Modeling Binary Outcomes Using Logistic Regression with a Predictor

Part 3 – Estimating the Probability of Success

Learning Objectives

In this lecture, you will learn how to:

- Predict the probability of success using the logistic regression model
- Calculate its confidence interval

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** 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**.

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

Fitting a Logistic Regression Model to Data using R

We fit the logistic regression model to the data and obtain the following output.

```
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
```

- Once we have determined that the **number of hours** is **significant predictor** of a student's probability of passing the exam,,
- we can use the **regression equation** to predict this **probability** based on the **number of hours studied.**
- To do that, we read the **coefficient table** from the regression output and use the values in the "Estimate" column

```
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 ***
```

A (Intercept) = -2.8984, B (Slope) = 0.6734

$$\ln\left(\frac{p}{1-p}\right) = -2.8984 + 0.6734 * Hours$$

Let's re-write the regression equation in terms of p

$$\ln\left(\frac{p}{1-p}\right) = -2.8984 + 0.6734 * Hours$$

$$p = \frac{e^{-2.8984+0.6734*Hours}}{1+e^{-2.8984+0.6734*Hours}}$$

Now, we want to estimate the **probability** that a student will pass the exam if the student studies for **0 hours.**

$$p = \frac{e^{-2.8984+0.6734*\underbrace{Hours}_{0}}}{1+e^{-2.8984+0.6734*\underbrace{Hours}_{0}}}$$

$$= \frac{e^{-2.8984}}{1+e^{-2.8984}}$$

$$\approx 0.055$$

If a student studies for **0 hours**, the estimated **probability of passing the exam** is 0.055.

Confidence Interval for the Probability

- The **estimated probability** we calculated always differs from its **true value**.
- Therefore, we always use a **confidence interval** that provides a range of possible value for the **probability**.
- The **confidence interval** for the true **probability** at a given x-value, x_0 is calculated in four steps.

Step 1 – Calculate the estimated log-odds from the regression equation

$$Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds} = \hat{A} + \hat{B}\, x_0$$

Confidence Interval for the Probability

Step 2 – Calculate the variance of the estimated log-odds

$$\boldsymbol{Var}\left[Estimated \underbrace{\ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right)}_{\boldsymbol{log-odds}}\right] = \boldsymbol{Var}[\hat{A} + \hat{B}\, x_0]$$

$$= \boldsymbol{Var}\left[\hat{A}\right] + \boldsymbol{Var}\left[\hat{B}\right](x_0)^2 + 2x_0\boldsymbol{COV}\left(\hat{A}, \hat{B}\right)$$

OR written in a **Matrix form if you are good in Matrix algebra**

$$= [1, x_0\,] \underbrace{\begin{bmatrix} \boldsymbol{Var}\left[\hat{A}\right] & \boldsymbol{COV}\left(\hat{A}, \hat{B}\right) \\ \boldsymbol{COV}\left(\hat{A}, \hat{B}\right) & \boldsymbol{Var}\left[\hat{B}\right] \end{bmatrix}}_{\text{Covariance Matrix}} \begin{bmatrix} 1 \\ x_0 \end{bmatrix}$$

We will use R to obtain the variance of $\hat{A}$ / $\hat{B}$ and covariance of $\hat{A}$ and $\hat{B}$

There is no need to do any hand calculation.

Confidence Interval for the Probability

Step 3 – Get the Confidence Interval for log-odds

The Confidence Interval for **log**-odds is given by:

$$Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds} \pm z_c \times \sqrt{Var\left[Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds}\right]}$$

$$= \left(\hat{A} + \hat{B}\, x_0\right) \pm z_c \times \sqrt{Var\left[\hat{A}\right] + Var\left[\hat{B}\right](x_0)^2 + 2x_0\, COV\left(\hat{A}, \hat{B}\right)}$$

Confidence Interval for the Prediction

Last Step

- Let (L, U) represent the lower and upper bounds of the confidence interval for the ***log-odds*** calculated in the previous step.
- Recall that the probability of success is given by:

$$p = \frac{e^{A+Bx}}{1+e^{A+Bx}} = \frac{e^{\boldsymbol{log-odds}}}{1+e^{\boldsymbol{log-odds}}}$$

The confidence interval for the true odds is given by:

$$\left(\frac{e^{L}}{1+e^{L}}, \quad \frac{e^{U}}{1+e^{U}}\right)$$

Confidence Interval for the Prediction

From the previous analysis, we estimated that the **probability of passing** is **0.055** if a student studies for **0 hours**.

Question: Construct a 95% confidence interval for the **true probability**

Step 1 – Calculate the estimated log-odds from the regression equation

$$\ln\left(\frac{p}{1-p}\right) = -2.8984 + 0.6734 * \underbrace{Hours}_{0}$$

$$= -2.8984$$

Confidence Interval for the Prediction

Step 2 – Calculate the variance of the estimated log-odds

First, we need to use R to get the variance of $\hat{A}$ / $\hat{B}$ and the covariance of $\hat{A}$ and $\hat{B}$.

The following is the R command:

```
cov.matrix = vcov( fitted.model )
```

R Console

```
> cov.matrix = vcov( fitted.model)
> cov.matrix
             (Intercept)           x
(Intercept)   0.9397409 -0.16273699
x            -0.1627370  0.03459715
```

R Console

```
> cov.matrix = vcov( fitted.model)
> cov.matrix
             (Intercept)           x
(Intercept)   0.9397409 -0.16273699
x            -0.1627370  0.03459715
```

In the matrix,

- Var($\hat{A}$) = Var(intercept) = 0.93974
- Var($\hat{B}$) = Var(slope) = 0.34597
- Cov($\hat{A}$, $\hat{B}$) = -0.16274

So, $Var\left[Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds}\right] = Var[\hat{A} + \hat{B}\, x_0]$

$= \underbrace{Var[\hat{A}]}_{0.93974} + \underbrace{Var[\hat{B}]}_{0.34597}(x_0)^2 + 2x_0 \underbrace{COV(\hat{A}, \hat{B})}_{-0.16274}$

0 hours

$= 0.93974$

Note: You can use R to perform matrix calculation of variance.

$$Var\left[Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds}\right] = [1, x_0]\begin{bmatrix} Var[\hat{A}] & COV(\hat{A},\hat{B}) \\ COV(\hat{A},\hat{B}) & Var[\hat{B}] \end{bmatrix}\begin{bmatrix} 1 \\ x_0 \end{bmatrix}$$

The following is the command:

```
x0 = 0
xvector = matrix( c(1, x0) )
var.log.odds = t(xvector) %*% cov.matrix %*% xvector
var.log.odds
```

R Console

```
> x0 = 0
> xvector = matrix( c(1, x0) )
> var.log.odds = t(xvector) %*% cov.matrix %*% xvector
> var.log.odds
            [,1]
[1,] 0.9397409
```

Step 3 – Get the Confidence Interval for log-odds

$$Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds} \pm z_c \times \sqrt{Var\left[Estimated \underbrace{\ln\left(\frac{p}{1-p}\right)}_{log-odds}\right]}$$

$$= -2.8984 \pm \mathbf{1.96} \times \sqrt{0.93974}$$

$$= (-4.7984, -0.9983)$$

Last Step – Get the Confidence Interval for odds

$$\left(\frac{e^{-4.7984}}{1+e^{-4.7984}}, \frac{e^{-0.9983}}{1+e^{-0.9983}}\right) = (0.0082, \quad 0.2693)$$

If a student studies for **0 hours**,

We are 95% confident that there is between **0.0082 and 0.2693**

probability that this student will pass the exam

Confidence Interval for the Prediction

Exercise: Suppose a student studies for 1 hour.

a) Estimate the probability that the student will pass the exam.
   Please perform the calculation manually (by hand)

b) Use R to construct a 95% confidence interval for the true probability of a student passing the exam