STAT 331 Final Project

Deep Antala, Thushar Ishwanthlal, Vanessa Li, Harry Qu

August 4th, 2021

1 Summary

The objective of the report is to investigate the extent to which removing outliers from the pollution data set improves prediction accuracy on birthweight. This is done through fitting and validating models with a holdout sample to determine prediction error. Then, outliers are removed from the data set. Finally, the models are fit using the reduced data set and validated once again to determine their new prediction error. The analysis finishes with a simulation that performs this test on a large number of linear models, producing a histogram of the changes in prediction error as a result of removing outliers.

The report concludes that removing outliers does not necessarily improve prediction accuracy. For a random model, the expected error reduction is normally distributed with a negative mean, which suggests that removing outliers worsens prediction accuracy in general. However, the report also concludes that splitting the covariates into domains changes the results. When regressing on the chemical domain, removing outliers generally worsens the prediction accuracy, whereas when regressing on the lifestyle or outdoor exposures domains, removing outliers generally improves the prediction accuracy.

2 Objective

To investigate the extent to which removing outliers from a data set improves prediction accuracy. We wish to see if removing outliers and refitting the model is a good strategy to create a model with lower prediction error when tested.

3 Exploratory Data Analysis

This section will report only relevant final numbers and figures resulting from the exploratory data analysis; all R code and methodology used to arrive at the numbers and plots are shown in the appendix.

We started by looking at some summary statistics and plotting of the response variable, birthweight. The sample mean of birthweight is 3378 grams and the sample variance is 259317. A histogram of birthweight is shown in Figure 1 below. Birthweight appears to be unimodal, and slightly left skewed. There is a longer tail to the left, representing some outliers with very low birthweights. A boxplot of birthweight is shown in Figure 2 below.

Figure 1: Histogram of birthweight

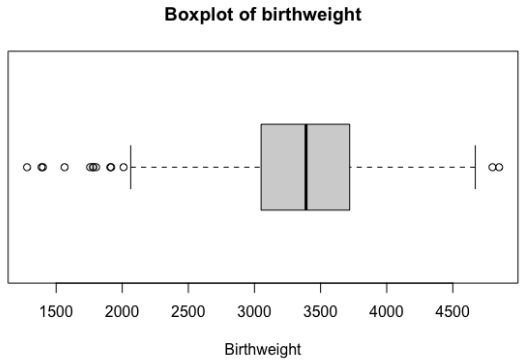

Figure 2: Boxplot of birthweight

These further show that there are a number of outliers, especially on the left of the distribution. These findings motivated the investigation as to whether removing outliers could improve prediction accuracy.

We examined a correlogram to look for pairwise correlations between the variables. A sample of correlogram data is shown in the heat map (Figure 3). In general, there was not strong pairwise correlation between the variables, except for between measurements of different PCBs in the Chemical domain of covariates. The individual variable with the highest correlation with the birthweight response variable was gestational period ($\rho = 0.543$). Figure 4 shows a scatterplot of birthweight and gestational period, split by gender.

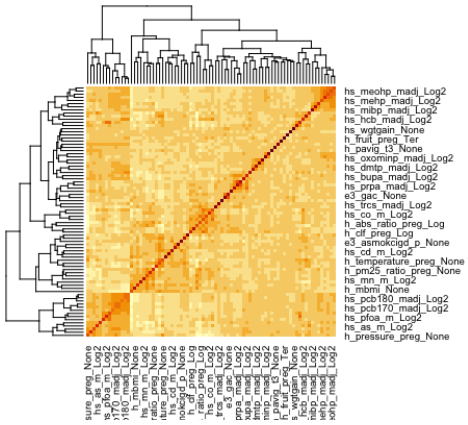

Figure 3: Sample of correlogram

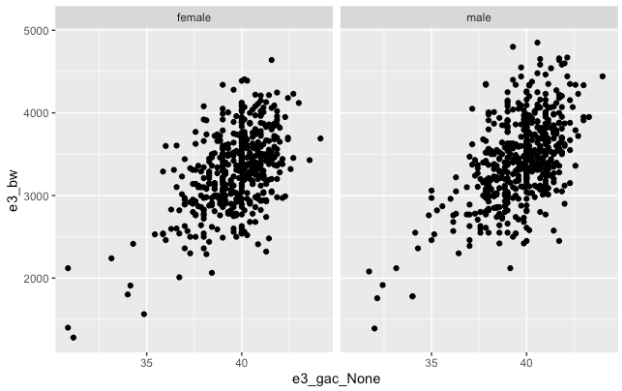

Figure 4: Scatterplot of birthweight and gestational period

Although gestational period is not a pollution-based predictor, the strength of correlation suggests that it would be prudent to include it or interactions with it in a model.

3.1 Initial Data Analysis

As part of our exploratory data analysis, we fit an initial model with all covariates included, without any selection or optimization. Figure 5 shows a histogram of studentized residuals with a N(0,1) distribution overlaid. Figure 6 shows a Q-Q plot of the studentized residuals. These suggest normality under the model.

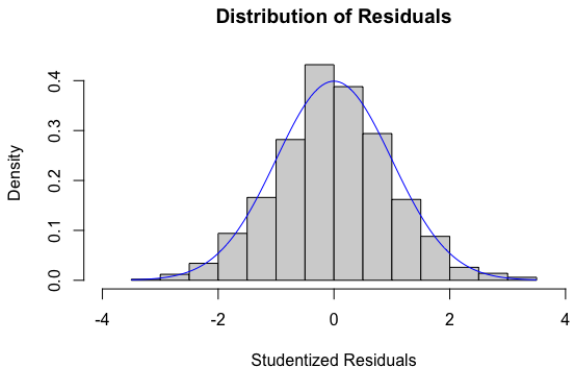

Figure 5: Histogram of studentized residuals

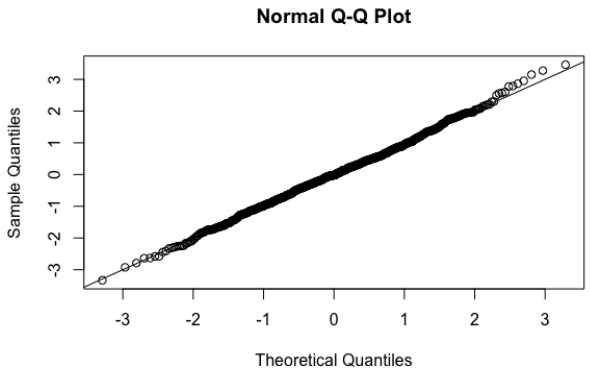

Figure 6: Q-Q plot

Figure 7 shows a sample of some added-variable plots generated by partial regression. The added-variable plots suggest linearity under the model.

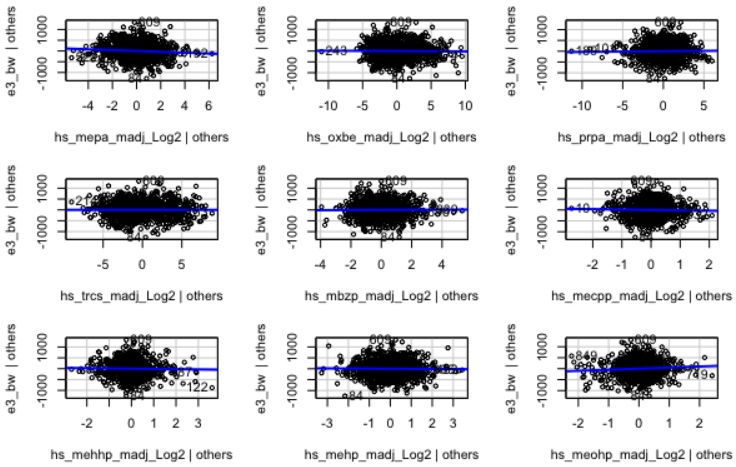

Figure 7: Sample of added-variable plots

Figure 8 shows a plot of residuals against fitted values to assess heteroskedasticity. The uniform distribution of residuals suggests that the observations have equal variance.

Without the methodology of data collection, it is difficult to assess whether observations are independent, so we will assume independence for the purposes of this report. Subsequent models used for prediction will be tested for these assumptions in a similar fashion, however for brevity the plots will not be shown in the report.

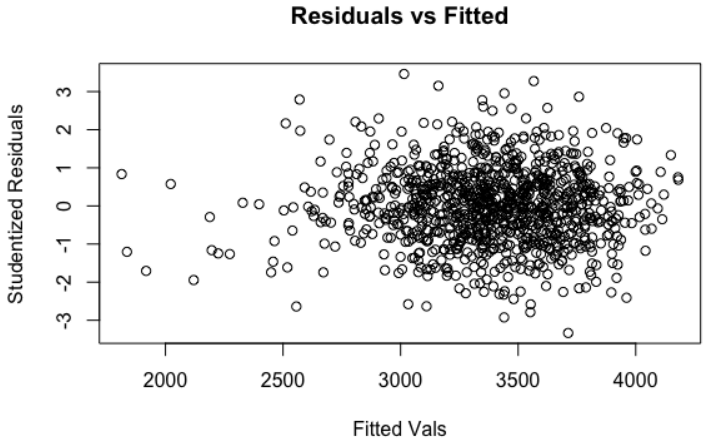

Figure 8: Residuals against fitted values

4 Methods

Our procedure is as follows:

1. Create a holdout sample. Here, our training set contains the first 600 observations and the test set contains the last 400.
2. Fit Model 1 using this training set using the respective selection method.
3. Use DFFITS to pick out outliers, using the statistical rule of thumb $|DFFITS_i| > 2 \times \sqrt{\frac{p+1}{n}}$, where $p =$ number of covariates indicated by the best model fit and $n = 600$.
4. Create a second training set excluding the outliers found above.
5. Fit Model 2 on the second training set using the same selection method in step 2.
6. Measure prediction accuracy with mean squared prediction error on the test set using Model 1.
7. Measure prediction accuracy with mean squared prediction error on the test set using Model 2.
8. Compare the two MSPEs.

We used the following selection methods: automatic selection (without interactions), manual selection (with interactions), and random selection (simulation).

First, we fit the full model, then removed each covariate with the maximum VIF that satisfy VIF>10 each time we refit the model. This way, we were able to minimize multicollinearity among individual variates. Then, we used each of the mentioned selection methods to generate the best fitting model based on AIC, with data set being the training set. DFFITS are the scaled differences between the fitted value for y_i and what we would have gotten if we hadn't observed y_i. A large value of DFFITs suggests the fitted values change substantially.

We chose DFFITS as our standard to pick out outliers because it incorporates information about both y-outliers and x-outliers. Furthermore, we do not need to refit the model for each y_i. Although both DFFITS and Cook's Distance can be generated using functions built in in R, DFFITS is more favorable over Cook's Distance because its statistical "rule of thumb" is more straightforward and yielded more reasonable outliers in our procedure.

For the full model in our selection methods, we considered both including the interactions and excluding the interactions to examine whether important interactions affect our results significantly.

4.1 Analysis

No Interaction

Minimal Model: lm(e3_bw ~ 1, data = training_set)

Full Model (no interactions): lm(e3_bw ~ ., data = training_set)

Full Model (with interactions): lm(e3_bw ~ (.)^ 2, data = training_set)

Minimal model and full model is used in selection methods as initial and threshold model respectively. *training_set* was defined as the first 600 observations and *test_set* was defined as the last 400 observations.

Forward Selection

We performed forward selection with the minimal and full models with *train_set* using the *step* function built-in to R. We call this Model 1. Figure 9 is a plot that highlights high influential points in red. Then we performed forward selection again with the same minimal model and full model but with the new training data set. We call this Model 2. The new training data set does not include high influential points (outliers). We used MSPE to assess the prediction accuracy of the two models.

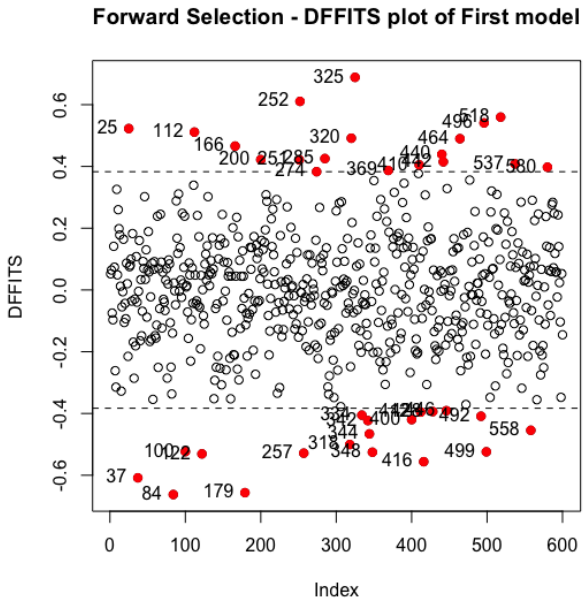

Figure 9: Forward Selection - DFFITS plot with Outliers Marked

Backward Selection

Next we performed backward selection with the minimal and full models with *train_set* using the *step* function built-in to R. We call this Model 1. Figure 10 is a plot that highlights high influential points in red. Then we performed backward selection again with the same minimal model and full model but with the new training data set. We call this Model 2. The new training data set does not include high influential points (outliers). We used MSPE to assess the prediction accuracy of the two models.

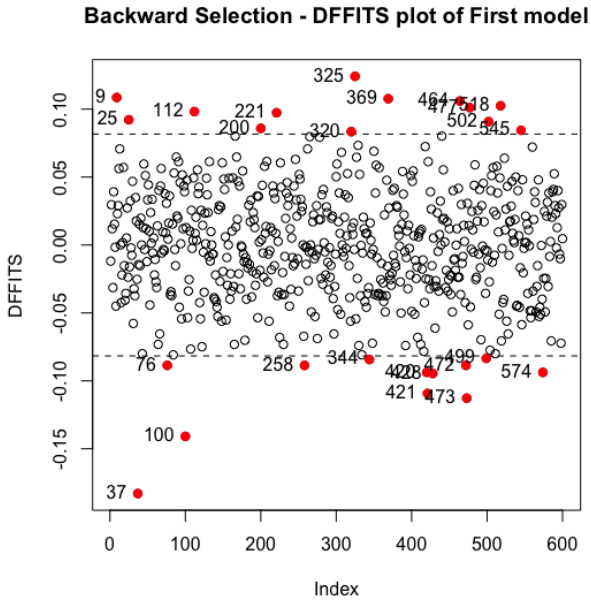

Figure 10: Backward Selection - DFFITS plot with Outliers Marked

Stepwise Selection

Finally, we performed step-wise selection with the minimal and full models with *train_set* using *step* function built-in to R. We call this Model 1. Figure 11 is a plot that highlights high influential points in red. We then performed step-wise selection again with the same minimal model and full model but with the new training data set. We call this Model 2. The new training data set does not include high influence points (outliers). We used MSPE to assess the prediction accuracy of the two models.

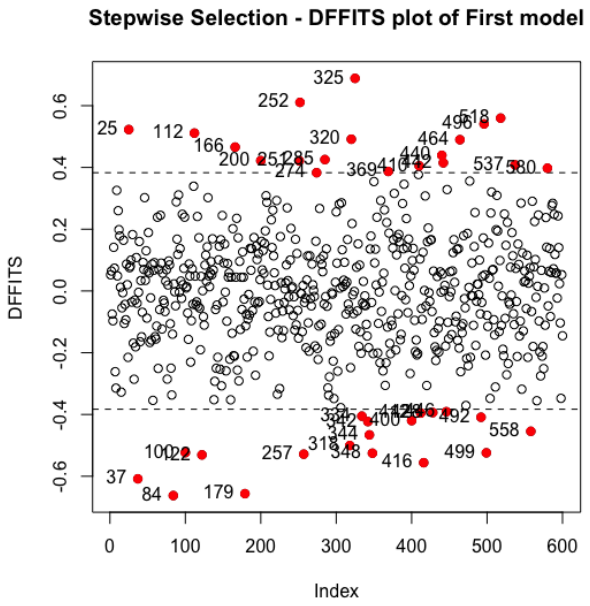

Figure 11: Stepwise Selection - DFFITS plot with Outliers Marked

4.2 Simulation

During the process of selecting certain methods and comparing the models created before and after removing outliers, there are occasions in which we see an increase in MSPE after removing outliers and at other times we see a decrease. To verify which happens more often, we create a simulation to test many randomly selected models to identify a trend. We split the data into 4 sections to describe the effect of removing the outliers in each: chemicals, outdoors, lifestyles and others. The simulation runs as follows:

1. Split the data into a training and a test set. The way the simulation is presented, the training sets that are used are [1:600], [101:700], [201:800], [301:900], [401:1000]. In each case, the other observations are used as the test case. This form of cross validation ensures that the results are not restricted to just removing the outliers in one of the sets.

2. Generate a random set of 1s or 0s. Each one indicates that the column should be included in the model and the zero means not. We can also control the probability that a column is chosen or not. We generate as many of these sets of 1s and 0s as we want models in our sample. From these sets, we generate a list of vectors indicating which parameters to include in each model. i.e. If the vector c(1,3,7) appears in the list, then one of the models in the sample will be the model created by including the 1st, 3rd, and 7th covariate as features (all models in the sample have the same dependent variable.)

3. For each of the models created by the sample above, we perform the method described previously, noting the DFFITS outliers and removing them to create a second model. We record the number of features, number of outliers, the MSPEs of the model against the current test set and the difference between the MSPEs (model with outliers – model with outliers removed).

5 Results

Forward Selection (no interactions)

There were 20 outliers for Model 1.

1. MSPE for first forward selected model: 189595.1.
2. MSPE for no outliers forward selected model: 187941.4.

Stepwise Selection (no interactions)

There were 38 outliers for Model 1.

1. MSPE for first stepwise selected model: 189595.1.
2. MSPE for no outliers stepwise selected model: 185221.9.

The results above show that when the outliers were removed, the MSPE decreases. Also, we see that the data is more scattered in the second plot (after removing outliers). This is what we wanted.

Backward Selection (no interactions)

There were 35 outliers for Model 1.

1. MSPE for first forward selected model: 196500.4.
2. MSPE for no outliers forward selected model: 200789.4.

Manual Selection (with interactions)

1. Fit: Chemicals + others + interactions
    - MSPE for first model: 224859.6.
    - MSPE for no outlier model: 279121.2.
2. Fit: Outdoors + others + interactions
    - MSPE for first model: 189053.6.
    - MSPE for no outlier model: 188276.7.
3. Fit: Lifestyles + others + interactions
    - MSPE for first model: 199510.4.
    - MSPE for no outlier model: 198149.

From the backward selected model and the model fit on interactions of chemical and demographic covariates, we see that the MSPE increases after outliers are removed. However, for the other models, we see a decrease in MSPE when outliers are removed.

5.1 Results of the simulation

Running the simulation, we get a table of observations of different sets of covariates and the MSPE of both the normal model and the model with the outlier removed. If taking out the outliers from the training data set would improve the quality of the model, we would expect most of the observations to satisfy (MSPE_original $>$ MSPE_no_outliers) in other words the difference between the original and the removed outlier version to be positive however that is not the case. The results give us a normal distribution. The sample mean of the 14995 observations gathered in the simulation is -3130.807 with a standard deviation of 5072.353. The test statistic that we observe in a hypothesis test that true mean is greater than 0 is -75.5823, low enough for R to describe that probability as 0. However if we dissect the columns into the domains chemicals, outdoors, lifestyles and others, we notice that the chemicals and the miscellaneous domains contribute the most to the negative values while the other domains boast at least an expected decrease in MSPE after removing the outliers. In addition, combining the domains seems to combine the net expected outcome of removing the outliers.

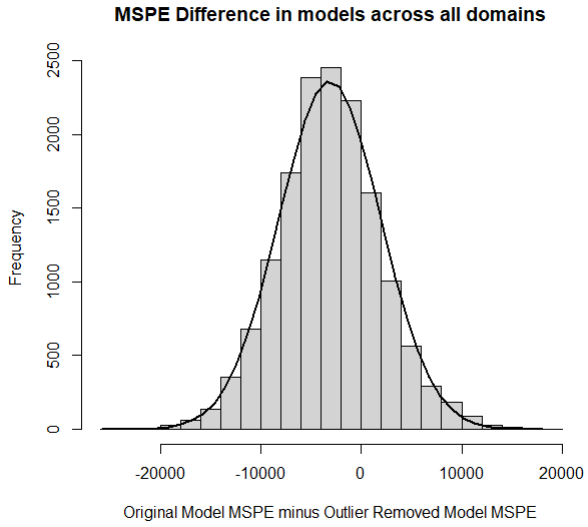

Figure 12: MSPE Difference in models across all domains

Statistical Summary for different domains:

Chemicals: Observations: 15000, Mean: -2710.446, Standard Deviation: 4967.106, Test-Statistic for probability that the mean is above 0: -66.83177, Probability that the mean is above 0: 0 (With underflow).

Outdoors: Observations: 4644, Mean: 549.979, Standard Deviation: 1436.867, Test-Statistic for probability that the mean is above 0: 26.08409, Probability that the mean is above 0: 1 (With underflow).

Lifestyle: Observations: 9362, Mean: 697.1062, Standard Deviation: 1841.945, Test-Statistic for probability that the mean is above 0: 36.61902, Probability that the mean is above 0: 1 (With underflow).

Misc. : Observations: 600, Mean: -1079.204, Standard Deviation: 1695.341, Test-Statistic for probability that the mean is above 0: -15.59273, Probability that the mean is above 0: 4.078588e-55.

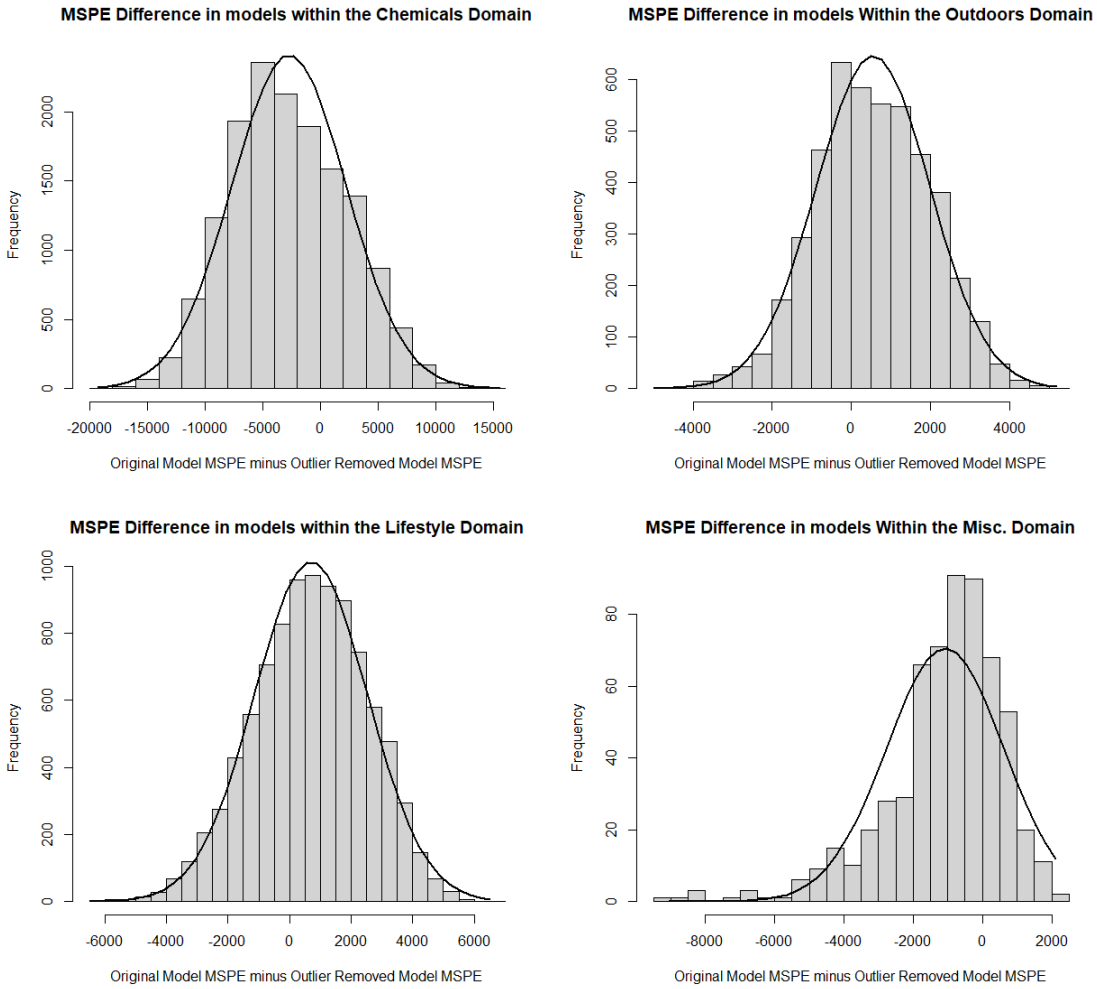

Figure 13: MSPE Difference in models limited to each individual domain

6 Discussion

From our results we observe that, although removing outliers sometimes produced a lower prediction error on our models, it was not always the case. One cause of this could be the fact that we used the original test set for our outlier-removed training sets. Since we have stripped the outliers from the original training set, the predicted new observations were generated based on the model fitted using the smaller data set. This means that these predictions could deviate more significantly from the observations in the test set, therefore resulting in larger MSPEs.

A limitation of this analysis is that it does not distinguish legitimate extreme values from errors. In particular we found that, if the outliers from the Chemicals domain were removed, the MSPE from the interaction-included Model 2 (after the removal of outliers) increased. In other words, removing these outliers yielded a worse model.

Therefore, we conclude that the extreme values from the Chemicals domain should not be treated as outliers in default, but instead should have their impacts on our outcomes (birthweight) examined first. Extending from this conclusion, we also conclude that for any data set, we should always examine the impact of extreme values on our results before deciding to remove them. This way, we can avoid removing statistically significant extreme values and consequentially avoid producing a worse model yielding a higher MSPE.