Intro to Data Science

Project 1 Part 2 (updated 07MAR2015)

Resubmitted after revision (07APR2015)

Resubmitted after revision (11APR2015)

Section 1 Statistical test

1.1

In PS3, we used the Mann-Whitney U test; the p value returned was one-tailed.
The null hypothesis is that in a randomly sampled pair of values (a, b) drawn from each of the sets (A, B) being compared, the likelihood of 'a' being ranked higher than 'b' is 0.5.

or $H_o : P(a > b) = 0.5$

The p critical can vary but is often taken to be 0.05 or 0.01.
I chose the 0.05 threshold for my interpretation.

1.2

The Mann-Whitney U test can be used to analyze populations of unknown distributions provided the sample size is large enough (> 20 from each population). There were 44104 & 87847 samples from the rainy & not rainy populations, respectively. So we are ok there.

1.3

The results from this application of the Mann-Whitney U test on these populations are:

The population means were 'rainy' 1105.4 'ENTRIESn_hourly' and 'not rainy' 1090.3 'ENTRIESn_hourly'.
The U statistic: 1924409167.0
The one-tailed p value: 0.024999912…

1.4

Using a p-critical of 0.05, we can reject the null hypothesis.

Note added in response to second review:

I appreciate the hint regarding rounding off the one-tailed p-value. That is in fact what I did, rounding to 0.025 and leading to a 2-tailed p-value of 0.05.
If we do not round up, then the 2-tailed p-value is formally below the threshold value of 0.05 and the null hypothesis can be rejected.

That said, the 0.05 threshold for statistical significance is a convention not a mathematical constant underpinning the structure of the universe. As such, it is a bit arbitrary and arguing that a p-value that is a 9th decimal point shy of 0.025 seals the question of significance still makes uncomfortable. Or am I missing your point?

I agree that using a 2-tailed test is the right way to go here. That is in fact what I did all along.

Randomly picking an 'ENTRIESn_hourly' value from a rainy day has a 0.5 (50/50) chance of being larger than a randomly picked 'ENTRIESn_hourly' value from a non rainy day.

Section 2 Linear Regression

2.1

I used gradient descent as in exercise 3.5

2.2

The input variables I started off with were: 'rain', 'precipi', 'Hour', 'meantempi', 'maxtempi', 'fog'. These were chosen on the intuition that they might affect ridership.

However, trying to simplify the model a bit, I removed one of these at a time & reran the regression w/ & w/o 'UNIT' as a dummy variable and gauging the effect on the basis of the resulting R^2.

Results are in the following Table.

Features	**R^2 w/ dummy UNIT**	**R^2 w/o dummy UNIT**
'rain', 'precipi', 'Hour', 'meantempi', 'maxtempi', 'fog'	0.4648	0.0335
'precipi', 'Hour', 'meantempi', 'maxtempi', 'fog'	0.4648	0.0335
'Hour', 'meantempi', 'maxtempi', 'fog'	0.4648	0.0335
'meantempi', 'maxtempi', 'fog'	0.4266	0.0018
'maxtempi', 'fog'	0.4258	0.0010
'fog'	0.4255	0.0007
None…	0.4251	0.0000
rain', 'precipi', 'Hour'	0.4633	
rain', 'Hour'	0.4632	
rain	0.4251	
Hour	0.4632	0.0317

Having only the UNIT as dummy feature gives an R^2 of 0.4251 which is very close to the more complex model with the additional features 'rain', 'precipi', 'Hour', 'meantempi', 'maxtempi', 'fog'.

This suggests that the best predictor of 'ENTRIESn_hourly' is the UNIT/ turnstile. This makes sense, there are likely stations that have more traffic for reasons unrelated to weather, e.g. location. It is also consistent with a quick look at the data showing low to zero entries at certain turnstiles irrespective of time of day.

The coefficient of the feature 'Hour' is 468.2.

The simplest model I found in this quick scan is "Hour" and "UNIT" as dummy. It had an R^2 of 0.4632.

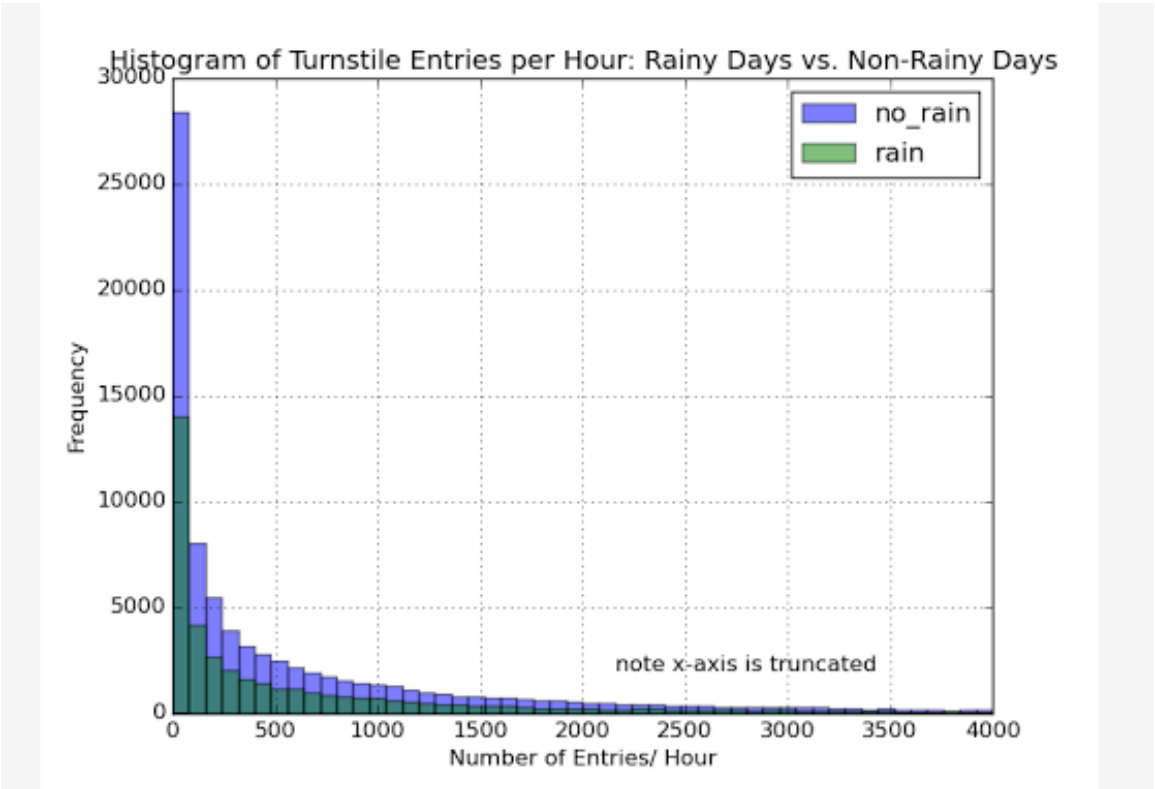
The R^2 value is the fraction of the variation explained by the model.
As a percentage, the model explains 46.3% of the variation in hourly entries.

Section 3. Visualization

3.1

http://pastebin.com/DiX2dm63

```
plt.figure()

   turnstile_rain = turnstile_weather[turnstile_weather['rain']==1]
   entries_rain = turnstile_rain['ENTRIESn_hourly']
   turnstile_norain = turnstile_weather[turnstile_weather['rain']==0]
   entries_norain = turnstile_norain['ENTRIESn_hourly']

   df = pandas.DataFrame({'no_rain':entries_norain, 'rain':entries_rain}, columns =
['no_rain', 'rain'])

   df.plot(kind='hist', bins= 50, range = (0,4000), alpha=0.5)
   plt.ylabel('Frequency')
   plt.xlabel('Number of Entries/ Hour')
   plt.title('Histogram of Turnstile Entries per Hour: Rainy Days vs. Non-Rainy Days')
   plt.text(2160, 2000, r'note x-axis is truncated')


   return plt
```



The distributions of number of entries/hour and their respective frequency in the dataset are similar for rainy and non-rainy days. The value of 0 entries/hour is the most frequent and the frequency diminishes with the increase of the entries/hour. The distributions are not normal and have a very long tail.

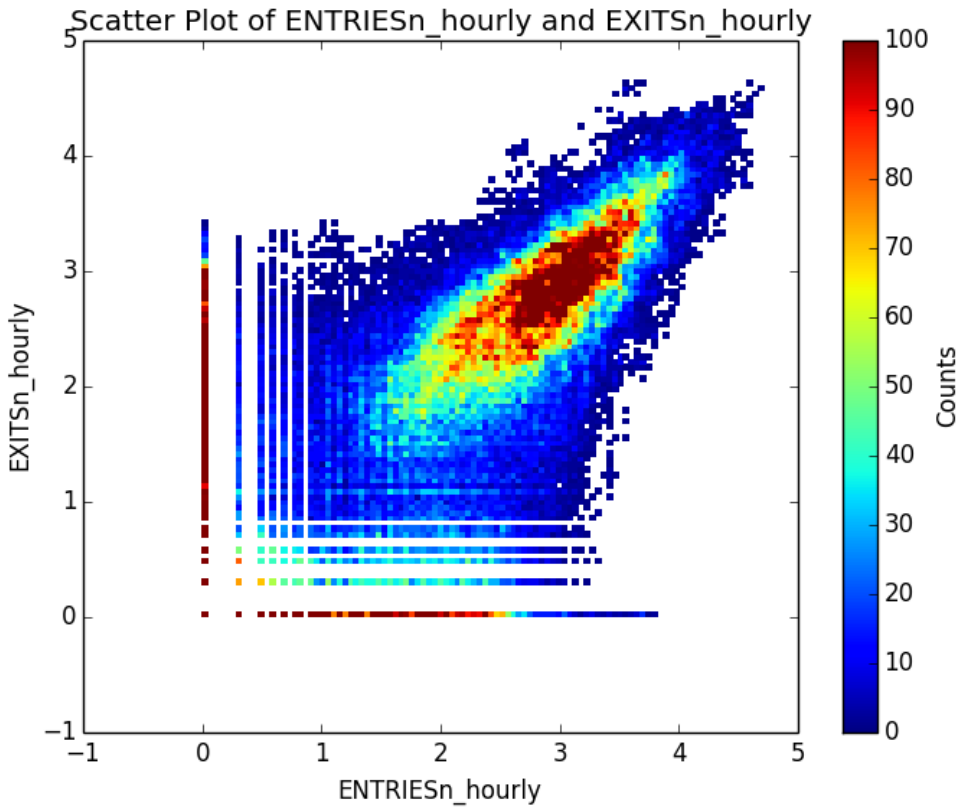
3.2

free form visualization

http://pastebin.com/wZCbwCha

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
import pandas

def entries_histogram(turnstile_weather):

    turnstile_weather['ENTRIESn_hourly'].replace(0,1, inplace = True)
    turnstile_weather['EXITSn_hourly'].replace(0,1, inplace = True)

    x = np.log10(turnstile_weather['ENTRIESn_hourly'])
    y = np.log10(turnstile_weather['EXITSn_hourly'])


    nbins = 100
    H, xedges, yedges = np.histogram2d(x,y,bins=nbins)

    H = np.rot90(H)
    H = np.flipud(H)

    Hmasked = np.ma.masked_where(H==0,H)

    fig2 = plt.figure()
    plt.pcolormesh(xedges,yedges,Hmasked)
    plt.xlabel('ENTRIESn_hourly')
    plt.ylabel('EXITSn_hourly')
    plt.ylim(-1,5)
    plt.xlim(-1,5)
    plt.title('Scatter Plot of ENTRIESn_hourly and EXITSn_hourly')
    cbar = plt.colorbar()
    plt.clim(0,100)
    cbar.ax.set_ylabel('Counts')
```

```
    plt.show()

turnstile_weather = pandas.read_csv('turnstile_data_master_with_weather.csv')
entries_histogram(turnstile_weather)
```



In general, the plot shows a correlation between level of entries and exits/ hour.

I thought it was neat that the data seems to reveal groups which are high exit/low entry, others the reverse. There isn't time to examine this but some could represent stations that are net destinations at a certain time of day (many exits & few entries, e.g. Manhattan) and the reverse at others times of day. Then again, some turnstiles might be like the roach motel.

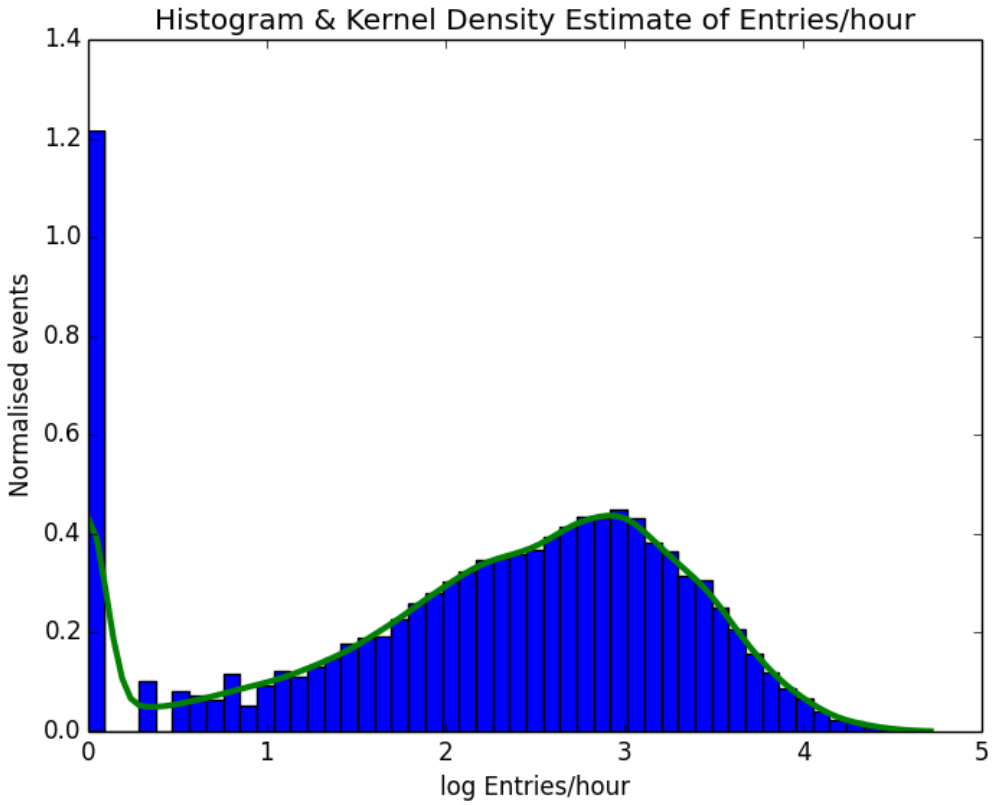
I did get around to plotting this log/log to provide a better separation at the origin. This was done by replacing '0's with '1's.

Note added in response to review:

The above plot proposes to address the overplotting limitation by adding a z-dimension giving the local plot density. This was done by using a 2D-histogram. I should add that I also took the invitation to do a more "free form" graph as license to try coming up with something I find visually appealing.

I don't know how to address the point about log-scale interpretation. Log-scale graphing is a standard tool to represent datasets spanning long ranges. Can you be more explicit regarding the interpretation problem?

I did give the kernel density estimation a whack on this dataset. The result, using 'Entries/hour', is below.
It is a neat tool. I could not however get a kde representation of x/y data.



Section 4

Conclusion

4.1

If we assume/accept that 'ENTRIESn_hourly' is a good measure of ridership, then more people ride the NYC subway when it is raining.
A simple way to answer this is to split 'ENTRIESn_hourly' into a 'rain' series and a 'non_rain'series and get descriptive statistics on these 2 sets. Doing so gives mean 'ENTRIESn_hourly' of 1105 and 1090 for the "rain' and 'no_rain' sets, respectively. However that difference is relatively small (~ 1.38%).
The sum of the number of datapoints in these 2 sets is equal to the total number of datapoints in the "turnstile_data_master_...", which suggests we are not missing a fraction of the dataset, which could lead us astray.

It is possible to think of scenarios where 'ENTRIESn_hourly' would not be a good measure of ridership. For example, if on rainy days, passengers took shorter trips than on non rainy days, then by some measures (say 'miles traveled', or some instantaneous survey of people in the subway) the ridership on non rainy days could be higher.

4.2

From descriptive statistics, as stated in 4.1, the mean 'ENTRIESn_hourly' for rainy days is larger than that for non rainy days, 1105 and 1090, respectively. However that difference is small, ~ 1.38%.
In fact, the standard deviation and CV for both these values are very large, which is a indication that we have to be cautious about the meaningfulness of this difference.

	Rain	**No rain**
Mean (entries/hour)	1105	1090
St. dev. (entries/hour)	2370	2320
CV (%)	214	212

From inferential statistics using the Mann-Whitney test we get a U statistic of 1924409167.0 and a one-tailed p value of 0.024999912....
This value has to be doubled since we are applying a 2-tailed test. The result is formally < 0.05. Therefore, we can reject the null hypothesis.

From the linear regression, using a [['rain', 'Hour', 'fog']] model with "Units" as dummy variable, I get thetas of -1.39767820e+00, 4.68330215e+02 , and 4.94197397e+01, respectively. The theta of the 'rain' feature is small (-1.39) indicating that 'rain' as a feature does not have a strong effect on the value 'ENTRIESn_hourly'.

Note added in response to review:

As I see it, the evaluation of the effect of rain on 'Entries_hourly' using descriptive statistics, inferential statistics and linear regression are consistent.
Descriptive statistics indicate a small difference in the means tempered by a large variation in the groups compared (evidenced by the CV). With a commonly used graphing convention where groups are plotted with their means +/- their standard deviation, these groups would substantially overlap.
Inferential statistic analysis returns a p value below the lowest generally accepted level of confidence ($p < 0.05$). This would indicate that we can reject the null hypothesis.
Linear regression analysis indicates that the theta for rain as a feature is small, which is again consistent with a small (emphasis small) effect of rain on ridership. Interestingly, the theta is negative, which is in the opposite direction of the nominal difference between group means. A confidence interval around this theta would likely include 0.
In conclusion, there appears to be a real though small effect of 'rain' on 'ridership' in this dataset.
It still would be interesting to try trimming the dataset to exclude some of the outliers (e.g. stations datapoints with 0 entries and exits) and retest the effect of rain on ridership.

Section 5

Reflection

5.1

Potential shortcomings of the methods of analysis as they relate to:

1. The dataset

How adequate the dataset is depends on the inferences sought from it. This dataset covers a single month in a single year. If we are looking to understand May 2011 we can have a greater degree of confidence than if we extrapolate to, say, another month, a year in the future or another in the dim distant past.
Every feature can and should be examined critically. Rain is not really a binary, is it cats or is it dogs today? How and where is it measured, how often updated?
Someone more familiar with the dataset may understand the geography of the measurements, I did not.
Professionally, I take measurements and interpret them. I also get information from technical publications in which authors have a responsibility to explain their methods and math as clearly as possible/practical. Even then, questions remain (abound).
Data analysis as it appears at the moment seems to work from datasets that are not necessarily well characterized.
Depending on the goals of the analysis, there may be instances where it is important to do background work to validate or at least understand the limits of the data.

2. The analysis

One of the analytical shortcomings that strikes me is that we used a linear model for each feature. Many dose/response curves are, well, curves. Many biological responses in particular are linear only in a restricted range and sigmoid in a broader one.
There were hints of polynomial regression models to come. I look forward to those. I imagine there are models where one of the 'variables' when fitting a feature to value relationship is which equation form to use and include these outcomes in the cost function.

Regarding statistical analysis, my lack of experience with the strengths and weaknesses of different tests precludes much of a critical understanding of its limitations. Hope to remedy that sometime soon. I did take "Inferential Stats' but still.

Sources:

I consulted a long list of sites in the course of this class and did not make a note of them over these past weeks.
The list would include Stackoverflow, scipy.org, previous classes in Udacity, pandas.org , Wikipedia, Graphpad, Prism, and many others.

Resubmission used http://oceanpython.org/2013/02/25/2d-histogram/
As a starting point for the 2 D histogram.