Modeling Binary Outcomes Using Logistic Regression with a Predictor

Part 4 – Interpreting Regression Coefficients and Calculating its Confidence Intervals

Learning Objectives

In this lecture, you will learn how to:

- Interpret the Regression Coefficients in a Logistic Regression Model
- Calculate its confidence interval and interpret it.

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 exam.

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

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

Introduction to Regression Coefficients

In this logistic regression model, there are two parameters:

- Intercept A
- Slope B

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{log-odds function of } p} = A + B * Hours$$

Estimated Regression Coefficients

We have fitted the logistic regression model to data and obtained the following 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 ***
```

Estimated A (Intercept) = -2.8984,

Estimated B (Slope) = 0.6734

Interpreting the Regression Coefficients – Intercept (A)

Let's interpret the **intercept (A)** in the model.

First, we set the **hours** to **ZERO**.

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{log-odds of passing the exam}} = A \cancel{+ B * \underbrace{Hours}_{0}}$$

$$\Rightarrow \ln\left(\frac{p}{1-p}\right) = A \Rightarrow \frac{p}{1-p} = e^{A}$$

If the student studies for **0 hours**, we estimate that

- the **log**-odds of a student **passing the exam** are given by A

OR

- the odds of a student **passing the exam** are given by e^{A}

Interpreting the Regression Coefficients – Intercept (A)

Recall, The **intercept A** is estimated to be $\mathbf{-2.8984}$

So, the **odds** of passing the exam is estimated to $e^{-2.8984} \approx 0.055$

Interpretation:

Interpreting the Regression Coefficients - Slope (B)

Now, we interpret the **"slope" coefficient for the hours (B)** in the model.

First, we set the hours to *k*

p_k is the probability that a student will pass the exam if they studied for k hours

$$\underbrace{\ln\left(\frac{p_k}{1-p_k}\right)}_{\text{Log-odds of passing the exam } \textit{if hours} = k} = A + B * k$$

Log-odds

of **passing the exam** *if* ***hours*** = *k*

Interpreting the Regression Coefficients - Slope (B)

Second, we set the hours to *k + 1*

p_{k+1} is the probability that a student will pass the exam if they studied for k + 1 hours

$$\underbrace{\ln\left(\frac{p_{k+1}}{1-p_{k+1}}\right)}_{\text{Log-odds of passing the exam}} = A + B * (k+1)$$

Log-odds
of passing the exam *if hours = k+1*

$$\ln\left(\frac{p}{1-p}\right) =$$

Now, we take the difference between two **log**-odds between studying for k hours and k+1 hours

$$\ln\left(\frac{p_{k+1}}{1-p_{k+1}}\right) = A + B * k + B$$

$$- \quad \ln\left(\frac{p_k}{1-p_k}\right) = A + B * k$$

$$\underbrace{\ln\left(\frac{p_{k+1}}{1+p_{k+1}}\right) - \ln\left(\frac{p_k}{1+p_k}\right)}_{\text{Mathematically, we combine two log terms into a single log.}} = B$$

$$\ln(x) - \ln(y) = \ln\left(\frac{x}{y}\right)$$

odds of passing the exam *if hours* = $k+1$

$$\ln\left(\frac{\dfrac{p_{k+1}}{1+p_{k+1}}}{\dfrac{p_k}{1+p_k}}\right) = B$$

odds of passing the exam *if hours* = k

odds of passing the exam *if* **hours** = k+1

$$\ln\left(\frac{\dfrac{p_{k+1}}{1+p_{k+1}}}{\dfrac{p_k}{1+p_k}}\right) = B$$

odds of passing the exam *if* **hours** = k

Odds ratio

$$\Rightarrow \quad \frac{\dfrac{p_{k+1}}{1+p_{k+1}}}{\dfrac{p_k}{1+p_k}} = e^{B}$$

So, we interpret the **"slope" (B)** associated **with Hours** as follow:

For **every additional hour** spent studying for the exam,

We predict the **log**-odds of a student **passing the exam increases** by **B**

OR we predict the odds of a student **passing the exam increases** by

a factor of e^{B} **(if** $B > 0$**)**

Question: If $B < 0$, does the odds increase or decrease?

Recall, the slope B is estimated to be 0.6734

$$\longrightarrow e^{\hat{B}} = e^{0.6734} = 1.96072$$

Interpretation:

For every additional hour spent studying for the exam, we predict

E.g. Recall from the previous calculation.

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

  (This mean for every 100 students who fail, we expect about 5.5 students who pass)

- If a student studies one hour, the odds of **passing the exam** is 0.108

  (This mean for every 100 students who fail, we expect about 10.8 students who pass)

- It is almost double the expected number of students who **passing the exam**

  for every 100 students who fails.

Calculating the Confidence Interval for Regression Coefficients

The Confidence Interval for the Regression Coefficient is given by:

$$\begin{matrix} Estimated \\ Regression \\ Coefficient \end{matrix} \pm z_c \times SE\begin{bmatrix} Estimated \\ Regression \\ Coefficient \end{bmatrix}$$

Both the estimated regression coefficient and its standard error can be found in the coefficient table.

```
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 ***
```

Confidence Interval for Intercept(A)

Let's calculate the 95% confidence interval for the **intercept A**.

```
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 ***
```

$$\underbrace{\begin{matrix} Estimated \\ Intercept \end{matrix}}_{} \pm \underbrace{z_c}_{} \times \underbrace{SE\begin{bmatrix} Estimated \\ Intercept \end{bmatrix}}_{}$$

Let's interpret the confidence interval for the **intercept (A)**

The **odds** of a student **passing the exam** is predicted to be e^{A} if the student studies for **0 hours**.

$$95\%\ CI\ for\ A = (-4.798, \quad -0.9983)$$

$$95\%\ CI\ for\ e^{A} =$$

Interpretation:

Confidence Interval for Slope(B)

Let's calculate the 95% confidence interval for the **Slope (B)**

```
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 ***
```

$$\underbrace{\begin{matrix} Estimated \\ Slope \end{matrix}} \pm \underbrace{z_c} \times \underbrace{SE\begin{bmatrix} Estimated \\ Slope \end{bmatrix}}$$

Let's interpret the confidence interval for the **Slope (B)**

For **each additional hour** spent studying for the exam, we predict the **odds** of a student **passing the exam**

- **increases** by a factor of $e^{\boldsymbol{B}}$ (if **B > 0**)
- **decreases** by a factor of $e^{\boldsymbol{B}}$ (if **B < 0**) .

$$95\%\ CI\ for\ \boldsymbol{B} = (0.3088, \quad 1.038)$$

Since every number is **positive**, the **Slope (B)** should be **positive** and the **odds** should **increase** by a factor of $e^{\boldsymbol{B}}$

$$95\%\ CI\ for\ \boldsymbol{e^{B}} =$$

Interpretation: