

# Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

## Chapter 20, part 1: Logistic Regression Models

Jinko Graham

2018-11-25

# Introduction to Logistic Regression

- ▶ In logistic regression we study the effect of explanatory variables on the odds of a binary outcome.
- ▶ This is a generalization of the analyses of odds ratios we have studied before.
- ▶ Think of the binary outcome  $Y$  as disease status (0=non-disease; 1=disease).
- ▶ The explanatory variables  $X_1, \dots, X_q$  could be categorical (e.g., exposures), or quantitative variables.

## Example Data

- ▶ In 223 low-birthweight infants sampled from the neonatal ICU of a large hospital, 76 were diagnosed with bronchopulmonary dysplasia (BPD;  $Y = 1$ ) and 147 were not ( $Y = 0$ ).
- ▶ Birth weight ( $X_1$ ) is believed to influence the risk of BPD. Grouping the birth weights into 3 categories, we obtain:

birthweight(in g)	BPD	no BPD	odds BPD	log-odds* BPD
0-950	49	19	$49/19 = 2.58$	0.95
951-1350	18	62	$18/62 = 0.29$	-1.24
1351-1750	9	66	$9/66 = 0.14$	-1.99

\* Use the natural logarithm.

- ▶ Consider the ratio of the odds at two values of  $X_1$ .
- ▶ To get the log of this odds ratio (the log-OR), we take the difference between the two log-odds.
- ▶ E.G. Consider the odds of BPD in babies with birthweight  $< 950g$  relative to babies with birthweight between 1351-1750g. The log-OR is  $0.95 - (-1.99) = 0.95 + 1.99 = 2.94$ .

# The Logistic Regression Model

- ▶ We may model the log-odds of  $Y = 1$  (e.g. BPD) as a function of  $X_1$  (e.g. birthweight):

$$\log \left[ \frac{p}{1-p} \right] = \alpha + \beta_1 X_1, \quad \text{where}$$

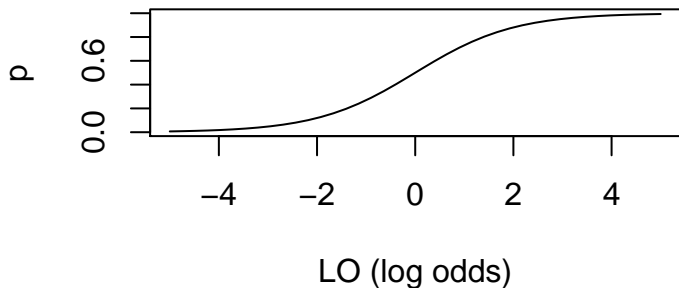
- ▶ log is the *natural logarithm* and
  - ▶  $p$  is the probability of  $Y = 1$  given  $X_1$ .
- ▶ Let  $LO = \alpha + \beta_1 X_1$  be the linear predictor for the log-odds.
  - ▶ The logistic-regression parameters are  $\alpha$  and  $\beta_1$ .
- ▶ It can be shown that

$$p = \frac{e^{LO}}{1 + e^{LO}}.$$

- ▶ i.e., the probability  $p$  is the *logistic function* of the log-odds.

## Graph of the Logistic Function

- ▶  $y$ -axis is  $p$ ;  $x$ -axis is  $LO$ ; curve is function  $p = \frac{e^{LO}}{1+e^{LO}}$ .



- ▶ On the  $y$ -axis,  $p$  is constrained to be between 0 and 1 and, on the  $x$ -axis,  $LO$  is unconstrained.
  - ▶ As  $LO$  gets large and negative,  $p$  approaches 0.
  - ▶ As  $LO$  gets large and positive,  $p$  approaches 1.
  - ▶ At  $LO = 0$ ,  $p = 1/2$ . ( $LO = 0 \iff \text{odds} = 1$ .)

# Fitting the model to the data

- ▶ To fit this logistic-regression model to the data and get the parameter estimates, we use a technique called the method of *maximum likelihood*.
  - ▶ Likelihood methods for fitting statistical models to data and obtaining parameter estimates are beyond the scope of this course.
  - ▶ Instead, see STAT 475 on Applied Discrete Data Analysis, for which STAT 305 is a pre-requisite.
- ▶ The slope parameter  $\beta_1$  for  $X_1$  summarizes the association between  $Y$  and  $X_1$  in the population.
- ▶ For large sample sizes, the CLT kicks in and allows us to make approximate inference about the slope parameter  $\beta_1$  using our fitted model.

# Review of Natural Logarithms and Exponents

- ▶ Recall that if  $a$  is the natural logarithm of  $z$ , written  $a = \log(z)$ , then  $e^a = e^{\log(z)} = z$ .
- ▶ The logarithm of 1 is always zero;
  - ▶ e.g,  $0 = \log(1)$  and  $e^0 = 1$ .
- ▶ Sums of exponents are multiples; that is,  $e^{a+b} = e^a e^b$ .
- ▶ Differences of exponents are ratios; that is,  $e^{a-b} = e^a / e^b$ .
  - ▶ We will make use of this as  $e^a / e^b = e^{a-b}$ .

## Interpretation of $\beta_1$

- ▶ A one-unit (insert relevant units) increase in  $X_1$  is associated with a change of  $\beta_1$  in the log-odds of the outcome, or a  $e^{\beta_1}$ -fold change in the odds of the outcome.
- ▶ Need to take care with negative parameter values.
- ▶ E.G. Let's say that  $\beta_1 = -2$  and  $X_1$  is measured in grams. Then:

*A one-gram increase in  $X_1$  is associated with a change of  $-2$  in the log-odds of the outcome, or a  $e^{-2} = 0.135$ -fold change in the odds of the outcome.*

- ▶ Rephrase to be shorter. E.G., if  $X_1$  is birthweight and the outcome is BPD:

*A one-gram increase in birthweight is associated with a 0.135-fold change in the odds of BPD.*



## Mathematical justification of interpretation

- ▶ Let  $p_1$  be the probability of  $Y = 1$  given  $X_1 = x_1$ .
- ▶ When  $X_1 = x_1$ , we have log-odds

$$\log \left[ \frac{p_1}{1 - p_1} \right] = \alpha + \beta_1 x_1.$$

- ▶ Let  $p_2$  be the probability of  $Y = 1$  given  $X_1 = x_1 + 1$ .
- ▶ When  $X_1 = x_1 + 1$ , we have log-odds

$$\log \left[ \frac{p_2}{1 - p_2} \right] = \alpha + \beta_1 (x_1 + 1) = \alpha + \beta_1 x_1 + \beta_1.$$

- ▶ The odds at  $X_1 = x_1$  and  $X_1 = x_1 + 1$  are, respectively,

$$\frac{p_1}{1 - p_1} = e^{\alpha + \beta_1 x_1} \quad \text{and} \quad \frac{p_2}{1 - p_2} = e^{\alpha + \beta_1 x_1 + \beta_1},$$

- ▶ Hence the odds-ratio for  $X_1 = x_1 + 1$  relative to  $X_1 = x_1$  is

$$\left( \frac{p_2}{1 - p_2} \right) / \left( \frac{p_1}{1 - p_1} \right) = e^{\alpha + \beta_1 x_1 + \beta_1 - (\alpha + \beta_1 x_1)} = e^{\beta_1}.$$

# Interpretation of $\beta_1$ for a Binary Exposure

- ▶ If  $X_1$  is a binary exposure that takes values 1 for exposed and 0 for unexposed, a one-unit increase in  $X_1$  means going from unexposed to exposed.
- ▶ Set  $x_1 = 0$  on the previous slide to find that  $e^{\beta_1}$  is the odds ratio for the exposed subjects relative to the unexposed subjects.
- ▶ In the homework assignment, you will be asked to interpret fitted coefficients from a logistic regression on a binary exposure variable.

# BPD Example

- Let's read in the BPD data and look at it:

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_18/Data/bpd.csv")
bpd <- read.csv(uu)
head(bpd)
```

```
##      bpd birthwt  gestage  toxemia  steroid
## 1      1      850      27         0         0
## 2      0     1500      33         0         0
## 3      1     1360      32         0         0
## 4      0      960      35         1         0
## 5      0     1560      33         0         0
## 6      0     1120      29         0         1
```

## Fit the Logistic Regression of BPD on Birth Weight

- ▶ To fit a logistic-regression model to the data, we use the `glm()` function in R.
- ▶ Similar to the `lm()` function, `glm()` also requires a model formula.
  - ▶ The model formula is `bpd ~ birthwt`.
  - ▶ Response `bpd` on left-hand side of the `~` in the formula and explanatory variable `birthwt` on the right-hand side are columns in the dataframe `bpd`.

```
bfit <- glm(bpd~birthwt,data=bpd,family=binomial)
coefficients(bfit)
```

```
## (Intercept)      birthwt
##  4.03429128 -0.00422914
```

- ▶ To three significant digits, the estimated parameters are  $\hat{\alpha} = 4.03$  and  $\hat{\beta}_1 = -0.00423$

## Software Notes

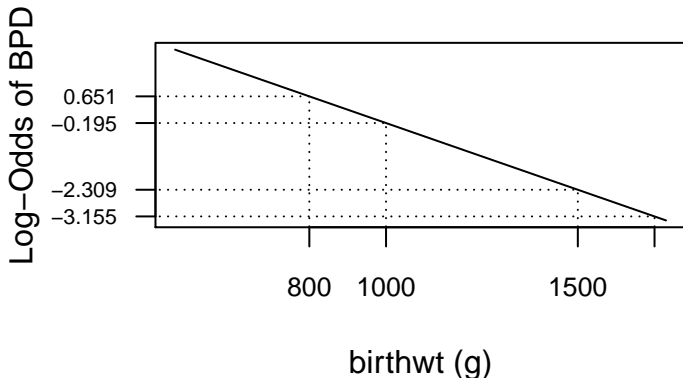
- ▶ When we fitted linear-regression models to data using the method of least squares, we used the *linear model* or `lm()` function in R.
- ▶ When we fit logistic-regression models to data using the method of maximum-likelihood, we use the *generalized linear model* or `glm()` function in R.
- ▶ In fact, `glm()` fits several types of models, including linear and logistic regression.
- ▶ Specify the type of model in `glm()` with the `family` option.
  - ▶ Regular linear regression is `family=gaussian` (default); same as using `lm()`.
  - ▶ Logistic regression is `family=binomial`.
- ▶ **BEWARE:** Omitting the `family=binomial` argument in `glm()` will fit a linear regression to binary-outcome data.
  - ▶ *For binary-outcome data, the fitted model will be nonsense.*

## Interpretation of Birth-Weight Effect

- ▶ To three significant digits,  $\hat{\beta}_1 = -0.00423$
- ▶ Model the log-odds, but interpret in terms of the odds.
- ▶ We estimate that a one-gram increase in birth weight is associated with a  $-0.00423$  change in the log-odds of BPD.
- ▶ Report: “A one-gram increase in birth weight is associated with an estimated  $e^{-0.00423} = 0.996$ -fold change in the odds of BPD.”
- ▶ As one-gram units are too fine-grained, can instead work with a 100-gram increase in birth weight.
  - ▶ Then a 100-gram increase in birth weight is associated with an estimated  $100 \times -0.00423 = -0.423$  change in the log-odds of BPD.
  - ▶ Report: “A 100-gram increase in birth weight is associated with an estimated  $e^{-0.423} = 0.655$ -fold change in the odds of BPD.”

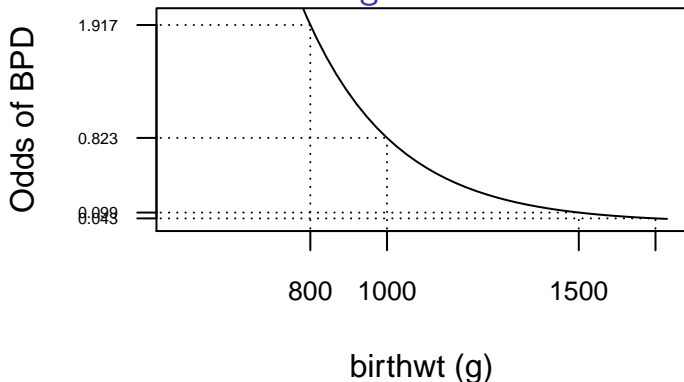
## Log-Odds of BPD vs. Birthweight

- ▶ The logistic-regression model specifies a straight-line relationship between the log-odds of BPD and `birthwt`.



- ▶ E.G.: A 200g increase in `birthwt` is associated with an estimated  $200 \times \hat{\beta}_1 = 200 \times -0.00423 = -0.846$  change in log-odds of BPD.
  - ▶ From plot, estimated log-odds of BPD in an 800g baby is 0.651.
  - ▶ So, for a 1000g baby, it is  $0.651 - 0.846 = -0.195$

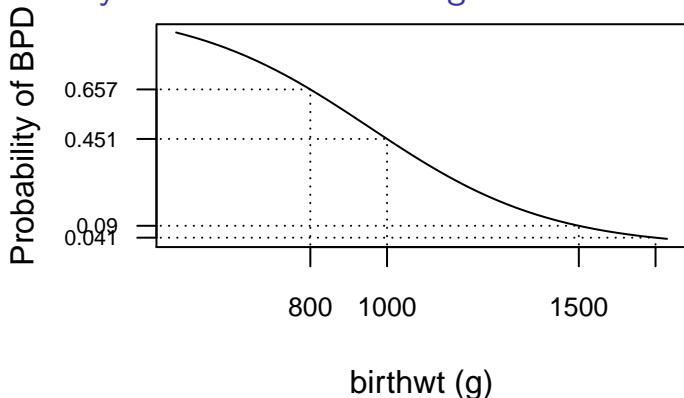
## Odds of BPD vs. Birthweight



- ▶ Exponentiate to get odds. E.G. A 200g increase in birthwt is associated with an estimated  $e^{200 \times -0.00423} = e^{-0.846}$  or 0.429-fold change in the odds of BPD.
  - ▶ From plot, estimated odds of BPD in an 800g baby are  $e^{0.651} = 1.92$ .
  - ▶ For a 1000g baby, the odds are  $e^{0.651 - 0.846} = e^{-0.195} = 0.823$



## Probability of BPD vs. Birthweight



- ▶ Saw that estimated log-odds of BPD in an 800g and 1000g baby are, respectively, 0.651 and  $0.651 - 0.846 = -0.195$ .
  - ▶ Therefore, corresponding probabilities of BPD are  $e^{0.651}/(1 + e^{0.651}) = 0.657$  for 800g babies and  $e^{-0.195}/(1 + e^{-0.195}) = 0.451$  for 1000g babies.

# Predicted Log-Odds and Probability of BPD

- ▶ We can use the `predict()` function to estimate the log-odds or the probability of the outcome at new values of the explanatory variable.
- ▶ The range of `birthwt` (in grams) in the `bpd` dataset is:

```
range(bpd$birthwt)
```

```
## [1] 450 1730
```

- ▶ Let's consider new values of `birthwt` towards the extremes of this range, for values 450.5g and 1729.5g

```
newdat <- data.frame(birthwt = c(450.5,1729.5))
library(dplyr)
newdat <- mutate(newdat,
  logodds = predict(bfit,newdata=newdat,type="link"),
  probability = predict(bfit,newdata=newdat,type="response"))
newdat
```

```
##   birthwt   logodds probability
## 1   450.5  2.129064  0.89369610
## 2  1729.5 -3.280006  0.03626351
```

# Software Notes

In the above calls to `predict()`:

- ▶ specifying the `type` argument as **`type=link`** requests predictions on the scale of the linear predictor;
  - ▶ i.e. on the **log-odds scale**,
  - ▶ possible values of the log-odds are between  $-\infty$  and  $\infty$ .
- ▶ specifying the `type` argument as **`type=response`** requests predictions on the scale of the response;
  - ▶ i.e. on the **probability scale**,
  - ▶ possible values are between 0 and 1.

# Fitting a Logistic Regression to Case-Control Data

- ▶ The study of low-birthweight babies takes a simple random sample from a hospital ICU to see which babies have BPD.
- ▶ But what if, instead, we had a case-control study.
  - ▶ A case-control study does not take a SRS from the population but rather separate SRS's from cases and from controls.
  - ▶ Cases are typically over-sampled relative to their frequency in the population.
- ▶ This is called *biased sampling* and the case-control study design is called a *biased sampling design*.
- ▶ The biased sampling leads to biased estimates of the intercept parameter  $\alpha$  in the linear predictor and therefore of the log odds, odds and probabilities.
  - ▶ We can't estimate any of these on an absolute scale.
- ▶ Fortunately, we **can** estimate the *changes* in the log odds and odds because estimates of the slope parameter  $\beta_1$  turn out to be unbiased.

- ▶ Since the estimates of  $\beta_1$  are not biased by the case-control sampling:
  - ▶  $\hat{\beta}_1$  can still be interpreted as the estimated effect of a one-unit increase in  $X_1$  on the log-odds of the disease outcome.
  - ▶  $e^{\hat{\beta}_1}$  can still be interpreted as the estimated odds-ratio describing the multiplicative change resulting from a one-unit increase in  $X_1$ .
- ▶ The association between the binary disease outcome  $Y$  and the explanatory variable  $X_1$  is our main interest, and  $e^{\hat{\beta}_1}$  estimates an odds-ratio that describes this association.