**STAT239 Homework 1**

**Linear regression using R**

**Tarek Tohme**

**PART 1:** Studying the effects of car attributes on retail price

**Dataset:** cardata.txt

1)

The variables of interest were the size of the engine, number of cyinders, horsepower, highway MPG, weight, wheelbase and whether the car is hybrid or not. A linear fit of the variables against the suggested retail price yields the following results:

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

Fig 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 with the added variable plots.

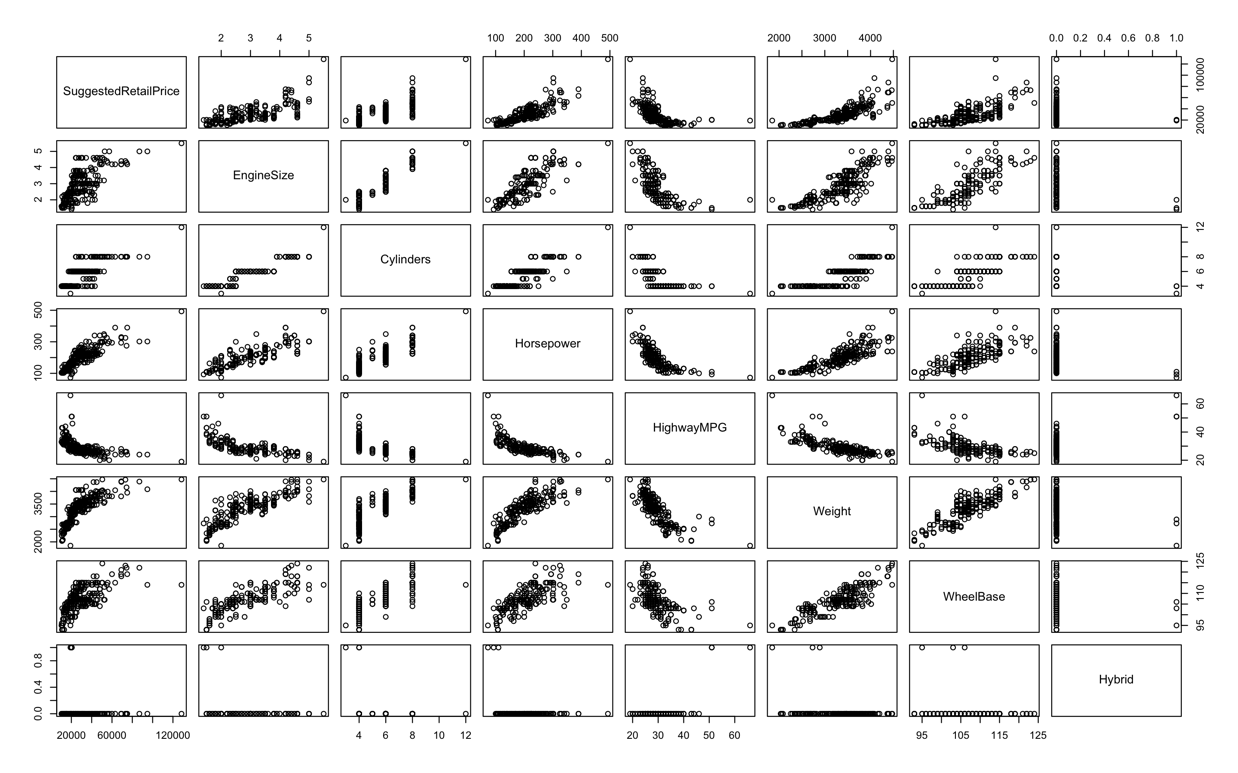


Fig 2. pairs plot for model 1

As we can see in the pairs plot (Fig 2), some variables such as HighwayMPG need to be transformed to fit a linear relationship with the response.

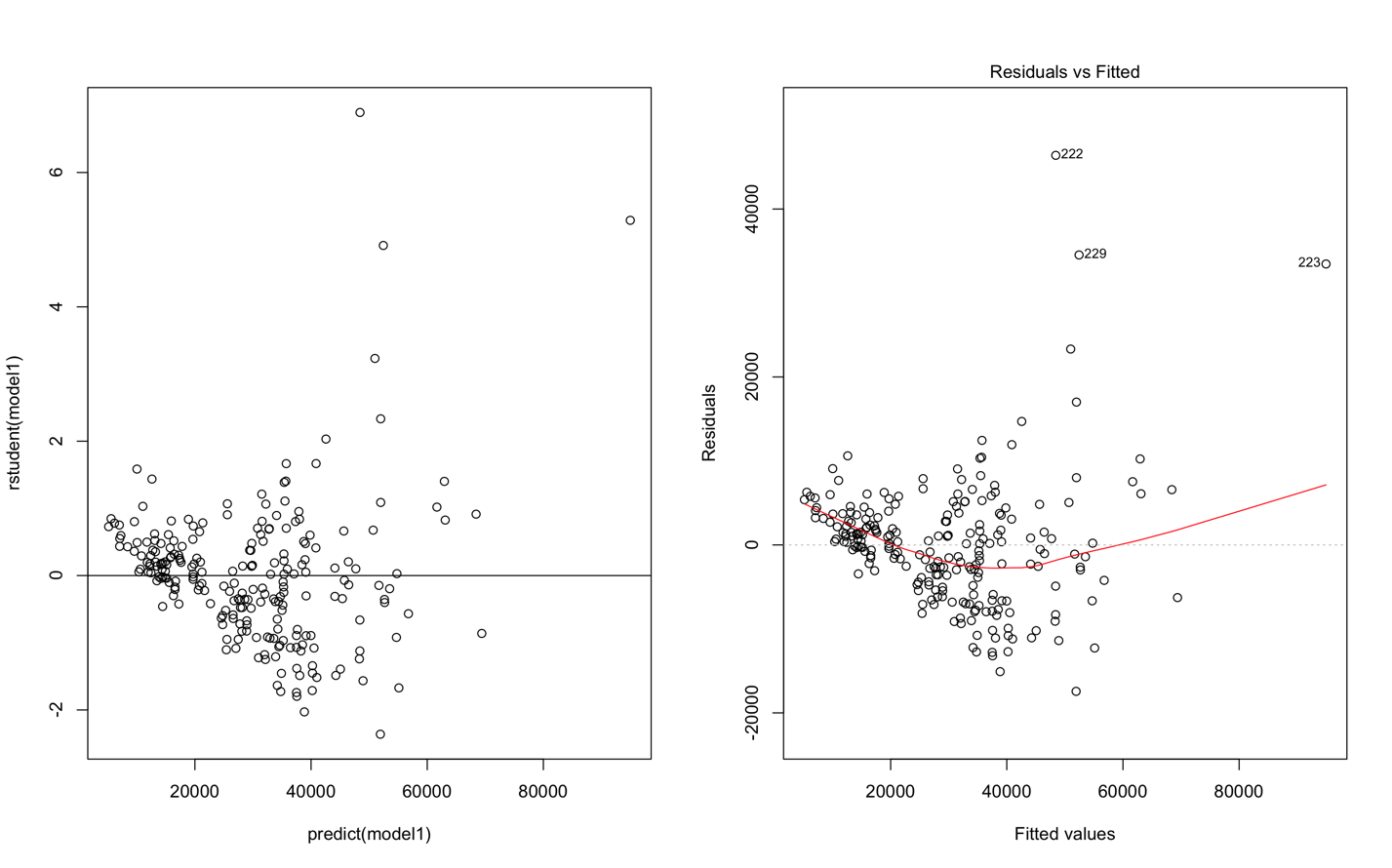


Fig 3. studentized residual plot (left) and regular residual plot (right) for model 1

Fig 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.

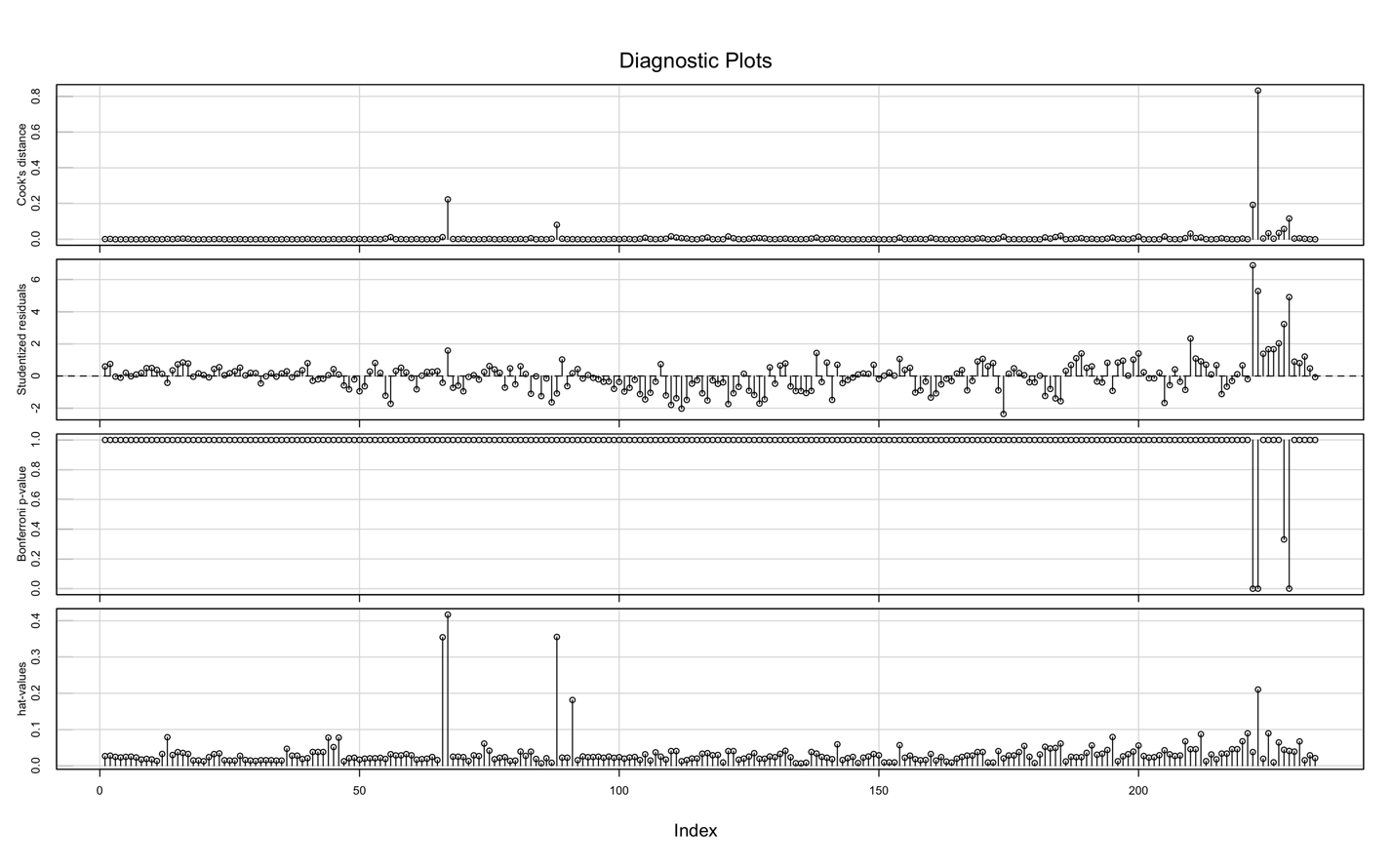


Fig 4. Influence / Index plot for model 1

From the studentized residuals plot, we see that points 222, 229 and 223 are above 4 standard deviations away from the predicted value. However the hat values plot in Fig 4 tells us that only point 222 is significantly influent on the response, so we might need to remove it. Let's transform the data first and see if that's still the case.

The figures above 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 Fig. 5.

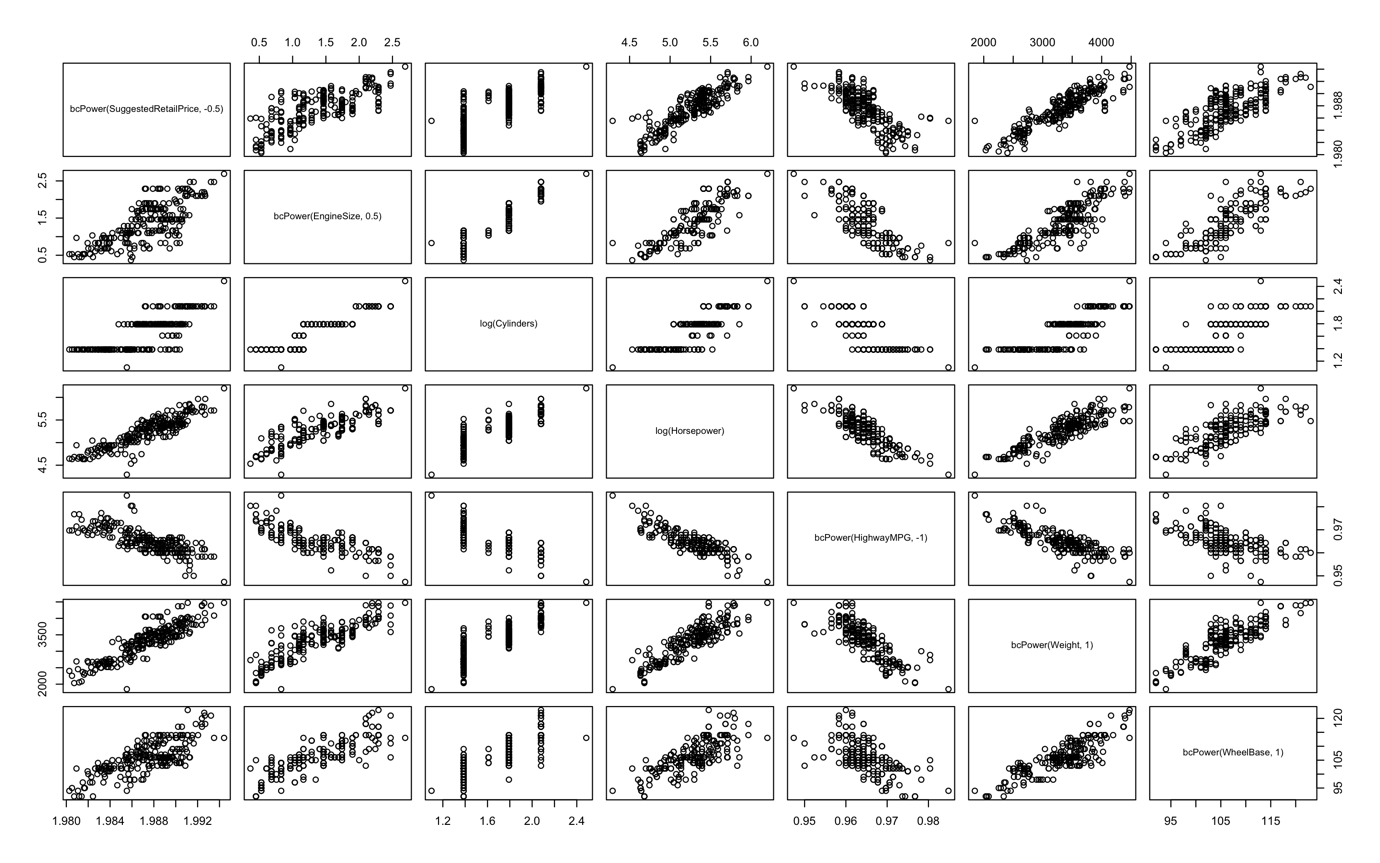


Fig 5. pairs plot for model 2.

After the power transformation, a new leverage analysis (shown in figure 5) 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.

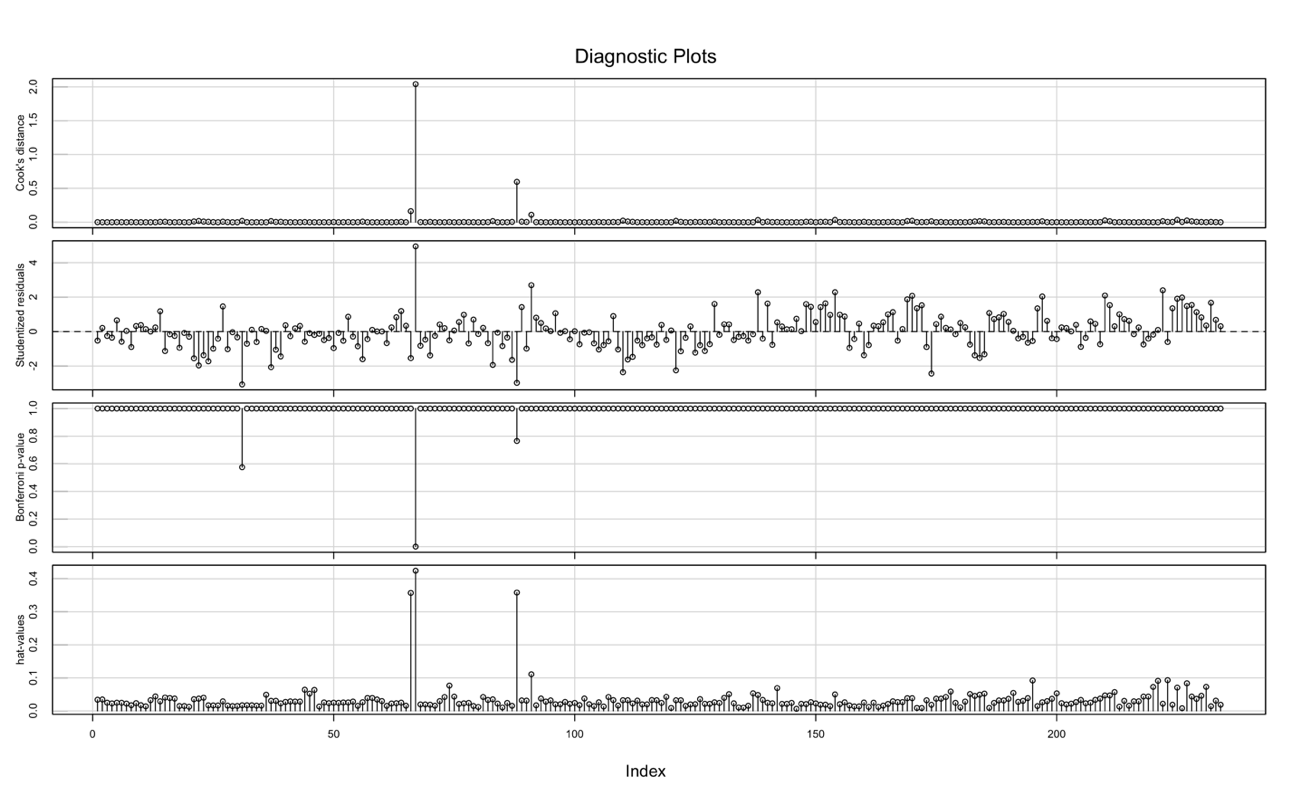


Fig. 6 influence / index plot for model 2

Here is the entry in cardata.txt at index 67:

Manufacturer Hybrid SuggestedRetailPrice DealerCost EngineSize Cylinders Horsepower

67 Honda 1 19110 17911 2 3 73

CityMPG HighwayMPG Weight WheelBase Length Width

67 60 66 1850 95 155 67

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 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 6 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 7 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 6 is significantly higher than that of model 2. We can therefore drop the predictors WheelBase and HighwayMPG.

2)

We could introduce a term in the model that accounts for categorical variables using the factor() function. As we can see in figure 7, 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.

**PART 2:** Studying the effects of Italian restaurant attributes on dinner price in new york

**Dataset:** nyc.csv

1)

We start by making a linear model of the data. Model 1 contains the variables Price, Food, Decor, Service and East.

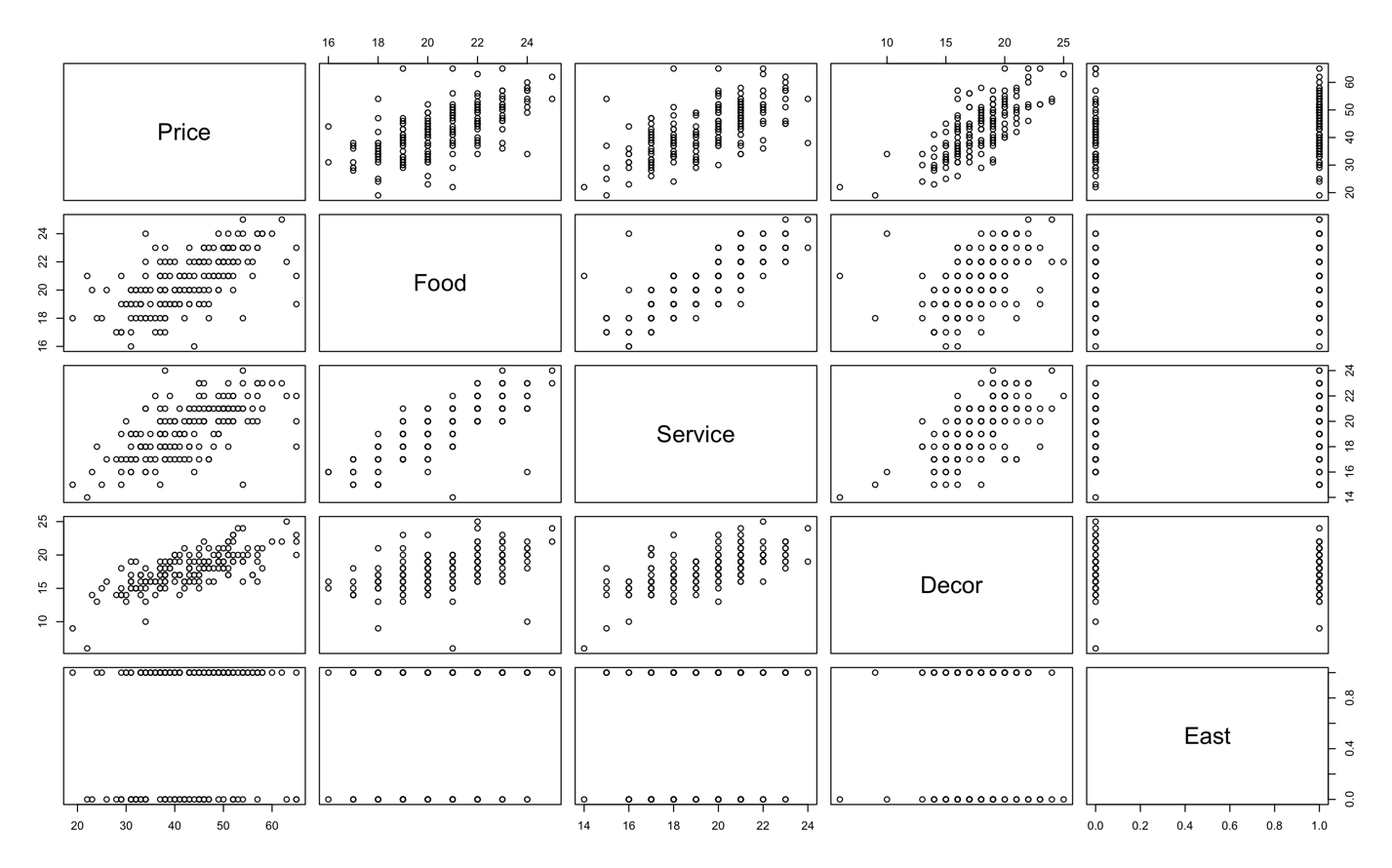


Fig 8. Pairs plot of variables

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 9) shows us that points number 56, 30 and 130 are outliers, but they all have very little leverage, so we keep them.

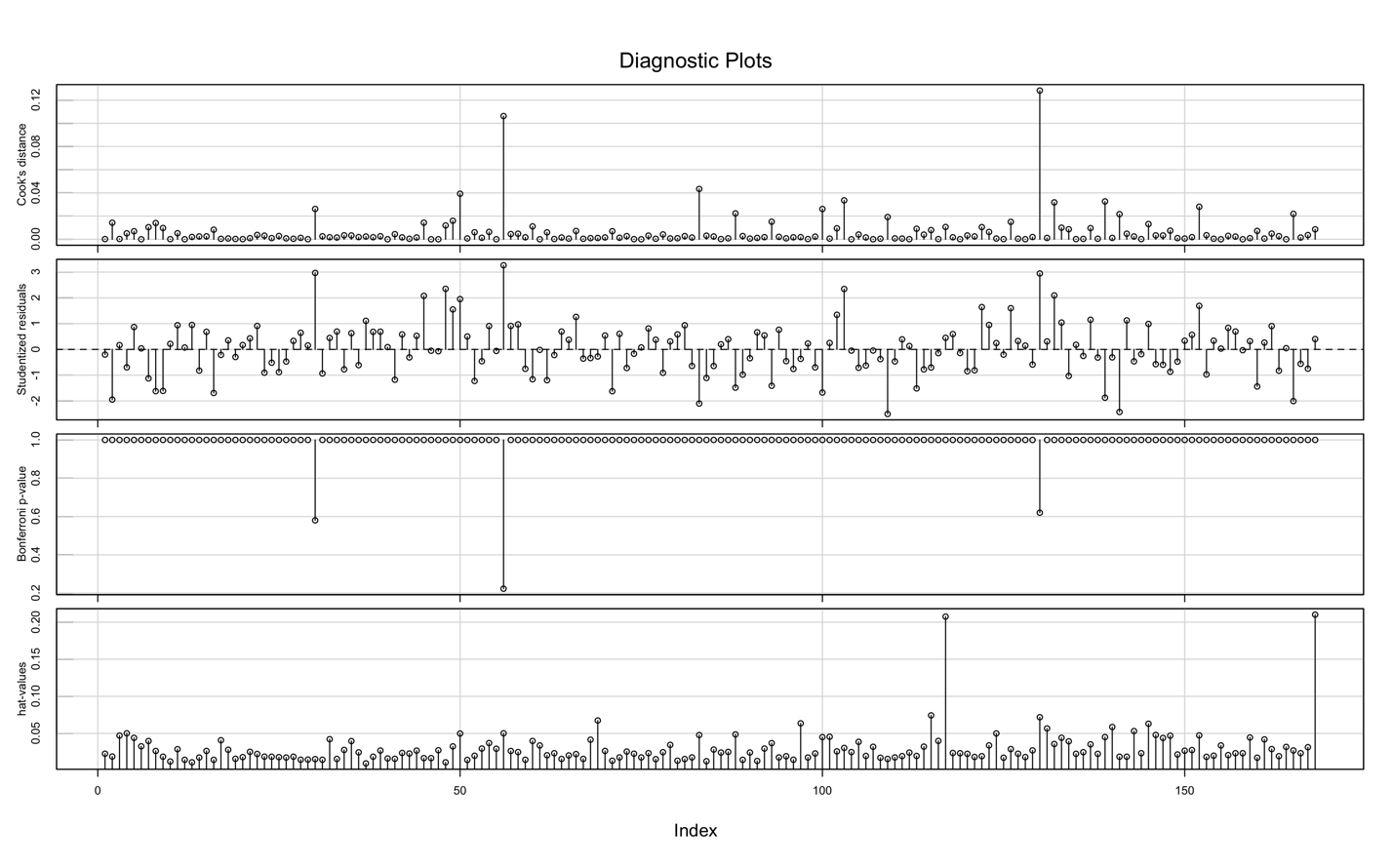


Figure 9. Influence / index plot for restaurant variables

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 are shown in fig. 10.

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 10. Summary of model without Service predictor

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.

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.

