

B365 Homework 7

1. Consider the Vocab.csv data on Canvas (in the data folder), containing the number of years of education and the performance on a vocabulary test for a number of individuals. For this problem we want to perform simple linear regression by modeling the relationship between the response (the test score) and the predictor (the education level). As with the simple linear regression done in class, case we want to model

$$y_i \approx \alpha x_i + \beta$$

However, we want to do this according to the generic method for regression presented in class in which we solve the normal equations using an appropriate choice of X matrix.

- (a) In R, read in the data and create the X matrix that would be appropriate for estimating α and β .
 - (b) Solve the normal equations with your choice of X and report the values you get for α and β .
 - (c) Does it appear that people with more education tend to have larger vocabularies?
 - (d) It is usually hard to make quantitative statements about the value of a year of education, however, you can do so here in the context of this particular vocabulary test. Make such a quantitative statement here.
2. This problem deals with the ais.csv data (also in the Canvas data folder) describing various measurements taken on a collection of Australian athletes. These data can be read in using the

```
dat = read.csv("ais.csv",stringsAsFactors=FALSE, sep=",")
```

command, unlike previous datasets where we have used “read.csv2”. For each athlete, the data contain a number of numeric variables we will consider, though we will ignore the gender and sport variables. We are interested in trying to predict the red blood cell count (rcc) from the other numeric variables.

- (a) In R, create the X matrix using variables 3 through 12 as the predictors, while creating the response variable from the 2nd variable, rcc. Solve the normal equations and report the values you get for the regression coefficients.
 - (b) Compute and your predicted values of the rcc level, \hat{y} , the errors ($e = y - \hat{y}$), and the give the sum of squared errors (sse) , defined by $\sum_i e_i^2$.
 - (c) In a loop, perform a regression by omitting a single variable from the collection of predictors used above. The first time through the loop you should omit variable 3, the 2nd time through the loop you should include 3 but omit 4, etc. In each case compute your sum of squared errors. Which variable’s omission causes the greatest increase in sse? Which variable appears to be the most important in predicting rcc?
3. This problem uses the Nottingham beer sales data which you can include into your R program with

```
data(nottem)
y = nottem
n = length(y)
x = 1:n;
```

Here y is a vector containing the montly beer sales from the Nottingham company for 20 consecutive years ($n = 240$ months), numbered by the vector x .

- (a) Plot the sales data for the various months using both lines and points using lines and points, as below

```
plot(x,y,type="b")
```

- (b) Construct the X matrix for simple linear regression ($y_i \approx ax_i + b$) and solve for a and b . Plot the resulting line on top of the original plot.

- (c) Of course the linear model has no chance to capture the seasonal pattern of sales. We will this periodic component by including a sine wave and a cosine wave, both n points long, oscillating with 12 points per period (there are 12 months in a year). Specifically, we use the model:

$$y_i \approx a \cos(2\pi x_i/12) + b \sin(2\pi x_i/12) + c$$

Find the values for a, b, c that minimize the sum of squared errors in predicting the $\{y_i\}$. Plot your fitted model on the same plot as the original data, using a different color for each. Be sure to use both lines and points in your plot.

- (d) The company hopes that, in addition to the obvious seasonal pattern, their sales grow over time. Consider the model

$$y_i \approx a \cos(2\pi x_i/12) + b \sin(2\pi x_i/12) + c + dx_i$$

that accounts for a possible linear trend in sales. Fit this model to that data and plot the fitted results as you did in the previous part, showing both original data and fitted model. Do the results suggest that the company is experiencing growth in sales?