RA module 23: Logit and probit models practice problems

(The attached PDF file has better formatting.)

Fox Regression analysis Chapter 14 Logit and Probit Models

** Exercise 23.1: Categorical response variables

Classical regression analysis is not appropriate for models with dichotomous response variables for three reasons: normal distribution, constant variance, range of response variable.

A. What is a dichotomous response variable?
B. Why is the distribution of the error terms not normal?
C. Why is the variance of the error terms not constant?
D. What is the range of the response variable?

Part A: A dichotomous response variable takes one of two values, such a True vs False. Many medical studies have dichotomous response variable. A researcher tests the optimal dosage of a new medication. The response variable may be

- patient dies or does not die
- patient recovers or does not recover

Jacob: Does the model predict whether the patient will die or not die (recover or not recover)?

Rachel: The model predicts a death rate or a recovery rate, such as a 30% death rate. The observed values are 0 (lives) or 1 (dies). The error term has only two possible values: 0 – 30% = –30% and 1 – 30% = 70%. This distribution is not normal.

Jacob: Is this a Bernoulli distribution?

Rachel: For a single patient (a single trial), this is a Bernoulli distribution. The study may have 1,000 patients, and the distribution is a binomial distribution.

Part C: Suppose the death rate for a dosage of 10 is 30% and for a dosage of 20 is 50%. The distribution of the error term is a binomial distribution with a mean of 30% or 50%. The variance of the error term is

$(\pi \times (1 - \pi)) / N$, where π is the death rate and N is the number of patients. This variance differs for each value of the explanatory variable.

Part D: The range of the response variable is [0, 1]: the death rate ranges from 0% to 100%. If the response variable had a normal distribution, its range would be ($-\infty$ to $+\infty$).

** Exercise 23.2: Link functions and conditional distributions

An actuary examines the relation of retention rates (renewal rates) to several explanatory variables (the time since the policy was first issued to this insured, the attributes of the insured, such as sex and age, and so forth).

A. What conditional distribution should the actuary use?
B. What four link functions might the actuary use?
C. How would you choose among these four link functions?

Part A: The response variable has two values: renew or not renew; this is a Bernoulil distribution. If the data point has N exposures, the response variable has N+1 possible values: 0 renewals, 1 renewal, …, N renewals: this is a binomial distribution.

Part B: The textbook recommends four link functions: logit, probit, log-log, and complementary log log.

Part C: The logit and probit link functions are symmetric and give about the same conditional distribution of the response variable. Many statisticians prefer the logit link function because it has a simple interpretation: it is the log odds of the probability. For a likelihood of P%, the logit is *ln*(P% / (1 – P%)). But this rationale doesn't mean the logit link function is better or worse than the probit link function.

The log log and complementary log log link functions are skewed: one to the right and one to the left. If the observed values are skewed, one of these link functions may be better.

** Exercise 23.3: Variance

An actuary uses a generalized linear model to relate the retention rate (the probability that the policyholder renews the policy) to the time since the policy was first issued and characteristics of the policyholder.

- The dependent variable has a Bernoulli distribution: the policyholder either renews or does not renew.
- The expected value of the dependent variable is a probability of renewal.

The actuary uses a binomial distribution with a logit link function. Policyholders with longer durations since the policy was first issued have higher retention rates. The data have 1,000 policyholders at each duration since the policy was first issued with 20,000 total policyholders.

- The average observed retention for all durations is 82%.
- The predicted retention for all durations is 80%.
- The average observed retention for duration = 10 years is 90%.
- The predicted retention for duration = 10 years is 92%.

What is the variance of the retention rate for duration = 10 years?

Solution 23.3: The variance of a binomial distribution is (p × (1 – p)) × N

92% × (1 – 92%) × 1,000 = 73.6

The variance of the retention rate (= the binomial distribution / N) is (p × (1 – p)) / N

92% × (1 – 92%) / 1,000 = 0.00007360

See Fox, *Regression analysis*, Chapter 15, Structure of Generalized linear models, page 381