Inference and Multiple Explanatory Variables

Logistic regression – Stat 230

Recap: Framingham heart study

Goal: Determine whether participant experienced coronary heart disease (CHD) in 10-year window after their exam

- Y = CHD (0 = no, 1 = yes)
- X = participant's age, sex, total cholesterol, and systolic blood pressure

Strategy: model the probability of CHD given these factors

Binary logistic regression model

If Y follows a Bernoulli distribution

$$\mathrm{E}(Y|X) = \pi(X)$$

We link this mean function to the explanatory variables using the logit link

$$\begin{aligned}\eta &= \operatorname{logit}(\pi(X)) \\ &= \log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\end{aligned}$$

Framingham model

$$\log\left(\frac{\hat{\pi(X)}}{1-\hat{\pi(X)}}\right) = -8.08 + 0.061\texttt{age} + 0.686\texttt{male} + 0.002\texttt{totChol} - 0.018\texttt{sysBP}$$

$\hat{\beta}_2$

- $e^{0.686} = 1.986$
- The odds of having a heart attack in the next ten years are nearly twice as high for males as females, after accounting for age, total cholesterol, and systolic blood pressure.

$\hat{\beta}_4$

- $e^{-0.018} = 0.9822$
- The odds of having a heart attack in the next ten years decrease by about 1.8% (i.e., a factor of 0.982) for a one-unit increase is systolic blood pressure, after accounting for age, sex, and total cholesterol.

Wald-based inference for logistic regression

Maximum likelihood (ML) estimation

The coefficients in logistic regression are estimated by finding the $\hat{\beta}_0, \ldots, \hat{\beta}_p$ that maximize the probability of the observed outcomes

$$L(\beta) = P(Y_1 = y_1, \ldots Y_n = y_n | \beta, X) = \prod_{i=1}^{n} \pi(X)^{y_i} \left[1 - \pi(X)\right]^{1-y_i}$$

where

$$\pi(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Properties of ML estimators

Large-sample properties of ML estimators, $\widehat{\beta}_i$s (if the model is correct):

1. Essentially unbiased
2. SEs can be computed and are about as small as any other unbiased estimator
3. The sampling distributions for the estimators are approximately normal

Wald test for a coefficient

Hypotheses: $H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$

Test statistic:

$$z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Reference distribution: $N(0, 1)$

CI for a coefficient

A normal-based confidence interval for β_i

$$\hat{\beta}_i \pm z^*_{1-\alpha/2} SE(\hat{\beta}_i)$$

CI for the multiplicative effect on odds of success for a 1- unit change in x (odds ratio of Success for $x+1$ vs. x):

$$e^{\hat{\beta}_i - z^*_{1-\alpha/2} SE(\hat{\beta}_i)} \quad \text{to} \quad e^{\hat{\beta}_i + z^*_{1-\alpha/2} SE(\hat{\beta}_i)}$$

… for a C-unit change in x

$$e^{C \cdot \hat{\beta}_i - z^*_{1-\alpha/2} C \cdot SE(\hat{\beta}_i)} \quad \text{to} \quad e^{C \cdot \hat{\beta}_i + z^*_{1-\alpha/2} C \cdot SE(\hat{\beta}_i)}$$

Framingham example

What impact does age have on the odds of having a heart attack in the next 10 years?

```
# A tibble: 5 × 5
  term        estimate std.error    conf.low conf.high
  <chr>          <dbl>     <dbl>       <dbl>     <dbl>
1 (Intercept) -8.08      0.412   -8.90        -7.29
2 age          0.0610    0.00579  0.0497       0.0724
3 male         0.686     0.0930   0.505        0.869
4 totChol      0.00201   0.00102  0.00000310   0.00401
5 sysBP        0.0176    0.00200  0.0137       0.0215
```

Let's construct a 95% confidence interval for β_1:

$$\widehat{\beta}_1 \pm z^*_{1-\alpha/2} SE(\widehat{\beta}_1)$$

$z^*_{1-0.05/2} = z^*_{0.975} =$ 0.975 quantile from $N(0, 1)$

```r
qnorm(0.975)
```

```
[1] 1.959964
```

$$0.0610 \pm 1.96 \cdot (0.00579) = (0.0497,\ 0.0724)$$

Framingham example

$0.0610 \pm 1.96 \cdot (0.00579) = (0.0497,\ 0.0724)$

Exponentiating the endpoints to get the CI for the odds ratio for a one-year increase in age:

$$e^{0.0497} = 1.051 \quad \text{to} \quad e^{0.0724} = 1.075$$

We are 95% confident that a one-year increase in age is associated with an increase in the odds of having a heart attack in the next 10 year of between 5.1% (a 1.051 factor

Framingham example

How do the odds of having a heart attack in the next 10 years change for someone 10 years older?

```
# A tibble: 5 × 5
  term        estimate std.error    conf.low conf.high
  <chr>          <dbl>     <dbl>       <dbl>     <dbl>
1 (Intercept) -8.08      0.412   -8.90         -7.29
2 age          0.0610    0.00579  0.0497        0.0724
3 male         0.686     0.0930   0.505         0.869
4 totChol      0.00201   0.00102  0.00000310    0.00401
5 sysBP        0.0176    0.00200  0.0137        0.0215
```

95% CI: $C \cdot \left[\hat{\beta}_1 \pm z^*_{1-\alpha/2} SE(\hat{\beta}_1)\right]$

$$
\begin{aligned}
10\,[0.0610 \pm 1.96 \cdot (0.00579)] &= (10(0.0497),\ 10(0.0724)) \\
&= (0.497,\ 0.724)
\end{aligned}
$$

We are 95% confident that a 10-year increase in age is associated with an increase in the odds of having a heart attack in the next 10 year of between a factor of 1.644 (a 64.4% increase) and a factor of 2.063 (a 106.3% increase), holding all other variables constant.

Likelihood-based inference for logistic regression

Likelihood function

Recall that the likelihood function gives the plausibility of the observed data given our parameter values

$$L(\beta) = P(Y_1 = y_1, \ldots Y_n = y_n | \beta, X) = \prod_{i=1}^{n} \pi(X)^{y_i} \left[1 - \pi(X)\right]^{1-y_i}$$

Idea:

- "better" model explains more of the variation in our data set
- "better" model makes our data more plausible

Likelihood ratio test

Full model: $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_p x_p$

Reduced model: $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

Hypotheses: $H_0 : \beta_{k+1} = \cdots = \beta_p = 0$ vs. H_a : at least one $\beta_j \neq 0$

Test statistic:

$$G = 2 \cdot \text{log-likelihood(full model)} - 2 \cdot \text{log-likelihood(reduced model)}$$

Reference distribution: χ^2 distribution

d.f. = # βs in full model – # βs in reduced model

Deviance

> **R implementation**
>
> R reports the **deviance** rather than the log-likelihood

In GLM, deviance is used to measure "unexplained" variation in the response

Alternate representation of the LRT test statistic

$$
\begin{aligned}
G &= 2 \cdot \text{log-likelihood(full model)} - 2 \cdot \text{log-likelihood(reduced model)} \\
&= \text{deviance(reduced model)} - \text{deviance(full model)}
\end{aligned}
$$

The LRT is sometimes called the drop-in-deviance test

Framingham example

R's default output gives

```
# A tibble: 5 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) -8.08      0.412       -19.6  9.65e-86
2 age          0.0610    0.00579      10.5  5.32e-26
3 male         0.686     0.0930        7.38 1.56e-13
4 totChol      0.00201   0.00102       1.98 4.83e- 2
5 sysBP        0.0176    0.00200       8.80 1.39e-18

    Null deviance: 3564.8  on 4189  degrees of freedom
Residual deviance: 3214.1  on 4185  degrees of freedom
```

Full model: $\eta = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{male} + \beta_3 \texttt{totChol} + \beta_4 \texttt{sysBP}$

Reduced model: $\eta = \beta_0$

Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_a : \text{at least one } \beta_j \neq 0$

$G = \text{deviance(reduced model)} - \text{deviance(full model)}$

$G = 3564.8 - 3214.1 = 350.7$

$\text{d.f.} = 5 - 1 = 4$

Framingham example

Hypotheses: $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_a$: at least one $\beta_j \neq 0$

$G = \text{deviance(reduced model)} - \text{deviance(full model)}$

$G = 3564.8 - 3214.1 = 350.7$

d.f. $= 5 - 1 = 4$

```r
1 - pchisq(350.7, df = 4)
```

```
[1] 0
```

There is overwhelming evidence that at least one of the explanatory variables helps explain the odds of having a heart attack in the next ten years ($G = 350.7$, $\text{d.f.} = 4$, p-value < 0.001).