CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Worksheet 07

Name (Print):________________________________ **Section:**_______________

1. Identify the correct type (Descriptive or Inferential statistics):

Type	**Objective**
	Makes inference about the population
	Provides data summary

2. Fill in the blanks:
i. When the p-value is less or equal to 0.05, you ____________________ the null hypothesis.
ii. If the p-value is greater than 0.05, you ____________________ the null hypothesis.

3. Fill in the blanks:
i. The variable to be predicted is called the ____________________ variable. It is usually represented as the letter ____________.
ii. Variables used to predict are called ____________________ variables. They are usually represented as the letter ____________.

4. Provide the general form of linear regression:

5. Provide alternative names:

Independent Variables	**Error**
1.	
2.	
3.	
4.	

6. Match the regression type with its characteristics:

Regression Type	**Characteristics**
1. Simple	a. Predicts a binary outcome (e.g., success/failure)
2. Multiple	b. More than one independent variable predicting a continuous outcome
3. Logistic	c. Handles multicollinearity by introducing a penalty term
4. Ridge	d. Performs variable selection and regularization
5. Lasso	e. A single independent variable predicting a continuous outcome

7. Fill in the blanks:
i. A regression used to predict a count variable is called ___________________________.
ii. When the response variable has more than two nominal categories, ___________________________ regression is appropriate.
iii. ___________________________ regression is used for predicting an ordered response.

8. Based on the regression types discussed, suggest which type of regression is appropriate for the following scenarios:

   a. Predicting house prices based on size, location, and age
   Regression Type:

   b. Determining the likelihood of a student passing an exam (Pass/Fail) based on study hours and attendance
   Regression Type:

9. Which statement is false:

   a. Logistic regression can be used for continuous dependent variables.

   b. Ridge regression is suitable when predictor variables are highly correlated (multicollinearity).

   c. Lasso regression performs variable selection by identifying a simpler model (=it eliminates irrelevant predictors).

10. Simple Regression assumptions. Complete the table:

Assumption Name	**Characteristics**
1. Variable type 2. Linear 3. Outliers 4. Independence 5. Equal variance (homoskedasticity)	

11. How do you determine how well the model fits data? Describe in your own words to see if you understand this concept. See slide 8