Statistics and Data Analysis, Fall 2015
Homework 4: Regression Analysis

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

This homework is due ***6:35 pm, December 21, 2015***. Each student should submit her/his own ***hard copy*** to the instructor at the beginning of the class. All the data for this homework are contained in the file "SDA-Fa15_hw03_data.xlsx," i.e., the same as Homework 3. Discuss with your classmates but NEVER copy one's work.

Just to make grading easier, for all the problems below, please construct your answers with ALL the data given to you. DO NOT try to remove any potential outliers.

1. (10 points) Our ultimate objective is to build a model to predict future availability given a site and a time point. For example, we want want to know the number of available bikes at "MRT Gongguan Sta.(Exit 2)" during 8 am to 9 am on December 25, 2015. To do so, we will try to build a regression model. Note that the rental patterns at different sites can be quite different. Therefore, let's do our analysis once for a site. Imagine that now one site has been assigned to you. Look at each variable (column) provided to you and comment on whether it may be useful in a regression model. If it is, comment on whether it is quantitative or qualitative; otherwise, explain why. The variable *available_bike_num* is obviously our dependent variable.

2. (30 points) Consider the site "MRT Gongguan Sta.(Exit 2)."

   (a) (10 points) Look at all the *quantitative* variables that may be included in the regression model. Use descriptive statistics to give some directions to do variable selection and transformation.

   (b) (5 points) Construct a regression model

   $$available\_bike\_num = \beta_0 + \beta_1 temp.$$

   Find the coefficients and validate the model.

   (c) (5 points) Construct a regression model

   $$available\_bike\_num = \beta_0 + \beta_1 temp + \beta_2 pressure + \beta_3 humidity + \beta_4 wind\_speed.$$

   Find the coefficients and validate the model.

   (d) (10 points) Limit yourself to quantitative variables, give the best regression model that you may find. Find the coefficients and validate the model.

3. (35 points) Consider the site "N.T.U.S.T."

   (a) (10 points) Look at all the *qualitative* variables that may be included in the regression model. Use descriptive statistics to give some directions to do variable selection and transformation.

   (b) (5 points) Construct a regression model

   $$available\_bike\_num = \beta_0 + \beta_1 weekday.$$

   Find the coefficients and validate the model.

   (c) (5 points) Let's include *hour* into our model. Obviously, *hour* is a qualitative variable, so let's divide a day into 12 two-hour periods (0-1, 2-3, ..., and 22-23). If we choose 0:00-2:00 as the reference level, we may construct a regression model

   $$available\_bike\_num = \beta_0 + \beta_1 hour^{(2\text{-}3)} + \beta_2 hour^{(4\text{-}5)} + \cdots + \beta_{22\text{-}23} hour^{(22\text{-}23)},$$

   where $hour^{(i\text{-}i+1)} = 1$ if $hour \in \{i, i+1\}$ or 0 otherwise, $i = 0, 2, 4, ..., 22$. Find the coefficients and validate the model.

(d) (5 points) Continue from Part (c). Change the reference level to 14-15 and then redo the regression analysis. Compared to the model in Part (c), are there more or fewer significant variables? Intuitively explain why.

(e) (10 points) Limit yourself to quantitative variables, give the best regression model that you may find. Find the coefficients and validate the model.

4. (25 points) For a given site, the numbers of available bikes at different time points can be ordered in time to form a *time series*. For example, the time series of available bikes at the site "NTU Information Bldg." is depicted in Figure 1.[1]

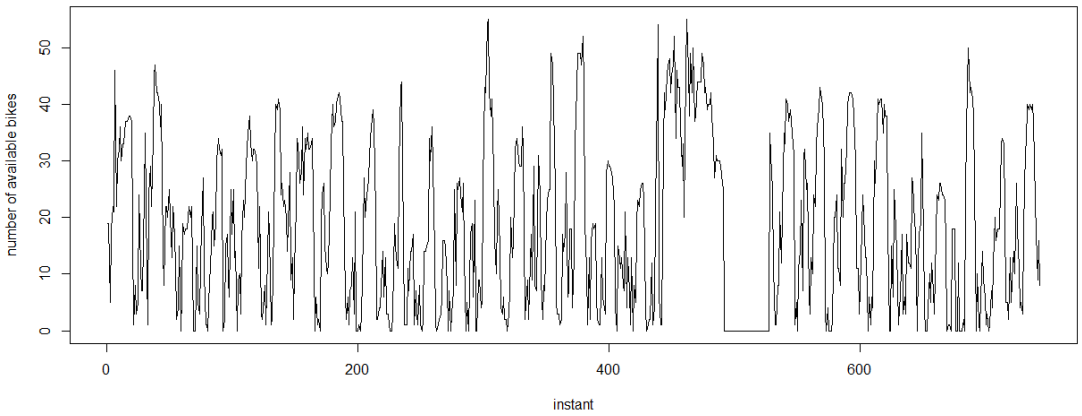

Figure 1: Time series of available bikes at "NTU Information Bldg."

For a time series data set, we may consider *autoregression*, i.e., using the dependent variable *in previous periods* to be independent variables. For example, if we let $available\_bike\_num_t$ as the number of available bikes at time t, then we may add $available\_bike\_num_{t-1}$ as an independent variable. For "NTU Information Bldg.," these variables are shown below.

t	$available\_bike\_num_t$	$available\_bike\_num_{t-1}$
1	19	N/A
2	19	19
3	5	19
4	18	5
5	22	18
⋮	⋮	⋮
741	9	17
742	16	9
743	8	16

For the numbers of available bikes, it is intuitive that $available\_bike\_num_{t-1}$ may be a good predictor for $available\_bike\_num_t$ (why?).

(a) (5 points) Intuitively explain why $available\_bike\_num_{t-1}$ may be considered a good predictor for $available\_bike\_num_t$. How about $available\_bike\_num_{t-2}$, $available\_bike\_num_{t-3}$, etc.?

(b) (10 points) For "NTNU Library," construct a regression model

$$available\_bike\_num_t = \beta_0 + \beta_1 available\_bike\_num_{t-1}.$$

Find the coefficients and validate the model. Note that the first row ($t = 1$) must be excluded.

(c) (10 points) Give the best regression model that you may find. Find the coefficients and validate the model.

[1]The consecutive zeros are due to a system failure. While in practice we will remove them, please simply ignore this failure and assume that they are not outliers in this homework.