters

Exercise

The raw material used in the production of a certain synthetic fiber is stored in a location without a humidity control. Measurements of the relative humidity in the storage location and the moisture content of a sample of the raw material were taken over 15 days with the following data (in percentages) resulting.

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Least Squares Estimators of the Regression Parameters



Least Squares Estimators of the Regression Parameters

Exercise

The following are the average scores on the mathematics part of the Scholastic Aptitude Test (SAT) for some of the years from 1994 to 2009.

Year	SAT Score
1994	504
1996	508
1998	512
2000	514
2002	516
2004	518
2005	520
2007	515
2009	515

Assuming a simple linear regression model, predict the average scores in 1997, 2006 and 2008.

Distribution of the Estimators

Suppose that if Y_i is the response corresponding to the input value x_i, then $Y_i, ..., Y_n$ are independent and

$$Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$$

Since $B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$, thus

$$E[B] = \frac{\sum_{i=1}^{n}(x_i - x)E[Y_i]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - x)(\alpha + \beta x_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \beta$$

So B is an unbiased estimator of β.

Distribution of the Estimators

$$Var\left(B\right)=\frac{Var\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)Y_i\right)}{\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)^2}=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$$

We have

$$A=\sum_{i=1}^{n}\frac{Y_i}{n}-B\bar{x}$$

Therefore,

$$E\left(A\right)=\sum_{i=1}^{n}\frac{E\left[Y_i\right]}{n}-\bar{x}E\left[B\right]=\sum_{i=1}^{n}\frac{\left(\alpha+\beta x_i\right)}{n}-\bar{x}\beta=\alpha$$

$$Var\left(A\right)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)}\text{ (exercise)}$$

Distribution of the Estimators

The quantities $Y_i - A - Bx_i$, $i = 1, \cdots, n$, which represent the differences between the actual responses (that is, the Y_i) and their least squares estimators (that is, $A + Bx_i$) are called the residuals. Let

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

It can be shown that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}, E\left[\frac{SS_R}{\sigma^2}\right] = n - 2$$

or,

$$E\left[\frac{SS_R}{n-2}\right] = \sigma^2$$

Thus $SS_R/(n-2)$ is an unbiased estimator of σ^2.

Notation

If we let

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

$$S_{YY} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

then the least squares estimators can be expressed as

$$B = \frac{S_{xY}}{S_{xx}}, A = \overline{Y} - B\overline{x}$$

Notation

Computational Identity for SS_R

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Distribution of the Estimators

Sums up the results

Suppose that the responses $Y_i, i = 1, ..., n$ are independent normal random variables with means $\alpha + \beta x_i$ and common variance σ^2. The least squares estimators of β and α:

$$B = \frac{S_{xY}}{S_{xx}}, A = \overline{Y} - B\overline{x}$$

are distributed as follows:

$$A \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right), B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

In addition,

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}; \frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

Least Squares Estimators of the Regression Parameters

Example

We re-consider example in slide #9 and suppose that we are interested in estimating the value of σ^2.

Worker	Training time (= x)	Time to complete project (= Y)
1	22	18.4
2	18	19.2
3	30	14.5
4	16	19.0
5	25	16.6
6	20	17.7
7	10	24.4
8	14	21.0

Solution

$$S_{xx} = 281.875, S_{YY} = 61.0806, S_{xY} = -125.35$$

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} = 5.3375$$

The estimate of σ^2 is $\frac{SS_R}{n-2} = \frac{5.3375}{6} = 0.8896$.

Least Squares Estimators of the Regression Parameters

Exercise

The following data relate the speed of a particular typist and the temperature setting of his office. The units are words per minute and degrees Fahrenheit.

Temperature	Typing speed
50	63
60	74
70	79

(a) Compute, by hand, the value of SS_R.
(b) Estimating the value of σ^2.

Inferences Concerning β

An important hypothesis to consider regarding the simple linear regression model $Y = \alpha + \beta x + e$ is the hypothesis that $\beta = 0$.

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0$$

Fact: $\sqrt{\frac{SS_R}{(n-2)S_{xx}}} B \sim t_{n-2}$.

- Reject H_0 if $\sqrt{\frac{(n-2)S_{xx}}{SS_R}} |B| > t_{\gamma/2,n-2}$, where γ is a significance level.
- Accept H_0 otherwise.

Inferences Concerning β

Remark: This test can be performed by first computing the value of the test statistic

$$TS = \sqrt{\frac{(n-2)\,S_{xx}}{SS_R}}\,|B| > t_{\gamma/2,n-2}$$

and then rejecting H_0 if the desired significance level is at least as large as

$$p - Value = P\left(|T_{n-2}| > TS\right) = 2P\left(T_{n-2} > TS\right)$$

Inferences Concerning β

Example

An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 70 miles per hour. The miles per gallon attained at each of these speeds was determined, with the following data resulting:

Speed	Miles per Gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

Do these data refute the claim that the mileage per gallon of gas is unaffected by the speed at which the car is being driven?

Inferences Concerning β

Suppose that a simple linear regression model $Y = \alpha + \beta x + e.$ relates Y, the miles per gallon of the car, to x, the speed at which it is being driven. Now, the claim being made is that $\beta = 0$.

We test:

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0$$

We compute:

$$S_{xx} = 700, S_{YY} = 21.757, S_{xY} = -119$$

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} = 1.527; B = \frac{S_{xY}}{S_{xx}} = \frac{-119}{700} = -0.17$$

$$TS = \sqrt{5\,(700)\,/1.527}\,(0.17) = 8.139 > t_{0.005,5} = 4.032$$

$\rightarrow$ Reject H_0 at the 1 percent level of significance. Thus, the claim that the mileage does not depend on the speed at which the car is driven is rejected.

Inferences Concerning β

Exercise

Test the hypothesis that $\beta = 0$ for the following data.

x	Y
3	7
8	8
10	6
13	7

Use the 5 percent level of significance.

Hint: Do not reject $H_0 : \beta = 0$.

Confidence Interval for β

Confidence Interval for β

A $100(1-a)$ percent confidence interval estimator of β is

$$\left(B - \sqrt{\frac{SS_R}{(n-2)\,S_{xx}}}\, t_{a/2,n-2},\ B + \sqrt{\frac{SS_R}{(n-2)\,S_{xx}}}\, t_{a/2,n-2} \right)$$

Example

Derive a 95 percent confidence interval estimate of β in the example fuel consumption in slide 24.

Solution

$$t_{0.25,5} = 2.571$$

$$-0.170 \pm 2.571\sqrt{\frac{1.527}{3500}} = -0.170 \pm 0.54$$

That is, we can be 95 percent confident that β lies between -0.224 and -0.116.

Selected exercises

Chapter 9: 1,3,5,6,10,11,12,20.

–END OF CHAPTER 7. THANK YOU!–