Analyzing the NYC Subway dataset

By Jaroslav Klen

Section 1. Statistical test

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?**

For analyzing the NYC Subway dataset I used the Mann Whitney U-test. With this test we can answer the question, if these 2 samples have identical distribution, or in other words, if they came from the same population. Another way to interpret the results of this test, is if differences in our samples occured due to chance, or are statisticaly significant. Alternative hypothesis is, that they have not identical distribution, so I performed 2 tail test with p-critical = .05.

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 U-test is a non-parametrical statistical test, that means it does not make assumtions about distribution parameters of populations, from which the samples came from. In particular, it does not assume any certain form of distribution. Altough for bigger samples (n > 50), due to application of the central limit theorem, it is possible with proper handling of outliers and extremes to use parametrical statistical tests as well, no matter how are the values distributed.

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.**

U	1924409167.0
p[1]	0.0386
Mean – rain	1105.4463
Mean – no rain	1090.2788

4. **What is the significance and interpretation of these results?**

Significance level (alpha), which we choosed at the begining of our test is 0.05, or 5 %. Because our p value is less than our p-critical, we can reject null hypothesis, that our samples came from the same population. To be concrete, results of our test exactly means, that if our null hypothesis is true, there is only 3.86 % probability, that we will obtain at least as extreme results. We can say also, that with confidence of 95 % is the difference between our samples not due to chance, or sampling error, and our results are statistically significant at the 0.05 level. These results together with mean statistics for both samples could indicate, that there might be relationship between the independent variable (rain) and dependent variable (ENTRIESn_hourly), that people ride subway more when it is raining, rather when it is not raining.

Section 2. Linear Regression

1. **What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?**

[1] P value for two tailed test, which was obtained as the output p from Scikit Learn one tailed test multiplied by 2.

I used OLS from Statsmodels.

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

As input non-dummy variables I used normalized rain, mean temperature, mean pressure and mean wind speed.
For my model I used dummy variables created from 3 variables. First, 'UNIT' variable, which was categorical variable with 465 distinct values representing the unit, which is collecting turnstile information. Since there could be a big difference in utilization between different turnstile units, this information may be a important input for our regression model. To make this categorical variable usable, it was necessary to split it into 465 binary variables (dummy encoding).

Another categorical variable, from which I created dummy variables, was 'Hour'. This change (to use 'Hour' with dummy encoding instead of using the variable without dummy encoding) led to improving R^2 by approx. 0.05. Third categorical variable which I used in my model via dummy encoding was 'Weekday', representing the day of week, when was the turnstile information collected.

Maintaining all dummy variables and including constant in model resulted in multicollinearity problems, so it was neccesary to drop at least one dummy variable from each 'dummy' group, plus constant, to get condition number within acceptable value. This, and normalizing non-dummy features had biggest impact on models R^2.

3. Why did you select these features in your model?

For selecting mentioned features, I tried different approaches. First, I followed my own assumtions, or reasoning, what variables may have causal relationship with the subway ridership. Some of my assumptions include:

- People may use subway more when it is raining (this was also confirmed with the Mann Whitney U test), when precipitation is higher, when it is more windy or atmospheric pressure is low
- Subway ridership may differ with the time of day, based on peoples daily life cycle
- Subway ridership may differ with the day of week, based on peoples weekly life cycle
- People may use subway more, when it is too hot, or too cold

Secondly, I experimented, how the R squared behaves, when I remove, or add another variables with default parameters settings of the algorithm. Using non-dummy variables alone resulted in very low predictive power of the model, so it was neccesary to include more information about location, time and weekday when was the turnstile information collected in form of dummy variables.

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

Rain	Meantempi	Meanpressurei	Meanwindspdi
-76.90	-8.24	-271.56	-8.35

[2]

5. What is your model's R^2 (coefficients of determination) value?

R^2	0.6019

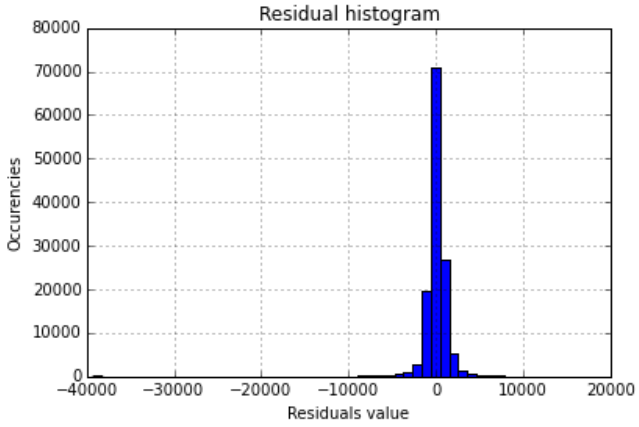
6. What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

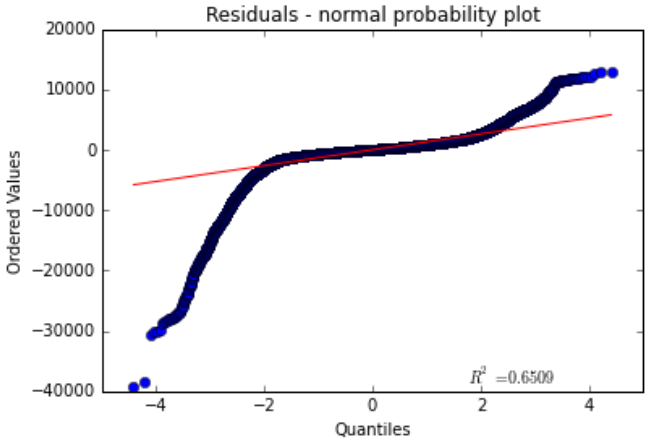
R^2 is the 'variance explained' by the model, or is the fraction by which the variance of the errors is less than the variance of the dependend variable. We can say then, the model explains arround 60 % of variance of the target (dependent) variable 'ENTRIESn_hourly'. Simply said, it measures, how well our model predicts the target variable.

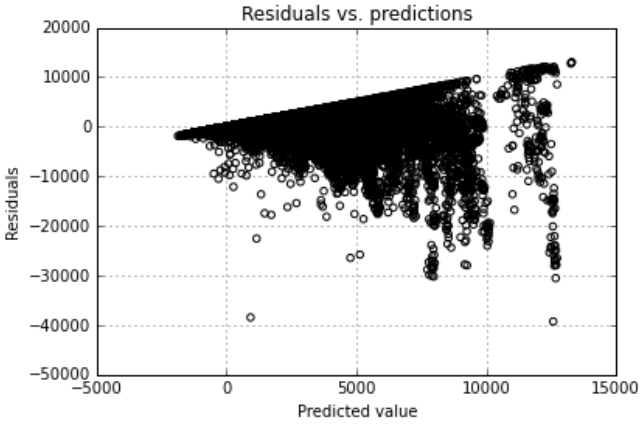
[2] Weights are after denormalizing.

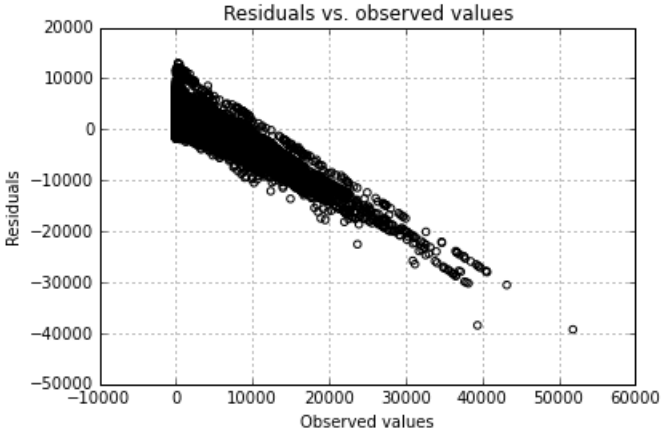
The value of R^2 alone is insufficient to answer question, if our model is appropriate and useful. It depends also on furhter decisions, which will be based on its predictions and models complexity (number of independent variables used as inputs).

To evaluate, if our model is appropriate for this dataset and if it meets regression assumptions, we have to perform analysis of residuals, to see if there are any violations, or space for improvement of the model.







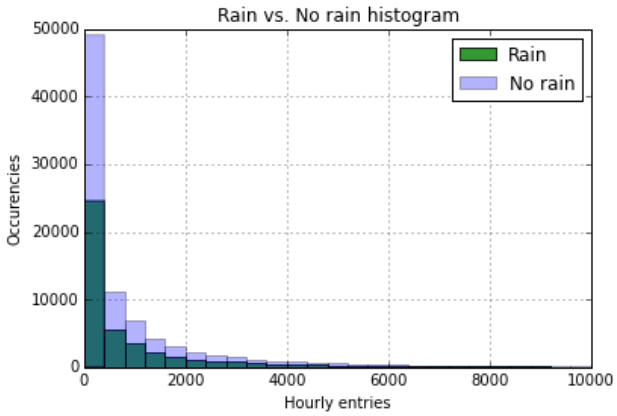


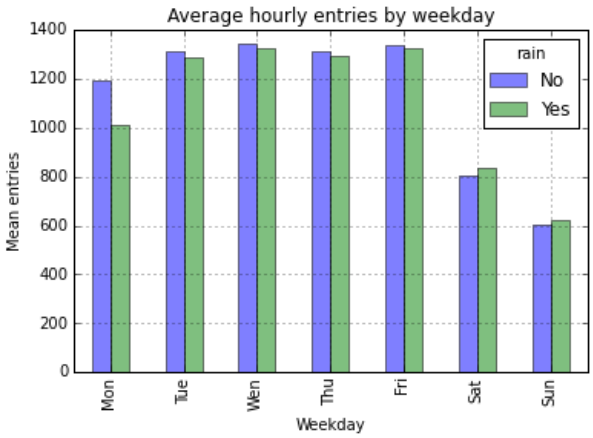
Based on above charts, we can conclude, that our model violates some of the regression assumptions. One of them is that residuals should be normally distributed with the mean around zero. From first chart (Residual histogram) we can see, that its distribution has too long tails, more on the negative side, which suggest few very big residuals. More information regarding normality we can get from the next plot (Residuals – normal probability plot), which shows us relative position of residual quantiles to quantiles of theoretical normal distribution. Big departures from straight red line are indications of abnormalities in residual distribution.

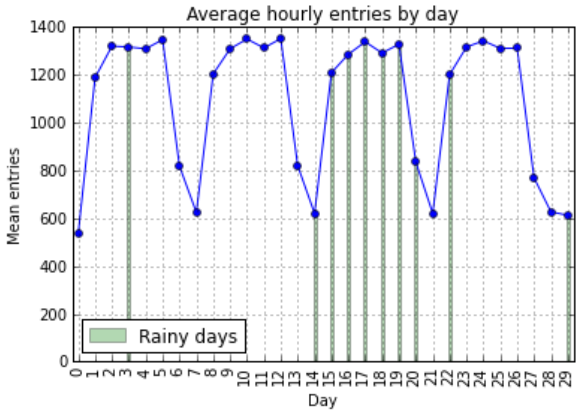
Residuals should display about the same degree of variability throughout the range of the dataset (constant variance, or homoscedasticity), but plot Residuals vs. predictions shows us, that the variance is increasing with the predicted value. This could indicate violation of linearity or additivity of the relationship between dependent and independent variables.

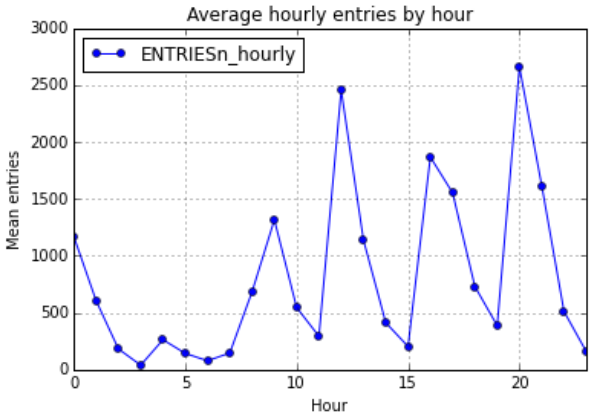
On Residuals vs. observed values plot we can see, that with smaller observed values, the model has tendency to overpredict, and with higher values to underpredict, but this behaviour is not violating any assumptions.

Section 3. Visualizations









Description

From first chart (unstacked Rain vs. No rain histogram), we can see, that values of both samples are not normally distributed, but have exponential distribution. Second thing what can we read from this histogram is, that our dataset contains more data points (each data point is per turnstile machine, date and hour) on days when it was not raining. Range of ENTRIESn_hourly values reaches overall up to 52 000, but for purpose of this visualization were the values bigger than 10 000 removed.

Second chart (Average hourly entries by weekday) shows the mean of hourly entries for day of week separately for rainy and non rainy days. As we can see, weekends are the only days, where is the mean for rainy days slightly higher, but except for Monday, there are no obvious differences. Eventually, this could be caused by an outlier or extreme value.

Third chart (Average hourly entries by day) shows the mean of hourly entries for day including the whole data set, which is only for May of 2011. Green bars indicate if it was a rainy day. We can see clearly cyclic patterns in the subway hourly entries, where during working days is the average subway ridership more than double than in weekends. Other interesting information is, that our dataset contains only 10 rainy days from overall 30 days.

Fourth chart (Average hourly entries by hour) shows us the peaks in subway ridership during the day. These are probably in relation with peoples 8 hours working hours cycle (9 – 16 and 12 - 20), but without having more data, we can only assume.

Section 4. Conclusion

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?**

According to results of Mann-Whitney U test, which was testing, if the number of hourly entries on rainy days is statistically diferent from hourly entries on non-rainy days, and the mean of both samples, we can conclude, that people ride subway more on rainy days. However, based on p value of the statistical test, in rejecting the null hypothesis, we could make mistake in our decision with 3.86% probability. Important is also to note, that we do not know if the rain 'intervention' caused our statistically significant results between samples in terms of subway ridership. I think, that we might get more insights and find evidence to support our decision by splitting existing data set into 2 parts (working days and weekends) and perform Mann-Whitney U test separately on both.

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.**

Created linear regression model, on the contrary, does not support the results of Mann-Whitney U test. Negative coefficient of -76.90 indicates, that when taking into account all variables used as model inputs, presence of rain decreases the subway ridership by 76.90 entries. Interpreting variables coefficients in multiple regression without consideration of other variables, can be misleading, because each variable coefficient is influenced by the presence of other variables in the model. This does not mean, that we took wrong decision by rejecting the null hypothesis, or that result of statistical test was false.

Section 5. Reflection

1. **Please discuss potential shortcomings of the methods of your analysis, including: dataset, analysis, such as the linear regression model or statistical test.**

Dataset

Dataset, which I used in this project has in my opinion few issues, which could result in wrong conclusions and outputs of analyses created from it. Mainly, that rain variable does not reflect the location of turnstile machines, which are calculating hourly entries and exits, and time when it was collected. So it is not obvious what information it represents, for example if it was some particular day raining whole day, or at least 20 minutes, in whole city, or only in some part. To sum it, rain variable has a different grouping level than turnstile part of the dataset.

Grouping level of turnstile information is by date -> trunstile machine -> hour, but our dataset does not contain all entries according this grouping, a lot of entries are missing.

Last issue is the size of the dataset. It covers only May 2011. Since weather conditions vary during whole year cycle, I think, that we would draw conclusion better reflecting the real world relationships when possibility to use data for whole year or more.

Linear regression and statistical test

Results of linear regression model and statistical tests based on our dataset will be probably not applicable in real world, due to mentioned dataset issues and violations of regression assumptions. Altough, we could possibly achieve normality and homoscedasticity of residuals by applying sort of transformation to dependent variable, which simplifies relationships with independent variables, like square root, or logarithm.

Non parametric statistical tests, like used Mann-Whitney U test in general do not make any asumptions about distribution of the data. This makes them more usable, but results in loss of information, which gives them lower

power than parametric tests. Because our dataset has big enough rain and norain sample sizes to perform parametric statistical tests based on central limit theorem, this way we could get additional evidence to support our current conclusions, which are based on Mann-Whitney U test.

References

http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients
http://stats.stackexchange.com/questions/47594/how-should-i-interpret-the-p-values-i-e-t-tests-in-regressions-and-can-i-use
http://www.stat.ucla.edu/~rgould/x401f01/mannwhitney.html
https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
http://www.itl.nist.gov/div898/handbook/prc/section1/prc131.htm
http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/
http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/
http://stattrek.com/regression/linear-transformation.aspx?Tutorial=AP
http://carbon.ucdenver.edu/~mas/coursemtls/resids.pdf