Logistic Regression

Problem:

A researcher is interested in how GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school.

- What are the dependent and independent variables?
- What range of values can they take on?

Source: http://www.ats.ucla.edu/stat/r/dae/logit.htm

Problem:

A researcher is interested in how GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school.

- **What are the dependent and independent variables?**

  Dependent: Admission into graduate school
  Independent: GRE, GPA, Prestige

- **What range of values can they take on?**

  Admission into graduate school: Yes, No
  GRE: Exam scores from 0-800
  GPA: 0.00 – 4.00
  Prestige: School ranking

Use simple logistic regression when you have one nominal variable with two values (male/female, dead/alive, etc.) and one measurement variable.

-> The nominal variable is the dependent variable, and the measurement variable is the independent variable.

-> Separating Simple logistic regression, with only one independent variable, from multiple logistic regression, which has more than one independent variable.

Source: http://www.biostathandbook.com/simplelogistic.html

Subtle distinction between logistic regression and ANOVA or Students t-test

-> Clue: Logistic regression allows you to predict the probability of the nominal variable.

Example:

- **Measured the cholesterol level in the blood of a large number of 55-year-old women, then followed up ten years later to see who had had a heart attack.**

IF (Test the null hypothesis that cholesterol level is not associated with heart attacks)

THEN (You could do a two-sample *t*–test, comparing the cholesterol levels of the women who did have heart attacks vs. those who didn't)

IF (*Predict* the probability that a 55-year-old woman with a particular cholesterol level would have a heart attack in the next ten years, so that doctors could tell their patients *If you reduce your cholesterol by 40 points, you'll reduce your risk of heart attack by X%*)

THEN (You would have to use logistic regression.)

+ *Situations where measurement is set and nominal variable is set to vary.*

What makes logistic regression different from linear regression is that you do not measure the Y variable directly; it is instead the probability of obtaining a particular value of a nominal variable.

$$\ln[Y/(1-Y)]=a+bX$$

Problem:

A researcher is interested in how GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school.

- **What are the dependent and independent variables?**

  Dependent: Admission into graduate school
  Independent: GRE, GPA, Prestige

- **What range of values can they take on?**

  Admission into graduate school: Yes, No
  GRE: Exam scores from 0-800
  GPA: 0.00 – 4.00
  Prestige: School ranking

```r
## read in data
mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

## view the first few rows of the data
head(mydata)

## summary statistics for variables (3M, Quantiles)
summary(mydata)

## find standard deviation of variables
sapply(mydata, sd)

## declare rank as a categorical variable
mydata$rank <- factor(mydata$rank)

## build logistic regression model
mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")

## produce output for logistic regression model
summary(mylogit)
```

```
## hold GRE and GPA at their means
newdata1 <- with(mydata, data.frame(gre = mean(gre), gpa = mean(gpa), rank = factor(1:4)))

## view new data frame
newdata1

## introduce new variable rankP: probability of acceptance into grad school
newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response")

#view probability table
newdata1
```