**PH 539: Biostatistical Methods II**

**HOMEWORK 1 (Linear Regression) [Due Feb 14, 2017]**

**Dr. Fares Qeadan: fqeadan@salud.unm.edu**

|  |
| --- |
| **PART I (MULTIPLE CHOICE AND FILLIN THE BLANK QUESTIONS):** |

**Question 1 [1 points]:**

In general, the variable whose value is fixed by the experimenter will be denoted by x and will be called the independent, predictor, or \_\_\_\_\_\_\_\_\_\_ variable. For fixed x, the second variable will be random; we denote this random variable and its observed value by Y and y, respectively, and refer to it as the dependent or \_\_\_\_\_\_\_\_\_\_ variable.

**Answer:**

**Question 2 [1 points]:**

A first step in a regression analysis involving two variables is to construct a \_\_\_\_\_\_\_\_\_\_. In such a plot, each (x, y) is represented as a point plotted on a two-dimensional coordinate system.

**Answer:**

**Question 3 [1 points]:**

The simple linear regression model is Y = β0 +β1 x +ε, where the quantity ε is a random variable assumed to be \_\_\_\_\_\_\_\_\_\_ distributed, with E (ε) = \_\_\_\_\_\_\_\_\_\_ and V (ε) =\_\_\_\_\_\_\_\_.

**Answer:**

**Question 4 [1 points]:**

The estimated regression line or least squares line for the simple linear regression model is the line whose equation is given by \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 5 [1 points]:**

In simple linear regression analysis, the \_\_\_\_\_\_\_\_\_\_, denoted by \_\_\_\_\_\_\_\_\_\_, can be interpreted as a measure of how much variability in y left unexplained by the model - that is, how much cannot be attributed to a linear relationship.

**Answer:**

**Question 6 [1 points]:**

In simple linear regression analysis, SST is the total sum of squares, SSE is the error sum of squares, and SSR is the regression sum of squares. The coefficient of determination r2 is given by R2 = \_\_\_\_\_/ SST or R2 = 1− (\_\_\_\_\_/ SST).

**Answer:**

**Question 7 [1 points]:**

If SSE = 36 and SST = 500, then the proportion of total variation that can be explained by the simple linear regression model is\_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 8 [1 points]:**

In a simple linear regression, the most commonly encountered pair of hypotheses about β1 is:

H0: \_\_\_\_\_\_\_\_\_\_ vs. H1: \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 9 [1 points]:**

The null hypothesis H0:β1 = 0 can be tested against H1: β1 ≠ 0 by constructing an ANOVA table, and rejecting H0 at α level of significance using the \_\_\_\_\_\_\_\_\_\_ test statistic.

**Answer:**

**Question 10 [1 points]:**

In testing H0:β1 = 0 versus H1: β1 ≠ 0, the t test statistic value is found to be t = 2.15. Should the null hypothesis be tested by constructing an ANOVA table, the F test would result in a test statistic value F = \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 11 [1 points]:**

A confidence interval refers to a parameter, or population characteristic, whose value is fixed but unknown to us. In contrast, a future value of Y is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future Y as a \_\_\_\_\_\_\_\_\_\_ rather than a confidence interval.

**Answer:**

**Question 12 [1 points]:**

The \_\_\_\_\_\_\_\_\_\_ is a measure of how strongly related two variables x and y are in a sample.

**Answer:**

**Question 13 [1 points]:**

Given n pairs of observations (x1, y1), (x2, y2),......... ,(xn, yn), if large x’s are paired with large y’s and small x’s are paired with small y’s, then a \_\_\_\_\_\_\_\_\_\_ relationship between the variables is implied. Similarly, it is natural to speak of x and y having a \_\_\_\_\_\_\_\_\_\_ relationship if large x’s are paired with small y’s and small x’s are paired with large y’s.

**Answer:**

**Question 14 [1 points]:**

The value of the sample correlation coefficient r is always between \_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 15 [1 points]:**

The sample correlation coefficient r equals 1 if and only if all (xi, yi) pairs lie on a straight line with \_\_\_\_\_\_\_\_\_\_ slope.

**Answer:**

**Question 16 [1 points]:**

If the sample correlation coefficient r equals -.80, then the value of the coefficient of determinations is \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 17 [1 points]:**

Which of the following statements are not true?

1. The objective of regression analysis is the exploit the relationship between two (or more) variables so that we can gain information about one of them through knowing values of the other(s).
2. Saying that variables x and y are deterministically related means that once we are told the value of x, the value of y is completely specified.
3. Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a deterministic fashion.

**Answer:**

**Question 18 [1 points]:**

Which of the following statements are not true?

1. In regression analysis, the independent variable is also referred to as the predictor or explanatory variable.
2. In regression analysis, the dependent variable is also referred to as the response variable.
3. A first step in a regression analysis involving two variables is to construct a scatter plot.
4. The simple linear regression model is Y = β0 +β1 x +ε, where the quantity ε is a random variable, assumed to be normally distributed with E (ε) = 0 and V (ε) =1.

**Answer:**

**Question 19 [1 points]:**

In simple linear regression model Y = β0 +β1x +ε, which of the following statements are not required assumptions about the random error term ε?

1. The expected value of ε is zero.
2. The variance of ε is the same for all values of the independent variable x.
3. The error term is normally distributed.
4. The values of the error term are independent of one another.
5. All of the above are required assumptions about ε.

**Answer:**

**Question 20 [1 points]:**

Which of the following statements are not correct?

1. The coefficient of determination, denoted by R2, is interpreted as the proportion of observed y variation that cannot be explained by the simple linear regression model.
2. The higher the value of the coefficient of determination, the more successful is the simple linear regression model in explaining y variation.
3. If the coefficient of determination is small, an analyst will usually want to search for an alternative model (either a nonlinear model or a multiple regression model that involves more than a single independent variable).
4. The coefficient of determination can be calculated as the ratio of the regression sum of squares (SSR) to the total sum of squares.

**Answer:**

**Question 21 [1 points]:**

Which of the following statements are not true?

1. The slope β1 of the population regression line is the true average change in the independent variable x associated with a 1– unit increase in the dependent variable y.
2. The slope of the least squares line is the best estimate for β1 of the population regression line.
3. Inferences about the slope β1 of the population regression line are based on thinking of the slope $\hat{β}\_{1}$ of the least squares line as a statistic and investigating its sampling distribution.

**Answer:**

**Question 22 [1 points]:**

Which of the following statements are not true about the sample correlation coefficient r?

1. The value of r depends on which of the two variables under study is labeled x and which is labeled y.
2. The value of r is independent of the units in which x and y are measure.
3. The value of r is always between -1 and +1, inclusive.
4. The value of r = 1 if all (xi, yi) pairs lie on a straight line with positive slope.
5. The value of r = -1 if all (xi, yi) pairs lie on a straight line with negative slope.

**Answer:**

**Question 23 [1 points]:**

Which of the following statements are not true?

1. The proportion of variation in the dependent variable explained by fitting the simple linear regression model does not depend on which variable is treated as the dependent variable.
2. A value of the sample correlation coefficient r near 0 is not evidence of the lack of a strong relationship, but only the absence of a linear relation, so that such a value of r must be interpreted with caution.
3. The sample correlation coefficient r can be used to make various inferences about the population correlation coefficient ρ.
4. The square root of the sample correlation coefficient gives the value of the coefficient of determination that would result from fitting the simple linear regression model.

**Answer:**

**Question 24 [1 points]:**

The following statement is \_\_\_\_\_\_\_\_\_\_. (TRUE/ FALSE).

Multivariate regression deals with the case where there are more than one dependent variables (DVs) while multiple regressiona deals with the case where there is one DV but more than one IV.

**Answer:**

**Question 25 [1 points]:**

If two or more explanatory variables have a linear relationship with the dependent variable, the regression is called:

1. Multiple Linear Regression
2. Simple Linear Regression
3. Multivariate Linear Regression

**Answer:**

*a In some textbooks, multiple regression is called multivariable regression which it totally not recommended.*

**PART II (STATA):**

**Question 1 [45 points (3 points for each part)]:**

Data on age and cholesterol levels [mg/ml] in 18 individuals are provided:

**ID Age Chol (mg/ml)**

1 46 3.5

2 20 1.9

3 52 4.0

4 30 2.6

5 57 4.5

6 25 3.0

7 28 2.9

8 36 3.8

9 22 2.1

10 43 3.8

11 57 4.1

12 33 3.0

13 22 2.5

14 63 4.6

15 40 3.2

16 48 4.2

17 28 2.3

18 49 4.0

1. Construct a scatter plot of the data and decide whether it supports the use of the simple linear regression model? Please provide both syntax and output

**Answer:**

1. Calculate the sample Pearson’s correlation coefficient and comment on the strength and direction of association?

**Answer:**

1. Use STATA to conduct simple linear regression to examine the effect of age on cholesterol levels. Please fill in the Parameter’s estimates and ANOVA tables respectively according to the following format:

|  |
| --- |
| **Table 1:** Parameters’ Estimates of the Simple Linea Regression Model for Predicting Cholesterol Levels |
|   | Estimate ($\hat{β})$ | Std. Error | t value | Pr(>|t|) |   |
| (Intercept) |  |  |  |  | \*\*\* |
| Age |  |  |  |  | \*\*\* |
| Residual standard error: …………… on ……….. degrees of freedom |
| Multiple R-squared: …………, Adjusted R-squared: ……….  |
| F-statistic: ………… on 1 and 16 DF, p-value: <0.0001  |
| \*\*\* p<0.0001, \*\* p<0.001, \* p<0.05 |

|  |
| --- |
| **Table 2:** ANOVA table of the Simple Linea Regression Model for Predicting Cholesterol Levels |
|   | SS |  DF |  MS | F value | P value |
| Regression |  |  |  |  |  |
| Error |  |  |  |  |  |
| Total |  |  |  |  |  |

 **Answer:**

1. Provide the Scatter-plot of age and cholesterol levels with the best fitted line, 95% confidence band and prediction intervals?

**Answer:**

1. What is the prediction equation from the conducted simple linear regression?

**Answer:**

1. What is the best estimate of σ2?

**Answer:**

1. What proportion of the observed variation in cholesterol levels can be attributed to the simple linear regression relationship between cholesterol levels and age?

**Answer:**

1. Calculate a point estimate of the true average cholesterol level when Age is 50?

**Answer:**

1. Is there sufficient evidence to conclude that there is a significant linear relationship between cholesterol levels and age? State the null and alternative hypotheses of interest before answering this question.

**Answer:**

1. Could one reach the same conclusion in I just by examining the 95% confidence interval of $\hat{β}\_{1}$?

**Answer:**

1. Was the assumption of normality for the residuals violated? Support your answer by graphical illustration and a formal test (use Studentized Residuals)? State the null and alternative hypotheses of interest before answering this question.

**Answer:**

1. Was the assumption of constant varince for the residuals violated? Support your answer by graphical illustration and a formal test (use Studentized Residuals)? State the null and alternative hypotheses of interest before answering this question.

**Answer:**

1. Does the conducted simple linear regression model suffer from outliers? Explain? Before answering this question recall that *outliers are not necessarily influential but they can be so (depending on leverage) yet high leverage points are not always influential and influential points are not necessarily outliers.*

**Answer:**

**Figure 5**: The plots of the standardized and studentized residuals indicate the presence of one outlying point.

1. Does the conducted simple linear regression model suffer from points with high leverage? Explain? Before answering this question recall that *outliers are not necessarily influential but they can be so (depending on leverage) yet high leverage points are not always influential and influential points are not necessarily outliers.*

**Answer:**

1. Does the conducted simple linear regression model suffer from influential points? Explain? Before answering this question recall that *outliers are not necessarily influential but they can be so (depending on leverage) yet high leverage points are not always influential and influential points are not necessarily outliers.*

**Answer:**

**PART III (STATA+TEXTBOOK):**

**Question 1 [25 points]:**

**2.24 Exercises 1-8 page 93 using the 2.ex.vonHippelLindau.dta data set.**

1. Please provide STATA syntax, output and answer asked questions.
2. Please provide STATA syntax, output and answer asked questions.

The required scatter plot is the plot of studentized residuals against y\_hat. The 95% prediction interval for these residuals is computed according to

**display \_N**

**display invttail(\_N-3,.025)**

So, instead of adding the reference lines -2 and +2 to the scatter-plot as shown in the following syntax:

twoway scatter ti y\_hat, yline(-2 2, lcolor(gray) lpattern(dash)) yline(0) title("Studentized Residuals vs. Fitted values")

you need to replace -2 and +2 with ± the value you obtain from **display invttail(\_N-3,.025)**. This is the value of the prediction interval for the studentized residuals.

To comment on the adequacy of the model, one need to examine:

1. Normality assumption
2. Constant variance assumptions
3. Presence of influential outliers.
4. Please provide STATA syntax, output and answer asked questions.

Lowess, is a non-parametric regression technique (locally weighted scatterplot smoothing). In the required figure, please use the option “||” twice so you could utilize ***lfitci***, ***scatter*** and ***lowess*** all together in one line of code.

1. Please provide STATA syntax, output and answer asked questions.

Answer may vary.

1. Please provide STATA syntax, output and answer asked questions.

When comparing your answer to question 1, please comment on the changes of:

1. The p-value of the F test
2. The p-value of the t-test for β1
3. R-square
4. Validity of assumptions.
5. Please provide STATA syntax, output and answer asked questions.

Remember that loge (100) =ln (100) = 4.6051702. So, don’t use the value 100 when replacing the value at line 38, instead use 4.6051702.

Hint: You syntax for this problem should start and end with

di \_N

set obs 38

replace logtumorvol = 4.6051702 in 38

predict newht

…

…

…

list logp\_ne logtumorvol newht lowerCI upperCI lowerPI upperPI in 38

1. Please provide STATA syntax, output and answer asked questions.

This question is the same as Question 2 but on the new model from Question 5.

1. Please provide STATA syntax, output and answer asked questions.

Note that the disease variable in the data set is numeric so when using the **if statement** one should use firstly **if disease==0** and secondly **if disease==1**. (0 denotes von Hippel-Lindau, 1 denotes multiple endocrine neoplasia)

To test the hypothesis, just compare the constructed confidence intervals and test if they overlap, there is no need to construct a confidence interval for the difference of the parameters’ estimates.