Logistic Regression

1 Introduction

Up to this point, you have been learning about linear regression, which is used when we have one quantitative response variable and at least one predictor. When we have a binary response variable and at least one predictor, we use logistic regression.

Common ways of summarizing binary variables include using probabilities and odds. For example, you may want to estimate the probability of college students who haven driven while drunk based on characteristics such as how often they party and how often they drink alcohol. You may consider using a linear regression model, with the probability of driving while drunk as the response variable. However, a linear regression model may end up having estimated probabilities that are less than 0 or greater than 1. A logistic regression model is set up to guarantee the estimated probability is always between 0 and 1.

Typical questions that a logistic regression model can answer include how does partying more often increase the odds of driving while drunk? Do characteristics such as gender help us better predict the odds of driving while drunk when we already know how often the student parties? How confident are we of our estimated odds?

In this module, you will learn about the logistic regression model. Similar to what you learned when studying the linear regression model, you will learn how to interpret the coefficients of a logistic regression model, perform various inferential procedures to answer various questions of interests, and learn how to assess the accuracy of the model. As you study the logistic regression model, it will be helpful for you to compare and contrast what you are learning with what you learned about the linear regression model.

2 Logistic Regression

We will introduce the notation of terms commonly used in logistic regression:

- π: **probability** of “success” (belonging in class coded 1 when using indicator variable to denote the binary response variable).
- $1 - \pi$: probability of “failure” (belonging in class coded 0 when using indicator variable to denote the binary response variable).
- $\frac{\pi}{1-\pi}$: **odds** of “success”.
- $\log\left(\frac{\pi}{1-\pi}\right)$: **log-odds** of “success”.

Using the driving while drunk example from the introduction section, the response variable is whether the student has driven drunk, which is binary with levels “yes” or “no”. We could use indicator variable to denote this binary response, with 1 denoting “yes”, and 0 denoting “no”. In this example:

- π denotes the probability that a student has driven while drunk,
- $1 - \pi$ denotes the probability that a student has not driven while drunk,
- $\frac{\pi}{1-\pi}$ denotes the odds that a student has driven while drunk,
- $\log\left(\frac{\pi}{1-\pi}\right)$ denotes the log odds that a student has driven while drunk.

Notice in these definitions, probability and odds are not exactly the same. If the probability is small, then probability and odds are approximately the same.

Because the response variable is binary, it can no longer be modeled using a normal distribution (which is assumed in linear regression). We will now assume the response variable follows a **Bernoulli** distribution. Another assumption we make is that observations are independent of each other.

- A Bernoulli variable has a probability distribution $P(y_i = 1) = \pi_i$ and $P(y_i = 0) = 1 - \pi_i$.
- The expectation is $E(y_i) = \pi_i$.
- The variance is $Var(y_i) = \pi_i(1 - \pi_i)$.

It may be tempting to try to fit a linear regression model in this framework, i.e.

$$E(y_i) = \pi_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

However, this formulation does not work. The estimated value for $\hat{\pi_i}$ has to be between 0 and 1. Nothing in this formulation ensures this.

2.1 The logistic regression equation

Instead, the logistic regression equation is written as

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k = \boldsymbol{X\beta}, \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ denote the design matrix and vector of parameters respectively. The formulation in (1) has the following characteristics:

- The estimated log odds, $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$, will be any real number.
- The estimated probability, $\hat{\pi}$, will be between 0 and 1.
- The log odds is expressed as a linear combination of the predictors.
- The log odds is a transformed version of the mean of the response.

- This transformation $\log\left(\frac{\pi}{1-\pi}\right)$ is called a **logit link**, and is denoted as $logit(\pi)$.

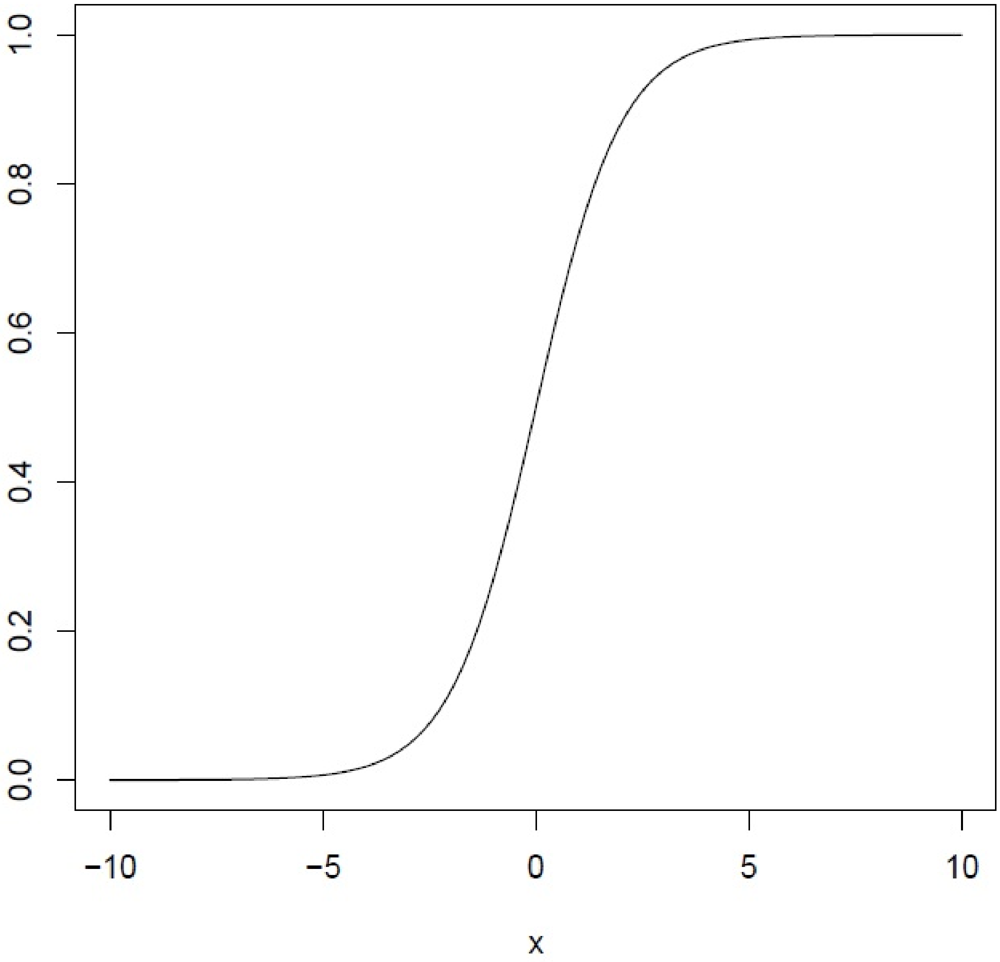
Two algebraically equivalent expressions of the logistic regression equation (1) are

$$\text{Odds: } \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k} = \exp(\boldsymbol{X\beta}) \tag{2}$$

and

$$\text{Probability: } \pi = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} = \frac{\exp(\boldsymbol{X\beta})}{1+\exp(\boldsymbol{X\beta})}. \tag{3}$$

Suppose we have just one predictor, x, and we wish to plot probability against x, we will have (assuming β_1 is positive)



So we see that with a positive coefficient for the predictor, the probability increases as the predictor increases. However, the increase is not linear. It is the log odds that increases linearly with the predictor, not the probability.

2.1.1 Thought question

How are probability and odds related? In other words, given the odds, how can we quickly calculate the probability?

3 Coefficient Estimation in Logistic Regression

Recall that we used the method of least squares to estimate the coefficients in a linear regression model. So a reasonable thought would be to apply a similar idea in a logistic regression framework, i.e. minimize

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}})^2$$

with respect to $\hat{\boldsymbol{\beta}}$, and using a rule that the fitted response $\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$ is rounded to either 0 or 1.

From a purely predictive standpoint, this idea could work. However, statistical inference (intervals, hypothesis tests) is not reliable as the distribution of the response variable is now Bernoulli and not normal. Although the linear regression model is robust to the normality assumption, we still need the distribution to be continuous, not discrete.

For logistic regression, the coefficients are estimated using a different method, called the method of **maximum likelihood**.

3.1 Maximum likelihood estimation for logistic regression

This subsection will give a very brief overview of maximum likelihood estimation in logistic regression.

The motivation behind **maximum likelihood estimation** is that model parameters are estimated by maximizing the likelihood that the process described by the model produced the observed data.

Since we assume the response variable follows a Bernoulli distribution, the probability distribution of each observation is

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{4}$$

where y_i is either 0 or 1. Since we assume the observations are independent, the likelihood function is

$$L(y_1, \cdots, y_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{5}$$

We want to maximize the likelihood function (5) with respect to $\boldsymbol{\beta}$. Since maximizing a function is the same as maximizing the log of a function, we can instead maximize the log-likelihood function

$$\log L(y_1, \cdots, y_n, \boldsymbol{\beta}). \tag{6}$$

It turns out that for logistic regression, it is computationally more efficient to maximize the log-likelihood function (6) instead of the likelihood function (5).

Using $\pi_i = \frac{\exp(\boldsymbol{x_i'\beta})}{1+\exp(\boldsymbol{x_i'\beta})}$ from (3) and after some algebra on the log-likelihood function (6), we maximize the following with respect to $\boldsymbol{\beta}$

$$\log L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{x_i'\beta} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x_i'\beta})] \tag{7}$$

There is no closed-form solution to maximizing (7). A numerical method called iteratively reweighted least squares is used. The details of which are beyond the scope of this class.

Please view the video for a more thorough explanation on maximum likelihood estimation.

4 Interpreting Coefficients in Logistic Regression

There are a few equivalent interpretations of the coefficient of a predictor. For β_1, they are:

- For a one-unit increase in the predictor x_1, the **log odds changes by** β_1, while holding other predictors constant.
- For a one-unit increase in the predictor x_1, the **odds are multiplied by a factor of** $\exp(\beta_1)$, while holding other predictors constant.

Let us go back to the driving while drunk example. The response variable is whether the student has driven while drunk, with 1 denoting "yes" and 0 denoting "no". Let us consider two predictors: x_1: `PartyNum` which denotes the number of days the student parties in a month, on average, and a categorical predictor `Gender`. The estimated coefficients are shown below:

```
result2<-glm(DrivDrnk~PartyNum+Gender, family=binomial, data=train)
result2
```

```
##
## Call:  glm(formula = DrivDrnk ~ PartyNum + Gender, family = binomial,
##     data = train)
##
## Coefficients:
## (Intercept)     PartyNum   Gendermale
##     -1.2433       0.1501       0.4136
##
## Degrees of Freedom: 121 Total (i.e. Null);  119 Residual
##   (2 observations deleted due to missingness)
## Null Deviance:       169.1
## Residual Deviance: 151.9     AIC: 157.9
```

So the estimated logistic regression equation is

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.2433 + 0.1501x_1 + 0.4136I_1,$$

where I_1 is 1 for male students and 0 for female students.

The estimated coefficient for x_1 is $\hat{\beta}_1 = 0.1501$, and can be interpreted as

- The estimated log odds of having driven while drunk for college students increases by 0.1501 for each additional day of partying, when controlling for gender.
- The estimated odds of having driven while drunk for college students is multiplied by $\exp(0.1501) = 1.1620$ for each additional day of partying, when controlling for gender.

The estimated coefficient for I_1 is $\hat{\beta}_2 = 0.4136$, and can be interpreted as

- The estimated log odds of having driven while drunk for college students is 0.4136 higher for males than females, when controlling for the number of days they party.
- The estimated odds of having driven while drunk for male college students is $\exp(0.4136) = 1.5123$ times the odds for female college students, when controlling for the number of days they party.

Please view the video for a more thorough explanation on interpreting the coefficients in logistic regression.

5 Inference in Logistic Regression

There are a number of hypothesis tests that we can conduct in logistic regression. These tests have analogous versions with linear regression.

	Logistic	Linear
Drop Single Term	Wald (Z) test	t test
Is model useful?	Likelihood ratio test	ANOVA F test
Full vs reduced models	Likelihood ratio test	General linear F test

5.1 Wald test

We can assess whether to drop a single term, with a **Wald test**, which is basically a Z test. We can test $H_0 : \beta_j = 0$ with the Z statistic

$$Z = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}, \tag{8}$$

which is compared with a standard normal distribution. A standard normal distribution is denoted by $N(0, 1)$, which is read as a normal distribution with mean 0 and variance 1. Let us look at the Wald tests in our earlier drink driving example:

```
summary(result2)
```

```
##
## Call:
## glm(formula = DrivDrnk ~ PartyNum + Gender, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1357  -1.0211  -0.1941   1.0701   1.7302
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.24333    0.38795  -3.205  0.00135 **
## PartyNum     0.15012    0.04216   3.560  0.00037 ***
## Gendermale   0.41363    0.40466   1.022  0.30670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 169.13  on 121  degrees of freedom
## Residual deviance: 151.93  on 119  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 157.93
##
## Number of Fisher Scoring iterations: 3
```

To assess the coefficient for `PartyNum`, we have

- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- Test statistic $Z = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{0.15012}{0.04216} = 3.560$.
- P-value of 0.00037 (or using `2*(1-pnorm(abs(3.560)))` in R).
- Reject the null. Do not drop `PartyNum` from the logistic regression model.

To assess the coefficient for `Gender`, we have a large p-value, which means we can drop `Gender` from the model and leave `PartyNum` in.

5.1.1 Pratice

Verify the Z statistic and p-value when testing $H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$ in this example.

5.2 Confidence intervals for coefficients

A $(1 - \alpha) \times 100\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm Z_{1-\alpha/2} se(\hat{\beta}_j). \tag{9}$$

Going back to our drink driving example, the 95% CI for β_1 is:

$$\begin{aligned} \hat{\beta}_1 &\pm Z_{0.975} se(\hat{\beta}_1) \\ = 0.15012 &\pm 1.96 \times 0.04216 \\ = (0.0675 &, \quad 0.2328). \end{aligned}$$

Note that $Z_{0.975}$ is found using `qnorm(0.975).` Based on this confidence interval, we can say that:

- We are 95% confident the true β_1 is between 0.0675 and 0.2328.
- CI excludes 0, so coefficient is significant. Do not drop term from model.
- Consistent conclusion with 2-sided hypothesis test at $\alpha = 0.05$.
- The log odds of having driven while drunk for college students increases between 0.0675 and 0.2328 for each additional day of partying, when controlling for gender.
- The odds of having driven while drunk for college students is multiplied by a factor between exp(0.0675) = 1.0698 and exp(0.2328) = 1.2621 for each additional day of partying, when controlling for gender.

5.3 Likelihood ratio tests in logistic regression

Likelihood ratio tests (LRTs) allow us to compare between a full and reduced model, denoted by F and R respectively. The test statistic measures the difference in deviances of the models, i.e.

$$\Delta G^2 = D(R) - D(F), \tag{10}$$

where $D(R)$ and $D(F)$ denote the deviance of the reduced model and deviance of the full model respectively. The deviance of a model is analogous to the SS_{res} of a linear regression model.

The test statistic ΔG^2 is then compared with a chi-squared distribution, denoted by χ^2_{df}, where df denotes the number of parameters you are dropping to get the reduced model.

The deviance of a model is labeled as **residual deviance** in R. The **null deviance** in R is the deviance of an intercept-only model. Going back to our drink driving example, with predictors `PartyNum` and `Gender`, the residual deviance is 151.93, and the null deviance is 169.13. So we can compare our model with an intercept-only model.

- $H_0 : \beta_1 = \beta_2 = 0, H_a :$ at least one coefficient in null is nonzero.
- $\Delta G^2 = D(R) - D(F) = 169.13 - 151.93 = 17.2$.
- P-value is `1-pchisq(17.2,2)` which is close to 0.
- Critical value is `qchisq(0.95,2)` which is 5.9915.
- We reject the null hypothesis and support our two predictor over the intercept-only model.