Regression and Diagnostics with Categorical Covariates

Askar Mukhanov, Christopher Qian, Siddhanth Sabharwal And Suyoung Park

Abstract

A common decision people with continuous covariates face is how to treat them. In our study we show that binning a continuous covariate into a categorical variable and then doing regression can outperform not doing so. We show this via some simulation studies where we test binning against polynomial regression. We consider both equal width and quantile binning. Our results indicate in certain scenarios polynomial regression overfits but binning does not. Furthermore, binning does not violate OLS diagnostics. These findings are then confirmed on a real-world dataset. We confirm that binning a continuous variable into a categorical one can increase your MSPE and BIC while satisfying all standard diagnostics.

1 Introduction

In statistics courses we have learned that data might have different forms, and we usually categorize data as continuous or categorical. Once we see data it may seem easy to see which variables are continuous and which are categorical. However, in practice, someone working with social-economics data may consider household income as categorical predictor dividing it into different income groups. The same logic might be applied to many other continuous covariates. The question is how the model after "discretization" is compared to a model with the continous covariate left as is.

The aim of this project is to investigate potential benefits of treating continuous variables as categorical. First, we do simulation studies on generic data to compare the performance of an OLS model with a continuous covariate binned into a categorical covariate (bucketing) against a polynomial regression model. We demonstrate that in certain cases bucketing performs better than the polynomial regression model. Not only does bucketing have higher prediction power, but also it does not violate any of the Gauss-Markov assumptions.

We apply the same technique of bucketing to a real-world dataset. It is a dataset collected by the Illinois government describing their prison population (crim dataset). This dataset has two continuous covariates: age and booking time. It seems natural to group observations by age and booking time. Minors are known to get lighter penalties for the same crimes as compared to adults. We check if binning these continuous predictors positively affects the prediction power. Comprehensive diagnostics are also done to conclude that the binning model does not violate any of our standard assumptions.

2 Simulation study

In this section, we investigate the situations in which bucketing is superior to polynomial regression. Our simulations are carried out in the following manner: first, we generate n observations of the predictor X, where X is distributed uniformly on the interval $[0, 85]$ (the idea is the common continuous predictors like the age). Then, we generate Y according to the relationship $Y = f(X)+\epsilon$, where $\epsilon \sim N(0, \sigma^2)$. We consider two specific functions for this study: $f_1(x) = 100\frac{(1+0.5x-10)}{1+(x-15)^2+(x-20)^2}$ and $f_2(x) = 100\frac{(1+0.5x-10)}{1+(x-15)^2+(x-50)^2} - 1$. The main difference between the two functions is that f_1 rises very sharply on $[18, 20]$, whereas f_2 increases much more gradually (see figure 4). The range of the two functions on $[0, 85]$ is roughly the same. Given this data, we fit a binning model with equal size boundaries and quantile boundaries and evaluate their performances by estimating the expected mean squared error (MSE). To compute this MSEs of each model, we generate 1,000 new observations and take the average of MSEs from 100 different testing sets. Additionally, we also consider different number of samples (from 50 to 800) and magnitude of variance of noise (from 0.1 to 0.6) in our simulation.

2.1 Model selection

We assume f_1 as the underlying function and perform an analysis on model selection. First, we consider the binning model using quantile binning and analyze how varying the number of bins affects AIC and BIC. From figure 1, we can see AIC tends to choose the larger model as it does not penalize the complex model severely like BIC. Secondly, as the number of bins increases, both AIC and BIC decrease at some point then increases then decrease sharply. We find that the best model is obtained by picking the lowest AIC/BIC model before the AIC/BIC starts increasing for the first time.

Next, we examine the performance of different model selection methods: AIC, BIC, and validation MSE. We compute the validation MSE by splitting the data 80-20 into a training and validation set. To see what models we should choose based on AIC, BIC or validation MSE in

the modeling process, we measure the performance of the best models selected by different criteria with different sample size by computing the actual expected MSE. We also include the results from the testing set considering practical situations. In the bottom of two plots from Figure 1, red dots represent the best model selected by each criteria (AIC, BIC or MSE) based on validation set and black dots show the average of 100 MSEs values calculated by $1,000$ new observations. As we can see in the $n = 50$ case, each criteria picks very different number of bins, all black dots lie above red dots, and the model chosen by MSE generally performs worse on the new data points because the MSE model was chosen based on only 10 data points. Therefore, the model based on the validation set is unlikely to be reliable and can mislead the user. Meanwhile, the models selected by AIC/BIC behave similarly on the new data points. When $n = 800$, we can see AIC/BIC/MSE criteria choose similar number of bins with similar MSE and the black dots are well below the red dots. We confirm that with sufficient large n the three criteria choose the exact same model. In conclusion, when n is small choosing the model based on MSE would be a bad choice and one should use either AIC or BIC. If n is very large, the number of buckets converges regardless of criteria. So, we recommend BIC considering computational cost. Lastly, if the n is neither very small nor large, we recommend using MSE to pick the best model.

2.2 Comparison with polynomial regression

We start with f_1, and analyze the performance of the min-BIC-bucket model in comparison to the min-BIC-polynomial model as we vary σ from 0.1 to 0.6 and n from 100 to 400. For each $\sigma \in \{0.1, 0.15, \ldots, 0.6\}$, we generate a training dataset, compute the difference in MSE between the min-BIC bucket model and the min-BIC polynomial model. We repeat this 100 times and plot the average. Likewise, for each $n \in \{100, 200, 300, 400\}$, we compute the difference 50 times and average the result. We observe that the bucketing model using the equal bucket size becomes better relative to the polynomial model as σ increases and n decreases (see figure 3a and 3b). We obtain similar results when we use the quantile boundaries, the performance of the bucketing model is better for much higher variance of noise.

We hypothesize that this result is because of the large slope from $[18, 20]$ from f_1. Visually, from Figure 4a, we can see that the high noise/low number of sample points causes the polynomial model to overfit drastically. Increasing σ and decreasing n are both similar in that doing so makes it harder to detect the underlying function accurately. Thus, varying both in this manner causes the polynomial model to overfit more and more, whereas the bucketing model is not as effected.

Next, we will perform the same analysis with f_2. We find the exact opposite results with f_1. As shown on Figure 3a and 3b, increasing the variance causes the polynomial model to perform better compared to the bucketing model and increasing the training data sample size causes the bucketing model to perform relatively better as well. The reason for this is that with f_2, even with high noise/small number of points, the polynomial model is able to pick out a reasonable curve, which the bucketing model is unable to do (see Figure 4b). This is likely because the underlying function is much shallower than f_1.

Therefore, if our data has high noise, and/or a small number of observations, it would be instructive to determine whether or not the underlying function is more like f_1 or f_2 to help guide whether to use bucketing or polynomial regression.

2.3 The Effect of Binning on Diagnostics

In this part we take a look at diagnostics plots of best binning models from the previous sections. In particular, we check the optimal models for homoscedasticity and normality. Moreover, we investigate how varying a number of bins may affect validity of these assumptions.

First we check homoscedasticity assumption, in other words, we check if errors have equal variance. Since we cannot observe errors we look at the residuals. Scale-Location plot shows if residuals are equally spread along the range of fitted values. Figure 5 demonstrates Scale-Location plots for f_1 and f_2 with different number of bins: minimum, optimal and maximum. It can easily be seen that both optimal binning models satisfy homoscedasticity assumption: the lines are almost perfectly flat.

We can do Breusch Pagan (BP) test to check for homoscedasticity more formally. The null hypothesis of the test is homoscedasticity. For both optimal binning models we fail to reject the null hypothesis at 5% significance level with p-values of 0.1 and .87 respectively.

Figure 5 also shows that as we increase the number of bins the red line goes from having a great slope to being flat around optimal n (# of bins) and gets curvy as we increase n furthermore.

The next diagnostic plot of our interest is Normal Q-Q plot. We want residuals to be aligned with theoretical quantiles in order to satisfy the normality assumption. Figure 6 shows that though some f_1 residuals are off the line, the optimal binning models (in the middle) satisfy the assumption. We see that as we increase the number of bins, residuals look more normal. But having too much buckets may have adverse effect. The right-most subplots demonstrate that a significant number of residuals are off the theoretical quantiles.

3 Applications

3.1 Description of Dataset

The dataset being used for this case study is arrest data from Champaign County Sheriff Office (CCSO) Booking in 2011. We refer to this dataset as "crime". The goal of this analysis is to investigate whether or not the amount of time spent in jail can be predicted better or worse by treating continuous explanatory covariates as-is or by treating them as categorical variables via binning. Before we go through the results of the experiments we will first walk the reader through the dataset.

We have 60,484 rows of data, 1 response variable (total seconds in jail), 7 categorical explanatory variables (STATUTE_CATEGORY, SEX, CITIZENSHIP_US, SERVED_IN_MILITARY, IN_MARRIAGE, RESIDENT_IL, WORKING, ACTUAL_RACE) and 2 continuous explanatory variables (Age at Arrest, Booking Time).

Our diagnostics on our first model model with no transformations and all variables thrown in has a Q-Q plot indicating serious non-normality of residuals. In an attempt to resolve this issue, we log-transformed our response variable, total_seconds_in_jail. This fix worked so our response variable is log_total_seconds_in_jail. (see Figure 8)

3.2 Binning Booking Time

All results summarized in Figure 7a.

Now that the data is ready to be experimented on, we are going to try binning the Booking_Time_Formatted and Age at Arrest variables with two binning strategies: equal length of bins [equal width binning], equal number of observations per bin [quantile binning]. In the appendix we display our plots of out of sample MSPE values and in sample BIC values vs # of bins, which is how we determined the best models.

The experiments show that binning the continuous variable, Booking_Time_Formatted, can in fact improve train BIC and test MSPE. The experiments show that the optimal number of bins according to train BIC is 3 based on equal width binning and 5 based on quantile binning. The train BIC of the optimal equal width binning configuration is below the baseline established by leaving the continuous booking time variable as is.

The experiments show that the optimal number of bins according to test MSPE is 7 based on equal width binning and 5 based on quantile binning.

Furthermore, the optimal test MSPE with both binning strategies is below the test MSPE established by the baseline model where the variable booking time was left as continuous. We will choose the model with equal width binning and 7 bins as optimal since both its test MSPE and train BIC are better than their respective baselines.

Between, the equal width and quantile binning models for this variable both configurations perform quite similarly. We suspect this is due to the uniform nature of this variable.

3.3 Binning Age

All results summarized in Figure 7b.

The experiments for the Age at Arrest variable, show that the optimal number of bins according to train BIC is 8 based on equal width binning and 10 based on quantile binning. The train BIC of both the optimal equal width and quantile binning configuration is below the baseline established by leaving the continuous age variable as is.

The experiments show that the optimal number of bins according to test MSPE is 9 based on equal width binning and 10 based on quantile binning. The optimal test MSPE with both binning strategies is below the test MSPE established by the baseline model where the variable age was left as continuous in its raw form.

We will choose the quantile binning model with 3 bins as the optimal since both its train BIC and test MSPE are significantly better than the baseline, and the test MSPE itself is very close to the optimal test MSPE across all models. Furthermore, we gain a lot of degrees of freedom by choosing this model.

For this variable, the quantile model tends to consistently outperform the equal width binning model across lower number of bins and vice-versa across higher number of bins. We suspect this is due to the right skewed distribution of this variable, so for lower number of bins the quantile binning is able to better approximate the true underlying continuous data because it is not as coarse as equal width binning where the number of observations per bin can be quite uneven. The uneveness is not as profound when we have more bins, because any one bin does not dominate when we have skewed underlying data.

Our experiments suggest that if the underlying continuous variable has an approximately uniform distribution like age at arrest, the equal width and quantile binning strategies perform about the same. If the underlying continuous variable has a skewed distribution like the booking time variable, then using a quantile binning strategy is better because it prevents class imbalance amongst the levels of the categorical variable.

Our experiments on this real-world crime data confirm that binning a continuous variable can improve prediction power.

3.4 Crime Dataset Diagnostics

We had to decide if our baseline model would have higher order terms for our continuous covariates, ie polynomial regression. Using BIC as our criterion, our results indicated that the baseline model should be each of the continuous variables raised to the power of 1. This means we should not be doing polynomial regression with higher powers for any of the continuous variables in our dataset. A visual inspection of the variables plotted against each other with a smoother overlaid confirmed this conclusion.

Our first pass suggested to drop the STATUTE_CATEGORY variable because the VIFs for its levels were far greater than 10. We dropped that variable and then reran the model, and still got improvements over our baseline so we decided to remove that variable from our final models because it had too many levels. This suggest that for categorical covariates with a lot of levels, one should test if removing it helps decrease the standard errors for the coefficients of the other variables.

In the second pass with that STATUTE_CATEGORY variable removed, binning the variables doesn't increase the variance of the other coefficients by much. The mean VIF went from 1.37 to 1.40. Additionally, none of the variables alone had a VIF above our rule of thumb of 10, indicating our optimal model doesn't suffer from multicollinearity.

For the next diagnostic, we checked if there are influential observations according to Cook's distance in the model and if we should remove them. When we computed the values, all the Cook's distances were very close to 0, none of them are above our rule of thumb of 1, indicating no observations are being overweighted in our analysis. Furthermore, the relative magnitudes of the Cook's distances are about the same, no single observation is far above the others. Our model passed this influential observations diagnostic check.

The final diagnostic checks were done on our baseline model with unbinned continuous variables and on our optimal model which used the equal width binning strategy on the booking time variable with 7 bins. The standard four diagnostic plots indicated nothing unusual. (see Figure 2)

4 Conclusion

This paper attempted to study potential benefits of "discretizing" continuous variables and treating them as categorical predictors. We did simulations to better understand advantages and limitations of binning over leaving continuous variables as it is. In particular, we found that selecting a binning model on the basis of BIC is preferred over AIC and validation MSE. Also, we compared the minimum BIC binning model with the minimum BIC polynomial model and found that depending on the shape of the underlying function, the standard deviation of the noise, and the sample size, the binning model may be superior. Moreover, in both cases the binning models did not violate the Gauss-Markov assumptions: homoscedasticity and normality of the errors. Therefore, we concluded that binning can be a reasonable choice for modeling data.

In the next section we used arrest data from CCSO Booking in 2011 to confirm findings from the simulation part. Two continuous predictors, age and booking time, were treated as categorical variables via two different binning strategies: quantile and equal width binning. Figure 7 demonstrates the models selected by different binning strategies having higher prediction power than the baseline model. For both predictors, MSPE of binning models are smaller when # of bins is greater than 5. We also concluded that quantile binning is preferred for this dataset since its MSPE is comparable with that of the optimal model but has less predictors (more degrees of freedom). Finally, we did some diagnostics to check the optimal binning model for equal variance and normality of residuals, multicollinearity and influential observations. Figure 2 displays the standard diagnostics plots which suggest normality and homoscedasticity of residuals. None of the observations had Cook's distance value larger than 1, and we concluded that there were no multicollinearity via inspecting the VIFs.

To conclude, this paper demonstrated benefits of "discretizing" continuous covariates into categorical covariates in certain scenarios. We saw that treating age and booking time as categorical predictors significantly increases prediction accuracy. However, further studies must be done to see if the binning strategy works for predictors that cannot be intuitively grouped. Moreover, performance of binning models must be compared to that of other linear models as regression splines etc.

Appendices

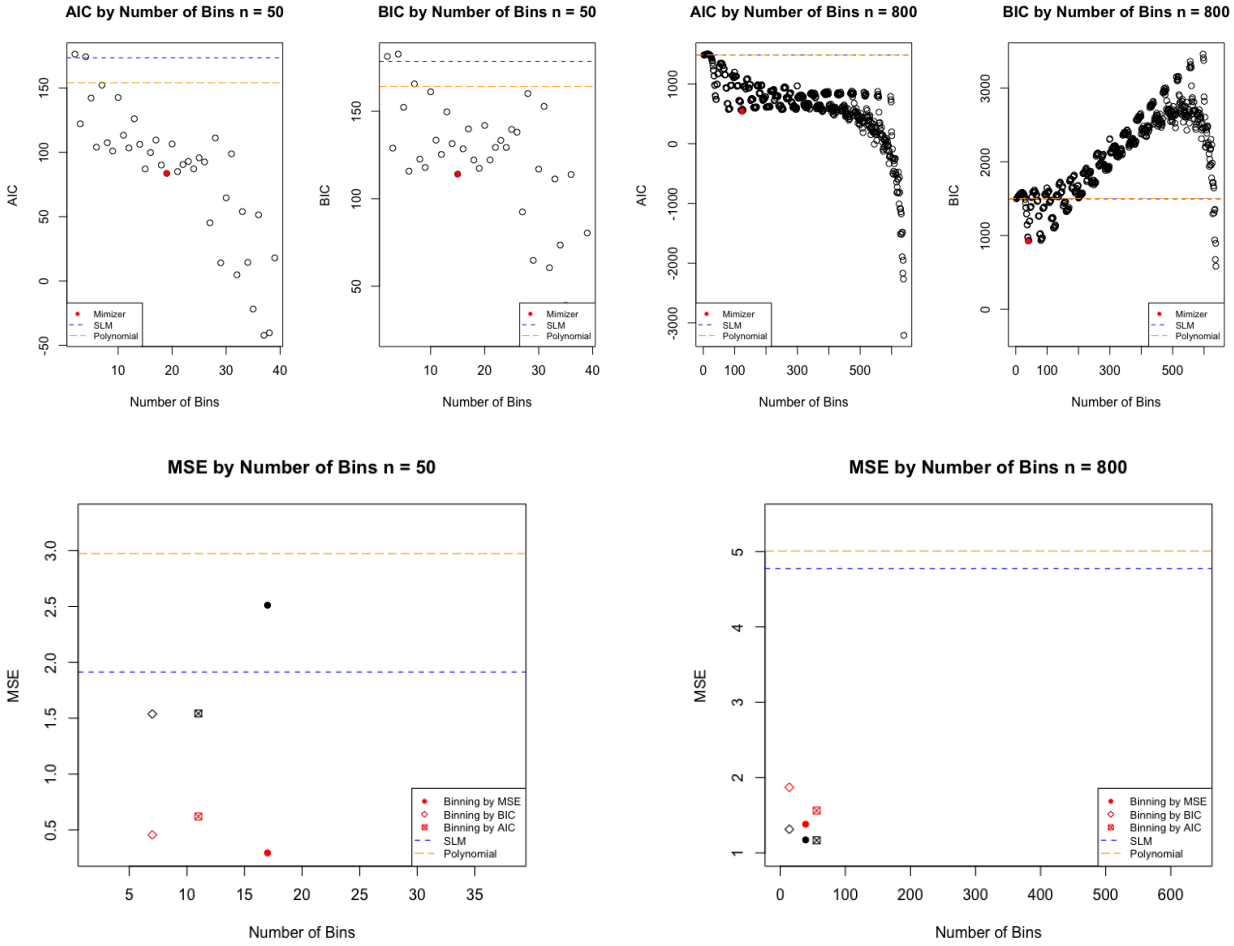


Figure 1: Optimal Model by Different Criteria, Bins and Size. $\sigma = 0.1$

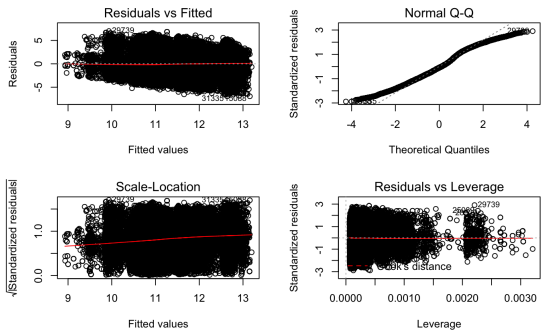
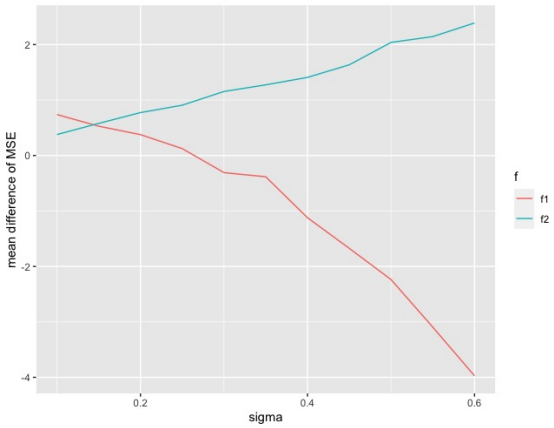
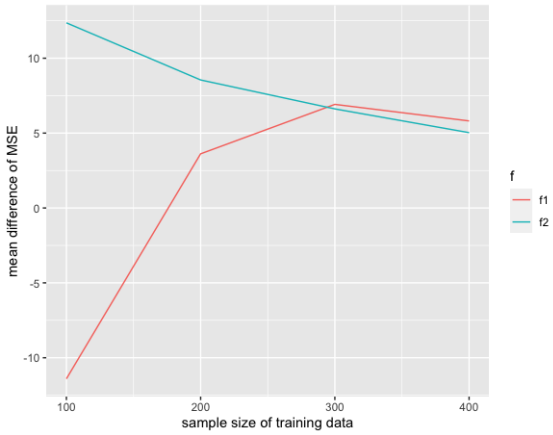


Figure 2: Final Selected Model Diagnostic Plots

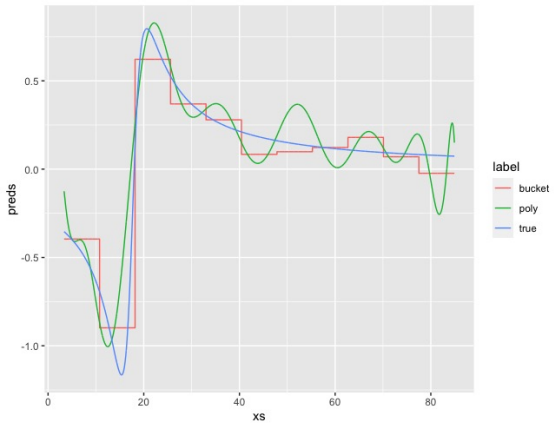


(a) Difference in mean MSE by σ, $n = 200$

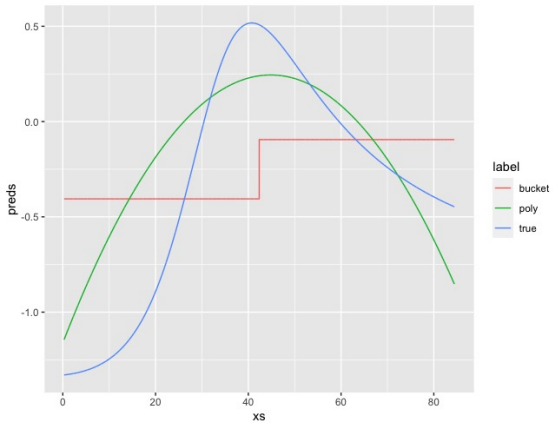


(b) Difference in mean MSE by n, $\sigma = 0.2$

Figure 3: Effects of sample size and noise

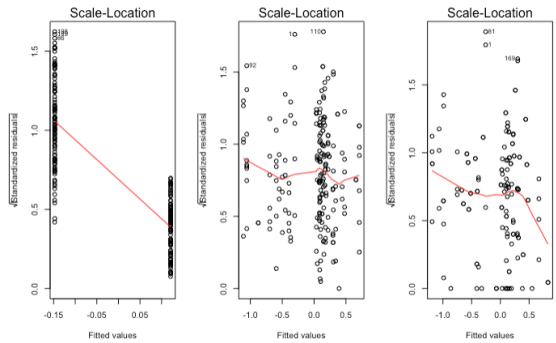


(a) Fitted Lines in f_1 $n = 100$, $\sigma = 0.2$

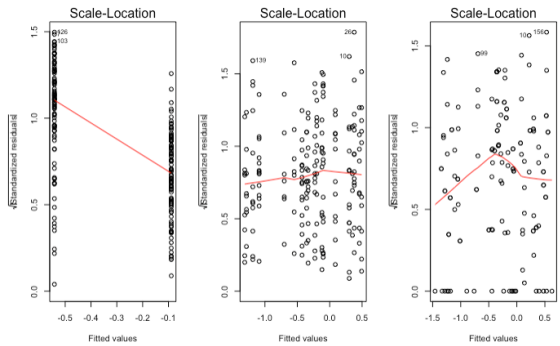


(b) Fitted Lines in f_2 $n = 200$, $\sigma = 2$

Figure 4: Fitted Lines

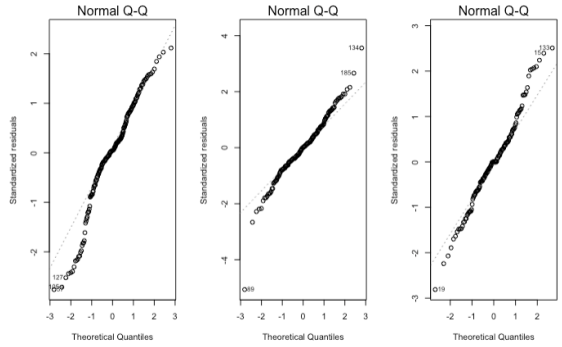


(a) Scale-location for f_1 $n = 2$, $n = 16$, $n = 200$

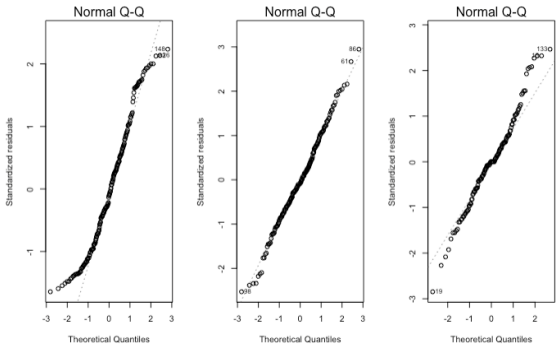


(b) Scale-location for f_2 $n = 2$, $n = 18$, $n = 200$

Figure 5: Comparison of scale-location

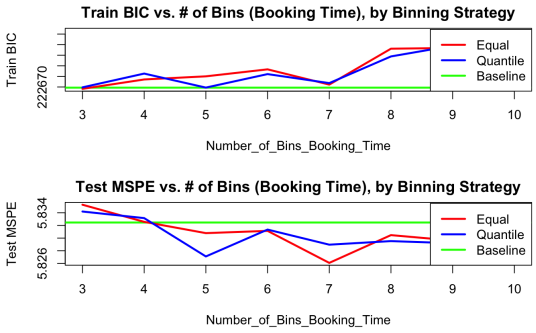


(a) Q-Q plots for f_1 $n = 2$, $n = 16$, $n = 200$

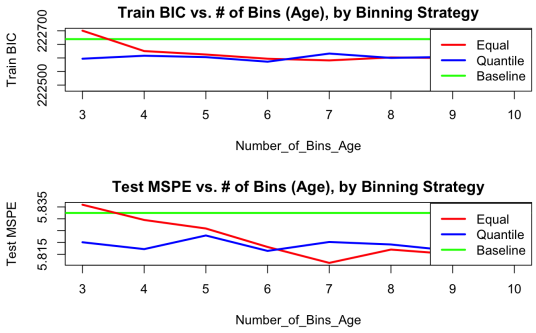


(b) Q-Q plots for f_2 $n = 2$, $n = 18$, $n = 200$

Figure 6: Comparison of Q-Q plots



(a) Booking Time Binning Comparison vs Baseline



(b) Age Binning Comparison vs Baseline

Figure 7: Test MSPE by Model

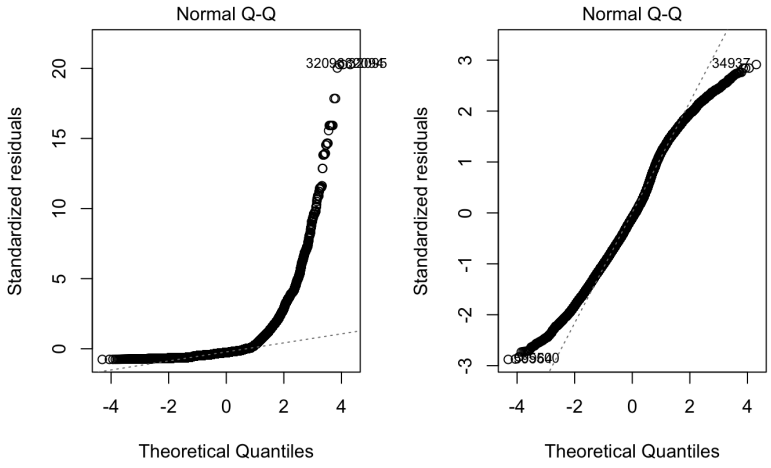


Figure 8: Q-Q Plot Comparison, Untransformed (left) vs Transformed (right)