MSAN 601: Linear Regression Analysis
Homework 2

Due: Thursday, September 20th by 9:00 AM

Computational Problems

For each of these problems, use `R`. The datasets considered here are loaded in the `faraway` package. Make sure that you have that package loaded before you begin.

1. (**From Ch. 2 of *Linear Models with R***). The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as a response and the sex, status, income and verbal score as predictors. Present the output.

   (a) Which observation has the largest (positive) residual? Give the case number

   (b) Compute the mean and median of the residuals

   (c) Compute the correlation of the residuals with the fitted values

   (d) Compute the correlation of the residuals with the income

   (e) Plot a scatterplot of the residuals against all of the predictors considered in the model. Comment on what you observe

   (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

2. Based on the same fitted model above, answer the following questions:

   (a) Which variables are statistically significant?

   (b) What interpretation should be given to the coefficient for `sex`?

   (c) Predict the amount that a male with average (based on these data) status, income and verbal score would be gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values (for this data) of status, income, and verbal score. Which CI is wider and why is this result expected?

3. The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a regression model with `lpsa` as the response and the other variables as predictors.

   (a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary?

   (b) Compute and display a 95% joint confidence region for the parameters associated with `age` and `lbph`. Plot the origin in your figure. The location of the origin tells us the outcome of a certain hypothesis test. State the test and its outcome.

(c) Suppose that a new patient with the following values arrives `lcavol` = 1.44692, `lweight` = 3.62301, `age` = 65.000, `lbph` = 0.30010, `svi` = 0, `lcp` = -0.79851, `gleason` = 7.000, `pgg45` = 15.000.
Predict the `lpsa` for this patient and provided a 95% prediction interval.

(d) Repeat (c) for a patient with the same values except that he or she is 20 years old. Explain why the prediction interval is wider.

Conceptual Problems

For the problems below, we will consider fitting a linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{1}$$

to observed data and calculating the ordinary least squares estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$. Let $\widehat{y}_i$ be the ith fitted value.

1. Construct an estimator for $13y_i + 2$ and justify it. Calculate your estimator's mean and variance.

2. Answer the following questions about residuals:

   (a) Let $e_i = y_i - \widehat{y}_i$ be the residual of the ith data point. What is the mean and variance of e_i? If we assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, what is the distribution of e_i?

   (b) If we assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, what is the distribution of $\sum_{i=1}^{n} e_i$?

   (c) Using the properties of correlation and an appropriate assumption for model (1)(please state which one you need), calculate $\text{Corr}(y_i, e_i)$

   (d) Based on the calculations in (a) and (c) above, what do we expect a scatter plot of e_i against y_i to look like if our model is correctly specified?

3. (**The Triangle Problem**) Suppose that you want to measure the three angles β_1, β_2, and β_3 of a triangle shown in Figure 1. Elementary geometry shows that $\beta_1 + \beta_2 = \beta_3$, but we nevertheless decide to measure all three angles, with measurement error. This suggests the following model

$$\begin{aligned} y_1 &= \beta_1 + \epsilon_1 \\ y_2 &= \beta_2 + \epsilon_2 \\ y_3 &= \beta_1 + \beta_2 + \epsilon_3 \end{aligned}$$

   in which ϵ_1, ϵ_2, and ϵ_3 are assumed to be independent with mean 0 and common variance σ^2.

   (a) For the model above, write out / calculate X, $X^T y$, and $(X^T X)^{-1}$

   (b) Using the above calculations, write out the least squares estimators $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\beta}_3$

(c) Calculate the variance of each angle estimate. Note that if we were to simply estimate each β_i by its corresponding measurement y_i, the corresponding variance would be σ^2. How does this compare with your answer here? And what does this suggest about using the least squares approach? (this is kinda cool)

4. Let $\theta = c^T\beta$ where $\beta = (\beta_0, \ldots, \beta_p)$ be a linear combination of the coefficient parameters as discussed in class. And let $\widehat{\theta} = c^T\widehat{\beta}$. From our derivations in class we know that

$$\frac{\widehat{\theta} - \theta}{s\sqrt{c^T(X^TX)^{-1}c}} \sim t_{n-p}, \tag{2}$$

where $X = (\mathbf{1}, \mathbf{x_1}, \ldots, \mathbf{x_p})$ is the design matrix and c is just a vector of numbers specified by the user.

(a) Use the distributional result in (2) to fill in the gaps of my lecture and show that a (1-α)*100% confidence interval for θ is given by

$$\widehat{\theta} \pm t_{n-p,1-\alpha/2}s\sqrt{c^T(X^TX)^{-1}c},$$

where $t_{n-p,1-\alpha/2}$ is a critical value satisftying

$$\Pr(t_{n-p} \leq t_{n-p,1-\alpha/2}) = 1 - \alpha/2.$$

(b) What does the above confidence interval simplify to if $c_1 = 1$ and all other values in the vector are 0? (Note that you can input any values to c, which makes this very flexible!)

(c) Use the distributional result in (2) to derive a level α decision rule that tests the null hypothesis $H_o : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$.