Statistics and Data Analysis
Suggested Solution for Homework 6

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

Below we will demonstrate the regression steps by R. The data stored in the two MS Excel sheets are first stored in two plain text files "Bike_Day.txt" and "Bike_Month.txt." To use MS Excel for the analysis, use "Data Analysis" and then "Regression."

1. (a) By the R codes

```
M <- read.table("Bike_Month.txt", header = TRUE, sep = "\t")
plot(M$cnt, type = "l", xlab = "instant", ylab = "cnt")
```

we may generate the line chart as Figure 1. We can see an increasing trend. Moreover, the demands during summer time seem to be higher than those during winter time.

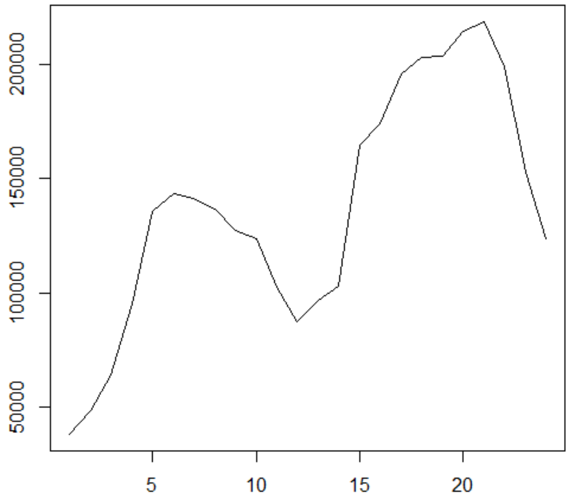

Figure 1: Monthly rentals

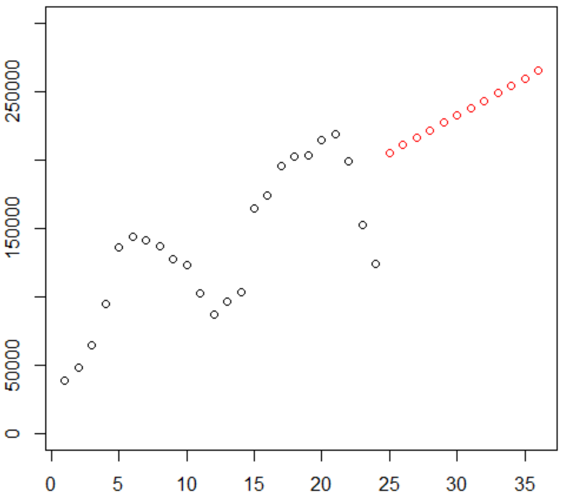

Figure 2: Prediction by *instant*

(b) Continue from Part (a), by the following R codes

```
fit <- lm(M$cnt ~ M$instant)
summary(fit)
```

we may find the intercept as 69033 and the slope as 5453. The regression line is $y = 69033 + 5453x$, where y is the monthly rental and x is the number of months since the beginning. In fact, once one gets `fit` one may use `fit$coefficient[1]` and `fit$coefficient[2]` to get the intercept and slope.

(c) By the report we get in Part (b), the R^2 is 0.5442 and the p-value of the slope is 3.89×10^{-5}. Around 54.42% of the rental volume in a month can be explained by the number of months since the beginning. Moreover, at a 99% confidence level, the number of months since the beginning has a nonzero impact on the monthly rental.

(d) To predict future monthly rentals, we only need to plug in x by an integer that represents the number of months since the beginning. For example, if we want to predict for the next March, we should set $x = 27$. Following this way, the predicted monthly rental of the next year are 205356.7, 210809.6, ..., and 265339. These predicted values are depicted as red points in Figure 2 (with historical values depicted as black points).

2. (a) By the following R codes

```
M <- read.table("Bike_Month.txt", header = TRUE, sep = "\t")
fit <- lm(M$cnt ~ M$instant + M$month)
summary(fit)
```

we may find the regression line as $y = 71045.5 + 5600.5x_1 - 593.4$, where y is the monthly rental, x_1 is the number of months since the beginning, and x_2 is the month number. This is not a valid regression model because month numbers are *ordinal* data rather than *quantitative* data. In fact, if we predict monthly rentals for the next year (as green points in Figure 3), we will not really get any seasonal effects.

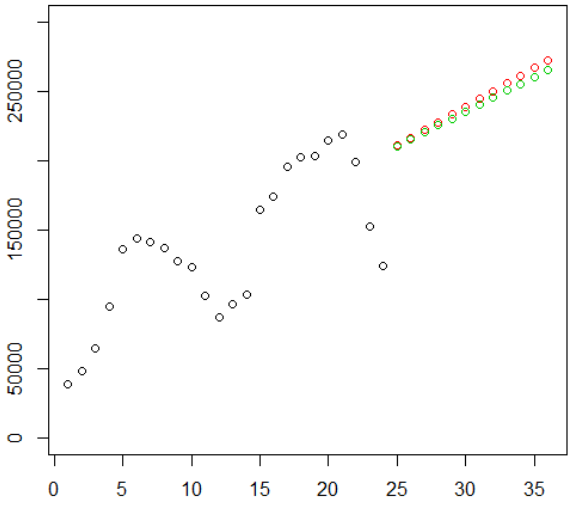

Figure 3: Prediction by *instant* and *month*

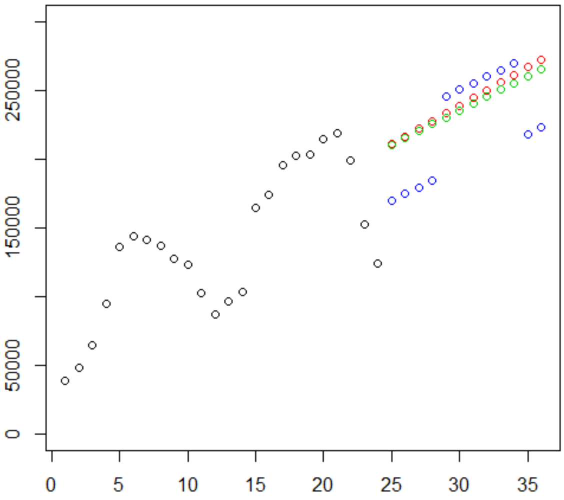

Figure 4: Prediction by *instant* and *high*

(b) The following R codes

```
avg1 <- mean(M$cnt[which(M$year == 0)])
avg2 <- mean(M$cnt[which(M$year == 1)])
month1 <- M$month[which(M$year == 0 & M$cnt > avg1)]
month2 <- M$month[which(M$year == 1 & M$cnt > avg2)]
```

find those months. For year 1, months 5 to 10 have above average monthly rentals. For year 2, months 4 to 10 have above average monthly rentals.

(c) The following R codes

```
high1 <- rep(0, 12)
high2 <- rep(0, 12)
high1[month1] <- 1
high2[month2] <- 1
high <- high1 * high2
high <- rep(high, 2)
```

sets up the `high` vector according to our definition. Then the following R codes

```
fit <- lm(M$cnt ~ M$instant + high)
summary(fit)
```

find the regression line $y = 48227.5 + 4865.5x_1 + 56298x_2$, where y is the monthly rental, x_1 is the number of months since the beginning, and $x_2 = 1$ if the month is a high-demand month and 0 otherwise.

The long-term increasing trend is captured by the coefficient of x_1: When one month passes by, we expect to get 4865.5 more monthly rentals. The seasonal effect is captured by the coefficient of x_2: If a month is in the peak season, we expect to get 56298 more monthly rentals.

(d) The predicted monthly rentals (as blue points) are depicted in Figure 4. We can see that the seasonal effect is somewhat captured (and the prediction seems to be more reasonable).[1]

3. (a) The following R codes

```
D <- read.table("Bike_Day.txt", header = TRUE, sep = "\t")
fit <- lm(D$cnt ~ D$instant)
summary(fit)
```

find the regression model $y = 2392.96 + 5.7688x$, where y is the daily rental and x is the number of days since the beginning. As the p-value of *instant* is almost 0, the coefficient of *instant* is significantly greater than 0. There is indeed an increasing trend.

(b) For the regression line for daily rentals, the slope is 5.7688, which means that we expect to get 5.7688 more daily rentals when one day passes by. Equivalently, we expect to get around $5.7688 \times 30 = 173.0645$ more daily rentals, or $173.0645 \times 30 = 5191.936$ for each month passes by. As $5191.936 < 5453$, the slope we found in Problem 1b, the regression line for daily rentals is flatter. This may because that the last few days have unreasonably low daily rentals, but when they are aggregated into the last monthly rental, such an extreme impact disappear. Another possible reason is that for daily rentals there are weekly fluctuations, but weekly fluctuations do not appear for monthly rentals.

(c) The following R codes

```
fit <- lm(D$cnt ~ D$instant + D$holiday)
summary(fit)
```

find the regression model $y = 2414.8471 + 5.7804x_1 - 909.9990x_2$, where y is the daily rental, x_1 is the number of days since the beginning, and $x_2 = 1$ if it is a holiday and 0 otherwise. Both independent variables are significantly not zero at a 99% confidence level. In average, the daily rental in a holiday is smaller than that in a non-holiday by 909.99. This is probably a hint that people tend to choose other recreation activities rather than bicycling.

(d) The following R codes

```
fit <- lm(D$cnt ~ D$instant + D$workingday)
summary(fit)
```

find the regression model $y = 2210.081 + 5.7714x_1 - 266.0111x_2$, where y is the daily rental, x_1 is the number of days since the beginning, and $x_2 = 1$ if it is a working day and 0 otherwise. Both independent variables are significantly not zero at a 5% confidence level. In average, the daily rental in a working day is larger than that in a non-holiday by 266.0111. This is probably a hint that most people rent bicycles for communicating between home and working places, not for recreation.

4. (a) The following R codes

```
D <- read.table("Bike_Day.txt", header = TRUE, sep = "\t")
fit <- lm(D$cnt ~ D$temp + D$atemp + D$hum + D$windspeed)
summary(fit)
```

find the regression model as $y = 3860.4 + 2111.8x_1 + 5139.2x_2 - 3149.1x_3 - 4528.7x_4$, where y is the daily rental, x_1 is the temperature, x_2 is the adjusted temperature, x_3 is the humidity, and x_4 is the wind speed.

(b) The regression model in (a) has $R^2 = 0.4638$, which is not satisfactory. Moreover, the p-value for *temp* is 0.3551, which suggests that *temp* may not have a significant impact (when put in the model with the other three variables).

(c) We take away *temp* and redo the regression analysis by

[1] Of course there are other methods to capture the seasonal effect. Courses like "Operations and Service Management" may introduce these methods.

```
fit <- lm(D$cnt ~ D$atemp + D$hum + D$windspeed)
summary(fit)
```

This time all three independent variables are significant. The regression model is $y = 3774 + 7504.1x_1 - 3167.5x_2 - 4411.7x_3$, where y is the daily rental, x_1 is the adjusted temperature, x_2 is the humidity, and x_3 is the wind speed. The R^2 is 0.4632, almost identical to that with four independent variables. This suggests that the three variables are as good as the four variables in explaining daily rentals. This is probably because that people decide whether to rent bicycles based on the adjusted temperature (which is the temperature that they feel), and temperature is just close to adjusted temperature.

(d) The following R codes

```
D <- read.table("Bike_Day.txt", header = TRUE, sep = "\t")
fit <- lm(D$cnt ~ D$instant + D$atemp + D$hum + D$windspeed)
summary(fit)
```

find the regression model as $y = 2044.488 + 4.8908x_1 + 6622.3024x_2 - 2976.2982x_3 - 3164.0346x_4$, where y is the daily rental, x_1 is the number of days since the beginning, x_2 is the adjusted temperature, x_3 is the humidity, and x_4 is the wind speed. All independent variables are significant at a 99% confidence level. Moreover, $R^2 = 0.7385$ and adjusted $R^2 = 0.7371$. The model is greatly better than the one without *instant* for capturing the increasing trend.