1 Binary Choice, or dummy variables as dependent variables

However, what we have discussed has only been cases where the explanatory variable are binary.

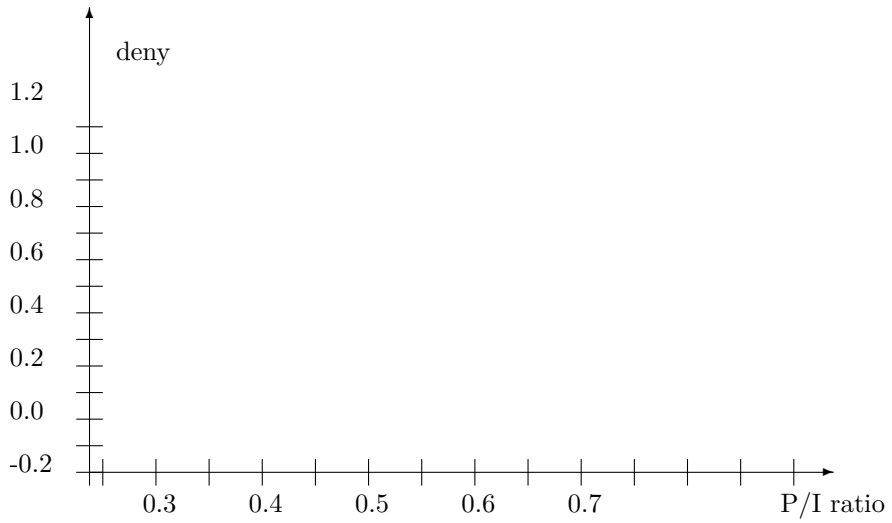
Just as useful are regressions where the *dependent* variable is binary, it can take on one out of two possible states, such as an yes-no answer.

Let us discuss this in terms of a simple example, discussed in Stock and Watson (2019) and taken from Munnell, Tootell, Browne, and McEneaney (1996).

Unlike the typical joke starting with a man coming into a bar, this example starts with a man (or woman) coming into a bank and asking for a loan. The bank can say yes or no. What does the bank base this answer on? Well, one typical question is the amount of income of the applicant. How big is the monthly payment of the loan compared to the monthly income, or the ratio of debt payments to income (P/I ratio).

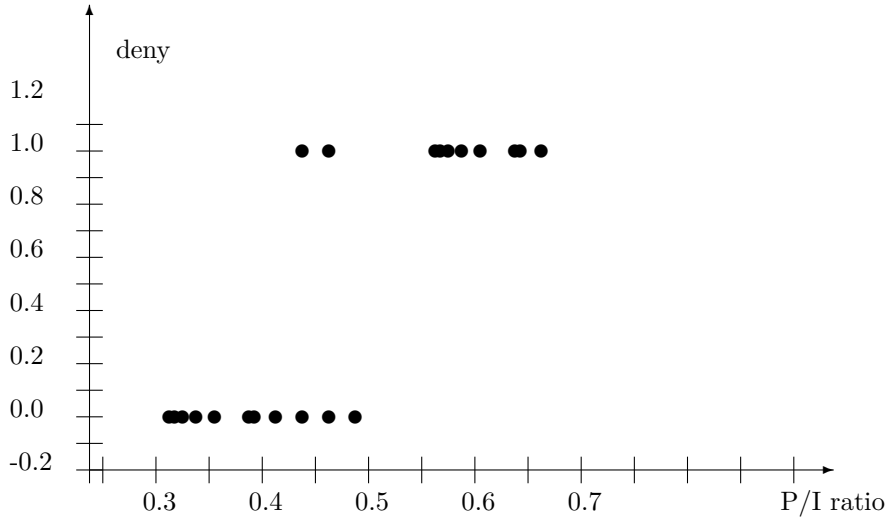
Suppose this is the only number thought relevant for the banks decision. Data: Bunch of decisions (0/1) and P/I ratios.

Illustrate as follows.



Now, what is data in this case?

A bunch of observations of deny=0 or 1, and P/I



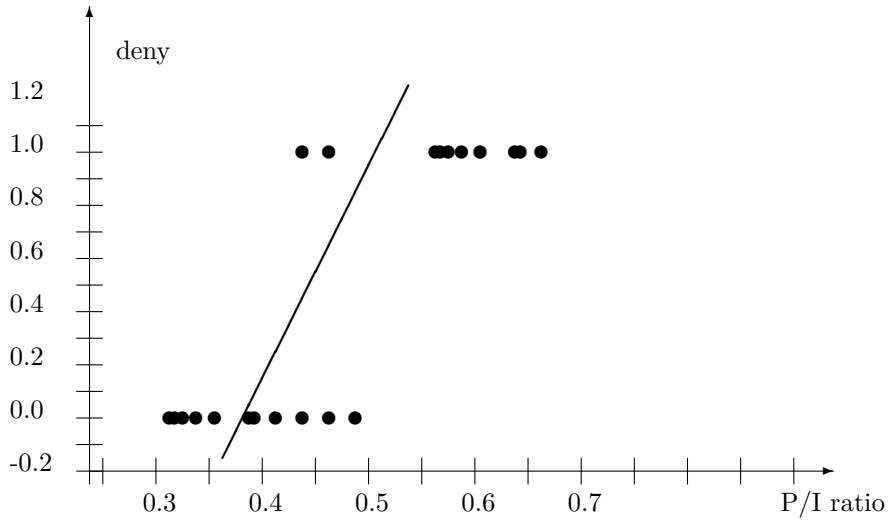
1.1 Regression framework

What is the relevant regression to run

Consider

$$\text{Deny} = a + b\text{P/I ratio} + e$$

Plotting the regression into the picture we just had.



How should the regression be interpreted?

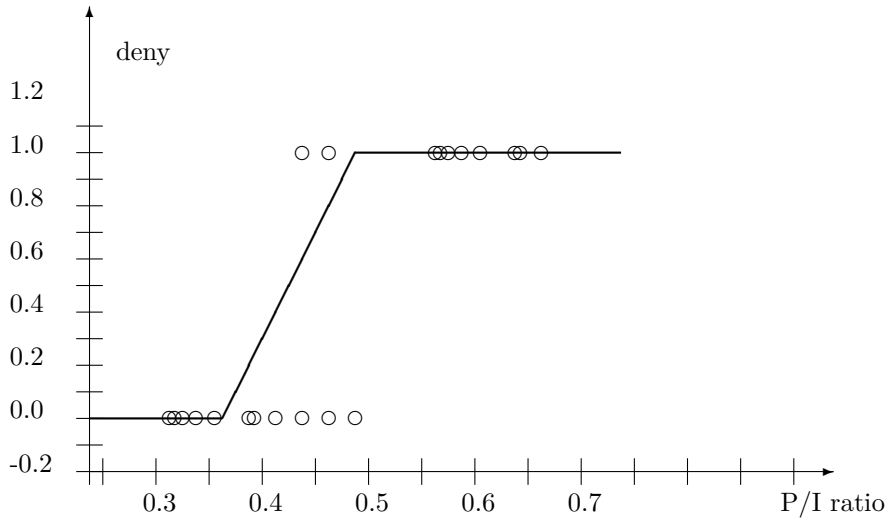
As a probability. We predict the probability that the loan will be denied as a function of the P/I ratio.

The figure shows that the larger the P/I ratio the less likely that the bank is willing to give the loan, and deny.

The figure illustrates an important problem with the formulation we have here.

The regression line will predict probabilities above 1 and below 0, which is impossible.

One can of course truncate the predicted relationship at 0 and 1, giving the following predicted relationship



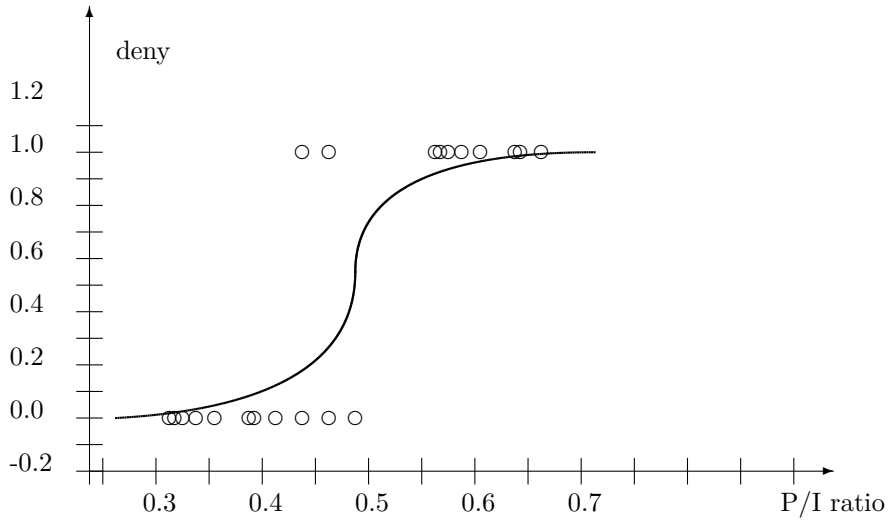
However, this becomes harder when one have more than one explanatory variable, which we typically have. The example we are discussing actually looked at racial discrimination. Controlling for everything else, did the applicants race affect the decision to accept or deny the loan?

Ensuring that the probability is proper in such a setting becomes more difficult.

1.2 Logit and Probit

In practical use of models like the ones we discussed above, binary choice, we will use methods which ensure that the estimated probability is proper. Logit and Probit are the typical methods. These methods are based on maximum likelihood.

The idea is that one want to fit a nonlinear relationship like the illustrated.



The interpretation of such models is similar to the linear regression model, the dependent variable is the probability of deny in this case, and the explanatory variables look at whether the probability increases or falls as we change the explanatory variable.

2 Qualitative and Limited Dependent Variables.

In this section we more formally show the estimation of such models.

We will look at one example of *binary response model*, the Probit model. In the typical regression type models we have looked at so far, eg

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

Note that $\mathbf{y}$ can take on any kind of value. A different type of model is needed in a case where the possible outcomes (y) is one among only a few variables. We call this type of analysis *quantal response analysis*.

Typical examples

- Own or rent a house.
- Drive your own car or take the train.
- Choose between garbage man or economics professor as a profession.

We will look at the simplest possible example, where y can take on only one out of two values,

$$y = \begin{cases} 1 \\ 0 \end{cases}$$

The economic issue is the fact that the outcome of this decision will depend on characteristics of the decision maker. For example, the decision on buying/renting a house can among other depend on

- Income
- Married
- Tenured/Nontenured
- Number of Kids
- Tax status
- etc ....

We want to estimate how the probability of choosing $y = 1$ (say) depends on the exogenous variables X.

Let us assume the decision maker can summarize all his (or her) decision variables into one variable W, which can be viewed as his utility index. He will choose $y = 1$ if W is above some critical level, (say $W \geq 0$).

Thus,

$$y_i = \begin{cases} 1 & \text{if } W_i \geq 0 \\ 0 & \text{if } W_i < 0 \end{cases}$$

is the decision problem of decision maker i. The problem is that we do not observe W_i, we only observe the exogenous variables X_i, which we assume is related to the true W as follows:

$$W_i = X_i\beta + u_t$$

Note that this looks very much like a regression, but you can not perform this regression since we do not observe the variable W_i.

If u_i is normally distributed

$$\begin{aligned} P(W_i > 0) &= P(X_t\beta + u_t > 0) \\ &= P(X_t\beta > u_t) \\ &= 1 - P(-X_t\beta > u_t) \\ &= 1 - P(u_t < -X_t\beta) \\ &= 1 - \Phi(-X_t\beta) \\ &= \Phi(X_t\beta) \end{aligned}$$

To estimate the binary response model, we use maximum likelihood.

Remember that we calculated

$$P(W_i > 0) = \Phi(X_t\beta)$$

If we observe $y_i = 1$, we use $P(W_i > 0)$ as the contribution to the likelihood function, and if we observe $y_i = 0$, we use $P(W_i \leq 0)$ as the contribution to the likelihood function.

We can write this as

$$\Phi(X_t\beta)^{y_t}(1 - \Phi(X_t\beta))^{1-y_t}$$

To find the probit estimates for β, we maximize the likelihood function

$$L(\beta, y, X) = \prod_{i=1}^{n} \Phi(X_t\beta)^{y_t}(1 - \Phi(X_t\beta))^{1-y_t}$$

or the log-likelihood function

$$\ell(\beta, y, X) = \sum_{i=1}^{n} y_t \log(\Phi(X_t\beta)) + (1 - y_t)\log(1 - \Phi(X_t\beta))$$

with respect to the parameters β.

2.1 Readings

(Davidson and MacKinnon, 1993, 15.1, 15.2)

Amemiya (1985)

References

Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

Russel Davidson and James G MacKinnon. *Estimation and Interference in Econometrics*. Oxford University Press, 1993.

Alicia H Munnell, Geoffrey M B Tootell, Lynne E Browne, and James McEneaney. Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review*, pages 25–53, 1996.

James H Stock and Mark W Watson. *Introduction to Econometrics*. Pearson, 4th edition, 2019.