STA 302f20 Assignment Nine[1]

The following problems are not to be handed in. They are preparation for the quiz in tutorial and the final exam. Please try them before looking at the answers. Use the formula sheet. Please remember that the R part (Question 8) is *not a group project*. You may compare numerical answers, but do not show anyone your code or look at anyone else's.

1. In an extended version of the SAT data, the dependent (response) variable is first-year university Grade Point Average (GPA) again. The independent (predictor) variables are

   $x_1 =$ Verbal SAT score

   $x_2 =$ Math SAT score

   $x_3 =$ High school Grade Point Average

   $x_4 =$ Mother's education, in years

   $x_5 =$ Father's education, in years

   $x_6 =$ Total family income,

   and also Location of the family home: City, Suburbs or Country.

   (a) First, write the regression equation. Use indicator dummy variables with an intercept, and make the regression planes parallel.

   (b) Make a table with one row for each location of the family home, showing how your dummy variables are defined. Make one more column showing $E(y|\mathbf{x})$ for each location. Note that the *symbols* for your dummy variables will not appear in this column. The lecture slides have examples.

   (c) For each of the following questions, do three things: Give the null hypothesis in the form of a statement about the β values, Give the $\mathbf{C}$ and $\mathbf{t}$ matrices in $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$, and Give $E(y|\mathbf{x})$ for the reduced model (note that expected y for the full model is always the same).

      i. Correcting for all other variables, is location of the family home related to first-year GPA?

      ii. Controlling for all other variables, is either Verbal SAT score or Math SAT score (or both) related to GPA?

      iii. When you allow for all the other variables, is family income a useful predictor of GPA?

iv. Controlling for all other variables, does expected GPA change faster as a function of Verbal SAT, or does it change faster as a function of Math SAT? No full versus reduced model for this one.

v. Once you correct for the two SAT scores and High School marks, do any of the family variables matter?

vi. Correcting for all other variables, does expected GPA change faster as a function of Mother's education, or does it change faster as a function of father's education? No full versus reduced model for this one.

vii. Holding all the other variables constant at fixed values, is Math SAT related to first-year university GPA?

viii. Controlling for the other variables, is average GPA of students from the suburbs different from average GPA of students from the city?

ix. Once you allow for location of the family home, do any of the other predictors matter?

(d) Now consider a model with cell means coding (indicator dummy variables and no intercept). The regression planes are still parallel.

i. Write $E(y|x)$.

ii. Make a table.

iii. What is the null hypothesis you would test to answer this question: Controlling for the other variables, does average GPA differ by location of the family home?

iv. What is the null hypothesis you would test to answer this question: Controlling for the other variables, is average GPA of students from the suburbs different from average GPA of students from the city?

2. It was suggested in lecture that using a different dummy variable coding scheme is just a linear transformation of the $\mathbf{X}$ matrix: $\mathbf{W} = \mathbf{XA}$, where $\mathbf{A}$ is a $(k+1) \times (k+1)$ matrix with an inverse, and $\mathbf{W}$ is the new $\mathbf{X}$ matrix. Suppose you want to switch from cell means coding to indicators with an intercept. Consider the specific case of a single categorical independent variable with three categories, and a single quantitative independent variable. Making the last category the reference category, there is a 4×4 matrix $\mathbf{A}$ such that

$$\begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ 1 & 0 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{pmatrix} \mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 1 & 0 & 0 & x_3 \\ 1 & 1 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_n \end{pmatrix}$$

Give the matrix $\mathbf{A}$. It is a matrix of specific numbers.

3. When there is more than one categorical explanatory variable in a regression model, there is no problem if the model has an intercept. But if two categorical variables are represented separately with cell means coding, there is potential trouble. Why?

4. Linear transformations of the $\mathbf{X}$ matrix are not limited to switching dummy variable schemes. In general,

$$\begin{aligned} & \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \iff \quad & \mathbf{y} = \mathbf{X}\mathbf{A}\mathbf{A}^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \iff \quad & \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

   where $\mathbf{A}$ is a $(k+1) \times (k+1)$ matrix, $\mathbf{W} = \mathbf{X}\mathbf{A}$ and $\boldsymbol{\alpha} = \mathbf{A}^{-1}\boldsymbol{\beta}$. There is a new vector of regression coefficients because the *meaning* of the regression coefficients changes when the predictor variables are transformed.

   (a) Denoting the least-squares estimate of $\boldsymbol{\alpha}$ by $\widehat{\boldsymbol{\alpha}}$, find a formula for $\widehat{\boldsymbol{\alpha}}$. Simplify. What is its connection to $\widehat{\boldsymbol{\beta}}$?

   (b) What is the vector of predicted y values for the transformed model? How does it compare to $\widehat{\mathbf{y}}$ from the original model?

   (c) Give a null hypothesis equivalent to $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$, but in terms of the transformed model. It's $H_0 : \mathbf{C}_2\boldsymbol{\alpha} = \mathbf{t}$. What is $\mathbf{C}_2$?

   (d) Compare the F^* statistics for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ and $H_0 : \mathbf{C}_2\boldsymbol{\alpha} = \mathbf{t}$. One would hope they are the same. Are they? Show your work.

5. You know that if a regression model has an intercept, the residuals add to zero. This yields $SST = SSR + SSE$, and makes $R^2 = \frac{SSR}{SST}$ meaningful.

   (a) When a regression model does not have an intercept, software authors apparently still feel the need to report an R^2. To do this, they partition not $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$, but the sum of squared deviations of the y_i around zero: $\sum_{i=1}^{n}(y_i - 0)^2 = \sum_{i=1}^{n} y_i^2$. For a model that might or might not have an intercept (it doesn't matter),

      i. Prove $\sum_{i=1}^{n} y_i^2 = SSE + \sum_{i=1}^{n} \widehat{y}_i^2$.

      ii. Write a formula for the proportion of $\sum_{i=1}^{n} y_i^2$ that is explained by the regression. This is the alternative definition of R^2 used by R and other software when the model does not have an intercept. It is often quite high compared to typical R^2 values.

(b) It turns out that for some models that do not have intercepts, the residuals still add up to zero. This is attractive because in this case the usual definition of R^2 is meaningful, and we are not stuck with the weird modified definition of R^2 from Question 5a.

Here is an easy condition to check. Let $\mathbf{j}$ denote an $n \times 1$ column of ones. Show that if there is a $(k+1) \times 1$ vector of constants $\mathbf{v}$ with $\mathbf{Xv} = \mathbf{j}$, then $\sum_{i=1}^{n} \widehat{\epsilon}_i = 0$. Another way to state this is that if there is a linear combination of the columns of $\mathbf{X}$ that equals a column of ones, then the sum of residuals equals zero. Clearly this applies to a model with a categorical explanatory variable and cell means coding.

6. 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: North Central, Northeast, 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 *no intercept* and parallel regression lines. 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 $E(y|x)$.

   (c) For each of the following questions, give the null hypothesis in terms of the β_j 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. Controlling for income, is average crime rate different in the North Central and Northeast regions?

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

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

      v. Controlling for 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?

(d) State why each of the following is a bad way to ask the question.

i. Controlling for income, does geographic region affect the average crime rate?

ii. Allowing for geographic region, does average income have any effect on crime rate?

(e) Write $E(y|\mathbf{x})$ for a regression model in which the regression lines might not be parallel. This time, use a model with an intercept. Make North Central the reference category; that's what R would do, since it's alphabetically first.

(f) Make a table showing how the 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 $E(y|\mathbf{x})$.

(g) For this new model with possibly unequal slopes, give the null hypothesis you would test in order to answer each question. Write it in scalar form, in terms of the β_j parameters.

i. Are the four regression lines parallel in the population?

ii. Is there an interaction between average income and geographic region?

iii. Does the relationship of average income to crime rate depend on geographic region?

iv. Do regional differences in average crime rate depend on the average income in the census tract?

v. Is the slope of the line relating average income to expected crime rate different for the North Central and Northeast regions?

vi. Is the slope of the line relating average income to crime rate different for the North Central and South regions?

vii. Is the slope of the line relating average income to crime rate different for the North Central and West regions?

viii. Is the slope of the line relating average income to crime rate different for the Northeast and South regions?

ix. Is the slope of the line relating average income to crime rate different for the Northeast and West regions?

x. Is the slope of the line relating average income to crime rate different for the South and West regions?

xi. Is average income related to crime rate for the South region? This is equivalent to asking if the slope of the regression line for the South region is different from zero.

xii. Is average income related to crime rate for the Northeast or South region (or both)? This is one test.

7. The usual advice is that when your regression model contains product terms to represent interactions, you should be sure to include the variables you are multiplying together. Usually, this keeps you out of trouble.

   (a) In the Census Tract problem (Question 6), suppose you have product terms to represent the interaction as in part 6e, but you omit the x variable. Make a table. What is this model saying?

   (b) Again in the Census Tract problem, suppose you have product terms to represent the interaction as in part 6e, but you omit the dummy variables for geographic region. Make a table. What is this model saying?

   (c) Staying with the Census Tract problem, suppose you use a model with no intercept and cell means coding. Try representing the interaction with four product terms. Write $E(y|\mathbf{x})$.

      i. The regression coefficients of this model cannot possibly be one-to-one with the regression coefficients of the model with product terms and an intercept. Why?

      ii. $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. Why?

      iii. Try using only three of the product terms. Just leave one out. Make a table. Is this better?

      iv. Instead of omitting one of the product terms, omit the x variable (income) from the model, leaving the product terms in. Make a table. Does this work?

8. For this problem, you will *not* be asked to upload a file with your complete R input and output. Instead, you'll do the R work during the quiz, and upload just what you did. Of course, if you do the problem below in advance and have your code handy, it will be a lot faster and easier.

   Telephone sales representatives use computer software to help them locate potential customers, answer questions, take credit card information and place orders. Twelve sales representatives were randomly assigned to each of three new software packages the company was thinking of purchasing. The data for each sales representative include the software package (1, 2 or 3), sales last quarter with the old software, and sales this quarter with one of the new software packages. Sales are in number of units sold.

   The data are available here. The URL is

   `http://www.utstat.toronto.edu/~brunner/data/legal/sales.data.txt`

   The response variable is sales this quarter.

   (a) Fit a model in which sales last quarter is ignored. This is very different from controlling for it. We want to know whether software package has any effect on sales. Why is it okay to use the word "effect?"

      i. Write $E(y|\mathbf{x})$.

      ii. What proportion of the variation in sales this quarter is explained by software package? The answer is a number from the output of `summary`.

      iii. What is the null hypothesis for testing whether software package has any effect on sales? Give the answer in terms of Greek letters from the regression model.

      iv. Give the test statistic. The answer is a number from the output of `summary`.

      v. Give the p-value. The answer is a number from the output of `summary`. The p-value is not the same as the test statistic.

      vi. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

      vii. Are the results statistically significant at the 0.05 level? Answer Yes or No.

      viii. Give the p-value for each pairwise comparison of software packages: That's 1 vs. 2, 1 vs. 3 and 2 vs. 3. Don't bother with a Bonferroni correction.

      ix. In plain, non-statistical language, what do you conclude from this analysis?

(b) Now fit a model with software package and sales last quarter as the explanatory variables, and sales this quarter as the response variable. There are no interaction terms yet.

i. Write $E(y|\mathbf{x})$. Make sure the variables are in the same order here and in your R program.

ii. What is the null hypothesis for testing whether software package has any effect on sales this quarter once you control for sales last quarter? Give the answer in terms of Greek letters from the regression model.

iii. Give the test statistic. The answer is a number. There is more than one good way to compute it.

iv. Give the p-value. The answer is a number.

v. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

vi. Are the results statistically significant at the 0.05 level? Answer Yes or No.

vii. What proportion of the *remaining* variation in sales this quarter is explained by software package once you allow for sales last quarter?

viii. Give the p-value for each pairwise comparison of software packages: That's 1 vs. 2, 1 vs. 3 and 2 vs. 3. Don't bother with a Bonferroni correction.

ix. In plain, non-statistical language, what do you conclude from this analysis?

(c) Fit a full model in which the slopes and intercepts of the regression lines relating sales last quarter to sales this quarter might depend on the kind of software the sales representatives are using.

i. Write $E(y|\mathbf{x})$. Make sure the explanatory variables are in the same order here and in your R code.

ii. What is the null hypothesis for testing whether the three slopes are equal? Give the answer in terms of Greek letters from the regression model.

iii. What is the null hypothesis for testing whether the effect of software program on sales this quarter depends on sales last quarter? Give the answer in terms of Greek letters from the regression model.

iv. Carry out an F-test to determine whether the effect of software type on sales depends on the representative's performance last quarter. Be able to state your conclusion in plain, non-statistical language.

A. Give the test statistic. The answer is a number.

B. Give the p-value. The answer is a number.

C. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

D. Are the results statistically significant at the 0.05 level? Answer Yes or No.

v. Estimate the slopes and intercepts of the three regression lines. You could base the estimates on numbers from `summary`, but instead please use the `coefficients` function to obtain $\widehat{\boldsymbol{\beta}}$ with more numerical accuracy. I don't see how you can do this without making a table.

vi. Test whether the slope is different from zero for software package two.
   A. State the null hypothesis in Greek letters.
   B. Give the test statistic. The answer is a number.
   C. Give the p-value. The answer is a number.
   D. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
   E. Are the results statistically significant at the 0.05 level? Answer Yes or No.

vii. Carry out tests to answer these questions. If they are already on the output of `summary`, use that.
   A. Are the slopes for Software 1 and 2 different? Give the uncorrected p-value.
   B. Are the slopes for Software 1 and 3 different? Give the uncorrected p-value.
   C. Are the slopes for Software 2 and 3 different? Give the uncorrected p-value.

   Protecting the three tests with a Bonferroni correction at the joint 0.05 significance level, what do you conclude? Plain language is not necessary, but you should say what happened.

9. Here is one last example of a model with a quantitative explanatory variable called x, and a categorical explanatory variable with three categories. This time we want a polynomial model, with a potentially different quadratic equation in x for each category of the categorical predictor. Accordingly, we form interaction terms by multiplying both x and x^2 by each dummy variable, as follows.

$$\begin{aligned} E(y|\mathbf{x}) &= \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x + \beta_4 x^2 \\ &+ \beta_5 d_1 x + \beta_6 d_2 x + \beta_7 d_1 x^2 + \beta_8 d_2 x^2 \end{aligned}$$

   (a) Please complete the table below. Collect terms and make it look nice. The *symbols* d_1 and d_2 should not appear in your answer, because they are either zero or one.

Category	d_1	d_2	$E(y\|\mathbf{x})$
A	1	0	
B	0	1	
C	0	0	

   (b) Suppose you wanted to test whether the three quadratic curves are parallel. What is the null hypothesis?