GMBA 7098: Statistics and Data Analysis (Fall 2014)

Feedback for Case Study 3

Ling-Chieh Kung

Department of Information Management
National Taiwan University

December 22, 2014

Road map

- **Testing an intuition**.
- Appropriateness of independent variables.
- Significance of independent variables.
- Skewed data and outliers.

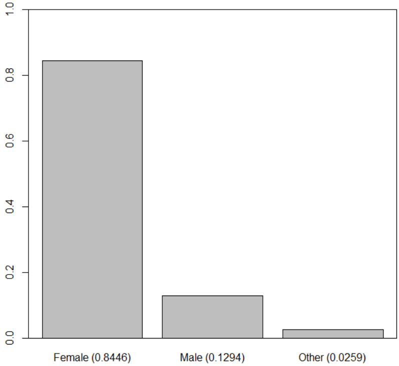
Problem 1 in Case Study 3

- Problem 1 in Case Study 3:
  - According to the USER table, around 82% of the users are female.
  - If all users in average post at the same frequency, the proportion of articles that are posted by female users should also be around 82%.
  - Do a descriptive study to find the sample proportion of articles posted by female users.
  - Then conduct an appropriate **statistical test** on the population proportion of articles posted by female users with respect to 82%.
- Two tasks:
  - Sample proportions (descriptive statistics).
  - Testing the population proportion (inferential statistics).

Sample proportion

- Sample proportion (without removing any row):

Gender	Count	Proportion
Female	25847	0.8446
Male	3960	0.1294
Other	794	0.0259
Total	30601	1



- Obviously, $0.8446 > 0.82$. Is this difference "significant"?
- Our data is just a sample!

Testing population proportion

- What we really want to see is:
  - Is the sample proportion $\hat{p} = 0.8446$ significantly different from the hypothesized population proportion.
- The sample size matters:
  - If $n = 100$, we are not so confident.
  - If $n = 100000$, we are highly confident.
- The statistical test:

$$\begin{aligned} H_0 &: p = 0.82 \\ H_a &: p \neq 0.82. \end{aligned}$$

Testing population proportion

- In R, run

  ```
  prop.test(x = Female, n = Total, p = 0.82,
    alternative = "t", correct = FALSE)
  ```

  we get p-value ≈ 0.
- We reject H_0 at any practical level of significance.
- We are confident to conclude that $p \neq 0.82$ (or $p > 0.82$ if we do a one-tailed test).
- Girls in average post more frequently than boys.

Road map

- Testing an intuition.
- **Appropriateness of independent variables**.
- Significance of independent variables.
- Skewed data and outliers.

Appropriateness of independent variables

- Regression is hard!
  - Selecting independent variables is hard.
- A general guideline:
  - An independent variable **should not** be affected by the dependent one.
  - We use independent variables to **predict**, **estimate**, or **explain** the dependent variable.

Problem 5 in Case Study 3

- Problem 5 in Case Study 3:
  - Nancy and Jay also wonder what factors decide a user's posting frequency.
  - For example, does one's age affects her/his posting frequency?
  - How about one's occupation, gender, frequency of viewing and liking articles, etc.?
  - Try to find a regression model to answer this question.
  - Then **interpret** your regression model and explain **how to predict** whether a user will actively post articles given the factors you identify.
- The dependent variable: posting frequency.
  - Number of articles posted per day/week/month since the registration.
  - Proportion of days that at least one article is posted.
  - There may be other definitions.

Potential independent variables

- Demographic information (occupation, gender, age, etc.):
  - Good! They are not affected by one's posting frequency.
- Non-posting Behaviors:
  - Registration date.
  - Whether one registers during a promotion period.
  - Frequencies of viewing/liking others' articles.
  - Frequencies of messaging with others.
- Posting-related behaviors:
  - Time between registration and the last post.
  - Frequencies of being viewed/liked.

Time between registration and the last post

- The posting frequency somewhat affects the timing of the last post.
- The timing of the last post **cannot** be used to **predict** the posting frequency.
- Nevertheless, some things are reasonable:
    - At any moment, predict **how many** articles one will post in the next **period** given the timing of the last post.

Frequencies of being viewed/liked

- The posting frequency somewhat affects the frequencies of being viewed/liked.
- The frequencies of being viewed/liked **cannot** be used to **predict** the posting frequency.
- Nevertheless, some things are reasonable:
  - At any moment, predict **how many** articles one will post in the next **period** given the up-to-now frequencies of being viewed/liked.

Road map

- Testing an intuition.
- Appropriateness of independent variables.
- **Significance of independent variables**.
- Skewed data and outliers.

Significance of independent variables

- When we have a set of candidate independent variables, how to decide whether one is significant?
- We do a multiple regression and look at their p-values.
  - Suppose the significance level is set to 95%.
  - Those variables whose p-values are less than 5% will be considered significant.
- How about this:
  - For each variable, run a simple regression to see if its p-value < 0.05.
  - Repeat this for all variables.

Significance of independent variables

- Testing each variable's p-value with simple regressions is not right!
- A variable may **become** insignificant when another variable is added.
  - E.g., temperature vs. adjusted temperature.
- This does not consider the **interaction** among variables.
  - In regression, the interaction between two variables can be tested by adding the **product** of these two as a new variable.
  - E.g., if we consider

    $$price = \beta_0 + \beta_1 size + \beta_2 bedroom + \beta_3 size \times bedroom,$$

    the last variable captures the interaction between *size* and *bedroom*.
  - For large houses, having more bedrooms is good; for small houses, having more bedrooms can be bad.

Significance of independent variables

- Sometimes you have **too many** variables.
  - In practice (typical for engineering applications), there may be hundreds or even thousands of variables.
  - Especially an issue in the age of big data.
- In one problem about predicting the yield rate of semiconductor manufacturing:
  - Each lot of chips goes through hundreds of stages.
  - In each stage, there are tens or hundreds of steps.
  - In each step, tens or hundreds or censored values (e.g., temperature, humidity, and many parameters that controls the machine) are recorded.
- There are many methods to **reduce the dimension**:
  - E.g., principal component analysis (PCA).
  - Take courses for data mining or multivariate analysis!

Road map

- Testing an intuition.
- Appropriateness of independent variables.
- Significance of independent variables.
- **Outliers**.

Outliers

- Before we do any fitting, we should first identify outliers.
- There is no standard way to define outliers.
- Use common intuitions and experiences.

Merry Christmas