# STA 302f13 Assignment Ten[1]

**Please bring your printout for Question 5 to the quiz.** The other questions are just practice for the quiz, and are not to be handed in.

1. In the general linear regression model, let $cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a *known* symmetric and positive definite matrix. As usual, $\sigma^2$ is an unknown constant.

   (a) What is the $cov(\mathbf{Y})$ for this unequal variance model?

   (b) Multiply both sides of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by $\mathbf{V}^{-1/2}$, obtaining what we will call the "transformed" model. What is the variance-covariance matrix of the error term for the transformed model?

   (c) Write down and simplify a formula for $\widehat{\boldsymbol{\beta}}$ under the transformed model.

   (d) If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, what is the distribution of $\widehat{\boldsymbol{\beta}}$ under the transformed model? Show your work.

2. The Wisconsin Power and Light Company studied the effectiveness of two devices for improving the efficiency of gas home-heating systems. The electric vent damper (EVD) reduces heat loss through the chimney when the furnace is in the off cycle by closing off the vent. It is controlled electrically. The thermally activated vent damper (TVD) is the same as the EVD except it is controlled by the thermal properties of a set of bimetal fins set in the vent. Ninety test houses were randomly assigned to have a free vent damper installed; 40 received EVDs and 50 received TVDs. For each house, energy consumption was measured for a period of several weeks with the vent damper active ("vent damper in") and for an equal period with the vent damper not active ("vent damper out". Here are the variables:

   House Identification Number

   Type of furnace (1=Forced air 2=Gravity 3=Forced water 4=Steam)

   Chimney area

   Chimney shape (1=Round 2=Square 3=Rectangular)

   Chimney height in feet

   Type of Chimney liner (0=Unlined 1=Tile 2=Metal)

   Type of house (1=Ranch 2=Two-story 3=tri-level 4=Bi-level 5=One and a half stories)

   House age in yrs

   Type of damper (1=EVD 0=TVD)

   Energy consumpt with damper active (in)

[1]Copyright information is at the end of the last page.

Energy consumpt with damper inactive (out)

Consider a model in which the response variable (Y) is average energy consumption with vent damper in and vent damper out, and the explanatory variables are age of house ($X_1$), chimney area ($X_2$) and furnace type (4 categories). There should be no interactions in your model.

(a) Write $E[Y|\mathbf{X}]$ for your model. This would be the *full* model for any $F$-test that uses the full versus reduced approach.

(b) Make a table with four rows, one for each type of furnace. Make columns showing how your dummy variables are defined, and include one wider column at the end, showing $E[Y|\mathbf{X}]$ for each furnace type.

(c) You want to test whether, controlling for age of house and chimney area, average energy consumption depends on furnace type.

   i. Give the null hypothesis in terms of the $\beta$s.

   ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(d) You want to test whether, controlling for furnace type and chimney area, average energy consumption depends on age of house.

   i. Give the null hypothesis in terms of the $\beta$s.

   ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(e) You want to test whether, controlling for age of house and chimney area, average energy is different for Forced air furnaces and Gravity furnaces.

   i. Give the null hypothesis in terms of the $\beta$s.

   ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(f) You want to test whether, controlling for age of house and chimney area, average energy consumption is different for Forced air and Forced water furnaces.

   i. Give the null hypothesis in terms of the $\beta$s.

   ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(g) You want to test whether, controlling for age of house and chimney area, average energy consumption is for Steam furnaces is different from the average of Forced air and Forced water furnaces. (You are comparing an expected value with the mean of two expected values.)

   i. Give the null hypothesis in terms of the $\beta$s.

   ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

3. High School History classes from across Ontario are randomly assigned to either a discovery-oriented or a memory-oriented curriculum in Canadian history. At the end of the year, the students are given a standardized test and the median score of each class is recorded. Please consider a regression model with these variables:

   $X_1$ Equals 1 if the class uses the discovery-oriented curriculum, and equals 0 the class it uses the memory-oriented curriculum.

   $X_2$ Average parents' education for the classroom

   $X_3$ Average parents' income for the classroom

   $X_4$ Number of university History courses taken by the teacher

   $X_5$ Teacher's final cumulative university grade point average

   $Y$ Class median score on the standardized history test.

   The full regression model has $E[Y|\mathbf{X}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$. Give $E[Y|\mathbf{X}]$ for the reduced model you would use to answer each of the following questions. Don't re-number the variables. Also, for each question please give the null hypothesis in terms of $\beta$ values.

   (a) If you control for parents' education and income and for teacher's university background, does curriculum type affect test scores? (And why is it okay to use the word "affect?")

   (b) Controlling for parents' education and income and for curriculum type, is teacher's university background (two variables) related to their students' test performance?

   (c) Controlling for teacher's university background and for curriculum type, are parents' education and income (considered simultaneously) related to students' test performance?

   (d) Controlling for curriculum type, teacher's university background and parents' education, is parents' income related to students' test performance?

4. In a study of recovery from spinal cord injury, patients were randomly assigned to four different physical therapy programmes, which will be called $A$, $B$, $C$ and $D$. The dependent variable is "mobility" (basically how well the patients can move around on their own) after two months, and severity of the initial injury is a covariate. Call the covariate $x$, and call the dummy variables $p_j$ for $j = 1, \ldots, ?$.

   (a) Write the equation for a regression model that includes the possibility of regression lines that are not parallel.

   (b) Make a table with columns showing how the dummy variables are defined. Make $D$ the reference category. Include a wider column in which you show $E(Y|x)$ for each treatment programme.

   (c) In terms of the $\beta$ coefficients of your model, what null hypothesis would you test to answer each of the following questions?

i. Are the four regression lines parallel?

ii. Are the slopes for treatments $A$, $B$ and $C$ equal?

iii. Are the slopes for treatments $A$, $B$ and $D$ equal?

iv. Is there an interaction between treatment programme an initial severity of the injury?

v. Holding initial severity of the injury constant at $x = 5$ (the definition of a "moderate" injury), do the treatments differ in their effectiveness?

vi. Holding initial severity of the injury constant at $x = 5$, which is more effective, treatment $A$ or treatment $C$?

(d) Write the last three null hypotheses in matrix form as $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

5. Please return to the Census Tract data from Assignment Seven. Fit a regression model in which crime rate is a function of `area`, `urban`, `docs`, `beds`, `hs`, and `region`. Remember that for `region`, 1=Northeast, 2=North Central, 3=South and 4=West. Make Northeast the reference category. Don't include interactions, for now. Carry out tests to answer the following questions. In each case, be able to give the value of the test statistic ($t$ or $F$), the $p$-value, state a conclusion in plain, non-technical language.

   (a) Allowing for the other variables, is percent High School graduates related to the crime rate?

   (b) Once you take regional differences into the account, are any of the other variables in the model related to crime rate? This is a simultaneous test of several regression coefficients.

   (c) Controlling for the other variables, do the four regions differ in their crime rates?

   (d) There are $\binom{4}{2}$ possible pairwise comparisons of the regions. Controlling for the other variables in the model, carry out $t$-tests for the ones that are not part of the default output. Guided by the $\alpha = 0.05$ significance level, what do you conclude. State your conclusions in plain, non-statistical language.

**Bring your R printout to the quiz.**

---