Statistics and Data Analysis
Suggested Solution for Homework 7

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

1. (a) By executing the following R codes

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

we get the regression model as

$$cnt = 3822.38 + 5.71 instant - 1009.74 weathersit. \tag{1}$$

(b) Because *weathersit* is a categorical variable with values not 0 and 1, using it in building a regression model is problematic. Even though the p-value of *weathersit* is small and the R^2 of the model is not small, the coefficient -1009.74 has no physical meaning.

2. (a) By executing the following R codes

```
D <- read.table("Bike_Day2.txt", header = TRUE, sep = "\t")
fit <- lm(D$cnt ~ D$instant + D$cloudy + D$rainy)
summary(fit)
```

we get the regression model as

$$cnt = 2776.28 + 5.7 instant - 813.69 cloudy - 2889.1 rainy. \tag{2}$$

For a cloudy day, we expect to get 813.69 fewer rentals than a sunny day. For a rainy day, we expect to get 2889.1 fewer rentals than a sunny day. Please note that we have no idea about the difference between the rentals in a cloudy and a rainy day. This is because that we use the sunny day as the reference level (the one encoded with all 0s).

(b) The regression model is fine. All variables that we are using are either quantitative variables (*instant*) or indicator variables (*cloudy* and *rainy*). The coefficients all have physical meanings.

3. (a) We get the model in (1). It is a bad regression model.

(b) We get the model in (2). It is a good regression model.

(c) We also get the model in (2). It is good.

4. (a) By executing the following R codes

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

we get the regression model as

$$cnt = 41323.9 + 5568 instant + 55282.3 season2 + 54413.6 season3 - 4609.3 season4.$$

Compared to months in season 1 (the reference level), for months in season 2 we expect to get 55282.3 more rentals, for months in season 3 we expect to get 54413.6 more rentals, and for months in season 4 we expect to get 4609.3 fewer rentals. However, because the p-value for *season4* is 0.749, there is no significant difference between seasons 1 and 4. The $R^2 = 0.8569$ shows that around 86% of monthly rentals can be explained by the increasing trend and seasonal impact.

(b) By executing the following R codes

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

we get the regression model as

$$\begin{aligned} cnt = 28263 &+ 5600.5 instant + 2609 month_2 + 35792.5 month_3 \\ &+ 50279 month_4 + 75974.5 month_5 + 77702.0 month_6 \\ &+ 71404.5 month_7 + 68927.0 month_8 + 60724.9 month_9 \\ &+ 43304.9 month_{10} + 3943.9 month_{11} - 23554.1 month_{12}. \end{aligned}$$

For each month, the coefficient is the expected difference on monthly rentals between January and that month. However, because the p-value for *month2*, *month11* and *month12* are all larger than 0.1, there is no significant difference between January and one of the three months at the 90% significance level. The $R^2 = 0.9712$ shows that around 97% of monthly rentals can be explained by the increasing trend and monthly differences.

5. (a) One possible way to answer this question is to construct a linear regression model with *year* and *price*. By executing the R codes

```
C <- read.table("Car.txt", header = TRUE, sep = "\t")
fit <- lm(C$Price ~ C$Year)
summary(fit)
```

we get the regression model as

$$price = 440.54 - 47.6 year.$$

When one year passes, the expected price of a house decreases by \$476,000. The variable is significant and the R^2 is 0.7212. The residuals for this model are depicted in Figure 1.

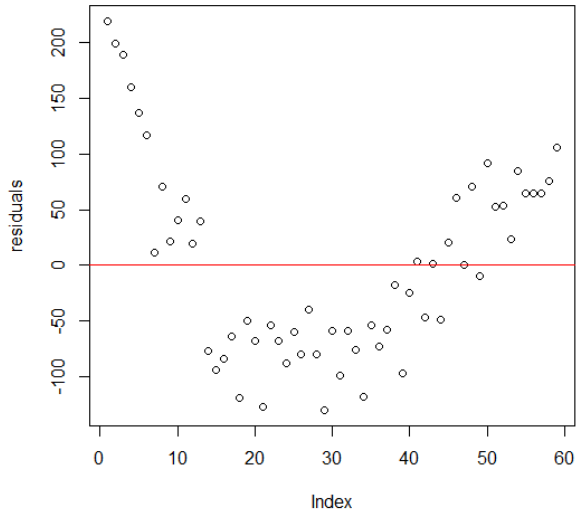

Figure 1: Residuals for Problem 5a

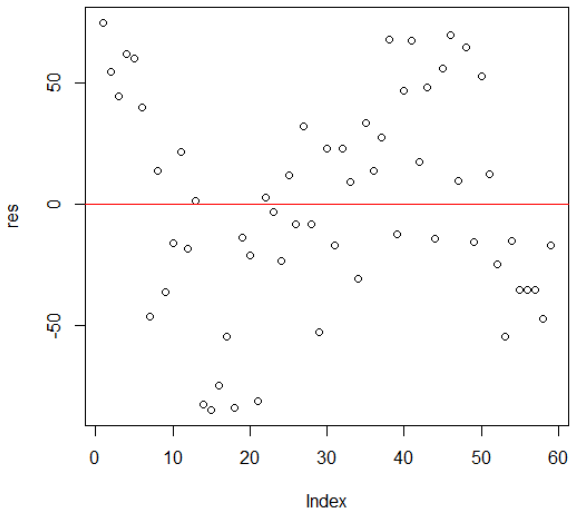

Figure 2: Residuals for Problem 5b

(b) By executing the R codes

```
YearSq <- C$Year^2
fit <- lm(C$Price ~ C$Year + YearSq)
summary(fit)
```

we get the regression model as

$$price = 585.535 - 145.798 year + 10.354 year^2.$$

Both variables are significant. Moreover, the R^2 goes up to 0.9296 (and the adjusted R^2 is also higher than that with no $year^2$). The new regression curve better fits the sample data. The residuals for this model are depicted in Figure 2.

(c) The residuals of the first model has an unnatural pattern: The model underestimates the prices of new and old cars but overestimates the prices of those cars with moderate ages. The residuals of the second model does not have such an obvious pattern. Therefore, the second model is better.