Lab Report 3: Multiple Linear Regression

MA 575 Fall 2021 - C3 Team #2

Ali Taqi, Hsin-Chang Lin, Huiru Yang, Ryan Mahoney, Yulin Li

11/15/2021

1 Abstract

The prevalence of bikesharing has been steadily increasing over the years; as the industry grows, there is an increasing interest in predicting the daily frequency in which users utilize these bikesharing services. Using the data provided by UIAAD, we seek to use the daily counts and a variety of seasonal and envrionmental predictors to estimate 2012 counts by using 2011 data. By selecting two sets of predictors (one simple, one complex), an OLS model is fit to demonstrate the reliability of a core set of predictors: `atemp`, `atemp^2`, `weathersit`, `windspeed`, and `hum`. Additionally, a technique of "growth-discounting" response variables on the testing data is motivated in order to solve the issue of non-constant size of the bikesharing user-base. Significant improvements in testing performances (11-12x) were observed when the response was adjusted. Small, but considerable improvements (0.4-4.6%) were also observed when increasing the complexity of the model, indicating room for more complex models.

2 Introduction

In this lab report, we narrow our scope from last time and perform Multiple Linear Regression (MLR) on the `cnt` variable and a selected subset of predictors chosen from the 2011 Bike Sharing dataset [1]. Then, we will be testing our model fit of the 2011 data on the 2012 data.

The model should help answer the following question: what are the **daily** bike rentals under different conditions? Business owners may like to know the daily bike rentals in 2012 so that they could optimize the inventory to reduce costs, and they may also wonder whether it is worth leaving the bike-sharing system open on days with extreme weather conditions. This can be done by performing predictive modeling on the daily rental variable based on data given in 2011.

3 Background

First, we take a peek at our dataset In the appendix, attached is the explanation of the column names. The response variable is `cnt`, the number of users. Our selection of predictors include three numerical predictors: `atemp`, `hum`, `windspeed` and one categorical: `weathersit`. All our chosen predictors are weather-related. That being said, here is a sample of our data to obtain a sense of the values.

```
##   weathersit    atemp      hum windspeed  cnt
## 1          2 0.363625 0.805833  0.160446  985
## 2          2 0.353739 0.696087  0.248539  801
## 3          1 0.189405 0.437273  0.248309 1349
```

Preprocessing: Data Type & Value Conversion

Next, we need to preprocess our data before we can fit our models. We mainly do two things: (i) perform a few type changes on our predictors, and (ii) rescale our numerical predictors for interpretability. For the

purposes of being concise, we only cover the variables we end up using.

Typically, all variables whose numerical values are not attached to actual physical meanings are treated as **categorical** variables. Note that for the `weathersit` variable, weather conditions "worsen" as the index gets higher. Furthermore, the normalized weather condition measurements are also converted to their original values, so that the numerical values being used "make more sense" to us. This makes it easier for commonsense and real-life experience to be applied in later analysis.

```
# Other categorical variables (from int to factor type)
weathersit <- as.factor(bikedata$weathersit)     #1 to 4
# Re-scale the normalized measurements
temp <- bikedata$temp * 41
atemp <- bikedata$atemp * 50
hum <- bikedata$hum * 100
windspeed <- bikedata$windspeed * 67
```

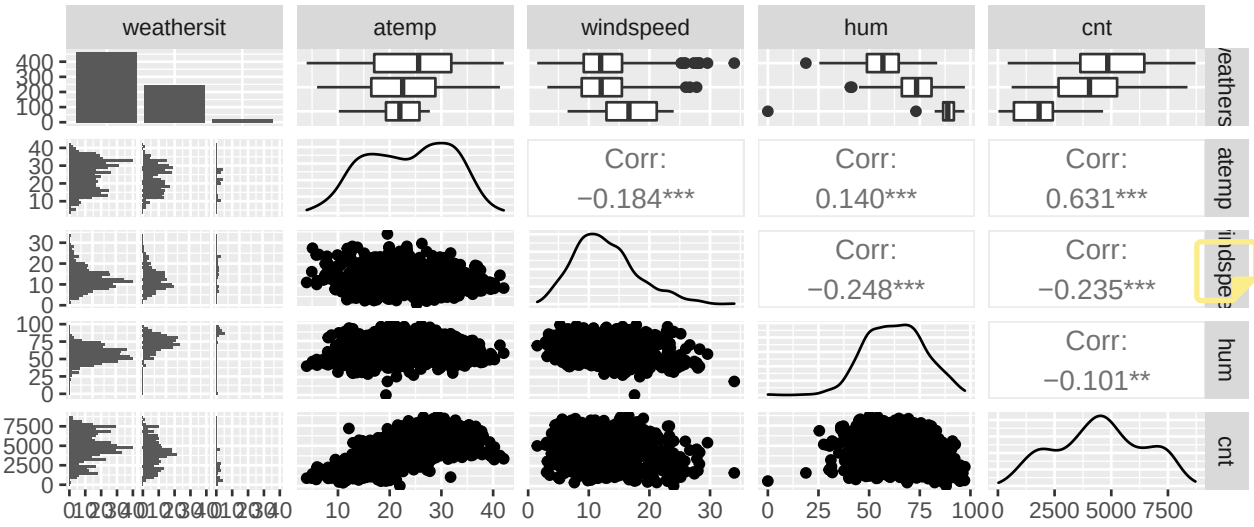
4 Modeling & Analysis

4.1 Variable Selection

The first step to fitting a model starts with selecting our variables. To keep it simple, we decide to fit a full preliminary model and select a core set of very important, statistically significant variables. Of course, since models of high complexity are prone to overfitting, we do not get "fooled" by the low p-values on predictors we expect to have little predictive power.

To stay concise, we deicded to omit the time-class variables (due to autocorrelation and generally being inappopriate for our OLS model) and the variables `workingday`, `weekday`, and `holiday` due to their weak predictive power and general instability. Note that `temp` is almost perfectly correlated to `atemp`, and is thus omitted.

4.2 Pairwise Variable Plot



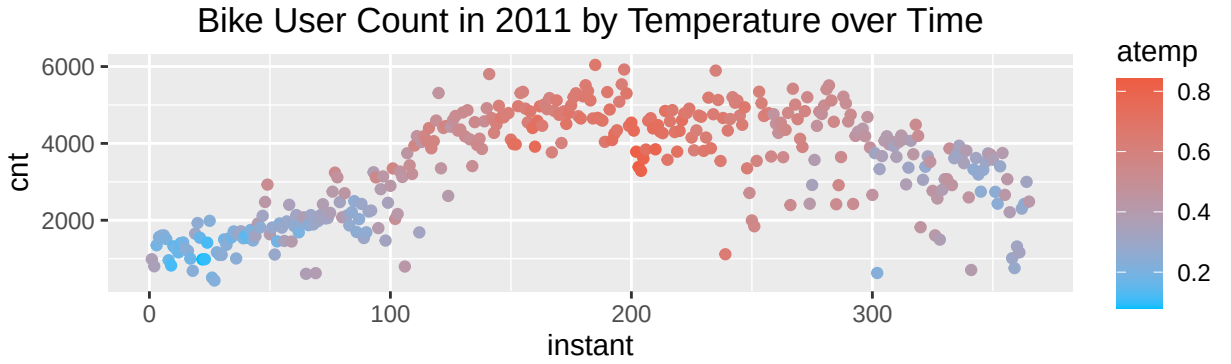
As we can see we have a wide array of variables with reasonable amounts of explanatory power. That being said, we will go forward with two classes of models: a simple model, and a complex model. As we know, simple models tend to perform better since there is much less risk of overfitting to the data as exists in the complex model paradigm. That being said, we propose the following models by variable subset:

Simple Model: `cnt ~ weathersit + atemp + atemp^2`

Complex Model: `cnt ~ weathersit + atemp + atemp^2 + windspeed + hum`

In both models, we use the same core subset of variables containing "most" of the explanatory power: `weathersit`, and `atemp` (along with its second power) since we expect them to encode the most predictive power. This decision was made primarily to be conservative regarding model complexity for the reasons mentioned above (regarding testing performance). Additionally, one might surmise the predictive power of `weathersit` may be somewhat correlated to `windspeed` and `hum`; as such, to avoid the risk of any collinearity, we distinguish it from the simple model.

Additionally, with regards to the second order term of `atemp`, we choose to confidently always use the quadratic model from preliminary findings in Lab Report 2, where we found that temperature and count tended to have a parabolic shape. Choosing a linear model wouldn't capture the full complexity, and choosing orders of 3 or above would be prone to overfitting. The plot below demonstrates the parabolic relation mentioned earlier.



4.3 The Response Variable

4.3.1 Accounting for Growth in Bikeshare Users

Recall that our response variable is `cnt`, the count of users on a given day. There is a bit of a problem if we are to use the 2011 data as training data and 2012 as testing data. Namely, it assumes the number of users stays constant, which is not necessarily true. **This is critical to account for since our response variable is a raw count**. As such, treating the 2011 and 2012 data disjointly may be appropriate. Let's do a hypothesis test to see if $\mu_{2011} - \mu_{2012} = 0$. Indeed, as we see below we find evidence suggesting there is a significant difference between the two populations.

```
##
##  Welch Two Sample t-test
##
## data:  data2011$cnt and data2012$cnt
## t = -18.578, df = 685.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2426.069 -1962.276
## sample estimates:
## mean of x mean of y
##  3405.762  5599.934
```

That being said, we must account for the growth in the number of users between 2011-12. This is because in theory, the true parameter estimate $\hat{\beta}_{atemp}$ should be independent of how many users there. As such, one solution to consider is **discounting the 2012 responses by the average user growth observed between 2011-12, which we call** r. That being said, the factor r is given by $r = \frac{\mu_{2011}}{\mu_{2012}}$, where μ_Y is the average `cnt` in year Y.

To summarize, we want to test the model against the "discounted" response variables $\tilde{c}_i = r \cdot c_i$. While this

solution is elegant, there is one clear issue. Validating on 2012 data implies we do not have μ_{2012}, and thus the factor r. This means we must somehow estimate r without the testing data. The estimation of r is not within the scope of this report. To justify this cost, we demonstrate the benefits of estimating r by observing the effects on $\hat{\beta}$ assuming we perfectly estimate r. In our data, we observe $r = 0.608$, meaning the 2011 bikesharing community is about 60.8% the size of the 2012 community.

4.3.2 Parameter Estimate Stability

Now, we fit the same (simple) model on the training data and the testing datasets (adjusted and unadjusted response). By looking at the parameter estimates, we indeed find the parameter fits on the adjusted-response 2012 data resoundingly agrees with the 2011 estimates relative to the unadjusted-response 2012 data.

```
# Fit model to 2011 data (as a baseline for parameter estimates)
m0$coefficients
```

```
##  (Intercept)        atemp   I(atemp^2)  weathersit2  weathersit3
## -1348.564979   341.949059    -4.809216  -550.400289 -1906.670408
```

```
# Fit model to adjusted response variable
m1$coefficients
```

```
##  (Intercept)        atemp   I(atemp^2)  weathersit2  weathersit3
## -1429.131797   357.467345    -5.583859  -486.128275 -1878.708135
```

```
# Fit model to unadjusted response variable
m2$coefficients
```

```
##  (Intercept)        atemp   I(atemp^2)  weathersit2  weathersit3
## -2349.854507   587.766820    -9.181278  -799.317964 -3089.071833
```

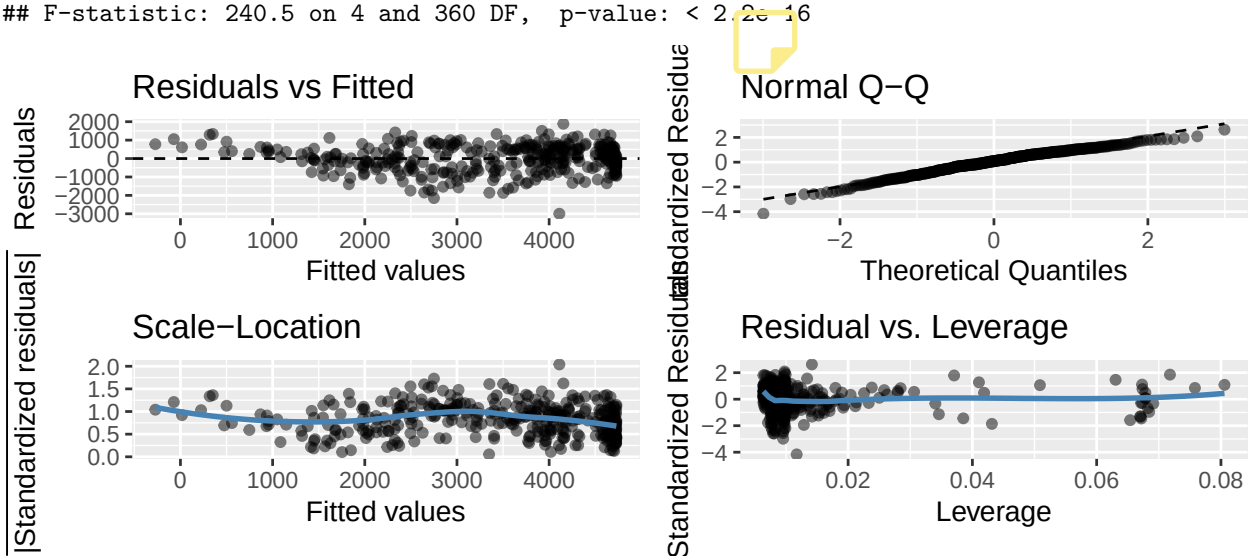
4.4 Models

Below, we fit all the models on the 2011 data and obtain our parameter estimates. We find that both models are statistically significant given their F-statistics yield near-zero p-values. All parameters across both models are statistically significant, with the exception of $\hat{\beta}_0$ in the complex model, which is a curious fact. As expected, the more complex model had a higher R^2 value of 0.75 compared to 0.72. However, given that we have added two predictors, this doesn't necessarily imply better testing performance. On the same note, we note $RSE_1 = 723.5$ and $RSE_2 = 688.1$ indicating the RSE diminishes with model complexity as we expect. However, this is most likely suspect to the training overfitting, which we will verify in the next sections.

Model 1: Simple Model

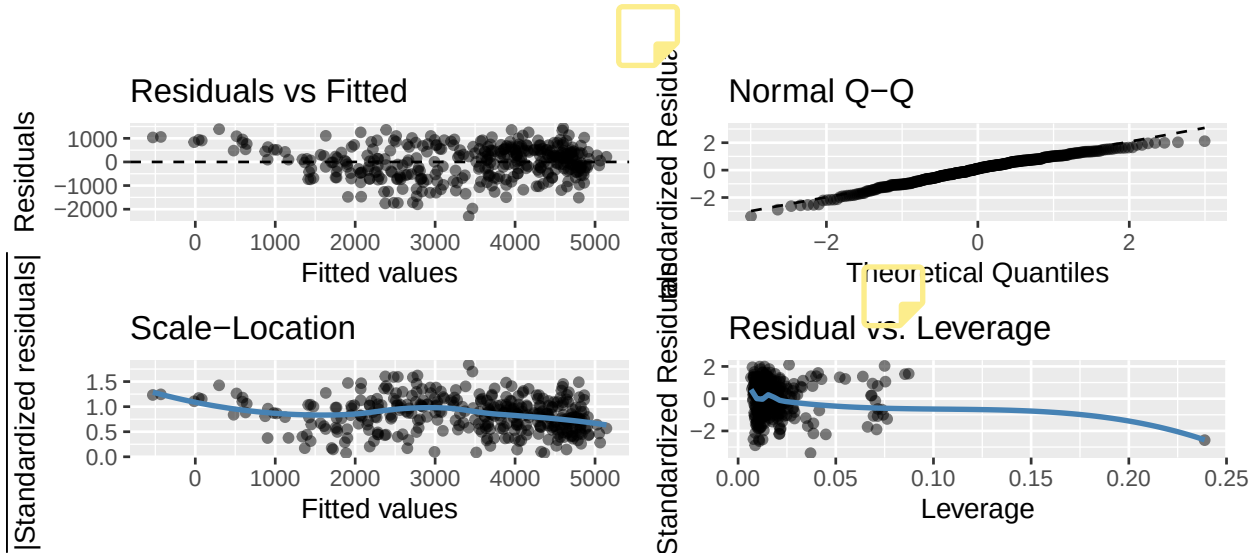
```
##
## Call:
## lm(formula = cnt ~ weathersit + atemp + I(atemp^2), data = data2011)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2995.94  -454.83    58.33   533.34  1888.59
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1348.5650   268.1421  -5.029 7.78e-07 ***
## weathersit2  -550.4003    82.2339  -6.693 8.39e-11 ***
## weathersit3 -1906.6704   195.2411  -9.766  < 2e-16 ***
## atemp         341.9491    25.4663  13.427  < 2e-16 ***
## I(atemp^2)     -4.8092     0.5498  -8.746  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 723.5 on 360 degrees of freedom
## Multiple R-squared:  0.7277, Adjusted R-squared:  0.7247
## F-statistic: 240.5 on 4 and 360 DF,  p-value: < 2.2e-16
```



Model 2: Complex Model

```
##
## Call:
## lm(formula = cnt ~ weathersit + atemp + I(atemp^2) + windspeed +
##     hum, data = data2011)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2304.89  -433.60    72.38   498.40  1441.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -336.2996   309.0530  -1.088 0.277256
## weathersit2  -360.3435    94.9620  -3.795 0.000174 ***
## weathersit3 -1486.7769   210.7201  -7.056 8.94e-12 ***
## atemp         361.1477    25.3330  14.256  < 2e-16 ***
## I(atemp^2)     -5.2218     0.5409  -9.653  < 2e-16 ***
## windspeed     -46.8690     7.5397  -6.216 1.42e-09 ***
## hum           -10.6740     3.3820  -3.156 0.001734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 688.1 on 358 degrees of freedom
## Multiple R-squared:  0.7551, Adjusted R-squared:  0.7509
## F-statistic: 183.9 on 6 and 358 DF,  p-value: < 2.2e-16
```



With regards to diagnostics, both models seem to be reasonably well-behaved. Both Normal Q-Q plots seem to display a good fit. Roughly speaking, the Scale-Location plot seems to show a constant line, with a slight bend towards the lower fitted values, however. Additionally, the residuals in both models seem to be independent of the fitted values, with the exception of the lower values being consistently overestimated. When it comes to the ranges, the simple model has a larger range which makes sense since $RSE_1 > RSE_2$. Lastly, the simple model seems to show no outliers or badly-behaved points from the leverage plot. On the other hand, there is one outlier with considerable leverage in the complex model.

5 Prediction

Now, we come to test the performance of our models. Recall that in our discussion regarding the response variable, we talked about rescaling the response variable in the testing data to account for the growth in the user base. We may finally concretely show this strategy to be powerful by observing the MSE when considering both the scaled and unscaled response variable for both Model 1 and Model 2. Consider the following table, summarizing our results below:

```
##         MSE   Model   Response
## 1 5258686.8  simple unadjusted
## 2  434167.1  simple   adjusted
## 3 5042661.5 complex unadjusted
## 4  431570.7 complex   adjusted
```

As we can see, the most profound result comes from switching from the unadjusted responses to the growth-adjusted ones. Namely, when we do so, we observe that the MSE improved and decreased by 1211% and 1164% in the simple and complex models, respectively. Keeping the response type constant, when we go from the simple to the complex models, we observe the MSE improves and decreases by by 4.3% and 0.6% in the unadjusted and adjusted response types, respectively. **All in all, it seems as though in our case, the complex model did indeed perform better (0.6-4.3% decrease). The game-changer comes when we adjust the response variable, in which we observe amazing reduction in the MSE (11-12x decrease).**

6 Discussion

Overall, we achieved a few things in this report. One of our main objectives from last time is to narrow the scope (and length) of our results and focus on finding a core subset of variables and test our models to ensure they perform well. Indeed, we did so and found that the model `cnt ~ weathersit + atemp + I(atemp^2)` is a great core model. It seems like the complexity introduced by `windspeed` and `hum` have improved the

model, meaning it is very possible the model could benefit from adding even more variables.

Additionally, we really focused on the main problem we are trying to solve: predicting the 2012 counts well. We have identified a core problem which got us much closer to doing so. Namely, last time, we failed to consider that the number of users using bikes is not necessarily constant across 2011-2012. By introducing the "growth-discount factor" $r = \frac{\mu_{2011}}{\mu_{2012}}$, we were able to "renormalize" our 2012 data to behave under the same underlying logic of the 2011 data. This allowed us to glimpse the true potential of $\hat{\beta}$ on our testing data. The results were remarkable, showing a 11-12x improvement in MSE. That being said, we only demonstrated this in the scenario where we "perfectly estimated" $\hat{r}$. This is our primary shortcoming – while we demonstrated the power of estimating r, we did not provide a method, technique, or even data to do so. However, as we know, the world is generating more and more data every day, making this less of an issue than in 2011.

When it came to model selection, we decided to take the simple approach to keep the report digestible, and chose a simple model and a complex one. With everything in place, there is certainly more room to explore more complex models. As it turns out, we have still not reached a point in model complexity where we are overfitting the data, but we are probably close. That is in lieu of the improvements of 0.4-4.3% in terms of the MSE. Additionally, the scope of this report was very focused on our OSE models. Now that we have achieved this, we may begin to slowly increase our scope to our previous goals of studying the `registered` and `casual` variables to observe any patterns or differences in predictability. Additionally, we only made use of our weather predictors, tossing away all of our *time-based* predictors. So for next time, we are considering bringing those variables back to wield the power of time-series models and/or autocorrelation models.

7 Appendix

- `instant`: record index
- `dteday`: date
- `season`: season (1:spring, 2:summer, 3:fall, 4:winter)
- `yr`: year (0:2011, 1:2012)
- `mnth`: month (1 to 12)
- `hr`: hour (0 to 23)
- `holiday`: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- `weekday`: day of the week
- `workingday`: if day is neither weekend nor holiday is 1, otherwise is 0.
- `weather-sit`:
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp`: Normalized temperature in Celsius. The values are divided to 41 (max)
- `atemp`: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- `hum`: Normalized humidity. The values are divided to 100 (max)
- `windspeed`: Normalized wind speed. The values are divided to 67 (max)
- `casual`: count of casual users
- `registered`: count of registered users
- `cnt`: count of total rental bikes including both casual and registered

8 References

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

[2] MA 575 Fall 2021 C3 Team #2, "Lab Report 2: Ordinary Least Squares".