

## Lesson 27. Multiple Logistic Regression – Part 1

### 1 The multiple logistic regression model

- Binary response variable  $Y$
- Quantitative or categorical explanatory variables  $X_1, \dots, X_k$
- Logit form of the model:

- Probability form of the model:

- The explanatory variables can include transformations or interaction terms, like we saw for multiple linear regression

### 2 Interpreting the model

- The fitted model is

- Plug values of  $X_1, \dots, X_k$  into the fitted model  $\implies$  solve for odds( $\hat{\pi}$ ) =  $\frac{\hat{\pi}}{1 - \hat{\pi}}$  or  $\hat{\pi}$

- The estimated slope  $\hat{\beta}_i$  for explanatory variable  $X_i$  is

- Therefore,  $e^{\hat{\beta}_i}$  is

- In other words:

### 3 Formal inference for multiple logistic regression

Test for single $\beta_i$	z-test (Wald test)
CI for $\beta_i$	$\hat{\beta}_i \pm z_{\alpha/2} SE_{\hat{\beta}_i}$
Test for overall model Compare nested models	LRT test Nested LRT test

#### 3.1 z-test (Wald test) for the slope of a single predictor

- Question: after we account for the effects of all the other predictors, does the predictor of interest  $X_i$  have a significant association with  $Y$ ?
- Formal steps:

1. State the hypotheses:

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_A : \beta_i \neq 0$$

2. Calculate the test statistic:

$$z = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

3. Calculate the  $p$ -value:

- If the conditions for logistic regression hold, then test statistic follows  $N(0,1)$

$$p\text{-value} = 2(1 - P(N(0,1) < |z|))$$

4. State your conclusion, based on the given significance level  $\alpha$

**If we reject  $H_0$  ( $p\text{-value} \leq \alpha$ ):**

We reject  $H_0$  because the  $p$ -value is less than the significance level  $\alpha$ . We see evidence that, after accounting for the other explanatory variables,  $X_i$  is significantly associated with  $Y$ .

**If we fail to reject  $H_0$  ( $p\text{-value} > \alpha$ ):**

We fail to reject  $H_0$  because the  $p$ -value is greater than the significance level  $\alpha$ . We do not see evidence that  $X_i$  is significantly associated with  $Y$  after accounting for the other explanatory variables.

#### 3.2 Confidence intervals for the slope of a single predictor

- The  $100(1 - \alpha)\%$  confidence interval for the slope  $\beta_i$  is

$$(\hat{\beta}_i - z_{\alpha/2} SE_{\hat{\beta}_i}, \hat{\beta}_i + z_{\alpha/2} SE_{\hat{\beta}_i})$$

- The  $100(1 - \alpha)\%$  confidence interval for the odds ratio  $e^{\beta_i}$  is

$$(e^{\hat{\beta}_i - z_{\alpha/2} SE_{\hat{\beta}_i}}, e^{\hat{\beta}_i + z_{\alpha/2} SE_{\hat{\beta}_i}})$$

### 3.3 Likelihood ratio test (LRT) for model utility

- Question: Is the overall model effective?
- Formal steps:

1. State the hypotheses:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad \text{versus} \quad H_A : \text{at least one } \beta_i \neq 0$$

2. Calculate the test statistic:

$$G = \underbrace{-2 \log(L_0)}_{\text{null deviance}} - \underbrace{(-2 \log(L))}_{\text{residual deviance}}$$

3. Calculate the  $p$ -value:

- If the conditions for logistic regression hold, then test statistic follows  $\chi^2(k)$

$$p\text{-value} = 1 - P(\chi^2(k) < G)$$

4. State your conclusion, based on the given significance level  $\alpha$

**If we reject  $H_0$  ( $p\text{-value} \leq \alpha$ ):**

We see significant evidence that the model is effective.

**If we fail to reject  $H_0$  ( $p\text{-value} > \alpha$ ):**

We do not see significant evidence that the model is effective.

### 3.4 Nested likelihood ratio test (LRT) to compare models

- Question: is the full or reduced model better?

$$\text{Full model: } \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \dots + \beta_{k_1} X_{k_1} + \beta_{k_1+1} X_{k_1+1} + \dots + \beta_{k_1+k_2} X_{k_1+k_2}$$

$$\text{Reduced model: } \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \dots + \beta_{k_1} X_{k_1}$$

- Formal steps:

1. State the hypotheses:

$$H_0 : \beta_{k_1+1} = \beta_{k_1+2} = \dots = \beta_{k_1+k_2} = 0 \quad (\text{reduced model is more effective})$$

$$H_A : \text{at least one } \beta_i \neq 0 \quad (i \in \{k_1 + 1, \dots, k_1 + k_2\}) \quad (\text{full model is more effective})$$

2. Calculate the test statistic:

$$G = (\text{residual deviance for reduced model}) - (\text{residual deviance for full model})$$

3. Calculate the  $p$ -value:

- If the conditions for logistic regression hold, then the test statistic follows  $\chi^2(k_2)$

$$p\text{-value} = 1 - P(\chi^2(k_2) < G)$$

4. State your conclusion, based on the given significance level  $\alpha$

**If we reject  $H_0$  ( $p\text{-value} \leq \alpha$ ):**

We see significant evidence that the full model is more effective.

**If we fail to reject  $H_0$  ( $p\text{-value} > \alpha$ ):**

We do not see significant evidence that the full model is more effective.