

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

Chapter 20, part 2: Inference in Logistic Regression

Jinko Graham

2018-11-28

Inference in Logistic Regression

- ▶ In logistic regression, the log-odds of the outcome is modeled by the straight-line relationship $\alpha + \beta_1 X_1$.
- ▶ The intercept α is not typically of interest; instead we focus on β_1 because it summarizes the effect of X_1 on Y .
- ▶ It turns out that the sampling distribution of $\hat{\beta}_1$ is approximately normal with mean β_1 and SD that depends on α and β_1 .
- ▶ Let $SE(\hat{\beta}_1)$ denote the SE of $\hat{\beta}_1$, obtained by inserting parameter estimates $\hat{\alpha}$ and $\hat{\beta}_1$ into the SD formula for $\hat{\beta}_1$.
- ▶ For large samples, the pivotal quantity has an approximate standard-normal distribution; i.e.,

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0, 1)$$

- ▶ Hypothesis tests and CIs follow in the usual way.

Dataset

- ▶ We'll be working with the `bpd` dataframe of low-birthweight babies from the neonatal ICU of a large hospital:

```
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
```

- ▶ The `bpd` dataframe has a variable `bpd` indicating whether the baby had bronchopulmonary dysplasia.
 - ▶ This condition results from damage to the lungs caused by a respirator and long-term use of oxygen.
 - ▶ Most infants recover, but some may have long-term breathing difficulty.

Logistic Regression of BPD on Birth Weight

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

```
##              Estimate Std. Error  z value    Pr(>|z|)  
## (Intercept)  4.03429128 0.6957120604  5.798795 6.679332e-09  
## birthwt      -0.00422914 0.0006407678 -6.600112 4.108460e-11
```

- ▶ β_1 is the increase in the log-odds of BPD associated with a one-gram increase in birthweight.
- ▶ If $\beta_1 = 0$, then the log-odds of BPD don't change with birthweight and so BPD and birthweight are not associated.
- ▶ To assess whether BPD is associated with birthweight, we test the hypotheses that $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- ▶ The estimate of β_1 is $\hat{\beta}_1 = -0.0042$ with SE 0.00064.
- ▶ The test statistic is

$$\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.0042 - 0}{0.00064} = -6.6,$$

which gives a tiny p value. Strong statistical evidence that BPD is associated with birth weight.

Approximate 95% CIs

- ▶ We will obtain 95% CI's for the following:
 - ▶ β_1 , the increase in the log-odds of BPD associated with a one-gram increase in birthweight, and
 - ▶ e^{β_1} , the factor by which the odds of BPD changes with a one-gram increase in birthweight (an OR)
- ▶ An approximate 95% CI for β_1 is

$$\hat{\beta}_1 \pm z^* \times SE(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1);$$

i.e., estimate \pm margin of error.

- ▶ For the BPD data, the 95% CI for β_1 is

$$-0.0042 \pm 1.96 \times 0.00064 = (-0.0055, -0.0029)$$

- ▶ The 95% CI for e^{β_1} is obtained by exponentiating and so is

$$(e^{-0.0055}, e^{-0.0029}) = (.995, .997)$$

But CIs from R are obtained differently

- ▶ The `confint()` function in R, when applied to a `glm()`-fitted object (such as `bfit`), gets the CIs for logistic-regression coefficients differently.
- ▶ Its CIs are based on inverting hypothesis tests directly:
 - ▶ E.G., A 95% CI is the set of all β_1 values, b , retained in a test of $H_0 : \beta_1 = b$ vs. $H_a : \beta_1 \neq b$ at the 5% level, with the data at hand.
 - ▶ Recall: Same approach used by `mantelhaen.test()` to get its CI (see Ch16 notes, pg 23).

```
confint(bfit) # CIs for alpha and beta1
```

```
##                2.5 %          97.5 %  
## (Intercept)  2.727548536  5.466632033  
## birthwt      -0.005565106 -0.003040923
```

```
exp(confint(bfit)["birthwt",]) # CI for OR parameter  $e^{\beta_1}$ 
```

```
##          2.5 %          97.5 %  
## 0.9944504 0.9969637
```

Comparing CI methods

- ▶ In this example, the CI for e^{β_1} from applying `confint()` to the `glm()`-fitted object is (0.994,0.997), which is very similar to the CI of (0.995,0.997) for the pivotal-quantity method.
- ▶ Know how to extract what is needed from the coefficients summary below to calculate a CI for the logistic-regression coefficients using the pivotal-quantity method.

```
summary(bfit)$coefficients
```

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	4.03429128	0.6957120604	5.798795	6.679332e-09
## birthwt	-0.00422914	0.0006407678	-6.600112	4.108460e-11

- ▶ Need the coefficient estimate, its SE and the critical value z^* .
- ▶ The critical value will be provided; for 95% CIs it is 1.96.

Interpreting the CI

- ▶ Also, know how to interpret the CI from `confint()`.
 - ▶ e.g., the `birthwt` row of `confint(bfit)` gives us the 95% CI for β_1 , the slope coefficient for `birthwt`.

```
confint(bfit)["birthwt",]
```

```
##          2.5 %          97.5 %  
## -0.005565106 -0.003040923
```

- ▶ To get the 95% CI for the OR, e^{β_1} , exponentiate the above CI for β_1 :

```
exp(confint(bfit)["birthwt",])
```

```
##          2.5 %          97.5 %  
## 0.9944504 0.9969637
```

- ▶ Interpretation: “With 95% confidence, an increase in birthweight of 1 gram is associated with an estimated change in the odds of BPD by a factor of between 0.994 and 0.997.”

Example with Binary Explanatory Variable

- ▶ The bpd dataframe also has a column for the toxemia status of the baby's mother (1 if she was toxic and 0 if not)

```
head(bpd, n=3)
```

```
##      bpd birthwt gestage toxemia steroid
## 1      1      850      27        0        0
## 2      0     1500      33        0        0
## 3      1     1360      32        0        0
```

- ▶ Logistic regression of BPD on the binary variable toxemia:

```
bfit2 <- glm(bpd~toxemia,data=bpd,family=binomial)
summary(bfit2)$coefficients
```

```
##              Estimate Std. Error  z value    Pr(>|z|)
## (Intercept) -0.5717863   0.1494999 -3.824660 0.0001309529
## toxemia      -0.7719484   0.4821774 -1.600964 0.1093849689
```

CI for the Toxemia Effect

- ▶ Recall: For a binary exposure, X_1 , e^{β_1} is the odds-ratio (OR) for exposed vs. unexposed groups.
- ▶ In this example, e^{β_1} is the ratio of the odds of BPD given toxemia divided by the odds of BPD given no toxemia

```
confint(bfit2)["toxemia",] #CI for beta1, the log-OR
```

```
##      2.5 %      97.5 %  
## -1.8057010  0.1162591
```

```
exp(confint(bfit2)["toxemia",]) # CI for OR
```

```
##      2.5 %      97.5 %  
## 0.1643592  1.1232868
```

- ▶ An approximate 95% CI for this OR is (0.164, 1.12):