Lecture 15 Classification Pt. 2

Hypotheses testing

Each estimated logit coefficient $\hat{\beta}_j$ comes with an asymptotic standard error. The formula for it is somewhat complicated and we will not discuss it here. Once we have the standard errors, we can construct asymptotic t-tests and confidence intervals, just as with OLS. In particular, to test $H_0 : \beta_j = 0$ we form the t-statistic $\hat{\beta}_j / \text{se}(\beta_j)$ and carry out the test in the usual way, once we have decided on a one- or two-sided alternative.

We can also test multiple restrictions in the logit model. Here we will focus only on the exclusion restrictions. Our *unrestricted* model is

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

We will denote the log-likelihood from the unrestricted model as $\mathcal{L}_{ur}$.

The null hypothesis in the test of q exclusion restrictions can be written, as before, as

$$H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0,$$

without loss of generality. The alternative hypothesis is that at least one of these coefficients is non-zero. If the null hypothesis is true, we obtain the *restricted* model

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-q} X_{k-q}).$$

We will denote the log-likelihood from the restricted model as $\mathcal{L}_r$.

We can use the *likelihood ratio* (LR) test to test the exclusion restrictions. The LR test is based on the same idea as the F test in a linear model. The F test measures the increase in the sum of squared residuals when we go from the unrestricted to the restricted model. The LR test, instead, is based on the difference in the log-likelihoods for the unrestricted and restricted models. Because the MLE maximizes the log-likelihood function, dropping variables generally leads to a smaller log-likelihood. This is similar to how the sum of squared residuals never decreases when variables are dropped from a regression. The question is whether the fall in the log-likelihood is *large enough*.

To conduct the LR test, we first compute the *likelihood ratio* statistic:

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r).$$

Since $\mathcal{L}_{ur} \geqslant \mathcal{L}_r$, the LR statistic is non-negative (typically positive). Note, however, that each of the log-likelihoods *can* be negative.

To make statistical inference, we need to know the sampling distribution of our computed LR statistic. The following result holds.

Sampling distribution of the LR statistic

If H_0 holds, the LR statistic follows an asymptotic chi-square (χ^2) distribution with q degrees of freedom:

$$LR \dot{\sim} \chi^2_q$$

Since the LR statistic is non-negative, to reject the null hypothesis, the value of the statistic has to be *large enough*. The general decision rule is

$$\begin{cases} LR > c, \text{ reject } \mathrm{H}_0 \\ LR \leqslant c, \text{ do not reject } \mathrm{H}_0. \end{cases}$$

The critical value c will be the $1-\alpha$ percentile of the χ^2 distribution with q degrees of freedom, where α is a chosen significance level.

As with the F test, if we reject the null hypothesis, we would say that $\beta_{k-q+1}, \ldots, \beta_k$ are *jointly statistically significant*. If we do not reject the null, we say that the coefficients are *jointly insignificant*. However, we cannot tell which coefficient exactly is significant and which one is not.

Instead of using critical values to conduct the LR test, we can also use p-values. The p-value of an LR test is defined as

$$p \equiv \mathbb{P}(Z > LR),$$

where Z is a random variable following a χ^2 distribution with q degrees of freedom.

The interpretation of the p-value for an LR test is identical to its interpretation for an F test. To reject the null, the p-value has to be *low enough*. Our decision rule for a chosen α significance level will be

$$\begin{cases} p \leqslant \alpha, \text{ reject } H_0, \\ p > \alpha, \text{ do not reject } H_0. \end{cases}$$

Example

In this example, we will be estimating a model of a woman's labor force participation. Our outcome variable ($inlf$) will be equal to one if a woman participates in the labor force and zero otherwise. Our predictors will be a woman's age, education, work experience, number of kids smaller than 6yo, and other family income.

The first model we will try is the linear probability model:

$$inlf_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 kidslt6_i + \beta_5 nwifeinc_i + u_i.$$

We will also estimate the logit model:

$$\mathbb{P}(inlf_i = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 kidslt6_i + \beta_5 nwifeinc_i)$$

The table below shows the coefficient estimates from the linear probability model (LPM) and logit. The logit estimates are presented as the raw coefficients ("conditional") and as

average marginal effects ("marginal"). We can observe that the LPM estimates are very similar to the average marginal effects from the logit.

	LPM	Logit (coef)	Logit (AME)
(Intercept)	0.770	1.153	
	(0.135)	(0.742)	
age	-0.019	-0.095	-0.017
	(0.002)	(0.013)	(0.003)
educ	0.039	0.223	0.041
	(0.007)	(0.043)	(0.009)
exper	0.022	0.118	0.021
	(0.002)	(0.013)	(0.003)
kidslt6	-0.275	-1.464	-0.266
	(0.033)	(0.200)	(0.043)
nwifeinc	-0.003	-0.020	-0.004
	(0.001)	(0.008)	(0.002)
Num.Obs.	753	753	753

Predictions

Our fitted values $\hat{Y}$ will be the predicted probabilities of $Y = 1$ given the predictors. Recall, though, that our original outcome is in terms of belonging to either class A or class B. How do we get predictions in terms of classes? We just need to specify a threshold probability $\bar{p}$ that will determine whether we assign an observation to class A or B based on the predicted class probability.

$$\tilde{Y} = \begin{cases} \text{class A}, \text{ if } \hat{Y} \geqslant \bar{p}, \\ \text{class B}, \text{ if } \hat{Y} < \bar{p}. \end{cases}$$

A natural choice for $\bar{p}$ is 0.5. In other words, if an observation has a more than 50% chance to belonging to class A, we assign it to class A, and vice versa. However, $\bar{p} = 0.5$ is not the only possibility. The threshold probability often depends on the application. For example, in some cases the cost of misclassifying an observation as belonging to one class may be higher than the cost of misclassifying it as the other class, which would lead to a different threshold probability.

Goodness-of-fit

The goodness-of-fit measures for binary classifiers are closely linked to how we get the fitted values (or predictions) in terms of classes. It is often helpful to designate one of the classes as the "positive" class or an *event*. What is considered an event depends on the application. For example, in a medical setting, an event could be a patient developing a decease.

After we make class predictions, we can create a *confusion matrix*, which calculates the number of observations in each cell of a 2x2 table as follows

	Truth: Event	Truth: No event
Prediction: Event	A	B
Prediction: No event	C	D

From the confusion matrix we can derive a bunch of performance measures. One of the most basic performance measures we can compute is *accuracy* :

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}.$$

Accuracy is simply the proportion of observations we classified correctly. The opposite of accuracy is the *error rate*:

$$\text{Error rate} = 1 - \text{Accuracy}.$$

It is useful to compare accuracy to the probability of the most common class. If one class is much more prevalent than the other, the accuracy measure can be misleading.

Those aggregate measures can be broken down further by class. For example, we can ask how many events we classified correctly. The resulting measure is called *sensitivity*:

$$\text{Sensitivity} = \frac{A}{A + C}.$$

Sensitivity is also known as the *true positive rate*.

A related measure is how many non-events we classified correctly. This measure is called *specificity*:

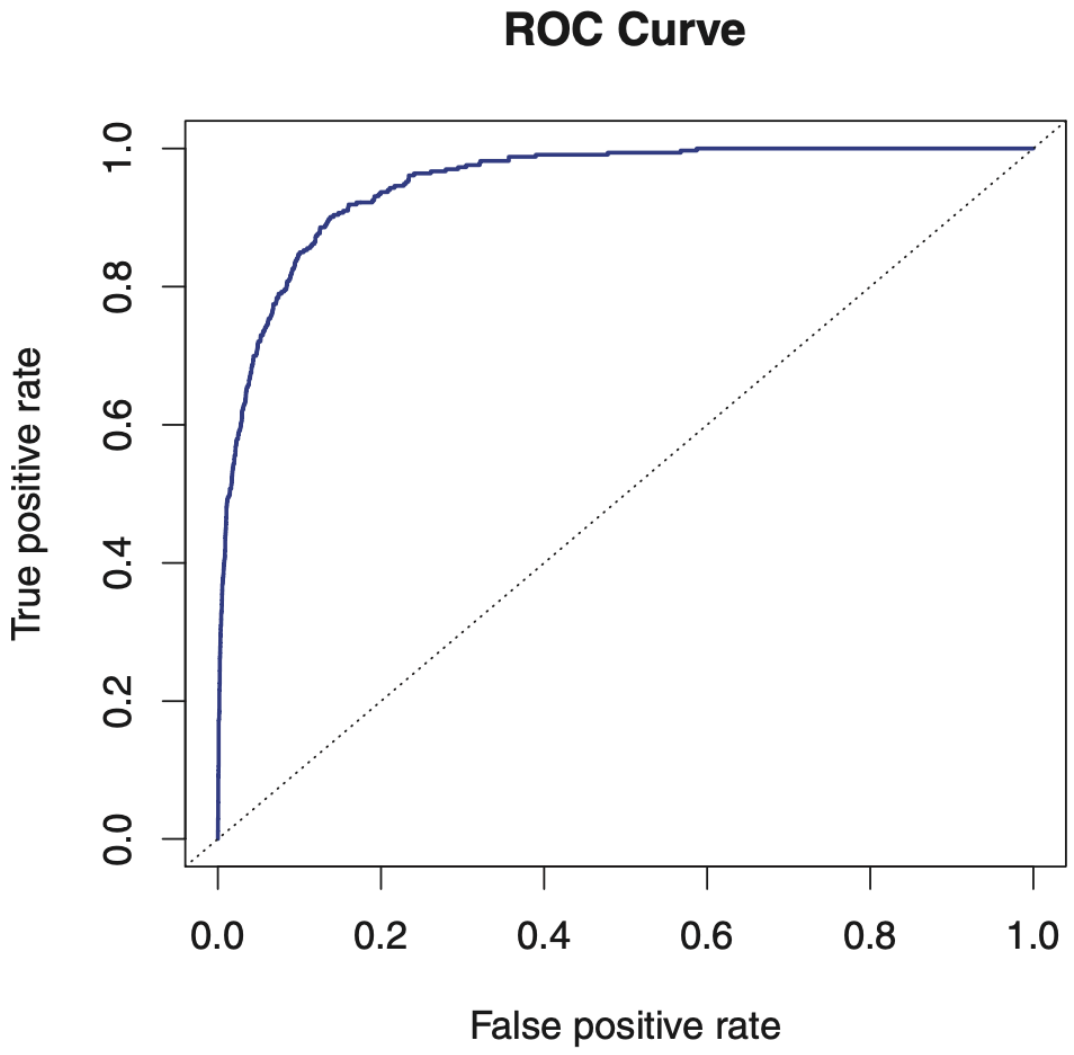
$$\text{Specificity} = \frac{D}{B + D}.$$

Specificity is also known as (1 - the *false positive rate*).

Recall that our predictions in terms of classes, and hence events, depend on the threshold probability we set. The threshold affects true and false positive rates and creates a trade-off. A high threshold for classifying an observation as an event will lead to fewer predicted events. This will imply a low sensitivity but also a high specificity (low false positive rate). On the other hand, a low threshold will lead to more predicted events, which will imply a high sensitivity and a low specificity (high false positive rate).

We can summarize this trade-off using the *ROC curve*, which plots the true and false positive rates for each possible threshold. A typical ROC curve looks like this. The thresholds are implicit in this graph. As we move away from the origin we are decreasing

the threshold for an event.



The area under the ROC curve (*AUC*) is a summary measure of a classifier's performance. The smallest possible value is 0.5, which corresponds to making predictions at random. The highest possible value is 1, which would correspond to the AUC hugging the top left corner.