**STAT 43000/STAT 53001 Applied Statistics Spring 2023**

**Test 1- Part I (Solution) Name:**

**Please answer the following questions**.

**1.Short answer questions (10 points)**

1. Below you are given a summary of the output from a simple linear regression analysis for a sample of size 15:   
    SS (total) = 165  
    SS(regression) =125  
   Calculate the coefficient of determination .

*Solution: We have*

1. In a simple linear regression model, the slope coefficient measures.
2. the elasticity of Y with respect to X
3. the change in Y which the model predicts for a unit change in X
4. the change in X which the model predicts for a unit change in Y
5. the value of Y for any given value of X
6. A fitted regression equation for a dataset . What is the value of the residual for the observation (*x,y*)=(100, 90)?
7. 95 b) 5 c) 10 d) -5 e) -10

*Solution:*

1. If the correlation coefficient between two variables is 0.9, the coefficient of determination is
2. 0.9 b) 0.81 c) -0.81 d) 0.95
3. If all else is held constant, compared to the prediction interval for a particular value y, the confidence interval for a mean value of y will be
4. Narrower b) Wider c) Equal width d) No relation
5. The following table was generated from the sample data of 10 newborn babies regarding the weight of the mother at birth, the weight of the father at birth, and the weight of the baby at birth.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard*  *Error* | *t Stat* | *P-value* |
| **Intercept** |  |  |  |  |
| **Mother's Weight** |  |  |  |  |
| **Father's Weight** |  |  |  |  |

Write the equation of a multiple regression model based on the computer output above.

*Solution:*

1. The measure used to describe the case when the regressor variables in a multiple linear regression model are correlated is
2. Coefficient of determination
3. Multicollinearity
4. Correlation
5. Regression
6. The estimated parameters in multiple linear regression model with design matrix X in the matrix form is expressed as
8. Which of the following is not a quantity to measure the quality of the model?
9. R2
10. PRESS
11. AIC
12. First Difference
13. To test autocorrelation, we use
14. Durbin Watson test
15. Box-Cox test

c) Cook’s distance test

d) Hildreth-Lu test

**Q.N. 2)** Mark as TRUE or FALSE (**5 Points**)

a) and Adjusted are used to measure the quality of fitted regression model. (True)

b) For a given data set, the equation of the least square regression line will always pass through . (True)

c) Box-Cox transformation yields . This means no transformation is needed. (False)

d) The sampling distribution of is normal if the usual regression assumptions are satisfied. (True)

e) While performing the lack of fit test we decompose the error sum of squares (SSE). (True)

**Q.N. 3)** In order to study the relationship between age and muscle mass a set of data was collected from many patients of diﬀerent ages (in years) and attached is the computer output from muscle mass (y) and the age (x). Answer the following question based on the output. **(15 Points)**

1. How many patients were included in this study?

Answer: There are 60 observations in this dataset.

1. What is the equation of the estimated regression line?

Answer:

1. Obtain the point estimate of the muscle mass if the age of the patient is changed by one year.

Answer: Since the slope of the model is 1.19 the estimated muscle mass change is 1.19 .

1. Obtain a point estimate of the muscle mass for a patient of age 60 years.

Answer: The estimated muscle mass is

1. What is the 95% conﬁdence interval for the β1?

Answer: 95% CI is ( -1.370545, -1.009446)

1. Test the hypothesis for the significance of the age to determine the muscle mass.

Answer: Since p-values <2e-16 \*\*\*, we reject the null hypothesis and conclude that the muscle mass is a significance variable.

1. What is a point estimate of σ2 ?

Answer: Note that MSE is an unbiased estimator of . From the computer output MSE is 66.8. Therefore, point estimate of is 66.8.

1. What is the coefficient of determination?

Answer: The coefficient of determine is 0.7501.

> model=lm(y~x)

> model

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

156.35 -1.19

> summary(model)

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-16.1368 -6.1968 -0.5969 6.7607 23.4731

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 156.3466 5.5123 28.36 <2e-16 \*\*\*

x -1.1900 0.0902 -13.19 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.173 on 58 degrees of freedom

Multiple R-squared: 0.7501, Adjusted R-squared: 0.7458

F-statistic: 174.1 on 1 and 58 DF, p-value: < 2.2e-16

>confint(model)

2.5 % 97.5 %

(Intercept) 145.312572 167.380556

x -1.370545 -1.009446

> anova(model)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 11627.5 11627.5 174.06 < 2.2e-16 \*\*\*

Residuals 58 3874.4 66.8

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Q.N. 4) Match the following: (5 points)**

|  |  |
| --- | --- |
| Principal Component Analysis | Time Series Data Analysis |
| Autocorrelation | Dimension Reduction Technique |
| VIF | Test for Multicollinearity |
| Stepwise Regression | Model Selection Technique |
| Akaike Information Criteria | Variable Selection Method |

**Choose one question (5 or 6) (15 points)**

**Q.N. 5)** Suppose we have the following summary information about a predictor variable(x) and the response variable (y)



1. Estimate the equation of the simple linear regression line.

We have,



Therefore, the resulting simple linear regression model is 

1. Calculate the predicted value of the response variable (y) when x = 2.

We have 

1. Estimate the value of.

*We have *

*Therefore, MSE=SSE/(n-2)=53.4566/8=6.682.*

d) Estimate the standard deviation of b1.

We know that . However, MSE is an unbiased estimator of . Therefore

Hence, the standard deviation of b1 is 0.234.

Q.N. 6) For a random sample of patients we have recorded the information of presence (1) or absence (0) of chronic disease based on their age. Answer the questions using the computer output



> summary(model)

Call:

glm(formula = Disease ~ Age, family = "binomial")

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3712 -0.8485 -0.6730 1.1311 1.8151

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.62437 0.40575 -4.003 6.25e-05 \*\*\*

Age 0.03183 0.01204 2.644 0.00819 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 122.32 on 97 degrees of freedom

Residual deviance: 114.91 on 96 degrees of freedom

AIC: 118.91

Number of Fisher Scoring iterations: 4

> predict(model, data.frame(Age=40))

1

-0.3510305

> predict(model, data.frame(Age=40), type="resp")

1

0.4131325

1. How many patients are included in the dataset?

Answer: There are 98 patients in this study

1. State the equation of the logistic regression model.

Answer: The logistic regression model is

OR

1. Does the age appear to be a significant variable?

Answer : We would like to test .

Note that the p-value of the test is 0.00819, which is less than 0.05 indicating that the age is indeed a significant variable to determine the presence or absence of chronic disease.

1. What is the probability that a 40 years patient is suffering from chronic diseases?

Answer: Based on the computer output the probability is 0.413

1. Please interpret the value 0.03183 (The value of the parameter corresponding to Age)

Answer : Note that exp(0.03183)=1.0323. This means that for a one-year increase in age the odds of presence of chronic diseases has increased by 3.23 %