STA 302f13 Assignment Seven[1]

This assignment uses the data file CensusTract.data, given in *Applied Linear Statistical Models* (1996), by Neter et al.. The data are used here without permission. There is a link on the course home page in case the one in this document does not work.

The cases (there are n cases) are a sample of census tracts in the United States. For each census tract, the following variables are recorded.

`area`	Land area in square miles
`pop`	Population in thousands
`urban`	Percent of population in cities
`old`	Percent of population 65 or older
`docs`	Number of active physicians
`beds`	Number of hospital beds
`hs`	Percent of population 25 or older completing 12+ years of school
`labor`	Number of persons 16+ employed or looking for work
`income`	Total Total before tax income in millions of dollars
`crimes`	Total number of serious crimes reported by police
`region`	Region of the country: 1=NE, 2=NC, 3=S, 4=W

1. First, fit a regression model with `crimes` as the dependent variable and just one independent variable: `pop`.

   (a) In plain, non-statistical language, what do you conclude from this analysis? The answer is something about population size and number of crimes.

   (b) What proportion of the variation in number of crimes is explained by population size? The answer is a number between zero and 1.

   Bring your printout to the quiz.

2. Based on that last analysis, we will create a new dependent variable called crime *rate*, defined as number of crimes divided by population size. Now fit[2] a new regression model in which crime rate is a function of `area`, `urban`, `old`, `docs`, `beds`, `hs`, `labor` and `income`.

   Based on this model,

   (a) What is k? The answer is a number.

   (b) What is $\widehat{\beta}_4$? The answer is a number.

   (c) Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.

      i. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_8 = 0$

      ii. $H_0 : \beta_6 = 0$

      iii. $H_0 : \beta_0 = 0$

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

[2] To "fit" a model means to estimate the parameters.

(d) What proportion of the variation in crime rate is explained by the independent variables in this model? The answer is a number.

(e) What is the smallest value of $\widehat{\epsilon}_i$? The answer is a number.

(f) What is the largest value of $\widehat{\epsilon}_i$? The answer is a number.

(g) Look at the output of `summary`. For the first entry under "`t value`" (that's `2.057`), what is the null hypothesis? The answer is a symbolic statement involving one or more Greek letters.

(h) Look at the F test at the end of the `summary` output. What is the null hypothesis? The answer is a symbolic statement involving one or more Greek letters.

(i) Controlling for all the other variables in the model, is number of hospital beds related to crime rate?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

iii. Give the p-value. The answer is a number from your printout.

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. Allowing for other variables, census regions with more hospital beds tend to have _____ crime rates.

(j) Controlling for all the other variables in the model, is number of physicians related to crime rate?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

iii. Give the p-value. The answer is a number from your printout.

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. Allowing for other variables, census regions with more physicians tend to have _____ crime rates.

(k) Predict the crime rate for a new census tract with an area of 2,500 square miles, 50 percent urban, 10 percent senior citizens, 2,000 doctors, 6,000 hospital beds, 50 percent finished high school, a labour force of 450 thousand, and a total income of 6,500 million dollars. Give both a predicted value (a single number) and a 95% prediction interval.

Bring your printout to the quiz.

3. The general linear model with normal error terms is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the columns of $\mathbf{X}$ are linearly independent, and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. You know that

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$
- $SSE/\sigma^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/\sigma^2 \sim \chi^2(n-k-1)$, independent of $\widehat{\boldsymbol{\beta}}$.

Derive the $(1-\alpha) \times 100\%$ prediction interval for a new observation from this population.
