STAT 425 — Fall 2018

Homework 2

(Due at 11 am on Tue, September 25 at 9:30 AM)

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

1. Using R, create a 10×4 matrix X:

$$X = \begin{pmatrix} 1 & 0 & 2 & -2 \\ 1 & 0 & -1 & -2 \\ 1 & 0 & 3 & -2 \\ 1 & 0 & -2 & 3 \\ 1 & 0 & 2 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

   Now create a 4×1 matrix β whose entries are 0, 1, -1, and 2. Next create a 10×1 matrix ϵ whose entries are IID standard normal (useful command: "rnorm"). Finally, set $Y = X\beta + \epsilon$.

   (a) Calculate $(X'X)^{-1}X'Y$ to estimate β. What do you get?
   (Don't use the "lm" command. Do the computation directly. You can use the "solve" command to compute a matrix inverse.)

   (b) What is the true variance of $\hat{\beta}$? What is the true variance of $\hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$?
   (Remember that the variance of $\hat{\beta}$ is a 4×4 matrix.) ("true" variance because, in this example, we know the true value of σ^2.)

   (c) Use the residuals to estimate σ^2. What do you get?

   (d) Now create a new ϵ and re-estimate β. Do this 500 times using an R loop, and save all the answers in memory. Make a histogram of the 500 values of $\hat{\beta}_1$. Do the same for $\hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$. Also calculate the sample variance for $\hat{\beta}_1$ and $\hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$ based on the 500 replications. Do your answers match with question (b)?

   (e) Repeat (d), but instead of using a normal distribution for ϵ use some other distribution that also has expectation 0 and variance 1. Do your answers change much? Explain. Summarize what you learn from this problem.

2. Consider the following simple linear regression models for data (x_1, y_1), $(x_2, y_2), \ldots (x_n, y_n)$:

$$M1: \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$
$$M2: \quad y_i = \alpha_0 + \alpha_1 x_i^* + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where $x_i^* = \frac{\sqrt{n}(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Model M2 is a simple linear regression model with the predictor standardized to have sample mean zero and variance 1. Assume that the errors satisfy Gauss-Markov assumptions.

(a) Find the least squares estimators for α_0 and α_1. How are they related to β_0 and β_1?

(b) What is $var(\hat{\boldsymbol{\alpha}})$?

(c) Consider the estimated mean response for a new value at $x = x_{new}$, given by $\hat{\mu} = \hat{\alpha}_0 + \hat{\alpha}_1(x_{new} - \bar{x})$. What is the variance of $\hat{\mu}$? Provide a 95% confidence interval for the mean parameter $\mu = \alpha_0 + \alpha_1(x_{new} - \bar{x})$.

3. The data set `salmonella` is from an experiment to examine the relationship between number of bacterial colonies on a plate and applied dosage level of quinoline. Consider the (simple) linear regression of `colonies` (dependent variable) upon `dose` (independent variable).

(a) Examine the data, and determine how many *replications* r there are for each dosage level.

(b) Fit the simple linear regression model, and produce a summary of your results.

(c) Consider the following new dose levels $x = (20, 40, 60, 80, 120, 150, 180, 300, 500)$. Provide the predicted values along with their 95% confidence intervals in a figure by plotting the dose levels on X-axis and the predicted values, lower and upper confidence limits on Y-axis.

4. For the `prostate` data set, fit a model with `lpsa` as the response, and the other variables as predictors.

(a) Suppose a new patient with the following values arrives:

`lcavol = 1.5`, `lweight = 3.6`, `age = 60`, `lbph = 0.3`,
`svi = 0`, `lcp = -0.8`, `gleason = 7`, `pgg45 = 15`.

Predict the `lpsa` for this patient along with an appropriate 95% prediction interval.

(b) Repeat the questions in (a) for a patient with the same values except that his age is 20. Explain why the prediction interval is wider.

(c) For the model of the previous question, remove all the predictors that are not significant at the 5% level. Using the reduced model recompute the predictions for the x values given in the previous questions (a) and (b). Are the new prediction intervals wider or narrower than in parts (a) and (b)? Which predictions would you prefer? Explain.

5. Consider the simple linear regression model $Y = \alpha + \beta X + \epsilon$ with the standard assumptions on ϵ with $n = 2m$ observations. Partition X and Y as $X = (X_1, X_2), Y = (Y_1, Y_2)$, where

(X_1, Y_1) and (X_2, Y_2) correspond to equal halves of the data with m observations each. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the least squares estimates obtained by regressing Y_1 on X_1 and Y_2 on X_2, respectively, and let $\tilde{\beta} = \frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2)$.

(a) If $\bar{X}_1 = \bar{X}_2$, state a condition such that the variance of $\tilde{\beta}$ is the same as the least squares squares estimate $\hat{\beta}$ obtained by regressing Y on X. Then state whether when this condition holds, $\tilde{\beta}$ is the same as the least squares estimate, or if it is a different estimate with the same variance.

(b) Now consider the more general case where $\bar{X}_1 \neq \bar{X}_2$. Show that in this case the variance of $\tilde{\beta}$ is always greater than the variance of $\hat{\beta}$.

6. Recall that the variance of a random vector $z_{m\times 1}$ is an $m \times m$ matrix with its $(i,j)^{th}$ element given by $cov(z_i, z_j)$. Similarly, the covariance matrix of two random vectors $z_{m\times 1}$, $w_{k\times 1}$ is an $m \times k$ matrix with its $(i,j)^{th}$ element given by $cov(z_i, w_j)$. Based on these definitions, prove the following:

- For non-random matrices $B_{n_1\times m}$ and $C_{n_2\times m}$, $Cov(Bz, Cz) = BVar(z)C^T$.

Based on these results, Find the covariance matrix of

- the residuals $\hat{\epsilon}$ and the fitted values $\hat{y}$ (the covariance matrix will have a dimension of $n \times n$).
- the response y and the fitted values $\hat{y}$.

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.