

# 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

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

to observed data and calculating the ordinary least squares estimators  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ . Let  $\hat{y}_i$  be the  $i$ th fitted value.

- Construct an estimator for  $13y_i + 2$  and justify it. Calculate your estimator's mean and variance.
- Answer the following questions about residuals:
  - Let  $e_i = y_i - \hat{y}_i$  be the residual of the  $i$ th data point. What is the mean and variance of  $e_i$ ? If we assume  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ , what is the distribution of  $e_i$ ?
  - If we assume  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ , what is the distribution of  $\sum_{i=1}^n e_i$ ?
  - Using the properties of correlation and an appropriate assumption for model (1) (please state which one you need), calculate  $\text{Corr}(y_i, e_i)$
  - 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?

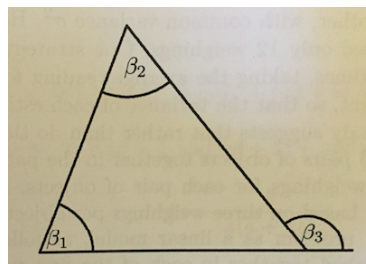


Figure 1: The triangle problem figure for problem 3.

- (The Triangle Problem)** Suppose that you want to measure the three angles  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  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  $\epsilon_1, \epsilon_2$ , and  $\epsilon_3$  are assumed to be independent with mean 0 and common variance  $\sigma^2$ .

- (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  $\hat{\beta}_1, \hat{\beta}_2$ , and  $\hat{\beta}_3$
  - (c) Calculate the variance of each angle estimate. Note that if we were to simply estimate each  $\beta_i$  by its corresponding measurement  $y_i$ , the corresponding variance would be  $\sigma^2$ . 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  $\beta = (\beta_0, \dots, \beta_p)$  be a linear combination of the coefficient parameters as discussed in class. And let  $\hat{\theta} = c^T \hat{\beta}$ . From our derivations in class we know that

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

where  $X = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p)$  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  $(1-\alpha)*100\%$  confidence interval for  $\theta$  is given by

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

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

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

- (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  $\alpha$  decision rule that tests the null hypothesis  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta < \theta_0$ .