Machine learning with Python

Logistic regression

Outline

1. Linear Probability Model
2. Probit Model
3. Logit Model
4. Estimation

Binary response variable

- If a dependent variable can take only two values, say, 1 and 0 we call it a binary, or dichotomous, variable
- One example of such variable is university admission decision(1- admitted, 0- not admitted)
- Another example is vote choice in US (1- Democratic, 0-Republican)
- It is important to note that dependent variable is qualitative

Binary response variable

- Binary response variables have a Bernoulli probability function:

$$f(Y_i|X_i) = P_i^{Y_i}(1-P_i)^{1-Y_i}, \quad Y_i = 0, 1 \tag{1}$$

- where P_i is short notation for $P(Y_i = 1|X_i)$, the conditional probability of observing outcome one, given the regressors
- $E(Y_i|X_i) = 0(1-P_i) + 1(P_i) = P_i$, the conditional expectation is equal to the conditional probability of observing outcome one.
- The standard models used in practice define P_i as a monotonic transformation of a linear index function:

$$P_i = G(X_i'\beta) \tag{2}$$

- $X_i'\beta = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$

Binary response variable

- P_i denotes the probability so we usually require $0 \leqslant G(X_i'\beta) \leqslant 1$
- This is very important when probability is to be modeled
- For $G()$ such transformations are
  1. The cumulative distribution function of the standard normal distribution - the probit model
  2. the cumulative distribution function of the logistic distribution - the logit model
- It is possible to treat $G()$ as the identity function and in this case we get the linear probability model but this model violates the requirement of $0 \leqslant G(X_i'\beta) \leqslant 1$
- We include this model because it is commonly used in practice

Multinomial response variable

- We can also find variables with more than two categories
- The response variable can be three or multiple category
- Returning to example with vote voice suppose that there are three or more parties
- In a model where dependent variable is quantitative our objective is to estimate expected value given the values of the regressors
- In models where Y is qualitative, our objective is to find the probability of something happening are often known as probability models.

Linear Probability Model

- LPM model looks like a typical linear regression model but the dependent variable is binary, or dichotomous

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + u_i \tag{3}$$

- Let's assume that $Y = 1$ represents the family with house and X family income
- $E(Y_i|X_i)$ can be interpreted as the conditional probability that the event will occur given X_i
- Assuming $E(u_i) = 0$ we get:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} \tag{4}$$

Linear Probability Model

- Defining P_i as the probability that $Y_i = 1$ and $1 - P_i$ -probability that $Y_i = 0$ we can show that:

$$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i \tag{5}$$

- Combining with model equation:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} = P_i \tag{6}$$

- The conditional expectation can be interpreted as the conditional probability of Y_i

Linear Probability Model

- The conditional expectation must lie between 0 and 1.

$$0 \leqslant E(Y_i|X_i) \leqslant 1 \tag{7}$$

- The OLS method cannot be easily extended to binary dependent variable regression model due to several problems:
  1. Non-normality of the u_i
  2. Heteroskedastic Variances of the Disturbances
  3. Nonfulfillment of $0 \leqslant E(Y_i|X_i) \leqslant 1$

Non-normality of the u_i in LMP

- u_i like Y_i take only two values that is also follow the Bernoulli distribution
- Nonfulfillment of the normality assumption is not so critical because the OLS point estimates still remain unbiased
- As the sample size increases indefinitely, CLT shows that the OLS estimators tend to be normally distributed.

Heteroscedastic Variances in LMP

- u_i follows Bernoulli distribution where the theoretical mean and variance are, respectively, p and $p(1-p)$, where p is the probability of success

$$var(u_i) = P_i(1 - P_i) \tag{8}$$

- The variance of the error term in the LPM is heteroskedastic
- Knowing that $P_i = E(Y_i|X_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$ the variance of u_i depends on the values of X and is heteroskedastic
- In the presence of heteroscedasticity, the OLS estimators, are not efficient that is, they do not have minimum variance.

Handling the heteroscedastic problem

- One way to resolve the heteroscedasticity problem is to transform the model by dividing it through by $\sqrt{P_i(1-P_i)} = \sqrt{w_i}$

$$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_o}{\sqrt{w_i}} + \frac{\beta_1 X_{i1}}{\sqrt{w_i}} + ... + \frac{\beta_k X_{ik}}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \tag{9}$$

- As you can notice this transformation is weighted least squares

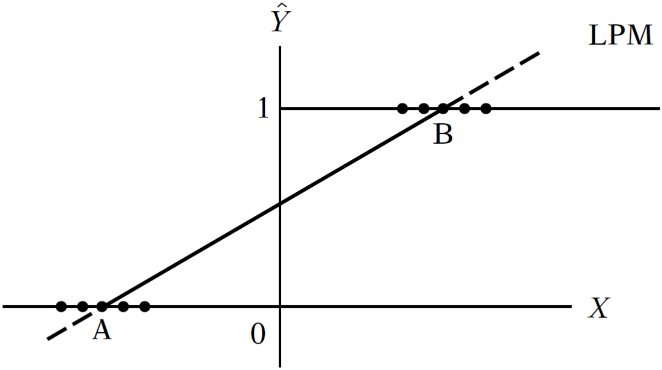
Handling the heteroscedastic problem

- The weights w_i are unknown and to estimate them we can use simple procedure:
    1. Run the OLS to obtain $\hat{Y}_i$ that is an estimate of the true $E(Y_i|X_i)$
    2. Calculate weights as $\hat{w}_i = \hat{Y}_i(1 - \hat{Y}_i)$
    3. Transform the data using $\hat{w}_i$
    4. Estimate the transformed equation by OLS

Nonfulfillment of $0 \leqslant E(Y_i|X_i) \leqslant 1$

- There is no guarantee that $\hat{Y}_i$ will necessarily fulfill this restriction, and this is the real problem with the OLS estimation of the LPM.
- One way to check that is to estimate the LPM by the usual OLS method and find out empirically the fullfilment of that constraint.
- The second procedure is a different estimating technique that will guarantee that estimated probabilities will lie between 0 and 1.

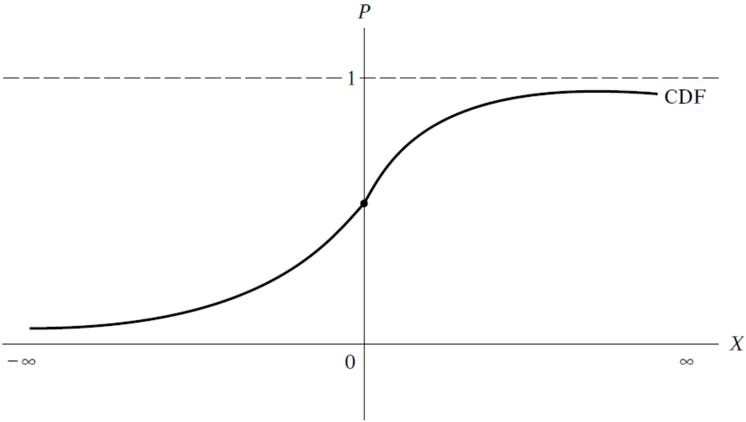
Nonfulfillment of $0 \leqslant E(Y_i|X_i) \geqslant 1$



Alternatives to LPM

- LPM has several problems and the worst of them is not the possibility of $\hat{Y}_i$ lying outside the 0-1 range.
- We can use restricted least-squares to solve 0-1 range problem
- One more concern with this model is that it assumes that $P_i = E(Y_i = 1|X_i)$ increases linearly with X, the marginal effect of X is constant
- In reality we would expect that P_i is nonlinearly related to X_i
- At a very high and low income any increase in income should have little effect on the probability of owning a house. Therefore we need a probability model that has two features:
  1. P_i increases as X_i increases but in interval 0-1
  2. the relationship between P_i and X_i is nonlinear, approaches zero at slower and slower rates as X_i gets small and analogous other way

Geometrical translation



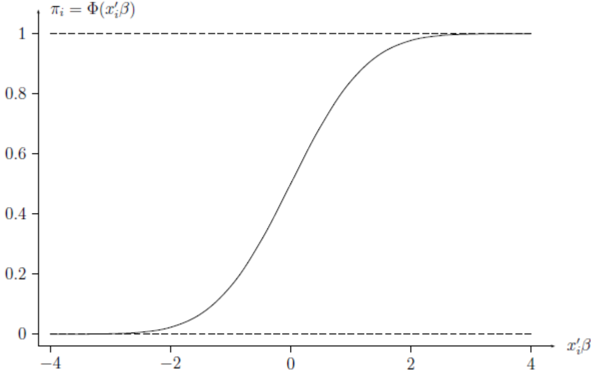
Geometrical translation

- S-shaped, curve is very similar to the cumulative distribution function (CDF) of a random variable.
- For practical reasons, the CDF's commonly chosen to represent the 0-1 response models are:
    1. the logistic - logit model
    2. the normal - probit model
- The differences between the two models are subtle

The Probit model

- Recall the linear probability model, which can be written as $P_i = E(Y = 1|X_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$
- In the probit model we use the cumulative density function of the standard normal distribution
- $P_i = G(X_i'\beta) = \Theta(X_i'\beta) = \int_{-\infty}^{X_i'\beta} \frac{1}{\sqrt{2\pi}} exp(-(t^2/2)dt$
- Θ is the cumulative density function of the standard normal distribution.

Probability function in the Probit model



The Logit model

- Recall the linear probability model, which can be written as $P_i = E(Y = 1|X_i) = \beta_0 + \beta_1 X_i$
- An alternative is to model the probability as a function $G(\beta_0 + \beta_1 X)$
- In the logit model for $G(z)$ we use the logistic function which is the CDF for standard logistic random variable
- $G(z_i) = \frac{exp(z)}{1+exp(z)}$
- Rewriting that for ease of exposition we may show that

$$P_i = \frac{exp(z_i)}{1 + exp(z_i)} = \Lambda(z_i) \tag{10}$$

- where $z_i = \beta_0 + \beta_1 X_i$

The Logit model

- Knowing formula for P_i the probability for $(1 - P_i)$ is:

$$1 - P_i = \frac{1}{1 + exp(z_i)} \tag{11}$$

- Therfore we can write:

$$\frac{P_i}{1 - P_i} = exp(z_i) \tag{12}$$

- $\frac{P_i}{1-P_i}$ is simply the odds ratio

The Logit model

- If we take the natural log of odds ratio we obtain:

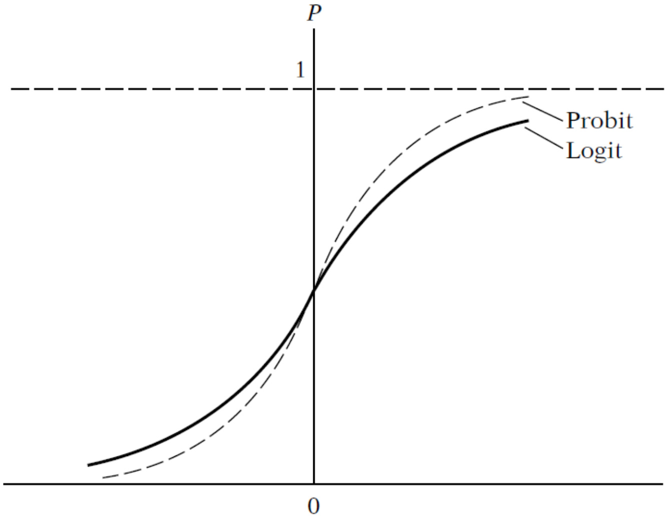
$$L_i = ln(\frac{P_i}{1-P_i}) = Z_i = \beta_0 + \beta_1 X_i \tag{13}$$

- L is called the logit and as we can see is not only linear in X but also in the parameters

Logit vs Probit model

- Logit, and Probit models give qualitatively very similar results
- The main difference between these two models is that the logistic distribution has slightly fatter tails
- The conditional probability P_i approaches zero or one at a slower rate in logit than in probit.
- In practice many researchers choose the logit model because of its comparative mathematical simplicity.
- Models are similar but you have to be careful in interpreting the coefficients estimated by the two models.

Logit vs Probit model



Estimation of binary response models

- The parameters of binary response models can be estimated by the method of maximum likelihood(we assume random sampling)
- The log likelihood function can be written as:

$$logL(\beta, Y, X) = \sum_{i=1}^{n} Y_i log P_i + (1 - Y_i) log(1 - P_i) \quad (14)$$

- Substituting P_i by an appropriate CDF function we can estimate:
  1. $P_i = G(X_i'\beta) = \Theta(X_i'\beta)$ - probit model
  2. $P_i = G(X_i'\beta) = \Lambda(X_i'\beta)$ - logit model
- Parameters of the model are obtained by using iterative optimization algorithms as a closed form solution for $\hat{\beta}$ is not available

Interpretation of parameters

- Estimated coefficients relate X to Z
- Interpreting parameters involves 3 aspects:
  1. Statistical significance
  2. Sign
  3. Magnitude

Interpretation of parameters

- In most applications of binary response models, the primary goal is to explain the effects of the x_k on the response probability $P(Y = 1|X)$.
- In both the logit and probit models all the regressors are involved in computing the changes in probability, whereas in the LPM only the j-th regressor is involved
- The direction of the effect of x_k variable on probability has the same sign as β_k

Interpretation of parameters

- The effect of the change in x_k on the response probability $P(Y = 1|X)$ is not β_k.
- It is more complicated by the nonlinear nature of $G()$ function.
- To find the partial effect of roughly continuous variables on the response probability, we must rely on calculus.
- The effect of X on $P(Y = 1|X)$ varies depending on the values of all explanatory variables.
- In practice we evaluate that effects at the mean values for each X

Statistical significance

- The statistical significance of X_k is determined by whether we can reject $H0 : \beta_k = 0$ at a sufficiently small significance level.
- Insted of t-statistic Stata provides in that case z-statistic
- The z-statistic reported for probit or logit is analogous to OLS's t-statistic
- The procedure of verification is unchanged. We use empirical level of significance to decide.

Likelihood Ratio test - LR

- Usually used as F test in the linear regression models to test significance of the estimated model.
- Also used to test exclusion restrictions
- The LR test is based on the difference in the log-likelihood functions for the unrestricted and restricted models.
- MLE maximizes the log-likelihood function, dropping variables generally leads to a smaller—or at least no larger—log-likelihood.
- Test statistic is $LR = 2(lnL_{ur} - lnL_r) \sim \chi^2_q$
- lnL_{ur} is the log likelihood of unrestricted model and lnL_r is the log likelihood of restricted model. q is the number of exclusion restrictions

Goodness of fit

- Contrary to the linear regression model, there is no single measure for the goodnessof-fit in binary choice models and a variety of measures exists.
- Often, goodness-of-fit measures are implicitly or explicitly based on comparison with a model that contains only a constant as explanatory variable.
- The larger the difference between the two loglikelihood values, the more the extended model adds to the very restrictive model.

Goodness of fit

- A first goodness-of-fit measure Pseudo R^2 is defined as:

$$PseudoR^2 = 1 - \frac{1}{1 + 2(logL_1 - logL_0)/N} \tag{15}$$

- An alternative measure is suggested by McFadden:

$$McFaddenR^2 = 1 - \frac{logL_1}{logL_0} \tag{16}$$

- Both measures take on values in the interval $[0, 1]$

Goodness of fit

- An alternative way to evaluate the goodness-of-fit is comparing correct and incorrect predictions.

		$\hat{y}_i$		
		0	1	Total
y_i	0	n_{00}	n_{01}	N_0
	1	n_{10}	n_{11}	N_1
	Total	n_0	n_1	N