MAT 265
Final Project
Ilia Nikiforov

Modeling item circulation using linear regression

In this report I try to model the duration between check-out and check-in using several variables that I constructed and a sample of 2813 observations. The linear regression showed that adult items, CDs, and DVDs tend to be returned faster. However, the regression method with my dataset failed several important diagnostics, so I conclude that these preliminary findings should be tested using a different method, more appropriate for this data.

Report:

Query 1 below creates a dataset for my analysis using the following specifications:

1. Checkout dates are limited to two days – March 1-2, 2015. This is done to reduce the size of the dataset and speed up the query.
2. The difference in days between check-in and check-out is called “duration” and will be used as the main response variable.
3. The first predictor variable is called “adult” – this is a binary variable that is coded as 1 if item type starts with “ac” and 0 otherwise.
4. The second predictor variable is called item_type – this is a factor variable with 4 levels: book, cd, dvd, and other.
5. The third predictor variable is the hour when an item was checked out – this is a continuous variable.
6. The fourth predictor variable is called “dewey_exists” and is coded as 1 if an item has a dewey class in the database and coded 0 otherwise.
7. Finally, to further reduce the size of the dataset, I use RAND() to draw a sample of 10% of data fetched by the query. The resulting dataset has 2814 rows.

Query 1:

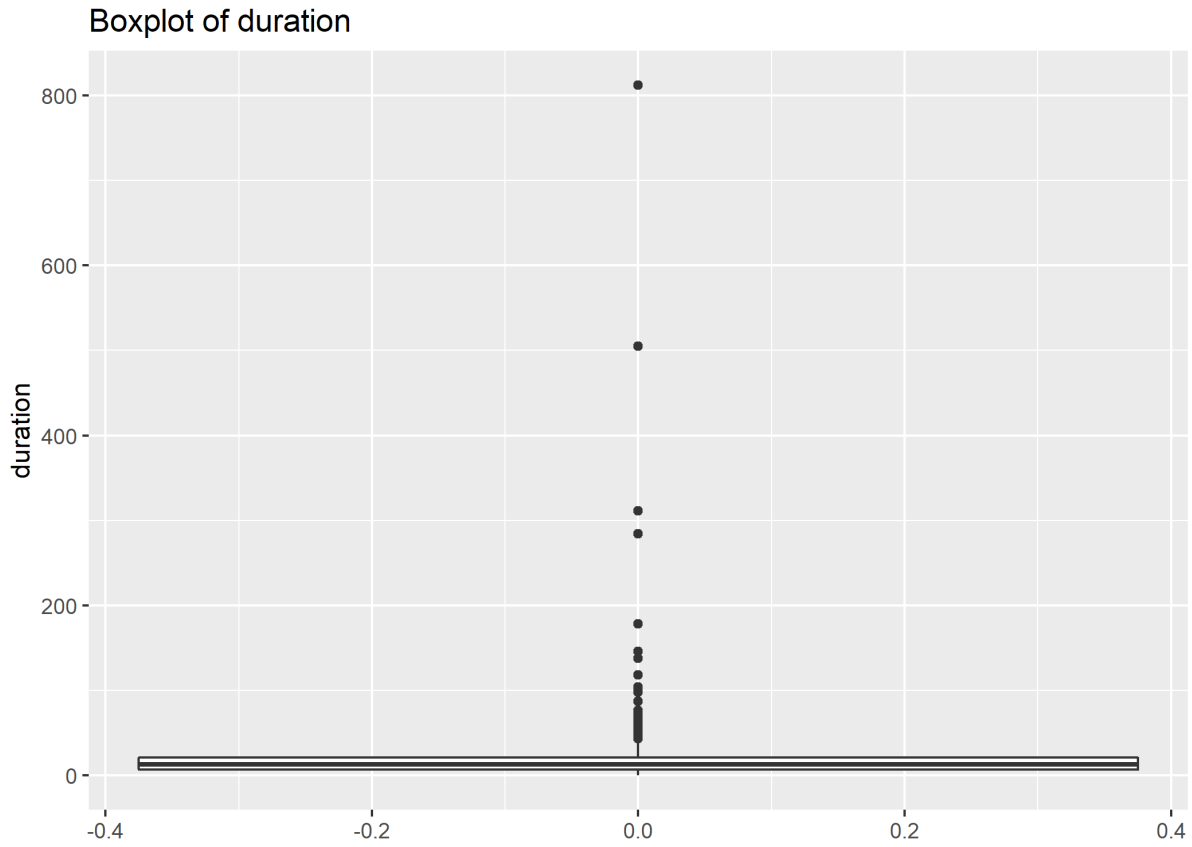
```
select

timestampdiff(day,cout,cin) as duration,

title,

case when itemtype like 'ac%' then '1'

        else '0' end as adult,

case when itemtype like '%bk' then 'book'

        when itemtype like '%cd' then 'cd'

   when itemtype like '%dvd' then 'dvd'
```

```
   else 'other' end as item_type,
hour(cout) as checkout_hour,
case when deweyClass='' then '0'
          else '1' end as dewey_exists
from spl_2016.inraw
where date_format(cout, '%Y-%m-%d') between '2015-03-01' and '2015-03-02'
and RAND() < 0.1
```

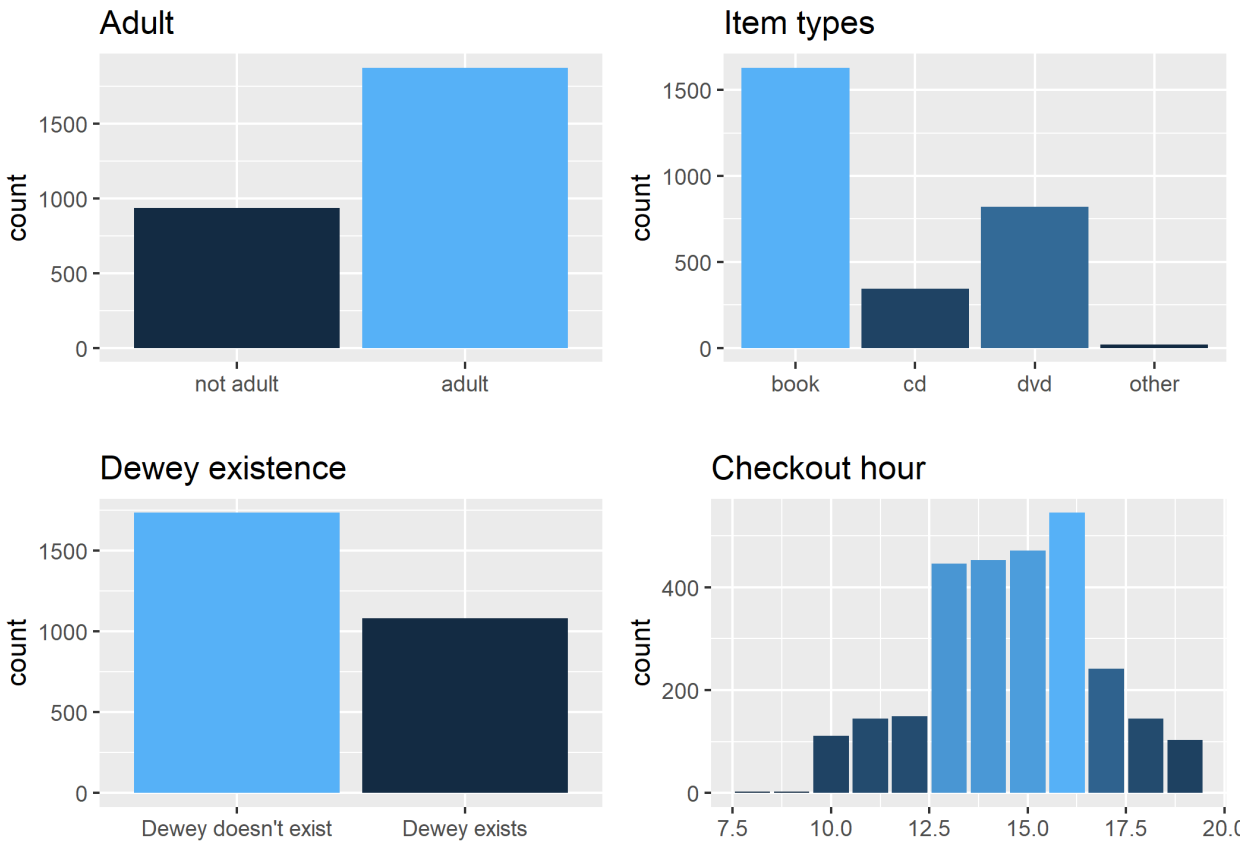
Results

The results are stored in “checkouts_sample_custom_variables.csv”.

I begin with some exploratory data analysis:

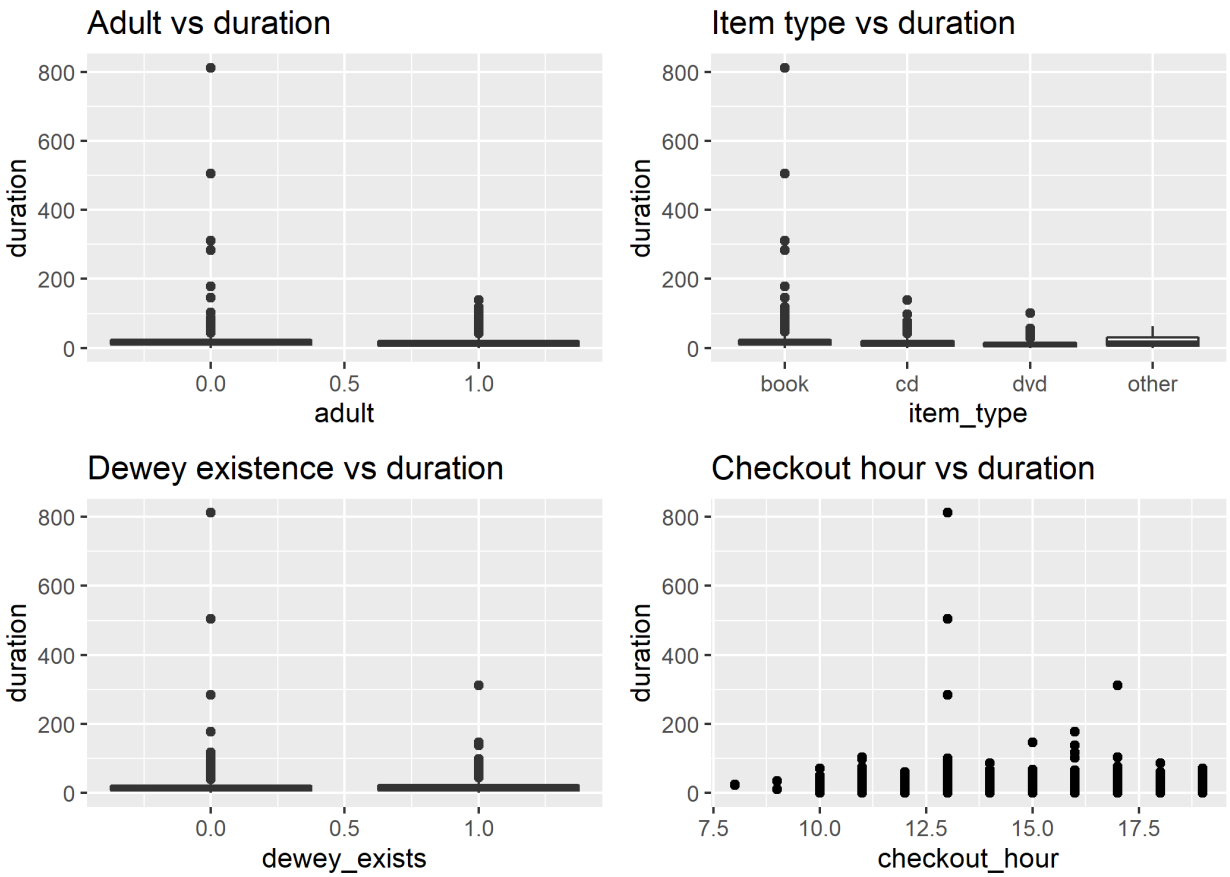


The graph above shows the boxplot of the duration (in days). It is clear that most of the data in the sample has the duration within approximately 25 days; and that there exists a large number of extreme outliers, which will definitely be a problem for the model.



This graph visualizes the distribution of all predictors in the data.

1. Most of the checkouts were in the adult category
2. Majority of checkouts were books, followed by dvd, cd, and a very small number of other categories
3. Interestingly, most items do not have a dewey class in the database
4. Most checkouts occurred in the afternoon



This graph visualizes the duration against each predictor variable in the data. Extreme outliers make it difficult to visualize this relationship and notice any meaningful differences. It does, in fact, seem that most predictors do not create a meaningful difference in the response variable. We may notice that a slight decrease in duration happens in the adult category vs the non-adult category, but it is hard to estimate that difference based on the visual alone, due to outliers. A regression model should be informative.

Modeling

I start with a full regression model, using every predictor. R studio output for the model can be seen below:

```
Call:
lm(formula = duration ~ adult + factor(item_type) + checkout_hour +
    dewey_exists, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-23.41  -9.36  -3.08   3.33 790.79

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             19.2186     3.1658   6.071 1.44e-09 ***
adult                   -2.5652     1.0761  -2.384   0.0172 *
factor(item_type)cd     -5.8676     1.4730  -3.983 6.96e-05 ***
factor(item_type)dvd    -8.4369     1.1823  -7.136 1.22e-12 ***
factor(item_type)other  -0.3418     5.4896  -0.062   0.9504
checkout_hour            0.1529     0.2105   0.727   0.4675
dewey_exists             1.7445     1.0720   1.627   0.1038
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.77 on 2806 degrees of freedom
Multiple R-squared:  0.03666,   Adjusted R-squared:  0.0346
F-statistic:  17.8 on 6 and 2806 DF,  p-value: < 2.2e-16
```

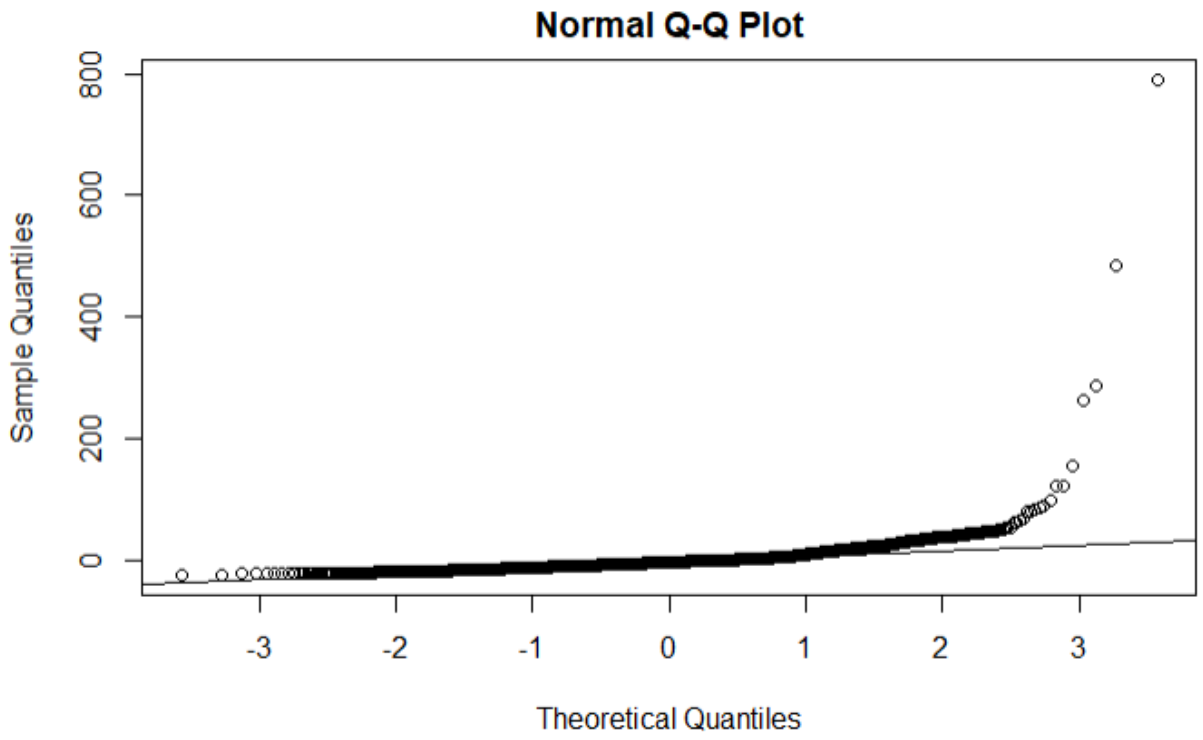
A brief interpretation is as follows:

1) Holding everything else constant, being in adult category reduces the duration by 2.56 hours on average
2) Holding everything else constant, being a cd reduces the duration by 5.86 hours on average (compared to books), whereas being a dvd reduces the duration by 8.43 hours on average (compared to books). Makes sense – as it realistically takes longer to read the book than to listen to a cd or watch a dvd.
3) All other variables have the p-value over 0.05, which means that there is a high chance of observing their coefficients while the null hypothesis (no association between them and the response variable is true). We therefore fail to reject said null hypothesis for item type "other", checkout hour, and dewey existence and must assume these predictors have no effect on the duration.

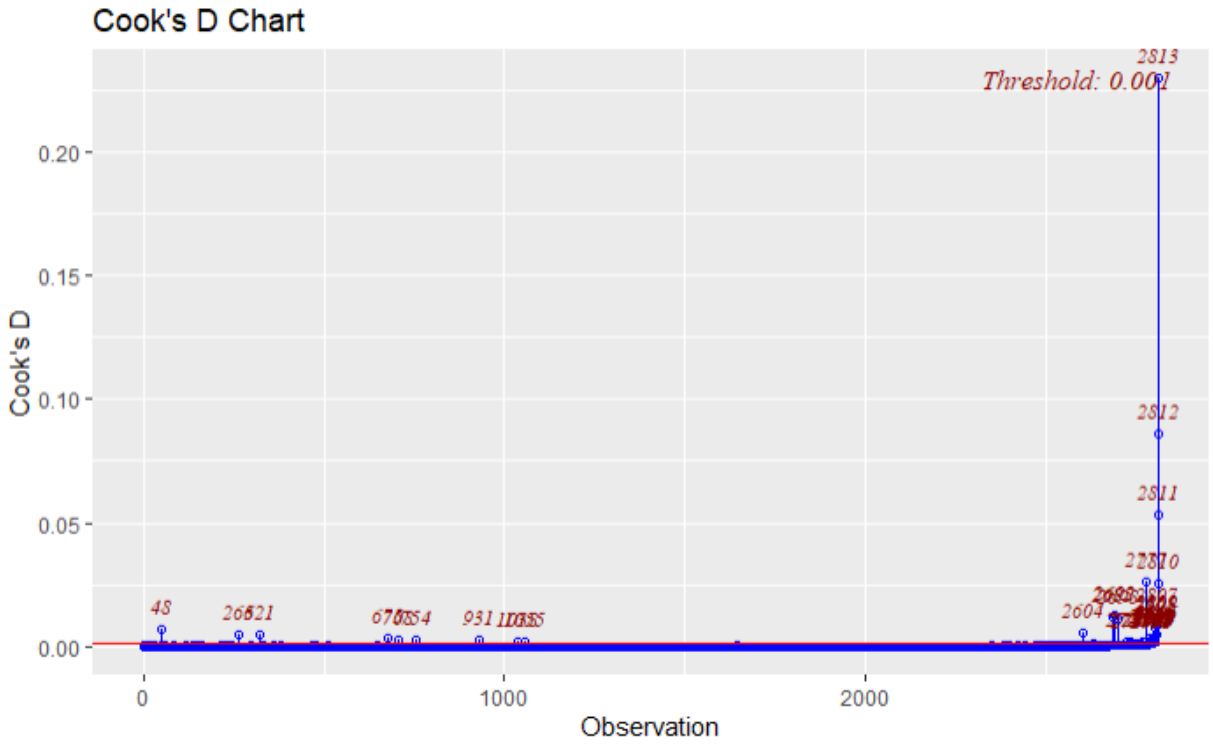
These results generally make sense, however, we must perform a round of diagnostics to check the assumptions associated with linear regressions. The full code and output for diagnostics is in the final_analysis file. Here, in order to conserve space, I will provide a brief summary.

Diagnostics summary:

1. No transformation of variables required
2. No non-linear relationship between predictors and response detected
3. Residuals (the distance between data points and line fitted by the model) are independent.
4. Residuals slightly deviate from normality (see QQ plot below)
5. A large number of influential outliers affect the fit heavily (see Cook's distance plot below)



QQ plot above is used to test the assumption of normality of residuals – the points on the plot should be aligned with the line (regardless of the slope of the line). Here we can see on the right, that a fairly large number of points deviate from the line – this, theoretically, indicates a problem with the fit.



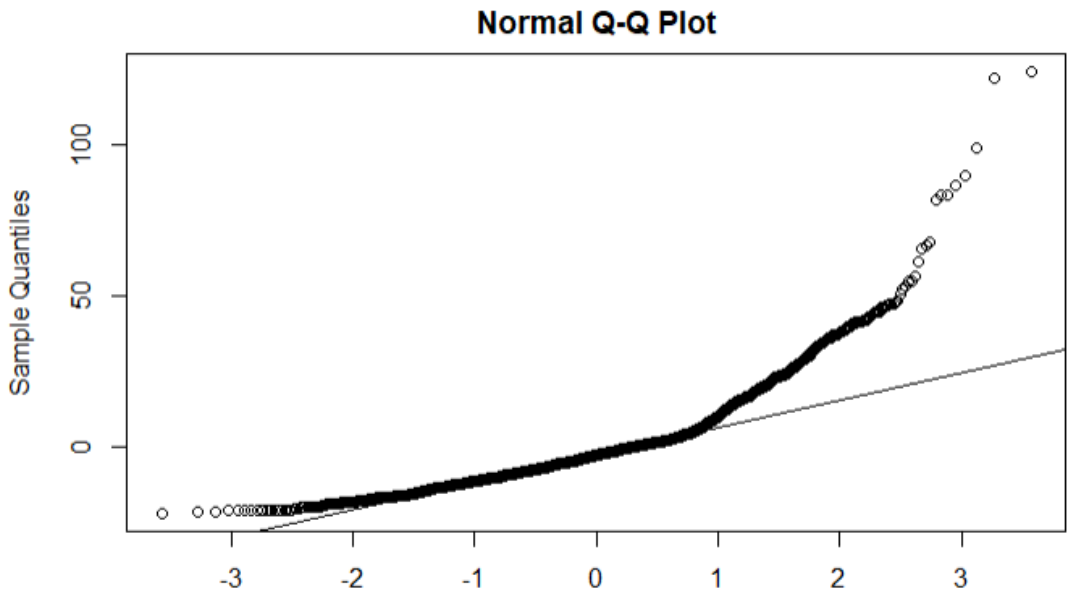
Cook's distance plot above identifies observations that are disproportionately affecting the results of the model – influential observations. Ideally, we would want as few of those as possible. Everything above the threshold line (in this case 0.001, a barely visible line at the very bottom) is considered an influential observation. The plot demonstrates that there is a lot of influential observations, and some observations (such as observation 2813) are extremely influential. Naturally, those observations are outliers:

2806	118	Assassins assignment Jerusalem target Antichrist	1	book	16	0
2807	138	Finnish the short course	1	cd	16	1
2808	146	Dreaming up a celebration of building	0	book	15	1
2809	178	Danny is done with diapers a potty ABC	0	book	16	0
2810	284	Max and Zoe at school	0	book	13	0
2811	311	Rapunzel	0	book	17	1
2812	505	Nohohonzoku kyo mo yurayura nonbiri raifu	0	book	13	0
2813	812	Nohohonzoku kyo mo yurayura nonbiri raifu	0	book	13	0

Generally, statisticians are against removing outliers from the model and prefer changing variables or adding new variables instead, unless there is a clear technical issue with data entry. Out of interest, I tried removing outliers using 3 methods:

1) Manually removing the biggest outliers: 2810,2811,2812,2813
2) Removing outliers based on the threshold: Cook's distance > 4 / n, which removed approximately 35 observations.
3) Removing everything above the initial Cook's distance threshold of 0.001, which removed approximately 80 observations.

The resulting datasets were used for regression and then went through the same rounds of diagnostics repeatedly, which I omit here to conserve space. The results were not good – even after removing **everything** above the initial Cook's distance threshold, the next Cook's distance plot showed about 200 observations that were extremely influential in the new iteration. Furthermore, the more outliers I removed from the model, the higher was the deviation of residuals from normality, see this QQ plot for example:



These results show that indeed, removing outliers is not a good way of dealing with linear regression. It also demonstrates that linear regression is generally not the best method for this data. Linear regression works best with a combination of numerical, interval predictors and response, whereas most of my predictors were categorical.

Conclusion:

Overall, while the first model is flawed, it still points us to some valid ideas that can be explored using different methods. Linear regression seems to be inappropriate for this particular task, because of mostly categorical predictors and the fact that response is heavily affected by outliers. Perhaps a generalized linear model or another tool would be more useful.