



UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu

The University of Edinburgh Business School

A photograph of a modern, multi-story building with a glass and metal facade, identified as the University of Cambridge Business School. The building features a prominent corner with large windows and a dark metal frame. The sky is blue with some light clouds.

Estimation and Comparison

Likelihood ratio test

Wald test

Recap

- Logistic regression: binary and multi-level outcomes
- Main difference between logistic and linear regression: target variable
- The expected value of a binary variable is its proportion, which can be interpreted as a probability of an event of interest.
- The probability and predictors is not linear, although it is monotonic → link function:
 - logit or log-odds transformation of the probability of the event
 - Probit: the inverse normal link function
- MLE

Recap

- Interpret the logistic regression coefficients: exponentiating the corresponding coefficients that have a multiplicative effect on odds of y .
 - The exponentiated coefficients are said to have no effect when they are equal to 1.
 - <1 produce a negative effect, that is, decrease the odds of the event of interest
 - >1 produce positive effect, that is, increase the odds of the event of interest
- For linear regression, there are two types of tests:
 - F-test - a general one to assess the whole model
 - t-test - used to assess the significance of each coefficient separately

Likelihood ratio test

A model is said to be **nested** inside another one if it includes only a subset of the predictors included in the second model. For example:

- The complete model (M_C): $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- The reduced model (M_R): $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1$

More general

The **complete** model M_C : $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q + \beta_{q+1} X_{q+1} + \dots + \beta_p X_p$

The **reduced** model M_R : $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q$

We indicate with $L(M_C)$ and $L(M_R)$ the values of the (maximized) likelihood function for the complete and the reduced models respectively.

Likelihood ratio test

The likelihood ratio test (LRT) corresponds in the logistic regression context to the F test for nested models in linear regression.

$$M_C: \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q + \beta_{q+1} X_{q+1} + \dots + \beta_p X_p; \quad M_R: \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q$$

- The hypotheses compared by the LRT are:

$$H_0: \beta_{q+1} = \dots = \beta_p = 0 \quad \text{vs} \quad H_1: \text{at least one of the } \beta_j \text{ in } H_0 \text{ is } \neq 0.$$

The test statistic of the LRT compares the deviances for the two models. The deviance of a logistic regression model M is defined as

$$DEV.RES(M) = -2 \ln L(M)$$

The deviance is a measure of the response variability left unexplained by the model. It turns out that the **LRT test statistic** is given by

$$\begin{aligned} \Lambda &= DEV.RES(M_R) - DEV.RES(M_C) && -2(l_0 - l_1) \\ &= -2 \ln L(M_R) + 2 \ln L(M_C) = -2 \ln \left(\frac{L(M_R)}{L(M_C)} \right) = -2(l_R - l_C), && l_R = \ln(L(M_R)) \end{aligned}$$

This test statistic follows χ^2 distribution with degrees of freedom equal to the difference in the number of parameters tested between the two models.

Likelihood ratio test

$$H_0 : \beta_{q+1} = \dots = \beta_p = 0 \quad \text{vs} \quad H_1 : \text{at least one of the } \beta_j \text{ in } H_0 \text{ is } \neq 0.$$

The **LRT** test statistic is given by $\Lambda = -2 \ln \left(\frac{L(M_R)}{L(M_C)} \right)$

The test will reject the null hypothesis H_0 if the ratio $\frac{L(M_R)}{L(M_C)}$ is sufficiently small, that is the test statistic value Λ is large enough, suggesting a significant reduction in the model deviance.

As usual, we reject the null hypothesis at the α level (e.g. $\alpha=0.05$) if the corresponding p-value is smaller than α .

If the null hypothesis comprises all the coefficients(excluding the intercept), this corresponds to the global usefulness F test in linear regression.



Likelihood ratio test: the credit default example

Model 1:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1$$

Since X_1 is a binary dummy variable, we can simplify it further:

For $X_1 = 0$: $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 \rightarrow$ the baseline or null model

For $X_1 = 1$: $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1$

Test a hypothesis that the model with the Telephone Known is statistically significant in comparison to the baseline one:

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

The corresponding likelihood-ratio test statistic:

$$\Lambda = -2 \ln\left(\frac{L(M_{null})}{L(M_1)}\right) \sim \chi^2_1$$

Variable	Description
Y	Credit Status: 1 = Default/Bad customer; 0 = Non-default/Good customer
X_1	Telephone Known?: 1=yes; 0 = no
X_2	Own Housing: 1= Owner; 0 = otherwise
X_3	Free Housing: 1= does not pay for the housing; 0=otherwise
X_4	Age in years

Df Model: 1	# degrees of freedom
Log-Likelihood: -418.77	# alternative model
LL-Null: -419.70	# baseline
LLR p-value: 0.1728	

Do not reject H_0

Likelihood ratio test: the credit default example

What happens if we add Housing?

Model Null:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0$$

Model 2:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs $H_1: \text{at least one of } \beta_1, \beta_2, \beta_3 \neq 0$
Degrees of freedom = 3

Variable	Description
Y	Credit Status: 1 = Default/Bad customer; 0 = Non-default/Good customer
X ₁	Telephone Known?: 1=yes; 0 = no
X ₂	Own Housing: 1= Owner; 0 = otherwise
X ₃	Free Housing: 1= does not pay for the housing; 0=otherwise
X ₄	Age in years

Df Model: 3 # degrees of freedom
Log-Likelihood: -410.27 # alternative model
LL-Null: -419.70 # baseline
LLR p-value: 0.0002907

Reject H_0

Likelihood ratio test: the credit default example

What about Age?

Model Null:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0$$

Model 3:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ vs}$$

$$H_1: \text{at least one of } \beta_1, \beta_2, \beta_3, \beta_4 \neq 0$$

Degrees of freedom = 4

Variable	Description
Y	Credit Status: 1 = Default/Bad customer; 0 = Non-default/Good customer
X ₁	Telephone Known?: 1=yes; 0 = no
X ₂	Own Housing: 1= Owner; 0 = otherwise
X ₃	Free Housing: 1= does not pay for the housing; 0=otherwise
X ₄	Age in years

Df Model: 4

degrees of freedom

Log-Likelihood: -407.00

alternative model

LL-Null: -419.70

baseline

LLR p-value: 4.162e-05

Reject H_0

Likelihood ratio test: the credit default example

How can we test if Age adds significantly to the model?

We can compare Model 3 to Model 2 using the same likelihood ratio test.

Model 2:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Model 3:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$\Lambda = -2 \ln\left(\frac{L(M_2)}{L(M_3)}\right) = -2(l_2 - l_3) = -2(-410.27 + 407.00) = 6.54$$

Degrees of freedom are $4-3=1$, and the probability for the value of 6.54 from a χ^2 distribution with 1 df is 0.0105. Significant at 5% level, so Age does bring some useful information about future credit behaviour.

Wald test

The Wald test, which checks the significance of each parameter separately. The Wald test statistic is computed by dividing the estimated coefficient of interest by its standard error (similar to t-test). This test statistic follows approximately a standard normal or Z distribution in large samples.

$$Z = \frac{\hat{\beta}}{s_{\hat{\beta}}} \sim N(0, 1) \qquad Z^2 = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}} \right)^2 \sim \chi^2_{df=1}$$

The square of Z follows approximately a χ^2 distribution with one degree of freedom.

Note that Wald test works well for larger samples. For small samples, likelihood ratio test is preferred.

Wald test

The Wald test is convenient because you can evaluate at least an approximate significance of many variables from just one run.

Model 3:

	coef	std err	z	P> z	[0.025	0.975]
const	0.2822	0.307	0.919	0.358	-0.320	0.884
Telephone	-0.2215	0.178	-1.248	0.212	-0.569	0.126
Own House	-0.6054	0.214	-2.830	0.005	-1.025	-0.186
Free House	0.3613	0.326	1.108	0.268	-0.278	1.000
Age	-0.0210	0.008	-2.495	0.013	-0.037	-0.004

Note that this layout gives us the linear predictor part or the right-hand side of logistic regression in linear form. In other words we are talking about zero or no-zero effect of variables on logit. To get the interpretation in terms of odds, we need to exponentiate the estimated coefficients.

A photograph of a modern, multi-story building with a glass and metal facade, identified as the University of Cambridge Business School. The building features large windows and a prominent corner structure. A sign on the ground floor reads "UNIVERSITY OF CAMBRIDGE Business School".

Model evaluation

Evaluating models: McFadden pseudo R- squared

Pseudo R-squared: they do not really measure the proportion of variance explained by the model, they are just relative indices forced to take values between 0 and 1; The higher values do indicate better model fit.

- 'Statsmodels' package in Python reports **McFadden pseudo R- squared**:

is the log likelihood of the full model and l_0 is the log likelihood of the baseline (constant-only) model.

Evaluating models: McFadden pseudo R- squared

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{the sum of squared errors of the fitted model}}{\text{total sum of squares of baseline model}}$$

the degree to which the alternative model parameters improve upon the prediction of the baseline/null model.

$$R_{McF}^2 = 1 - \frac{l_1}{l_0} = \frac{\text{The log likelihood of the full model}}{\text{The log likelihood of the baseline model}}$$

indicates the level of improvement achieved by the alternative model over the null one.

When comparing two models trained on the same sample, McFadden R-squared will be higher for the model with the larger likelihood.

Evaluating models: McFadden pseudo R-squared

$$R_{McF}^2 = 1 - \frac{l_1}{l_0}$$

For Model 1: $R_{McF}^2 = 1 - \frac{-418.77}{-419.7} = 0.002216$

For Model 3: $R_{McF}^2 = 1 - \frac{-407}{-419.7} = 0.03026$

Model 3 has higher R_{McF}^2 , therefore is better than Model 1.

Model1

Df Model: 1 # degrees of freedom

Log-Likelihood: -418.77 # alternative model

LL-Null: -419.70 # baseline

LLR p-value: 0.1728

Model3

Df Model: 4 # degrees of freedom

Log-Likelihood: -407.00 # alternative model

LL-Null: -419.70 # baseline

LLR p-value: 4.162e-05

Evaluating models

Cox and Snell pseudo R - squared: $R_{CS}^2 = 1 - \left(\frac{L_0}{L_1}\right)^{2/n}$

R_{CS}^2 builds on the likelihood ratio test statistic $-2\ln\left(\frac{L_0}{L_1}\right)$. The ratio of likelihoods indicates the improvement of the alternative model over the baseline. We can interpret the likelihood as the conditional probability of the outcome variable given the predictors. If there are n observations in the sample, the likelihood is the product of n such probabilities. Therefore, taking the n -th root of this product gives an estimate of the likelihood for a single observation.

If the alternative model predicts the outcome perfectly, it should have a likelihood of 1, R_{CS}^2 is then $1 - L_0^{2/n}$, which is less than one.

Nagelkerke or Cragg and Uhler adjustment: $R_N^2 = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - L_0^{2/n}}$

AIC and BIC

- When experimenting with different models, it is possible to increase the likelihood just by adding parameters, however, this may lead to overfitting. There are two measures that try to address this problem by introducing a penalty term for the number of parameters in the model. These are **Akaike's Information Criterion (AIC)** and **Schwarz or Bayesian Information Criterion (BIC)**. Both measures originate from the Information Theory, and they are very popular for predictive models that are based on maximum likelihood estimation.

- Akaike's Information Criterion is given by:

$$AIC = -2l + 2p$$

where p is the number of parameters in the fitted model.

- Bayesian Information Criterion is given by:

$$BIC = -2l + p \ln(n)$$

where p is the number of parameters in the fitted model, and n is the sample size.

AIC and BIC

$$AIC = -2l + 2p, \quad BIC = -2\ell + p\ln(n)$$

The AIC and BIC measures provide two different ways of adjusting the statistic for the number of predictors in the model and BIC additionally - for the number of observations used. Therefore, these measures can be used when comparing different models for the same data, but not necessarily nested models. Lower values of these measures indicate a better model.

Let's calculate AIC and BIC for Models 1 and 3 considered above. Note that the intercept is also included into the number of parameters, and we use the training sample with 700 observations.

AIC

$$\text{For Model 1: } AIC = -2(-418.77) + 2 \times 2 = 841.54$$


$$\text{For Model 3: } AIC = -2(-407) + 2 \times 5 = 824$$


BIC

$$\text{For Model 1: } BIC = -2(-418.77) + 2 \times \ln(700) = 850.64$$

$$\text{For Model 3: } BIC = -2(-407) + 5 \times \ln(700) = 846.75$$

Both measures indicate that Model 3 is better.

- 
- You have seen how to interpret the results from logistic regression. Now, you have the opportunity to read more about a particular application of logistic regression in an academic paper that investigates the consequences of not being able to use certain information in a scoring model.
 - Read the following paper about the consequences of removal of 'Gender' variable from a scoring model:
 - [Andreeva G., Matuszyk A. \(2019\) 'The Law of Equal Opportunities or Unintended Consequences: the impact of unisex risk assessment in consumer credit', Journal of Royal Statistical Society, Series A](#)
 - You should read up to section 4.2.1 (you only need the part on logistic regression, although, of course, you are more than welcome to read the whole paper!)

- 
- A photograph of a modern building with large glass windows and a sign that reads 'UNIVERSITY OF CAMBRIDGE Business School'.
- Try the following exercise:
 - 14 - old - Activity 8 - Coding for logistic regression.ipynb

Activity: logistic regression exercise



UNIVERSITY OF EDINBURGH
Business School