Linear Regression 5

Iris Kim (6757827) and Ted Tinker (3223468)

May 14, 2017

Problem 1A

To multiply an $a \times b$ matrix by a $c \times d$ matrix (to produce an $a \times d$ matrix), b must equal c. Since X is of size $n \times 2$ and X^T has size $2 \times n$, this condition is satisfied, and $X^T X$ has size 2×2. Using $x_{i,j}$ to mean the entry of X in the i^{th} row and j^{th} column, $X^T X$ is of the form

$$\begin{bmatrix} \sum_{k=1}^{n} x_{1,k}^T \cdot x_{k,1} & \sum_{k=1}^{n} x_{1,k}^T \cdot x_{k,2} \\ \sum_{k=1}^{n} x_{2,k}^T \cdot x_{k,1} & \sum_{k=1}^{n} x_{2,k}^T \cdot x_{k,2} \end{bmatrix}$$

by the definition of matrix multiplication. As X's first column is entirely 1, $\sum_{k=1}^{n} x_{1,k}^T \cdot x_{k,1} = \sum_{k=1}^{n} 1 = n$. $\sum_{k=1}^{n} x_{1,k}^T \cdot x_{k,2}$ is equivalent to $\sum_{k=1}^{n} x_{2,k}^T \cdot x_{k,1}$, and both equal $\sum_{k=1}^{n} x_{k,2}$. Finally, $\sum_{k=1}^{n} x_{2,k}^T \cdot x_{k,2} = \sum_{k=1}^{n} x_k^2$:

$$\begin{bmatrix} n & \sum_{k=1}^{n} x_{k,2} \\ \sum_{k=1}^{n} x_{k,2} & \sum_{k=1}^{n} x_k^2 \end{bmatrix}$$

Factoring out an n, we find $X^T X$ is equal to

$$n \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \frac{\sum_{k=1}^{n} x_k^2}{n} \end{bmatrix}$$

Part B

The inverse of 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. In this case,

$$\begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{k=1}^{n} x_k^2 \end{bmatrix}^{-1} = \frac{1}{n \sum_{k=1}^{n} x_k^2 - n^2 \overline{x}^2} \begin{bmatrix} \sum_{k=1}^{n} x_k^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix}$$

which is equal to

$$\begin{bmatrix} \frac{\sum_{k=1}^{n} x_k^2}{n \sum_{k=1}^{n} x_k^2 - n^2 \overline{x}^2} & \frac{-n\overline{x}}{n \sum_{k=1}^{n} x_k^2 - n^2 \overline{x}^2} \\ \frac{-n\overline{x}}{n \sum_{k=1}^{n} x_k^2 - n^2 \overline{x}^2} & \frac{n}{n \sum_{k=1}^{n} x_k^2 - n^2 \overline{x}^2} \end{bmatrix}$$

Part C

First, let us find $X^T Y$. X^T has size $2 \times n$ and Y has size $n \times 1$ (having size $X \times \beta + e$), so $X^T Y$ is well-defined and has size 2×1. By the definition of matrix multiplication, its value should be

$$\begin{bmatrix} \sum_{k=1}^{n} x_{1,k}^T \cdot y_k \\ \sum_{k=1}^{n} x_{2,k}^T \cdot y_k \end{bmatrix}$$

$x_{1,k}^T$ is always 1 because X's first column is all 1. So,

$$X^T Y = \begin{bmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k \cdot y_k \end{bmatrix}$$

$(X^TX)^{-1}$ is size 2×2 and X^TY is size 2×1, so their product is well-defined and has size 2×1. Using the answer from part B, we see

$$\begin{bmatrix} \frac{\sum_{k=1}^n x_k^2}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} & \frac{-n\overline{x}}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} \\ \frac{-n\overline{x}}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} & \frac{n}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} \end{bmatrix} \times \begin{bmatrix} \sum_{k=1}^n y_k \\ \sum_{k=1}^n x_k \cdot y_k \end{bmatrix} = \begin{bmatrix} \frac{\sum_{k=1}^n y_k \times \sum_{k=1}^n x_k^2 - n\overline{x} \times \sum_{k-1}^n x_k \cdot y_k}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} \\ \frac{-n\overline{x} \times \sum_{k-1}^n y_k + n\sum_{k=1}^n x_k \cdot y_k}{n\sum_{k=1}^n x_k^2 - n^2\overline{x}^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{n\overline{y}\sum_{k=1}^n x_k^2 - n\overline{x}\sum_{k=1}^n x_k \cdot y_k}{n\sum_{k=1}^n x_k^2 - n\left(\sum_{k=1}^n x_k\right)^2} \\ \frac{-n^2\overline{x}\overline{y} + n\sum_{k=1}^n x_k \cdot y_k}{n\sum_{k=1}^n x_k^2 - n\left(\sum_{k=1}^n x_k\right)^2} \end{bmatrix} = \begin{bmatrix} \frac{\overline{y}\sum_{k=1}^n x_k^2 - \overline{x}\sum_{k=1}^n x_k \cdot y_k}{\sum_{k=1}^n (x_k - \overline{x})^2} \\ \frac{-n\overline{x}\overline{y} + \sum_{k=1}^n x_k \cdot y_k}{\sum_{k=1}^n (x_k - \overline{x})^2} \end{bmatrix} = \begin{bmatrix} \overline{y} - \overline{x}\frac{S_{XY}}{S_{XX}} \\ \frac{S_{XY}}{S_{XX}} \end{bmatrix}$$

Problem 2

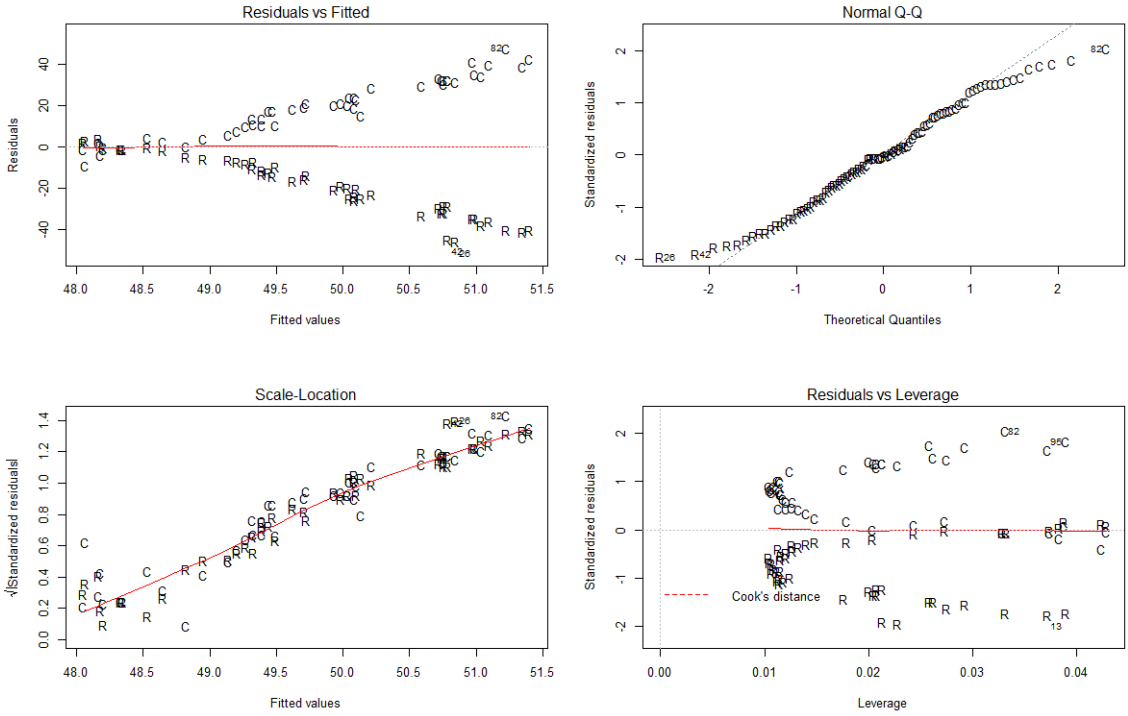
Abstract

In this problem, we develop a regression model for overdue bills from the Quick Stab Collection Agency and evaluate whether how late a bill is paid is correlated to its cost. Each bill is either commercial or residential and has listed the amount overdue and the duration the bill has been overdue. In our analysis, we find that the best models account for whether each bill is commercial or residential.

Method

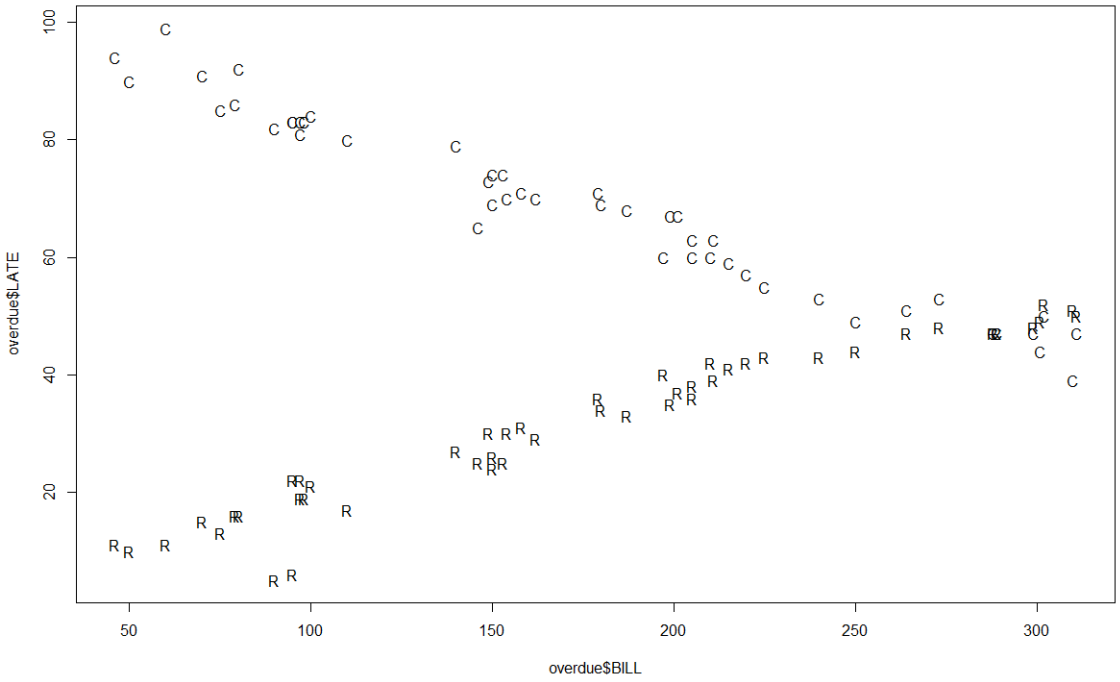
Before loading the data into R-Studio, we manually added the boolean variable R for Residential bills and C for Commercial bills. Then, we make produce diagnostic plots to evaluate whether a linear regression model is appropriate:

```
billLine <- lm(overdue$LATE~overdue$BILL) # Makes a linear regression model of the data
par(mfrow=c(2,2)) # Tells R we want four plots, 2 by 2
plot(billLine,pch=as.character(overdue$TYPE)) # Plots the diagnostics, using R and C accordingly
```



These graphs show several violations of the assumptions of the linear regression model. The errors aren't normally distributed, and the C points and R points appear to group together. When we plot the points, we understand why:

```
plot(overdue$LATE~overdue$BILL,pch=as.character(overdue$TYPE))
    # Makes a scatterplot for the data using R and C
```

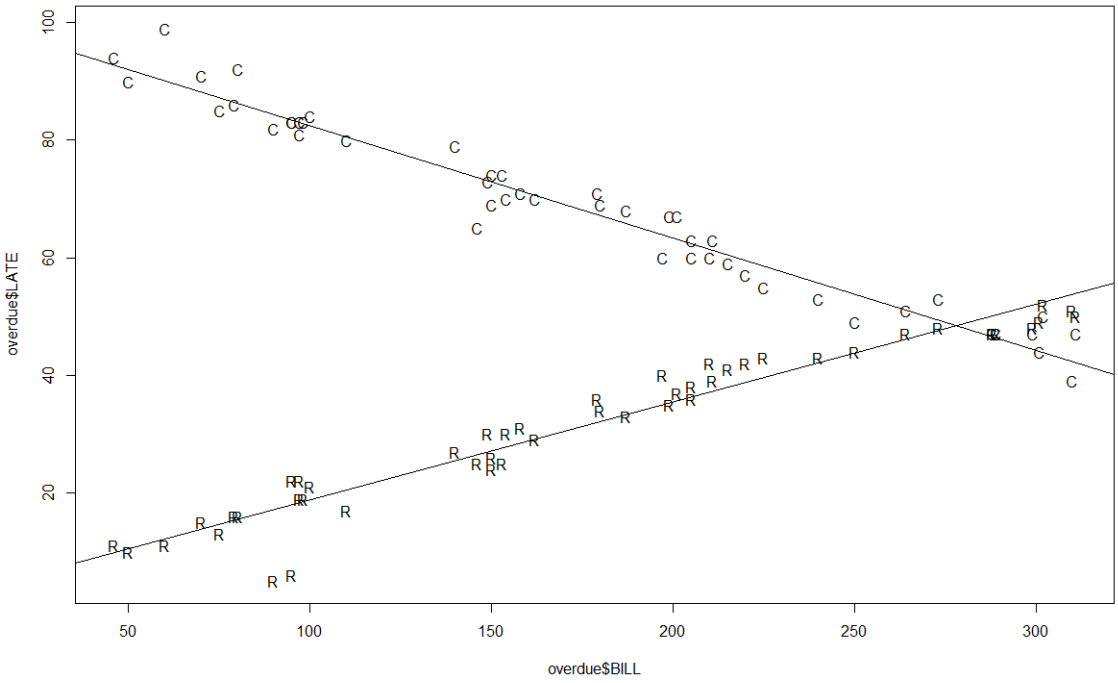


There are two distinct bands of data. Commercial and Residential bills behave differently, and this invalidates a simple linear model. Rather, we must use a dummy variable. Given the position of the bands, we see we should use two unrelated regression lines, as opposed to parallel lines or lines with equal intercepts. We may do this by first adding another line to our data, 0 for R and 1 for C, and with the following code:

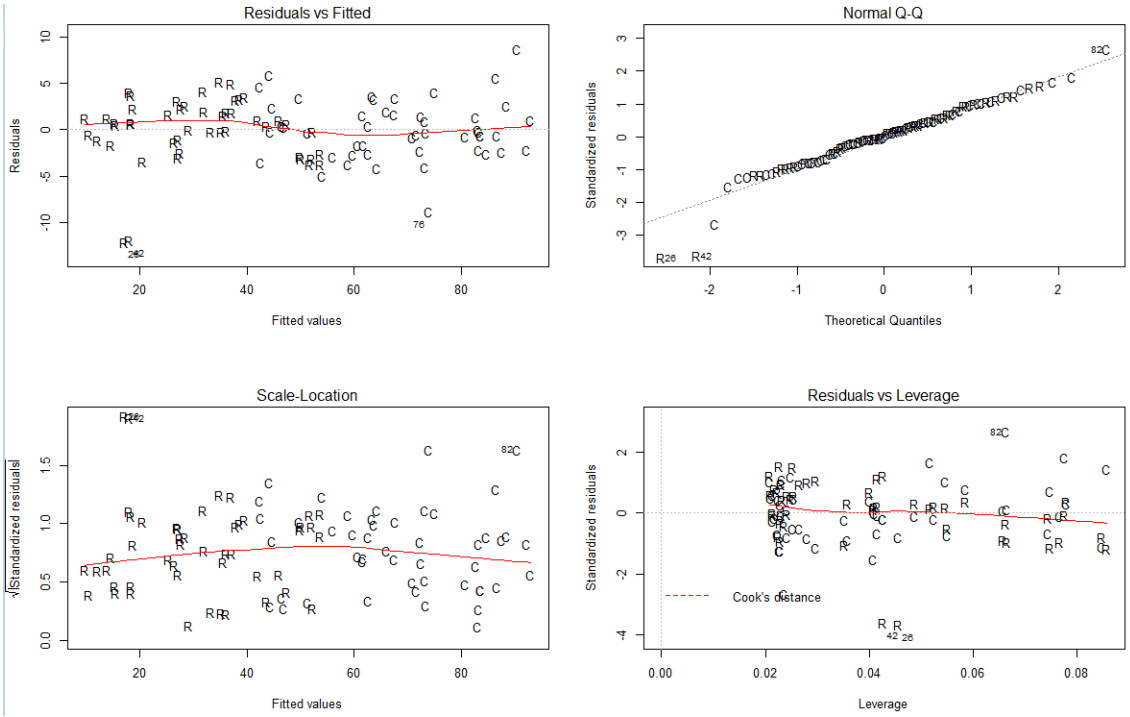
```
newBillLine <- lm(overdue$LATE~overdue$BILL+overdue$B + overdue$B:overdue$BILL)
    # Includes dummy variable in column B
summary(newBillLine) # Asks R for summary of newBillLine

Call:
lm(formula = overdue$LATE ~ overdue$BILL + overdue$B + overdue$B:overdue$BILL)

Residuals:
    Min       1Q   Median       3Q      Max
-12.1211  -2.2163   0.0974   1.9556   8.6995

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.209624   1.198504   1.844   0.0685 .
overdue$BILL              0.165683   0.006285  26.362   <2e-16 ***
overdue$B                99.548561   1.694940  58.733   <2e-16 ***
overdue$BILL:overdue$B   -0.356644   0.008888 -40.125   <2e-16 ***
```

This gives us two regression lines, $LATE = 2.21 + BILL(.166)$ for residential bills and $LATE = 101.56 + BILL(-.191)$ for commercial bills. Plotting these against the scatterplot above seems promising:

```
plot(overdue$LATE~overdue$BILL,pch=as.character(overdue$TYPE)) # Make the scatterplot
abline(2.21,.166) # Add first line
abline(101.56,-.191) # Add second line
```



Using the plot() function again to find the diagnostic plots, they are much more in keeping with the assumptions of a linear model. The errors are well distributed and although there are a few outliers, this should be a appropriate for analysis:



Analysis

In the summary on page 3, in the $Pr(>|t|)$ column, the values for both overdue$BILL and overdue$B are $< 2e-16$, so we confirm that the relationships uncovered by the model are statistically significant at the 95% level. The model predicts that a residential bill will take 2.21 days plus .166 days for each dollar overdue to collect, while a commercial bill will take 99.55 days minus .191 days for each dollar overdue.

Conclusion

This model fits the data quite well and provides estimates for the duration it will require to collect on a bill and the value of the overdue bill. The slogan which Marketing suggested, "Under 60 days or your money back!!", is a safe bet for residential bills under $350, but only for commercial bills above $200.

Problem 3

We load the .txt into R-Studio and use the following code to compare wine quality to its harvesting date (in days after August 31st), using 0 for wines collected without rain and 1 for wines collected with rain:

```
wineLine <- lm(Latour$Quality~Latour$EndofHarvest+Latour$Rain + Latour$Rain:Latour$EndofHarvest)
    # Make wineLine into a linear regression using Rain as a boolean dummy
summary(wineLine) # To plot the regression line, we need entries from the summary
plot(Latour$Quality~Latour$EndofHarvest,pch=as.character(Latour$Rain))
    # Plots the data using 0 for rainless harvests, 1 otherwise
```

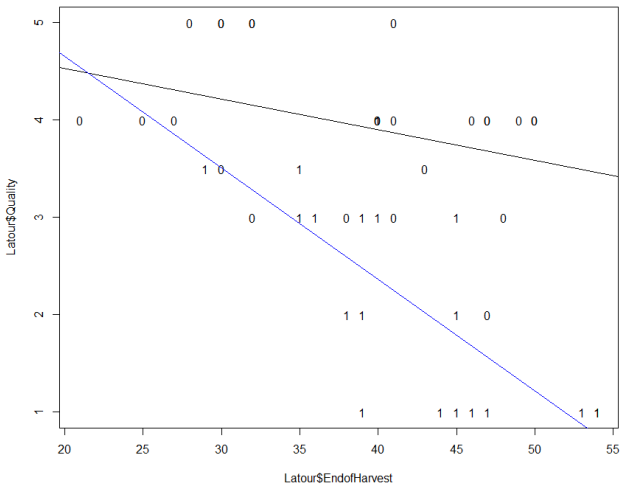
Use these entries of the summary to plot the regression lines:

```
    Coefficients:
                                  Estimate
(Intercept)                        5.16122
Latour$EndofHarvest               -0.03145
Latour$Rain                        1.78670
Latour$EndofHarvest:Latour$Rain   -0.08314
```

So, using

```
abline(5.16122,-0.03145) # Rainless harvest line
abline(5.16122 + 1.78670,-0.03145 -0.08314,col="blue") # Rainy harvest line
```

we address the first bullet point:



For the second bullet point, a comparison of variance between this model and a model which does not account for rain shows that rain has a significant effect on the quality of the wine:

```
wineLine <- lm(Latour$Quality~Latour$EndofHarvest+Latour$Rain + Latour$Rain:Latour$EndofHarvest)
    # Regression model including rain
badwineLine <- lm(Latour$Quality~Latour$EndofHarvest)
    # Regression model without rain
anova(badwineLine,wineLine) # Compares variance between the two models

Analysis of Variance Table

Model 1: Latour$Quality ~ Latour$EndofHarvest
Model 2: Latour$Quality ~ Latour$EndofHarvest + Latour$Rain + Latour$Rain:Latour$EndofHarvest
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     42 53.587
2     40 22.970  2    30.616 26.657 4.388e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because the residual sum of squares (RSS) is 53.587 for the simpler model and 22.970 for the more complex model, it is clear the more complicated model reflects the data with greater accuracy. This is statistically significant, because the F-statistic is high enough for the $Pr(> F)$ value to be $4.388e^{-8} < .05$, which means we may reject the null hypothesis (model 1) in favor of the alternative (model 2, accounting for rain).

Finally, using the slopes for the regression lines we found ($-.03145$ for rainless harvests, -0.11459 for rainy harvests), we expect the quality of a wine to decrease by 1 point for each $\frac{1}{.03145} \approx 32$ days if it's *not* raining, or $\frac{1}{.11459} \approx 9$ days if it is raining.