# STA 442/1008 Assignment 4

The Quiz will be on Friday March 15th. 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.**

1. First, we continue analysis of the TV data. Some of this is repeated from Assignment Three.

   (a) Do a single `proc reg` to answer these questions. In each case, be able to give the value of the test statistic, a $p$-value, say whether the results are significant, and answer the question Yes or No.

      i. Controlling for assessed valuation of home and number of people in the household (a single variable, of course), does amount of TV watched per household vary significantly by location?

      ii. Once you control for assessed valuation of home and number of people in the household, what proportion of the *remaining* variation is explained by location?

      iii. Controlling for location and number of people in the household, is assessed valuation of home related to amount of TV watched?

      iv. Once you control for location and number of people in the household, what proportion of the *remaining* variation is explained by assessed valuation of home?

   (b) Now we return to the attempt to predict number of TV hours from number of people in the household and location.

      i. In the following table, $x_1$ represents number of people in the household, and $d_1$ and $d_2$ are dummy variables for location. Please fill in the table.

| Location | $d_1$ | $d_2$ | $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2$ |
|---|---|---|---|
| Rural | | | $\mu_1 =$ |
| Small Town | | | $\mu_2 =$ |
| Urban | | | $\mu_3 =$ |

      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 $\beta$ 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 $\beta$ 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 $\beta$ 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 $\beta$ coefficients.

vi. 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

vii. 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.

viii. Once you take number of people in the household into account, what proportion of the remaining variation is explained by location?

ix. 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 | $\widehat{Y} = b_0 + b_1 x_1 + b_2 d_1 + b_3 d_2$ |
|---|---|
| Rural | |
| Small Town | 68.54925 |
| Urban | |

x. 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 $\beta$ values, but let's stick to $b$s for now.)

xi. Using your answer to the last question as a guide, conduct all pairwise comparisons of location, *controlling for number of people in the household*. (By the way, you've just done *post hoc* tests for an analysis of covariance.) State your conclusions in language that is as non-technical as possible.

xii. In the table below, the model includes *products* between $x_1$ and the two dummy variables, treated as additional independent variables in the regression. Find the expected value of $Y$ for each location.

| Location | $d_1$ | $d_2$ | $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \beta_4 x_1 d_1 + \beta_5 x_1 d_2$ |
|---|---|---|---|
| Rural | | | $\mu_1 =$ |
| Small Town | | | $\mu_2 =$ |
| Urban | | | $\mu_3 =$ |

A. The model says that for each location, there is a straight-line relationship between number of people in the household and number of television hours watched. For each location, what is the slope of the line? What is the intercept?

B. What would $\beta_4 = \beta_5 = 0$ mean?

xiii. Use proc reg to test the parallel slopes assumption of the analysis of covariance. What is the value of the $F$ statistic? Is it significant? What do you conclude? Is the parallel slopes assumption okay in this case, or not? If not, what should you do?

2. An experiment in dentistry seeks to test the effectiveness of a drug (HEBP) that is supposed to help dental implants become more firmly attached to the jaw bone. This is an initial test on animals. False teeth were implanted into the leg bones of rabbits, and the rabbits were randomly assigned to receive either the drug or a saline solution (placebo). Technicians administering the drug were blind to experimental condition.

   Rabbits were also randomly assigned to be "sacrificed" after either 3, 6, 9 or 12 days. At that time, the implants were pulled out of the bone by a machine that measures force in newtons and stiffness in newtons/mm. For both of these measurements, higher values indicate more healing. A measure of "pre-load stiffness" in newtons/mm is also available for each animal. This may be another indicator of how firmly the false tooth was implanted into the bone, but it might even be a covariate. Nobody can seem to remember what "preload" means, so we'll ignore this variable for now.

   The data are available in the file `bunnies.dat` (see course home page for a link). The variables are

   - Identification code
   - Time (3,6,9,12 days of healing)
   - Drug (1=HEBP, 0=saline solution)
   - Stiffness in newtons/mm
   - Force in newtons
   - Preload stiffness in newtons/mm

   Please do the following.

   (a) Use `proc freq` to find out how many rabbits are in each experimental condition.

   (b) Using `proc glm`, conduct separate univariate two-way ANOVAs on stiffness and force. Use the `means` statement to get cell means and marginal means. Be prepared to answer all the standard questions about the significance tests produced by default here.

(c) I know I said never to graph a non-significant interaction, but do it anyway just for the dependent variable `force`. A rough hand-drawn plot is fine.

(d) Being guided by just the standard tests with *no* Bonferroni corrections or other use of multiple comparisons, what do you conclude from this analysis? In particular, what would you say to the dental researchers?