Distributions and Models

Wonjae

1. Key Summary

- Without the normality assumption, under the other assumptions about the classical linear regression model, the Gauss-Markov theorem shows that the OLS estimators are best linear unbiased estimators (BLUE).
- With the additional assumption of **normality** about the probability distribution of μi, the OLS estimators are not only best unbiased estimators but also follow well-known **probability distributions**. The OLS estimators of the intercept and slope are themselves normally distributed and the OLS estimator of the variance of μi is related to the chi-square distribution.
- The theoretical justification for the normality assumption is the **Central Limit Theorem**. If we sample from a normal population, Yhat (the mean of Y) has a normal sampling distribution. But, what can we say about the sampling distribution of Yhat *if the variables **Yi** are not normally distributed*? **Yhat will have a sampling distribution that is approximately normal if the sample size is large** (at least 25). This statement is called the central limit theorem.
- Under the normality assumption, the ML and OLS estimators of the intercept and slope parameters of the regression model are identical. However, the OLS and ML estimators of the variance of μi are different. In large samples however, these two estimators converge.

2. Distributions

A. When we infer about **means or proportions**

- The binomial distribution
- The normal distribution → "z-test" ((when we know the variance of a population))
- The t-distribution → "t-test" ((when we do not know the variance of a population))
  - Usage: T-test is useful for identifying a difference between the means of two groups
    The two groups can come from the same population or different populations.
  - Assumptions: 1) The populations each have a ***normal distribution***.
    2) Each population has the ***same variance***.

B. When we infer about **variances**

- The chi-square distribution → "chi-square test"
  - We use this distribution in making inferences about a single population variance.
  - The variance of a sample is compared with the variance of its population.

- The F distribution → "F-test"
  - It is used for comparing two population variances.
  - The variances of two samples from two populations are compared.

C. ETC.

Classifying Models according to the dependent variable

A. When Y is continuous

→ the linear regression model (They are mostly estimated by OLS)

→ the nonlinear regression model (log-linear model, lin-log model, or log-log model)

B. When Y is categorical and limited

(Most models here are nonlinear, thus they are estimated by ML)

1. Binary variables → binary logit or probit
   - Bernoulli: Two possible events with nonzero probability
   - Binomial: When we only observe the sum of Bernoulli random variables.
     The probability of oberving the number of successes in N trials.
   - Geometric distribution: The probability of observing the first time to get a success or the waiting time to get the first success in infinite sequence of independent trials.
   - Negative binomial distribution: the waiting time for the rth success with the probability of p.
   - Extended beta-binomial (EBB): Probability varies according to a beta distribution

2. Ordinal variabes → the ordered logit or probit
3. Nominal variables → multinomial logit or probit

4. Censored variables → the tobit model

5. Count variables → Poisson or negative binomial regression model

- Poisson distribution: It is for a count with no upper bound. Mean and variance are identical. The probability of oberving the number of success in N intervals.
- NB probability distribution (Gamma distribution): Variance is greater than the mean. (overdispersion)

C. Statistical Packages (Gujarati p.804)

1. SPSS, SAS
2. STATA → except multinomial probit
3. LIMDEP, GAUSS → multinomial probit,
4. Time-series data → S-plus, E-view (graphing), R

Cf. “CLAEIFY” file

Rare event analysis → “relogit”

Conditional logit → “clogit”