

Statistics and Data Analysis

Homework 7: Regression (2)

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

1. The “Bike_Day” sheet in “SDA-Fa14_cp13_data.xlsx” contains the daily number public bike rentals in a city and other related information. The data were collected in the last two years.
 - (a) For the variable *weathersit*, 1 means sunny, 2 means cloudy, and 3 means rainy or snowy. Construct a regression model for *instant*, *weathersit*, and *cnt*. What are the coefficients?
 - (b) What are the problem of doing this?

2. The “Bike_Day2” sheet contains two new columns *cloudy* and *rainy*, whose values satisfy

$$cloudy = \begin{cases} 1 & \text{if } weathersit = 2 \\ 0 & \text{otherwise} \end{cases}$$

and

$$rainy = \begin{cases} 1 & \text{if } weathersit = 3 \\ 0 & \text{otherwise} \end{cases}.$$

Obviously, if *weathersit* = 1, we have *cloudy* = *rainy* = 0.

- (a) Construct a regression model for *instant*, *cloudy*, *rainy*, and *cnt*. How do you interpret the result?
 - (b) Do we still have the previous problem?
3. Consider the plain text file “Bike_Day2.txt” with columns *weathersit*, *cloudy*, and *rainy*. Read the data into a data frame called D.
 - (a) Run `lm(D$cnt ~ D$instant + D$weathersit)`. What do we get? Is it good or bad?
 - (b) Run `lm(D$cnt ~ D$instant + D$cloudy + D$rainy)`. What do we get? Is it good or bad?
 - (c) Run

```
factorWeather <- factor(D$weathersit)
lm(D$cnt ~ D$instant + factorWeather)
```

What do we get? Is it good or bad?

4. The “Bike_Month” sheet in “SDA-Fa14_cp13_data.xlsx” contains aggregated monthly rentals.
 - (a) Construct a regression model for *instant*, *season*, and *cnt*. Write down the regression formula and then interpret the outcome.
 - (b) Construct a regression model for *instant*, *month*, and *cnt*. Write down the regression formula and then interpret the outcome.
5. The “Car” sheet in “SDA-Fa14_cp13_data.xlsx” contains the numbers of years since production and prices of 60 cars.
 - (a) How do *year* affect *price*?
 - (b) Based on the scatter plot of *year* and *price*, let's try to assume that

$$price_i = \beta_0 + \beta_1 year_i + \beta_2 year_i^2 + \epsilon_i,$$

i.e., we assume that *year* affects *price* in a quadratic way. Construct the regression model (by creating a new column for *year*²) and interpret the outcome.

- (c) For each model, find the *residuals* $y_i - \hat{y}_i$. Plot the residuals. In which model do we have *systematic errors*? As random errors are expected for a good regression model, which model is better?