STAT 425 — Spring 2016 — Sections 1UG, 1GR

Homework 2

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. Using the `stopping` data set (from package `alr4`), fit a simple linear regression model with the stopping distance as the response and speed as the predictor. Answer the following parts.

   In parts (b) through (g), you should draw a specific conclusion and clearly refer to the diagnostic tool(s) (plots or statistics) you used to draw your conclusion.

   (a) [2 pts] Produce the four default diagnostic plots given by R.

   (b) [2 pts] Using an appropriate diagnostic plot, check whether the assumed *mean* function is appropriate.

   (c) [2 pts] Check the assumption of constant variance.

   (d) [2 pts] What is the largest (most positive) least-squares residual value? What is the smallest (most negative) least-squares residual value?

   (e) [2 pts] Identify one observation that has the largest leverage value.

   (f) [2 pts] Check for outliers. (You should use diagnostic plots — a formal test is not necessary.)

   (g) [2 pts] Check for influential points.

2. [14 pts] Using the `drugcost` data set (from package `alr4`), fit a model with `COST` as the response and all of the other variables as predictors. Then answer the same parts as in Problem 1.

3. Using the `fuel2001` data set (from package `alr4`), fit a model with `FuelC` as the response and `Tax`, `Drivers`, and `Income` as predictors. Answer the following.

   (a) [2 pts] Produce a plot of the *standardized* residuals r_i versus the ordinary (least squares) residuals $\hat{e}_i$. (Show R code.)

   (b) [2 pts] The points in this plot *do not* exactly fall on a straight line. Briefly explain why. [Hint: What is the formula for the standardized residuals?]

   (c) [2 pts] List the *studentized* residuals t_i (which are used as test statistics in the Mean Shift Test).

   (d) [2 pts] Perform all Mean Shift Tests *without* Bonferroni adjustment, using $\alpha = 0.05$. Which states are identified as outliers?

   (e) [2 pts] Perform all Mean Shift Tests *with* Bonferroni adjustment, using $\alpha = 0.05$. Which states are identified as outliers?

4. Using R, produce a grid of 9 normal probability plots (`qqnorm`) for samples of size $n = 50$ simulated independently from a *geometric* distribution with parameter `prob` equal to 0.4.

   (Refer to the last section of the *Diagnostics in Linear Regression* slides. Use the R function `rgeom` to simulate from the geometric distribution.)

(a) [2 pts] Display your plots and the R code you used to produce them.

(b) [2 pts] Describe *two* distinct ways in which these plots tend to differ in appearance from what you would expect for normally-distributed data.

5. [GRADUATE SECTION ONLY] Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ under the *Gauss-Markov conditions* and with $\boldsymbol{X}$ having p' columns and full column rank. Consider the leverage values h_{ii} from the hat matrix $\boldsymbol{H}$ and the *random* residuals $\hat{e}_i$ (from ordinary least squares).

(a) [2 pts] Let $\text{tr}(\boldsymbol{M})$ be the *sum* of all diagonal elements of the square matrix $\boldsymbol{M}$, called the *trace* of $\boldsymbol{M}$. If the matrix product $\boldsymbol{AB}$ is square, show that

$$\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$$

(Hint: The ith diagonal element of $\boldsymbol{AB}$ is $\sum_j a_{ij}b_{ji}$ and the jth diagonal element of $\boldsymbol{BA}$ is $\sum_i b_{ji}a_{ij}$.)

(b) [2 pts] Using the previous part, show that

$$\sum_{i=1}^{n} h_{ii} \;=\; \text{tr}(\boldsymbol{H}) \;=\; p'$$

(Hints: Take $\boldsymbol{A} = \boldsymbol{X}$ and $\boldsymbol{B} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. What is the trace of an identity matrix?)

(c) [2 pts] Find an expression for $\text{E}\left(\hat{e}_i^2\right)$ in terms of h_{ii} and σ^2.

(d) [2 pts] Using results from the previous parts, show that the usual error variance estimator $\hat{\sigma}^2$ is unbiased.

Some reminders:

- Unless otherwise stated, all data sets are either automatically available or can be found in either the `alr4` package or the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.