Analyzing the NYC Subway Dataset

Submitted by Rebecca Mayer, August 31, 2015

Section 0. References

@anfego. "Improving R^2," Chat communication posted on slack study group, August 1, 2015. https://udajul15.slack.com/messages/project2/team/slackbot/.

@ballsdotballs. Reply to "Numpy - Fast (but Not Very Accurate) Method for Finding Distance between 2 Points Using Python and Pandas - Stack Overflow." Accessed August 21, 2015.

"Are the model residuals well-behaved?". NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/. Accessed August 27, 2015.

Fagerland, M. W. (2012). "t-tests, non-parametric tests, and large studies—a paradox of statistical practice?". BioMed Central Medical Research Methodology 12: 78. doi:10.1186/1471-2288-12-78 as cited by the Wikipedia entry, "Welch's t test". Accessed on August 29, 2015.

Giles, Dave. "More About Spurious Regressions." Econometrics Beat: Dave Giles' Blog, May 30, 2012. http://davegiles.blogspot.com/2012/05/more-about-spurious-regressions.html. Accessed August 25, 2015.

Nielsen, Heino Bohn. "Non-Stationary Time Series, Cointegration and Spurious Regression." Lecture materials from Econometrics 2 - Fall 2005, University of Copenhagen Department of Economics, 2005.

"Spatial Regression with GeoDa." In Spatial Structures in the Social Sciences. S4 Training Modules. Accessed August 22, 2015. UCLA: Statistical Consulting Group. http://www.s4.brown.edu/S4/Training/Modul2/GeoDa3FINAL.pdf.

"Uber NYC has launched". Uber.com, May 3, 2011. http://newsroom.uber.com/2011/05/uber-nyc-launches-service/. Accessed August 29, 2015.

Understanding the Mann-Whitney U Test. Udacity Data Analyst Nanodegree Events: Hangouts on Air. Broadcast July 14, 2015.

Ward, Michael D., and Kristian Skrede Gleditsch. An Introduction to Spatial Regression Models in the Social Sciences, 2007. https://web.duke.edu/methods/pdfs/SRMbook.pdf.

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used a Mann-Whitney test to examine whether subway population characteristics were different on rainy days using the original (unimproved) turnstile_weather dataset as the source of data. Because I was not sure which direction a difference would take, I chose a two-tailed p-value: subway ridership might plausibly increase or decrease when it rains. The null hypothesis for the test, with a p-critical value of .05, is as follows:

X = subway entries on rainy days
Y = subway entries on non-rainy days

Null hypothesis: Subway entries are the same on rainy and non-rainy days. In mathematical terms, this can be expressed as $P(x > y) = 0.5$.[i]

Alt: Subway entries differ on rainy and non-rainy days. $P(x > y) != 0.5$

The Mann-Whitney test evaluates the null hypothesis using a rank-ordered statistical technique to assess the likelihood that the probability distributions of the two populations are the same. The ranks of the data points in both samples are summed in order to produce the U test statistic. Because the U statistic is a normally distributed variable, whose mean and standard deviation depend only on the number of data points in each sample, the statistic can be standardized into a Z-score and assessed for statistical significance using the conventional p-value approach.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The Mann-Whitney test is applicable because it does not require any assumptions about the distribution of the underlying data. T-tests are based on the assumption that the data follows a normal distribution. The histogram of ENTRIESn_hourly (provided in Section 3.1) shows that the subway dataset is not normally distributed. The large size of the dataset, over 130,000 observations, makes it a good candidate for the Mann-Whitney test, which needs a large sample size to achieve sensitivity to small differences between population distributions. Such a large dataset indicates that a t-test would also be an acceptable approach[ii].

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

Using the scipy.stats implementation of Mann-Whitney I received the following results:

With-rain mean: 1105.45
No-rain mean: 1090.28
U statistic = 1924409167.0
one-tailed p-value: 0.0249999

The two-tailed probability of obtaining a test statistic more extreme than this result is just under .05, which I calculated by doubling the one-tailed p-value. The fact that the obtained p-value was so close to the boundary of the critical region led me to take a closer look at the original hypothesis and test formulation. I started to doubt whether the data was appropriate for the question asked. Specifically, the rain data was recorded on a daily basis while entries data was recorded every four hours or less. This meant that the data could misrepresent the true rain state.

In order to align the reporting intervals, I processed the data so that ENTRIESn_hourly were summed for each UNIT by date. This allowed me to test whether rain at any time during the day was associated with a net change in subway ridership for the day. The sample size remained large as the processed data comprised 9,270 observations with no rain and 4,642 observations with rain.

Using the same hypothesis described in Section 1.1, I ran a Mann-Whitney test on the summed daily ENTRIESn_hourly data. The results of the test were:

With-rain mean: 10502.9
No-rain mean: 10332.0
U statistic = 21345225.5
one-tailed p-value = 0.2227

1.4 What is the significance and interpretation of these results?

After adjusting the p-value to .45 to account for a two-sided test, it is clear that the p-value is far higher than the p-critical level of .05. With entries summed on a daily basis to match the reporting interval of the rain classification, the Mann-Whitney test does not provide evidence to reject the null hypothesis. Based on this second test, I conclude that total subway entries do not differ significantly on rainy and non-rainy days.

Related code: ‘Mann-Whitney_daily_entries_original_dataset.py'.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model<

Statsmodels OLS with improved dataset

Related code: 'OLS_with_constant_v1.py'

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

The features I used were *precipi*, *holiday*, *lag_4hr*, *lag_8hr*, *neighbor_hour* and *UNIT*. The model included a constant term.

precipi is a continuous variable indicating precipitation in inches at the time and location of the weather station.

holiday is a dummy variable representing weekends and public holidays (value=1) vs. business days (value=0). Memorial Day (May 30, 2011) was the only public holiday in the dataset's date range.

lag_4hr and ***lag_8hr*** are time lag variables to represent the predictive value of entries at the same UNIT location four and eight hours prior.

neighbor_hour is the product of *hour* and *neighbor_count*. *hour* is an integer representing the hour of the day. Possible values for hour in this dataset are [0, 4, 8, 12, 16, 20]. *neighbor_count* is the number of remotes located within 1 km of the unit. Distances for *neighbor_count* were calculated from latitude and longitude using the haversine formula[iii]. *neighbor_hour* represents time of day in the model while giving higher weight to remotes located in areas of high station density at later hours of the day.

UNIT represents the dummy variables for the 240 unique remotes in the improved dataset.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

Most of the selected features contributed to an increase in R-squared that justified their inclusion in the model. As illustrated in **Table 1**, the location-specific variable *UNIT* was responsible for the largest single increase in explanatory power, contributing 0.141 to R-squared when added to the model. Three other features serving as proxies for location effects - neighbor_count (station density) and the time lag terms - also contributed to R-squared.

Without any location-related variables the model's R-squared drops to 0.109. The high impact of location-based variables on the linear regression model's R-squared value confirms what one might guess about an urban transport network: station location is a major determinant of ridership patterns.

Table 1: Impact of removing individual features on the model's R-squared value

Features		**R-squared**
Complete model	*const, precipi, holiday, lag_4hr, lag_8hr, neighbor_hour, UNIT*	0.598
Model with indicated features removed (all other features retained)	*precipi*	0.597
	holiday	0.578
	lag_4hr, lag_8hr	0.519
	neighbor_hour	0.494
	UNIT	0.457

To put the statistical results in context, I developed an overview of the types of factors that might impact subway ridership and their mechanisms of influence, and linked them to specific proxy variables in the regression (**Table 2**). These factors were based on my personal experience riding urban transport systems as well as general business principles.

The hypothesized mechanisms of influence on subway ridership fall into two camps: factors that influence the supply and demand for urban travel in general, and competitive factors that influence whether the subway will be chosen from among alternative transport options. Factors that generate the demand for urban transport include things like the distance between work and home locations. Substitution effects are influenced by price, convenience, and the availability of competing transport options at a given time and place. Some examples of both types of factors are likely to be invariant during a short enough time period, such as the one month span of the turnstile_weather dataset. In this case they will not show up in a regression analysis (other than perhaps group-wise in the constant term) although they may be very important determinants of subway ridership.

I selected *precipi* as the sole feature from the set of weather indicators because it was the single data point that most accurately encoded the variable of interest, rain, in terms of timing and intensity of rainfall, and because it produced a coefficient with a low p-value. The other weather indicators that I tried did not improve R-squared so I saw no reason to include them.

As illustrated in the **Table 2** column 'proxy variables in regression', several potentially important factors, such as price and availability of competing transport options, were absent from the dataset, leaving them unrepresented in my model. I expect that these factors will contribute to residual errors and limit the explanatory power of the model.

Table 2: Mechanisms of influence on subway entries

Factor	Proxy variable in regression	Hypothesized underlying mechanisms of influence on subway entries
Time of day	*neighbor_hour*	• Timing of work commutes and leisure outings • Frequency of trains by time period
Location	*UNIT* *neighbor_hour*	• Geographical patterns of distribution of home, work and leisure destinations
Interaction of time & location	*neighbor_hour*	• Peak/off-peak demand interacting with gateway/hub stations
Date	*holiday*	• Work / holiday calendar
Station entrance closed/no trains	None	• Likely responsible for some data points showing zero entries (~2% of data)
Financial costs	None	• Substitution vis-a-vis other transport options
Non-financial costs (e.g. travel time and convenience)	*precipi* non-financial costs other than weather: None	• Substitution vis-a-vis other transport options, e.g. harder to find taxi when it's raining • Substitution between subway stations, e.g. choose closer local station instead of farther express station when raining
Social/cultural perceptions	None	• Substitution vis-a-vis other transport options, e.g. perception of subway travel generally, or a particular station in a particular time and place, as dirtier/cleaner, more/less dangerous than alternative transport.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

As shown in Table 3, all non-UNIT features included in the model had coefficients with p-values less than .001. This indicates that their effects were statistically unlikely to have been caused by chance if the model is properly specified.

Holiday has a negative impact on ridership, being associated with a decrease of 915.75 entries on average. Precipitation also has a negative effect.

The time-lagged dependent variables *lag_4hr* and *lag_8hr* hint at cyclic patterns in ridership over the course of the day. The four-hour lag term is associated with a positive effect on current entries while the eight-hour lag is associated with a negative effect. The eight-hour lag may be dominated by the effects of work commutes; if a station serves a high proportion of commuters it will have a high number of entries in the morning and relatively fewer at the end of the business day.

Table 3: Regression coefficients

feature	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	833.0718	19.597	42.511	0.000	794.662	871.482
holiday	-915.7521	19.896	-46.026	0.000	-954.749	-876.755
precipi	-3027.5953	355.537	-8.516	0.000	-3724.456	-2330.735
lag_4hr	0.3534	0.004	86.638	0.000	0.345	0.361
lag_8hr	-0.2011	0.004	-48.402	0.000	-0.209	-0.193
neighbor_hour	27.8893	0.267	104.631	0.000	27.367	28.412

2.5 What is your model’s R2 (coefficients of determination) value?

R-squared = 0.598

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

If the model meets all of the assumptions for OLS regression, then 0.598 is a good R-squared value as most of the variation in subway entries is explained. Based on the analytical framework in Sec. 2.3, it is clear that some predictive variables are missing from the model, while those that are included don't encode all of the relevant information. For example, there is only indirect information in the data regarding the start and end points of each rider's journey and the types of destinations reachable from each subway station. Given these limitations, one would not expect the model to explain all of the variation in such a complex dataset.

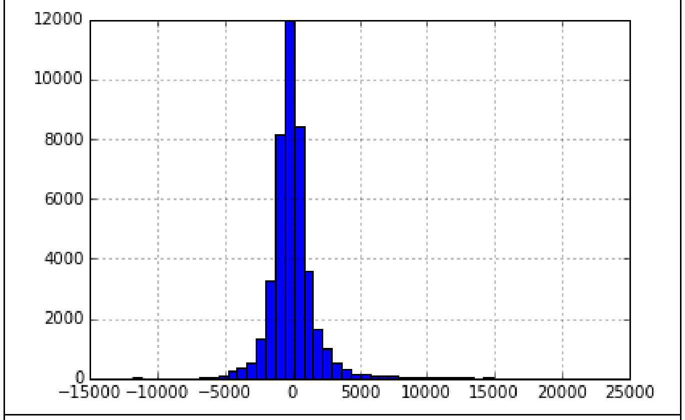

Figure 1: Histogram of residuals

The histogram of the model’s residual errors, shown in Figure 1, is roughly centered around zero and looks more or less normal. According to NIST’s Engineering Handbook, this suggests that the distribution of errors does not indicate a major structural element is missing from the model.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

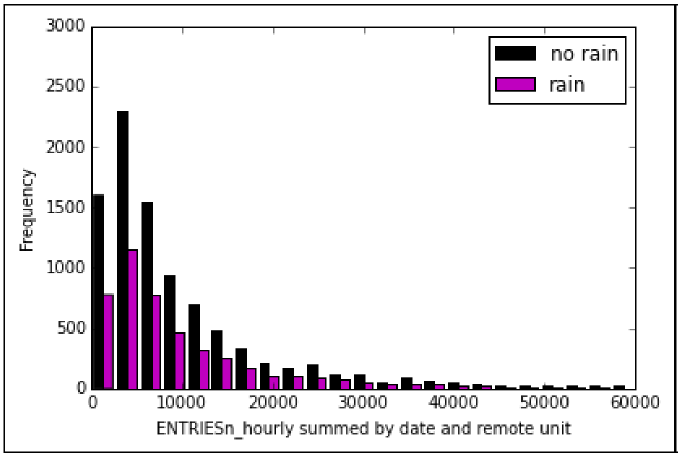

Figure 1: Histogram of daily subway entries

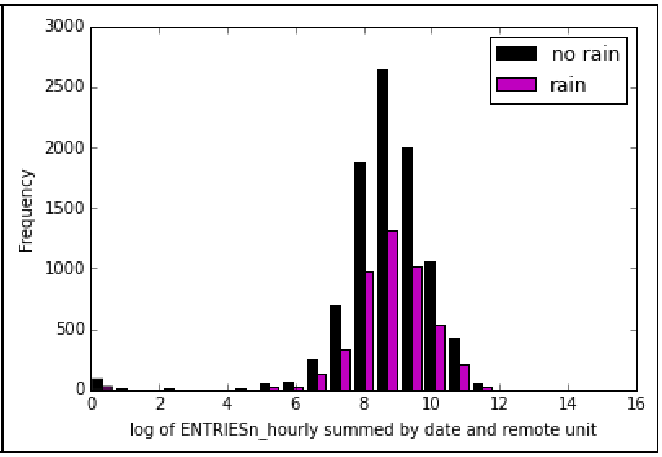

Figure 2: Histogram of log-transformed daily subway entries

The plots above offer two views of the subway data from the original (unimproved) turnstile_weather dataset. Figure 1 shows the frequency distribution of ENTRIESn_hourly data when summed by date and remote unit. Figure 2 shows the same data after a log transformation[iv]. Comparison of the two histograms suggests that the subway entries data follows a roughly log-normal distribution, although the transformed variable remains skewed.

Related code: ‘daily_entries_histogram_raw.py’, ‘daily_entries_histogram_log.py’

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

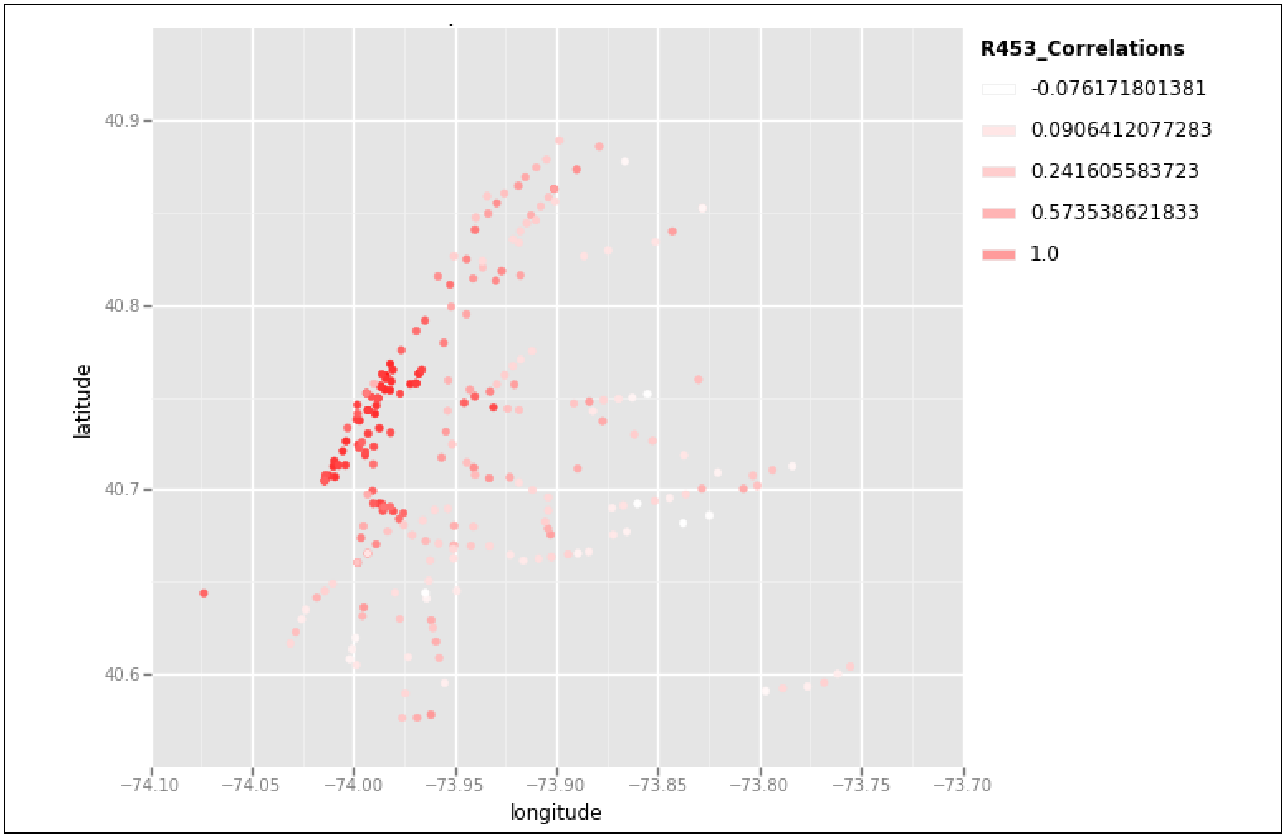

Figure 3: Correlations between ENTRIESn_hourly at R453 and all other units

Figure 3 sheds light on the issue of multicollinearity in the improved dataset by plotting the Pearson's correlation coefficients between entries at subway remote R453 and all other remote units. High correlation, shown using deeper shades of red, is a sign of possible linear dependence. Note that the data points are laid out on a latitude/longitude grid to give a rough sense of each unit's physical location. Although the scales of the latitude and longitude axes are distorted with respect to actual geographic distance, they are near enough to show a clear pattern of spatial concentration in the features correlation matrix.

R453 is a bank of turnstiles located at the 23rd Street station of the Port Authority of NY & NJ (PATH), a major gateway to New York City for commuters from New Jersey. Not surprisingly, this remote has a dense region of highly correlated stations on the Jersey side of the Holland Tunnel and in Manhattan, with its many draws for nearby residents: offices, shopping and culture.

Related code: 'Sec3.2_correlation_map.py'

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

Based on my analysis of the improved dataset, I conclude that more people ride the NYC subway when it is not raining, but the rain effect is temporary and doesn't significantly impact net ridership for the day.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

The decrease in riders when raining is supported by my linear regression analysis, in which the coefficient for precipitation had a statistically significant ($p < 0.001$) negative relationship with subway entries recorded over four hour intervals. The negative sign indicates that subway entries decrease when precipitation increases. However, because the regression model shows signs of multicollinearity (as discussed further in Section 5.2), the coefficient for precipitation may not be a reliable predictor of the magnitude of the effect.

In addition to the linear regression, I ran Welch's t-test on the difference between mean entries while raining (1743) and not raining (1897). The resulting t-statistic was -3.019 and the two-tailed p-value was .005, indicating rejection of the null hypothesis of equal means and supporting the interpretation that mean entries were, indeed, lower while raining.

The finding of fewer people riding the subway when it rains did not hold up when the time span was extended from four hours to a daily total. A Mann-Whitney test comparing the distribution of total daily entries on rainy and non-rainy days produced a one-tailed p-value of 0.383. This is far in excess of the p-critical value of .05 and provides no reason to reject the null hypothesis that the rain and no-rain samples come from the same population.

In the framework presented in Table 2 of Section 2.3, transient weather conditions are presumed to influence subway entries by impacting convenience. To validate this assumption and to find out more detail, I would suggest conducting a poll of subway riders and asking them how their decision to ride the subway is affected by rain. Such a poll could provide valuable clues as to how to segment and process the subway data in order to tease out the quantitative impacts of rain on ridership.

Related code: 'Welchs_t-test_improved_dataset.py', 'Mann-Whitney_daily_entries_improved_dataset.py'

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,

Limitations of the improved dataset include:

- Date range is limited to one month, May 2011, which does not reflect the extremes of weather throughout a typical year.
- The four-hour reporting interval is too coarse for precise time-of-day analysis. It is difficult to identify the boundaries of peak vs non-peak travel times, as well as the effect of rain that lasts much less than four hours.
- Many subway journeys consist of a round trip, which implies that the ENTRIES data points are not all independent.
- In the subway system, each remote is associated with a specific station and each station is associated with specific subway, train and/or bus lines. Multiple remotes can be associated with the same station. These connections are not well represented in the current dataset.
- Lack of indicators for price and availability of alternative transport options, limiting the ability to analyze how weather effects are mediated through the availability of alternative transport options.
- The improved dataset was truncated to include 240 of the original 465 remote units. This processing no doubt improved the dataset by making it more consistent and comparable along some dimensions, but are the remaining stations representative of the full network?

2. Analysis, such as the linear regression model or statistical test.

Review of online statistical guides suggests that the linear regression model I developed may violate some of the requirements for OLS and therefore the coefficients and other statistics may not have the expected interpretation. The regression's condition number is extremely high at 1.64e+19, indicating multicollinearity in the data. Several of the UNIT coefficients have confidence intervals that span zero along with high p-values, suggesting that the UNIT dummies may be one source of multicollinearity. In this situation the coefficients provided by the model will not be numerically stable. Therefore, the individual coefficients may not provide reliable or accurate predictions[v].

The presence of peak and off-peak travel times suggests that subway entries do not have constant mean and variance at different hours. This indicates non-stationary properties of the data (Neilsen, 2005). Econometrics blogger Dave Giles suggests that applying OLS to non-stationary time-series data can give rise to “spurious regressions”, which are regression models that produce high R-squared values even when the independent and dependent variables are not related.

Many techniques have been developed to analyze datasets with significant spatial components such as this one. The concepts of spatially lagged dependent variables and neighbor effects described in Ward & Gleditsch (2007) and Spatial Regression with GeoDa inspired my experimentation with the *neighbor_hour* and time lag features in the model. There are many more techniques and tests that can be applied to model spatial connectivity that could make the model a better predictor of subway entries.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

To make the regression model more useful for real-world business decisions about the New York subway system, the model should be extended to better incorporate competitive factors and reflect the substitution effects from other transport options. It would be interesting to extend this analysis to incorporate geo-referenced data from the Uber API for New York City to examine the impact of localized supply increases and/or price declines in taxi services on subway ridership. Uber launched its taxi services in May 2011[vi], the same month this subway dataset is taken from. A spatially-localized regression model of joined Uber/subway data since that date could shed light on whether Uber has taken market share only from competing taxi services, or whether its effects have also been disruptive to New York City public transport on a local or city-wide scale.

End Notes

See Section 0: References for complete citations.

[i] Source: Understanding the Mann-Whitney U Test (2015).
[ii] Wikipedia cites Fagerland (2012) to argue that Welch's t-test "remains robust for skewed distributions and large sample sizes".
[iii] Code posted by @ballsdotballs on stackoverflow.
[iv] The use of a log transformation on the data was suggested by @anfego. In order to facilitate plotting, data points equal to zero in the ENTRIESn_hourly series were preserved as zero during the log transformation rather than being converted to -inf as per numpy.log.
[v] Interpretation in this paragraph is based on material from the Udacity webinar on multicollinearity.
[vi] Source: http://newsroom.uber.com/2011/05/uber-nyc-launches-service/