Lecture 14 Classification

Motivation

Up until now we typically worked with the outcome and predictor variables that can be called continuous, e.g., GDP, education, wages, test scores. In Lecture 8 Categorical predictors, polynomials, interactions we have introduced *binary* predictors, as well as more general *categorical* predictors. These variables take on two or more possible values that are qualitative in nature. We call their values *categories* or *classes*. For example, a person's gender can take values such as "male" or "female," or a person's race can take values such as "black," "hispanic," and "white." We learned that we can work with these qualitative categorical predictors by encoding them as binary indicator variables that take values of 0 and 1.

But what if our *dependent* variable is binary or categorical? In this case we are trying to explain why a given observation belongs into a given category. This type of problems is called the *classification problem* and the methods for dealing with it are, in general, different from the methods for dealing with the *regression problem* that we have been studying so far in case of a continuous outcome. Here we will only consider the classification problem for binary outcomes.

Just like with binary predictors, our first step would be to encode our outcome as a binary indicator variable. Let's say our outcome $\tilde{Y}$ can take two possible values class A and class B. For example, $\tilde{Y}$ could be whether a person has a college education, and the two classes are "has a college degree" and "does not have a college degree". We can convert our outcome into a numeric variable Y as follows:

$$Y = \begin{cases} 1, \tilde{Y} = \text{class A}, \\ 0, \tilde{Y} = \text{class B}. \end{cases}$$

Whether we assign a value of 1 to class A or B is completely arbitrary. However, the interpretation of the results will depend on that, so make sure to remember your encoding.

Linear probability model

Our first approach to the classification problem might be to use the linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$$

But let's consider the conditional expectation of the outcome Y given predictors $\mathbf{x} \equiv (X_0, X_1, \ldots, X_k)$:

$$\mathbb{E}[Y \mid \mathbf{x}] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

Recall the expectation of a binary random variable:

$$\mathbb{E}[Y \mid \mathbf{x}] = 1 \times \mathbb{P}(Y = 1 \mid \mathbf{x}) + 0 \times \mathbb{P}(Y = 0 \mid \mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x}).$$

In other words, the conditional expectation of our binary Y is the probability that it equals 1.

If we use the linear regression model on our binary outcome, we are effectively modeling the *probability* of $Y = 1$ conditional on predictors. Moreover, since our regression is linear, the probability will be modeled as a linear function of the predictors. The resulting model is called the *linear probability model* (LPM). Even though this is not the most appropriate tool for the classification problem, the LPM is often used in economics research.

We can estimate the coefficients of the LPM using OLS, as usual. The interpretation of each individual coefficient β_j is now the change in the probability of Y being 1 when X_j increases by 1 unit while keeping other variables fixed. To interpret the value of β_j correctly, however, you need to remember which class corresponds to $Y = 1$. For example, if $Y = 1$ corresponds to a class "has a college degree", then a positive coefficient β_j would imply that the predictor X_j has a positive effect on the probability of having a college degree.

There are a few issues with the LPM. The biggest one is that it is possible to generate predicted values $\hat{Y}$ that are either less than zero or greater than one. Since we are predicting probabilities that are by definition bounded between zero and one, it does not make sense to have predictions like that.

The second issue is that the estimated coefficients will be constant regardless of the values of the predictors. Suppose that the values of the predictors for an observation i are such that the predicted probability of class A is 0.99. Suppose also that some estimated coefficient $\hat{\beta}_j$ is 0.1. The model would predict that increasing the value of X_j by one unit would increase the predicted probability by 0.1. However, the predicted probability is already at 0.99 and can only increase by 0.01 at most. Thus the estimated coefficients from the linear model may be wrong. In other words, the linear probability model is likely to be misspecified.

These two issues cannot be fixed with a linear model. There is a third issue, which however, can be fixed relatively easily. When the outcome is binary, its variance is, by definition

$$\mathrm{Var}(Y \mid \mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})(1 - \mathbb{P}(Y = 1 \mid \mathbf{x})).$$

Therefore, the variance of the error term cannot be constant, which leads to heteroskedasticity.

The good news is that we at least know its shape and hence can use FGLS to correct for it. The estimated $\hat{h}_i$ for each observation will be

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}).$$

The bad news is that, the first issue with the LMP can mess up this strategy because the predicted $\hat{h}_i$ and thus the weights must be positive for FGLS to work. If our predicted probabilities fall outside of the unit interval, this will generate negative weights. In this case there are two possible fixes. The first one would be to abandon FGLS and simply use

heteroskedasticity robust standard errors. The second one would be to force all of the $\hat{h}_i$ to be between 0 and 1. For example, we can set $\hat{y}_i = 0.01$ if $\hat{y}_i < 0$ and $\hat{y}_i = 0.99$ if $\hat{y}_i > 1$.

Logit model

Set-up

The logit model is a tool specifically designed for classification problems. It is an example of a *generalized linear model* (GLM). As a starting point, recall that in the LPM we are modeling the probability of $Y = 1$ conditional on predictors.

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mathbf{x}'\beta.$$

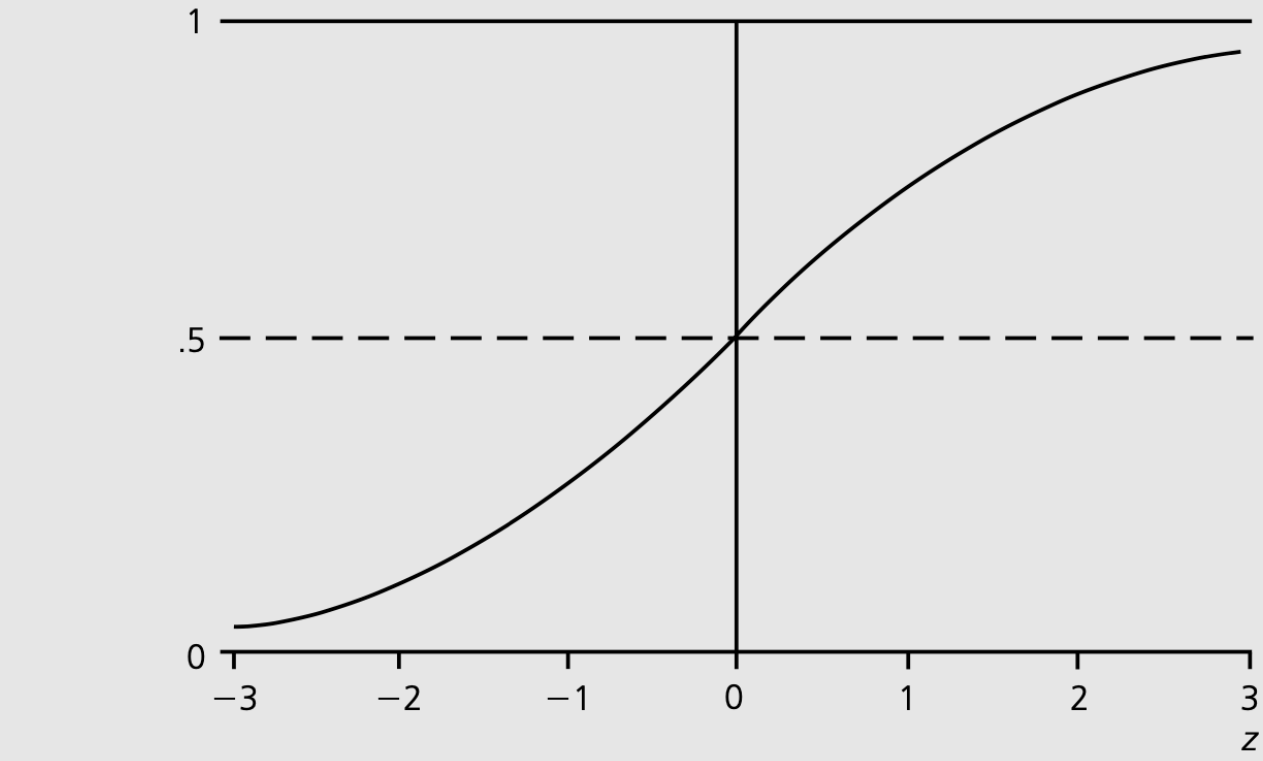
We will use the linear algebra notation $\mathbf{x}'\beta$ to make the formulas more compact. The main issue with the LPM is that the linear function $\mathbf{x}'\beta$ technically allows the predicted values to be outside of the unit interval. The idea of a GLM is to put some function G, called a *link function*, on top of the linear function $\mathbf{x}'\beta$ to ensure that the predicted values have the desired properties.

In a binary classification problem we are predicting probabilities. Hence, the G function should produce values strictly between 0 and 1. It should also be defined for any value between minus and plus infinity and be monotonically increasing. All of these properties are satisfied by a G function that is a CDF. One of the most commonly used G functions is the *logistic CDF* (denoted by the capital letter Λ):

$$G(z) = \Lambda(z) = \frac{e^z}{1 + e^z}.$$

The picture below shows the graph of the logistic CDF.

Graph of the logistic function $G(z) = \exp(z)/[1 + \exp(z)]$.



It looks similar to the standard normal CDF. The logistic CDF is also symmetric, meaning that

$$\Lambda(z) = 1 - \Lambda(-z).$$

Thus our model becomes

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \Lambda(\mathbf{x}'\beta).$$

This model is called the *logistic regression* or the *logit* model.

> **Note**
>
> Another popular link function is the standard normal CDF. Using this link function results in a *probit* model.

Latent variable model

The logit model can be derived using the so-called *latent variable model* that satisfies the classical linear model assumptions. A latent variable model has two parts. The first part specifies a model for the latent variable Y^*. The second part specifies how the observed variable Y is related to the latent Y^*. For the logit model, we assume that

$$\begin{aligned} Y^* &= \mathbf{x}'\beta + U, \\ Y &= \mathbb{I}[Y^* > 0], \\ U &\sim \Lambda(\cdot). \end{aligned}$$

The latent variable U^* can be thought of as the unobserved utility associated with choosing class A over class B. Notice that this means that we model the utility as a linear function of predictors. When the utility of class A is positive, we will choose it over class B, hence the observed variable Y will become 1. For example, if our outcome variable is whether a person has a college degree, the latent variable will be the utility of going to college. If this utility is positive, the person will choose to go to college. If the utility is negative, the person will choose not to go to college.

Now, the probability of $Y = 1$ is

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid \mathbf{x}) &= \mathbb{P}(Y^* > 0 \mid \mathbf{x}) \\
&= \mathbb{P}(\mathbf{x}'\beta + U > 0 \mid \mathbf{x}) \\
&= \mathbb{P}(U > -\mathbf{x}'\beta \mid \mathbf{x}) \\
&= 1 - \mathbb{P}(U \leqslant -\mathbf{x}'\beta \mid \mathbf{x}) \\
&= 1 - \Lambda(-\mathbf{x}'\beta) \\
&= \Lambda(\mathbf{x}'\beta).
\end{aligned}$$

We derived the logit model!

Estimation

Estimating the logit model cannot be done using OLS. Instead, we need to use a different kind of estimator called the *Maximum Likelihood Estimator* (MLE). The idea of the MLE is the following. The *likelihood* of observing an outcome $Y_i = y_i$, conditional on the predictors $\mathbf{x}_i$ and parameters β is simply

$$f(Y_i = y_i \mid \mathbf{x}_i, \beta) = \Lambda(\mathbf{x}_i'\beta)^{y_i}(1 - \Lambda(\mathbf{x}_i'\beta))^{1-y_i}.$$

If we observe a random sample of size n, then the joint likelihood of observing this random sample is the product of individual likelihoods

$$\Pi_{i=1}^{n}\Lambda(\mathbf{x}_i'\beta)^{y_i}(1 - \Lambda(\mathbf{x}_i'\beta))^{1-y_i},$$

where $\Pi_{i=1}^{n}a_i$ denotes the product of all the elements from 1 to n: $a_1 \times a_2 \times \ldots \times a_n$.

We can view this expression as the *likelihood function* of the parameters β given the data $\mathbf{y}, \mathbf{X}$:

$$L(\beta \mid \mathbf{y}, \mathbf{X}) = \Pi_{i=1}^{n}\Lambda(\mathbf{x}_i'\beta)^{y_i}(1 - \Lambda(\mathbf{x}_i'\beta))^{1-y_i}.$$

Taking the logs, we get the *log-likelihood function*

$$\mathcal{L}(\beta \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n}\big[y_i \ln \Lambda(\mathbf{x}'\beta) + (1 - y_i)\ln(1 - \Lambda(\mathbf{x}'\beta))\big].$$

Then the maximum likelihood estimator of β is defined as the value that maximizes the log-likelihood function:

$$\widehat{\beta}^{\text{MLE}} = \arg\max_{\beta} \mathcal{L}(\beta \mid \mathbf{y}, \mathbf{X}).$$

Unfortunately, there is typically no closed-form solution for the ML estimator, unlike for OLS. The value of $\widehat{\beta}^{\text{MLE}}$ is usually found through numeric optimization, although, it is often quite fast. In general, it can be shown that the MLE is consistent, asymptotically normal, and asymptotically efficient.

Interpretation

The interpretation of the coefficients in the logit model is trickier than in the OLS. We first note that the logistic function has the following property

$$\frac{\Lambda(z)}{1-\Lambda(z)} = e^z.$$

We also need to define a statistic called the *odds*, which in our context equals

$$\frac{\mathbb{P}(Y=1 \mid \mathbf{x})}{1-\mathbb{P}(Y=1 \mid \mathbf{x})}.$$

The odds tell one how likely a given outcome is by computing the ratio of the probability of that outcome happening (in our case, $Y=1$) versus the probability of the outcome not happening (in our case, $Y=0$). If the odds are greater than one, then the outcome is more likely to happen than not. If we have odds, we can always convert it to probabilities and vice versa.

Since in the logit model, $\mathbb{P}(Y=1 \mid \mathbf{x}) = \Lambda(\mathbf{x}'\beta)$, we have that

$$\frac{\mathbb{P}(Y=1 \mid \mathbf{x})}{1-\mathbb{P}(Y=1 \mid \mathbf{x})} = e^{\mathbf{x}'\beta}.$$

Taking logs, we get

$$\ln \frac{\mathbb{P}(Y=1 \mid \mathbf{x})}{1-\mathbb{P}(Y=1 \mid \mathbf{x})} = \mathbf{x}'\beta.$$

The term on the left is called the *log-odds* or *logit*. For the logit model, the logit turns out to be a *linear* function of the predictors. Therefore, each coefficient β_j has the interpretation of the marginal effect of the predictor X_j on the logit.

Thinking about the marginal effect on the log-odds can be a little unintuitive. Instead, we can make use of exponentiated coefficients, e^{β_j}. Let's denote the odds as $\text{Odds}(\mathbf{x})$. Then

$$\begin{aligned}
\text{Odds}(X_1, \ldots, X_j+1, \ldots, X_k) &= e^{\beta_0+\beta_1 X_1+\ldots+\beta_j(X_j+1)+\ldots+\beta_k X_k} \\
&= e^{\beta_0+\beta_1 X_1+\ldots+\beta_j X_j+\ldots+\beta_k X_k} e^{\beta_j} \\
&= \text{Odds}(\mathbf{x}) e^{\beta_j}.
\end{aligned}$$

Therefore, the exponentiated coefficient e^{β_j} tells one by how much the odds change (in multiplicative terms) when X_j increases by one unit. Here, if $e^{\beta_j} > 1$, then X_j has a positive effect on the probability of $Y=1$, if $e^{\beta_j} < 1$, then X_j has a negative effect on the probability of $Y=1$, and if $e^{\beta_j} = 1$, then X_j has no effect.

Even the odds can often be a little hard to interpret. Instead, we might want to work directly with the marginal effect of X_j on the probability of $Y = 1$. Let's compute it.

$$\frac{\partial \mathbb{P}(Y = 1 \mid \mathbf{x})}{\partial X_j} = \frac{\partial \Lambda(\mathbf{x}'\beta)}{\partial X_j} = \beta_j \Lambda'(\mathbf{x}'\beta).$$

Clearly, β_j is *not* the marginal effect of X_j on the probability of $Y = 1$. Instead, the marginal effect of X_j now depends on the values of *all* other predictors, it is not constant.

It is easy to show that the logistic CDF has the following property:

$$\Lambda'(z) = \Lambda(z)(1 - \Lambda(z)).$$

Therefore, the marginal effect of X_j can be written as

$$\frac{\partial \mathbb{P}(Y = 1 \mid \mathbf{x})}{\partial X_j} = \beta_j \Lambda(\mathbf{x}'\beta)(1 - \Lambda(\mathbf{x}'\beta)).$$

Once we estimate the parameters, we have that the predicted probability of class A is $\hat{y}_i = \Lambda(\mathbf{x}_i'\hat{\beta})$, and therefore

$$\frac{\partial \mathbb{P}(Y = 1 \mid \mathbf{x}_i)}{\partial X_j} = \hat{\beta}_j \hat{y}_i(1 - \hat{y}_i).$$

This expression shows that the marginal effect of X_j depends on the predicted probability in a reasonable way. When the predicted probability is already high with $\hat{y}_i$ close to 1, then the marginal effect of X_j on the probability will be close to zero. The marginal effect of X_j on the probability will be largest when the probability is close to 0.5.

The marginal effect of X_j on the probability $\mathbb{P}(Y = 1 \mid \mathbf{x})$ is not a single number, it is a function of the values of all the predictors. But often we would like to get a single number that summarizes the effect of a given predictor. In this case, we can compute the *average marginal effect* (AME) by computing the marginal effects of X_j on the probability for each observation and then compute the average of those effects:

$$\text{AME}(X_j) = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_j \hat{y}_i(1 - \hat{y}_i).$$