

# IDS 702: MODULE 2.2

## LOGISTIC REGRESSION WITH ONE PREDICTOR

DR. OLANREWAJU MICHAEL AKANDE

# INTRODUCTION TO LOGISTIC REGRESSION

- Relative risk and odds ratio can be useful, but it would be great to be able to do either or both in more flexible settings, particularly when we have multiple predictors.
- Let's start small: suppose we want to use linear regression to predict binary  $y$  from some predictor  $x$ .
- Recall that the simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

- Also recall that this means the model implies that  $y$  could be any continuous value, when in fact for a binary outcome, it has to be exactly zero or one.
- Therefore, linear regression is not a reasonable model here.
- What distribution(s) do you think would be more ideal for  $y$ ?

# MORE APPROPRIATE DISTRIBUTION FOR $Y$

- Assume for any observation  $i$  that

$$\Pr[y_i = 1|x_i] = \pi_i \text{ and } \Pr[y_i = 0|x_i] = 1 - \pi_i$$

where  $\pi_i$  is some function of  $x_i$ .

- Notice that this is simply a **Bernoulli distribution** or a **Binomial distribution** (with number of trials = 1) where the probability  $\pi_i$  is allowed to be potentially different for each observation  $i$ .
- What then should the function that connects  $\pi_i$  to  $x_i$  look like?
- Some "not so ideal" options could be:

- Linear:**

$$\pi_i = \beta_0 + \beta_1 x_i; \quad \text{But } \pi_i \text{ can be outside } [0, 1]!$$

- Log-linear:**

$$\log(\pi_i) = \beta_0 + \beta_1 x_i \Rightarrow \pi_i = e^{\beta_0 + \beta_1 x_i}; \quad \text{But } \pi_i \text{ can be } > 1!$$

# LOGISTIC REGRESSION MODEL

- From the log-linear function, we can already see a potential solution to the  $\pi_i > 1$  problem: we can divide  $e^{\beta_0 + \beta_1 x_i}$  by a denominator that will always be greater than it.
- Thus, we can use the function

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \Rightarrow \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i.$$

$\log\left(\frac{\pi_i}{1 - \pi_i}\right)$  is called the **logit function**, also written as  $\text{logit}(\pi_i)$ .

Notice that the logit function is essentially the **log-odds**, i.e., log of the odds.

- We can then formally write the **logistic regression model** as

$$\Pr[y_i = 1|x_i] = \pi_i \text{ and } \Pr[y_i = 0|x_i] = 1 - \pi_i; \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i,$$
$$\text{OR } y_i|x_i \sim \text{Bernoulli}(\pi_i); \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i.$$

# SOLVING THE LOGIT EQUATION

- Let's see how to go from  $\text{logit}(\pi_i)$  back to  $\pi_i$ .

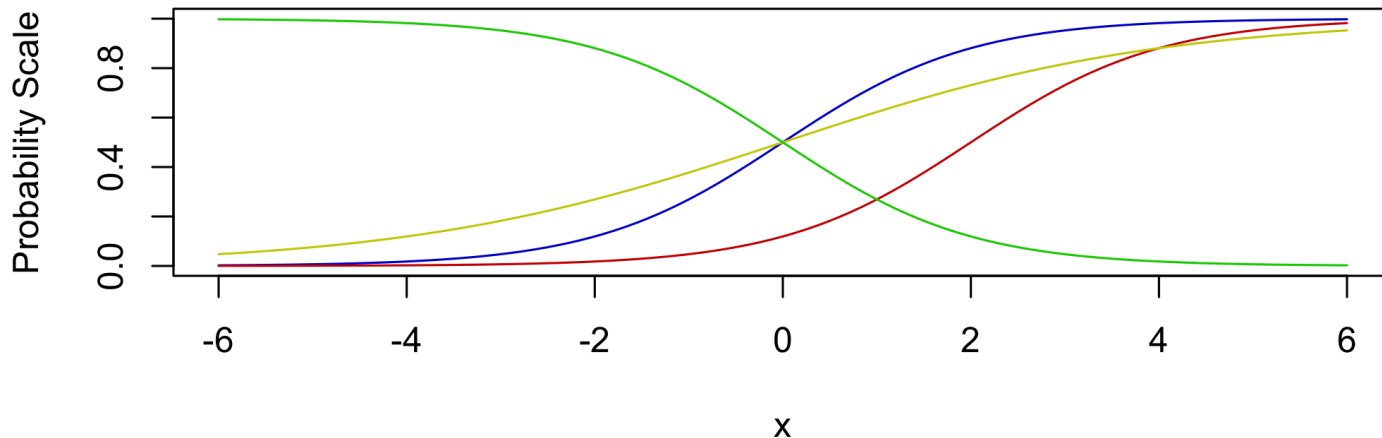
$$\begin{aligned}\text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i \\ \Rightarrow e^{\log\left(\frac{\pi_i}{1 - \pi_i}\right)} &= e^{\beta_0 + \beta_1 x_i} \\ \Rightarrow \frac{\pi_i}{1 - \pi_i} &= e^{\beta_0 + \beta_1 x_i} \\ \Rightarrow \pi_i &= e^{\beta_0 + \beta_1 x_i} (1 - \pi_i) \\ \Rightarrow \pi_i &= e^{\beta_0 + \beta_1 x_i} - \pi_i e^{\beta_0 + \beta_1 x_i} \\ \Rightarrow \pi_i + \pi_i e^{\beta_0 + \beta_1 x_i} &= e^{\beta_0 + \beta_1 x_i} \\ \Rightarrow \pi_i (1 + e^{\beta_0 + \beta_1 x_i}) &= e^{\beta_0 + \beta_1 x_i} \\ \therefore \pi_i &= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\end{aligned}$$

- By the way, another function that works well for linking  $\pi_i$  to  $x_i$  is the **probit function**; the quantile function (or inverse of the cumulative distribution function) associated with the standard normal distribution.
- We will formally explore the **probit regression model** later.

# THE INVERSE LOGIT FUNCTION

What does the **inverse logit function** look like?

```
curve(invlogit((x)),xlim=c(-6,6),ylim=c(0,1),col="blue3",ylab="Probability Scale")
curve(invlogit((-2+x)),col="red3",add=T)
curve(invlogit((0.5*x)),col="yellow3",add=T)
curve(invlogit((-x)),col="green3",add=T)
```



This will be useful for us, when doing EDA, as you will see later. Pay attention to the tails.

# INTERPRETING COEFFICIENTS

- From

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i$$

we can see that "as we increase  $x$  by 1 unit, we increase the log-odds of  $y$  being 1 by  $\beta_1$ ".

- Equivalently, from

$$\frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 x_i} = e^{\beta_0} e^{\beta_1 x_i}$$

we can see that "as we increase  $x$  by 1 unit, we increase the odds for  $y$  by a multiplicative effect of  $e^{\beta_1}$ ".

- With mean-centered  $x$ ,  $\beta_0$  is the log-odds for  $y$  at the mean of  $x$ , and  $e^{\beta_0}$  is the odds for  $y$  at the mean of  $x$ .
- Often also interesting to interpret results by graphing the (predicted) probabilities for values of  $x$ .

# INTERPRETING COEFFICIENTS: CATEGORICAL PREDICTORS

When  $x$  is binary,

- **Odds** of  $y = 1$  for level  $x = 1$ :  $e^{\beta_0} e^{\beta_1(1)} = e^{\beta_0 + \beta_1}$
- **Odds** of  $y = 1$  for level  $x = 0$ :  $e^{\beta_0} e^{\beta_1(0)} = e^{\beta_0}$
- **Odds ratio (OR)**:  $\frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$
- Thus,  $e^{\beta_1}$  has a nice interpretation as the odds ratio when  $x = 1$  versus  $x = 0$ , and  $\beta_1$  is the corresponding log odds ratio.

When  $x$  is categorical with  $K > 2$  levels, the corresponding  $e^{\beta_{1k}}$  is the odds ratio when  $x = k$  versus whichever level is set as the baseline.

It is also easy to calculate relative risk from the results of the logistic model.



# ESTIMATION OF COEFFICIENTS

- Use maximum likelihood estimation.
- Basic idea is to find the values of  $(\beta_0, \beta_1)$  that are most likely to have generated the  $Y$  we see.
- Requires multivariate calculus and numerical methods (Newton Raphson algorithm) for estimation.
- Beyond the scope of this class, so we will not get into it. If interested, take a look at the textbook readings.
- R to the rescue yet again: R does it for us!!!



# INTERVALS AND SIGNIFICANCE TESTS

- As with all coefficients, the standard errors represent chance deviations in the estimated values  $(\hat{\beta}_0, \hat{\beta}_1)$  from the actual values  $(\beta_0, \beta_1)$
- Confidence intervals is usually based on large-sample normal distribution approximations. For example,

- 95 CI for  $\hat{\beta}_1$ :

$$\hat{\beta}_1 \pm 1.96 \times \text{SE}_{\hat{\beta}_1}$$

- 95 CI for  $e^{\hat{\beta}_1}$ :

$$e^{\left[\hat{\beta}_1 \pm 1.96 \times \text{SE}_{\hat{\beta}_1}\right]}$$

- Confidence intervals can also be computed using the profile-likelihood approach. Also beyond the scope of this class.
- Although both methods can yield similar CIs with large sample sizes, the profile-likelihood limits can often have better small-sample properties than the asymptotic approximations. Note that R can compute both.

# WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!