STA 302f17 Assignment Six[1]

These problems are preparation for the quiz in tutorial on Thursday November 2nd, and are not to be handed in.

1. Show that if $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \Sigma)$, with Σ positive definite, then $y = (\mathbf{w} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w} - \boldsymbol{\mu})$ has a chi-squared distribution with p degrees of freedom.

2. Let $y_1, \ldots, y_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The sample variance is $s^2 = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n-1}$.
   (a) Show $Cov(\overline{y}, y_j - \overline{y}) = 0$ for every $j = 1, \ldots, n$.
   (b) How do you know that $\overline{y}$ and s^2 are independent?
   (c) Show that
   $$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$
   Hint: $\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \overline{y} + \overline{y} - \mu)^2 = \ldots$

3. Recall the definition of the t distribution. If $z \sim N(0,1)$, $w \sim \chi^2(\nu)$ and z and w are independent, then $t = \frac{z}{\sqrt{w/\nu}}$ is said to have a t distribution with ν degrees of freedom, and we write $t \sim t(\nu)$. As in the last question, let $y_1, \ldots, y_n$ be random sample from a $N(\mu, \sigma^2)$ distribution. Show that $t = \frac{\sqrt{n}(\overline{y} - \mu)}{s} \sim t(n-1)$.

4. For the general linear regression model with normal error terms, prove that the $(k+1) \times n$ matrix of covariances $cov(\mathbf{b}, \mathbf{e}) = \mathbf{0}$. Why does this show that $SSE = \mathbf{e}'\mathbf{e}$ and $\mathbf{b}$ are independent?

5. Calculate $cov(\mathbf{e}, \widehat{\mathbf{y}})$; show your work. Why should you have known this answer without doing the calculation, assuming normal error terms? Why does the assumption of normality matter?

6. In an earlier Assignment, you proved that
   $$(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{e}'\,\mathbf{e} + (\mathbf{b} - \boldsymbol{\beta})'X'X(\mathbf{b} - \boldsymbol{\beta}).$$
   Starting with this expression and assuming normality, show that $\mathbf{e}'\,\mathbf{e}/\sigma^2 \sim \chi^2(n-k-1)$. Use the formula sheet.

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

7. 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 ν 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 $\boldsymbol{\ell}$ be a $k+1 \times 1$ vector of constants.

   (a) What is the distribution of $\boldsymbol{\ell}'\mathbf{b}$? Your answer includes both the expected value and the variance.

   (b) Now standardize $\boldsymbol{\ell}'\mathbf{b}$ (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 $H_0 : \boldsymbol{\ell}'\boldsymbol{\beta} = \gamma$. Write down a formula for the test statistic. A statistic is a function of the sample data that is *not* a function of any unknown parameters.

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

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

   (h) Consider a data set in which there are n first-year students in ECO100. x_1 is High School Calculus mark, x_2 is High School grade point average, x_3 is score on a test of general mathematical knowledge, and y is mark in ECO100. You seek to estimate expected mark for a student with a 91% in High School Calculus, a High School GPA of 83%, and 24 out of 25 on the test. You are estimating $\boldsymbol{\ell}'\boldsymbol{\beta}$. Give the vector $\boldsymbol{\ell}$.

   (i) 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 $\boldsymbol{\ell}'\boldsymbol{\beta}$. "Derive" means show the High School algebra.

8. For the general linear model with normal errors,

   (a) Let C be an $m \times (k+1)$ matrix of constants with linearly independent rows.What is the distribution of $C\mathbf{b}$?

   (b) If $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$ is true, what is the distribution of $\frac{1}{\sigma^2}(C\mathbf{b}-\boldsymbol{\gamma})'(C(\mathbf{X}'\mathbf{X})^{-1}C')^{-1}(C\mathbf{b}-\boldsymbol{\gamma})$? Please locate support for your answer on the formula sheet. For full marks, don't forget the degrees of freedom.

   (c) What other facts on the formula sheet allow you to establish the F distribution for the general linear test? The distribution is *given* on the formula sheet, so of course you can't use that. In particular, how do you know numerator and denominator are independent?

9. Suppose you wish to test the null hypothesis that a *single* linear combination of regression coefficients is equal to zero. That is, you want to test $H_0 : \boldsymbol{\ell}'\boldsymbol{\beta} = 0$. Referring to the formula sheet, verify that $F = t^2$. Show your work.

10. The exact way that you express a linear null hypothesis does not matter. Let A be an $m \times m$ nonsingular matrix (meaning A^{-1} exists), so that $C\boldsymbol{\beta} = \boldsymbol{\gamma}$ if and only if $AC\boldsymbol{\beta} = A\boldsymbol{\gamma}$. This is a useful way to express a logically equivalent linear null hypothesis. Show that the general linear test statistic F for testing $H_0 : (AC)\boldsymbol{\beta} = A\boldsymbol{\gamma}$ is the same as the one for testing $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

11. For the general linear regression model with normal error terms, show that if the model has an intercept, then $\mathbf{e}$ and $\overline{y}$ are independent. If you can show that $\overline{y}$ is a function of $\mathbf{b}$, you are done (why?). Here are some ingredients to start you out. For the model with intercept,

    (a) What does $X'\mathbf{e} = \mathbf{0}$ tell you about $\sum_{i=1}^n e_i$?

    (b) Therefore what do you know about $\sum_{i=1}^n y_i$ and $\sum_{i=1}^n \widehat{y}_i$?

    (c) Now show that $\mathbf{e}$ and $\overline{y}$ are independent.

12. Carefully examine the formulas for $SST = SSE + SSR$ on the formula sheet. How do you know that SSR and SSE are independent if the model has an intercept?

13. Continue assuming that the regression model has an intercept. Many statistical programs automatically provide an *overall* test that says none of the independent variables makes any difference. If you can't reject that, you're in trouble. Supposing $H_0 : \beta_1 = \cdots = \beta_k = 0$ is true,

    (a) What is the distribution of y_i under H_0?

    (b) What is the distribution of $\frac{SST}{\sigma^2}$? Just write down the answer. Check Problem 2.

14. Still assuming $H_0 : \beta_1 = \cdots = \beta_k = 0$ is true and the model has an intercept, what is the distribution of SSR/σ^2? Use the formula sheet and show your work. Don't forget the degrees of freedom.

15. 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, $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$. Show that $F = \frac{SSR/k}{SSE/(n-k-1)}$ has an F distribution under $H_0 : \beta_1 = \cdots = \beta_k = 0$? Refer to the results of questions above as you use them.

16. The null hypothesis $H_0 : \beta_1 = \cdots = \beta_k = 0$ is less and less believable as R^2 becomes larger. Show that the F statistic of Question 15 is an increasing function of R^2 for fixed n and k. This means it makes sense to reject H_0 for large values of F.

This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/302f17