Statistics 515 - Fall 2000 - Exam 3 *(modified a bit)* Solutions

1) Define what is meant by the *p-value* (or the *observed significance level*) of a test. **The p-value is the probability of observing a statistic as extreme as the one observed, or more extreme, if the null hypothesis is true. -or- The p-value is the smallest α-level at which the null hypothesis would be rejected.**

2) Consider the following partial ANOVA table for a one-way analysis of variance. How many different treatment groups were there in this experiment? 4 (the number of treatment df + 1)

3) A one-way analysis of variance has five different treatment groups. The means for the population means for these five groups are μ_1, μ_2, μ_3, μ_4, and μ_5. Specify the null and alternate hypothesis that are tested by the p-value in the ANOVA table. **H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$ H_A: at least two means differ**

4) In performing a linear regression to predict *y* from *x*, the following four assumptions must be satisfied:

a) The mean of the errors is zero **c) The variance of the errors is equal**

b) The errors are normal **d) The errors are independent**

5) The basic regression equation is $y = \beta_0 + \beta_1 x + \varepsilon$.
Identify which term in the equation is the:
dependent variable ***y*** slope **β_1** independent variable ***x*** intercept **β_0** error **ε**

6) PROC INSIGHT produces three plots when performing linear regression. The **scatter plot** of the independent and dependent variable, the **residual** plot of the residuals versus the predicted values, and the **q-q plot of the residuals**. Which assumption(s) do you check by looking at the residual vs. predicted plot? **That the errors have mean zero, and that the errors have constant variance.**

7) In simple linear regression, the values of β_0 and β_1 are chosen so that they minimize the **SSE**.

8) If the assumptions of a regression model for predicting *y* from *x* are met, and we do not reject the null hypothesis that $\beta_1=0$, then we conclude that *x* **cannot** be used to predict *y*. If we do reject the null hypothesis that $\beta_1=0$ then we **may not** conclude that *x* causes *y*. **[Rejecting H_0 shows correlation, not causation.]**

1a) An increase in velocity of 10 km/sec would correspond to how many more megaparsecs of distance?
0.0006*10=0.0060

b) At α=0.05 do we accept or reject H_0: β_1=0? **Accept (fail to reject) since the p-value is 0.0696 > 0.05**

c) What percent of the variation in the estimated distance is explained by the estimated velocity?
R-square = 0.2312 = 23.12%

d) What is the 95% prediction interval for a velocity of 500 km/sec? **500 km/sec is the last observation, which has a 95% prediction interval of (0.0284 megaparsecs, 1.4031 megaparsecs)**.

e) Assumptions: **We can check if the errors are normally distributed from the Q-Q plot. It is a little off the straight line, but not too bad. We could accept this assumption because the ANOVA table F-test is a robust test. We can check if the errors have the same variance from the residual vs. predicted plot. This appears to be true as each slice of values seems equally spread out. We also check if the errors have mean zero from the residual vs. predicted plot. The mean seems above zero for the low predicted values, and below zero for the high predicted values. This assumptions is not met. (We cannot check if the errors are independent from the plots).**

2) The following is the incomplete work for a linear regression problem.

SS_{xx} = 10.000 average x = 5.000
SS_{yy} = 6.000 average y = 2.000
SS_{xy} = 7.000

Source	SS	DF	MS**(=SS/DF)**	F	Prob>F
Regression	4.900	1	**4.9000**	**13.364**	0.0354
Error	1.100	3	**0.3667**	**(=MSR/MSE)**	
Total	**6.000**	**4**			

b) Determine the estimated regression equation.

slope = SSxy/SSxx = 7/10 = 0.7
intercept = average y - slope * average x = 2 - 0.7*5 = 2-3.5 = -1.5
so, **y = -1.5 + 0.7 x**

c) What was the original sample size?

the total df = n-1 = 4 and the error df = n-2 = 3, so n, the original sample size, is **5**.

d) Determine a 90% interval for the slope β_1.

α=0.10, α/2=0.05, and df=3 so t= 2.353
0.7 ± 2.353 * sqrt(0.3667)/sqrt(10)
0.7 ± 0.45 or (0.25,1.15)