STAT 239 Homework 1

Tarek Tohme

March 8th, 2018

1 Studying the effects of different car attributes on retail price

In this part we perform linear regression on the cardata.txt dataset to find out how the car's features affect its retail price. The features of interest were the size of the engine, number of cylinders, horsepower, highway MPG, weight, wheelbase and whether the car is hybrid or not.

Question 1 Model 1 is a linear fit of the variables against the suggested retail price. The summary information is shown in figure 7.

```
Call:
lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
    Horsepower + HighwayMPG + Weight + WheelBase + factor(Hybrid),
    data = cardata)

Residuals:
   Min     1Q Median     3Q    Max
-17436  -4134    173   3561  46392

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -68965.793  16180.381  -4.262 2.97e-05 ***
EngineSize       -6957.457   1600.137  -4.348 2.08e-05 ***
Cylinders         3564.755    969.633   3.676 0.000296 ***
Horsepower         179.702     16.411  10.950  < 2e-16 ***
HighwayMPG         637.939    202.724   3.147 0.001873 **
Weight              11.911      2.658   4.481 1.18e-05 ***
WheelBase           47.607    178.070   0.267 0.789444
factor(Hybrid)1    431.759   6092.087   0.071 0.943562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7533 on 226 degrees of freedom
Multiple R-squared:  0.7819, Adjusted R-squared:  0.7751
F-statistic: 115.7 on 7 and 226 DF,  p-value: < 2.2e-16
```

Figure 1: Summary of model 1

At first glance it seems that WheelBase and Hybrid aren't important predictors because their p-values are 0.7894 and 0.9436 respectively. This will be confirmed later on in the added variable plots. As we can see in the pairs plot in figure 2, some variables such as HighwayMPG need to be transformed to fit a linear relationship with the response. Figure 3 suggests that the error term isn't constant, since there is a change in the residual's spread over the predicted values. This means that the linear model may not be well suited for the data as it is. From the studentized residuals plot in figure 4, we see that points 222, 229 and 223 are above 4 standard deviations away from the predicted value. However the hat values plot in figure 4 tells us that only point 222 is significantly influent on the response, so we might need to remove it. We'll transform the data first and see if that's still the case.

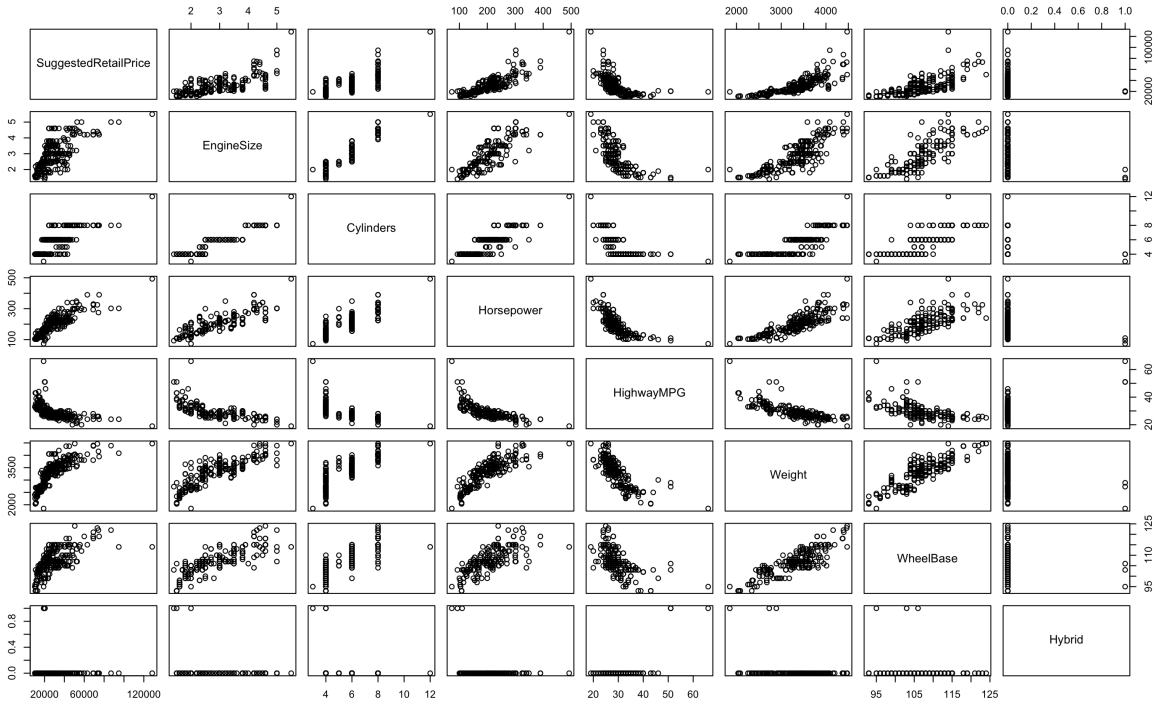

Figure 2: Pairs plot for model 1

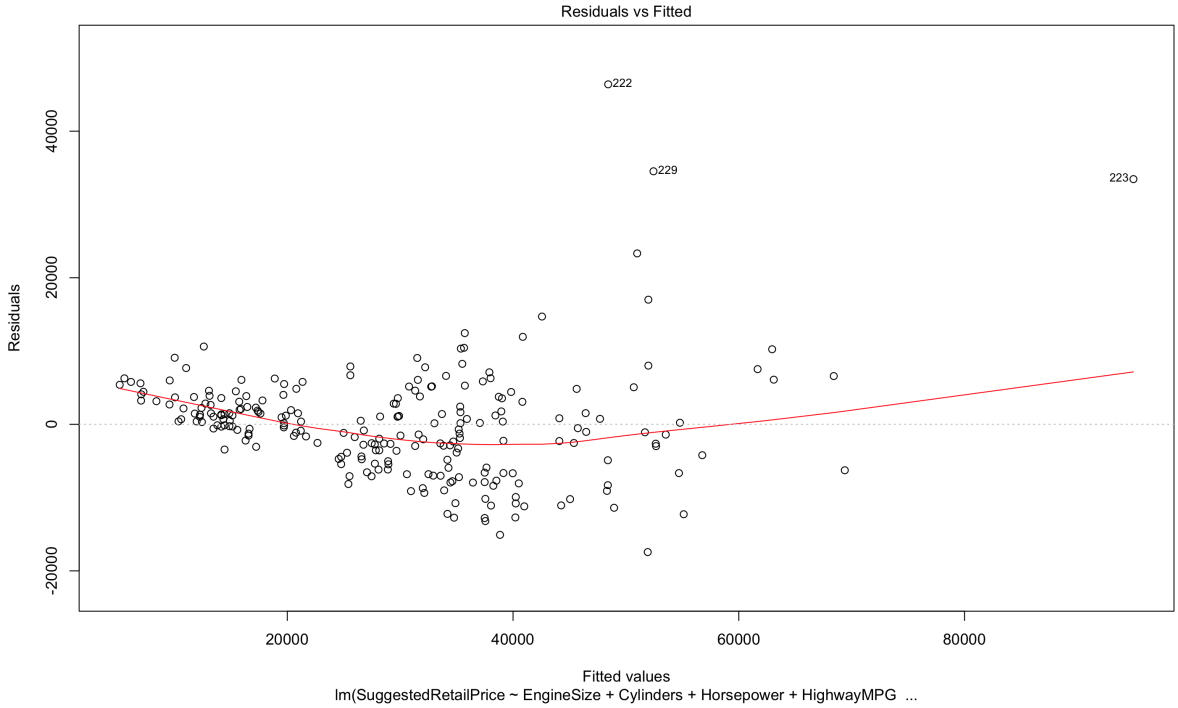

Figure 3: Residual plot for model 1

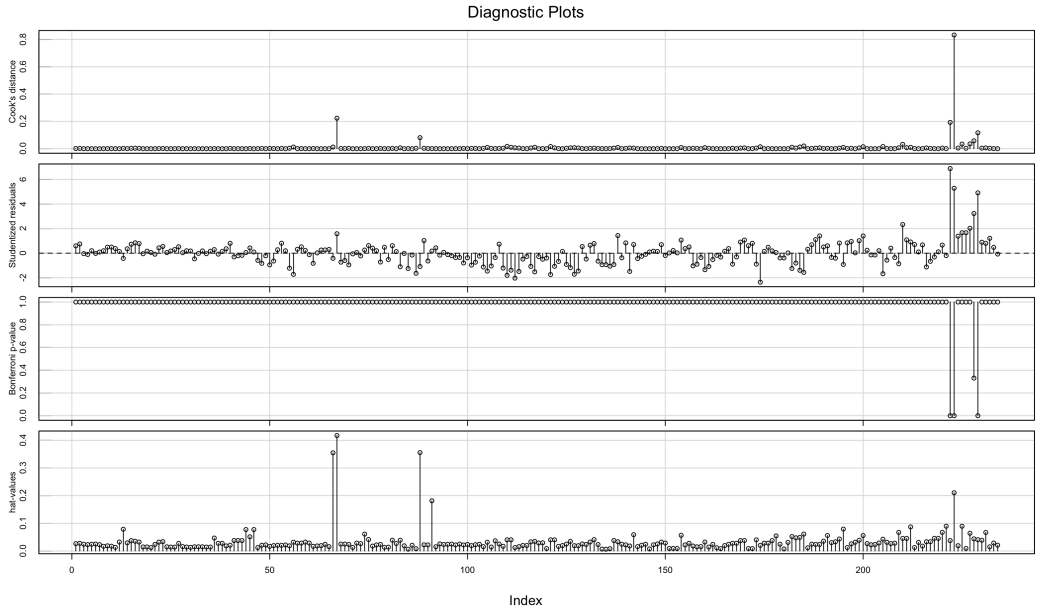

Figure 4: Influence/Index plot for model 1

The preceding figures strongly suggest that the model could be improved by transforming the variables. A power transformation of the model is performed using the powerTransform function. The transformed variables of model 2 are shown in figure 5. After the power transformation, a new leverage analysis (figure 6) tells us that point 67 is an outlier and is very influential, and point 222 (which hasn't been removed) is no longer an outlier.

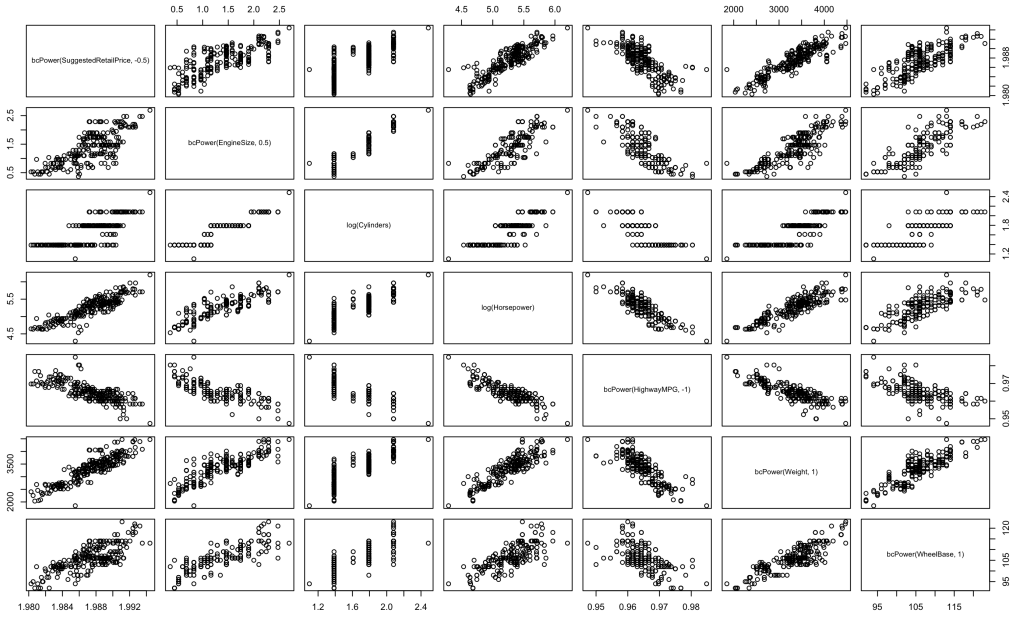

Figure 5: Pairs plot for model 2

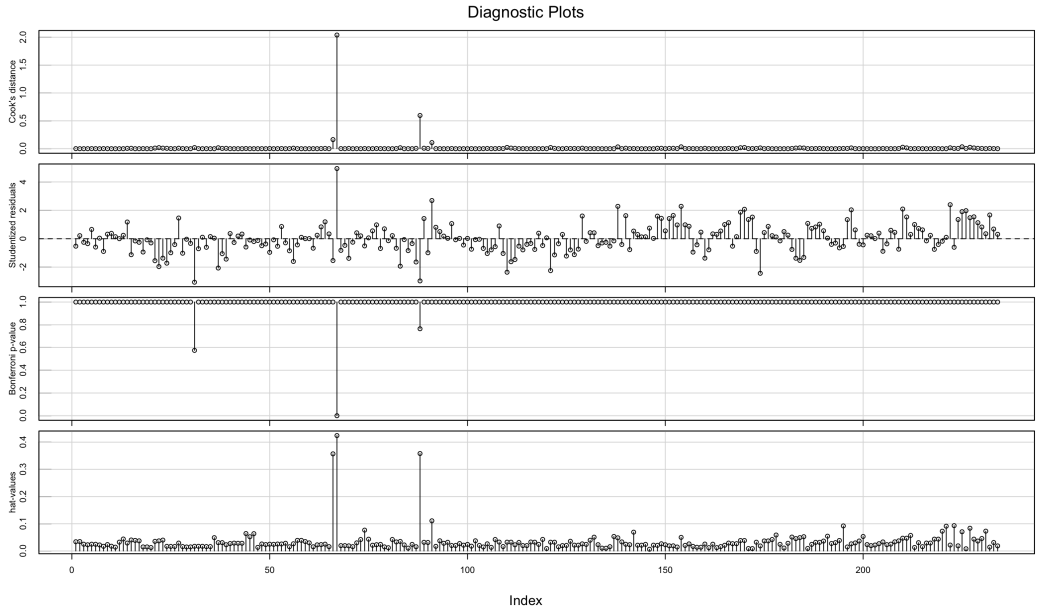

Figure 6: Influence/Index plot for model 2

The entry in cardata.txt at index 67 is a hybrid Honda with 3 cylinders, 73 horsepower, 66 highway miles per gallon (all of which are extrema in the dataset) and an engine size of 2.0. This is a very unusual combination of values: with one other exception, all other cars have horsepower values greater than 100, including those with engines of size 1.4. We can therefore assume that this car is a very unlikely occurence, and removing it from the model will not hinder its predictive or descriptive power.

After removing the outlier, a new power transformation is carried out (model 3). Weight, EngineSize and Horsepower have the lowest p-values, hence they are the most trustworthy predictors. Their regression coefficients tell us that a unit increase in each of log(EngineSize), log(Horsepower) and Weight causes an increase of -3.61e-03, 5.54e-03 and 3.72e-06 in bcPower(SuggestedRetailPrice, -0.5), respectively. The coefficient of Cylinders, with a p-value of 0.23, has a sizeable chance of being zero, and is already an order of magnitude smaller than the rest of the predictors, which hints that the number of cylinders isn't a good predictor for price. Weight, on the other hand, has a p-value of nearly zero, so the coefficient is very reliable, but still quite low, so not very indicative of a relation with price. Interestingly, the coefficient for EngineSize is negative, which means that it contributes negatively to the price, given all other predictors are fixed.

Figure 7 shows the comparison of the summaries of each model. We can see that the p-values of Wheelbase and HighwayMPG in model 3 are very high (0.841 and 0.866 respectively). The added variable plots in figure 8 also give strong evidence that these two variables aren't contributing to the retail price. In the fourth model we remove these two predictors and redo the power transformation. The F-test result for model 3 shown in figure 7 is significantly higher than that of model 2. We can therefore drop the predictors WheelBase and HighwayMPG.

Question 2 We could introduce a term in the model that accounts for categorical variables using the factor() function. Model 5 introduces such a term, and it was observed that some brands have larger p-values than others, which means that they are more likely to affect the price (given their beta coefficient is high enough in absolute value) than the rest. For example, Chevrolet, Dodge, Chrysler, Ford and Honda cars all have prices in the $18,000 to $25,000 range, and similar weights and horsepower (2.7 to 3.5 tons and 150 to 200 Hp respectively). They all have very negative coefficients, and very low p-values. On the other

hand, Saab, Volvo and BMW cars have prices roughly in the $30,000 to $60,000 range, similar weights and horsepower, and they have positive coefficients and acceptably low p-values. This model is therefore able to relate car brands to prices. We notice however these p-values are higher than those of the cheaper cars. This indicates that the brand correlates with the price much better with cheap cars than with expensive cars, given similar attributes.

```
Call:
lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
    Horsepower + HighwayMPG + Weight + WheelBase + factor(Hybrid),
    data = cardata)

Residuals:
   Min     1Q Median     3Q    Max
-17436  -4134    173   3561  46392

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -68965.793  16180.381  -4.262 2.97e-05 ***
EngineSize       -6957.457   1600.137  -4.348 2.08e-05 ***
Cylinders         3564.755    969.633   3.676 0.000296 ***
Horsepower         179.702     16.411  10.950  < 2e-16 ***
HighwayMPG         637.939    202.724   3.147 0.001873 **
Weight              11.911      2.658   4.481 1.18e-05 ***
WheelBase           47.607    178.070   0.267 0.789444
factor(Hybrid)1    431.759   6092.087   0.071 0.943562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7533 on 226 degrees of freedom
Multiple R-squared:  0.7819, Adjusted R-squared:  0.7751
F-statistic: 115.7 on 7 and 226 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = bcPower(SuggestedRetailPrice, -0.5) ~ log(EngineSize) +
    log(Cylinders) + log(Horsepower) + bcPower(HighwayMPG, -1) +
    bcPower(Weight, 1) + bcPower(WheelBase, 1) + factor(Hybrid),
    data = cardata)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0030682 -0.0006115 -0.0000166  0.0005617  0.0036740

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.954e+00  2.733e-02  71.513  < 2e-16 ***
log(EngineSize)      -2.519e-03  6.097e-04  -4.131 5.08e-05 ***
log(Cylinders)        1.236e-04  6.761e-04   0.183    0.855
log(Horsepower)       5.484e-03  5.109e-04  10.734  < 2e-16 ***
bcPower(HighwayMPG, -1) -3.953e-03  2.747e-02  -0.144    0.886
bcPower(Weight, 1)    3.547e-06  3.659e-07   9.693  < 2e-16 ***
bcPower(WheelBase, 1)   -1.174e-05  2.524e-05  -0.465    0.642
factor(Hybrid)1       4.176e-03  6.619e-04   6.308 1.47e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001029 on 226 degrees of freedom
Multiple R-squared:  0.8804, Adjusted R-squared:  0.8767
F-statistic: 237.6 on 7 and 226 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = bcPower(SuggestedRetailPrice, -0.5) ~ log(EngineSize) +
    log(Cylinders) + log(Horsepower) + bcPower(HighwayMPG, -1) +
    Weight + log(WheelBase) + factor(Hybrid), data = cardata,
    subset = -67)

Residuals:
       Min         1Q     Median         3Q        Max
-2.931e-03 -6.078e-04 -1.100e-07  5.312e-04  2.586e-03

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.950e+00  2.540e-02  76.772  < 2e-16 ***
log(EngineSize)      -3.616e-03  6.233e-04  -5.801 2.22e-08 ***
log(Cylinders)        7.896e-04  6.575e-04   1.201   0.2311
log(Horsepower)       5.535e-03  4.867e-04  11.374  < 2e-16 ***
bcPower(HighwayMPG, -1) -4.400e-03  2.606e-02  -0.169   0.8661
Weight                3.719e-06  3.503e-07  10.614  < 2e-16 ***
log(WheelBase)        5.242e-04  2.612e-03   0.201   0.8411
factor(Hybrid)1       1.993e-03  7.693e-04   2.590   0.0102 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0009793 on 225 degrees of freedom
Multiple R-squared:  0.8919, Adjusted R-squared:  0.8886
F-statistic: 265.3 on 7 and 225 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = bcPower(SuggestedRetailPrice, -0.33) ~ bcPower(EngineSize,
    0.5) + log(Cylinders) + bcPower(Horsepower, 0.33) + bcPower(Weight,
    1.42) + factor(Hybrid), data = cardata, subset = -67)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0147603 -0.0033634 -0.0001861  0.0029666  0.0173879

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.804e+00  5.075e-03 552.498  < 2e-16 ***
bcPower(EngineSize, 0.5)  -1.213e-02  1.979e-03  -6.130 3.85e-09 ***
log(Cylinders)            5.565e-03  3.854e-03   1.444   0.1501
bcPower(Horsepower, 0.33)  5.480e-03  4.407e-04  12.434  < 2e-16 ***
bcPower(Weight, 1.42)      6.935e-07  5.379e-08  12.893  < 2e-16 ***
factor(Hybrid)1           1.055e-02  4.079e-03   2.587   0.0103 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005542 on 227 degrees of freedom
Multiple R-squared:  0.8884, Adjusted R-squared:  0.886
F-statistic: 361.6 on 5 and 227 DF,  p-value: < 2.2e-16
```

Figure 7: Summaries of all models. Clockwise from top left: Model 1, Model 2, Model 4, Model 3

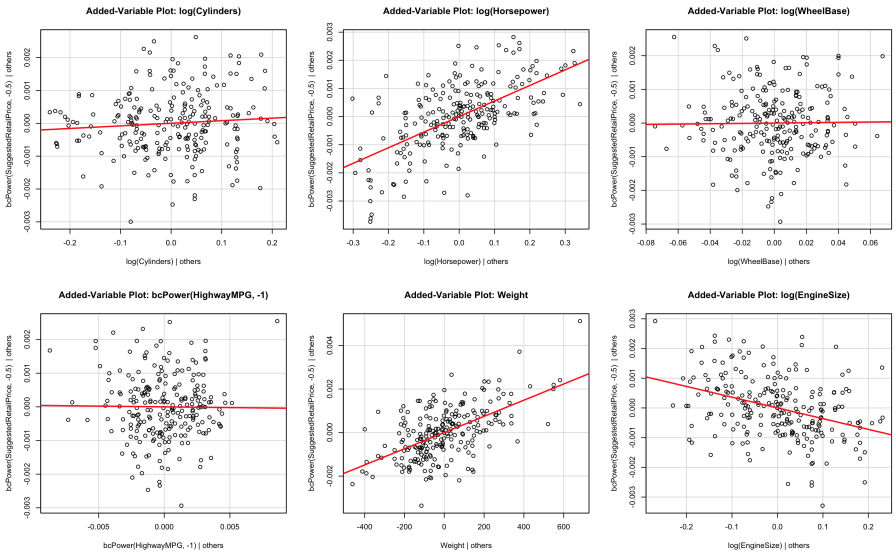

Figure 8: Added variable plots for model 3

2 Studying the effects of restaurant attributes on dinner prices

Question 1 We start by making a linear model of the data. Model 1 contains the variables Price, Food, Decor, Service and East. A pairs plot gives us a general idea of the relationship between the variables.

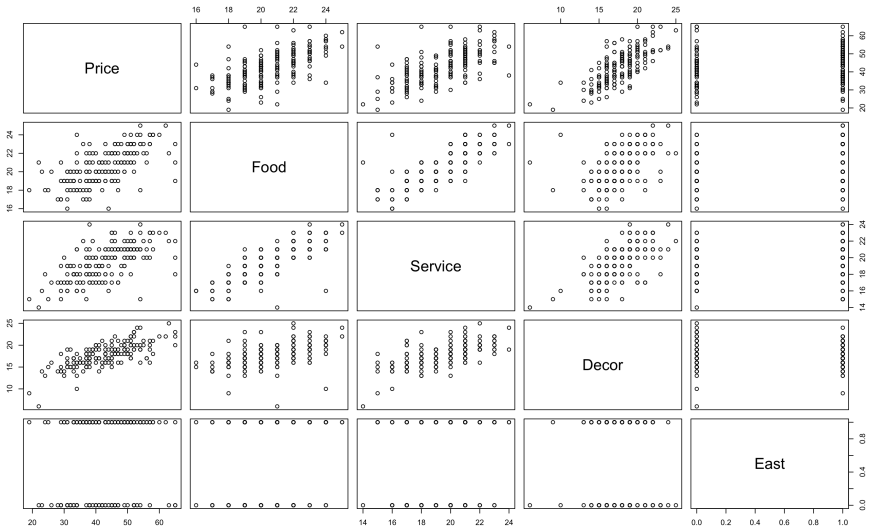

Figure 9: Pairs plot for model 1

It seems that a linear model would fit the data pretty well. A powerTransform() command yields coefficients of 1 for all predictors, which confirms our statement. An influenceIndexPlot() command (figure 10) shows us that points number 56, 30 and 130 are outliers, but they all have very little leverage, so we keep them.

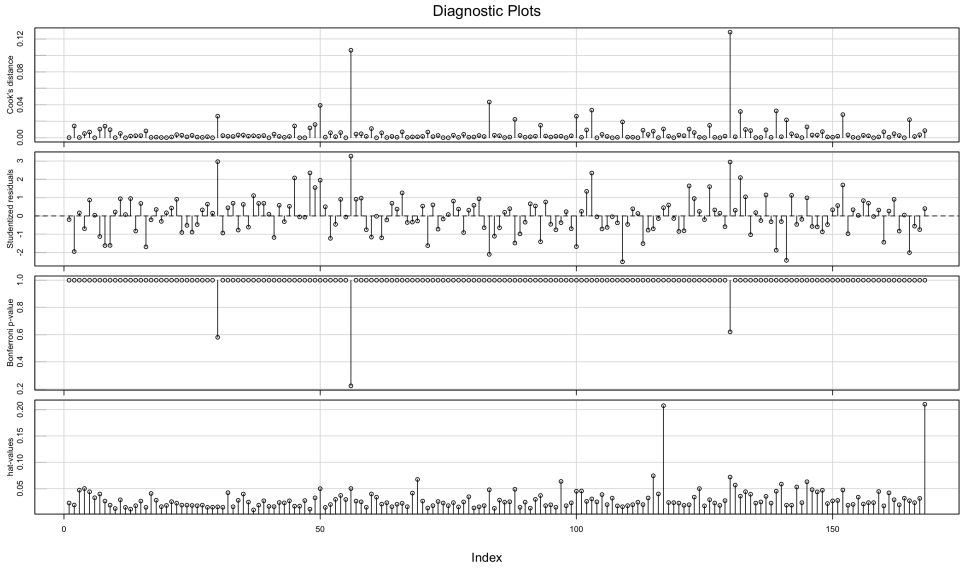

Figure 10: Influence / Index plot for model 1

The summary() command shows us that Service has a very high p-value (0.99) and so can safely be dropped from the model. The coefficients of the resulting model (model 2) are shown in figure 11. Predictably, we see that the restaurant price is highly correlated with the customer ratings for Food and Decor, and a little less so with the location.

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.0451  -3.8809   0.0389   3.3918  17.7557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,     Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

Figure 11: Summary for model 2

Question 2 Adding interaction terms Food_East, Decor_East and Service_East, we see that the model doesn't improve much: the F-value decreases from 92 to 40, the new variables have large p-values and their added variable plots are very spread out. Service is still a poor predictor, with a p-value of 0.88 and a beta estimate of -0.05. We can safely say that East doesn't interact much with other predictors, in other words, being on the East side of New York doesn't necessarily increase the effect of the Food, Decor or Service rating on the price. Therefore, there is little evidence for the need to make two separate models for East and West.

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.5099  -3.7996  -0.1413   3.6522  17.1656

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -24.32952    4.71898  -5.156 7.37e-07 ***
Food           1.76881    0.39172   4.516 1.22e-05 ***
Decor          1.73035    0.24459   7.074 4.43e-11 ***
East           1.90101    0.95434   1.992   0.0481 *
Service       -0.05871    0.39630  -0.148   0.8824
Food_East      1.20769    0.77427   1.560   0.1208
Decor_East    -0.25001    0.45701  -0.547   0.5851
Service_East  -1.27194    0.81706  -1.557   0.1215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.713 on 160 degrees of freedom
Multiple R-squared:  0.6379, Adjusted R-squared:  0.622
F-statistic: 40.27 on 7 and 160 DF,  p-value: < 2.2e-16
```

Figure 12: Summary for model 3

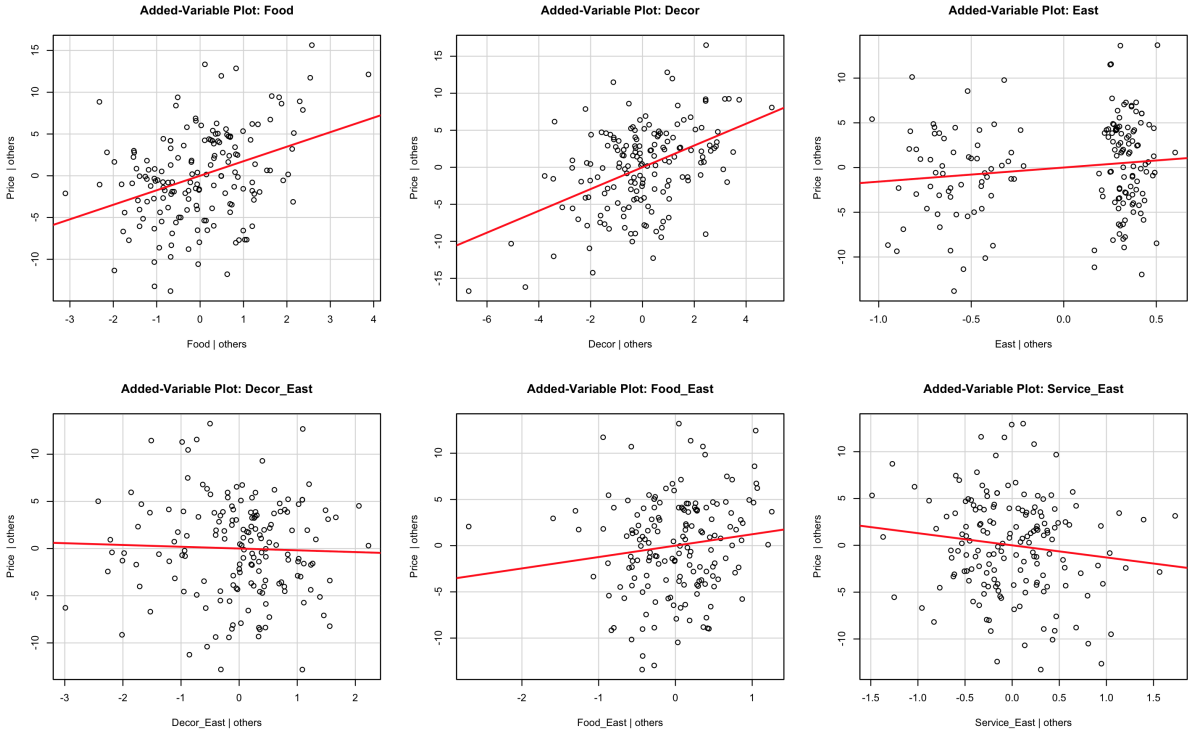

Figure 13: Added variable plots for model 3