5 Binary Dependent Variable Models

This chapter discusses a special class of regression models that aim to explain a limited dependent variable. In particular, we consider models where the dependent variable is binary. We will see that in such models, the regression function can be interpreted as a conditional probability function of the binary dependent variable. We will be looking at:

- Linear probability model;
- Probit model;
- Logit model.

5.1 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 we have

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

where

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

Thus, 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). Just as in common multiple regression, the β_j can be estimated using least squares approach.

In most linear probability models, R^2 has no meaningful interpretation since the regression line can never fit the data perfectly if the dependent variable is binary and the regressors are continuous. Also the ϵ in a linear probability model are always heteroskedastic. Because linear probability models are linear models with a binary dependent variable, they are easily estimated in `R` using the function `lm()`.

Let's consider the following example where we examine the question whether there is racial discrimination in the US mortgage market. This question is answered by looking at data that relate to mortgage applications filed in Boston in the year 1990. The variable we are interested in modelling is `deny`, an indicator for whether an applicant's mortgage application has been accepted ($\texttt{deny} = 0$) or denied ($\texttt{deny} = 1$). 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. It is straightforward to translate this into the simple regression model

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

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

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

The true coefficient on `pirat` is statistically different from 0. Its estimate can be interpreted as follows: a one unit increase in `pirat` leads to an increase in the probability of a loan denial by 0.6.

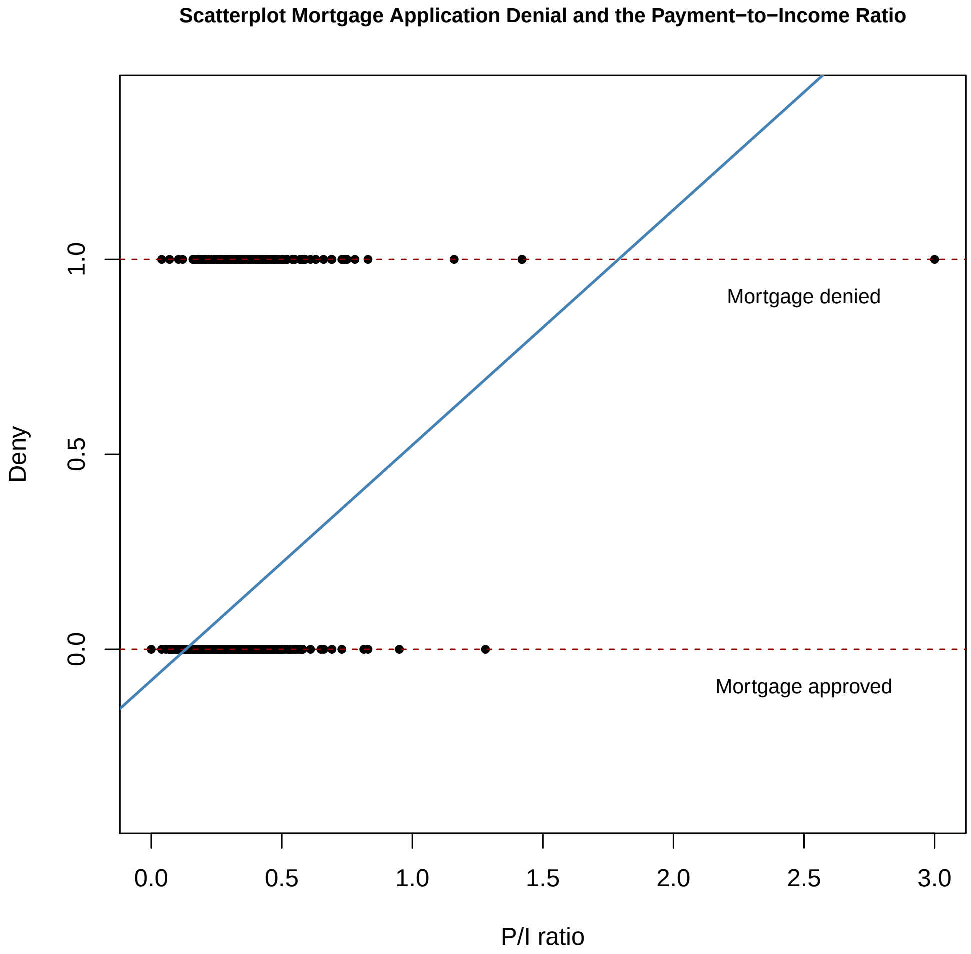

Figure 21: Scatterplot of mortgage application denial and the payment-to-income ratio.

According to the estimated model, Figure 21 shows that a payment-to-income ratio of 1 is associated with an expected probability of mortgage application denial of roughly 50%. The model indicates that there is a positive relation between the payment-to-income ratio and the probability of a denied mortgage application so individuals with a high ratio of loan payments to income are more likely to be rejected.

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: if being black has a significant (positive) influence on the probability of a loan denial when we control for factors that allow for an objective assessment of an applicants credit worthiness, this is an indicator for discrimination.

The new estimated regression function is

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

The coefficient of `black` is positive and significantly different from zero. The interpretation is that, holding constant `pirat`, being black increases the probability of a mortgage application denial by about 0.18. This finding is compatible with racial discrimination. However, it might be distorted by omitted variable bias so discrimination could be a premature conclusion.

Figure 22 shows plots the residuals. The "residuals versus fitted" plot shows two well-defined lines. Normally, such clear structure in a residual plot would indicate a problem with the model. However, in this case the observations are either 0 or 1. Any plot of residuals against linear predictors will therefore show points on one of two curves. Thus the apparent structure in this plot tells us essentially nothing about the (lack of) fit of the model. Similar conclusions can be applied to the remaining plots.

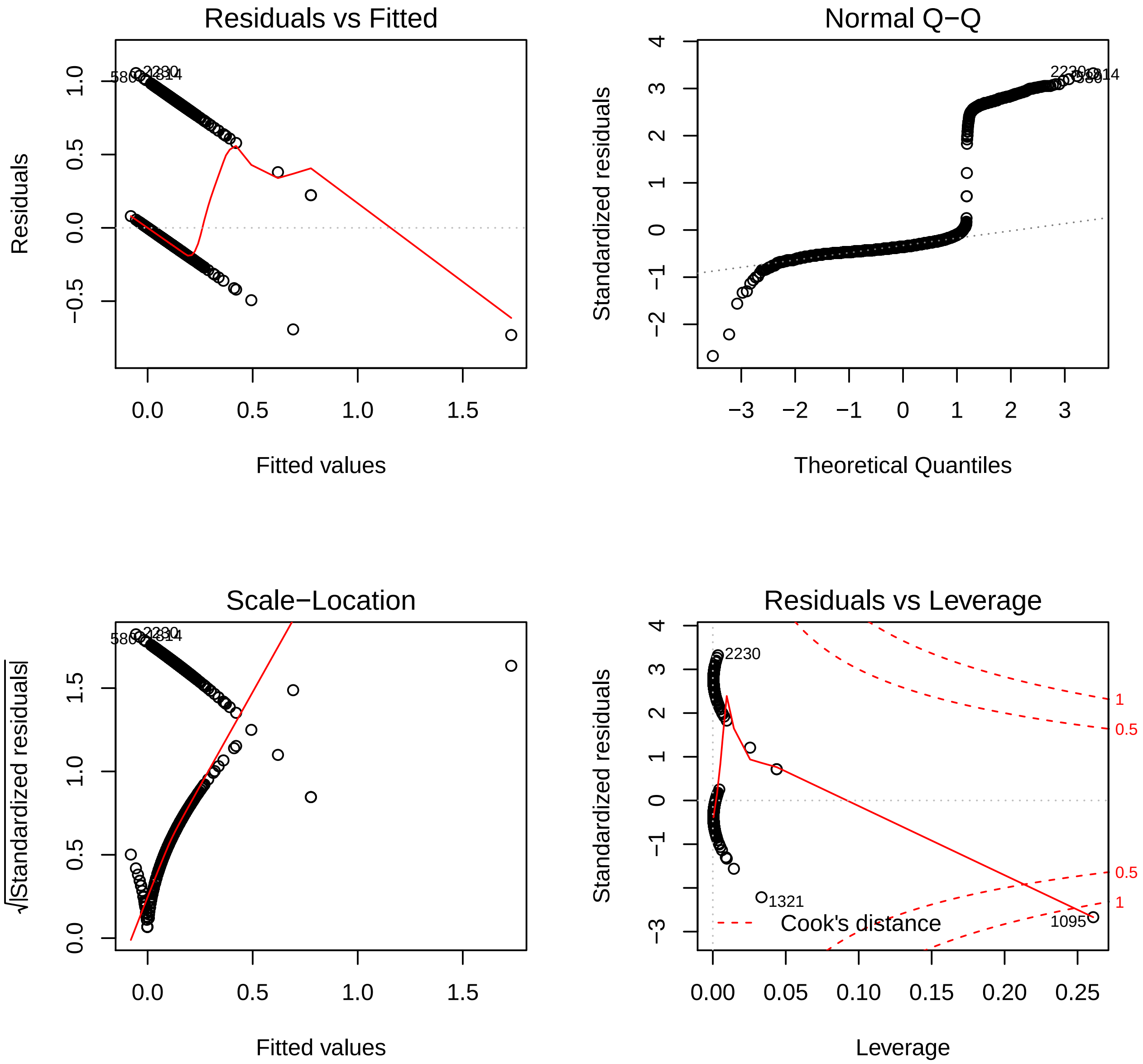

Figure 22: Residual plots.

5.1.1 Lab: Linear Probability Model

Let's consider the female labor force participation for a sample of 872 women from Switzerland. These data can be find in library `AER`. The dependent variable is `participation`, which we regress on all further variables `age`, `income`, `education`, numbers of younger and older children (`youngkids` and `oldkids`), and on the factor `foreign`, which indicates citizenship. For a detialed description of the variables edit `?SwissLabor`. Now we fit a linear probability model. We estimate this model just as any other linear regression model using `lm()`. Before we do so, the variable `participation` must be converted to a numeric variable using `as.numeric()` as `lm()` does not accepts the dependent variable to be of class factor. Note that will turn `participation=no` into `participation` = `0` and `participation` = `yes` into `participation` = 2, so using `as.numeric(SwissLabor$partecipation)-1` we obtain the values 0 and 1.

```
library(AER)
data("SwissLabor")
SwissLabor$participation <- as.numeric(SwissLabor$participation) - 1
swiss_lpm <- lm(participation ~ ., data = SwissLabor)
summary(swiss_lpm)
```

For the current model, all variables except `education` and `oldkids` are highly significant. As for the interpretation, we can see that if `age` is increases by one unit, keeping constant the remaining regressors, the probability of labour participation decreases by 0.11. Also, the probability of participating in the labor force for a non-Swiss is 0.28 higher than that of a Swiss.

5.2 Probit and Logit Models

The linear probability model has a major flaw: it assumes the conditional probability function to be linear. This does not restrict $\mathrm{P}(Y = 1|X_1, \ldots, X_p)$ to lie between 0 and 1. We can easily see this in our reproduction of Figure 21: for `pirat`≥ 1.75, the linear probability model predicts the probability of a mortgage application denial to be bigger than 1. For applications with `pirat` close to 0, the predicted probability of denial is even negative so that the model has no meaningful interpretation here. This circumstance 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.

5.2.1 Probit Models

In probit regression, the cumulative standard normal distribution function $\Phi(\cdot)$ is used to model the regression function when the dependent variable is binary, that is, we assume

$$\mathrm{E}(Y|X) = \mathrm{P}(Y = 1|X) = \Phi(\beta_0 + \beta_1 X).$$

$(\beta_0 + \beta_1 X)$ plays the role of a quantile z. Recall that

$$\Phi(z) = \mathrm{P}(Z \leq z) \ , \ Z \sim N(0, 1)$$

such that the probit coefficient β_1 is the change in z associated with a one unit change in X. Although the effect on z of a change in X is linear, the link between z and the dependent variable Y is non-linear since Φ is a non-linear function of X. Since the dependent variable is a non-linear function of the regressors, the coefficient on X has no simple interpretation.

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

$$\frac{\partial\left[\Phi(\beta_0 + \beta_1 X)\right]}{\partial X} = \phi\left(\beta_0 + \beta_1 X\right)\beta_1, \tag{28}$$

where $\phi(\cdot)$ is the probability density function of a standard normal distribution. Equation (28) is not a constant effect and varies with the values of the explanatory variable. For this reason researchers often report average marginal effects. This can be obtained by taking the average of the sample marginal effects:

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

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

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

where β_1 and β_1 are replaced with their estimated counterparts.

Of course we can generalize the simple probit regression to a case were we have multiple regressors to mitigate the risk of facing omitted variable bias. In this case a probit model with multiple regressors, $X_1, X_2, \ldots, X_p$ is

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

where β_j is the effect on $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_p$ of a one unit change in regressor X_j, holding constant all other $(p-1)$ regressors. The effect on the predicted probability of a change in a continuous regressor X_j can be quantified as

$$\frac{\partial \mathrm{E}\left[\Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)\right]}{\partial X_j} = \phi\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_p\right)\beta_j,$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip})\hat{\beta}_j.$$

For a binary variable X_j which takes value 0 and 1 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:

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

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

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

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. The estimated regression function is reported in Figure 23

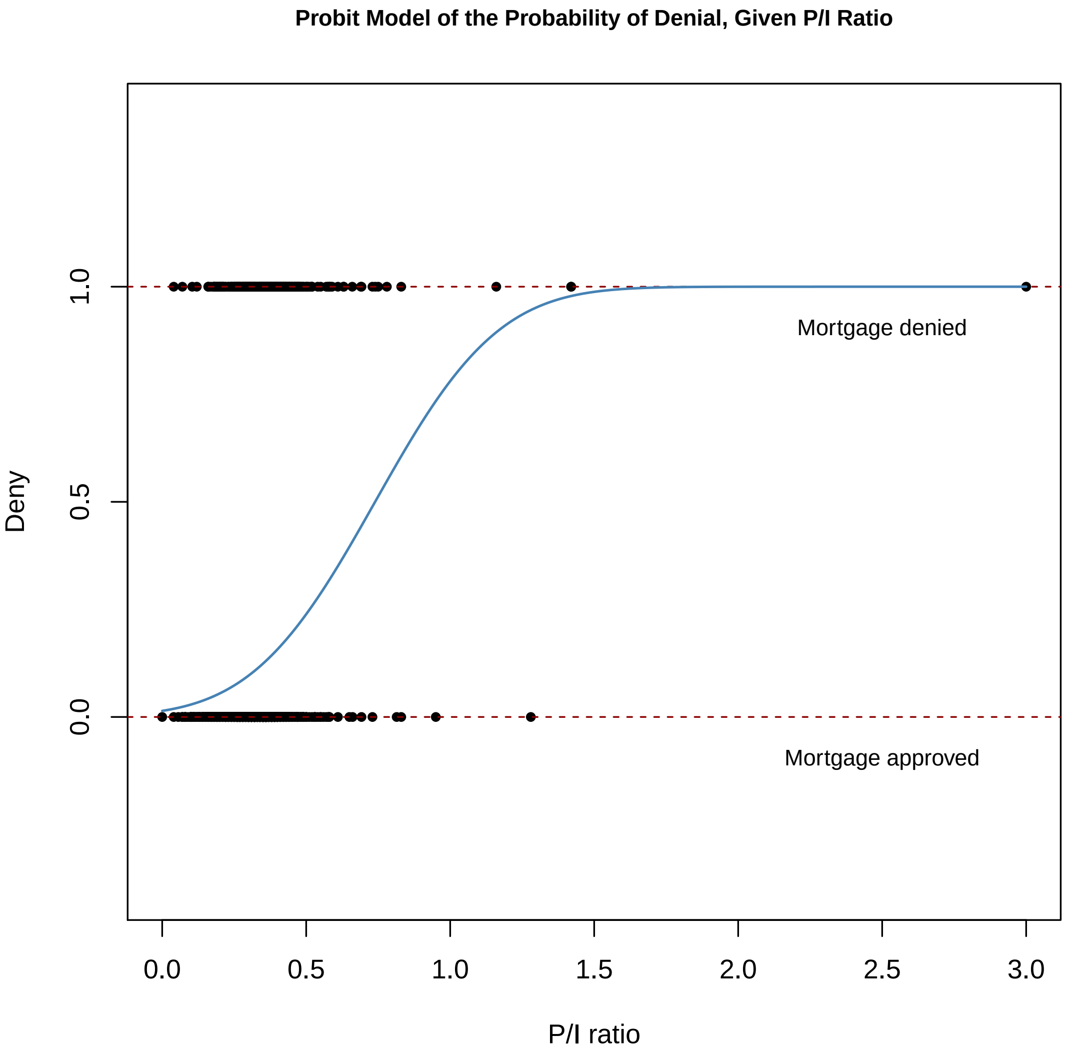

Figure 23: Probit model of the probability of `deny`, given `pirat`.

The estimated regression function has a stretched S-shape which is typical for the cumulative distribution function of a continuous random variable with symmetric probability density function like that of a normal random variable. The function is clearly nonlinear and flattens out for large and small values of `pirat`. The functional form thus also ensures that the predicted conditional probabilities of a denial lie between 0 and 1.

We continue by using an augmented probit model to estimate the effect of race on the probability of a mortgage application denial. The estimated model equation is:

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

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. Again, the coefficients are difficult to interpret but they indicate that, first, African Americans have a higher probability of denial than white applicants, holding constant the payments-to-income ratio and second, applicants with a high payments-to-income ratio face a higher risk of being rejected.

How big is the estimated difference in denial probabilities between black and non-black applicants? We may use the formula of the average of the sample marginal effect

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

which is equal to 0.17.

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

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

which is equal to 0.50. Both effects are comparable to the estimates obtained with the linear probability model.

5.2.2 Logit Models

The logit regression function is

$$\mathrm{P}(Y=1|X_1,X_2,\ldots,X_p)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}}.$$

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.

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 Figure 24, both models produce very similar results.

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 logit model, when comparing probabilities, a useful statistic is the odds ratio. In the logistic regression the odds multiplier $\exp(\beta_j)$ is actually an odds ratio since, for any value of x_j, it can be expressed as

  $$\exp(\beta_j)\;=\;\frac{\text{odds of success for }(x_j+1)}{\text{odds of success for }x_j},$$

  where the odds of success is

  $$\frac{\mathrm{P}(Y=1)}{1-\mathrm{P}(Y=1)}.$$

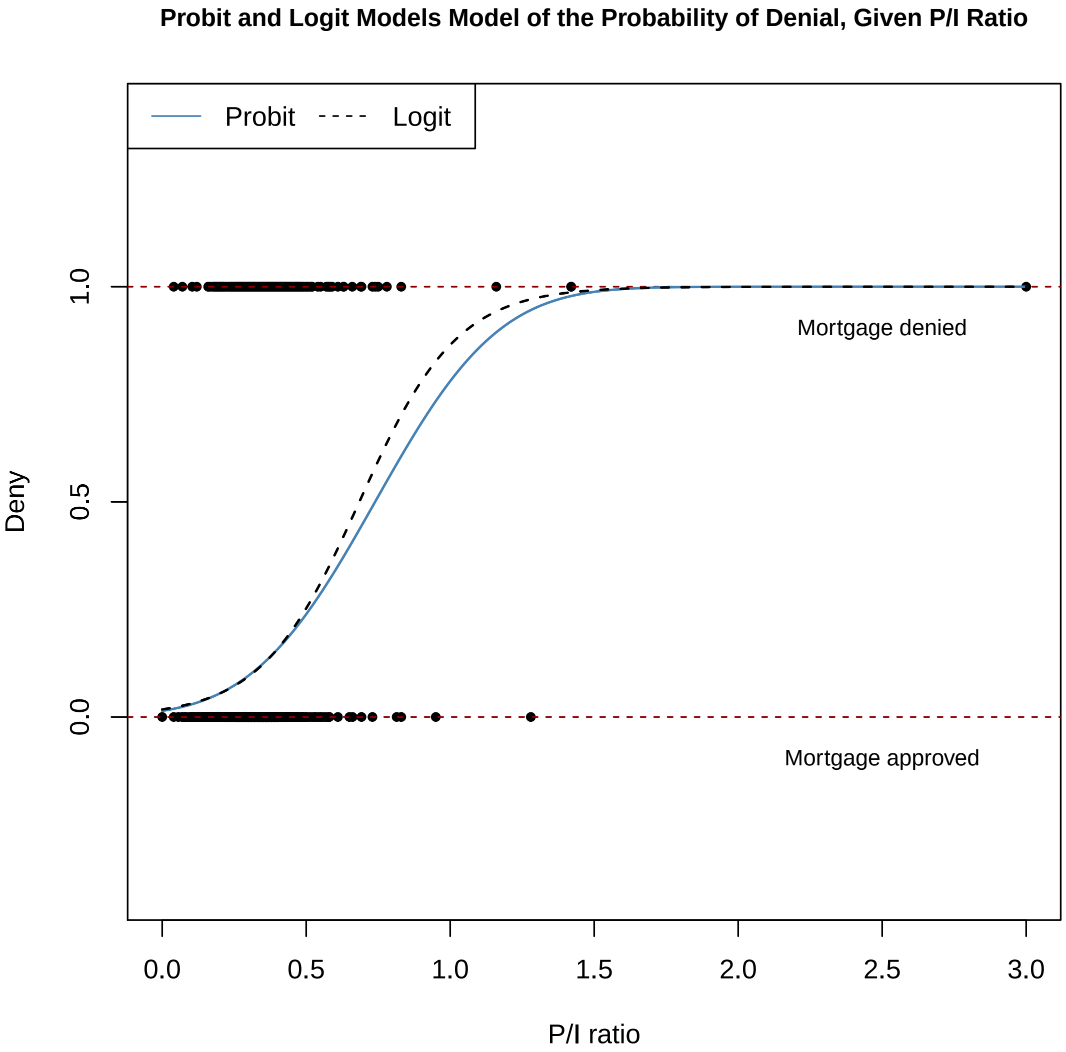

Figure 24: Probit and logit models of the probability of denial, given P/I ratio.

So for example, $\exp(1.27) = 3.56$ means that the odds of having a mortgage rejected for a non-white applicant is 3.56 times the odds (of having a mortgage rejected) for a white applicant,caeteris paribus.

- When using models for binary data, the observations are either 0 or 1: thus, if the model predicts a probability for the ith case in the data set, the residual will be a function of either probability of success or (1 - probability of success). Any plot of residuals against linear predictors (which are a 1-1 function of the fitted values) will therefore show points on one of two curves. The shape of the curves is determined by the linear predictors, not by the responses - thus the apparent structure in this plot tells us essentially nothing about the (lack of) fit of the model.
- Given the binary nature of response variable Y, 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 using a confusion matrix. Alternatively, the receiver operating characteristic (ROC) curve can be used (more on this in the lab section).

5.2.3 Lab: Probit and Logit Models

We again turn to labor economics data SwissLabor and fit the probit and logit models in R. Fitting logit or probit models proceeds using the function glm() with the appropriate family argument (including a specification of the link function). For binary responses (i.e., Bernoulli outcomes), the family is binomial, and the link is specified either as link = "logit" or link = "probit", the former being the default. (A look at ?glm reveals that there are further link functions available, but these are not commonly used in practice.)

We begin with a probit regression:

```
data("SwissLabor")
swiss_probit <- glm(participation ~ ., data = SwissLabor, family = binomial(link = "probit"))
summary(swiss_probit)
```

This shows that the summary of a "glm" object closely resembles the summary of an "lm" object: there are a brief summary of the residuals and a table providing coefficients, standard errors, etc., and this is followed by some further summary measures. Among the differences is the column labeled z value. It provides the familiar t-statistic (the estimate divided by its standard error), but since it does not follow an exact t-distribution here, even under ideal conditions, it is commonly referred to its asymptotic approximation, the normal distribution. Hence, the p-value comes from the standard normal distribution here. For the current model, all variables except education and oldkids are highly significant. The dispersion parameter is taken to be 1 because the binomial distribution is a one-parameter exponential family (see next chapter). The deviance resembles the familiar residual sum of squares. Finally, the number of Fisher scoring iterations is provided, which indicates how quickly the algorithm terminates.

Traditional scatterplots are of limited use in connection with binary dependent variables. When plotting participation versus a continuous regressor such as age using

```
plot(participation ~ age, data = SwissLabor, ylevels = 2:1)
```

R by default provides a so-called spinogram. It first groups the regressor age into intervals, just as in a histogram, and then produces a spine plot for the resulting proportions of participation within the age groups. Note that the horizontal axis is distorted because the width of each age interval is proportional to the corresponding number of observations. By setting ylevels = 2:1, the order of participation levels is reversed, highlighting participation (rather than non-participation). The same plot can be produced for education

```
plot(participation ~ education, data = SwissLabor, ylevels = 2:1)
```

The two plots indicate an approximately quadratic relationship between participation and age and slight non-linearities between participation and education (more on this in the last chapter).

The sample average effects for the continuous covariates can be obtained using simply

```
pav <- mean(dnorm(predict(swiss_probit, type = "link")))
pav * coef(swiss_probit)
```

From here we can say, for example, that the probability of participating in labour force increases by 0.7% for one year increase in education. The effect for the `foreign` variable can be calculated using

```
SwissLaborS <- SwissLaborNS <- SwissLabor

SwissLaborS$foreign <- ’yes’
SwissLaborNS$foreign <- ’no’

predictionsS <- predict(swiss_probit, newdata = SwissLaborS, type = "response")
predictionsNS <- predict(swiss_probit, newdata = SwissLaborNS, type = "response")

mean(predictionsS) - mean(predictionsNS)
```

This means that the probability of labour participation for a non-Swiss is 0.29 higher that that for a Swiss.

Let’s now fit the logit model in `R`

```
swiss_logit <- glm(participation ~ ., data = SwissLabor, family = binomial)
summary(swiss_logit)
```

The point estimates of the logit model are not directly comparable to those of the probit model because of the different (non-linear) link function. If we want to compare the results of the two model, we need to calculate the results in terms of marginal effects:

```
lav <- mean(dlogis(predict(swiss_logit, type = "link")))
lav * coef(swiss_logit)
```

As for the effect of the binary variable `foreign` we have

```
predictionsS <- predict(swiss_logit, newdata = SwissLaborS, type = "response")
predictionsNS <- predict(swiss_logit, newdata = SwissLaborNS, type = "response")

mean(predictionsS) - mean(predictionsNS)
```

The average effects of the logit model are similar to those of the probit model, this comes as no surprise; in fact the distribution of the normal and that of the logistic are rather similar. The results are also very close to those obtained using the linear probability model. Why?

One advantage of using the logit model is that $\exp \beta_j$ can be interpreted in terms of odds ratios. For example if we consider estimated coefficient of `foreign` and we take $\exp(1.31) = 3.71$, we could conclude that the odds of participating in the labor force for a non-Swiss is 3.71 times the odds of participating in the labor force for a Swiss. If we consider a continuous variable, such as `age`, then $\exp(-0.09) = 0.91$, which means that the odds ratio of participating in the labor force for one year increase in `age` is 0.91.

Let’s assess the fit of a binary regression models using both the confusion matrix and the ROC curve. The confusion matrix can be obtained using

```
table(true = SwissLabor$participation, pred = round(fitted(swiss_probit)))
table(true = SwissLabor$participation, pred = round(fitted(swiss_logit)))
```

corresponding to 67% correctly classified and 33% misclassified observations in the observed sample with both probit and logit models. However, this evaluation of the model uses the arbitrarily chosen cutoff 0.5 for the predicted probabilities.

Alternatively, the ROC curve can be used: for every cutoff $c \in [0, 1]$, the associated true positive rate (TPR(c), in our example the number of women participating in the labor force that are also classified as participating compared with the total number of women participating) is plotted against the false positive rate (FPR(c), in our case the number of women not participating in the labor force that are classified as participating compared with the total number of women not participating). Thus, $\text{ROC} = (\text{FPR}(c), \text{TPR}(c)) | c \in [0, 1]$. For a sensible predictive model, the ROC curve should be at least above the diagonal (which corresponds to random guessing). The closer the curve is to the upper left corner ($\text{FPR} = 0$, $\text{TPR} = 1$), the better the model performs. In R, visualizations of this (and many other performance measures) can be created using the ROCR package. In the first step, the observations and predictions are captured in an object created by prediction(). Subsequently, various performances can be computed and plotted:

```
library(ROCR)
pred <- prediction(fitted(swiss_probit), SwissLabor$participation)
par(mfrow=c(1,2))
plot(performance(pred, "acc"))
plot(performance(pred, "tpr", "fpr"))
abline(0, 1, lty = 2)
```

The first plot evaluates the performance for every conceivable cutoff; in this case it indicates that the best accuracy is achieved for a cutoff slightly above 0.4. The second indicates that the fit is reasonable but also that there is room for improvement. Same conclusions are obtained for the logit model.