Modeling Report

Benjamin Lee, Stephanie Trinh, Zhi Long Yeo

2023-05-03

Introduction

For our project, we analyzed various datasets related to bike sharing data from different geographical locations (e.g. Seoul, Washington D.C., London etc) containing information about the number of bikes rented on different days, along with weather conditions (temperature, humidity, etc) and miscellaneous information about the day these bikes were rented (weekends, holidays, etc).

Research question

The question we wish to answer using these datasets, is what various variables/factors can be used to predict the number of bikes that will be rented on a given day.

We also want to test the following hypotheses (since the test procedure for the hypotheses are similar, ie nonparametric bootstrap F-test, we will only be testing a few main ones):

- Is the effect of temperature different across the locations? We expect people in different areas to have different temperature preferences.
- Are any of the betas insignificant when tested as a group? We will test this after model selection
- Does the number of bike users follow any "common" distribution? We expect the ols model to be a poor fit, and that it follows a discrete probability distribution

Our findings could help inform bike sharing companies into making better economic decisions in regards to improving marketing decisions, scheduling, and conducting maintenance.

Primary focus

Our primary focus is causal inference. We aim to figure out what are the factors most likely to affect the count of bike users. We will use nonparametric bootstrap F-tests to achieve this goal.

Data Overview

We are using the following datasets (each bullet-point is a hyperlink) for data exploration and modeling:

- Bike Sharing in Washington D.C. Dataset (2011-2012)
- Seoul Bike Sharing Demand Data Set (2017-2018)
- London bike sharing dataset (2015-2017)

Each dataset contains the hourly count of rental bikes on each specific date, with additional information on weather and holiday schedules. Each observation corresponds to an hour of the day.

Modeling Introduction

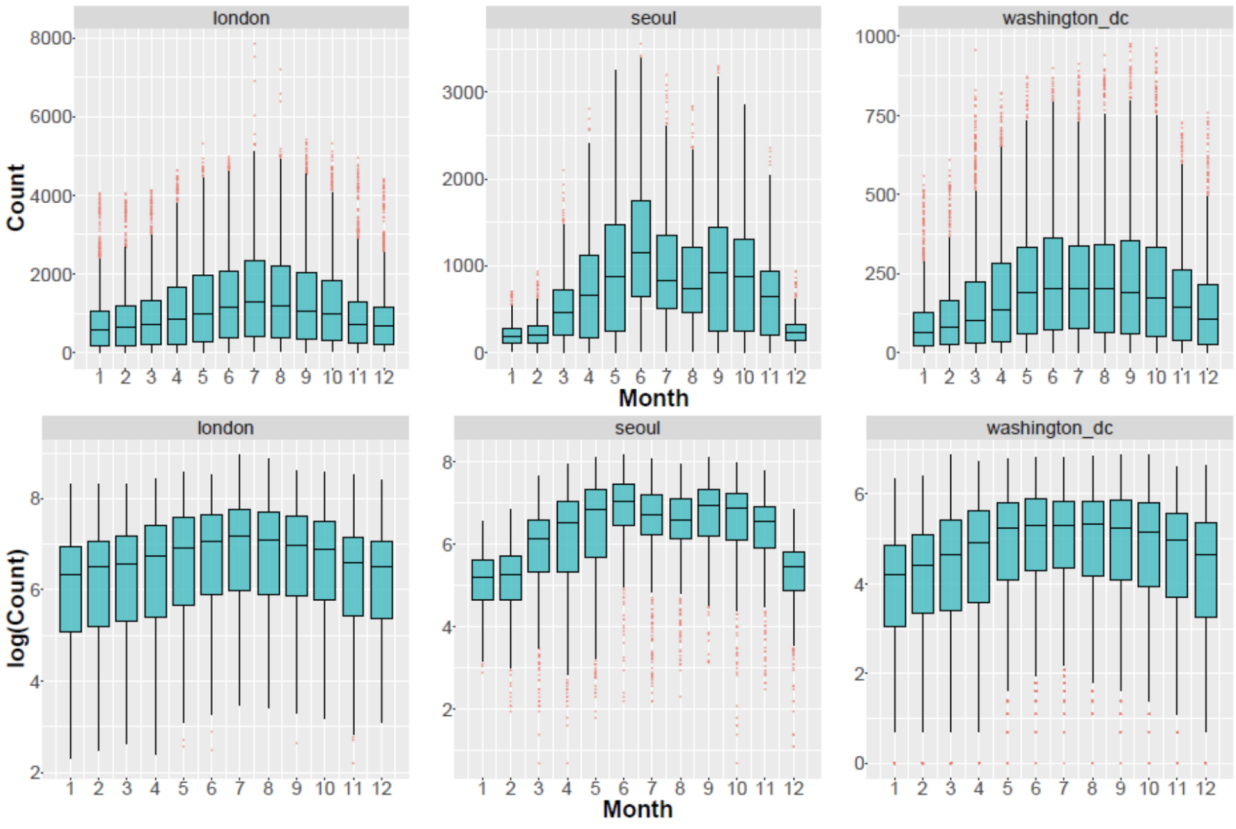
We will be utilizing model selection techniques (CV, AIC/BIC) and shrinkage methods (LASSO) to improve our model fit. We will also be testing various GLM fits (motivations for this listed below) and will be using bootstrapped F-testing to answer the research questions listed above.

We will be modeling our data using the negative binomial, poisson distribution and log-normal distributions.

Motivation for Modeling Tools

Motivation for GLMs

In our EDA, we observed heteroscedasticity in our data, and therefore concluded that the OLS model would not be a good fit for our data. We believe there are reasonable grounds to attempt to model our data using a poisson or negative binomial due to two main reasons. Firstly, our project deals with count-based data. Secondly, in our EDA, we observed that the variance of count in each month increases as mean increases, so a poisson or negative binomial model should fit our data better than an OLS model. With that said, we will also test log-normal distributions because it is possible that the transformed data might be normal, and the variance of log(cnt) are fairly homoscedastic. In fact, after running our tests, the log-ols model has better RMSE and log-likelihood than the poisson and negative binomial models.



Motivation for Model Selection

The inclusion of triple interaction terms in our full model means that all related double interaction terms are also in the model, which might lead to the model having many unrequired terms. Additionally, the triple interaction terms might also not be needed in our model. Since it is difficult to determine heuristically what

terms are important and what are not, given the large number of terms in the model, we will use model selection methods (AIC/BIC and CV) to select the variables to be included in the model. We will also be testing LASSO as a form of model selection, and comparing this against the models selected by AIC/BIC. We note that CV and LASSO are assumption-free model selection criterias, and AIC/BIC relies on our data following the assumed probability distribution. Our end goal is an interpretable model with reasonable performance that can be applied for inference.

Motivation for Nonparametric Bootstrap F-test

The nonparametric bootstrap will be employed on the log-ols model for parametric inference, since we found it to be the model with the best RMSE and log-likelihood after fitting our data. The main reason for using this is because the nonparametric bootstrap makes no assumption about the distribution of our data. A normal F-test might not be reliable because the EDA plot above suggests that even in the log-ols model, there still exists some heteroskedasticity.

Modeling Tool Assumptions

Assumptions for Distributions

OLS/log-OLS Assumptions

- Homoscedastic normal random errors
- Response variable is linear with respect to our data
- Continuous response variable
- We observe that variance of log-normal model is quite similar across months (figure above)

Poisson Assumptions

- Modeling count data (response variable is non-negative integers)
- Conditioning on all the covariates, the response variable (count in our project) should have equal mean and variance. [1].
- There is a linear relationship between the covariates and log(count).

Negative Binomial Assumptions

- Similar assumptions to a poisson model, since the poisson distribution is just a special case of the negative binomial distribution
- Does not make the assumption that the variance of the response variable is equal to the mean, but rather that the variance is proportional to the mean in addition to the inclusion of a dispersion parameter. This more general model might give us a better fit to our data.

Shared Assumptions

For all models, we must ensure the assumption of independence of rows, or independence of different observations, which motivates our decision to treat "hour" as categorical data to remove the row dependence associated with time series data.

[1] We note that this hypothesis is hard to test in our project. Due to the presence of continuous variables (temp, hum, windspeed), when we group the rows by all covariates, we find that none or almost none of the rows have the exact same values for all covariates. We cannot construct a distribution using only one (or very few) data point(s)

Assumptions for model selection and shrinkage methods

AIC/BIC assumptions

AIC and BIC are selection criterias based on the likelihood of the models, therefore the underlying assumption is that our data does indeed follow the probability distribution used in our model. The difference between AIC and BIC is that BIC penalizes the number of parameters more heavily than AIC, hence it will result in a more parsimonious model.

CV/LASSO

CV does not assume anything of our data, and is the most flexible approach, and is an approach meant to minimize the error metric that we define (in this case, the MSE). Lasso also does not assume much about our data and also aims to minimize the MSE, but with an added L1 regularization term on our parameters. Both CV and LASSO cannot tell us anything about the likelihood of our model being correct.

Assumptions of bootstrap and F-tests

In employing the nonparametric bootstrap F-test, we do not make any assumption about the distribution of the population data, but we assume that our data is representative of the population data. This could be a reasonable assumption as our sample size (>43k) is relatively large.

Testing GLMs

Shown below is the data that we are using for our analysis. Our data consists of 43,553 rows.

Table 1: Preview of data used

cnt	temp	hum	windspeed	location	hr	is_workday	season
182	3.0	93.0	6.0	london	0	FALSE	winter
138	3.0	93.0	5.0	london	1	FALSE	winter
134	2.5	96.5	0.0	london	2	FALSE	winter
72	2.0	100.0	0.0	london	3	FALSE	winter
47	2.0	93.0	6.5	london	4	FALSE	winter
46	2.0	93.0	4.0	london	5	FALSE	winter

We first test the various models with cross-validation to observe the errors.

Table 2: CV Train Results

model	rmse	mae	rSquared
ols_simple	568.7073	413.2720	0.5684431
ols_complex	468.8167	338.3552	0.7068644
log_ols_simple	479.5067	261.7669	0.7225723
log_ols_complex	344.4226	189.3736	0.7974307
poisson_simple	424.1024	250.5383	0.7601542
poisson_complex	282.6797	160.2553	0.8935990
nbm_simple	501.9530	278.1413	0.6704287
nbm_complex	385.6975	196.7931	0.8064936

We observe that the models with the triple interaction terms perform better (on RMSE) than the simpler models. This follows our expectations from our EDA where we see different trends when we conditioned on these triple interaction terms. We also see that the RMSE of the poisson model is better than the negative binomial model, which is interesting. This will be explored in the following section.

We train on full data in the following section.

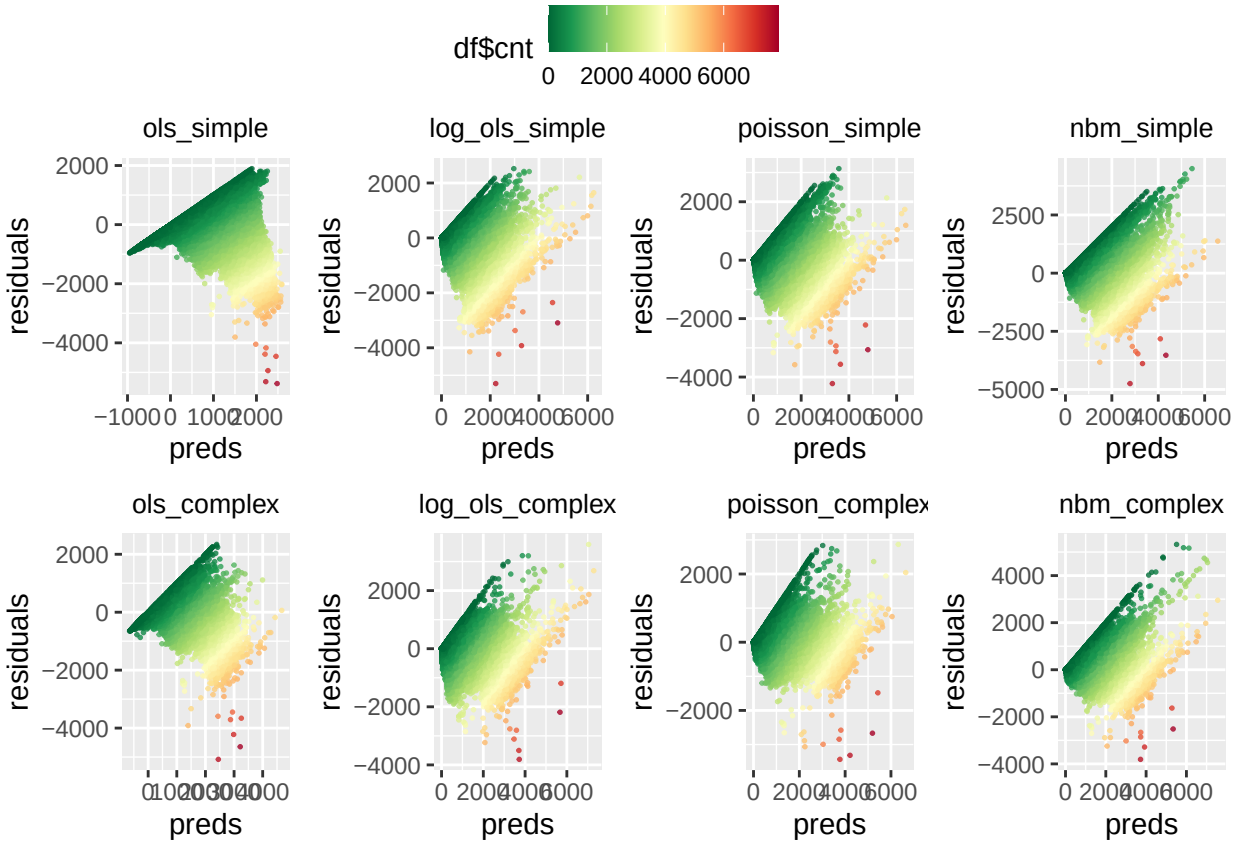
Table 3: Full Data Train Results

model	rmse	mae	loglikelihood
ols_simple	568.4313	412.9621	-338050.50
ols_complex	467.7445	337.4954	-329559.48
log_ols_simple	479.5067	261.7669	-55305.85
log_ols_complex	344.4226	189.3736	-48375.10
poisson_simple	423.6435	250.2636	-3868800.35
poisson_complex	281.5935	159.6263	-2056998.67
nbm_simple	501.6885	277.9098	-290420.22
nbm_complex	384.4704	196.1322	-280575.97

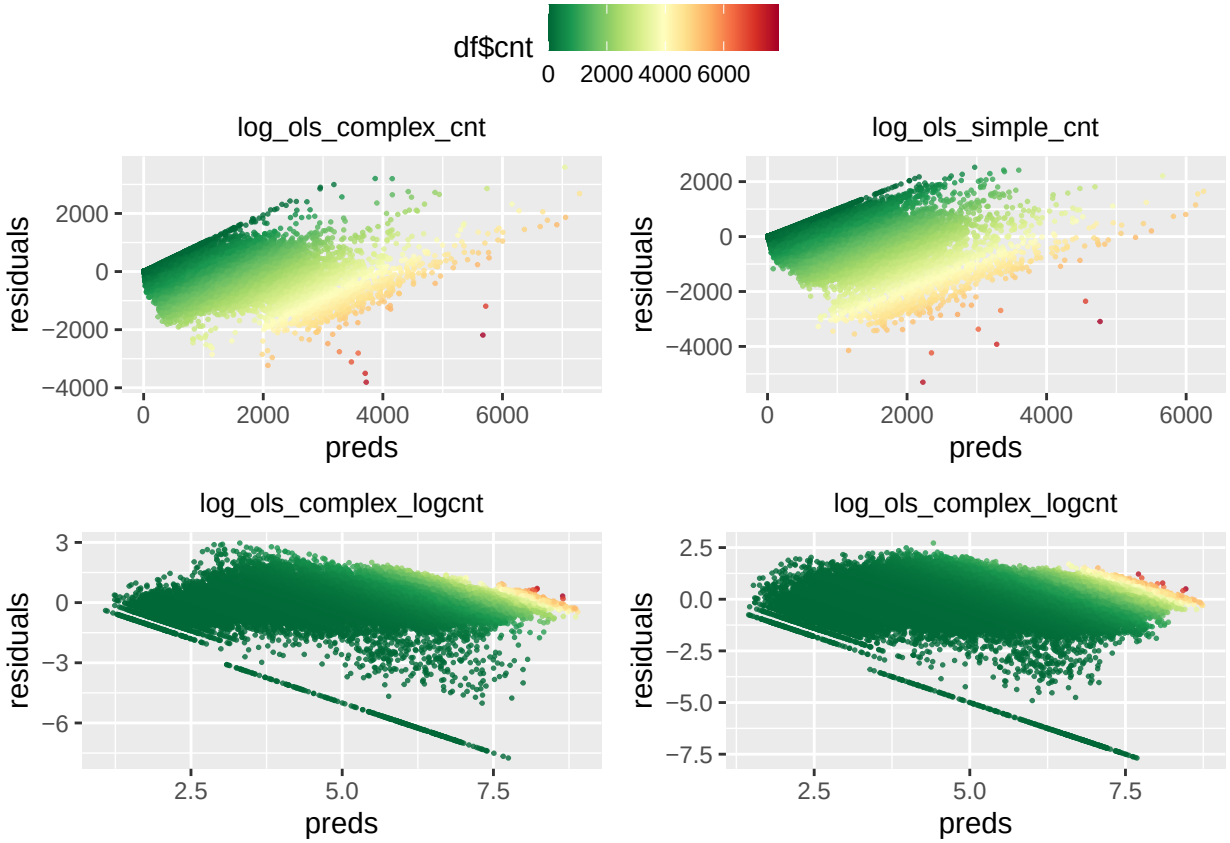
Again, we see that when training on full data, the RMSE (and log-likelihood) of the complex models are better than that of the simpler models. An interesting observation is that even though the RMSE of the poisson models are better than the equivalent negative binomial models, they have much worse log-likelihoods, approx 10x worse. Due to the extremely poor log-likelihood of the poisson model, we will not be using it for further analysis, although we will still investigate the fit of this model.

The best performing model based on log-likelihood is the log-ols model, by a significant margin. It also has the best RMSE and MAE of the models excluding the poisson models. This provides us justification for choosing this over the other models in our further analysis.

The following plots show us the relationship between predicted values and residuals. We observe that residuals are not evenly scattered about 0.



Here's a closer look at the log-ols model, with the fit against cnt and $log(1 + cnt)$ respectively. We observe that besides a group of errors when cnt is extremely low (seen by the negative downwards sloping lines), the log-ols model has a good fit against $log(cnt+1)$. We see that when cnt is extremely low, our model frequently overpredicts the actual value. Our residuals are evenly scattered around the x-axis for our log-normal model, with slight heteroskedasticity.



Model Selection

In the following, we apply both AIC and BIC model selection (both backwards and forwards) to select for a sparser set of variables from the log-ols model. We expect BIC to give us a sparser model because the penalty term of BIC scales with the number of parameters in the model, but AIC does not. Note that both AIC and BIC require a likelihood distribution of our coefficient vector, β, hence the methods we are currently applying are valid only if our assumption that our data follows a log normal distribution is true.

Table 4: Number of coefficients after selection

model	num_coeff
full	120
bic_forward	120
aic_forward	120
bic_backward	97
aic_backward	97
bic_bidirectional	97
aic_bidirectional	97

Table 5: Errors of models

model	rmse	mae
full	344.4226	189.3736
reduced	346.4640	189.7816

Forward selection does not shrink the model at all, while backwards and bidirectional selection shrinks both the AIC and BIC model by the same 23 parameters. The parameters removed are:

```
##  [1] "temp:hr1:is_workdayTRUE"  "temp:hr10:is_workdayTRUE"
##  [3] "temp:hr11:is_workdayTRUE" "temp:hr12:is_workdayTRUE"
##  [5] "temp:hr13:is_workdayTRUE" "temp:hr14:is_workdayTRUE"
##  [7] "temp:hr15:is_workdayTRUE" "temp:hr16:is_workdayTRUE"
##  [9] "temp:hr17:is_workdayTRUE" "temp:hr18:is_workdayTRUE"
## [11] "temp:hr19:is_workdayTRUE" "temp:hr2:is_workdayTRUE"
## [13] "temp:hr20:is_workdayTRUE" "temp:hr21:is_workdayTRUE"
## [15] "temp:hr22:is_workdayTRUE" "temp:hr23:is_workdayTRUE"
## [17] "temp:hr3:is_workdayTRUE"  "temp:hr4:is_workdayTRUE"
## [19] "temp:hr5:is_workdayTRUE"  "temp:hr6:is_workdayTRUE"
## [21] "temp:hr7:is_workdayTRUE"  "temp:hr8:is_workdayTRUE"
## [23] "temp:hr9:is_workdayTRUE"
```

The selection criteria means that the triple interaction term of $temp * hr * isWorkday$ is removed. The resulting model is:

```
##
## Call:
## lm(formula = log(cnt + 1) ~ temp + hum + windspeed + location +
##     hr + is_workday + season + temp:location + temp:season +
##     location:season + temp:hr + temp:is_workday + hr:is_workday +
##     temp:location:season, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7459 -0.2280  0.0554  0.3143  2.9690
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             6.7697412  0.0588389 115.056  < 2e-16
## temp                                    0.0544355  0.0034378  15.835  < 2e-16
## hum                                    -0.0122013  0.0002470 -49.399  < 2e-16
## windspeed                              -0.0106291  0.0005205 -20.422  < 2e-16
## locationseoul                          -0.1643099  0.0509350  -3.226 0.001257
## locationwashington_dc                  -1.9612318  0.0463642 -42.301  < 2e-16
## hr1                                    -0.1160606  0.0565480  -2.052 0.040134
## hr10                                    0.8217056  0.0573574  14.326  < 2e-16
## hr11                                    0.9550652  0.0583493  16.368  < 2e-16
## hr12                                    1.0723730  0.0591720  18.123  < 2e-16
## hr13                                    1.0771867  0.0598455  17.999  < 2e-16
## hr14                                    1.0478587  0.0602396  17.395  < 2e-16
## hr15                                    1.0296611  0.0602800  17.081  < 2e-16
## hr16                                    0.9490423  0.0600145  15.814  < 2e-16
```

```
## hr17                                  0.8578888  0.0593333  14.459  < 2e-16
## hr18                                  0.7202302  0.0587153  12.266  < 2e-16
## hr19                                  0.4815379  0.0583169   8.257  < 2e-16
## hr2                                  -0.3331057  0.0564403  -5.902 3.62e-09
## hr20                                  0.2338885  0.0579714   4.035 5.48e-05
## hr21                                  0.0777806  0.0576330   1.350 0.177157
## hr22                                 -0.0114547  0.0573439  -0.200 0.841673
## hr23                                 -0.1925318  0.0571124  -3.371 0.000749
## hr3                                  -0.7624444  0.0563113 -13.540  < 2e-16
## hr4                                  -1.4151817  0.0561570 -25.200  < 2e-16
## hr5                                  -1.6486289  0.0560336 -29.422  < 2e-16
## hr6                                  -1.2882450  0.0556472 -23.150  < 2e-16
## hr7                                  -0.6129186  0.0555053 -11.043  < 2e-16
## hr8                                   0.1598583  0.0557509   2.867 0.004141
## hr9                                   0.6078067  0.0563733  10.782  < 2e-16
## is_workdayTRUE                       -0.5833341  0.0388437 -15.017  < 2e-16
## seasonspring                         -0.2405095  0.0478087  -5.031 4.91e-07
## seasonsummer                          0.0346268  0.0689781   0.502 0.615672
## seasonwinter                         -0.0983043  0.0440845  -2.230 0.025759
## temp:locationseoul                   -0.0645609  0.0034443 -18.744  < 2e-16
## temp:locationwashington_dc           -0.0005295  0.0031642  -0.167 0.867102
## temp:seasonspring                     0.0137525  0.0038130   3.607 0.000310
## temp:seasonsummer                    -0.0041461  0.0040793  -1.016 0.309458
## temp:seasonwinter                    -0.0036381  0.0039891  -0.912 0.361771
## locationseoul:seasonspring           -0.3379634  0.0687723  -4.914 8.95e-07
## locationwashington_dc:seasonspring   -0.3488032  0.0621732  -5.610 2.03e-08
## locationseoul:seasonsummer            1.6742370  0.1194964  14.011  < 2e-16
## locationwashington_dc:seasonsummer    0.4305932  0.1024900   4.201 2.66e-05
## locationseoul:seasonwinter           -0.9050965  0.0591248 -15.308  < 2e-16
## locationwashington_dc:seasonwinter   -0.4784144  0.0552474  -8.659  < 2e-16
## temp:hr1                             -0.0078394  0.0030368  -2.581 0.009841
## temp:hr10                            -0.0150473  0.0029198  -5.154 2.57e-07
## temp:hr11                            -0.0128706  0.0029280  -4.396 1.11e-05
## temp:hr12                            -0.0129111  0.0029284  -4.409 1.04e-05
## temp:hr13                            -0.0123243  0.0029304  -4.206 2.61e-05
## temp:hr14                            -0.0113808  0.0029303  -3.884 0.000103
## temp:hr15                            -0.0103516  0.0029257  -3.538 0.000403
## temp:hr16                            -0.0069023  0.0029198  -2.364 0.018085
## temp:hr17                            -0.0050416  0.0029084  -1.733 0.083026
## temp:hr18                            -0.0018524  0.0029101  -0.637 0.524410
## temp:hr19                             0.0048809  0.0029387   1.661 0.096737
## temp:hr2                             -0.0115022  0.0030488  -3.773 0.000162
## temp:hr20                             0.0085489  0.0029663   2.882 0.003953
## temp:hr21                             0.0086560  0.0029904   2.895 0.003798
## temp:hr22                             0.0070714  0.0030043   2.354 0.018590
## temp:hr23                             0.0047128  0.0030195   1.561 0.118587
## temp:hr3                             -0.0155193  0.0030675  -5.059 4.23e-07
## temp:hr4                             -0.0170218  0.0030778  -5.531 3.21e-08
## temp:hr5                             -0.0097675  0.0030663  -3.185 0.001446
## temp:hr6                             -0.0064566  0.0030466  -2.119 0.034073
## temp:hr7                             -0.0111373  0.0029996  -3.713 0.000205
## temp:hr8                             -0.0191019  0.0029493  -6.477 9.47e-11
## temp:hr9                             -0.0203080  0.0029237  -6.946 3.81e-12
## temp:is_workdayTRUE                  -0.0141561  0.0009067 -15.613  < 2e-16
```

```
## hr1:is_workdayTRUE                       -0.3363352  0.0526574  -6.387 1.71e-10
## hr10:is_workdayTRUE                       0.4839443  0.0526703   9.188  < 2e-16
## hr11:is_workdayTRUE                       0.3097385  0.0526724   5.880 4.12e-09
## hr12:is_workdayTRUE                       0.3559695  0.0526724   6.758 1.41e-11
## hr13:is_workdayTRUE                       0.3167532  0.0527045   6.010 1.87e-09
## hr14:is_workdayTRUE                       0.2562131  0.0527305   4.859 1.18e-06
## hr15:is_workdayTRUE                       0.3389743  0.0527352   6.428 1.31e-10
## hr16:is_workdayTRUE                       0.6770539  0.0527164  12.843  < 2e-16
## hr17:is_workdayTRUE                       1.2995055  0.0526933  24.662  < 2e-16
## hr18:is_workdayTRUE                       1.4452083  0.0526796  27.434  < 2e-16
## hr19:is_workdayTRUE                       1.2174493  0.0526660  23.116  < 2e-16
## hr2:is_workdayTRUE                       -0.5514573  0.0527506 -10.454  < 2e-16
## hr20:is_workdayTRUE                       1.0981426  0.0526395  20.862  < 2e-16
## hr21:is_workdayTRUE                       1.0373993  0.0526288  19.712  < 2e-16
## hr22:is_workdayTRUE                       0.9544270  0.0526210  18.138  < 2e-16
## hr23:is_workdayTRUE                       0.8059833  0.0526648  15.304  < 2e-16
## hr3:is_workdayTRUE                       -0.4725656  0.0528215  -8.946  < 2e-16
## hr4:is_workdayTRUE                        0.1953792  0.0528245   3.699 0.000217
## hr5:is_workdayTRUE                        1.2864311  0.0528152  24.357  < 2e-16
## hr6:is_workdayTRUE                        2.2721827  0.0526640  43.145  < 2e-16
## hr7:is_workdayTRUE                        2.6684124  0.0526345  50.697  < 2e-16
## hr8:is_workdayTRUE                        2.5471412  0.0526438  48.384  < 2e-16
## hr9:is_workdayTRUE                        1.4238734  0.0526214  27.059  < 2e-16
## temp:locationseoul:seasonspring           0.0455448  0.0050118   9.088  < 2e-16
## temp:locationwashington_dc:seasonspring   0.0111619  0.0045257   2.466 0.013655
## temp:locationseoul:seasonsummer          -0.0119255  0.0057152  -2.087 0.036926
## temp:locationwashington_dc:seasonsummer -0.0257661  0.0051660  -4.988 6.13e-07
## temp:locationseoul:seasonwinter           0.0600855  0.0054206  11.085  < 2e-16
## temp:locationwashington_dc:seasonwinter   0.0387589  0.0049287   7.864 3.81e-15
##
## (Intercept)                              ***
## temp                                     ***
## hum                                      ***
## windspeed                                ***
## locationseoul                            **
## locationwashington_dc                    ***
## hr1                                      *
## hr10                                     ***
## hr11                                     ***
## hr12                                     ***
## hr13                                     ***
## hr14                                     ***
## hr15                                     ***
## hr16                                     ***
## hr17                                     ***
## hr18                                     ***
## hr19                                     ***
## hr2                                      ***
## hr20                                     ***
## hr21
## hr22
## hr23                                     ***
## hr3                                      ***
## hr4                                      ***
```

```
## hr5                                  ***
## hr6                                  ***
## hr7                                  ***
## hr8                                  **
## hr9                                  ***
## is_workdayTRUE                       ***
## seasonspring                         ***
## seasonsummer
## seasonwinter                         *
## temp:locationseoul                   ***
## temp:locationwashington_dc
## temp:seasonspring                    ***
## temp:seasonsummer
## temp:seasonwinter
## locationseoul:seasonspring           ***
## locationwashington_dc:seasonspring   ***
## locationseoul:seasonsummer           ***
## locationwashington_dc:seasonsummer   ***
## locationseoul:seasonwinter           ***
## locationwashington_dc:seasonwinter   ***
## temp:hr1                             **
## temp:hr10                            ***
## temp:hr11                            ***
## temp:hr12                            ***
## temp:hr13                            ***
## temp:hr14                            ***
## temp:hr15                            ***
## temp:hr16                            *
## temp:hr17                            .
## temp:hr18
## temp:hr19                            .
## temp:hr2                             ***
## temp:hr20                            **
## temp:hr21                            **
## temp:hr22                            *
## temp:hr23
## temp:hr3                             ***
## temp:hr4                             ***
## temp:hr5                             **
## temp:hr6                             *
## temp:hr7                             ***
## temp:hr8                             ***
## temp:hr9                             ***
## temp:is_workdayTRUE                  ***
## hr1:is_workdayTRUE                   ***
## hr10:is_workdayTRUE                  ***
## hr11:is_workdayTRUE                  ***
## hr12:is_workdayTRUE                  ***
## hr13:is_workdayTRUE                  ***
## hr14:is_workdayTRUE                  ***
## hr15:is_workdayTRUE                  ***
## hr16:is_workdayTRUE                  ***
## hr17:is_workdayTRUE                  ***
## hr18:is_workdayTRUE                  ***
```

```
## hr19:is_workdayTRUE                     ***
## hr2:is_workdayTRUE                      ***
## hr20:is_workdayTRUE                     ***
## hr21:is_workdayTRUE                     ***
## hr22:is_workdayTRUE                     ***
## hr23:is_workdayTRUE                     ***
## hr3:is_workdayTRUE                      ***
## hr4:is_workdayTRUE                      ***
## hr5:is_workdayTRUE                      ***
## hr6:is_workdayTRUE                      ***
## hr7:is_workdayTRUE                      ***
## hr8:is_workdayTRUE                      ***
## hr9:is_workdayTRUE                      ***
## temp:locationseoul:seasonspring         ***
## temp:locationwashington_dc:seasonspring *
## temp:locationseoul:seasonsummer         *
## temp:locationwashington_dc:seasonsummer ***
## temp:locationseoul:seasonwinter         ***
## temp:locationwashington_dc:seasonwinter ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7359 on 43456 degrees of freedom
## Multiple R-squared:  0.7983, Adjusted R-squared:  0.7979
## F-statistic:  1792 on 96 and 43456 DF,  p-value: < 2.2e-16
```

Most of our covariates have very significant coefficients, but we note that the following covariates do not have significant p-values. We will run a F-test to determine if they are significant as a group.

Table 6: Insignificant Coefficients

	pvals
hr21	0.1771565
hr22	0.8416732
seasonsummer	0.6156720
temp:locationwashington_dc	0.8671018
temp:seasonsummer	0.3094583
temp:seasonwinter	0.3617705
temp:hr18	0.5244099
temp:hr23	0.1185874

We also test variable selection with LASSO. Choosing $\lambda_{1\sigma}$, we find that lasso shrinks the model by the following 48 variables.

```
##  [1] "(Intercept)"
##  [2] "hr22"
##  [3] "hr23"
##  [4] "hr8"
##  [5] "seasonsummer"
##  [6] "temp:locationwashington_dc"
##  [7] "temp:seasonsummer"
##  [8] "temp:seasonwinter"
```

```
##  [9] "locationseoul:seasonspring"
## [10] "locationwashington_dc:seasonsummer"
## [11] "temp:hr1"
## [12] "temp:hr10"
## [13] "temp:hr11"
## [14] "temp:hr12"
## [15] "temp:hr13"
## [16] "temp:hr14"
## [17] "temp:hr15"
## [18] "temp:hr16"
## [19] "temp:hr2"
## [20] "temp:hr3"
## [21] "temp:hr5"
## [22] "temp:hr6"
## [23] "temp:hr7"
## [24] "temp:hr8"
## [25] "temp:hr9"
## [26] "hr10:is_workdayTRUE"
## [27] "hr11:is_workdayTRUE"
## [28] "hr12:is_workdayTRUE"
## [29] "hr13:is_workdayTRUE"
## [30] "hr14:is_workdayTRUE"
## [31] "hr15:is_workdayTRUE"
## [32] "temp:locationwashington_dc:seasonspring"
## [33] "temp:locationseoul:seasonsummer"
## [34] "temp:hr1:is_workdayTRUE"
## [35] "temp:hr10:is_workdayTRUE"
## [36] "temp:hr11:is_workdayTRUE"
## [37] "temp:hr12:is_workdayTRUE"
## [38] "temp:hr13:is_workdayTRUE"
## [39] "temp:hr15:is_workdayTRUE"
## [40] "temp:hr17:is_workdayTRUE"
## [41] "temp:hr18:is_workdayTRUE"
## [42] "temp:hr19:is_workdayTRUE"
## [43] "temp:hr2:is_workdayTRUE"
## [44] "temp:hr20:is_workdayTRUE"
## [45] "temp:hr21:is_workdayTRUE"
## [46] "temp:hr4:is_workdayTRUE"
## [47] "temp:hr8:is_workdayTRUE"
## [48] "temp:hr9:is_workdayTRUE"
```

However, this model shrinkage does not result in an easily understood model, and as seen below, the errors are significantly worse than the errors of the models selected by AIC/BIC. We will therefore not use this model.

Table 7: Errors of LASSO 1se model

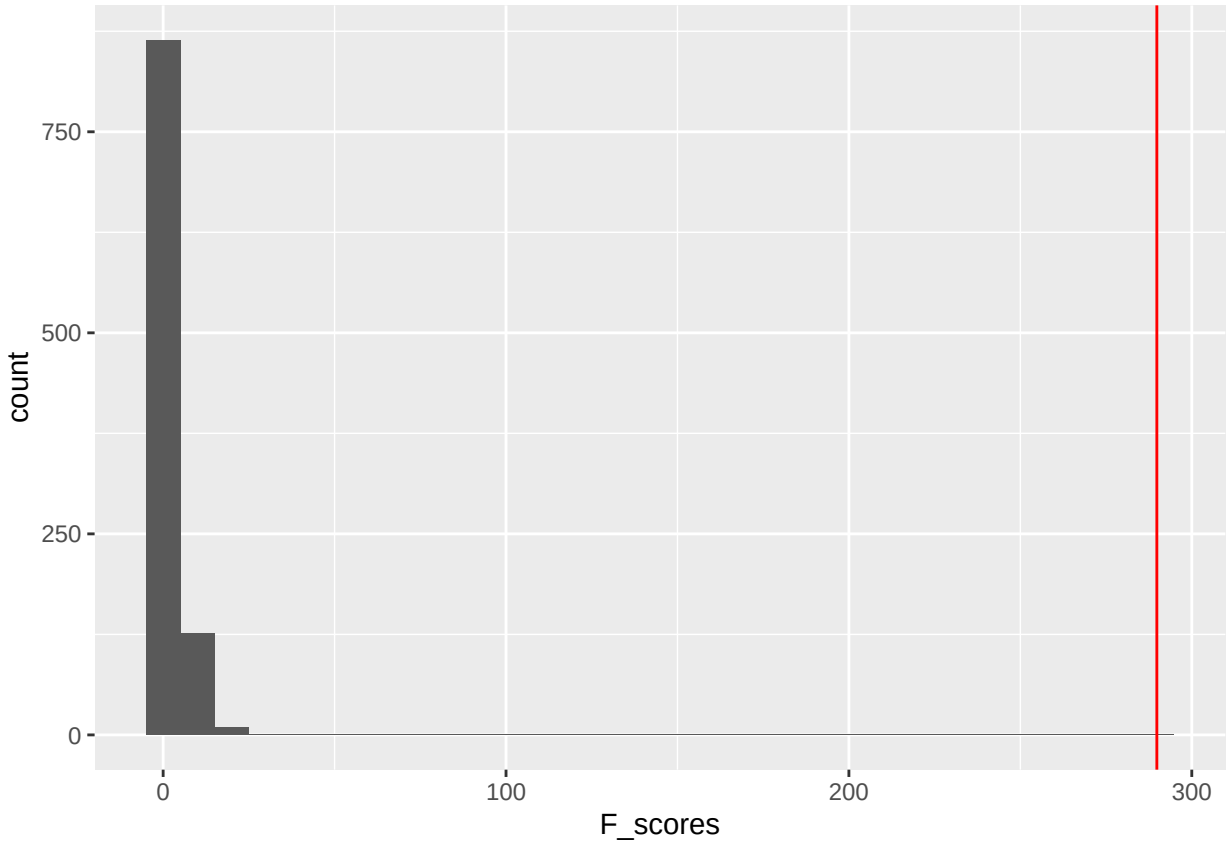
rmse	mae
905.8535	574.0437

Bootstrap F-test

Due to heteroskedasticity in our data, we will be employing the bootstrap F-test to answer our research questions.

Question 1: Is the effect of temperature different across the locations? We expect people in different areas to have different temperature preferences.

To test this hypothesis, we want to test whether the betas for temp:locationseoul and temp:locationwashington_dc are 0.

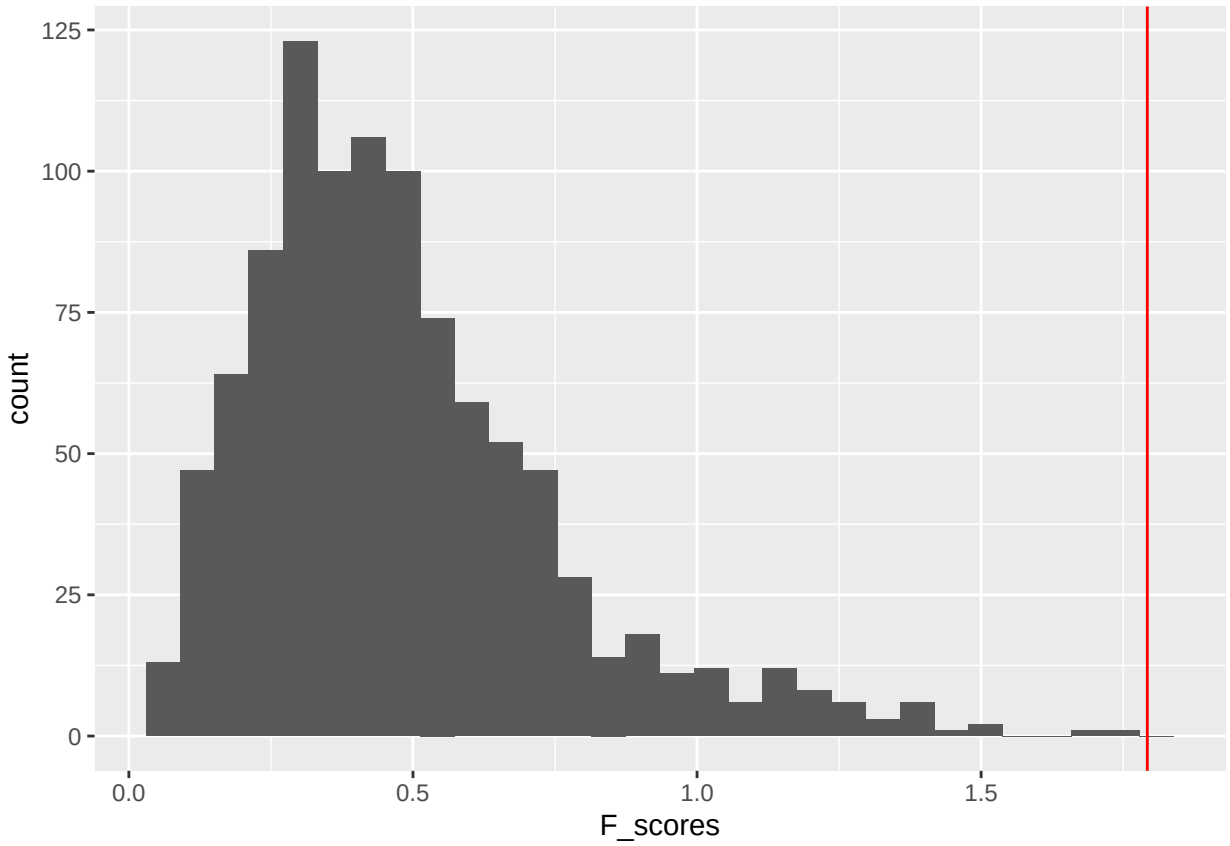



```
## P value of F-test temp:locationseoul = temp:locationwashington_dc = 0: 0
```

In the plot, the red vertical line represents $F(\hat{\beta}, \beta_0)$, which has a value of 289.8147 and is far away from the values of $(F(\hat{\beta_1}, \hat{\beta}), F(\hat{\beta_2}, \hat{\beta}), ..., F(\hat{\beta_B}, \hat{\beta}))$.

Our p value from bootstrap (setting B=1000) is 0, so we see that the interaction term $temp * location$ is significant, proving that our hypothesis was right.

Question 2: Are the covariates from our model that had low p-values insignificant when tested as a group?




```
## P value of F-test all insignficant betas = 0: 0
```

For this test, $F(\hat{\beta}, \beta_0) = 1.792518$, but the p value from bootstrap is also 0. Seeing that the p value from bootstrap is 0, we can also conclude that the the beta values β_{hr21}, β_{hr22}, $\beta_{seasonsummer}$, $\beta_{temp:locationwashington_dc}$, $\beta_{temp:seasonsummer}$, $\beta_{temp:seasonwinter}$, $\beta_{temp:hr18}$, $\beta_{temp:hr23}$ are significant as a group and reject the null hypothesis that all these coefficients are actually 0.

Furthermore, it is interesting to see the difference in p-value (and conclusion) one can obtain via running a nonparametric bootstrap F-test. Should we have assumed that our log-ols model satisfied the canonical assumptions of normal, homoscedastic errors, we would have computed our p-value as $P(F_{8,43456} > 1.792518) = 0.07335$, and we would not have been able to reject the null hypothesis at $alpha = 0.05$.

Conclusion

It is not surprising that the negative binomial distribution is a better fit than the ols model, but it is somewhat surprising that our data follows the log normal distribution more closely than both the poisson and negative binomial distributions. However, this is still reasonable since the log normal model also always gives us non-negative values, unlike the ols model.

One mode of errors observed in the log-ols model is that when the actual values are extremely low, but the model still predicts high values. We could possibly address this issue by implementing a 2 stage model, the first stage being a logistic regression to determine whether the actual output should be below some threshold

value, and the second stage being the models we tested. We did preliminary testing of this approach, and found that logistic regression works poorly, and more sophisticated methods might be required.

In regards to our research questions, we were able to conclude that our hypothesis that β_{temp} is different in different locations, at least for the fall season by using a nonparametric bootstrap f-test to determine that the interaction terms $\beta_{temp:locationseoul}$ and $\beta_{temp:locationwashingtonDC}$ are not 0. We also were able to conclude with a nonparametric bootstrap f-test that the covariates found to be insignificant in the model selected by BIC bidirectional selection are actually in fact significant as a group.