# STA 302f14 Assignment Six[1]

These problems are preparation for the quiz in tutorial on Friday October 24th, and are not to be handed in. For reference, 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 $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

1. In the general linear model with normal error terms, what is the distribution of $\mathbf{Y}$?
2. You know that the least squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. What is the distribution of $\widehat{\boldsymbol{\beta}}$? Show the calculations.
3. Let $\widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. What is the distribution of $\widehat{\mathbf{Y}}$? Show the calculations.
4. Let the vector of residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$. What is the distribution of $\hat{\boldsymbol{\epsilon}}$? Show the calculations. Simplify both the expected value (which is zero) and the covariance matrix.
5. Recall from an earlier homework problem that if $\mathbf{T}$ is a random vector with expected value $\boldsymbol{\mu}$, then $cov(\mathbf{T}) = E(\mathbf{T}\mathbf{T}') - \boldsymbol{\mu}\boldsymbol{\mu}'$. Using this fact, give expressions for
   (a) $E(\mathbf{Y}\mathbf{Y}')$
   (b) $E(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}')$

   These may be helpful in the next question.
6. For the general linear regression model, show that the $n \times (k+1)$ matrix of covariances $C(\hat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$. Why does this show that $SSE = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{\beta}}$ are independent?
7. In an earlier Assignment, you proved that
$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$
   Starting with this expression, show that $SSE/\sigma^2 \sim \chi^2(n-k-1)$. Use the formula sheet.
8. The $t$ distribution is defined as follows. Let $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$, with $Z$ and $W$ independent. Then $T = \frac{Z}{\sqrt{W/\nu}}$ is said to have a $t$ distribution with $\nu$ degrees of freedom, and we write $T \sim t(\nu)$.

   For the general fixed effects linear regression model, tests and confidence intervals for linear combinations of regression coefficients are very useful. Derive the appropriate $t$ distribution and some applications by following these steps. Let $\mathbf{a}$ be a $p \times 1$ vector of constants.

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

(a) What is the distribution of $\mathbf{a}'\widehat{\boldsymbol{\beta}}$? Show a little work. Your answer includes both the expected value and the variance.

(b) Now standardize the difference (subtract off the mean and divide by the standard deviation) to obtain a standard normal.

(c) Divide by the square root of a well-chosen chi-squared random variable, divided by its degrees of freedom, and simplify. Call the result $T$.

(d) How do you know numerator and denominator are independent?

(e) Suppose you wanted to test $H_0 : \mathbf{a}'\boldsymbol{\beta} = c$. Write down a formula for the test statistic.

(f) For a regression model with four independent variables, suppose you wanted to test $H_0 : \beta_2 = 0$. Give the vector $\mathbf{a}$.

(g) For a regression model with four independent variables, suppose you wanted to test $H_0 : \beta_1 = \beta_2$. Give the vector $\mathbf{a}$.

(h) Letting $t_{\alpha/2}$ denote the point cutting off the top $\alpha/2$ of the $t$ distribution with $n-k-1$ degrees of freedom, derive the $(1-\alpha) \times 100\%$ confidence interval for $\mathbf{a}'\boldsymbol{\beta}$. "Derive" means show the High School algebra.

9. Letting $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ and $SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$, show $SST = SSR + SSE$.

10. Show that $\overline{Y}$ is a function of $\widehat{\boldsymbol{\beta}}$. Why does this establish that $SSR$ and $SSE$ are independent?

11. If $H_0 : \beta_1 = \cdots = \beta_k = 0$ is true,

    (a) What is the distribution of $Y_i$?

    (b) What is the distribution of $\frac{SST}{\sigma^2}$? Just write down the answer. You already did it in Assignment 2, and again in Assignment 5.

12. Still assuming $H_0 : \beta_1 = \cdots = \beta_k = 0$ is true, what is the distribution of $SSR/\sigma^2$? Use the formula sheet and show your work.

13. Suppose $H_0 : \beta_1 = \cdots = \beta_k = 0$ were *false*. Would you expect $SSR$ to be bigger, or would you expect it to be smaller? Which one, and why?

14. Recall the definition of the $F$ distribution. If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ are independent, $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$. How do you know $F = \frac{SSR/k}{SSE/(n-k-1)}$ has an $F$ distribution under $H_0 : \beta_1 = \cdots = \beta_k = 0$? List the numbers of the questions that establish the necessary facts.

This rest of 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 |

15. First, fit[2] 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.**

16. Based on that last analysis, we will create a new dependent variable called crime *rate*, defined as number of crimes divided by population size. The `attach` function should help; type `help(attach)` at the R prompt. Now fit 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. $H_0 : \beta_0 = 0$

[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.

**Bring your printout to the quiz.**

---