STAT 425 — Fall 2018

Homework 1

(Posted on Fri Aug. 31; Due Tues Sep. 11)

Please submit your assignment *on paper*, following the Formatting Guidelines for Homework Submission. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. The simple linear regression model is specified by

$$y_i \;=\; \beta_0 \;+\; \beta_1 x_i \;+\; \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

with $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2 > 0$, and $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$.

The *ordinary least squares* estimates minimize the *residual sum of squares*

$$RSS(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

(a) Take partial derivatives of RSS with respect to each parameter (separately), and set the resulting expressions equal to zero. (These are equivalent to what are called the *normal equations.*)

(b) Solve the equations of the previous part simultaneously for β_0 and β_1. Simplify your expressions to the form of the least squares estimates given in lecture.

(c) To solve the equations in part (b), an assumption must be made about the values of X. What is that assumption, and why is it needed?

(d) Denote $\beta := (\beta_0, \beta_1)$ and rewrite RSS in the form of

$$RSS(\beta) = (\beta - b)^T A(\beta - b) + c,$$

identifying the vector b, matrix A and the scalar c.

(e) Show that A is a symmetric and positive definite matrix under the same assumption needed in part (c). As a result, show that the solution in part (b) is the unique minimizer of RSS.

2. Show that in simple linear regression, residuals have **sample** mean zero and that the **sample** correlation between the residuals and the predictor values is also zero.

3. For a constant matrix $\boldsymbol{A}$ and a random vector $\boldsymbol{z}$,

$$\mathrm{E}(\boldsymbol{A}\boldsymbol{z}) \;=\; \boldsymbol{A}\,\mathrm{E}(\boldsymbol{z}) \qquad \mathrm{var}(\boldsymbol{A}\boldsymbol{z}) \;=\; \boldsymbol{A}\,\mathrm{var}(\boldsymbol{z})\boldsymbol{A}^T$$

(assuming expectations and variances all exist).

Under the Gauss-Markov conditions, determine the mean vector and the variance-covariance matrix for each of the following vectors (in terms of $\boldsymbol{X}$, $\boldsymbol{\beta}$, and σ^2). Simplify, if possible.

(a) $\boldsymbol{y}$

(b) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$

(c) $\boldsymbol{y} - \hat{\boldsymbol{y}}$

(d) $\hat{\boldsymbol{y}}$

4. The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988.

   (a) Fit a regression model with weekly wages as the response and years of education as predictor. Plot the predictor on X axis and response on Y-axis and plot the fitted regression line on the same figure. Present the regression output.

   (b) What percentage of variation in the response is explained by these predictors? (Percentage variance explained is the same as coefficient of determination).

   (c) Which observation has the largest magnitude for the residual? Give the case number.

   (d) Compute the mean and median of the residuals. Explain what the difference between the mean and the median indicates.

   (e) For two people with one year difference in education, what would be the difference in predicted weekly wages?

   (f) Compute the correlation of the residuals with the fitted values. Plot residuals against fitted values. Explain the value of this correlation using the geometric (projection) interpretation of least squares.

Some reminders:

- Unless otherwise stated, all data sets are from the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.