STU22005 Laboratory Session 5

Professor Caroline Brophy

Week beginning 29th March 2021

Instructions

- In this lab sheet, there will be a series of notes for you to work through. For each section, you should run the R code yourself on your own computer and verify the output provided. Ask your lab demonstrator any questions you have as you work through the material.
- In the last section of the sheet, there are several questions where you will write your own code. Your lab demonstrator is there to help if you get stuck in this section.
- Create a script for all the code that you write in this lab session. Save the script into the folder you have for the module. Save your code regularly throughout the class. (Do not write your code into the "Console" window as this will not save in an R script.)

Introduction

We are going to fit multiple regression models in our lab today. We will start with the Franchise data that we looked at in class.

The Franchise data set

- Data on 27 "All Greens" franchise branches
- There are multiple variables recorded on each branch
  - net sales (in thousands)
  - shop size, inventory (in thousands)
  - advertising revenue (in thousands)
  - size of sales district (thousands of families)
  - the number of competing stores in the sales district.

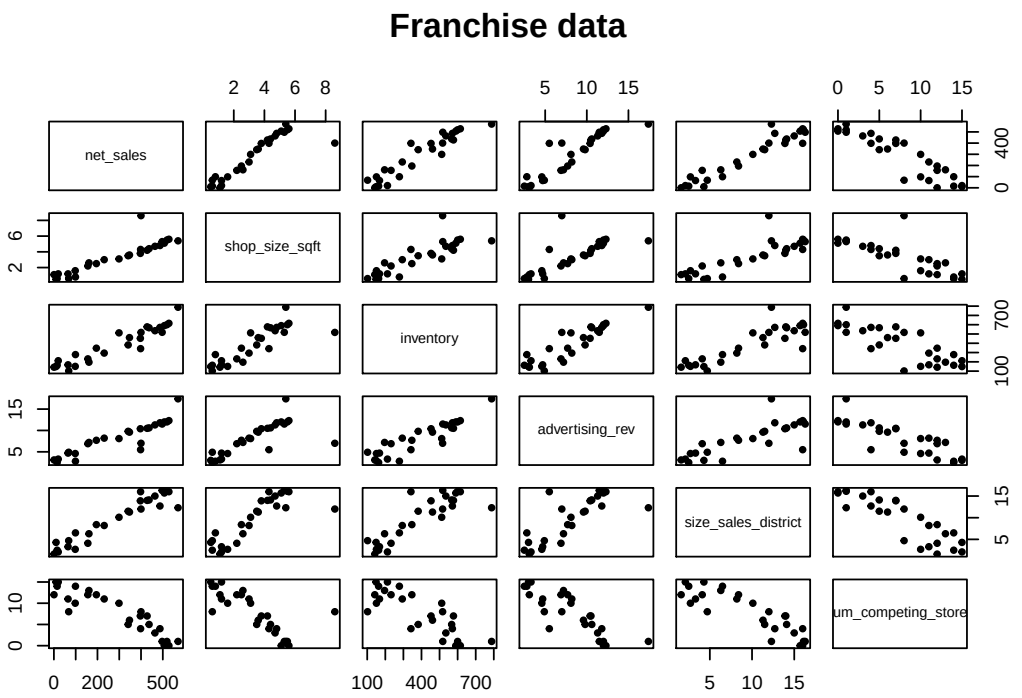
Explore the data

- Read the data into R.
- Generate a matrix plot of the data.

```
# Read in the greens franchise data
greens <- read.csv("Lab5_all_greens_franchise.csv")

# Generate a matrix plot
```

```
pairs(net_sales ~ shop_size_sqft + inventory + advertising_rev
      + size_sales_district + num_competing_stores,
      pch=20, data = greens, main="Franchise data")
```



- Find the correlation coefficient between each pair of variables

```
#Find the correlation coefficients
cor(greens)
```

```
##                       net_sales shop_size_sqft  inventory advertising_rev
## net_sales             1.0000000      0.8940921  0.9455036       0.9140241
## shop_size_sqft        0.8940921      1.0000000  0.8436158       0.7485872
## inventory             0.9455036      0.8436158  1.0000000       0.9062306
## advertising_rev       0.9140241      0.7485872  0.9062306       1.0000000
## size_sales_district   0.9536831      0.8380229  0.8639169       0.7954345
## num_competing_stores -0.9122364     -0.7657378 -0.8073804      -0.8412800
##                      size_sales_district num_competing_stores
## net_sales                      0.9536831          -0.9122364
## shop_size_sqft                 0.8380229          -0.7657378
## inventory                      0.8639169          -0.8073804
## advertising_rev                0.7954345          -0.8412800
## size_sales_district            1.0000000          -0.8695896
## num_competing_stores          -0.8695896           1.0000000
```

Analyse the data

- Fit a multiple regression model, including all predictors in the model.
- Go through each part of the output and interpret parts of interest. In particular, interpret the coefficients and the associated tests.

```r
# Fit a multiple regression model
lmGreens1 <- lm(net_sales ~ shop_size_sqft + inventory + advertising_rev
               + size_sales_district + num_competing_stores, data = greens)
summary(lmGreens1)
```

```
##
## Call:
## lm(formula = net_sales ~ shop_size_sqft + inventory + advertising_rev +
##     size_sales_district + num_competing_stores, data = greens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.338  -9.699  -4.496   4.040  41.139
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -18.85941   30.15023  -0.626 0.538372
## shop_size_sqft        16.20157    3.54444   4.571 0.000166 ***
## inventory              0.17464    0.05761   3.032 0.006347 **
## advertising_rev       11.52627    2.53210   4.552 0.000174 ***
## size_sales_district   13.58031    1.77046   7.671 1.61e-07 ***
## num_competing_stores  -5.31097    1.70543  -3.114 0.005249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.65 on 21 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9916
## F-statistic: 611.6 on 5 and 21 DF,  p-value: < 2.2e-16
```

- Advertising revenue and inventory are very strongly correlated, remove inventory and re-fit the model. Compare the coefficient estimates to the first model.

```r
# Fit a multiple regression model
lmGreens2 <- lm(net_sales ~ shop_size_sqft + advertising_rev
               + size_sales_district + num_competing_stores, data = greens)
summary(lmGreens2)
```

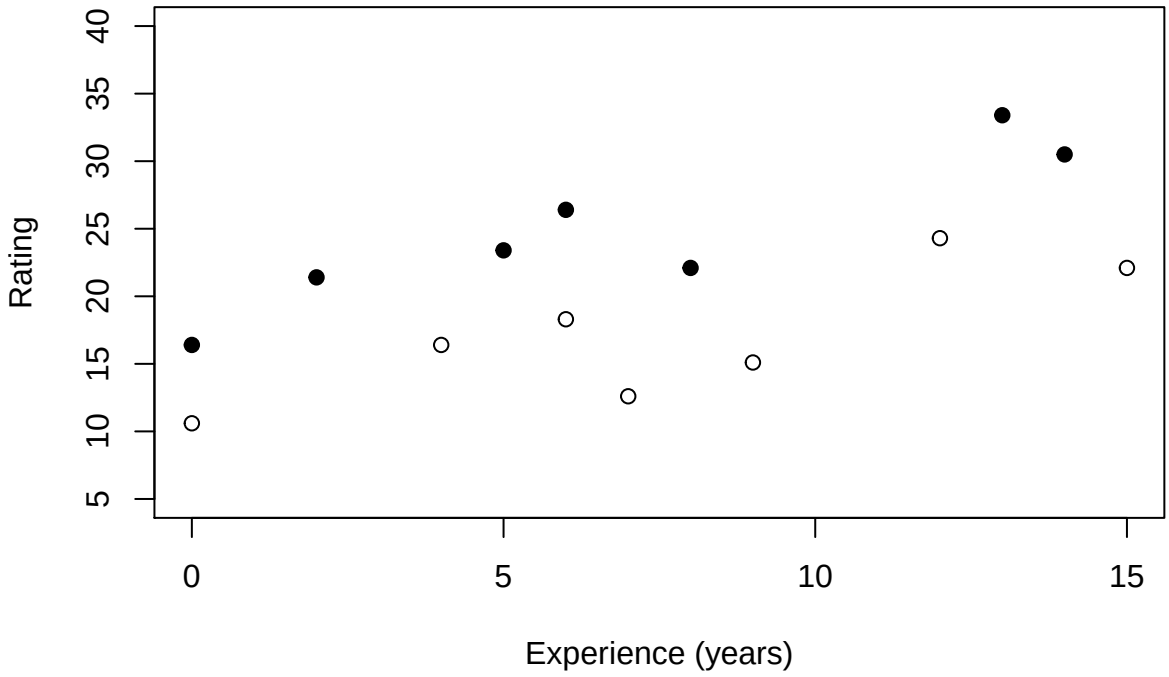
```
##
## Call:
## lm(formula = net_sales ~ shop_size_sqft + advertising_rev + size_sales_district +
##     num_competing_stores, data = greens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.422 -12.858  -6.477  16.160  45.255
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -39.460     34.411  -1.147   0.2638
## shop_size_sqft         20.444      3.815   5.359 2.22e-05 ***
## advertising_rev        16.966      2.093   8.107 4.73e-08 ***
## size_sales_district    15.673      1.910   8.206 3.86e-08 ***
## num_competing_stores   -4.043      1.937  -2.088   0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 20.68 on 22 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9884
## F-statistic: 555.4 on 4 and 22 DF,  p-value: < 2.2e-16
```

Questions

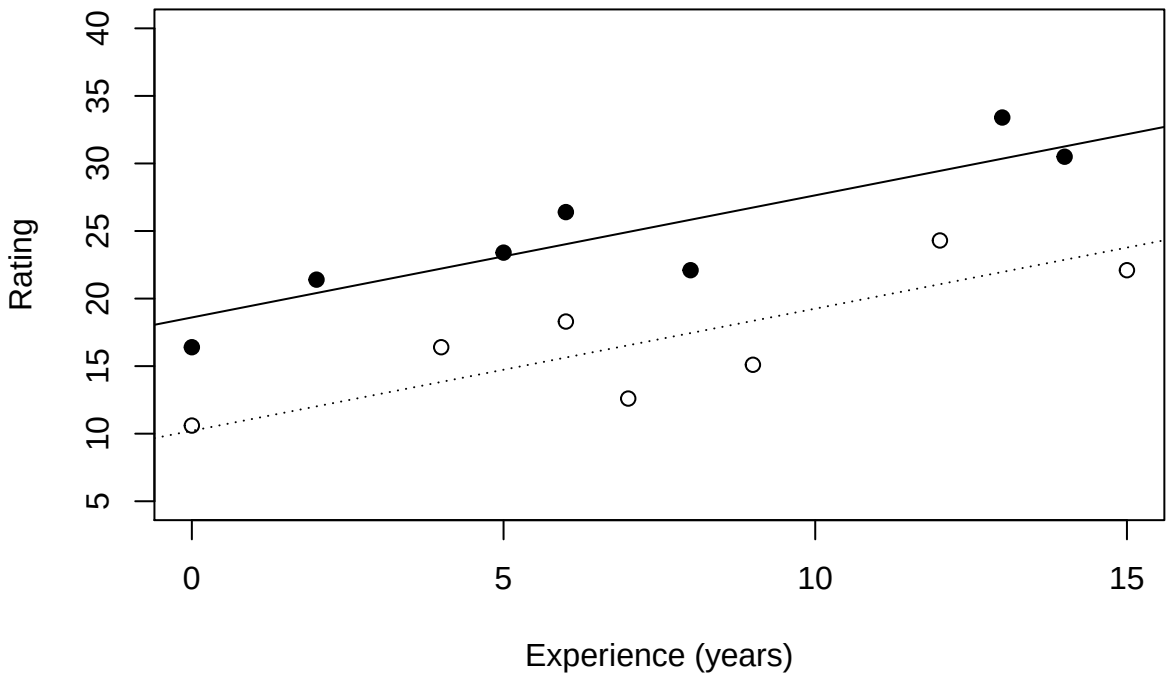
A study was carried out at a firm to assess two training methods for a task. Employees were randomly assigned to undertake the standard training (A), or a new training method (B) for a particular task and after a period of training were rated in the task. The dataset is stored in 'Lab5_training.csv'. The experience of the employee prior to the training was also recorded.

Here is a scatter plot of the data, rating versus experience, with training A points empty circles, and training B points filled circles:



1. Read the data into R.
2. Fit a multiple regression model with the response rating and predictors training method and experience. Note that training is a categorical variable.
3. Write out the algebraic equation of the model.
4. Explain what is going on in the output and what the coefficients are estimating. Are the hypothesis tests useful here?

Essentially we are fitting a simple linear regression model but with two intercepts and all in one go! See the fitted model, where the solid line is for training B and the dotted line for training A:



5. Write out the X matrix for the model.

6. Create a dummy variable coded 1 for when `training = A`, and 0 otherwise and call it `xA`. Create another dummy variable coded 1 for when `training = B`, and 0 otherwise and call it `xB`. Fit two further models replacing `training` with either `xA` and `xB` each time. Examine the output and compare to the earlier model fit.

```
# Create dummy variables
tr$xA <- ifelse(tr$training == 'A', 1, 0)
tr$xB <- ifelse(tr$training == 'B', 1, 0)
```

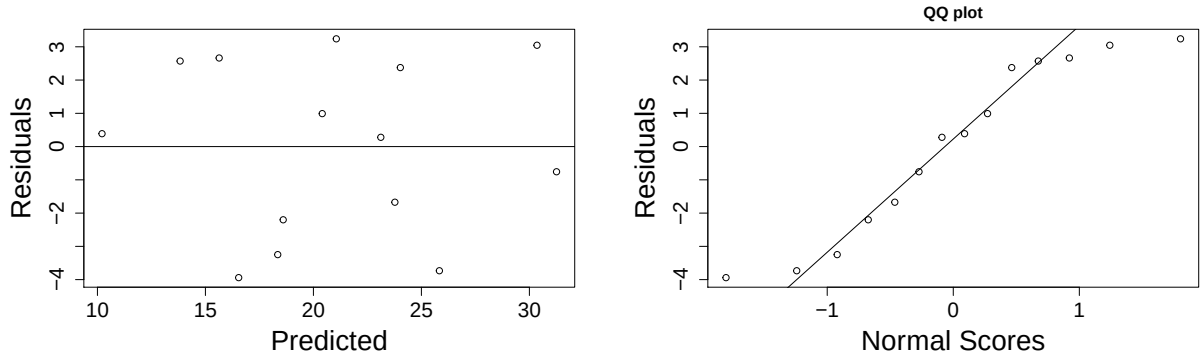
Here are the three sets of coefficients that you should have after doing this question:

```
## (Intercept)   trainingB  experience
##  10.2125989   8.3885688   0.9039964
```

```
## (Intercept)          xA  experience
##  18.6011678  -8.3885688   0.9039964
```

```
## (Intercept)          xB  experience
##  10.2125989   8.3885688   0.9039964
```

7. Here are the diagnostic plots. Discussion: Comment on the model assumptions and whether or not they are reasonable in this case.



8. Discussion: What can you say about the effects of training and experience on the task rating?