Suggested Solution for Homework 4

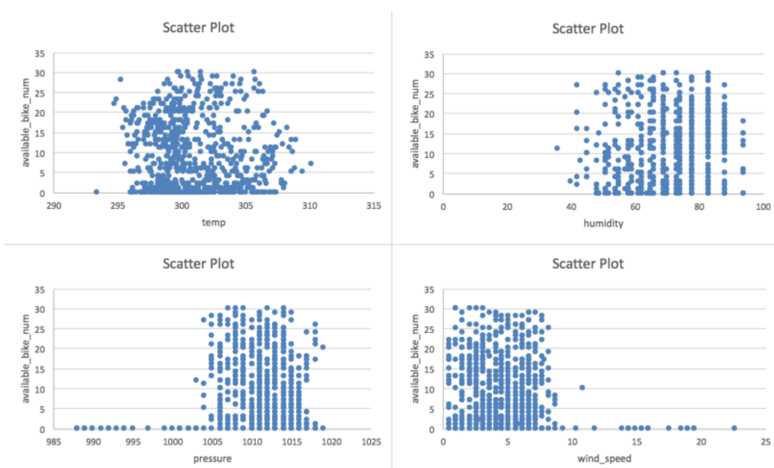
Statistics and Data Analysis, Fall 2015

1. (10 points)

Month is qualitative variable. It would be useful because seasonality may influence number of available bikes. *Day*, *weekday* and *hour* are qualitative variables. They would be useful since time factors may affect number of available bikes as well. For *station_id*, *location*, *station_name* and *station_address*, since we only focus on one specific site, these qualitative variables are not useful. *Total_parking_num* is quantitative variable. It also not useful since it is a fix number for a specific site. *Empty_parking_num* is equivalent to *available_bike_num* since it is the difference between *total_parking_num* and *available_bike_num*. *Latitude* and *longitude* are not suitable for simple regression model. *In_service* is qualitative variable. It is useful because if the station is out of service, there would be no available bikes at all. *Weather_type* is qualitative variable. *Temp*, *pressure*, *humidity* and *wind_speed* are quantitative variables. Intuitively, weather variables would have some impact on whether people want to ride a bike or not. They would be useful in a regression model.

2. (30 points)

(a) Use scatter plot to help us consider quantitative variable selection. *wind_speed* and *temp* seems to have slightly negative effect on *available_bike_num*. However, it seems that no obvious pattern can be observed. We may consider some transformation or interaction terms.



(b) (5 points)

$$available\_bike\_num = 66.5106 - 0.1899temp$$

Coefficient of *temp* is –0.1899. R-squared is 0.0044. *p*-value of temp is 0.06829. It seems the model need more variables or some transformation.

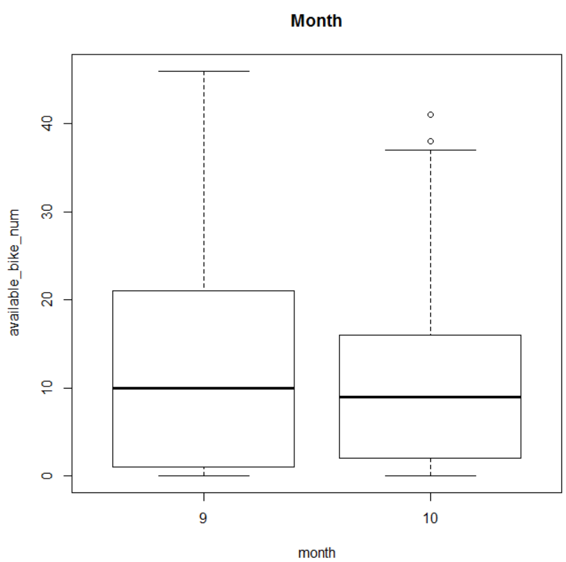
(c) (5 points)

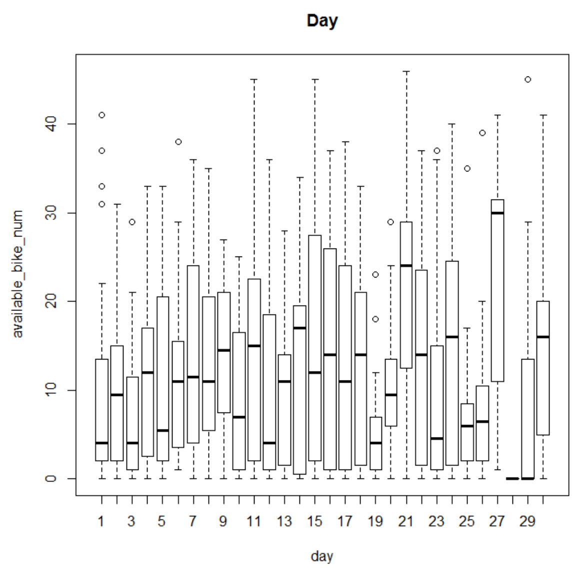
$$available\_bike\_num = 94.4060 - 0.1838temp - 0.0254pressure - 0.0224humidity - 0.5104wind\_speed$$

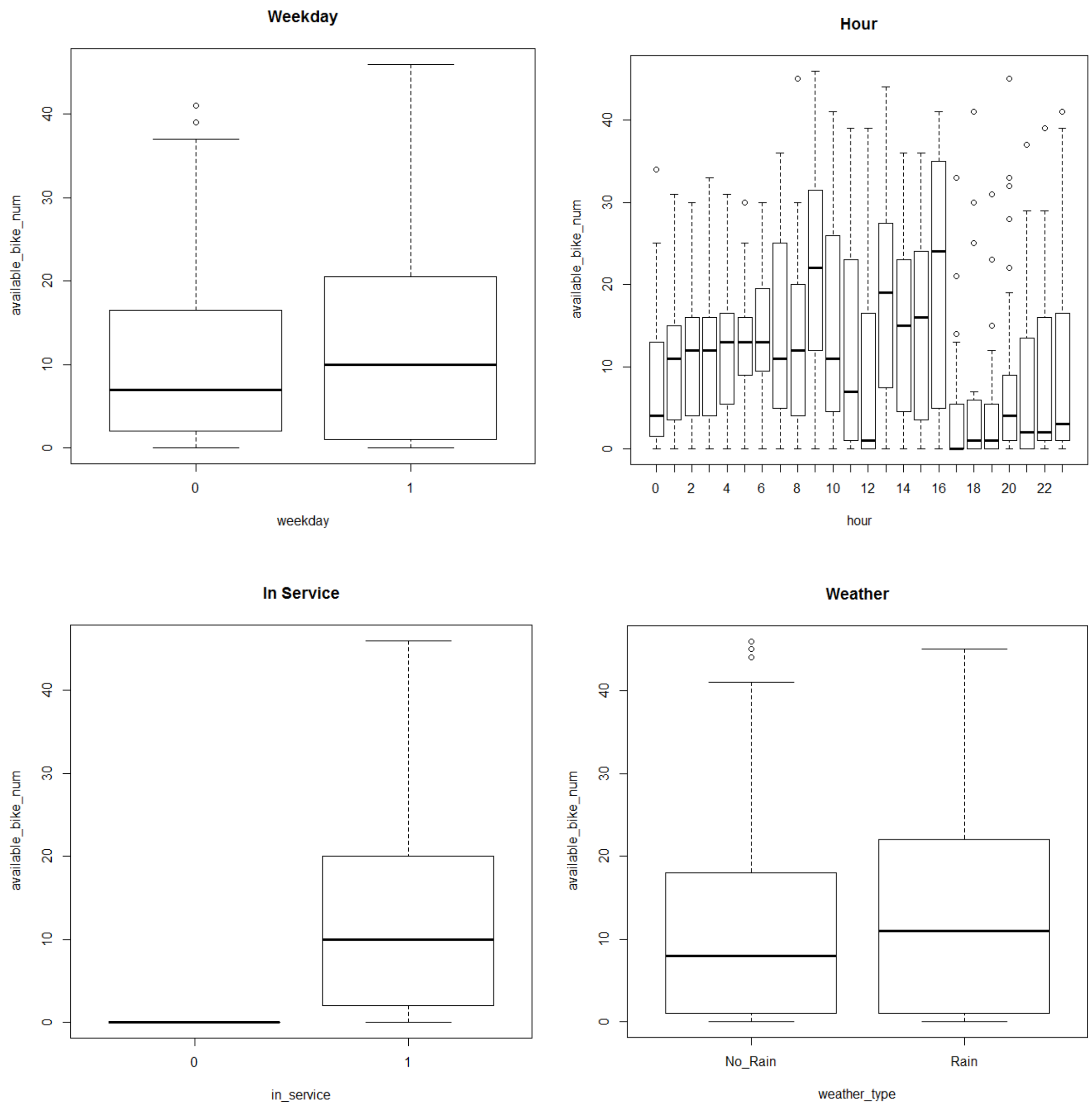
R-squared is 0.029 only. *wind_speed* is the only variable significant, whose *p*-value is approximately to 0. It seems that the model need some transformation and further refinement.

3. (35 points)

(a) (10 points)







We can see by the boxplots that the averages of $available\_bike\_num$ are not significantly different in different $month$, $weekday$, and $weather\_type$.[1] We might want to try day and $hour$ as our independent variable to do the regression.

(b) (5 points)

$available\_bike\_num = 10.64583 + 1.53384 weekday$

[1] A box plot is a way to describe the distribution of a set of numeric values. The box contains data points within the first and third quartiles, and the think bar indicates the location of the median. Small circles represent data points far from the median by more than 2.5 interquartile ranges. When two box plots are put together, a larger box typically means that the corresponding distribution is more disperse.

If weekday changes from 0 to 1, we expect the available bike number to increase by 1.53384. With R^2 = 0.00355, adjusted R^2 = 0.0022, which are really small; and the p-value is 0.1048 for $weekday$, which is not small enough, we know that this model should be modified.

(c) (5 points)

$$
\begin{aligned}
available\_bike\_num &= 9.225806 + 2.580645hour^{(2-3)} + 3.3870967hour^{(4-5)} \\
&+ 5.225806hour^{(6-7)} + 7.758065hour^{(8-9)} + 4.548387hour^{(10-11)} \\
&+ 4.080645hour^{(12-13)} + 5.516129hour^{(14-15)} + 3.354839hour^{(16-17)} \\
&- 4.48387hour^{(18-19)} - 0.82258hour^{(20-21)} - 0.50449hour^{(22-23)}
\end{aligned}
$$

If hour changes from $0-1$ to $i-(i+1)$ and all others remain the same, we expect the dependent variable to increase by β_{i-i+1}. With R^2 = 0.08356, adjusted R^2 = 0.06977, and more significant p-values for the independent variables, we know that this model does better than part (b), but should still be improved since R^2 is not good enough.

(d) (5 points)

$$
\begin{aligned}
available\_bike\_num &= 14.74194 - 5.51613hour^{(0-1)} - 2.93548hour^{(2-3)} - 2.12903hour^{(4-5)} \\
&- 0.29032hour^{(6-7)} + 2.241935hour^{(8-9)} - 0.96774hour^{(10-11)} \\
&- 1.43548hour^{(12-13)} - 2.16129hour^{(16-17)} - 10hour^{(18-19)} \\
&- 6.33871hour^{(20-21)} - 6.02062hour^{(22-23)}
\end{aligned}
$$

There are fewer significant variables than in the model in part (c). It might be so because hour 15-16 in around the middle of a day, the average available bike number would be somewhat similar to other hour ranges within a day; however, hour 0-1 is in the early morning of a day, the average available bike number would be very different from those in rush hours.

4. (25 points)

(a) (5 points)

Intuitively, $available\_bike\_num_{t-1}$ may be a good predictor for $available\_bike\_num_t$ since the number would be somewhat close to each. $available\_bike\_num_{t-2}$ and

$available\_bike\_num_{t-3}$ would also be good predictors since the time is not too far away from each other, and they will help you see the trend within those hours.

(b) (10 points)

$$available\_bike\_num_t = 6.338023 + 0.464603 available\_bike\_num_{t-1}$$

If the available bike number at time t – 1 increases 1, we expect the available bike number at time t to increase by 0.464603. With R^2 = 0.21582, adjusted R^2 = 0.21476, we know that $available\_bike\_num_{t-1}$ can explain 21.6%, and it is really significant since its p-value is small enough.