STA 442/1008 Assignment 3

The Quiz will be on Friday Feb. 8th. It will be open book and notes, but you are *not* allowed to bring in written answers to the questions below. Bring a calculator. Of course **bring your log and list files to the quiz.**

I realize that this assignment is rather long, and it was posted before we covered the material at the end. But please start doing it from the beginning. I may cut off a few questions at the end if necessary, but I hope it's not necessary.

As usual, answers to the questions on this assignment are not available. However, you can check some of your answers with me or Alison, and we will tell you if we agree. Please do this in office hours rather than by email.

1. First, here is some material that could well have appeared on Quiz One.

   (a) A correlation of -0.7 means that each variable explains what percentage of the variation in the other variable?

   (b) By hand, make a scatterplot where the correlation is zero but the points all lie exactly along a straight line. Yes, it's possible.

   (c) Is it possible to have a curvilinear relationship and a zero correlation? Answer yes or no. If the answer is yes, draw an example scatterplot by hand.

   (d) Is it possible to have a curvilinear relationship and a substantial positive correlation? Answer yes or no. If the answer is yes, draw an example scatterplot by hand.

   (e) A market research firm is interested in testing two versions of a television commercial. A very large sample of consumers is randomly divided into two groups, by tossing a fair coin. If the coin comes up Heads, the person sees commercial version One. If the coin comes up Tails, the person sees commercial version Two. Each consumer views the commercial alone, in a separate room, and then is given an opportunity to purchase the product. The dependent variable is binary – purchase versus non-purchase. Is there a problem here with potential confounding variables? Explain.

   (f) A market research firm is interested in testing the effect of an advertising campaign (consisting of radio and TV ads, billboards, coupons, etc.) for Jolt Cola. A very large random sample of consumers is interviewed before the campaign begins, and asked how much Jolt Cola they have purchased during the past seven days. Then the campaign runs for a month, and the same consumers are interviewed again. Once again, they are asked how much Jolt Cola they have purchased during the past seven days. Is there a problem here with potential confounding variables? Explain.

   (g) Answer the following questions T for true or F for false. Assume the significance level is $\alpha = .05$ in all cases. You must get at least 9 out of 10 right in order to get any credit at all on this question. No marks will be deducted if you get one wrong. This is supposed to be a easy. No tricks!

      i. In an experimental study, a statistically significant relationship between the independent variable and the dependent variable can provide some evidence of a causal relationship.

ii. In simple regression, a positive regression coefficient b_1 implies that high values of X tend to go with low values of Y and low values of X tend to go with high values of Y.

iii. We observe $r = -0.70, p = .009$. We conclude that X and Y are unrelated.

iv. We would like to predict fuel efficiency from type of automobile owned (North American vs. other). It makes sense to use a matched (paired) t-test.

v. An observational study is one in which cases are randomly assigned to the different values of an independent variable.

vi. If $p < .05$, we say the results are statistically significant at the .05 level.

vii. We seek to predict the dependent variable from the independent variable.

viii. In a study attempting to predict income from education and race, we observe substantial race differences in highest grade completed. This means that income cannot be correlated with education.

ix. We observe $r = 0.50, p = .002$. This means that 50% of the variation in the dependent variable is explained by a linear relationship with the independent variable.

x. When a relationship between the independent variable and the dependent variable is statistically significant, we conclude there is no evidence that the two variables are actually related.

2. Starting with your command file for the TV data used in Assignment 2, create a new variable called `location`. Location = 1 if the household is in a rural district, Location = 2 if the household is in a small-town district, and Location = 3 if the household is in an urban (city) district. Use `proc format` to set up printing formats for `location`. Write a SAS program to answer do the following; as you will see, in all cases the dependent variable is Total hours of TV watched last week. Please do not, repeat *not* put the data step in a separate file and use `%include`; it does not show up in the log file.

   (a) Obtain n, mean and standard deviation and number of TV hours watched for households in each location. Do a one-way ANOVA to test whether average TV hours watched differs significantly as a function of urban vs. rural vs. small town location. If the results are significant, follow up with Tukey tests. In plain language that could be understood by someone with no statistical training, what do you conclude?

   (b) What proportion of the variation in number of TV hours watched is explained by urban vs. rural vs. small town location?

   (c) Make indicator dummy variables to represent location. Set it up so that the reference category is `Urban`. Start by filling in the following table.

Location	d_1	d_2	$E[Y] = \beta_0 + \beta_1 d_1 + \beta_2 d_2$
Rural			$\mu_1 =$
Small Town			$\mu_2 =$
Urban			$\mu_3 =$

(d) What do the quantities β_0, β_1 and β_2 mean, in words? The term "population mean" should occur more than once in your answer.

(e) Now put the dummy variables in your program. Please call them something other than d_1 and d_2. Use `proc reg` to reproduce the main F-test of the one-way ANOVA you did with `proc glm`.

(f) What do the t-tests from `proc reg` tell you? Compare this to the conclusions you draw from the Tukey tests. What is the advantage of the Tukey tests?

(g) Explain why number of people in the household is a potential confounding variable that should be taken into account when one examines the relationship between location and number of TV hours watched.

(h) Fill in the following table. Use generic b symbols for the regression coefficients; don't calculate their numerical values yet.

Location	d_1	d_2	$\widehat{Y} = b_0 + b_1x_1 + b_2d_1 + b_3d_2$
Rural			
Small Town			
Urban			

i. What variable in the TV data set is x_1 intended to represent?

ii. We have a separate regression line for estimating number of TV hours watched in each location. For each line, what is the slope? What is the intercept? These answers are in terms of b coefficients. Are the lines parallel?

iii. For any fixed number of people in the household, the difference in predicted TV hours between households in Rural and Urban locations is represented by ...? Answer in terms of b coefficients.

iv. For any fixed number of people in the household, the difference in predicted TV hours between households in Small Town and Urban locations is represented by ...? Answer in terms of b coefficients.

v. For any fixed number of people in the household, the difference in predicted TV hours between households in Small Town and Rural locations is represented by ...? Answer in terms of b coefficients.

(i) Use `proc reg` to test location controlling for number of people in the household (it's a single variable). What is the numerical value of the test statistic? What is the p-value? Does amount of TV watched differ according to location once you control for number of people in the household? Answer Yes or No. It is understood that by "No," you would mean that there is insufficient evidence to conclude that a difference exists.

(j) Now try to reproduce the F-test you just did, this time using `proc glm`. Just put number of people as another independent variable in the `model` statement, right before `location`. Can you find the right F statistic? You've just done an analysis of covariance.

(k) What is the simple Pearson correlation r between number of people in a household and total number of TV hours watched by the people in that household? I got this from the `proc reg` output, because I used the `corr` option.

(l) What proportion of the variation in total number of TV hours watched is explained by number of people in the household? The answer is a number.

(m) Once you try to predict number of TV hours from number of people in the household, a certain amount of the variation in number of TV hours is explained. The remaining variation is *unexplained.* Once you take number of people in the household into account, what proportion of the remaining variation is explained by location?

(n) What is the mean number of people per household? I got this from the `proc reg` output, because I used the `simple` option.

(o) Now we need to *describe* the results, and say what happened. *How* is number of TV hours related to location once we control for number of people in the household? It's not safe to base the answer on the mean number of hours watched in each location, because this does not allow for number of people in the household. Remember, even the *direction* of results (what's bigger than what) can change when we introduce additional variables into a regression.

So what we want to look at is *predicted* Y for each location. But as you saw when you filled out the last table, this depends on the number of people in the household. The most natural thing to do is to hold the number of people in the household constant at the mean value, and calculate $\widehat{Y}$ for each location given that value of x_1. Please do this in the table below, calculating a *numerical value* for $\widehat{Y}$ in each location. You are being asked for three numbers, one on each line. Actually, I'm giving you the answer for Small Town. If your $\widehat{Y}$ matches this one, you're really on the right track.

Location	$\widehat{Y} = b_0 + b_1x_1 + b_2d_1 + b_3d_2$
Rural	
Small Town	68.54925
Urban	

(p) Granted that we can't do *post hoc* (follow-up, multiple comparison) tests in this situation yet (but wait till next chapter), what does the table seem to indicate about the pattern of results? Who watches the least TV?

(q) What do the t-tests from `proc reg` tell you?

(r) Suppose you wanted to test whether, controlling for number of people in the household, the average TV hours watched is different for rural and small-town areas. You want to test whether a particular linear combination of b coefficients differs significantly from zero. What is the linear combination? (By the way, there is an equivalent formulation in terms of β values, but let's stick to bs for now.)