CASMA575: Lab Report 3

Raul-Fikrat Azizli, Thomas Laz, Harrison Ofori, Clare Oudard, Zeid Yunis

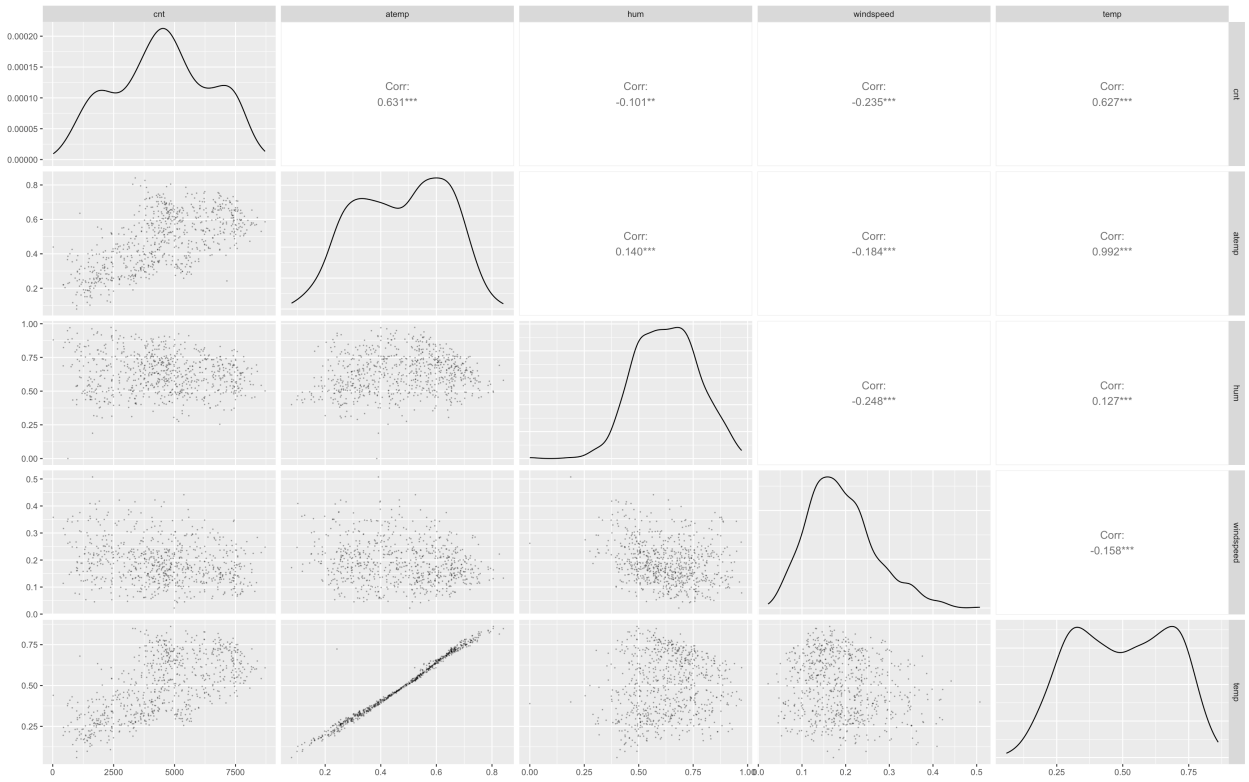
Introduction

We want to explore how weather conditions (such as weather classification, temperature, humidity, and windspeed), seasons, days of the week, and months of the year affect the number of bikes that are rented on a given day. By doing linear regression, we can see each variable will individually affect the count and how variables will affect the count when multiple are taken into account.

Data

We use the bike rentals dataset. This is a data set that records the cnt (number of bikes that are rented) as well atemp (temperature), hum (humidity), windspeed, season (season1, 2, 3, 4 representing Winter, Spring, Summer, Fall, respectively), year (year 0 as 2011 and year 1 as 2012), month (indexed 1-12 beginning at January), weekday (indexed 0-6 starting at Sunday), holiday (0 representing not a holiday and 1 representing a holiday), weathersit (1, 2, 3 representing clear days, less clear and light rain, and cold, rainy, and snowy), casual (representing first time users), and registered (representing repeat users).

We first checked the correlation between the variables:



We see extreme correlation between atemp and temp. We used atemp rather than temp. We don't use casual or registered because these variables are dependent variables and are heavily correlated with the cnt. We know that atemp is not independent of humidity in real life but the data does not show this, so we continue to use both.

Statistical Modeling Framework

Our dependent variable is cnt. Our independent variables: atemp, hum, windspeed, season, year, month, weekday, holiday, and weathersit.

From our previous labs we saw that the temperature alone was explaining about 30-40% of the model's variability, so we wanted to look at adding additional variables to temperature

After fitting models, we tried to plot the contribution of each variable.

Here is the AVS plot, which represents the relationship between the dependent variable given the added variable for a particular independent variable. We realized that some of the seasons created a positive, upward trend. This means they have some influence on the dependent variable. Some of the months are contributing, but we cannot pull out specific months from our data. We can see that the year greatly affects the number of bikes rented.

We cannot do the typical covariance matrices because our data is mostly categorical.

We created four models with the same independent variables other than specified:

1. **Model looking only at weekends**
2. **Model looking only at weekdays**
3. **Model looking at weekend vs weekday**
4. **Model with no weekday vs weekend distinction**

In all cases, there is a positive trend in bike rentals from 2011 to 2012. More people rented bikes in 2012 in every model.

In every case, season4 experiences the most significant, positive change between winter against each season. There are more bike rentals in the fall.

Additionally, in every case, weathersit1 (or clear, sunny days) is the reference, and we see a significant decline in the cnt when moving to both other weathersits, with weathersit3 showing a much more extreme negative jump, which we attribute to people not riding bikes in the snow.

Windspeed and humidity both have negative trends, but vary in significance among models, while temperature has a significantly positive trend across all cases.

In reference to month1, January, February to October have a positive difference while there is a negative difference between January and November and December. This is consistent with our other hypotheses that people do not want to ride bikes in the winter or on averagely colder, worse-weather days.

Questions

For every question, we assume independent, normally distributed variables.

- *How do bike rentals differ within weekends?*

Linear regression model describing how the cnt (number of bike rentals) is affected by all the aforementioned variables, excluding weekdays (weekdays 1-5) and using weekday0 as a reference.

When using Saturday as a reference, there is a significant change. By comparing Sunday to Saturday, there is a significant difference between the two. The cnt jumps up significantly when the model goes to Saturday.

Windspeed and humidity both have a negative trend, but humidity is not significantly contributing to the model. There is a significant correlation between humidity and atemp which can account for the fact that in the presence of atemp, the humidity is not significant.

- *How do bike rentals differ within weekdays?*

Linear regression model describing how the cnt (number of bike rentals) is affected by all the aforementioned variables, excluding weekends (weekdays 0 and 6) and using weekday1 as a reference.

By looking at the effect of which weekday we select at a 5% significant level, there is no significant change when we compare the other weekdays to Monday. Even though there is a difference in the bike rentals across the weekdays, the difference is not significant. If there is a five day work week (which we assume) it makes sense that there is no day with a significantly different amount of bike rentals.

Humidity, in this case, is significant and shows a negative trend.

- *How do bike rentals differ between a weekday vs a weekend?*

Linear regression model describing how the cnt (number of bike rentals) is affected by all the aforementioned variables, but classifying each day as either weekend or weekday and using weekday as a reference.

By using weekday as reference, moving from weekday to weekend shows a decline in bike rentals. On weekdays there are more rentals — perhaps because people use bikes more consistently for work than for leisure.

We observed that humidity, and wind speed are all significantly negatively correlated with the cnt bike rentals.

- *How do bike rentals differ between every day?*

Linear regression model describing how the cnt (number of bike rentals) is affected by all the aforementioned variables.

Very consistent with past models for all variables (excluding days of the week), with humidity being significant in this case.

With Sunday as a reference, there is a positive difference in bike rentals on all days compared to Sunday. However, only the change from Sunday and Tuesday to Saturday is significant. We hypothesize that this is because most days off are on Mondays.

Holidays show a negative trend and represent a decline in bike rentals as compared to days that are not holidays.

Data Analysis Results

Our model approximates about 85% of the variability.

- Present the key statistics from your regression model, including coefficient estimates, p-values, and R-squared. You can format this information in tables or include the model summary output directly from R.

- Create relevant visualizations, such as scatterplots with regression lines and residual diagnostic plots, to illustrate model fit.
- Interpret the results of your model in the context of your research questions. Clearly explain your interpretations and findings.

Contributions

- Raul-Fikrat Azizli: Formatted report snippets into 1 final report.
- Thomas Laz:
- Harrison Ofori:
- Clare Oudard: Wrote up report based on what group talked about when meeting
- Zeid Yunis: Helped formulate project concepts/ideas and submitted the final report on Blackboard.

Team Roles

- Coordinator: Harrison Ofori
- Recorder/Writer: Clare Oudard
- Modeler: Thomas Laz
- Monitor/Checker: Zeid Yunis
- Coder: Raul-Fikrat Azizli

Appendix