Regression analysis for categorical variables

Data to decisions

Discrete dependent variables

- In many cases, Y can only assume a finite number of values:
  - $\{0,1\}$
  - $\{Buy, Hold, Sell\}$
  - $\dots$

- You could still regress Y on X (you always can) but interpreting the results gets unwieldy

- It is easier to write a model which, given X, produces probabilities that Y will assume different values

- This matches the way we naturally think the data were generated

Example: Probit model

- Assume that people select Y as follows:
    1. Draw a *latent* random variable $Y^* = a + bX + \varepsilon$ where ε is a draw from a standard normal distribution
    2. Choose $Y = 1$ if $Y^* > 0$, $Y = 0$ otherwise

- Given a sample, a guess for a, b (the model's parameters) and observed X and Y, we can compute the likelihood of observing that particular sample

- The probit model chooses parameters to maximize that likelihood

Back to chapter 1: treatment effect

- Assume we want to measure the relationship between promotion spending ($S = 1 \; or \; S = 0$) and the likelihood that a customer will sign up

- We could also do this for a continuous S, but easier when S is a *dummy variable*

- A possible model:
  1. $Y^* = a + bX + cS + \varepsilon$ where X are observed customer characteristics
  2. Sign-up ($Y = 1$) if $Y^* > 0$, $Y = 0$ otherwise

Treatment effect: matching approaches

- Model selection is even tougher than normal in discrete contexts
- Less *parametric* alternative: for a given target X find in existing data a set of observations with "similar" X
- Some were targeted (the *treatment group*) some were not (the *control group*)
- A natural estimate of the treatment effect given X is:

$$\begin{gathered} Fraction\ of\ (Y=1)\ among\ treated \\ - \\ Fraction\ of\ (Y=1)\ in\ the\ control\ group \end{gathered}$$

- Key, strong assumption: random assignment to treatment

Overcoming selection problems: experiment

- Run promotion strategy on a sample from the target group, measure impact
- This gets costly, obviously

Logistic regressions

- Logistics regressions are similar to Probit except that ε is assumed to follow a *logistic distribution* rather than a normal distribution

Multivariate case

- What if $Y = 0{,}1{,}2$?
- Trivial extension of binary case
- Assume that people select Y as follows:
  1. Draw $Y_1^* = a_1 + b_1 X + \varepsilon_1$
  2. Draw $Y_2^* = a_2 + b_2 X + \varepsilon_2$
  3. Choose $Y = 0$ if both $Y_1^* < 0$ and $Y_2^* < 0$
  4. Otherwise choose the option with highest payoff
- More model choices to make:
  1. How are the ε's potentially correlated?
  2. What's their distribution?

Forecasting with a probit model: example

- Assume you are applying to grad school and want to know what your probability of admittance is given your undergraduate GPA and GRE score

- $Y = \{Admit, Don't\ admit\}, X = \{GPA, GRE\}$

- Run a probit of Y on X

- Use resulting model to plug in you scores and get a forecast and the associated confidence interval

- You could also use it to estimate the value of boosting your GRE scores (see HW)