STA 302f13 Assignment Eleven[1]

Please bring your printout for Question 6 to the quiz. The other questions are just practice for the quiz, and are not to be handed in.

1. For this question, the *uncentered* regression model refers to

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

and the *centered* regression model refers to

$$Y_i = \alpha_0 + \alpha_1(x_{i1} - \overline{x}_1) + \cdots + \alpha_k(x_{ik} - \overline{x}_k) + \epsilon_i.$$

   (a) Give $\alpha_0, \ldots, \alpha_k$ in terms of $\beta_0, \ldots, \beta_k$.

   (b) Give $\beta_0, \ldots, \beta_k$ in terms of $\alpha_0, \ldots, \alpha_k$.

   (c) When fitting the uncentered model by ordinary least squares, the quantity $Q(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$ reaches its unique minimum when $\beta_0 = \widehat{\beta}_0, \beta_1 = \widehat{\beta}_1, \ldots, \beta_k = \widehat{\beta}_k$. Show that this same minimum is reached for the centered model when $\alpha_0 = \overline{Y}, \alpha_1 = \widehat{\beta}_1, \ldots, \alpha_k = \widehat{\beta}_k$.

   (d) Why is it clear that you could estimate $\beta_1, \ldots, \beta_k$ by centering Y as well as the X variables, and then fitting a regression through the origin?

2. Consider again the `furnace` data set described in Assignment 10. The model will have Y = average energy consumption with vent damper in and vent damper out, and the independent variables are age of house (X_1), chimney area (X_2) and furnace type (4 categories). There should be no interactions in your model, and *this time the covariates X_1 and X_2 are centered*.

   (a) Write $E[Y|\mathbf{X}_c]$ for your model. Of course only X_1 and X_2 are centered.

   (b) Make a table with four rows, showing *estimated* expected energy consumption ($\widehat{Y}$) for houses of average (sample mean) age and average (sample mean) chimney area. There is one estimate for each furnace type. Give your answer in terms of $\widehat{\beta}$ values based on your model.

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

3. As in Assignment 10, the performance of High School History students is the dependent variable in a regression with the following 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 variables X_2 through X_5 are centered this time.

   (a) Write the equation for a regression model that includes interaction terms allowing the possibility that the two regression planes (one for the discovery-oriented curriculum and one for the memory-oriented curriculum) are not parallel.

   (b) Make a table with two rows, showing the expected performance for each curriculum type.

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

      i. Are the two regression planes parallel?

      ii. Holding the covariates constant at their sample mean values, is average performance different for the two curriculum type?

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

   (e) In terms of $\widehat{\beta}$ values, give the estimated expected performance for students in classes that are average on X_2 through X_5. Give one answer for the discovery-oriented curriculum and one for the memory-oriented curriculum.

4. In the usual univariate multiple regression model, the $\mathbf{X}$ is an $n \times (k+1)$ matrix of known constants. But of course in practice, the independent variables are often random, not fixed. Clearly, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathbf{X} = \mathbf{x}$.

   (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased. You've done this before.

   (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally.

   (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an F-test comparing full and reduced models), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the independent variable values. That is, $P(F > f_c|\mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. Assume that the independent variables are discrete, so you can write a multiple sum.

5. Consider the following model with random independent variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \\ &= \alpha + \boldsymbol{\beta}'\mathbf{X}_i + \epsilon_i, \end{aligned}$$

   where

$$\mathbf{X}_i = \left(\begin{array}{c} X_{i1} \\ \vdots \\ X_{ik} \end{array} \right)$$

   and $\mathbf{X}_i$ is independent of ϵ_i.

   Here, the symbol α is used differently than in Question 1. This time it's the intercept of an uncentered model; and $\boldsymbol{\beta}$ does not include the intercept. The "independent" variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})'$ are not statistically independent. They have the symmetric and positive definite $k \times k$ covariance matrix $\boldsymbol{\Sigma}_x = [\sigma_{ij}]$, which need not be diagonal. They also have the $k \times 1$ vector of expected values $\boldsymbol{\mu}_x = (\mu_1, \ldots, \mu_k)'$.

   (a) Let $\boldsymbol{\Sigma}_{xy}$ denote the $k \times 1$ matrix of covariances between Y_i and X_{ij} for $j = 1, \ldots, k$. Calculate $\boldsymbol{\Sigma}_{xy} = C(\mathbf{X}_i, Y_i)$, obtaining $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_x \boldsymbol{\beta}$.

   (b) Solve the equation above for $\boldsymbol{\beta}$ in terms of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{xy}$.

   (c) Using the expression you just obtained and letting $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ denote matrices of *sample* variances and covariances, what would be a reasonable estimator of $\boldsymbol{\beta}$ that you could calculate from sample data?

(d) To see that your "reasonable" (Method of Moments) estimator is actually the usual one, first verify that the matrix $\frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c$ is a sample variance-covariance matrix. Show some calculations. What about $\frac{1}{n-1}\mathbf{X}_c'\mathbf{Y}_c$?

(e) In terms of $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$, what is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}_c$?

6. Please return to the Census Tract data of Assignments Seven and Ten. This time, fit a regression model in which crime rate is a function of just `docs` and `region`, but `docs` is centered and there are interactions in the full model. Remember that for `region`, 1=Northeast, 2=North Central, 3=South and 4=West. Make Northeast the reference category.

   (a) Estimate the expected crime rate for each region when the number of doctors is held constant at the sample mean level. Your answer is a set of four numbers.

   (b) 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 — except for the last one, where the answer is just Yes or No.

      i. For census tracts with an average (sample mean) number of doctors, is there a diference in expected crime rate between the Northeast and West regions?
      ii. For census tracts with an average (sample mean) number of doctors, is there a diference in expected crime rate between the Northeast and South regions?
      iii. For census tracts with an average (sample mean) number of doctors, is there a diference in expected crime rate between the North Central and South regions?
      iv. For census tracts with an average (sample mean) number of doctors, is there a diference in expected crime rate between the North Central and West regions?
      v. For census tracts with an average (sample mean) number of doctors, is there a diference in expected crime rate between the South and West regions?
      vi. Are the regression lines for the Northeast and South regions parallel?
      vii. Is there evidence that the regression lines for the four regions are not parallel?

Bring your R printout to the quiz.
