

Case-based Social Statistics II

CSSS 322

Spring 2002

Solutions to Practice Second Examination

Exam: Monday, June 10, 10:30am - 12:00pm

Professor: Mark S. Handcock

Name: Ronald Aylmer Fisher

1. Please write your name in the above space.
2. All questions are of equal value (but not necessarily of equal difficulty). Indicate on the table below the **5** problems that you attempted.
3. Do not turn the page until so instructed. (You will have 150 minutes to work after the examination has been discussed with you.)
4. You may use your crib sheet. and the statistical tables provided. Otherwise this is a closed book examination.
5. If you do not have enough room for your work in the place provided, use the back of a nearby page. (However, be sure to mark clearly which problem the material on the back of any page refers to.) If you pull the pages apart, sign all pages.
6. Answers should unambiguously state, in words, the approach taken. You should show your work so that partial credit can be given. Poorly described solutions will be penalized.
7. Good luck!

Question	Subject	Points available	Points earned
1	Zero Coupon Bonds	25	
2	Regression Assumptions	25	
3	Cherry trees	25	
4	Supply and Demand	25	
Total		100	

Question 1) Zero Coupon Bonds (25 points)

A zero-coupon bond pays a specified amount at maturity and nothing before maturity. The price paid for a bond depends on the time to maturity. Financial theory teaches that the volatility of a bond's price is proportional to its maturity. The volatility of a bond is defined to be the standard deviation in price for that bond. Below are the volatilities in price for bonds of four different maturities, based on U. S. Treasury bonds from the middle 1980's.

Maturity Time (in years)	Volatility in Price (in percent)
5	11
10	21
15	29
20	38

a) (10 points)

Calculate, by hand, the regression line of volatility against maturity. Be sure to show your work.

Solution: Let X be the maturity time and Y be the volatility in price. Then

$$\bar{X} = 12.5 \text{ years}, \quad \bar{Y} = 24.75 \text{ percent}$$

$$\sum_{i=1}^4 x_i y_i = 1460, \quad \sum_{i=1}^4 x_i^2 = 750$$

So

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^4 x_i y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^4 x_i^2 - n \bar{X}^2} \\ &= \frac{1460 - 4 \times 12.5 \times 24.75}{750 - 4 \times 12.5^2} \\ &= 1.78 \text{ percent per year} \end{aligned}$$

The intercept (or "constant") is,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 24.75 - 1.78 \times 12.5 = 2.5 \text{ percent}$$

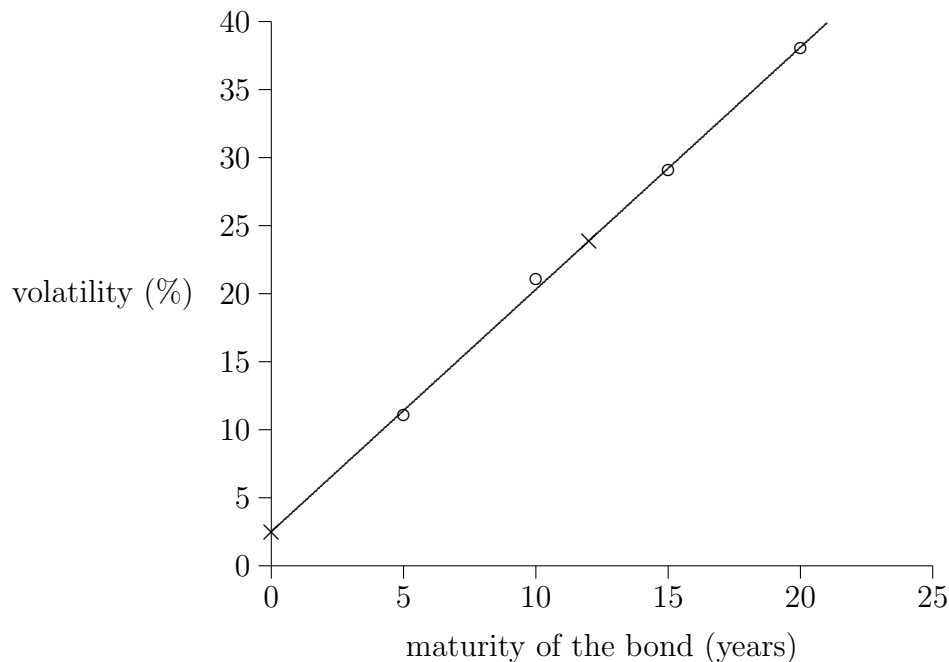
That is,

$$\hat{\beta}_0 = 2.5 \qquad \hat{\beta}_1 = 1.78$$

b) (4 points)

Graph the four points and the regression line on the axes provided on the top of the next page. Does the line appear to fit the data reasonably well?

Solution: The line and points are on the plot. Apparently the line fits the data extremely well.



c) (4 points)

Use the regression line to predict:

- i) The volatility in price of bonds with a maturity of 12 years.

Solution: The fitted value at $x = 12$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2.5 + 1.78 \times 12 = 23.86 \text{ percent}$$

- ii) The volatility in price of bonds that mature immediately. That is, with a maturity of 0 years.

Solution: The fitted value at $x = 0$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2.5 + 1.78 \times 0 = 2.5 \text{ percent}$$

Show these two points on the graph.

Solution: These points are shown on the graph using a “×”.

d) (2 points)

What is the increase in volatility in price for each year increase in maturity?

Solution: This is just the slope of the regression line, $\hat{\beta}_1$, that is 1.78 percent per year.

e) (5 points)

What is the correlation coefficient between volatility and maturity? Does it indicate that the model is an adequate fit?

Solution: We need to calculate this directly. The correlation is

$$r = \hat{\beta}_1 \cdot \frac{S_X}{S_Y}$$

where S_X , S_Y are the sample standard deviations of X and Y , respectively.

$$\begin{aligned} S_X^2 &= \frac{1}{n-1} \cdot \left(\sum_{i=1}^4 x_i^2 - n\bar{X}^2 \right) \\ &= \frac{1}{4-1} \cdot \left(750 - 4 \times 12.5^2 \right) \\ &= 41.67 \end{aligned}$$

Similarly,

$$\begin{aligned} S_Y^2 &= \frac{1}{n-1} \cdot \left(\sum_{i=1}^4 y_i^2 - n\bar{Y}^2 \right) \\ &= \frac{1}{4-1} \cdot \left(1847 - 4 \times 24.75^2 \right) \\ &= 132.25 \end{aligned}$$

The correlation is then

$$r = \hat{\beta}_1 \cdot \frac{S_X}{S_Y} = 1.78 \cdot \frac{6.455}{11.5} = 0.999$$

As this is very close to 1, the model fits extremely well.

Question 2) Regression Assumptions (25 points)

For each of the following descriptions of a residual plot, write one sentence (only) about what the plot reveals about regression assumptions. Be specific about which assumptions are implicated.

a) (4 points)

A plot of the standardized residuals versus the normal scores (i.e. a normal probability plot) is basically straight.

Solution: If the line is straight the standardized residuals are standard normal. Thus the plot provides no evidence against any of the assumptions.

b) (4 points)

A plot of the standardized residuals versus the independent variable resembles a sideways cone.

Solution: This means that the variance in the residuals increases with increasing values of the independent variable. Thus the assumption of constant variance is invalid.

c) (4 points)

A plot of the standardized residuals versus the independent variable has a U shape.

Solution: A pattern in the standardized residuals indicates what is missing from the model. An approximately U shaped pattern indicates that a quadratic pattern is missing. Thus the assumption of a linear expectation in Y with X is invalid.

Question 3) Cherry trees (25 points)

People in forestry need to estimate the amount of timber in a given area of forest. Therefore they need a quick and easy way to determine the volume of any given tree. Of course, it is difficult to measure the volume of a tree directly, but it is easy to measure the diameter. Thus the foresters would like to have an equation whereby the volume of a tree could be estimated from its diameter.

The data studied here pertain to a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania. For each tree, the foresters measured the diameter in inches at 4.5 feet above ground level, its height, and (after cutting up the tree) the volume in cubic feet.

A regression analysis of the logarithm of the volume against the logarithm of diameter for these trees has been performed using DataTools, the output from which is at the end of the question.

a) (6 points)

Write down the fitted regression line of the logarithm of volume on the logarithm of diameter? What are the slope and constant coefficients? What are their standard errors?

$$\text{fitted line : } \hat{y} = \text{LNVOL} = -2.353 + 2.200 \times \text{LNDIAM}$$

$$\text{slope} = 2.200 \quad \text{its standard error} = 0.090$$

$$\text{constant} = -2.353 \quad \text{its standard error} = 0.231$$

Solution: The values can be read off the output at the end of the question.

b) (3 points)

Draw the estimated regression line on the scatterplot given in the DataTools output.

Solution: See the graph at the end of the question. The two points used were the mean value: (2.557, 3.273) and the point at $x = 2$. The fitted value for the latter point is $\hat{y} = -2.353 + 2.200 \times 2 = 2.047$.

c) (3 points)

Does the model fit the data adequately? Why?

Solution: The correlation coefficient is $r = 0.977$, indicating a strong positive linear relationship between logarithm-volume and logarithm-diameter. The F-ratio is 599.717, with an associated p -value of 0.000, indicating that the line provides an adequate fit to the data.

d) (7 points)

One might conjecture that the volume of a tree would be proportional to the cube of its diameter: thus

$$\text{volume} = \text{constant} \times \text{diameter}^3$$

or, equivalently,

$$\ln(\text{volume}) = \ln(\text{constant}) + 3 \times \ln(\text{diameter}).$$

Is this conjecture supported by the data at hand? Use a 95 % confidence interval to support your conclusion.

Solution: The hypothesis test is for

$$H_0 : \beta_1 = 3$$

$$H_1 : \beta_1 \neq 3$$

A 95% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 \pm t_{p/2}(n-2) \times \widehat{\text{SE}}(\hat{\beta}_1) \\ = 2.200 \pm t_{0.025}(29) \times 0.090 \\ = (2.016, 2.384) \end{aligned}$$

where we have used the t-value $t_{0.025}(29) = 2.045$.

The null hypothesis is $\beta_1 = 3$. As 3 does not fall in the interval this hypothesis is rejected. That is, the conjecture is not supported by the data at hand.

e) (3 points)

Suppose the foresters measure a new tree (of the same species and in the same forest) and find the logarithm of its diameter to be 2.75. Predict the mean logarithm of the volume for trees of this diameter.

Solution: The predicted value at $x = 2.75$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2.353 + 2.200 \times 2.75 = 3.697 \text{ ln inches}$$

f) (3 points)

Convert the prediction in e) for the logarithm of volume to one for the volume itself?

Solution: As this previous value is in the logarithm of inches, we need to exponentiate to get inches. The predicted value for volume is then

$$\exp(3.697) = 40.326 \text{ inches}^3$$

```

Commands:-----
Summary statistics
Descriptive statistics
LNDIAM LNVOL
Output:-----

```

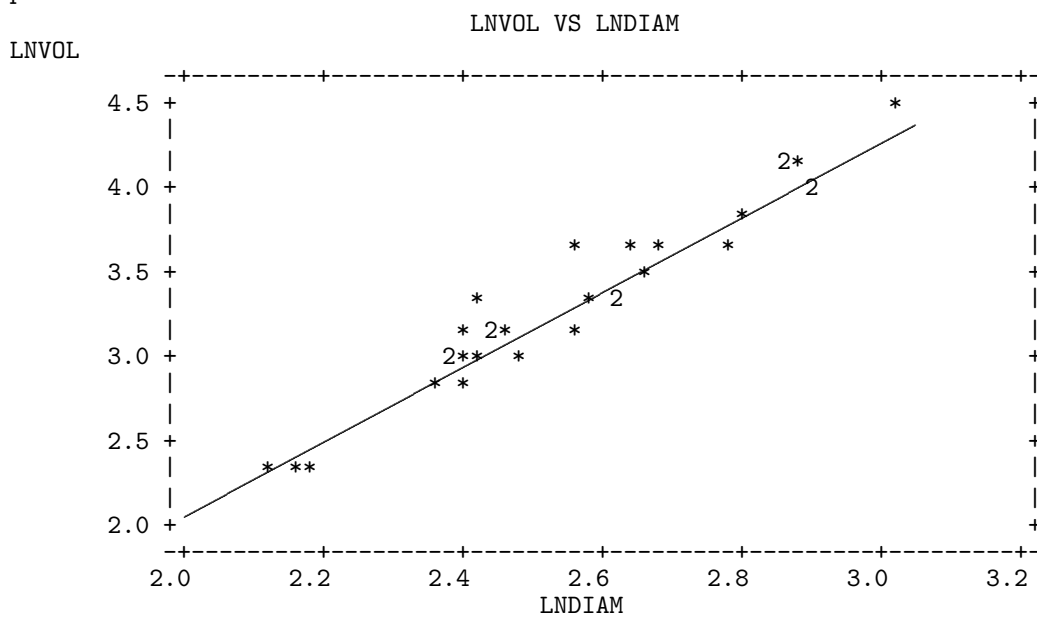
DESCRIPTIVE STATISTICS

	LNDIAM	LNVOL
CASES	31	31
LOWER 95.0% C.I.	2.475	3.088
MEAN	2.557	3.273
UPPER 95.0% C.I.	2.639	3.458
S.D.	0.234	0.526
S.E. (MEAN)	0.042	0.094
C.V.	9.151	16.07
MINIMUM	2.116	2.322
MEDIAN	2.570	3.333
MAXIMUM	3.025	4.344

```

Commands:-----
Stat
Summary statistics
Scatter plot
LNVOL LNDIAM
Output:-----

```



Commands:-----
 Linear models
 Multiple regression
 Dependent variable: LNVOL Independent variable: LNDIAM
 Output:-----

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF LNVOL

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	-2.353	0.231	-10.202	0.000
LNDIAM	2.200	0.090	24.489	0.000

R SQUARED 0.954 RESID. MEAN SQUARE (MSE) 0.013
 ADJUSTED R SQUARED 0.952 STANDARD DEVIATION 0.115

SOURCE	DF	SS	MS	F	P
REGRESSION	1	7.925	7.925	599.717	0.000
RESIDUAL	29	0.383	0.013		
TOTAL	30	8.308			

CASES INCLUDED 31 MISSING CASES 0

Question 4) Supply and Demand (25 points)

Many economists believe that increases in the supply of a product will lead to decreases in the price. To test this theory for the digital telephone market, data was collected for each state for 1992:

SUPPLY = number of digital phones in retail stores (in thousands)
PRICE = average sale price per phone (in dollars)

The DataTools output is provided below:

```
Commands:-----
Linear models
Linear regression
Dependent variable: PRICE Independent variable: SUPPLY
Output:-----
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRICE

PREDICTOR
VARIABLES      COEFFICIENT      STD ERROR      STUDENT'S T      P

CONSTANT      199.562          0.22251        896.87           0.0000
SUPPLY        0.01678          0.01228         1.37             0.1780

R-SQUARED      0.0375          RESID. MEAN SQUARE (MSE)  0.43789
ADJUSTED R-SQUARED 0.0174          STANDARD DEVIATION          0.66173

SOURCE      DF      SS      MS      F      P

REGRESSION    1      0.81833  0.81833  1.87  0.1780
RESIDUAL     48      21.0186  0.43789
TOTAL        49      21.8369

CASES INCLUDED 50  MISSING CASES 0
```

a) (5 points)

By interpreting the coefficient of supply, describe the relationship between supply and price.

Solution: From the interpretation of β , when the interest rate increases by 1 percent the inflation rate decreases by $\beta\%$. As $\hat{\beta} = -1.161$ we see that the inflation rate decreases by about 1.161%. The interpretation of the intercept is the inflation rate when the interest rate is zero. That is the inflation rate under full employment is 8.881%.

b) (4 points)

How much evidence is there in this data for a relationship between supply and price? Explain your answer with reference to a test.

Solution: A 95% confidence interval for β_2 the coefficient of INTEREST is

$$\begin{aligned} & \hat{\beta}_1 \pm t_{p/2}(n-2) \times \widehat{SE}(\hat{\beta}_1) \\ & = -1.161 \pm t_{0.025}(8) \times 0.255 \\ & = (-1.749, -0.573) \end{aligned}$$

where we have used the t-value $t_{0.025}(8) = 2.306$.

c) (6 points)

A student collects additional information on the population of each state in 1992:

POPULATION = population of the state in 1992 (in millions)

After adding this variable to the model the DataTools output is:

```

Commands:-----
Stat
  Simple Linear regression
  Dependent variable: PRICE Independent variable: SUPPLY POPULATION
Output:-----
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRICE

PREDICTOR
VARIABLES      COEFFICIENT      STD ERROR      STUDENT'S T      P      VIF

CONSTANT          199.899          0.09963          2006.42          0.0000
SUPPLY            -2.12252          0.14901          -14.24           0.0000      777.7
POPULATION         2.12773          0.14811           14.37           0.0000      777.7

R-SQUARED          0.8215          RESID. MEAN SQUARE (MSE)      0.08295
ADJUSTED R-SQUARED 0.8139          STANDARD DEVIATION              0.28801

SOURCE      DF      SS      MS      F      P

REGRESSION    2      17.9383      8.96915      108.13      0.0000
RESIDUAL      47      3.89864      0.08295
TOTAL         49      21.8369

CASES INCLUDED 50  MISSING CASES 0
  
```

By interpreting the coefficients of supply and population, describe what the model indicates about the relationships between the independent variables and price.

Solution: From the interpretation of β , when the supply increases by one thousand units (holding the population fixed) the price increases by -2.12 dollars. That is the price decreases by \$2.12. When the population increases by one million people (holding the supply fixed) the price increases by 2.13 dollars.

d) (3 points)

Is this model a good fit to the data? Explain your answer with reference to a test.

Solution: Their model does provide a good fit to the data. The multiple correlation coefficient is $r = \sqrt{0.8215} = 0.9064$, indicating a strong positive linear relationship between PRICE and the two variables. The F-ratio is 108.13, with an associated p -value of 0.000, indicating that the line provides an adequate fit to the data. Both variables are highly significant and the R^2 of the model is substantially larger than that with only SUPPLY in the model.

e) (7 points)

Is the economic theory correct? Explain what this data indicates about the theory by rationalizing the descriptions of the two models.

Solution: The economists theory is supported by the information, but in a very specific way. The model in a) suggests that there is only a weak relationship between price and supply - and even suggests that price increases with supply, which is very counter-intuitive. The second model suggests why. If the population level of the country is adjusted for then there is a clear negative effect on price of increasing supply. In addition population increases also increase price. This is inline with the natural interpretation of the economists claim: for countries with the same population, higher supply is associated with lower price. Note also that higher population is associated with higher supply - hence the spurious positive association between price and supply.

Question 5) Insurance (25 points)

It is of interest to determine if the responsiveness of an insurance company to new innovations is related to the size of the firm. A STERN economist collected data on 20 insurance firms, 10 of which were mutual firms and 10 of which were stock firms. To quantify the concept of responsiveness, she singled out a particular insurance innovation and measured the time between the announcement of the innovation and adoption by the firm. The measured variables are:

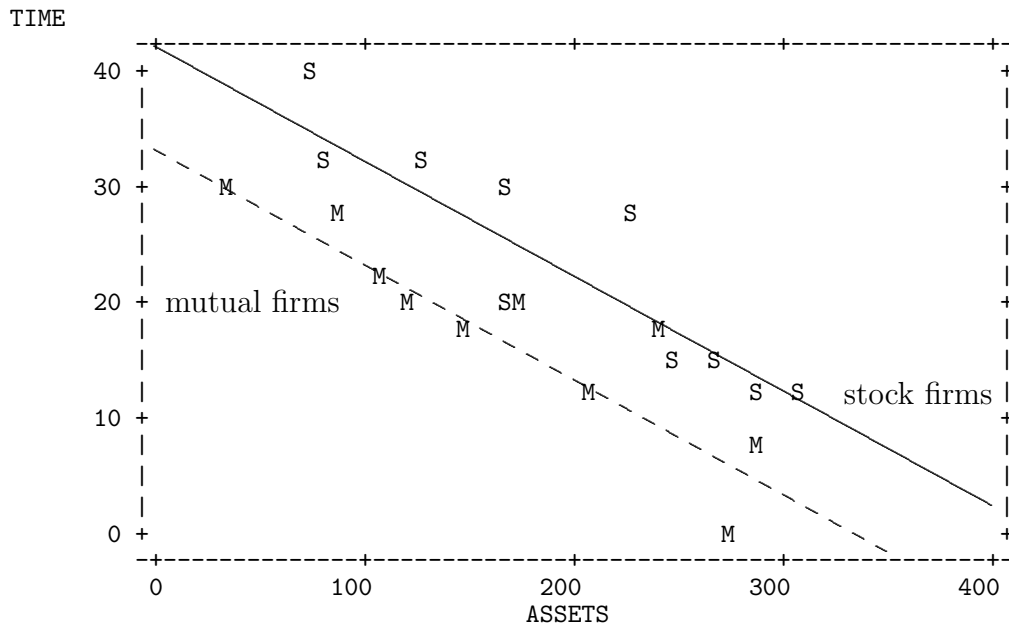
$$\begin{aligned} \text{TIME} &= \text{time (in months) to adopt the innovation} \\ \text{ASSETS} &= \text{total assets of the firm (in millions of dollars)} \\ \text{TYPE} &= \begin{cases} 1 & \text{if the firm is a mutual firm,} \\ 0 & \text{if the firm is a stock firm.} \end{cases} \end{aligned}$$

An analysis in DataTools is below:

```
Commands:-----
Stat
  Summary statistics
  Descriptive statistics
  X-Y pair: ASSETS TIME
```

Output:-----
DESCRIPTIVE STATISTICS

	TIME	ASSETS	TYPE
CASES	20	20	20
MEAN	19.700	180.800	0.500
S.D.	9.532	83.670	0.513
MINIMUM	0.000	31.000	0.000
MAXIMUM	38.000	305.000	1.000



```

Commands:-----
Linear models
Linear regression
Dependent variable: TIME Independent variable: ASSETS TYPE
Output:-----

```

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF LNVOL

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	42.251	2.312	18.274	0.000
ASSETS	-0.100	0.010	-9.704	0.000
TYPE	-9.033	1.677	-5.388	0.000

R SQUARED 0.865 RESID. MEAN SQUARE (MSE) 13.687
ADJUSTED R SQUARED 0.849 STANDARD DEVIATION 3.700

SOURCE	DF	SS	MS	F	P
REGRESSION	2	1493.524	746.762	54.561	0.000
RESIDUAL	17	232.676	13.687		
TOTAL	19	1726.300			

CASES INCLUDED 20 MISSING CASES 0

a) (3 points)

Write down the fitted multiple regression model for TIME on ASSETS and TYPE

Solution: From the DataTools output

$$\hat{y} = \text{TIME} = 42.251 - 0.100 \times \text{ASSETS} - 9.033 \times \text{TYPE}$$

b) (6 points)

Give an interpretation of the coefficients of ASSETS and TYPE

Solution: For every million dollar increase in assets the time to adoption decreases by 0.1 months.

The difference between mutual firms and stock firms in time to adoption is 9.033 months, in favor of the mutual firms.

c) (4 points)

Graphically represent the fitted multiple regression model by drawing the two fitted lines on the scatterplot of TIME against ASSETS given in the DataTools output. Clearly label the lines as “mutual” or “stock”, as appropriate.

Solution: The dash line corresponds to the mutual firms and the solid line to the stock firms.

d) (4 points)

Is ASSETS an important explanatory variable in the multiple regression? Justify your answer.

Solution: The hypotheses are

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Using the “P”-value < 0.05 rule ASSETS is an important variable as it has $p = 0.000$.

e) (3 points)

Is this a useful multiple regression model for TIME? In answering this question test the overall fit of the model.

Solution: The correlation coefficient is $r = 0.930$, indicating a strong positive linear relationship between TIME and the two variables. The F-ratio is 54.561, with an associated p -value of 0.000, indicating that the line provides an adequate fit to the data.

f) (5 points)

Construct a 95% confidence interval for the coefficient of the type of firm.

Solution: A 95% confidence interval for β_2 , the coefficient of TYPE, is

$$\begin{aligned} & \hat{\beta}_2 \pm t_{p/2}(n - k - 1) \times \widehat{SE}(\hat{\beta}_2) \\ & = -9.033 \pm t_{0.025}(17) \times 1.677 \\ & = (-12.571, -5.495) \end{aligned}$$

where we have used the t-value $t_{0.025}(17) = 2.110$.