Week 4: Model building, EDA, Diagnostics

Devesh Tiwari

June 1, 2016

Introduction

Point of this week is to dive in to model building and checking assumptions.

Methods of Moments

I posted an answer to the wall. To summarize, a moment is a mathematical or statistical property of a random variable. The basic idea is that we can estimate the parameters of interest of a random variable by setting its mathematical moments equal to its sample moments, and then solve for the parameters. Method of moment estimators are not perfect, but they are useful and often used when we are dealing with random variables or expressions that are more complicated than OLS.

Review Question [5mins]

Let's return to the example I posted on the wall. We are interested in understanding the relationship between the average daily number of steps taken by a user in the last year and the users' weight. We also have data on users' income and The dependent variable is thus *weight* and the independent variable of interest is *steps* and we are estimating the following equation:

$$weight_i = \beta_0 + \beta_1 steps_1 + \beta_2 income_i + whiteCollar_i + \epsilon_i$$

where whiteCollar is a dummy variable taking the value of one for users with a white collar (office) job and is zero otherwise. I am testing the hypothesis that people who take more steps are healthier (which here means weight). More formally, my null hypothesis is that β_1 equals zero and the alternative hypothesis is that it is not zero.

After running a linear regression, I find that β_1 is positive and that it has a p-value of 0.04 (so it is statistically significant).

(1) What assumptions do we make about the error term?

(2) How would you examine or test those assumptions?

(3) How would these results impact your conclusion that there is a positive relationship between the number of steps taken by a user and their weight? That people who walk more are healithier?

Modeling Overview

Remember that model building requires far more than generating some plots and running a linear model in R. Today, I want to spend some time on exploratory data analysis and post regression diagnostics.

Step 1: Question formation: Why are you creating a model?

1. What questions are you interested in answering?
2. What is the purpose of your model?
    1. Understand the relationship between X and Y
    2. Test specific hypotheses
    3. Understand (or establish) a causal relationship between X and Y
    4. Efficiently predict Y.

To what extent does the purpose of your model impact your modeling choice?

Step 2: Exploratory Data Analysis

1. To what extent are the underlying assumptions of the Classical Linear Model violated?
2. Which variables should be included in the model?
3. Do any of the variables require transformation?
4. What are your initial expectations, based on your EDA?
5. What is the functional form of your model?

Here are some numerical summaries to consider, though this list is not exhaustive. Remember, each plot is to answer a question that you have about the data and you should add/remove stuff from each summary based on what you think is best.

1. Tabulations and cross-tabulations for categorical variables.
2. Summary statistics for numerical variables.
3. Correlation and/or scatter plot matrix for numeric variables.
4. Numerical summaries segmented by a categorical variable.

Step 3: Model execution and evaluation

Step 4: Model diagnostics

1. Are residuals normally distributed?
2. Do they have constant variance?
3. Is the average value of the residuals zero for all values of the predicted value or other covariates?
4. Are there any unusual observations or outliers?

Step 5: Model scrutiny

Group Exercises

The dataset, *wageData.csv*, is a sample dataset containing information about individual's wages and other characteristics. The variable *wage* corresponds to a person's hourly wage, *educ* is a person's years of education, *exper* is years of job experience, *tenure* is the years of job experience with their current employer and *female* takes the value of one if the person is female.

Breakout Session: Steps 1 and 2

Load this dataset and begin working on the first two steps of the data modelling exericse.

Breakout Session: Step 4

Create a linear model to test the relationship between wage and education. Run one model (any model of your choice) and run some diagnostics. Be prepared to justify why you chose the model you did and evalute residuals. What charts did you create or use to diagnose the residuals? What does each chart tell you? What do these charts tell you about your model, should we be worried about any of the underlying CLM assumptions being violated?