Models for Qualitative Variables: LOGIT and PROBIT

by:
LI Jingxuan
YE Hongbo
ARIES Allen Jerry

Professor:
Catherine Bruneau

Empirical Application No. 1
submitted
as a requirement
in
Financial Econometrics II

7 February 2023

1 Introduction

Consider a dichotomous random variable Y that can take a value of either 0 or 1(i.e. representing two states of the world) which can be explained by a vector of X variables. A realization $y \in [0, 1]$ can be interpreted as probability $P(Y = 1)$ of taking one state, and the complement, $P(Y = 0) = 1 - P(Y = 1)$, of taking the other state. In theory, the standard OLS regression cannot be used to predict Y since the expected value from such an OLS can take any values in $\mathbb{R}$, which may not be in $[0, 1]$.

A latent variable Y^* can be introduced such that:

$$Y^* = \beta' X + \varepsilon$$

This variable is not observable and a quantitative continuous variable taking its values in $(-\infty, \infty)$. Taking a certain threshold S, the dichotomous Y can be explained by:

$$Y = 1 \text{ if } Y^* > S \text{ and}$$

$$Y = 0 \text{ if } Y^* \leq S$$

Suppose that $Y = 1$, it follows that $\beta' X + \varepsilon > S$. If $var(\varepsilon) = \sigma^2$, the standardized residual will have this relationship with the explanatory variables and the threshold:

$$\frac{\varepsilon}{\sigma} > \frac{S - \beta' X}{\sigma}$$

The model will be PROBIT if $\varepsilon \hookrightarrow \mathcal{N}(0, \sigma)$ and LOGIT if ε follows the logistic distribution whose cummulative density function is:

$$F(x; \mu, \sigma) = \frac{1}{1 + \exp(\frac{x-\mu}{\sigma})}$$

In probability theory, the *odds* of an event is easily computed as the ratio of the probability of the event happening against the probability of it not happening. The higher the probability $P(Y = 1|X = x)$, the higher will be the odds. For the case of Y:

$$odds(x) = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}$$

It can be deduced that $\ln(odds(x))$, defined as the logit transform of $P(Y = 1|X = x)$ is just equal to $\beta' X + \varepsilon$. The logit transform is the latent variable for the LOGIT model whose value can be compared against the the threshold S.

The performance of binary classification models can be measured in different ways. First is by looking at the the confusion matrix. Depending on the application, the decision maker can adjust the threshold S to minimize the proportions of the **False Positives** or **False Negatives**.

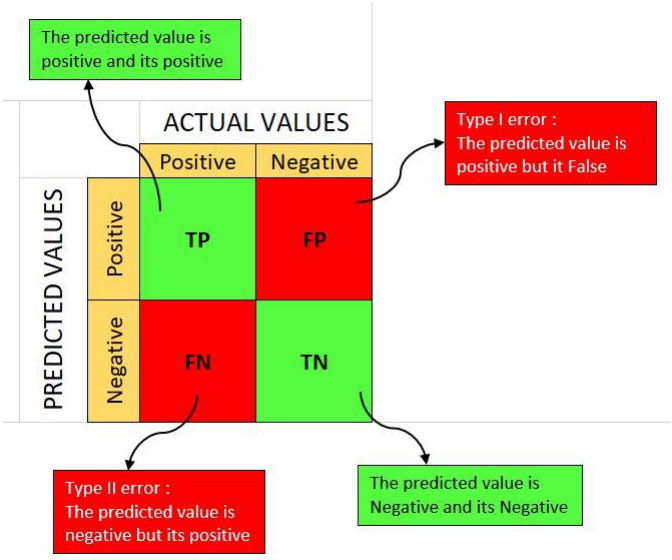

Figure 1: Confusion Matrix of Binary Classification

Another measure is the Receiving Operator Characteristic(ROC) Curve. it is the plot of the True Positive Rate (**Selectivity/ Recall / Probability of Detection**) against the False Positive Rate (**Fall-out/Probability of False Alarm**) as its discrimination threshold is varied. The Area under the ROC curve (AUC) can also be a measure of performance at various thresholds settings. It is the area between the ROC curve and the 45° line, and it represents the degree or measure of separability. The higher the separation, the better the model. Other performance measures include the log-likelihood ratio and the pseudo-R^2.

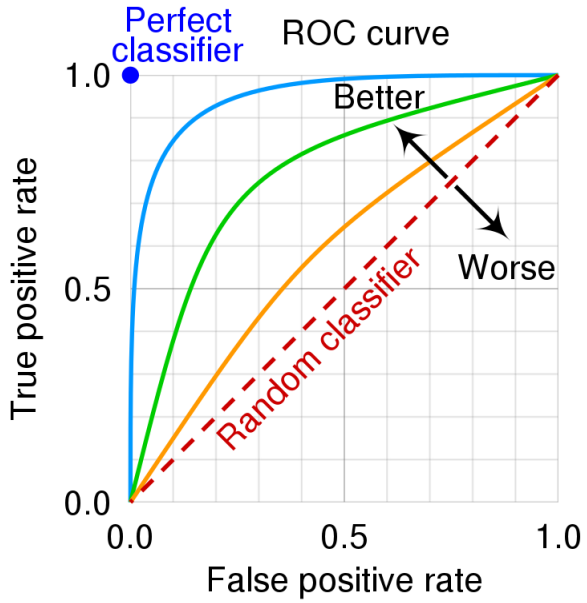

Figure 2: Receiver Operating Characteristic Curve

2 Objectives

3 Database and Exploratory Analysis

4 Model Selection

5 Conclusion

6 Python Codes