Statistics and Data Analysis
Homework 6: Regression Analysis (1)

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

1. The "Bike_Day" sheet in "SDA-Fa14_hw06_data.xlsx" contains the daily number public bike rentals in a city and other related information. The "Bike_Month" sheet contains aggregated monthly rentals. The data were collected in the last two years.

   (a) Draw a line chart to depict the trend and fluctuation for monthly rentals. Qualitatively describe the trend and seasonal effect.

   (b) Construct a simple linear regression model for *instant* and *cnt*. What is your regression line?

   (c) Check the R^2 and p-value and make some interpretations.

   (d) Predict monthly rentals for the next year with the model.

2. The previous regression model captures the trend but not the seasonal effect. Let's try to add the seasonal effect into the model.

   (a) Construct a multiple linear regression model for *month*, *instant*, and *cnt*. Try to interpret the model. Is there anything weird?

   (b) For year 1, which months have above average monthly rentals? How about year 2?

   (c) For those months whose monthly rentals are above average in both years, let's label them as high-demand months; for others, low-demand months. Create a new column *high* and enter 1 for high-demand months and 0 for low-demand months. Now construct a regression model for *instant*, *high*, and *cnt*.

   (d) Predict monthly rentals for the next year with the new model. How are the outcomes different from those with the old model?

3. Consider the daily rental data contained in the "Bike_day" sheet.

   (a) Construct a regression model for *instant* and *cnt*. Do you still see an increasing trend?

   (b) Recall that we have a regression line for *instant* and *cnt* for monthly rentals. Is the line for daily rentals flatter or steeper than that for monthly rentals? Why?

   (c) Add the column *holiday* into the regression model in (a). In average what is the impact of being a holiday?

   (d) Remove *holiday* and add the column *workingday* into the regression model in (a). In average what is the impact of being a working day? Compare the result with *holiday*.

4. Consider the daily rental data contained in the "Bike_day" sheet. The columns *temp*, *atemp*, *hum*, and *windspeed* in "SDA-Fa14_hw06_data.xlsx" record values *before* being normalized.

   (a) How do *temp*, *atemp*, *hum*, and *windspeed* affect *cnt*?

   (b) If you used a regression model with the five variables listed in (a), what are the potential drawbacks?

   (c) Try to take away *temp* and do the analysis again.

   (d) Try to add *instant* and do the analysis again.