Binary Dependent Variable Models

Rosalba Radice

Analytics Methods for Business

The Models

- We consider models where the dependent variable is binary.
- We will be looking at:
    - Linear probability model;
    - Probit model;
    - Logit model.

Linear Probability Model

- The multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  with a binary dependent variable Y is called the linear probability model.

- In the linear probability model

$$\mathrm{E}(Y|X_1, X_2, \ldots, X_p) = \mathrm{P}(Y = 1|X_1, X_2, \ldots, X_p)$$

  where

$$\mathrm{P}(Y = 1|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

- β_j can be interpreted as the change in the probability that $Y = 1$ if X_j increases by one unit (holding constant the other $p - 1$ regressors) and can be estimated using least squares approach.

US Mortgage Market Example

- We are interested in modelling `deny`, an indicator for whether an applicant's mortgage application has been accepted (deny $= 0$) or denied (deny $= 1$).

- A regressor that ought to have power in explaining whether a mortgage application has been denied is `pirat`, the size of the anticipated total monthly loan payments relative to the the applicant's income.

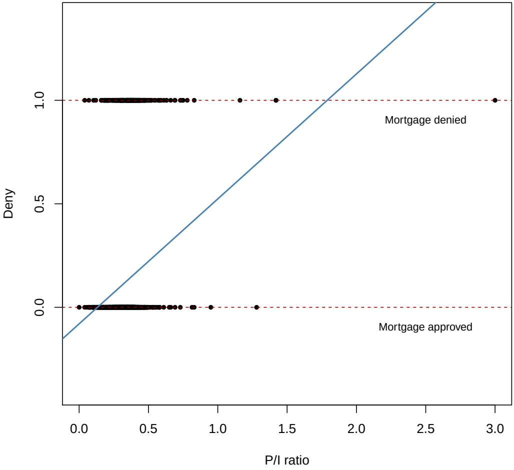
- The linear probability model is

$$\texttt{deny} = \beta_0 + \beta_1 \texttt{pirat} + \epsilon.$$

- The estimated regression line (with standard errors reported underneath the estimated coefficients) is

$$\widehat{\texttt{deny}} = \underset{(0.021)}{-0.080} + \underset{(0.061)}{0.604}\texttt{pirat}.$$

US Mortgage Market Example - continued



Scatterplot of mortgage application denial and the payment-to-income ratio.

US Mortgage Market Example - continued

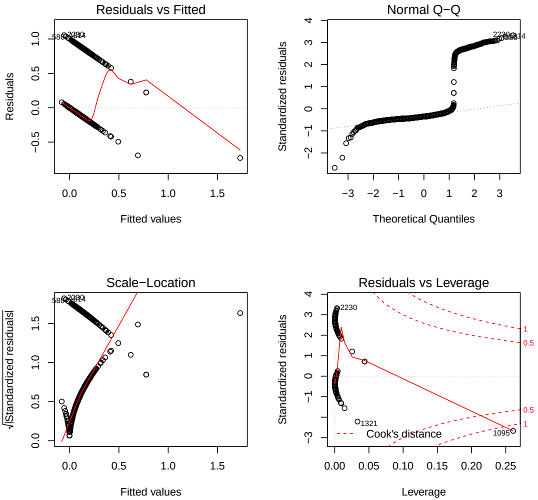
- We augment the simple model by an additional regressor `black` which equals 1 if the applicant is an African American and equals 0 otherwise.

- Such a specification is the baseline for investigating if there is racial discrimination in the mortgage market.

- The new estimated regression function is

$$\widehat{\texttt{deny}} = -\underset{(0.021)}{0.091} + \underset{(0.060)}{0.559}\texttt{pirat} + \underset{(0.018)}{0.177}\texttt{black}.$$

US Mortgage Market Example - continued



Residual plots.

Drawbacks of Linear Probability Model

- By construction the error therms ϵ_i have non-constant variance.
- This model assumes that the conditional probability function is linear. This does not restrict $\mathrm{P}(Y = 1|X_1, \ldots, X_p)$ to lie between 0 and 1.
- This second issue calls for an approach that uses a nonlinear function to model the conditional probability function of a binary dependent variable. Commonly used methods are probit and logit regression.

Probit Model

- The probit model is

$$E(Y|X) = P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X),$$

  where $\Phi(\cdot)$, the cdf of a standard normal distribution, ensures that the estimated probabilities are between 0 and 1.

- Since the dependent variable is a non-linear function of the regressors, β_1 has no simple interpretation.

Interpretation of the Model

- A way to quantify the effect of a continuous variable X on the probability that $Y = 1$ is to use:

$$\frac{\partial \left[\Phi(\beta_0 + \beta_1 X)\right]}{\partial X} = \phi\left(\beta_0 + \beta_1 X\right)\beta_1,$$

where $\phi(\cdot)$ is the pdf of a standard normal distribution.

- The formula is not a constant and varies with the values of the explanatory variable. For this reason researchers often report average marginal effects:

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\beta}_1.$$

- For a binary variable X which takes value 0 and 1 the effect can be quantified using

$$\Phi\left(\beta_0 + \beta_1\right) - \Phi\left(\beta_0\right).$$

Augmented Probit Model

- The model is:

$$P(Y = 1|X_1, X_2, \ldots, X_p) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p).$$

- The effect on the predicted probability of a change in a continuous regressor X_j can be quantified as

$$\frac{\partial \mathrm{E}\left[\Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)\right]}{\partial X_j} = \phi\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_p\right)\beta_j,$$

  which can be obtained by taking the average of the sample marginal effects:

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip})\hat{\beta}_j.$$

Augmented Probit Model - continued

- For a binary variable X_j the effect can be quantified using

$$\Phi\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_j + \ldots + \beta_p X_p\right) - \Phi\left(\beta_0 + \beta_1 X_1 + \ldots + 0 + \ldots + \beta_p X_p\right),$$

which can be calculated by taking the average of the sample marginal effects:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j + \ldots + \hat{\beta}_p x_{ip}\right) - \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + 0 + \ldots + \hat{\beta}_p x_{ip}\right)\right\}.$$

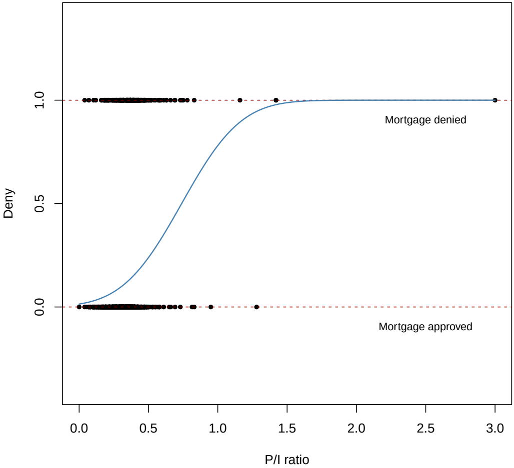
US Mortgage Market Example

- Now, we estimate a simple probit model of the probability of a mortgage denial. The estimated model is

$$\widehat{P(\texttt{deny}|\texttt{pirat})} = \Phi(-\underset{(0.14)}{2.19} + \underset{(0.39)}{2.97}\texttt{pirat}).$$

- Just as in the linear probability model we find that the relation between the probability of denial and the payments-to-income ratio is positive and that the corresponding coefficient is highly significant.

US Mortgage Market Example - continued



Probit model of the probability of `deny`, given `pirat`.

US Mortgage Market Example - continued

- Augmented probit model to estimate the effect of race on the probability of a mortgage application denial:

$$\widehat{\mathrm{P}(\texttt{deny}|\texttt{pirat}}, \texttt{black}) = \Phi(-\underset{(0.14)}{2.26} + \underset{(0.38)}{2.74}\,\texttt{pirat} + \underset{(0.08)}{0.71}\,\texttt{black}).$$

- While all coefficients are highly significant, both the estimated coefficients on the payments-to-income ratio and the indicator for African American descent are positive.

- How big is the estimated difference in denial probabilities between black and non-black applicants?

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Phi\left(-2.26 + 2.74\texttt{pirat}_i + 0.71\right) - \Phi\left(-2.26 + 2.74\texttt{pirat}_i\right)\right].$$

which is equal to 0.17.

US Mortgage Market Example - continued

- The marginal effect of `pirat`, keeping constant `black` can be estimated using

$$\frac{1}{n}\sum_{i=1}^{n}\phi(-2.26+2.74\texttt{pirat}_i+0.71\texttt{black}_i)\times 2.74,$$

which is equal to 0.50.

- Both effects are similar to the estimates obtained with the linear probability model.

Logit Model

- The logit regression function is

$$P(Y = 1 | X_1, X_2, \ldots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}.$$

- The idea is similar to probit regression except that a different cumulative distribution function is used:

$$F(z) = \frac{1}{1 + e^{-z}}$$

is the cumulative distribution function of a standard logistically distributed random variable.

US Mortgage Market Example

- The fitted logistic model is

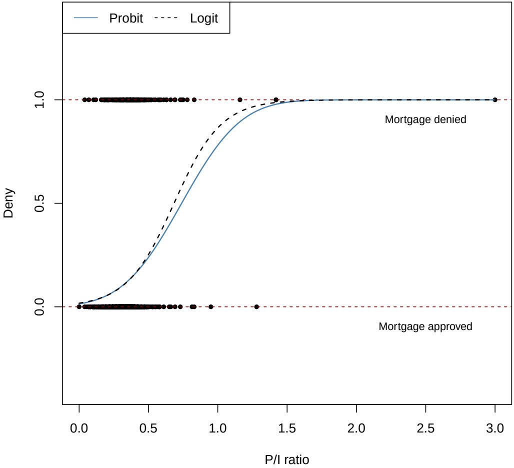
$$\widehat{\mathrm{P}(\texttt{deny} = 1|\texttt{pirat}, \texttt{black})} = F(-\underset{(0.27)}{4.13} + \underset{(0.73)}{5.37}\texttt{pirat} + \underset{(0.15)}{1.27}\texttt{black}).$$

- As for the probit model all model coefficients are highly significant and we obtain positive estimates for the coefficients on `pirat` and `black`.

- For comparison we compute the marginal effects of `pirat` and `black`; they are 0.17 and 0.52, respectively.

- By comparing the estimated effects and looking at the next figure, both models produce very similar results.

US Mortgage Market Example - continued



Logit and probit models of the probability of `deny`, given `pirat`.

Some Remarks

- In both logistic an probit models, t-statistics and confidence intervals based on large sample normal approximations can be computed as usual.
- In the logistic regression $\exp(\beta_j)$ is an odds ratio.
  So for example, $\exp(1.27) = 3.56$ means that the odds of having a mortgage rejected for a non-white applicant is 3.57 times the odds for a white applicant.
- Given the nature of the binary response variable, residual analysis is not meaningful in this case.
- R^2 is not a good measure of goodness of fit. A way of assessing the fit of a binary regression model is to compare the categories of the observed responses with their fitted values (more on this in the lab section).