Analytics Primer

Quiz 3

Honor Pledge:

I certify that I have not received or given unauthorized aid in taking this exam.

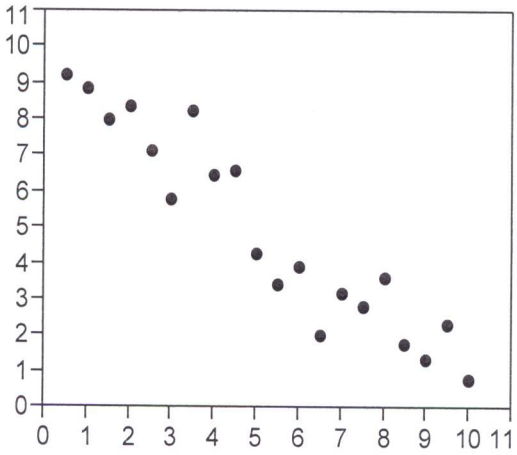
Signed: _____

Printed Name: KEY

Date: _____

There are 15 questions. The first 10 are worth 2 points each. The last 6 are worth different point values. There will be some True/False, Multiple Choice, and Short Answer questions.

Use the following scatter plot to answer questions 1 and 2.



1. From the scatter plot we say that the two variables have approximately the following relationship:
   a. Positive Nonlinear
   b. Positive Linear
   c. Negative Linear
   d. No Relationship

2. What would be the ***best*** estimate of the correlation coefficient between the two variables in the scatter plot?
   a. -0.89.
   b. -1.00.
   c. +0.64.
   d. 0.

3. In your own words, describe what correlation is. Do not give an equation, but explain what it means if we say two variables are correlated.

   Correlation is the strength of the linear relationship between two variables

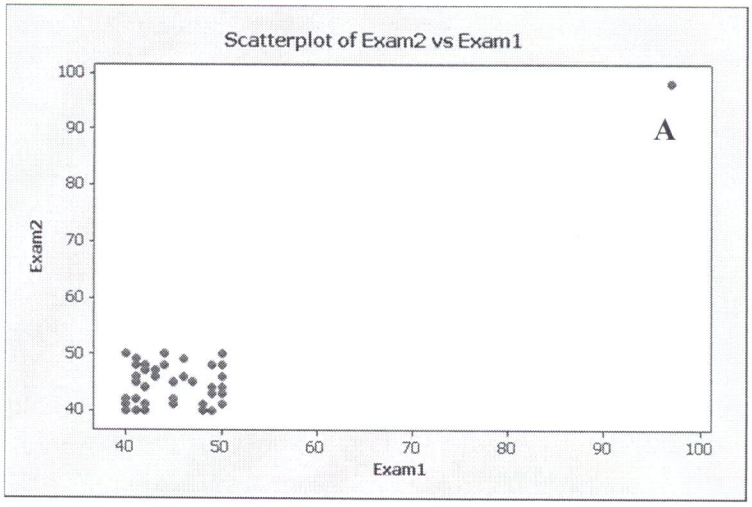
4. List two assumptions for the *simple* linear regression model.

   1. Linearity
   2. Independence of errors

   Normality of errors
   Homoskedasticity of errors (equal variance)

5. In simple linear regression, the sign of the slope coefficient must be the same as the sign of the correlation coefficient between the independent and dependent variables.
   a. True.
   b. False.

A teacher of a statistics course recorded the scores of 50 students on the two exams she gave this semester. She plotted the following scatter plot. Use this to answer question 6.



6. There is an outlier in this data (marked with the letter "A"). If we removed this outlier the following would happen to the correlation coefficient:
   a. Get closer to zero.
   b. Get closer to one.
   c. Stay the same.
   d. Need more information.

7. A marketing researcher develops two models to try and explain the usage of products at a convenience store. His first model has an $R^2 = 0.722$ and adjusted $R^2 = 0.703$. His second model is the same as the first model with one additional variable. His second model has an $R^2 = 0.727$ and adjusted $R^2 = 0.691$. What can we conclude?
   a. Something is calculated incorrectly because the adjusted R^2 is always higher than the R^2.
   b. The second model is probably better because it has a higher R^2 value.
   c. The additional variable in the second model probably hinders the overall model more than helps it.
   d. You cannot compare adjusted R^2 values between models.

8. Why do we calculate adjusted R^2 values in multiple regression instead of just R^2 values?
   a. Mathematically R^2 values always increase for extra variables added to a model.
   b. R^2 values only describe a simple linear regression, while adjusted R^2 values describe a multiple linear regression.
   c. The adjusted R^2 only increases if the addition of another variable out ways the loss of degrees of freedom.
   d. Both a and c, but not b.

A data analyst is trying to interpret the following results of a simple linear regression with an $R^2 = 0.84$. Use these results to answer question 9.

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	22.57	3.53	48	6.394	<0.0001
Slope	-9.82	1.32	48	-7.439	<0.0001

9. What is the value of the correlation coefficient r?
   a. 0.917.
   b. 0.84.
   c. -0.917.
   d. Not enough information.

10. If I were to refer to a regression line as the least squares regression line, what is meant by the term "least squares" here?

The line is chosen to minimize the sum of squared errors in the model.

An analyst in charge of mutual funds for a major financial firm is trying to use 2009 profits (in millions of dollar) from different corporations to help explain the prices of bonds (in dollars) issued by those same corporations in 2009. The analyst calculated the following simple regression model using a sample of 76 corporations. Use this information to answer questions 11 and 12.

$$\hat{y}_i = 14.98 + 3.54x_i$$

$$SSE = 854127, \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 1089$$

11. (10 points) Conduct a hypothesis test to see if the slope of the true regression line equals zero with a significance level of 0.05. Do not forget to include your statement of hypothesis, test statistic, p-value, decision rule, and conclusion.

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$t = \frac{b_1 - 0}{s_{b_1}}$

$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\Sigma(x_i - \bar{x})^2}}$ $\quad s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{859127}{74}} = 107.435$

$= \frac{107.435}{\sqrt{1089}} = 3.26$

$= \frac{3.54}{3.26} = 1.09$

P-value $= (0.2, 0.3) > \alpha \Rightarrow$ Do not REJECT H_0.

Not enough evidence to conclude profits predict bond prices.

12. (5 points) Calculate a 90% confidence interval for the slope of the regression line in the above problem.

$b_1 \pm t^* \cdot s_{b_1}$

$3.54 \pm 1.667 \cdot 3.26$ $\rightarrow$ 3.54 ± 5.43

$(-1.89, 8.97)$

A data analyst for a major car manufacturer is trying to develop a pricing model for the company's cars. The analyst came up with the following model from a sample of 54 cars. Use this information to answer questions 13 – 15.

$x_1 = Maximum\ Speed\ of\ Vehicle\ (mph)$
$x_2 = Estimated\ Highway\ Miles\ per\ Gallon$
$x_3 = Maximum\ Number\ of\ Occupants$
$x_4 = Weight\ of\ Vehicle\ (pounds)$

$$\hat{y}_i = 620.8 + 127.1x_{1,i} - 6.87x_{2,i} + 985.4x_{3,i} + 5.5x_{4,i}$$

$$SSE = 101578, \quad SSR = 236165$$

13. (5 points) Interpret the coefficient for x_3 in terms of the problem.

For every extra max. occupant a car can hold, holding all else constant, the avg. price increase is $985.40.

14. (5 points) Calculate the R^2 value and the adjusted R^2 value.

SSR+SSE = TSS

$$R^2 = \frac{SSR}{TSS} = \frac{236165}{101578 + 236165} = 0.699$$

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right) = 1 - (1 - 0.699)\left(\frac{54-1}{54-4-1}\right)$$

$$= 0.674$$

15. (5 points) Calculate the F-statistic for overall model significance in this model. Do not compute the entire hypothesis test, just the F-statistic for the hypothesis test of overall model significance.

$$F = \frac{MSR}{MSE} \qquad MSR = \frac{SSR}{k} = \frac{236165}{4} = 59041.3$$

$$MSE = \frac{SSE}{n-k-1} = \frac{101578}{54-4-1} = 2073.02$$

$$F = \frac{59041.3}{2073.02} = 28.48$$