# Logistic Regression

## STA 102: Introduction to Biostatistics

Yue Jiang

November 9, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

# Models for binary outcomes

Suppose we have a binary outcome (e.g., $Y = 1$ if a condition is satisfied and $Y = 0$ if not) and predictors on a variety of scales.

If the predictors are discrete and the binary outcomes are independent, we can use the Bernoulli distribution for individual 0-1 data or the binomial distribution for grouped data that are counts of successes in each group.

# Models for binary outcomes

Let's suppose we want to model $P(Y = 1)$.

One strategy might be to simply fit a linear regression model to the probabilities. E.g., model

$$P(Y = 1)_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i.$$

# Primary biliary cirrhosis

The Mayo Clinic conducted a trial for primary biliary cirrhosis, comparing the drug D-penicillamine vs. placebo. Patients were followed for a specified duration, and their status at the end of follow-up (whether they died) was recorded.

Researchers are interested in predicting whether a patient died based on the following variables:

- ascites: whether the patient had ascites ($1 = $ yes, $0 = $ no)
- bilirubin: serum bilirubin in mg/dL
- stage: histologic stage of disease (ordinal categorical variable with stages 1, 2, 3, and 4)

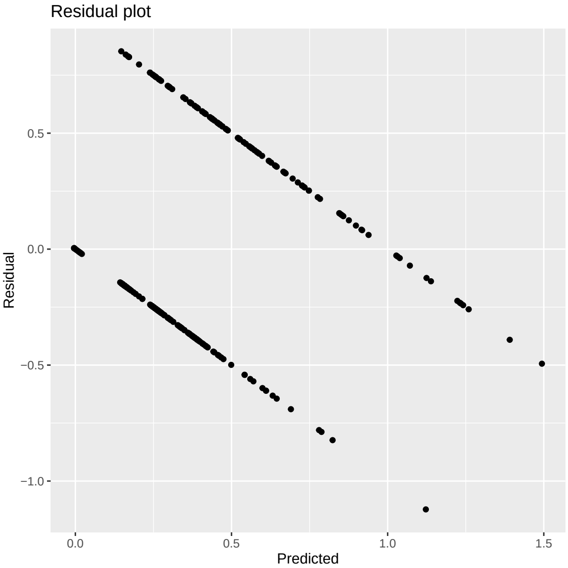
# What can go wrong?

Suppose we fit the following model:

$$P(Y = 1)_i = \beta_0 + \beta_1(\text{ascites})_i + \beta_2(\text{bilirubin})_i + \beta_3(\text{stage} = 2)_i + \beta_4(\text{stage} = 3)_i + \beta_5(\text{stage} = 4)_i + \epsilon_i$$

What can go wrong?

# What can go wrong?



# What can go wrong?

Additionally, as a probability, $p_i$ must be in the interval [0, 1], but there is nothing in the model that enforces this constraint, so that you could be estimating probabilities that are negative or that are greater than 1!

# From probabilities to log-odds

Suppose the probability of an event is $p$

Then the odds that the event occurs is $\frac{p}{1-p}$

Taking the (natural) log of the odds, we have the logit of $p$: the log-odds:

$$logit(p) = \log\left(\frac{p}{1-p}\right).$$

Note that although $p$ is constrained to lie between 0 and 1, the logit of $p$ is unconstrained - it can be anything from $-\infty$ to $\infty$

# Logistic regression model

Let's create a model for the logit of $p$:

$$logit(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

This is a linear model for a transformation of the outcome of interest, and is also equivalent to

$$p_i = \frac{exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}.$$

The expression on the right is called a *logistic function* and cannot yield a value that is negative or a value that is $> 1$. Fitting a model of this form is known as *logistic regression*.

# Logistic regression

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

Negative logits represent probabilities less than one-half, and positive logits represent probabilities above one-half.

# Interpreting parameters in logistic regression

Typically we interpret *functions* of parameters in logistic regression rather than the parameters themselves. For the simple model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1,$$

we note that the probability that $Y = 1$ when $X = 0$ is

$$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}.$$

# Interpreting parameters in logistic regression

Suppose that $X$ is a binary (0/1) variable (e.g., $X = 1$ for males and 0 for non-males). In this case, we interpret $\exp(\beta_1)$ as the odds ratio (OR) of the response for the two possible levels of $X$. For $X$ on other scales, $\exp(\beta_1)$ is interpreted as the odds ratio of the response comparing two values of $X$ one unit apart.

Why? The log odds of response for $X = 1$ is given by $\beta_0 + \beta_1$, and the log odds of response for $X = 0$ is $\beta_0$. So the odds ratio of response comparing $X = 1$ to $X = 0$ is given by $\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$.

In a *multiple logistic regression* model with more than one predictor, this OR is interpreted conditionally on values of other variables (i.e., controlling for them).

# Back to the PBC data

Fitting a logistic regression model, we obtain

| | Est. | SE | p-value |
|---|---|---|---|
| (Intercept) | -3.14 | 1.05 | 0.003 |
| ascites | 2.87 | 1.07 | 0.007 |
| bilirubin | 0.31 | 0.06 | $< 0.001$ |
| stage $= 2$ | 1.25 | 1.10 | 0.253 |
| stage $= 3$ | 1.72 | 1.07 | 0.109 |
| stage $= 4$ | 2.17 | 1.08 | 0.044 |

Remember, this is for the linear effect on the log-odds (the logit). How might we interpret these coefficients as odds ratios?

# Back to the PBC data

Remember, we are interested in the probability that a patient died during follow-up (a "success"). We are predicting the log-odds of this event happening.

- The $\hat{\beta}$ corresponding to ascites was 2.87. Thus, the odds ratio for dying is $\exp(2.87) \approx 17.6$. That is, patients with ascites have 17.6 times the odds of dying compared to patients that do not, holding all other variables constant.
- The $\hat{\beta}$ corresponding to bilirubin was 0.31. Thus, the odds ratio for dying for a patient with 1 additional mg/dL serum bilirubin compared to another is $\exp(0.31) \approx 1.36$, holding all other variables constant.
- The baseline stage was 1. The $\hat{\beta}$ corresponding to stage 3 was 1.72. Thus, patients in stage 3 have approximately 5.58 times the odds of dying compared to patients that do not, holding all other variables constant.

# Predicted probabilities

There is a one-to-one relationship between $p$ and logit($p$). So, if we predict logit($p$), we can "back-transform" to get back to a predicted probability.

For instance, suppose a patient does not have ascites, has a bilirubin level of 5 mg/dL, and is a stage 2 patient.

Their predicted log-odds are

$$-3.14 + 0.31 \times 5 + 1.25 = -0.34$$

Thus, the predicted probability of dying for this individual is

$$\frac{\exp(-0.34)}{1 + \exp(-0.34)} = 0.42.$$

(see slide 11 if you don't know where this expression came from).

# Hypothesis tests in logistic regression

Generally, we wish to know whether the OR $= 1$ or equivalently, whether the logit of $p$ (a $\beta$ coefficient) $= 0$. To test $H_0 : \beta_j = 0$, we can compare the ratio of a parameter estimate to its standard error using the standard normal distribution (reason we use $Z$ instead of $t$ is a bit technical).

# Confidence intervals in logistic regression

Confidence intervals for the effects on the logit scale,

$$\hat{\beta}_j \pm z^\star_{1-\alpha/2} \times \widehat{SE}(\hat{\beta}_j),$$

are typically translated into confidence intervals for ORs by exponentiating the lower and upper confidence limits:

$$\left(\exp\left(\hat{\beta}_j - z^\star_{1-\alpha/2} \times \widehat{SE}(\hat{\beta}_j)\right), \exp\left(\hat{\beta}_j + z^\star_{1-\alpha/2} \times \widehat{SE}(\hat{\beta}_j)\right)\right).$$

Don't worry about the computational details for now, just know that in order to get a confidence interval for ORs, the "back-transformation" must be performed.