MAST09104: A First Course in Statistical Learning

Week 5 Workshop and Lab

Workshop questions

Note that some of these involve R.

1. The following questions are about leverage

   (a) Suppose that H is an idempotent and symmetric matrix. Show that the diagonal entries of H are in $[0, 1]$ and that their sum is the rank of H.

   (b) Interpret part (a) in terms of leverage.

   (c) The formula for the variance of the ith residual is $\text{var}(\varepsilon) = \sigma^2(1 - H_{ii})$. If observation i has a leverage close to 1, how does this show that the response variable, y, for a row of the X matrix with large leverage *can* have a significant effect on the model parameters? When does such an observation have a significant effect?

2. For simple linear regression, $y = \beta_0 + \beta_1 x + \varepsilon$, show that a $100(1 - \alpha)\%$ confidence interval for the mean response when $x = x^*$ can be written as

$$b_0 + b_1 x^* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}}$$

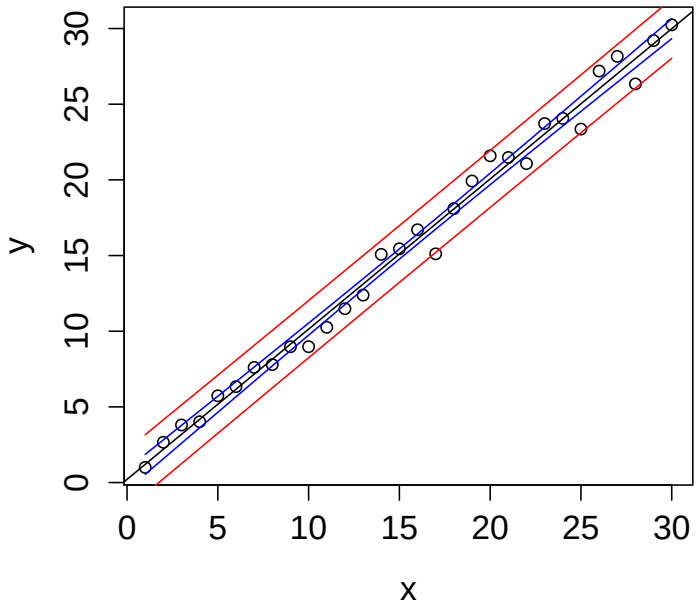
   where $s_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.
   Similarly, show that a $100(1 - \alpha)\%$ prediction interval for a new response when $x = x^*$ can be written as

$$b_0 + b_1 x^* \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}}.$$

3. We can generate some data for a simple linear regression as follows:

```
n <- 30
x <- 1:n
y <- x + rnorm(n)
```

   Construct 95% CI's for $\mathbb{E}y$ and 95% PI's for y, when $x = 1, 2, \ldots, n$. Join them up and plot them on a graph of y against x. Your plot should look something like this:



What proportion of the y's should you expect to lie beyond the outer lines?

4. In this exercise, we look at the dangers of overfitting. Generate some observations from a simple linear regression:

```
set.seed(3)
X <- cbind(rep(1,100), 1:100)
beta <- c(0, 1)
y <- X %*% beta + rnorm(100)
```

Put aside some of the data for testing and some for fitting:

```
Xfit <- X[1:50,]
yfit <- y[1:50]
Xtest <- X[51:100,]
ytest <- y[51:100]
```

(a) Using only the fitting data, estimate $\boldsymbol{\beta}$ and hence the residual sum of squares. Also calculate the residual sum of squares for the test data, that is, $\sum_{i=51}^{100}(y_i - b_0 - b_1 x_i)^2$.
Now add 10 extra predictor variables which we know have nothing to do with the response:

```
X <- cbind(X, matrix(runif(1000), 100, 10))
Xtest <- X[51:100,]
Xfit <- X[1:50,]
```

Again using only the fitting data, fit the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and show that the residual sum of squares has reduced (this has to happen). Then show that the residual sum of squares for the test data has gone up (this happens most of the time).
Explain what is going on.

(b) Repeat the above, but this time add x^2, x^3 and x^4 terms:

```
X <- cbind(X[, 1:2], (1:100)^2, (1:100)^3, (1:100)^4)
```

Lab questions

1. What will be the output of the following code? Try to answer this without typing it up.

```
fb <- function(n) {
    if (n == 1 || n == 2) {
        return(1)
    } else {
        return(fb(n - 1) + fb(n - 2))
    }
}
fb(8)
```

2. Let $A = (a_{i,j})_{i,j=1}^{n}$ be a square matrix, and denote by $A_{(-i,-j)}$ the matrix with row i and column j removed. If A is a 1×1 matrix then $det(A)$, the determinant of A, is just $a_{1,1}$. For $n \times n$ matrices we have, for any i,

$$det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \, det(A_{(-i,-j)}).$$

   Use this to write a *recursive* function to calculate $det(A)$.

3. This question writes a function to calculate leverages and generate histograms for them, and applies it to three different datasets in order to better understand leverage.

   (a) In R, generate 20 observations from a standard bivariate normal distribution with correlation 0.7. (You can do this using the definition and pairs of standard independent normal random variables or using the function `rmvnorm` in the package `mvtnorm`). The three data sets are pairs $(x1, y1), (x2, y2), (x3, y3)$ with the first data set having 20 pairs of (x, y) and the second two having 21. Specifically, the three data sets are:

      i. The 20 bivariate observations, labelling the first of each pair as `x1` in R, and the second as `y1`.

      ii. New variables, `x2`, `y2`, which include the pairs (`x1`,`y1`) together with an additional point $(10, 10)$.

      iii. New variables, `x3`, `y3`, which include pairs in (`x1`,`y1`) together with an additional point $(10, -0.3)$.

   (b) Fit simple linear regression models of the response $y$to the predictor x for each of the three data sets in part (a), storing the results in `model1`, `model2`, `model3`, including the options `x=TRUE` in order that it is possible to access the X matrix of the models subsequently.

   (c) Write a function, `histhat`, with argument, `model`, that uses matrix operations to find the leverage values in model and prints them out with the model formula as heading, as well as generates a histogram of the leverage values. To label the printout and the histogram, the following R commands may be helpful:

      i. `writeLines`

      ii. `noquote` to get text strings without quotes printed out

      iii. `paste` including the text string `"`
`n"` to get a new line before a subsequent `print`

   (d) Generate the printouts of leverages and histogram of leverages for the three models.

   (e) Plot the data in `y1` versus `x1` using the options `xlim` and `ylim` to make sure that the axes go from -2.1 to 11. Add the point $(10, 10)$ with an upper triangle and point $(10, -0.3)$ with a lower triangle. Add the lines from `model2` and `model3` using dashed and dotted lines. Add the line from `model1` in a solid line using the intercept and slope inputs.

4. The data set `ufc.csv` contains forest inventory observations from the University of Idaho Experimental Forest. In the experiment, scientists randomly selected a number of plots and then from each plot selected a number of trees. For each tree they measured its height and diameter (which are numeric), and also the species of tree (which is a character string). Answer the following questions:

   (a) What are the species of the three tallest trees? Of the five fattest trees? (Use the `order` command.)

   (b) What are the mean diameters by species?

   (c) What are the two species that have the largest third quartile diameters?

   (d) What are the two species with the largest median slenderness (height/diameter) ratios? How about the two species with the smallest median slenderness ratios?

   (e) What is the identity of the tallest tree of the species that was the fattest on average?