**I. True or false and multiple choices.**

Single answer - Select only one answer

(1) Simple linear regressions are used to compare two means.

1. True
2. False

(2) Correlation analysis is always used to evaluate the relationship between two continuous variables and can take values -1, +1 or any number between -1 and +1

1. True
2. False

(3) For linear regression analysis, which of these variables will not be used to predict the response?

1. Predictor variable
2. Explanatory variable
3. Independent variable
4. Risk factor
5. None of the above

(4) Assume that the sum of squared errors (SSE) of a simple linear regression model is 18 and the number of observations is 5, what will be the Mean Squared Error (MSE)?

1. 23
2. 13
3. 6
4. 11

(5) When we add one or more independent variables to a simple or multiple linear regression model, what happens to the value of coefficient of determination R²?

1. Decreases
2. Never decreases.
3. Unchanged
4. a and b

(6) Which of the following measures can be used in determining confounding?

1. Crude rates
2. Odds ratio
3. Relative risk
4. All of the above

(7) To minimize confounding by variables, stratification should be considered during the design of the study

1. True
2. False

(8) Randomization is a method to control confounding for a given data.

1. True
2. False.

(9) Which of the following is true?

1. The risk factors for disease can be potential confounders.
2. We can compare crude and adjusted measures of association between the exposure and the risk factor of disease to determine the confounder.
3. If the crude measure and the adjusted measure are close, the risk factor might be a confounder.
4. Both a and b.

(10) The following is not true about restriction?

1. It is cost effective
2. Results from the study provide reliable inference for the general populations.
3. It is straight forward
4. If you restrict the study to one category of a confounder, then compared groups cannot differ.

(11) Which of the following is not true about interaction

1. The causal eﬀect of a predictor on the outcome diﬀers according to the level of another predictor
2. It is the combined eﬀect of two or more independent variables acting simultaneously on a dependent variable
3. The eﬀect of two or more variables can be additive or multiplicative.
4. The interaction is called antagonism if the eﬀect is greater than would be expected.
5. The interaction is called negative if the eﬀect is less than would be expected.

(12) Which of the following can be used to control for confounding?

1. Multiple linear regression
2. Simple Logistic regression
3. Multiple Cox proportional hazards model.
4. All above, a, b and c.
5. a and c.

(13) It is easy to control for many confounding variables simultaneously because a large number of strata will be generated relative to the number of study subjects.

1. True
2. False.

(14) All are assumptions of a multiple regression analysis except? (Select one only)

1. Linearity
2. Normality
3. Dependence.
4. Homoscedasticity

(15) In detecting collinearity among independent variables, if VIF>10 then the pairwise collinearity among independent variables is detected.

1. True
2. False

(16) These rules must be adhered to in selecting a model.

1. Keep k < n/10 if possible
2. Include all variables that your theory tells you are necessary
3. Include all variables that you want to test
4. Include all variables that are statistically signiﬁcant
5. All of the above.

(17) In model selection, which of the following will never satisfy the selection criterion?

1. Lowest Adjusted R².
2. Lowest Malow’s Cp
3. Lowest Akaike Information Criterion (AIC)
4. None of the above

(18) Common mistakes in model selection include the following except

1. Not verifying that the relationship is linear
2. Relying on automated results without visual veriﬁcation
3. High R² is necessary but not suﬃcient for a good model
4. None of the above

(19) The following are important steps in the stepwise method of model selection, except,

1. The first step is to find the variable with largest partial F statistic value (or t-statistic value) and test for H0: =0.
2. If the null hypothesis is accepted, and then the variable in question is judged important and kept.
3. Next each of the remaining variables is added (separately) to the regression that already has the ﬁrst independent variable.
4. The variable with the highest Partial F statistic value is added. Now we have a two-variable model.

(20) For linear regression analysis, the residual mean square, MSE or s2 is always an unbiased estimator for σ2.

1. True
2. False.

**II. Fill in the blanks.**

The following is a partial ANOVA table for simple linear regression analysis and some values are missing. Complete the table from a simple linear regression model.

Source df sum squares mean square F p-value

Regression \_\_\_\_\_ \_\_\_\_\_ \_\_\_\_\_\_ \_\_\_\_\_\_ \_\_\_\_\_\_\_

Residual \_\_\_\_\_ 1600.00 \_\_\_\_\_\_

Total 19 1850.00

A simple linear regression model is fit on a computer spreadsheet and the following quantities are obtained:

Compute the estimated slope of the regression line (*b*) \_\_\_\_\_\_\_\_\_\_ and the estimated correlation coefficient (*r*) \_\_\_\_\_\_\_\_\_\_\_.

**III. Fill in the blanks and answer the questions**

**Table 1**. The simple linear regression of weight (Y) on height (X)

The REG Procedure

Model: MODEL1

Dependent Variable: Weight

|  |  |
| --- | --- |
| **Number of Observations Read** | 6 |
| **Number of Observations Used** | 6 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | ??? | ??? | ??? | ??? |
| **Error** | 4 | 186.18139 | ??? |  |  |
| **Corrected Total** | 5 | 386.83333 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | ??? | **R-Square** | ??? |
| **Dependent Mean** | 61.83333 | **Adj R-Sq** | 0.3984 |
| **Coeff Var** | ??? |  |  |

| **Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| **Intercept** | **1** | 1.69729 | 29.09713 | 0.06 | 0.9563 | -79.08929 | 82.48387 |
| **Height** | **1** | 1.19081 | ??? | 2.08 | ??? | ??? | ??? |

**Problems (1)-(8)** (based on Table 1)

(1) Compute the standard deviation of the errors.

(2) Compute the standard error of the slope estimate.

(3) Compute the percentage of the total variability in weightis explained by this model.

(4) Compute the sample correlation coefficient between weight and height.

(5) Write a sentence giving the interpretation of the slope.

(6) What is degrees of freedom for the overall F-test? Compute F-statistic for testing the linear relationship between weight and height, and the p-value.

(7) Estimate a 95% confidence interval for the slope.

(8) Compute p-value for testing the null hypothesis H0: slope =0 and discuss the relationship between weight and height. (Remark: Using t-statistic 2.96)

**IV. Answer the questions using SAS outputs**

**Table 2.**

The REG Procedure

Model: MODEL1

Dependent Variable: Weight

|  |  |
| --- | --- |
| **Number of Observations Read** | 5 |
| **Number of Observations Used** | 5 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | ?? | ?? | ?? | ?? | ?? |
| **Error** | ?? | 47.86447 | ?? |  |  |
| **Corrected Total** | ?? | 330.80000 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | ?? | **R-Square** | ?? |
| **Dependent Mean** | 63.20000 | **Adj R-Sq** | 0.7106 |
| **Coeff Var** | 7.74060 |  |  |

| **Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| **Intercept** | **1** | -33.46958 | 29.05523 | -1.15 | ?? | -158.48416 | 91.54500 |
| **Height** | **1** | 1.29635 | 0.41390 | 3.13 | ?? | -0.48451 | 3.07721 |
| **Age** | **1** | 3.48151 | 2.00787 | 1.73 | ?? | -5.15765 | 12.12067 |

**Problems (1)-(3)** (based on Table 2)

(1) Overall speaking, is there a linear relationship between Y and these two predictors? How do you know?

(2) What are the slopes of the variables height and age? Are they statistically significant at 5% or 1% significant level? How do you know?

(3) How much of the variability in Y is explained by the least squares regression of Y on X1 and X2?

**Table 3.**

The REG Procedure

Model: MODEL1

Dependent Variable: Weight

|  |  |
| --- | --- |
| **Number of Observations Read** | 5 |
| **Number of Observations Used** | 5 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 210.98272 | 210.98272 | 5.28 | 0.1051 |
| **Error** | 3 | 119.81728 | 39.93909 |  |  |
| **Corrected Total** | 4 | 330.80000 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 6.31974 | **R-Square** | 0.6378 |
| **Dependent Mean** | 63.20000 | **Adj R-Sq** | 0.5171 |
| **Coeff Var** | 9.99959 |  |  |

| **Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| **Intercept** | **1** | 1.59207 | 26.95337 | 0.06 | 0.9566 | -84.18559 | 87.36972 |
| **Height** | **1** | 1.22238 | 0.53184 | 2.30 | 0.1051 | -0.47018 | 2.91493 |

**Problems (4)-(5)**

(4) Is there confounding due to Age? Why?

(5)Interpret the slopes of Height and Age, respectively (based on Table 2)