Bike-Sharing Data Analysis: Prediction of Daily Bike Rental Counts Based on MLR

Final Project Report · MA 575 Fall 2021 · C3 · Team #2

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

12/10/2021

0.1 Prediction of the Yearly Growth Ratio

Our modeling paradigm is based on the assumption that the growth trend will remain the same in the future years as that in the year of 2011. Note that this is NOT saying that the user base is supposed to remain unchanged throughout the entire year; the fact that the same scaling factor works at all points in the entire year is due to the fact that the MLS model in the later part of the year, e.g., in fall and winter, are already trained to compensate for rental count growth due to user growth using the environmental and seasonal variables. As such, an estimation of the growth ratio, which we call $\hat{g}$ is very important and worthwhile. By construction, its true value is well estimated by the ratio $\frac{\mathbb{E}_{2012}(C)}{\mathbb{E}_{2011}(C)} \approx (0.608)^{-1} = 1.64$, which is the average counts in 2012 divided those in 2011. Note that this means our user base grew by around 64% between 2011-2012, which is a non-trivial amount. If unaccounted for, that magnitude of growth would (and does in fact) lead to chronic model underfitting. However, since we are not **allowed** to peek at the 2012 data, we must resort to other creative techniques in estimating $\hat{g}$. In this report, we propose two techniques for estimating $\hat{g}$: the environmental loss function technique (more involved) and the window technique (more simple).

0.1.1 Environmental Loss Function

The primary issue of relying solely on one years' worth of data to estimate the growth within that year is that there is great difficulty in factoring out the role environmental factors play in determining the difference in observed counts between any two given days. As such, if we are somehow able to find two days whose weather/environmental conditions are "similar" or "equivalent", the ratio of the counts between the later day and the earlier one is a reasonable "data point" useful for obtaining a sense of the growth. This follows because we asked those two days to be "environmentally equivalent" (we will formalize this notion soon), so *a priori*, if we were somehow able to marginalize out the environmental factors, then any observed difference in the count `cnt` **must** be attributed to user base growth! This motivates our technique, aptly named the "environmental loss function" technique. In brief, here is the strategy: by designing a loss function, we seek to find special pairs of days whose environmental factors are roughly equivalent. By taking the ratio of the counts, we obtain a preliminary estimate for the growth factor since in theory, we have marginalized the environmental factors in the selection process. Aggregating all these ratios and averaging them, we obtain an estimate for $\hat{g}$. That being said, we now slowly develop our technique.

Suppose we have two days (observations) d and d'. Consider the reduced dataset where we only have the three predictors: `atemp`, `hum`, and `windspeed`. In other words, our observations are vectors solely comprised of our *continous environmental variables*. Furthermore, suppose that we **standardize** these variables. This is critical since we don't want our loss functions' units to be arbitrarily inflated/deflated, but rather, unit-agnostic. This is one reason we might prefer continuous variables when implementing our loss function. Also, let's use `holiday` to focus only on non-holidays and removing holiday days from the dataset. We can adjust for environmental factors but comparing two "environmentally equivalent" days without accounting for `holiday` may very well throw off our estimates of $\hat{g}$. Losing the 11 days which are holidays prove to be trivial, since we are counting pairs and $\binom{365}{2} \approx \binom{354}{2}$. Note that we omit `weathersit` and `workingday` (albeit the latter is not environmental, it is still important) mainly since they are categorical variables. While loss functions can include categorical variables, we will opt for the simplicity, familiarity, and stability of continous variables. In any case, we can justify their omission anyway since `weathersit` is correlated to our chosen environmental variables, so the information loss is not catastrophic. Additionally,

since we are using `cnt` and not the subdivisions of `cas` and `reg`, the explanatory power of `workingday` is "canceled" out in the `cnt` variable in a sense since we observe inverse behavior of `cas` and `reg` with respect to `workingday`. Now, we may put the discussion of chosen predictors aside.

That being said, this means we may write $d = (x_1, x_2, x_3)$ and $d' = (x_1', x_2', x_3')$ where x_i represents the **standardized** value of predictor i. Furthermore, let $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ be a weight vector, more heavily weighing variables deemed more important in determining "environmental similarity". We deem this to be `atemp`. In practice, we end up weighing `atemp` by $\alpha_1 = \frac{2}{3}$ and the other two predictors, `hum` and `windspeed`, by $\alpha_2 = \alpha_3 = \frac{1}{6}$. Note that this is arbitrary, and subject to scrutiny, but for the sake of simplicity, we may all agree temperature is the major determinant in constituting a notion of "environmental equivalence".

Finally, let δ_i denote the weighted difference between in the values of predictor i between day d and d'. So, we write $\delta_i = \alpha_i(x_i - x_i')$. Observing the differences among all our predictors, we obtain the difference vector $\vec{\delta} = (\delta_i)_{i=1}^3$. Finally, a very natural notion of "loss function" arises: the magnitude of the difference vector! We restate this as such. Consider the loss function $\mathcal{L} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^+$, which takes two days as input, and outputs the magnitude of weighted vector of the differences in our normalized predictor variables between them. We write the *environmental equivalence loss* of d' with respect to d as follows:

$$\mathcal{L}_d(d') = |\vec{\delta}| \text{ where } \delta_i = \alpha_i(x_i - x_i')$$

Now, suppose we enumerate **all possible unique ordered pairs** of days, which we denote Π:

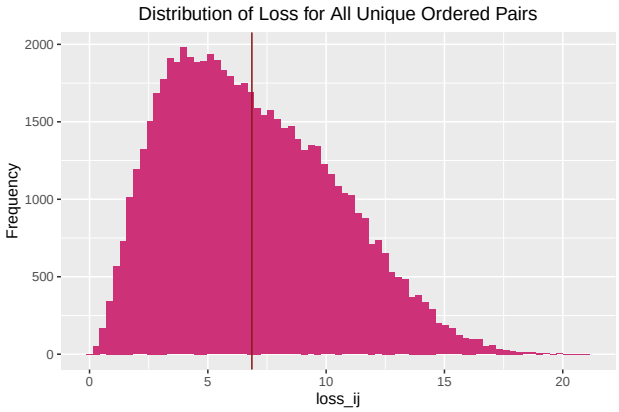
$$\Pi = \{\pi_{ij} = (d_i, d_j) \mid i < j\}$$

Note that we use the lower-triangle scheme, where we choose $i < j$ so that $i - j < 0$. This will prove useful later on since we want day i to preceed day j. We call this the "lower-triangular" scheme because this corresponds to the unique pairs of indices found in the "lower triangle part" of matrix. In any case, we compute the loss for every pair, and call it $l_{ij} = \mathcal{L}(\pi_{ij})$. Then, we obtain the following set:

$$\mathcal{L}(\Pi) = \{l_{ij} = \mathcal{L}(\pi_{ij}) \mid i < j\}$$

This is an exhaustive and computationally expensive procedure, since we compute the loss for $\binom{354}{2} \approx 62,000$ pairs of days! So, in practice we compute this once and store the data in `df_loss.csv`. When imported, we conventionally call the dataframe `df_loss`.

Now, here is the magic! We are not interested in taking pairs π_{ij} which are not very environmentally similar. After all, our quest is to estimate growth between days where we in theory expect a similar amount of users through the marginalization of environmental factors. So, there comes a question, which pairs do we discount as days too disparate to be considered environmentally equivalent? Well, our good foresight in normalizing the continuous variables and weighing them properly (to the best of our abilities) means that we can immediately consult distribution of l_{ij}, which we see below:



A quick note: it should not be surprising that we obtain a distribution that resembles a χ^2 distribution. After all, since we assume our predictors are (roughly) normal, their differences will be too; since our loss is essentially the norm of a multivariate normal, this explains our distribution shape! In any case, since we have done a good job by normalizing our variables (making our loss unit-less), a valid strategy is to **discount all pairs above a pre-specified upper bound**, which we will call B. An appopriate cutoff would probably be anywhere below the mean of our distribution $\mu \approx 6.86$, shown as the red line above. For example, we can take $B = \mu - 1\sigma \approx 3.35$ to be a reasonable cutoff for our upper loss bound B.

So, to summarize, we want to vet our exhaustive list of all possible day pairs Π and throw out bad pairs after doing some "quality control" to obtain a filtered set of pairs $\tilde{\Pi}$. Let us take a brief moment to recap everything. We are interested in estimating $\hat{g}$, and one way we are interested in doing so is observing the growth in pairs of days where the environmental effects are marginalized. A strategy for doing so is to consider "reasonably good" pairs of days, which we explicitly quantify through filtering Π through an upper bound on the loss B, which we may holistically choose based on the reasoning above.

So, we obtain our filtered/selected set of day pairs

$\tilde{\Pi}$, by cutting off the maximum loss at the selected upper bound B. That way, we can write our selected pair set as the output of a function f_{loss} that *filters* Π with respect to the chosen loss upper bound B:

$$f_{loss}(\Pi, B) = \tilde{\Pi}_B = \{\pi_{ij} \mid l_{ij} = \mathcal{L}(\pi_{ij}) \leq B\}$$

Next, we also filter for what we call an "index difference bound". This is our second round of "quality control". Namely, the idea is simple: we don't want to pick days too close to each other. We know days close to each other will have similar environmental conditions, but as to avoid pesky autocorrelation and vaccuous/misleading information, we impose a bound on how close the days can be.

Let $\rho := \min(i - j)$. In our lower-triangle matrix analogy (where we imagine indices as positions of matrix entries), ρ denotes which diagonal band we are considering. In any case, by imposing a lower bound on $i - j$, to pass the index difference filter, every pair must be atleast ρ days apart, or in other words, be at least ρ diagonal bands away from the corner or towards the main diagonal. In practice, we find low values of ρ to produce the most impactful filters, this makes sense, owing to the autocorrelation observed in days close to each other mentioned above. This makes the index difference bound much less critical than the loss bound, as it is much more impactful for low values of ρ. So, to summarize:

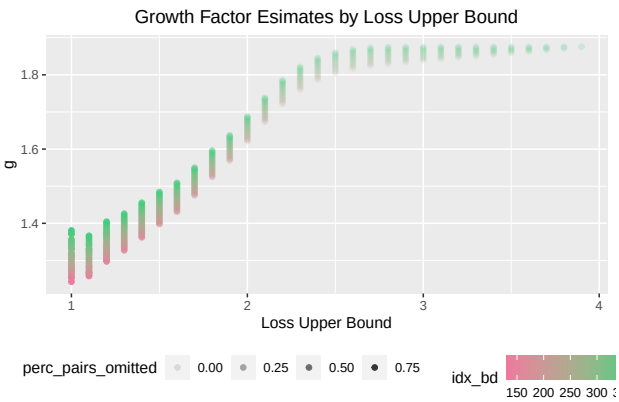
$$f_{idx\_diff}(\Pi, \rho) = \tilde{\Pi}_\rho = \{\pi_{ij} \mid i - j \geq \rho\}$$

Finally... we can compute the estimate derived from the filtered pair set $\tilde{\Pi}$ as follows. Let c_i denote the count of bike users in day i, and c_j those in day j. **Since we have, in theory, marginalized out the environmental factors (and holidays), we may obtain a reasonable estimate for g by taking the observed growth observed between day i and day j, then taking the average over all the pairs in our filtered pair set.** Since by construction, we used lower-triangular indices, this means day j is always after day i. As such, we can imagine each pair, which has marginalized environmental factors, to contain true information about user-base growth since day j is "environmentally equivalent" to day i and is **in the future** of day i! As such, we could in theory attribute this growth to none other than the user base growth itself. To be more explicit, consider a filtered set of pairs $\tilde{\Pi}_{B,\rho}$. Since every pair is considered "good", from each pair π_{ij} we obtain a preliminary estimate which we denote $\gamma_{ij} := \frac{c_j}{c_i}$. In general, the γ_{ij} are slightly unstable, and their variance may be worth studying; this remains outside the scope of the paper. However, since we are aggregating pairs from a set of $n = 62,500$ pairs (in practice we choose less), there is an abundance of data to counterbalance this. So, to summarize, given a filtered set of pairs $\tilde{\Pi}_{B,\rho}$, we obtain $\hat{g}$ as by taking average of the set of estimates γ_{ij} from every $\pi_{ij} \in \tilde{\Pi}_{B,\rho}$:

$$\hat{g} := \mathbb{E}[\gamma_{ij}] \text{ where } \gamma_{ij} = \frac{c_j}{c_i} \text{ and } \pi_{ij} \in \tilde{\Pi}_{B,\rho}$$

Estimator Analysis in Bound Paramaters Space. That being said, with our methodology explained, how does one estimate $\hat{g}$? This is our goal in the end! Well, as suggested by the notation above, the parameter estimate $\hat{g}$ is dependent on our filtering thresholds B and ρ. Of course, any choice of B_0 and ρ_0 would be arbitrary, so our next course of action is to observe and analyze our estimator over the bound parameters space. To be concrete, we say fix some ranges $[B_0, B_1]$ and $[\rho_0, \rho_1]$, then observe the behaviour of the resulting parameter estimate $\hat{g}$. To do so, we use the `df_loss` previously mentioned in conjuction with the `get_df_param` function which takes these ranges as an input and outputs `df_param`. A glimpse into what our table looks like is shown below. Note that n counts the number of pairs that "survive" the filtrations given the thresholds. In principle, we want our thresholds to pick exclusive pairs of days which are *unusually* environmentally similar (low B) but far apart enough (high ρ), with emphasis on the former. That being said, we now plot the results from this dataframe and obtain the exciting results below!

```
##   idx_bd loss_bd        g    n
## 1    130     1.0 1.242667  521
## 2    138     1.0 1.252098  506
## 3    138     1.1 1.258080 3198
## 4    146     1.0 1.256074  497
## 5    146     1.1 1.260398 3149
## 6    146     1.2 1.297205 8044
```



Again, as mentioned previously, the index difference bound is more sensitive when it comes to lower values. Greener points suggest that the index difference

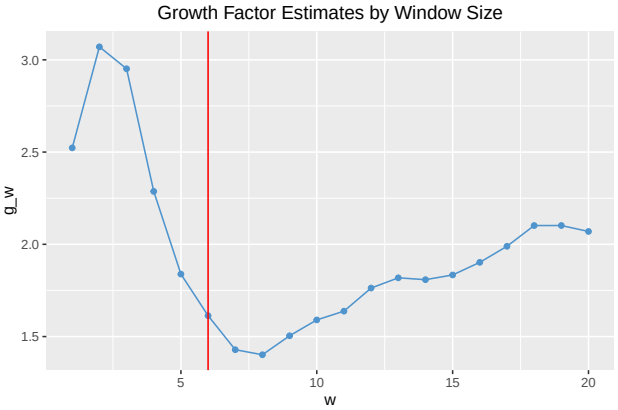
bound is high, so our days are quite far apart. Before we delve into the behaviour of the loss upper bound B, note that the opacity corresponds to the "percentage of pairs omitted". Recall that we want an exclusive set of pairs in which few pairs survive our strict filtrations or quality control. This is what the plot is suggesting, the more "faded" estimates of $\hat{g}$ are less well-founded, in fact completely unfounded as $p \to 1$ since we would indiscriminately be considering all pairs! As such, we conclude that the optimal range for B probably lies between $[1, 2]$, as these points have more "clarity" in their predictions due to their exclusively low loss. Otherwise, we also notice an interesting pattern with the index difference bound which makes sense, allowing days closer to each other will suggest "less growth". That being said, this means that in some sense, a potential optimal range for $\hat{g}$ could be that around the values where its variance starts to stabilize with respect to ρ. Visually, this is the range where the "length" of the streaks starts to hold still. For values before the critical $B \in [1, 2]$ range, this corresponds to the values around $\hat{g} \in [1.4, 1.6]$. That being said, this analysis is holistic, and not 100% conclusive. However, seeing this continous picture of estimating g and narrowing down its true value to be close to $\hat{g} \in [1.4, 1.6]$ is not bad!

0.1.2 Method II: The Window Technique

Jeez, that stuff was too technical. Thankfully, the window technique is much simpler. Remember our goal is to estiamte $\hat{g}$, which is the growth rate from 2011 to 2012. As such, the most elementary method to use based on this principle is to take the ratio of the counts in the first day of 2011 and those in the last day. This is justified because everyone would agree, there is some notion that Dec 31, 2011 $\approx$ Jan 1, 2012. However, the issue with this is that our estimate is based on one pair of points ($n = 1$).

As such, to preserve the spirit of the principle, but generalize as to instill more certainty and stability in our estiamte, we describe the "window technique". Firstly, we will say that the method mentioned above is the window technique, for a window size of $w = 1$. The principle of the window technique is the same as for the method described above, the days at the end of the year of 2011 will be similar to those in the beginning of 2012. So, the best we can do is take an equivalent "window" or sample of days from both the beginning and end of the year of size w. This way, we use more data and hence have more stability. Then, using those selected w days, we compute the average `cnt` in the beginning (first w days) and end (last w days), then take the ratio of those averages to estimate $\hat{g}$. We call this estimate $\hat{g}_w$. In our first example, we were describing $\hat{g}_1$.

In any case, just like before, $\hat{g}$ is an estimator based on a parameter, in this case w. We must note that as w increases, our assmuptions about "day similarity" between the end of 2011 and start of 2012 slowly starts to weaken. As such, we cannot go crazy and choose large w. Below, we exhaustively compute $\hat{g}_w$ for $w = 1, 2, \ldots, 20$, and plot our estimates with respect to w.



There are a few things to consider. As expected, the values $w \in [1, 4]$ exhibit highly unstable behaviour in estimating $\hat{g}$. However, it starts to stabilize around $w = 6$, highlighted in red. As it turns out, $w = 6$, holistically speaking is our best w for a few reasons. Firstly, going beyond $w = 6$ would mean including Christmas, which is a holiday and would thus throw off our estiamtes. Furthermore, the values of $\hat{g}$ start to stablize reasonably well around that value, indicating the marginal benefit of increasing w is almost fully realized at this point; remember, our assumptions about day similarity break quickly as w grows. Note that in the plot below, we compute $\hat{g}$ after omitting Jan 3rd, due to its high environmental loss value when paired with the other values. Lastly, one reason we consider $w = 6$ to be a good candidate is the fact that the estimate $\hat{g}$ starts to steadily increase again, indicating suspicious behaviour which could be attributed to the fact that as the window increases, we include days whose temperatures are changing in potentially different directions (generally both warming).

That being said, there is one curious piece of information corroborated by the environmental loss function technique. Namely, both graphs seem to suggest discounting the values $\hat{g} \gg 1.6$. In the second graph, this is suggested by the vacously increasing value of $\hat{g}$ with respect to w starting around $w = 10$, indicating that the point in which there is "good" information

about $\hat{g}$ probably lies before that value of $w = 10$, corroborated also by the simple principle that smaller w is generally better in terms of upholding assumptions. So, with all this information considered, we settle for $\hat{g} := \hat{g}_6 \approx 1.613$, which is the most compatible value from our holistic analysis of both of our estimation methods.