Modeling Binary Outcomes Using Logistic Regression with a Predictor

Part 2 – Fitting the Logistic Regression Model and Assessing the significance of the predictor in the model

Learning Objectives

In this lecture, you will learn how to:

- fit the logistic Regression Model to Data in R.
- Assess the significance of the predictor in the model

Example

A random sample of students is selected from a large statistics class.

The following variables are recorded:

- Number of hours studied
- Exam outcome, **pass (P)** or **fail (F)**

Hours	Grade
0	F
0	F
0.5	F
1.5	F
1.5	F
1.5	P
2	F
2.5	F
2.5	F
⋮	⋮
10.5	P
11	P
11	P

*The full dataset '**Hours-and-Grades**' can be downloaded from Brightspace*

Modeling

The objective of using the data is to use the number of hours studied as a predictor to model the probability of passing.

The model described is a **Logistic Regression model**, which relates the **log**-odds of passing to a linear function of study hours.

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{log-odds function of } p} = \underbrace{A + B * Hours}_{\text{the linear function of the predictor (e.g., hours) as used in a standard regression model}}$$

log-odds function of p

the linear function of the predictor (e.g., hours)

as used in a standard regression model

Fitting the logistic Regression Model to Data

- To fit the logistic regression model to data, we need to use the sample data to estimate two unknown parameters, **A (intercept)** and **B (slope)**.

$$\ln\left(\frac{p}{1-p}\right) = A + B * \boldsymbol{Hours}$$

- We will skip the details of estimating A and B, and
- simply use R to handle the estimations for us.

Fitting a Logistic Regression Model to Data using R

The R code has been saved in a separate file. Please open the file and run the code.

Below are the results:

```
Call:
glm(formula = y ~ x, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8984     0.9694  -2.990 0.002791 **
x             0.6734     0.1860   3.621 0.000294 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64.104  on 49  degrees of freedom
Residual deviance: 36.354  on 48  degrees of freedom
AIC: 40.354
```

Assessing the Predictive Effectiveness of the Logistic Regression Model

- After fitting the logistic regression model,
- it is important to assess whether the model is significantly useful for predicting a student's **probability** of passing the exam.
- In this model, the **number of hours studied** is used as the predictor.
- In other words, we need to test whether the **number of hours studied**
- is a **significant predictor** of the **probability** of passing the exam.

  (or is **significantly related** to the **probability** of passing the exam)

  (or has a **significant effect** on the **probability** of passing the exam)
- Let's conduct a hypothesis test.

Step 1 - State the null and alternative hypotheses

H_0: The number of hours studied is NOT a **significant predictor** of the **probability** of a student passing the exam.

$$\Rightarrow \quad p = \frac{e^{A + \overset{0}{\cancel{B}} * Hours}}{1 + e^{A + \overset{0}{\cancel{B}} * Hours}} \quad \Rightarrow \quad p = \underbrace{\frac{e^{A}}{1 + e^{A}}}_{\text{constant}}$$

H_a: The number of hours studied is a **significant predictor** of the **probability** of a student passing the exam

$$\Rightarrow \quad p = \frac{e^{A + B * Hours}}{1 + e^{A + B * Hours}} \quad \text{where} \quad B \neq 0$$

Step 2 – Obtain the test statistic and p-value

- To assess the significance of the predictor(H_o: $B = 0$ vs H_a: $B \neq 0$),
- we examine the coefficient table,
- focusing on the row corresponding to hours (the slope).
- The last columns, labeled **z-value** and **p-value**, provide the test statistic and p-value that we use to assess the significance of the predictor.
- This method is known as the **Wald test**.

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value
Intercept	-2.8984	0.9694	-2.99	0.002791
X (Hours)	0.6734	0.186	3.621	0.000294

Step 3 – State the conclusion

Term	Coef	SE Coef	Z-Value	P-Value
Intercept	-2.8984	0.9694	-2.99	0.002791
X (Hours)	0.6734	0.186	3.621	0.000294

Let's compare the p-value with the significance level (0.05).

Since the p-value (0.000294) is less than the significance level (0.05), the sample data provide sufficient evidence to conclude that the number of hours studied is a **significant predictor** of the **probability** of a student passing the exam.

Alternative Approach - Likelihood Ratio Test (LRT)

Under **Ho**, the logistic regression model only contains the intercept *A*

$$\boldsymbol{p} = \frac{e^{A}}{1 + e^{A}}$$

Under **Ha**, the logistic regression model contains the intercept *A* and predictor's term (**B** * **Hours**)

$$\boldsymbol{p} = \frac{e^{A + B * \boldsymbol{Hours}}}{1 + e^{A + B * \boldsymbol{Hours}}}$$

Likelihood Ratio Test (LRT) - Details

- We fit both models to the data and compute their **log-likelihoods**.
- Then we calculate the **likelihood ratio Chi-Square statistic**, G^2

$$G^2 = 2\left(log-likelihood \begin{pmatrix} Model \\ Under\ \boldsymbol{H_a} \end{pmatrix} - log-likelihood \begin{pmatrix} Model \\ Under\ \boldsymbol{H_0} \end{pmatrix} \right)$$

Note:

- The **alternative model** (**with the predictor**) will always have a **larger** **log-likelihood** than the **null model (intercept-only)**.
- This is because **adding predictors gives the model more flexibility** to fit the data, so it can never fit worse than a simpler model.

Likelihood Ratio Test (LRT) - Details

$$G^2 = 2\left(log{-}likelihood\begin{pmatrix}Model\\Under\ \boldsymbol{H_a}\end{pmatrix} - log{-}likelihood\begin{pmatrix}Model\\Under\ \boldsymbol{H_0}\end{pmatrix}\right)$$

- If the **null model (without the predictor)** predicts probabilities nearly **as well as** the **alternative model (with the predictor)**,
- then the G^2 is small, indicating little evidence against H_0

- If the **alternative model** predicts probabilities **significantly better** than **null model** then the G^2 is **large**, indicating **strong evidence** against H_0

- To assess whether G^2 provides evidence against $\boldsymbol{H_0}$, we convert G^2 to a p-value. How?
- G^2 follows a **Chi-Square Distribution** and the **degree of freedom** is given by **the number of regression coefficients being tested**.

Likelihood Ratio Test (LRT) in R

- We can carry the Likelihood Ratio Test in R.
- Please follow the commands below.

```
# Fit a logistic regression model with no predictors.
# The model includes only an intercept (denoted by "1").
fitted.model.no.predictor = glm( y ~ 1, family = binomial)

# Conduct the Likelihood Ratio Test (LRT)
# using the anova() function
# Compare the alternative model (with predictor)
# against the null model (intercept only)
anova(
  fitted.model.no.predictor, # the null model (without predictor)
  fitted.model,              # the alternative model (with predictor)
  test = "Chisq"             # specify the Chi-Square test
)
```

- The output of the likelihood ratio test is shown below.

```
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ x
  Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
1        49     64.104
2        48     36.354  1   27.749 1.381e-07 ***
```

$G^2 = 27.749$,
P-value $= \underbrace{0.000000}_{7\ zeros}1381$

Residual Deviance $= -2 \times \log\text{–}likelihood$

In the output,
$-2 \times \log\text{–}likelihood(H_0) = 64.104$
$-2 \times \log\text{–}likelihood(H_a) = 36.354$

Degree of freedom is 1 b/c only **one** predictor is being tested

Wald Test vs Likelihood Ratio Test (LRT)

- There are two tests available for assessing the significance of the predictor. Which one is preferable?
- Likelihood Ratio Test is **usually preferred** over the Wald test,
- Because LRT tends to give more **accurate p-values** and is **more robust**.
- even the sample size is small.
- The Wald test can be okay for quick checks if the sample is large.