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

I used OLS using "Statsmodels" to compute the coefficients theta and produce prediction for ENTRIESn_hourly in regression model.

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

**(Note: Below features are related to advance data set)**

**Features:**

1) ~~'rain' : Whether there was a rain or not at time and location~~
   I thought it should be a main feature but after adding other features there was no impact of 'rain' in the model.
2) 'tempi' : Temperature at time and location

**Dummy Variables:**

1) 'hour' : Hour of the timestamp
2) 'UNIT' : Remote unit that collects turnstile information
3) 'day_week' : Integer (0-6) corresponding to the day of the week

**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.**

**(Note: Below features are related to advance data set)**

**Features: Reason**

1) 'tempi' : I think the temperature at time and location affects the use of subway trains. As a human tendency, we avoid to go out if outside temperature is very high or low. Whereas in the pleasant atmosphere people plan to go out in the city.

2) 'hour'(dummy) : I think the use of subway is highly depend on the particular hour(time). From my own experience I can say the use of subway is really high during the office hours. Along with that, there are hardly few people commute during the late night.

3) 'UNIT'(dummy) : I believe the use of the subway differ from the station to station as the stations at the downtown location and corporate places will be crowdie where as in the suburb it won't be.

4) 'day_week'(dummy) : I think the use of subway would be very high during the week days due to office/work where as the use of subways on weekend would be low.

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

| Feature | Co-efficient |
|---|---|
| const | 629.9915 |
| tempi(normalized) | -119.8407 |

**2.5 What is your model's $R^2$ (coefficients of determination) value?**

**(Note: Below $R^2$ is related to advance data set)**
$R^2$ = 0.545

**2.6 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?**

$R^2$ value is a statistical measure of how close the data are to the fitted regression line. Which is between 0% to 100%.

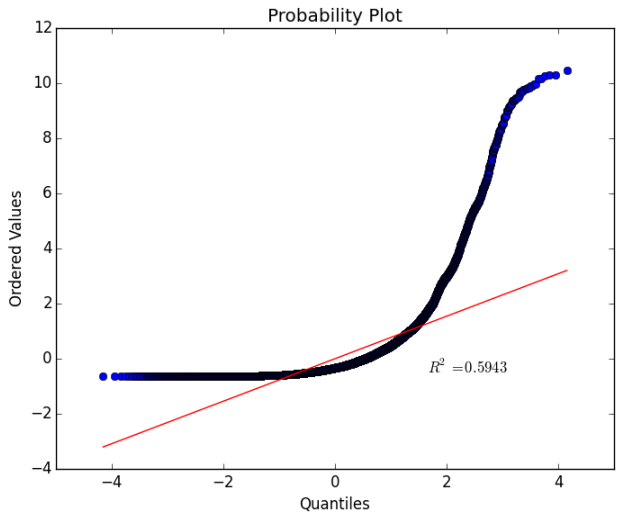
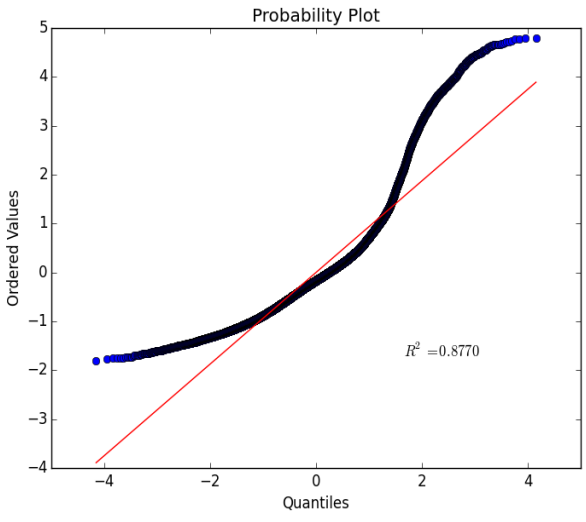
- 0% indicates that the model explains none of the variability of the response data from its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

Here the $R^2$ value for the prediction model is 0.545. Which means that the model explains the 54.5% variability of the response data from mean.

**The correctness of linear model:**

I believe it's difficult to judge any prediction model based upon the r-squared value. The acceptance/judgment parameters for r-squared value differ from problem to problem.

For the given dataset, I think the linear model is not appropriate to use. This can be concluded based upon the below QQ-plots. In the both QQ-plots it's clearly visible that it's not appropriate to represent the curved shape distribution by any best-fit line, as any linear line won't be able to mimic the curve representation.

| QQ Plot from the given number of entries per hour. | QQ Plot from the predicted number of entries per hour by linear regression model. |
|---|---|
|  |  |