HW 2

Elizabeth Hutton

September 6, 2018

1 3.10 Applied

a. A multiple regression model for *Sales* by *Price*, *Urban*, and *US*

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,        Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b. The negative coefficients for *Price* and *Urban* indicate that an increase in price or having an urban location tends decrease sales. Meanwhile, the larger, positive coefficient for *US* indicates that US locations make more in sales.

c. The model equation can be written as

$$Sales = -0.054Price + -0.022Urban + 1.2US + 13.04$$

Where 1 = yes and 0 = no for the qualitative variables *Urban* and *US*.

d. Both *Price* and *US* are significant predictors so we can reject the null hypothesis that their coefficients = 0.

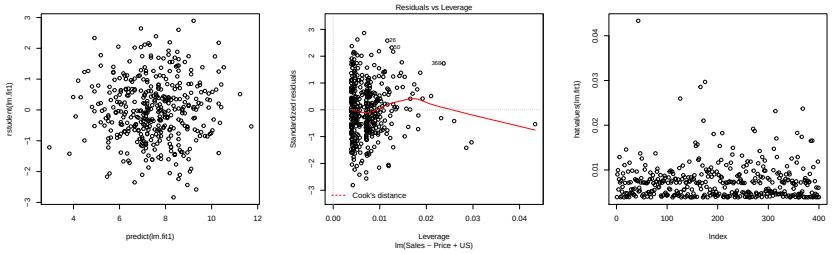
e. A smaller model using only *Price* and *US* as predictors for *Sales*:

```
Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,        Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f. Both models fit the data quite well, as both have p-values $<$ 2.2e-16 and the coefficients do not change much.

g. 95% confidence intervals for the coefficients in model (e)

```
> confint(lm.fit1,level=0.95)

                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

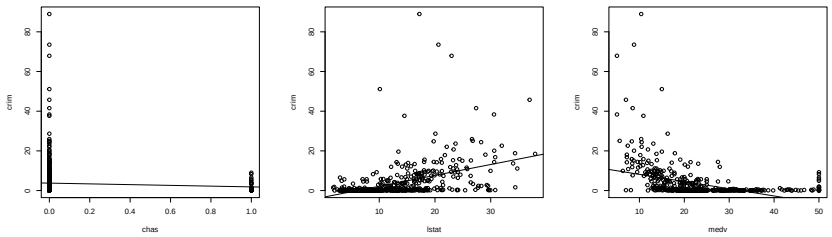


(a) Studentized Residuals vs. Fitted Values (b) Residuals vs. Leverage (c) Hat Values vs. Observation

Figure 1: Outliers and Leverage

h. From looking at the plots of residuals and rstudents, there do not appear to be any outliers as all observations lie within an rstudent value range of [-3,3]. However, the plots of leverage and hatvalues show a high leverage point in observation 43.

2 3.15 Applied



(a) Charles River (b) Low Status Pop. (c) Median Home Value

Figure 2: Crime Rate vs. Predictors

a. Plots of crime rate vs. individual predictors indicate that there is a significant relationship to features like *lstat* (percent low status population) and *medv* (median home value). However, none of the predictors have a clear linear relationship to crime rate, so the simple linear regression models don't fit the data well. Additionally, some predictors are categorical or qualitative (e.g. *chas* which is a boolean indicating whether the area is next to the Charles river), so regression models fail to capture the trend.

b. Aside from one outlying predictor, the coefficients of the predictors from the uni- and multi- variate models are very similar.

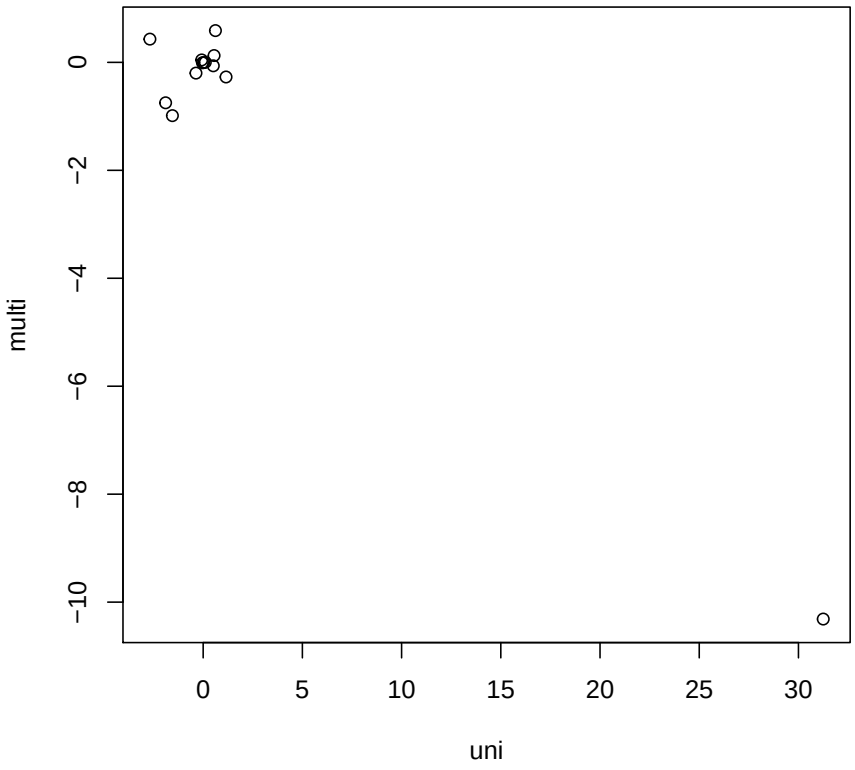


Figure 3: Comparison of Univariate and Multivariate Coefficients

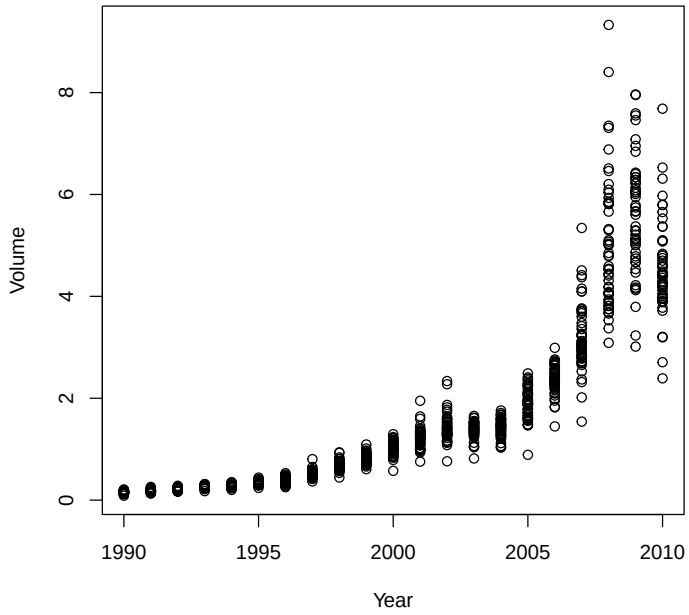
c. Incomplete

3 4.10 Applied

a. From the correlation matrix of predictors, there only seems to be a large positive correlation between *Year* and *Volume*, which has a value of 0.842. Plotting *Year* vs. *Volume* one can clearly see the increasing trend.

```
              Year         Lag1        Lag2        Lag3         Lag4
Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
```

```
Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
               Lag5      Volume        Today
Year   -0.030519101  0.84194162 -0.032459894
Lag1   -0.008183096 -0.06495131 -0.075031842
Lag2   -0.072499482 -0.08551314  0.059166717
Lag3    0.060657175 -0.06928771 -0.071243639
Lag4   -0.075675027 -0.06107462 -0.007825873
Lag5    1.000000000 -0.05851741  0.011012698
Volume -0.058517414  1.00000000 -0.033077783
Today   0.011012698 -0.03307778  1.000000000
```



b. The only significant predictor from the logistic regression model is *Lag2* with a p-value of 0.0296.

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Weekly)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.6949  -1.2565   0.9913   1.0849   1.4579

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1496.2  on 1088  degrees of freedom
Residual deviance: 1486.4  on 1082  degrees of freedom
AIC: 1500.4

Number of Fisher Scoring iterations: 4
```

c. The model severely overestimates the number of days when the market has gone "Up" vs. "Down," predicting "Up" 90% of the time when it is only "Up" 55% of the time in the data. The overall ratio of correct predictions is 0.56. Since the model is only correct about half the time and there is about an equal number of "Up" and "Down" days in the data, the model is performaing at chance. However, based on the chosen cutoff of 0.5, the model is biased to predict "Up" most of the time, as seen in the confusion matrix.

```
        Direction
glm.pred Down  Up
    Down   54  48
    Up    430 557
```

d. The model *Direction Lag*2 makes correct predictions on the test set (years 2008-2010) only 55% of the time. The confusion matrix for the test data can be found below.

```
         Direction.test
glm.pred2 Down Up
     Down    7  5
     Up     65 79
```

e. LDA produces nearly identical performance to logistic regression, correctly predicting test data with 55% accuracy and overestimating the number of days the market goes "Up."

```
         Direction.test
lda.class Down Up
     Down    6  5
     Up     66 79
```

```
[1] 0.5448718
```

f. QDA performs even more poorly than LDA or logistic regression. It correctly predicts only 54% of the test data and *never* predicts "Down."

```
         Direction.test
qda.class Down Up
     Down    0  0
     Up     72 84
```

```
[1] 0.5384615
```

g. Using KNN with $K = 1$ correctly predicts the test data 50% of the time, and has a more even distribution of "Up" vs. "Down" predictions (see confusion matrix).

```
        Direction.test
knn.pred Down Up
    Down   32 38
    Up     40 46
```

h. Even though the correct prediction rate is smaller, the KNN model predicts "Up" almost as often as "Down" which is more consistent with the data.

i. Incomplete.