STA 2101/442 Assignment 3[1]

Except for Question 8, the questions on this assignment are practice for the quiz on Friday September 29th, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz.

1. In simple regression through the origin, there is one explanatory variable and no intercept. The model is $y_i = \beta_1 x_i + \epsilon_i$.
   (a) Find the least squares estimator of β_1 with calculus.
   (b) This model is a special case of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. What is the $\mathbf{X}$ matrix?
   (c) What is $\mathbf{X}^\top\mathbf{X}$?
   (d) What is $\mathbf{X}^\top\mathbf{y}$?
   (e) What is $(\mathbf{X}^\top\mathbf{X})^{-1}$?
   (f) What is $\widehat{\beta}_1 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$? Compare this with your answer to 1a.

2. There can even be a regression model with an intercept and no explanatory variables. In this case the model would be $y_i = \beta_0 + \epsilon_i$. Again this is a special case of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
   (a) Find the least squares estimator of β_0 with calculus. What's a least-squares estimator again? Find the parameter value(s) that make the y_i observations as close as possible to their expected values.
   (b) What is the $\mathbf{X}$ matrix?
   (c) What is $\mathbf{X}^\top\mathbf{X}$?
   (d) What is $\mathbf{X}^\top\mathbf{y}$?
   (e) What is $(\mathbf{X}^\top\mathbf{X})^{-1}$?
   (f) What is $\widehat{\beta}_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$? Compare this with your answer to 2a.

3. The linear regression model with intercept can be written in scalar form as $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$.
   (a) Why does the presence of β_0 guarantee that the sum of residuals $\sum_{i=1}^n e_i = 0$?
   (b) Defining $SSTO = \sum_{i=1}^n (y_i - \overline{y})^2$, $SSR = \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2$ and $SSE = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$, show $SSTO = SSE + SSR$. I find it helpful to switch to matrix notation partway through the calculation.

4. The U.S. Census Bureau divides the United States into small pieces called census tracts; lots of information is collected about each census tract. The census tracts are grouped into four geographic regions: Northeast, North Central, South and West. In one study, the cases were census tracts, the explanatory variables were Region and average income, and the response variable was crime rate, defined as the number of reported serious crimes in a census tract, divided by the number of people in the census tract.

   (a) Write $E(Y|x)$ for a regression model with parallel regression lines. Use indicato dummy variables with an intercept. You do not have to say how your dummy variables are defined. You will do that in the next part.

   (b) Make a table showing how your dummy variables are set up. There should be one row for each region, and a column for each dummy variable. Add a wider column on the right, in which you show $E(Y|x)$. Note that the *symbols* for your dummy variables will not appear in this column. There are examples of this format in the lecture slides and the text.for each region.

   (c) For each of the following questions, give the null hypothesis in terms of the β parameters of your regression model. We are not doing one-tailed tests, regardless of how the question is phrased.

      i. Controlling for income, does average crime rate differ by geographic region?

      ii. Allowing for income, is average crime rate different in the Northeast and North Central regions?

      iii. Correcting for income, is average crime rate different in the Northeast and Western regions?

      iv. Holding income constant, is the crime rate in the South more than the average of the other three regions?

      v. For a given fixed value of income, is the average crime rate in the Northeast and North Central regions different from the average of the South and West?

      vi. Controlling for geographic region, is crime rate connected to income?

5. Now please answer Question 4 again using *cell means coding*. That's the dummy variable scheme with an indicator for each category and no intercept. Answer all the part of the question.

6. I know you did this already, but show that $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$.

7. The preceding problem implies that if a regression model has an intercept, the residuals add to zero. In Question 3, this was critical to showing $SSTO = SSE + SSR$, so that $R^2 = \frac{SSR}{SSTO}$ makes sense. What about a regression model with cell means coding (and maybe some covariates) and no intercept? Do the residuals still add to zero so that R^2 is meaningful? Show that if a linear combination of the columns of the $\mathbf{X}$ matrix equals a column of ones (true for cell means coding), the sum of residuals is zero. Denote the $n \times 1$ column of ones by $\mathbf{j}$, and assume there is some $p \times 1$ vector $\mathbf{a}$ with $\mathbf{Xa} = \mathbf{j}$. Start by writing $\sum_{i=1}^{n} e_i$ in matrix notation.

8. Before the beginning of the Fall term, students in a first-year Calculus class took a diagnostic test with two parts: Pre-calculus and Calculus. Data are in the file `math1.data.txt`. The variables are

   - Identification code
   - Course: 1=Catch-up 2=Mainstream 3=Elite 4=NoResponse
   - Score on pre-calculus part of diagnostic test
   - Score on calculus part of diagnostic test
   - High School GPA
   - High School Calculus mark
   - High School English mark
   - University Calculus mark
   - First language
   - Sex

   These are real data, with some rough edges. Fix them up your way for now. I strongly advise against manually editing the data file. Sooner rather than later, I will post the list of fixes we will all use. You will probably have to re-run your analysis, so of course save the code. Please start by fitting a full model with all the potential explanatory variables.

   (a) Make a table showing the dummy variable coding scheme for course — that is, the one R is using by default.

   (b) What proportion of the variation in university calculus mark is explained by the variables in this model? The answer is a number from your printout.

   (c) An F-test appears in the last line of output from `summary`. In symbols, what null hypothesis is being tested?

   (d) What was the original sample size? How many cases are being used to fit the full model?

   (e) For each statistically significant t test produced by `summary` (that is, H_0 is rejected at $\alpha = 0.05$), state a conclusion in plain, non-statistical language. The statements would begin with something like "Allowing for other variables, ..." I think the results for `hsengl` and `frstlangOther` are interesting.

   (f) Carry out a test of `course` controlling for other variables. Be ready to give the value of F, the degrees of freedom and the p-value. What, if anything, do you conclude?

   Please bring your printout to the quiz.