Logistic Regression

Adam J Sullivan, PhD

02/26/2018

Simple Logistic Regression

- We will begin with an example of simple logistic regression with a sample of the Western Collaborative Group Study.
- This study began in 1960 with 3154 men ages 39-59, who were employed in one of 11 California based companies.
- They were followed until 1969 during this time, 257 of these men developed coronary heart disease (CHD).
- You can read this data in with the code below.

Reading in the Data

```
library(haven)
wcgs <- read_dta("wcgs2.dta")
wcgs <- wcgs[, -16]
```

The Variables

NAME	DESCRIPTION
id	Subject identification number
age	Age in years
height	Height in inches
weight	Weight in lbs.
sbp	Systolic blood pressure in mm
dbp	Diastolic blood pressure in mm Hg

The Variables

NAME	DESCRIPTION
chol	Fasting serum cholesterol in mm
behpat	Behavior
	1 = A1
	2 = A2
	3 = B3
	4 = B4

The Variables

NAME	DESCRIPTION
ncigs	Cigarettes per day
dibpat	Behavior
	1 = type A
	2 = type B
chd69	Coronary heart disease
	1 = Yes
	0 = no

The Variables

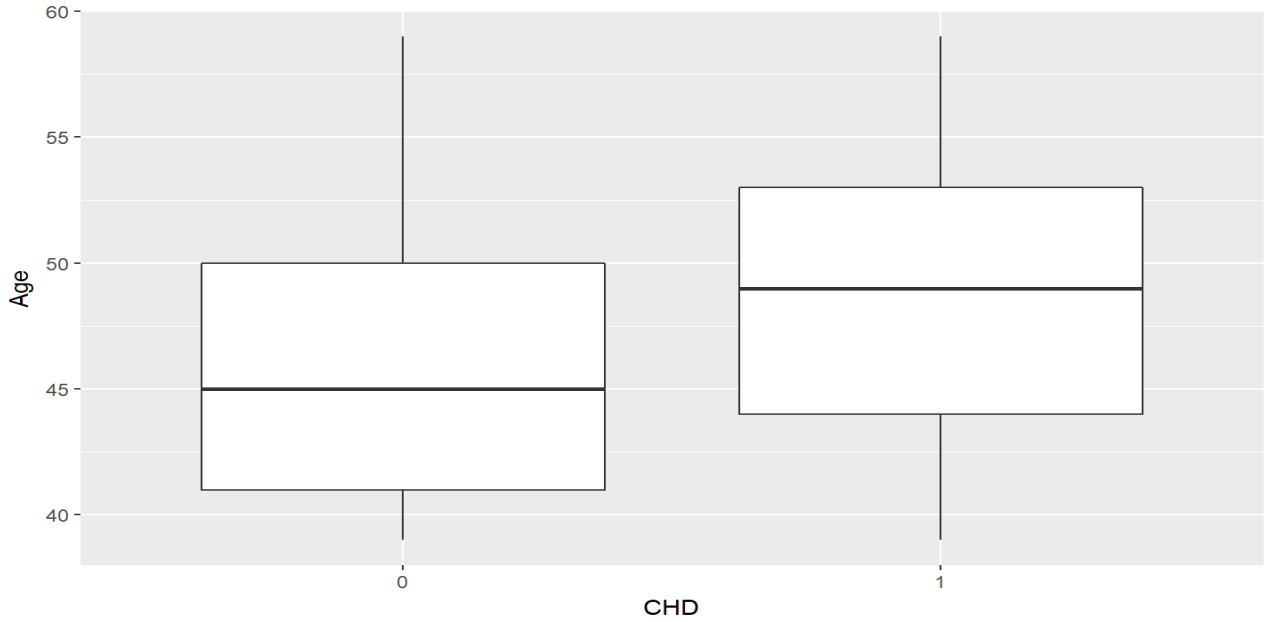
NAME	DESCRIPTION
typechd	Type of CHD
	1 = myocardial infarction or death
	2 = silent myocardial infarction
	3 = angina perctoris
time169	Time of CHD event or end of follow-up

The Variables

NAME	DESCRIPTION
arcus	Arcus senilis
	0 = absent
	1 = present
bmi	Body Mass Index

Continuous Covariate

- We will first consider the relationship between age and CHD.
- We could first consider boxplots



What do we see?

- This displays that the group with CHD appears older than the group without a CHD.
- To check if there is an increased probability of CHD due to being older we will fit the following logistic regression.

$$logit(p_{CHD}) = \hat{\beta}_0 + \hat{\beta}_1 Age$$

```
library(broom)

fit.cont <- glm(chd69 ~ age, data = wcgs, family = binomial(link = "logit"))
tidy(fit.cont, conf.int = TRUE)[, -c(3:4)]
```

```
##          term estimate  p.value conf.low conf.high
## 1 (Intercept)  -5.9395 3.00e-27  -7.0245   -4.8696
## 2         age   0.0744 4.56e-11   0.0523    0.0966
```

What can we conclude?

- From here he can reject the null hypothesis and conclude that $\beta \neq 0$ and further more show that the probability of having CHD increases with age.

Dichotomous Covariate

- We could then consider the relationship between CHD and Arcus Senilis:

CHD69	ARCUS	N
0	0	2058
0	1	839
1	0	153
1	1	102
1	NA	2

An Easier Format

	ABSENT	PRESENT
no	2058	839
yes	153	102

Analysis from Epidemiology Class

- We can see from this table that what we have here is a typical 2×2 table that many of you have seen.
- In epidemiology you learned how to work with these and probably learned that you can calculate the odds ratio as shown below:

$$\frac{2058 \cdot 102}{153 \cdot 839} = 1.63528$$

What do we have then?

We can then find the following:

- Pearson $\chi^2 = 13.64$, p-value=0.000221
- 95% Woolf CI (1.257, 2.127)

Why not Logistic Regression?

- We can answer this same type of question with logistic regression:

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

- This is nothing more than the log odds.
- Or we can list this as:

$$p_x = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

What do we have then?

This leads to

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 \Rightarrow p_1 = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$$

$$\log\left(\frac{p_0}{1-p_0}\right) = \beta_0 \Rightarrow p_0 = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

What does this mean?

- We then can find a way to estimate all probabilities associated with this problem.
- We can find the probability that someone with Arcus Senilis present has the disease or does not have the disease.
- We can also find the same for those with the absence of Arcus Senilis.

Saturated Model

- We refer to this as a , where we can define the probabilities of all relationships.
- This works in our models because we only need to estimate 2 probabilities and we have 2 coefficients so we can estimate everything:

The math

$$
\begin{aligned}
\beta_0 &= \log\left(\frac{p_0}{1-p_0}\right) \\
\beta_1 &= \log\left(\frac{p_1}{1-p_1}\right) - \beta_0 \\
&= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) \\
&= \log\left(\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_0}{1-p_0}\right)}\right) \\
&= \log\left(\frac{\text{Odds of CHD in Arcus Present Group}}{\text{Odds of CHD in Arcus Absent Group}}\right) \\
&= \log(\text{Odds Ratio})
\end{aligned}
$$

What does this mean?

-This shows us that β_1 is the log odds ratio comparing those with Arcus Senilis vs those without it.

$$\beta_1 = \log(\text{Odds Ratio})$$
$$\text{Odds Ratio} = \exp(\beta_1)$$

What about our Estimates?

- Thus with our data we can estimate:

$$\widehat{\log(OR)} = \hat{\beta}_1 \text{ or } \widehat{OR} = \exp\left(\hat{\beta}_1\right)$$

- We can fit this model now using out logistic regression model:

Lets Fit this!

```
fit.bin <- glm(chd69 ~ arcus, data = wcgs, family = binomial(link = "logit"))
tidy(fit.bin, exponentiate = TRUE, conf.int = TRUE)[, -c(3:4)]
```

```
##          term estimate   p.value conf.low conf.high
## 1 (Intercept)   0.0743 3.25e-211   0.0628    0.0873
## 2       arcus   1.6353  2.48e-04   1.2543    2.1241
```

What do we Notice?

- We can see that we have that $\hat{\beta}_1 = 0.492$.
- This is the log odds so if we exponentiate that we have that the Odds ratio is, 1.635.
- Then we can also exponentiate both the lower and upper elements of the confidence intervals to get a 95% confidence interval for the Odds Ratio of (1.254, 2.124).
- Aside from rounding differences this is the same as what you would have found before.

Interpretation of Coefficients

- Let's consider a 1 unit change in x:

$$p_x = \Pr(Y = 1|covariate x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$p_{x+1} = \Pr(Y = 1|covariate x + 1) = \frac{\exp(\beta_0 + \beta_1 (x+1))}{1 + \exp(\beta_0 + \beta_1 (x+1))}$$

Interpretation of Coefficients

- Then the Odds Ratio associated with a 1 unit increase in x is:

$$\begin{aligned}\text{OR} &= \frac{\frac{p_{x+1}}{1-p_{x+1}}}{\frac{p_x}{1-p_x}} \\ &= \exp[\beta_0 + beta_1(x+1) - \beta_0 - \beta_1 x] \\ &= e^{\beta_1}\end{aligned}$$

Interpretation of Coefficients

- If we want to consider a unit increase of 2 in x we would have:

$$\text{OR} = \exp(\beta_1 \cdot 2) = e^{2\beta_1} = e^{\beta_1} e^{\beta_1}$$

Continuous Covariate

- This means given our continuous covariate, age, we have the following interpretation:
    - If we have 2 people who differ in age by one year the older person would have on average an increased log odds of 0.074.

$$e^{\beta_1}$$

Continuos Covariate

- If there are 2 people who differ in age by one year the older person would on average have an odds of CHD 1.077 times that of the younger.
- On average a person has an odds of CHD 0.077 higher than someone a year younger.
- On average a person has an odds of CHD 7.726% higher than someone a year younger.

Binary Covariate

- Given our binary case we have the following interpretation.
  - On average a person with Arcus Senilis present has an odds of CHD 1.635 times that of the odds of CHD for another person with absence of Arcus Senilis.
  - On average a person with Arcus Senilis present has an odds of CHD 63.528% higher than the odds of CHD for another person with absence of Arcus Senilis.

Final Interpretation Notes

- We can see that there are multiple ways to interpret these just as in linear regression.
- so that it can be interpreted in more meaningful units.

Multiple Logistic Regression

When we work with clinical and epidemiological data we tend to focus on multiple predictors. For example:

- A few categorical predictors
  - Contingency Tables
  - Stratified Contingency Tables
- Continuous or many predictors
  - Logistic Regression

Our Model

- Our model is similar to before:

$$logit(\Pr(Y = 1|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- This implies that

$$\Pr(Y = 1|X_1, \ldots, X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

Interpeting the Model

- With multiple logistic regression we have the following interpretations:
  - β_0 which is the log odds that $Y = 1$ among those with $\beta_1 = \cdots = \beta_p = 0$.
  - β_j is the log odds ratio
    - This characterizes the association of X_j on the odds of $Y = 1$.
    - Conditional on all other covariates.

An Example

- We will continue exploring the WCGS data.
- This time we will use a multiple logistic regression model with the following covariates:
  - Age
  - Arcus Senilis
  - Cholesterol
  - Systolic Blood Pressure

Fitting the Model

- We fit this model using R in the same fashion as the simple logistic model.

```
fit.mult <- glm(chd69 ~ age + arcus + chol + sbp, data = wcgs,
    family = binomial(link = "logit"))
tidy(fit.mult, conf.int = TRUE, exponentiate = TRUE)[, -c(3:4)]
```

Fitting the Model

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	0.00	0.000	0.000	0.00
age	1.06	0.000	1.034	1.08
arcus	1.28	0.079	0.969	1.69
chol	1.01	0.000	1.008	1.01
sbp	1.02	0.000	1.014	1.03

What can we do?

- Calculate the log odds of CHD for a 60 year old with 253 mg/dL of total cholesterol, 136 mmHg SBP and the presence of Arcus Senilis.
- We achieve this solution by taking our regression equation and inserting the values from the problem.

```
log.odd <- fit.mult$coefficients[1] + fit.mult$coefficients[2] *
    60 + fit.mult$coefficients[3] * 1 + fit.mult$coefficients[4] *
    253 + fit.mult$coefficients[5] * 136
log.odd
```

```
## (Intercept)
##       -1.26
```

What can we do?

- Calculate the probability of CHD for a 60 year old with 253 mg/dL of total cholesterol, 136 mmHg SBP and the presence of Arcus Senilis.

```
exp(log.odd)/(1 + exp(log.odd))
```

```
## (Intercept)
##       0.222
```