Models with Binary Dependent Variables

Marcio Santetti | Fall 2023

Table of contents

Introduction

So far, we have applied OLS regression to models whose dependent variables were *quantitative* by definition. However, sometimes our research question involves estimating potential determinants of *qualitative* variables. A few weeks ago, we studied how to use and interpret binary (*dummy*) variables when these were located on the right-hand side of the regression model. Now, we investigate how to properly estimate and interpret models with this kind of variable on the left-hand side, as the explained variable.

In this lecture, we will further investigate three different techniques for the purpose above. First, we will see what happens when we use our traditional OLS estimation for the binary dependent variable case. We will conclude that this is not an appropriate method, since it does not fully capture the intrinsic limitations of the dependent variable. Then, we move on to non-linear techniques, the Binomial *Logit* and *Probit* models, which take care of the things that OLS cannot provide.

The Linear Probability Model

In case there were no better techniques to estimate models with *dummy* dependent variables, what would we do? Based on our previous knowledge, we would run OLS regression and see what happens. When this is the case, we call it the **Linear Probability Model** (LPM). But why do we call it by a different name, and not just the usual linear regression model?

Recall that binary variables take on either 1 or 0 values. When located on the left-hand side of an econometric model, our research question needs to change a bit. Since we are now dealing with *discrete choice* topics, the interpretation is no longer "a one-unit increase in x_i will change y_i by β_i units," but how changes in the independent variables affect the *likelihood* of success (i.e., the *dummy* variable being equal to 1). This is a similar reasoning as the one we are used to, though adapted to this new kind of dependent variable.

A *Linear Probability Model* is simply a linear-in-parameters equation that aims to explain a binary dependent variable:

$$D_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$$

where D_i is a *dummy* variable, and the right-hand side elements are those familiar parameter, variable, and error terms.

Let us now take the Expected Value on both sides:

$$\mathbb{E}(D_i|x_i) = P(D_i = 1|x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$

where $P(D_i = 1|x_i)$ is the probability of success, that is, the probability that the *dummy* variable equals 1 for the i^{th} observation, given the independent regressors. This term is also known as *response probability*.

To measure the change in the response probability from a one-unit change in the j^{th} variable, we use the partial derivatives method we saw a few weeks ago:

$$\Delta P(D_i = 1|x_j) = \beta_j$$

Lastly, the predicted probability of success, i.e., of D_i being equal to 1 is

$$\hat{D} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

Issues with the Linear Probability Model

Using OLS to estimate binary response models is straightforward. However, there are many problems associated with this practice, which will become clear as soon as we look at an example. But before that, a couple of things worth noting:

1. The estimated dependent variable, $\hat{D}_i$, is not bounded by 0 and 1. Even though a dummy variable can only take on 0 or 1 values, as soon as the regression model is estimated by OLS, we can no longer guarantee this fact. Depending on the values of the independent variables (x_i) and the estimated β coefficients ($\hat{\beta}_i$), the value of $\hat{D}_i$ may go below 0 or above 1, which does not make logical sense, since probabilities must lie between 0 and 1.

2. Usual goodness-of-fit measures, such as the R^2 and the adjusted R^2 will not reflect an accurate measure of overall fit. For a similar reason as the one for the item above, the nature of these measures do not allow for a precise measurement of model fit. Therefore, LPMs may return negative values for the adjusted R^2, even if the coefficients may make practical sense.

Dealing with this latter issue is not a big problem. We will learn how to calculate a better goodness-of-fit measure for these special models. And with respect to the first, we will look at more robust regression techniques for binary dependent variable models. Let us first look at a practical example of a Linar Probability Model.

An example

Suppose we want to determine the most relevant factors for labor force participation. An individual can either participate ($part = 1$) or not participate ($part = 0$) in the labor force. Thus, we have already defined our dependent variable, and it is a *dummy* variable. As independent variables, we will assume relevant the individual's *age* and its *squared* term, their *income*, years of *education*, the number of young children (*youngkids*), the number of children over 7 years of age (*oldkids*), and whether the individual is a foreigner or not (*foreign*). Our data set contains 872 observations, collected from a health survey conducted in Switzerland in 1981.

Our econometric model looks like this:

$$\begin{aligned} part_i = \beta_0 + \beta_1 income_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 education_i + \\ + \beta_5 youngkids_i + \beta_6 oldkids_i + \beta_7 foreign_i + u_i \end{aligned}$$

Table 1 illustrates this regression's estimates. This Linear Probability Model is interpreted in the same way as we are used to, only changing the effect on the dependent variable. Now, we are interpreting the likelihood of success, that is, likelihood of the *dummy* variable being equal to 1. In our case, we are investigating the likelihood of an individual participating in the labor force ($part_i = 1$), given our chosen independent variables.

First, let us analyze the coefficients' signs. Coefficients on *income*, age^2, *youngkids*, and *oldkids* are negative, meaning that the higher the income, the older, and the more children (regardless of age) an individual is/has, the *less likely* they are to participate in the labor force. Conversely, for the other factors, their positive signs indicate an *increased likelihood* of joining the labor force.

But what about marginal effects? In other words, how to make sense of the values returned by this regression? First, the table shows us that, except for years of *education*, all covariates are statistically significant at $\alpha = 0.01$. Moreover, the adjusted R^2 indicates that the model is responsible for explaining 18.6% of the predicted likelihood of an individual participating in the labor force. Lastly, on the estimated coefficients. The average age for this data set is close to 40 years old.[1] Thus, using this value we can interpret that, holding all other factors constant, an

[1] This data set has *age* in decades, thus we use 4 to calculate the respective marginal effect.

Table 1: A Linear Probability Model

	Dependent variable:
	part
income	-0.213^{***}
	(0.041)
age	0.683^{***}
	(0.130)
agesq	-0.097^{***}
	(0.016)
education	0.007
	(0.006)
youngkids	-0.241^{***}
	(0.031)
oldkids	-0.049^{***}
	(0.017)
foreign	0.250^{***}
	(0.040)
Constant	1.664^{***}
	(0.446)
Observations	872
R^2	0.193
Adjusted R^2	0.186
Residual Std. Error	0.450 (df = 864)
F Statistic	29.495^{***} (df = 7; 864)
Note:	*p<0.1; **p<0.05; ***p<0.01

additional year of age decreases the likelihood of an individual joining the labor force by 0.093 p.p., supposing this individual is 40 years old. Another interpretation we can easily draw from this model is that a foreign worker ($\texttt{foreign}_i = 0$) is 0.25 p.p. more likely to be part of the labor force, *ceteris paribus*. In summary, we pick the values given by the regression output and interpret these as probabilities of the dependent variable being equal to 1. As an exercise, try to interpret the effect of the remaining variables on the target likelihood.

So far so good, right? Actually, no. For several observations, the predicted probability, i.e., $\hat{D}_i$ will be either greater than 1 or less than 0, which does not make logical sense. Even though this may not be verified for other values, there is no way of having a consistent estimation technique if it delivers probabilities that lie outside the reasonable bounds. Lastly, let us look at Figure 1, showing a scatter plot between $\texttt{part}_i$ and $\texttt{income}_i$, including the respective regression line with slope $\hat{\beta}_{educ} = -0.213$. Notice that the values are bounded between 0 and 1, but OLS estimation delivers a regression line that goes beyond 1 for some observations. This fact does not make sense and therefore the Linear Probability Model is not the best option for estimating models with binary dependent variables. We thus look at more appropriate techniques to deal with this issue.

The Binomial Logit Model

One of the main issues associated with the Linear Probability Model is that some predicted probabilities end up being either greater than or lower than 0 or 1, the only accepted values for the dependent variable in models with *dummy* dependent regressors. Since a linear method such as OLS does not fulfill this basic requirements, we thus turn to *generalized* models, in which the dependent variable is no longer a linear function of the coefficients, independent variables, and error term. The

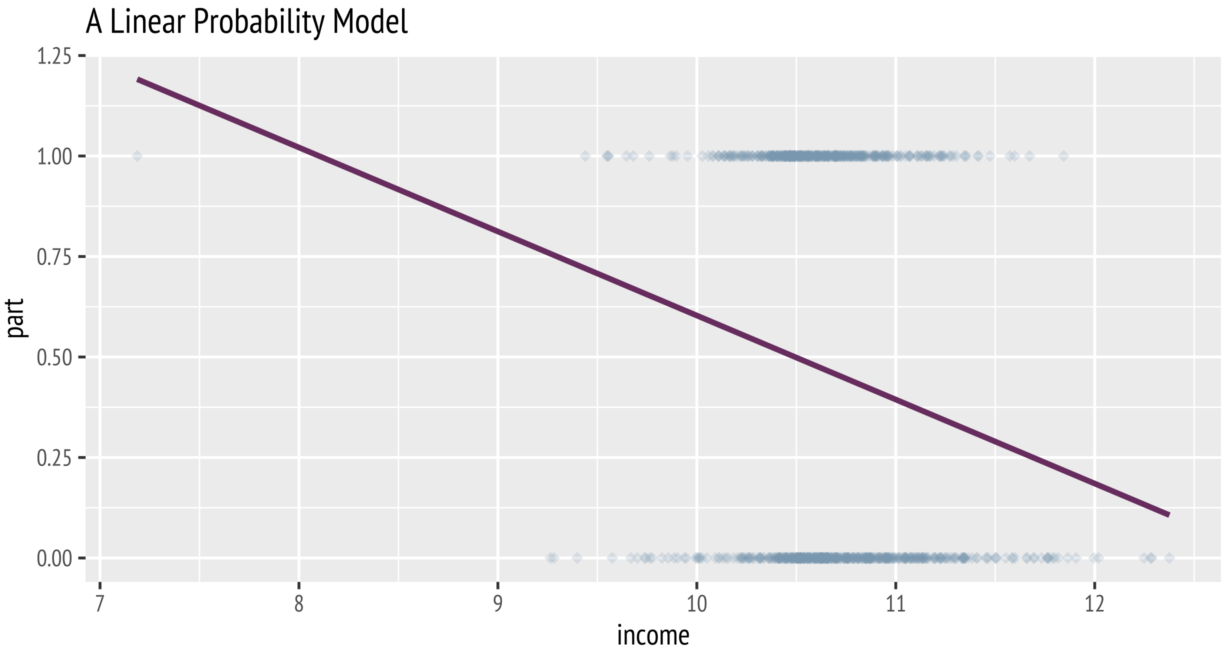

Figure 1: A linear probability model.

first version we will study is the **Binomial Logit Model**, which assumes that the probability of success, i.e. $P(D_i = 1|x_i)$ takes the form of a *logistic* function $\mathcal{L}(\cdot)$. Thus, we write:

$$P(D_i = 1|x_i) = \mathcal{L}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})$$

And $\mathcal{L}(\cdot)$ takes the following form:

$$\mathcal{L} = \frac{e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_k x_{ki})}}{1 + e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_k x_{ki})}}$$

This functional form is the Cumulative Distribution Function (CDF) for a *logistically* distributed random variable, and it assures that the predicted probabilities for the dependent variable will lie between 0 and 1.[2] Instead of estimating the model via OLS, as before, we use another estimator called *Maximum Likelihood*. Then, the *density* of $\mathcal{L}$, given the β coefficients and the set of independent variables, becomes

[2] Believe it or not, but *Wikipedia* has pretty decent pages on probability distributions. You can check out its Logistic Distribution page here.

$$f(D_i|x_i, \beta_i) = \Big[\mathcal{L}(\beta_0{+}\beta_1 x_{1i}{+}\beta_2 x_{2i}{+}...{+}\beta_k x_{ki})\Big]^{D_i}\Big[1{-}\mathcal{L}(\beta_0{+}\beta_1 x_{1i}{+}\beta_2 x_{2i}{+}...{+}\beta_k x_{ki})\Big]^{1-D_i}$$

where $f(\cdot) = \mathcal{L}(\cdot)$ when $D_i = 1$ and $f(\cdot) = 1 - \mathcal{L}(\cdot)$ when $D_i = 0$.

How to interpret Logit coefficients?

As with the LPM, the signs of the estimated coefficients are enough to assess whether a variable increases or decreases the likelihood of the dependent variable being equal to 1. However, in terms of their magnitudes, Logit coefficients are not directly interpretable.

Fortunately, there are several alternatives to interpret marginal effects derived from Logit models. In this course, we will focus on one of the most common approaches: *average* marginal effects (AME). These are basically

$$\frac{\partial P(D_i = 1|x_i)}{\partial x_{ij}} = \frac{\partial \mathcal{L}(\cdot)}{\partial x_{ij}} = \frac{\sum_{i=1}^{n} f(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)}{n} \cdot \hat{\beta}_j$$

where $f(\cdot)$ is the probability density function evaluated at the estimated coefficient values, multiplied by the independent variables. This partial derivative is the mean effect of a one-unit change in the j^{th} independent variable. Its application is easily implemented in a software like R or Stata, as we will see soon.

An example

Now that we know that there is a different estimator for binary dependent variable models, let us compute the same model as before, this time using Logit regression. Table 2 illustrates the results.

Table 2: A Binomial Logit Model

	Dependent variable:
	part
income	-1.104^{***}
	(0.226)
age	3.437^{***}
	(0.688)
agesq	-0.488^{***}
	(0.085)
education	0.033
	(0.030)
youngkids	-1.186^{***}
	(0.172)
oldkids	-0.241^{***}
	(0.084)
foreign	1.168^{***}
	(0.204)
Constant	6.196^{***}
	(2.383)
Observations	872
Log Likelihood	−508.785
Akaike Inf. Crit.	1,033.570
Note:	*p<0.1; **p<0.05; ***p<0.01

First, notice that the signs are the same as the ones delivered by the Linear Probability Model. In addition, evaluating the statistical significance of Logit coefficients is done in the same way as we are used to, using either the test statistic or the p-value. Observe that the table does not report any goodness-of-fit measure. For Logit and Probit models (which will be seen in the next section), these measures are no longer valid. However, we will look at a different way of assessing model fit later in these notes, which will use the *Log-Likelihood* informed by the regression output table. Lastly, the *Akaike Information Criterion* is a measure used for model comparison. Here, we will not study this criterion in detail.

The coefficients' magnitudes cannot be directly interpreted, as stated before. Thus, we can compute average marginal effects. Details on these computations will be provided in the applied lecture, but let us interpret the effect of *income* on the likelihood of an individual participating in the labor force.

$$\frac{\partial P(part_i = 1|x_i)}{\partial income_i} = \frac{\sum_{i=1}^{n} f(\hat{\beta}_0 + \hat{\beta}_1 income_i + ... + \hat{\beta}_k foreign_i)}{872} \cdot (-1.104) = -0.2203$$

The average marginal effect of *income* on *part* is -.2203. This means that if an individual's income increases by 1%, she is 0.2203 p.p. less likely to join the labor force, *ceteris paribus*.[3]

Lastly, Figure 2 illustrates the logit regression *curve* between *part* and *income*. Unlike before, now this estimation technique returns a curve that respects the 0 and 1 bounds, and this slope changes as 0 or 1 are approached. That is why we turn to these generalized techniques to estimate *dummy* dependent variable models, so it is possible to stay within logical limits for probability values.

[3] For this data set, income is log-transformed.

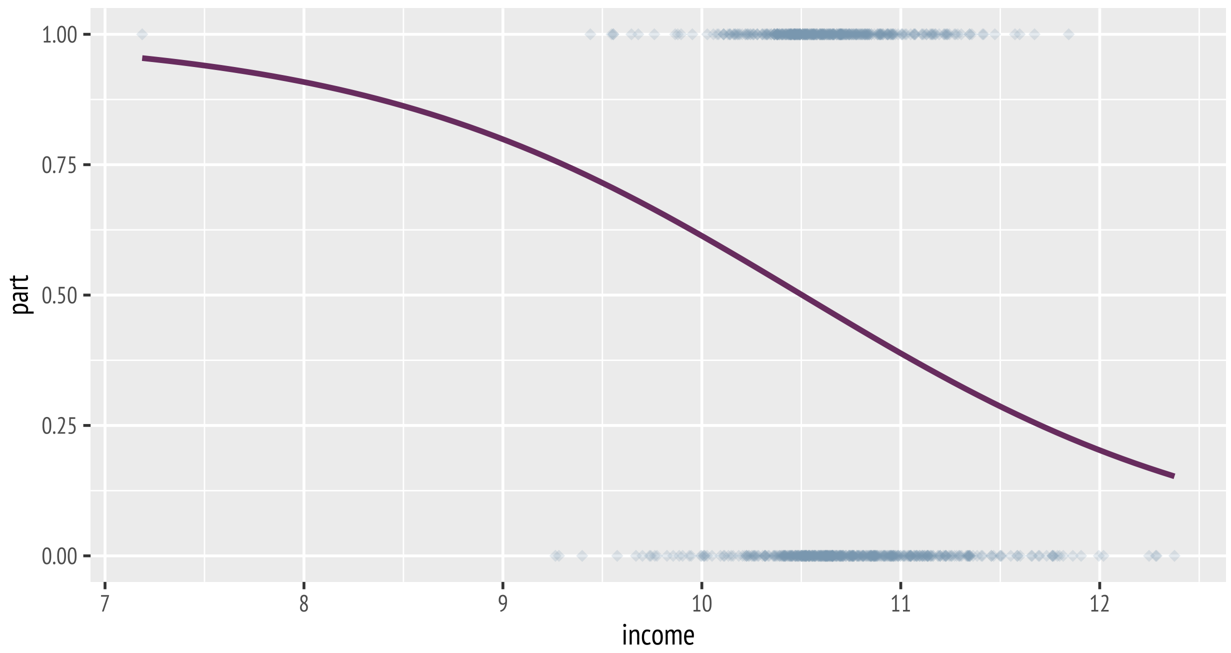

Figure 2: A binomial logit model.

The Binomial Probit Model

Another popular option for binary dependent variable models is the **Binomial Probit Model**. Its intuition is very similar to the Logit model's, with the main difference concerning the assumed cumulative density function (CDF). For Probit models, the latter is *standard Normal*, and not logistic, as assumed for Logit models. Hence, we write

$$\mathbb{E}(D_i|x_i) = P(D_i = 1|x_i) = \Phi(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})$$

where $\Phi(\cdot)$ is a variant of the cumulative Normal distribution. More specifically,

$$P(D_i = 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2}\, ds$$

where $Z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$ and s is a standardized normally distributed variable. Do not worry about these ugly maths. We will easily practice these in our applied lecture.

An example

Let us estimate the labor force participation model, this time using a Binomial Probit procedure. Table 3 summarizes its results.

Table 3: A Binomial Probit Model

	Dependent variable:
	part
income	-0.667^{***}
	(0.132)
age	2.075^{***}
	(0.405)
agesq	-0.294^{***}
	(0.050)
education	0.019
	(0.018)
youngkids	-0.714^{***}
	(0.100)
oldkids	-0.147^{***}
	(0.051)
foreign	0.714^{***}
	(0.121)
Constant	3.749^{***}
	(1.407)
Observations	872
Log Likelihood	−508.577
Akaike Inf. Crit.	1,033.155
Note:	*p<0.1; **p<0.05; ***p<0.01

As we can see, the output is similar to Logit's, and signs and significances are still the same as before. Computing average marginal effects follows the same procedure as shown for Logit models, with the only difference of using a Normal probability density function. Let us evaluate the average marginal effect of *income* on the likelihood of an individual participating in the labor force:

$$\frac{\partial P(part_i = 1|x_i)}{\partial income_i} = \frac{\sum_{i=1}^{n} f(\hat{\beta}_0 + \hat{\beta}_1 income_i + ... + \hat{\beta}_k foreign_i)}{872} \cdot (-1.104) = -0.2209$$

Holding all other factors constant, a 1% increase in an individual's income decreases the likelihood of participating in the labor force by .2209 p.p.. The Probit estimation returns a similar result as the one seen for the Logit estimation, and the next figure illustrates its curve.

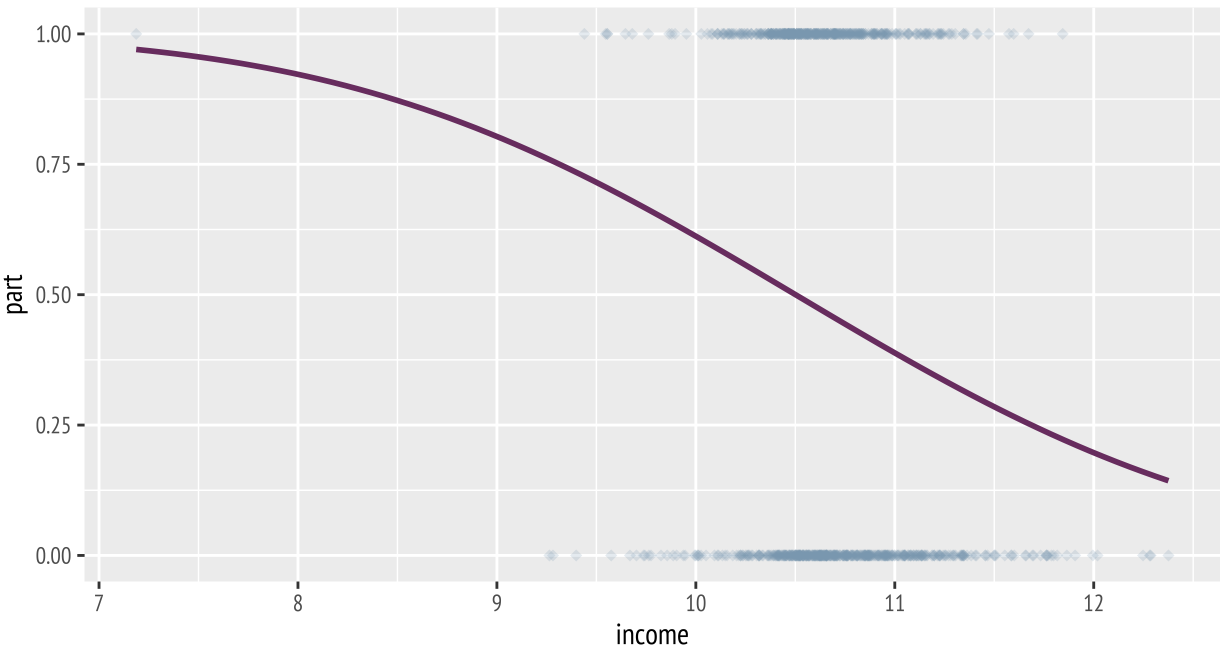

Figure 3: A binomial probit model.

Lastly, the next figure compares the regression lines for the Linear Probability Model (purple), Logit (orange), and Probit (green) estimations. We can see that the two generalized methods are bounded between 0 and 1, and yield very similar curves for this specific example.

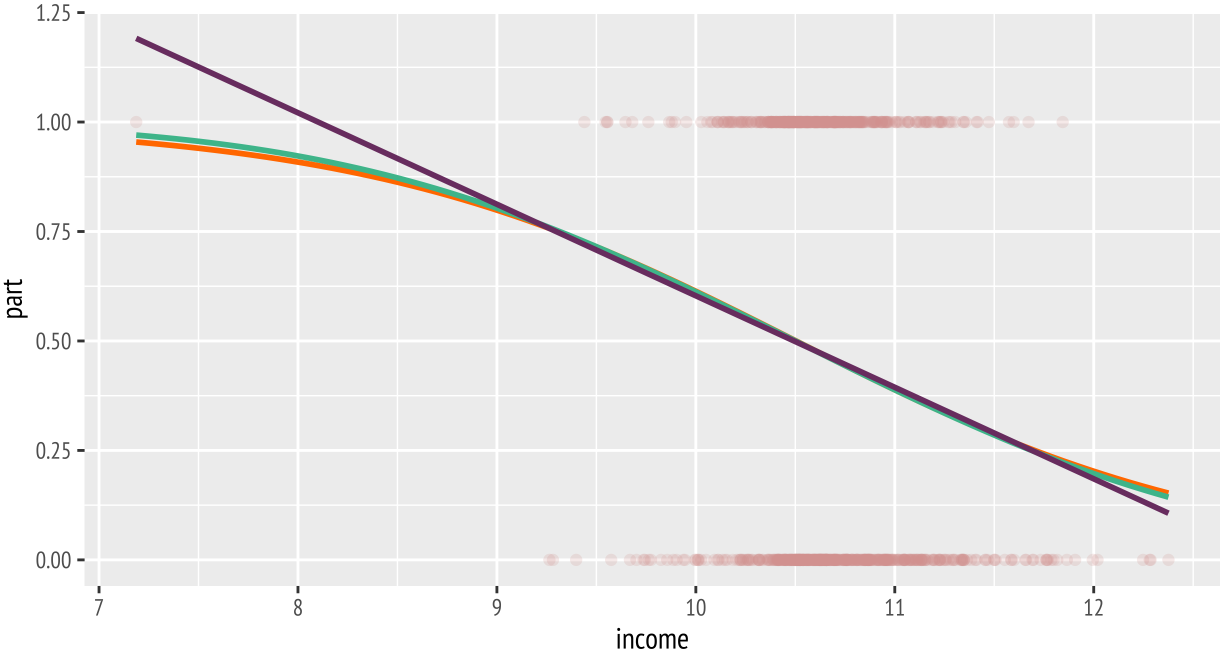

Figure 4: Model comparison.

Measuring goodness-of-fit in Logit and Probit models

A few pages ago, we saw that goodness-of-fit measures, such as R^2 and $\bar{R}^2$, are not effective for generalized linear models. Many times, this is not even a matter of interest when estimating these models, since the central interest lies in the signs and marginal effects returned by Logit and Probit models.

Nevertheless, in case we want to explain how much the estimated model explains the likelihood of success for our chosen dependent variable, there are several techniques to achieve this goal. Here, we will look more closely at the *McFadden's pseudo*-R^2. This measure is defined as

$$R^2 = 1 - \frac{\ell(\hat{\beta})}{\ell(\bar{y})}$$

where $\ell(\hat{\beta})$ is the log-likelihood of the fitted model, and $\ell(\bar{y})$ is the log-likelihood of a restricted model, only containing an intercept term. In these cases, the model will be only the dependent variable's mean, which explains the $\bar{y}$ term. The log-likelihood is simply the logarithm of a likelihood function, which is nothing but a goodness-of-fit measure. In more detail, it is the combination of parameter values that maximize the probability of observing a specific sample. And since we are using Maximum Likelihood instead of OLS, this maximization is obtained via this method.

From the Logit and Probit regression output tables, the informed Log-Likelihood values is $\ell(\hat{\beta})$, while $\ell(\bar{y})$ can be obtained by estimating these models without any dependent variable. For the Logit model, the pseudo-R^2 is 0.1542, while for the Probit model, its value is 0.1546. The pseudo-R^2 is better suited for model comparison purposes rather than direct interpretation. Therefore, if we are trying out different Logit and Probit models with varied variables and specification, this measure may be helpful in deciding which model to stick with. We will learn how to compute these values in our applied lecture as well.