Chapter 22: Logistic Regression

Regression Analysis
STA4234

Dr. Seals

* Binary Logistic Regression *

22.2: The Logistic Model

Suppose we now have outcomes that are binary (only two possible responses).

For example, suppose Y_i was "the student passes the class," then:

$$Y_i = \begin{cases} 1 & \text{if student passes} \\ 0 & \text{if student does not pass} \end{cases}$$

Binary variables do not always take on yes/no answers!

e.g., "Do you prefer cats or dogs?"

e.g., "Is the pug fawn or black?"

We model binary outcomes using *logistic regression*.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

* log odds *

where $\pi = P[Y = 1]$ = the probability of the outcome.

How is this different from linear regression?

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Why can't we use OLS estimation?

1. The residuals are not normally distributed.
2. The residuals do not have constant variance. (scatter vert)
3. The predicted values (probabilities) do not always fall between 0 and 1, the only possible values for the probability of success.

→ interval for probability

* No Proof !!! *

22.2: The Logistic Model

* Binary Outcome - Logistic Regression ⇒ Know how to apply/interpret *

Example:

A researcher is interested in how the GRE, college GPA, and prestige of the undergraduate institution affect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

We are modeling graduate school admission as a function of GRE, college GPA, and prestige of the undergraduate institution. Use R to construct this model.

Our model →

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.99 + 0.0026\text{GRE} + \dots$$

22.3: Odds Ratios

Logistic regression uses the *logit* function,

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

We can solve for the probability:

$$\pi = \frac{\exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}}{1 + \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}}$$

In linear regression, we interpret $\hat{\beta}_i$:

> For a 1 [unit of predictor] increase in [predictor name], we expect [outcome] to [increase or decrease] by [$|\hat{\beta}_i|$] [unit of outcome].

In logistic regression, we interpret $e^{\hat{\beta}_i}$ (the odds ratio):

> For a 1 [unit of predictor] increase in [predictor name], we expect the odds of [outcome] to be multiplied by [$e^{\hat{\beta}_i}$].

> For a 1 [unit of predictor] increase in [predictor name], we expect the odds of [outcome] to [increase or decrease] by [$100 \times$ (1–OR)%].

22.3: Odds Ratios

Example:

Convert all $\hat{\beta}_i$ to odds ratios and provide brief interpretations for the graduate school admissions data.

* Don't care about intercept b

1/+ : multiplicative factor

- # >1 is an increase
- # <1 is a decrease

Integers:

- for a 1 point ↑ in GRE, odds are multiplied by the odds ratio (GRE)
- 1 pt. ↑ GPA, odds ↑ 2.23#...

Categorical predictors:

- Ex, rank2 = 0.508..., then 100 - 51 (x 100 & round up)
  = 49% (lower odds in this case)
- rank3 = 100 - 26 = 74% (decrease)

* We are comparing to a reference group, ALWAYS state in interpretation!

22.X_1: Inference in Logistic Regression

Statistical Test for Individual Slopes, Adjusting for Covariates

Hypotheses

usually test against 0

$$H_0: \beta_i = \beta_i^{(0)} \mid \beta_i \geq \beta_i^{(0)} \mid \beta_i \leq \beta_i^{(0)}$$
$$H_1: \beta_i = \beta_i^{(0)} \mid \beta_i < \beta_i^{(0)} \mid \beta_i > \beta_i^{(0)}$$

$\beta_i^{(0)} = 0$, $\beta_i^{(0)} \neq 0$

Test Statistic "Wald's Test Statistic"

$$z = \frac{\hat{\beta}_i - \beta_i^{(0)}}{SE_{\hat{\beta}_i}}$$

Rejection Region

Reject H_0 if $p < \alpha$.

* Note 1: The hypotheses can be written in terms of odds ratios:

$$H_0: OR_i = OR_i^{(0)} \mid OR_i \geq OR_i^{(0)} \mid OR_i \leq OR_i^{(0)}$$
$$H_1: OR_i = OR_i^{(0)} \mid OR_i < OR_i^{(0)} \mid OR_i > OR_i^{(0)}$$

$e^0 = 1$

· usually, $OR_i^{(0)} = 1$.

· * I don't change value = No slope

If dealing with odds ratios, we change the test statistic as follows:

$$z = \frac{\hat{\beta}_i - \ln(OR_i)^{(0)}}{SE_{\hat{\beta}_i}}$$

Note 2: The Z statistic is called a Wald statistic.

Sometimes, this is given as a χ^2 with 1 df,

$$\chi^2 = Z^2$$

22.X_1: Inference in Logistic Regression

Example:

Determine which, if any, are significant predictors of graduate school admission. Test at the $\alpha = 0.05$ level.

22.X_1: Inference in Logistic Regression

Confidence Interval for β_i

$$\hat{\beta}_i \pm z_{1-\alpha/2}\text{SE}_{\hat{\beta}_i}$$

$\hat{\beta}_i$ (95% CI)

Confidence Interval for OR_i

$$\left(e^{\hat{\beta}_i - z_{1-\alpha/2}\text{SE}_{\hat{\beta}_i}},\ e^{\hat{\beta}_i + z_{1-\alpha/2}\text{SE}_{\hat{\beta}_i}}\right)$$

OR (95% for OR)

22.X_1: Inference in Logistic Regression

Example:

Find the 95% confidence intervals for the odds ratios.

"multi-nomial"

22.X_2: Nominal Logistic Regression

When we have a response variable with c categories, we can create multicategory logistic models simultaneously.

(review categorical predictors in models)

We will choose a reference category and create $c-1$ models.

The baseline-category logit model (or the multinomial logit model):

$$\ln\left(\frac{\pi_j}{\pi_c}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

π_j: probability of being in category j; π_c: " of being in cat c; $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$: model parameters

where $j = 1, \ldots, c-1$.

• exp. the parameters

$e^{\hat{\beta}_i}$

← ORs are interpreted similarly to what we've seen before:

(=control)

Continuous predictor:

> For a 1 [unit of predictor] increase in [predictor name], the odds in favor of [response category j] over [reference category of the response] are multiplied by [the odds ratio].

> For a 1 [unit of predictor] increase in [predictor name], we expect the odds of [response category j] to [increase or decrease] by [$100 \times$ (1–OR)%] as compared to the [reference category of the response].

Categorical predictor:

> As compared to [reference category of the response], the odds of [predictor group of interest] in favor of [response category j] over [reference category of the response] are multiplied by [the odds ratio].

> As compared to [the predictor reference group], we expect the odds of [response category j] over [reference category of the response] to [increase or decrease] by [$100 \times$ (1–OR)%] for [the predictor group of interest].

22.X_2: Nominal Logistic Regression

Example:

Let us examine foods that alligators in the wild choose to eat. For 59 alligators sampled in Lake George, Florida, the alligator data shows the primary food type (in volume) found in the alligator's stomach. Primary food type has three categories: Fish, Invertebrate, and Other. The invertebrates were primarily apple snails, aquatic insects, and crayfish. The "other" category included amphibian, mammal, plant material, stones or other debris, and reptiles (primarily turtles, although one stomach contained the tags of 23 baby alligators that had been released in the lake during the previous year!). The data also shows the alligator length, which varied between 1.24 and 3.89 meters. Use R to construct a model that models primary food choice as a function of alligator length.

Use R to model primary food choice as a function of alligator length in meters.

multinom (food ~ length, data) * → no tests output

- we can ignore the output last interested if converged or not?
  - we were able to estimate parameters

* C-1 models

Is length a significant predictor of food choice?

(general f-test)
anova (m2, m1)
- remove predictor from model
+ use 1 as predictor to use as intercept
↑ reduced model

* Report final model:

- $\ln\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) = 5.69 - 2.47\,\text{length}$ (Invertebrates)
- $\ln\left(\frac{\hat{\pi}_F}{\hat{\pi}_0}\right) = 1.62 - 0.11\,\text{length}$ (Fish model)

↑ full models

22.X_2: Nominal Logistic Regression

Example:

Find and interpret the odds ratios for length of alligator.

use · coefficients(model) → $\hat{\beta}_i$

· exp(coefficients(model)) → $\hat{OR}_i$

↳ interested in length not intercept

* for a 1 meter increase in alligator length, the odds of choosing I over O food are [multiplied by 0.08 (length of category) / decreased by 92%. 100 - 8% = 92%]

Construct the 95% confidence interval for the odds ratios.

* Same thing for gators that eat fish *

* confint(model) for OR & $\hat{\beta}$:

exp(confint(model))

22.X_3: Ordinal Logistic Regression

Suppose our response variable has c ordered categories (e.g., classification of student: freshman, sophomore, junior, senior).

We again will create $c - 1$ models.

- The $\hat{\beta}_i$ will be the same across the models.
- The $\hat{\beta}_0$ will change for each category.

This is called the cumulative logit model,

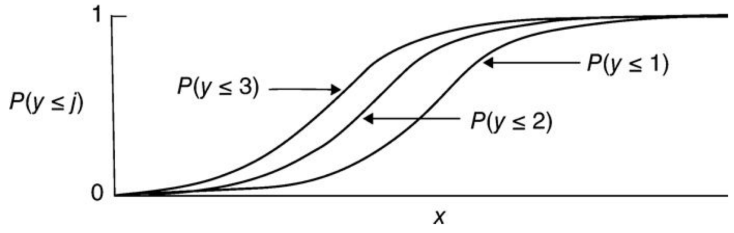
$$\text{logit}\left(P[Y \leq j]\right) = \hat{\beta}_{0j} + \hat{\beta}_{1j}X_1 + \ldots + \hat{\beta}_{kj}X_k$$

Note that the logit function is as follows:

$$\text{logit}\left(P[Y \leq j]\right) = \log\left(\frac{P[Y \leq j]}{1 - P[Y \leq j]}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots \pi_c}\right)$$

As noted above, the intercept depends on j.

- This means that curves will have the same shape $\forall\ j$.
- We are just shifting the model along the x-axis, depending on the response category.
- Graphically,



Also note that this assumes *proportional odds*.

- For each predictor included in the model, the slopes across two outcomes response levels are the same, regardless of which two responses we consider.

22.X_3: Ordinal Logistic Regression

Odds ratios are interpreted a little bit differently due to the model being cumulative.

For a one unit increase in x, the odds in favor of the response category j or lower are multiplied by $e^{\hat{\beta}_i}$.

> Again, the change in odds does not depend on the category of the response.

Note that we can convert to probabilities:

$$P[Y \leq j|X_1, X_2, ..., X_k] = \frac{\exp\{\beta_{0j} + \beta_1 X_1 + \ldots + \beta_k X_k\}}{1 + \exp\{\beta_{0j} + \beta_1 X_1 + \ldots + \beta_k X_k\}}$$

22.X_3: Ordinal Logistic Regression

Example:

Consider the following data from a General Social Survey, relating political ideology to political party affiliation. Political ideology has a five-point ordinal scale, ranging from very liberal ($Y = 1$) to very conservative ($Y = 5$). Let X be an indicator variable for political party, with $X = 1$ for Democrats and $X = 0$ for Republicans.

		Political Ideology				
Sex	Party	V. Lib.	Lib.	Mod.	Cons.	V. Cons.
F	Dem.	44	47	118	23	32
	Rep.	18	28	86	39	48
M	Dem.	36	34	53	18	23
	Rep.	12	18	62	45	51

Use R to construct an ordinal logistic regression model that models political ideology as a function of political party and sex. Are either political party or sex significant predictors of political ideology?

22.X_3: Ordinal Logistic Regression

Example:

Find and interpret the odds ratios for political party and sex.

Construct the 95% confidence interval for the odds ratios.