Logistic regression

Inference & Model selection

Prof. Maria Tackett

04.01.20

STA 210

Click for PDF of slides

Risk of coronary heart disease

This dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to predict if a patient has a high risk of getting coronary heart disease in the next 10 years.

Response:

`TenYearCHD`:

- 0 = Patient is not high risk of having coronary heart disease in the next 10 years
- 1 = Patient is high risk of having coronary heart disease in the next 10 years

Predictors:

- `age`: Age at exam time.
- `currentSmoker`: 0 = nonsmoker; 1 = smoker
- `totChol`: total cholesterol (mg/dL)

Modeling risk of coronary heart disease

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.111	0.077	-27.519	0.000	-2.264	-1.963
ageCent	0.081	0.006	13.477	0.000	0.070	0.093
currentSmoker1	0.447	0.099	4.537	0.000	0.255	0.641
totCholCent	0.003	0.001	2.339	0.019	0.000	0.005

Hypothesis test for β_j

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

The test of significance for the coefficient β_j is

Hypotheses: $H_0 : \beta_j = 0 \ \text{ vs } \ H_a : \beta_j \neq 0$

Test Statistic:

$$z = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

P-value: $P(|Z| > |z|)$,

where $Z \sim N(0, 1)$, the Standard Normal distribution

Confidence interval for β_j

- We can calculate the **C\% confidence interval** for β_j using the following:

$$\hat{\beta}_j \pm z^* SE(\hat{\beta}_j)$$

where z^* is calculated from the $N(0, 1)$ distribution

> We are $C\%$ confident that for every one unit change in x_j, the odds multiply by a factor of $\exp\{\hat{\beta}_j - z^* SE(\hat{\beta}_j)\}$ to $\exp\{\hat{\beta}_j + z^* SE(\hat{\beta}_j)\}$, holding all else constant.

STA 210

Modeling risk of coronary heart disease

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.11059	0.07670	-27.51865	0.00000	-2.26360	-1.96285
ageCent	0.08150	0.00605	13.47651	0.00000	0.06973	0.09344
currentSmoker1	0.44743	0.09862	4.53705	0.00001	0.25461	0.64134
totCholCent	0.00254	0.00108	2.33919	0.01933	0.00040	0.00465

1. How is the test statistic for `currentSmoker1` calculated?
2. Is `totCholCent` a statistically significant predictor of the odds a person is high risk for coronary heart disease?
   - Justify your answer using the test statistic and p-value.
   - Justify your answer using the confidence interval.

Model Selection

Comparing Nested Models

- Suppose there are two models:
  - Reduced Model includes predictors $x_1, \ldots, x_q$
  - Full Model includes predictors $x_1, \ldots, x_q, x_{q+1}, \ldots, x_p$
- We want to test the hypotheses

$$
\begin{aligned}
H_0 &: \beta_{q+1} = \cdots = \beta_p = 0 \\
H_a &: \text{ at least 1 } \beta_j \text{ is not} 0
\end{aligned}
$$

- To do so, we will use the **Drop-in-Deviance test** (also known as the Nested Likelihood test)

Deviance residual

- The deviance residual is the a measure of how much the observed data differs from what is measured using the likelihood ratio
- The deviance residual for the i^{th} observation is

$$d_i = \text{sign}(y_i - \hat{\pi}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{\pi}_i}\right)\right]}$$

where $\text{sign}(y_i - \hat{\pi}_i)$ is positive when $y_i = 1$ and negative when $y_i = 0$.

STA 210

Drop-in-Deviance Test

- The **deviance statistic for Model *k*** is $G^2_k = \sum\limits_{i=1}^{n} d_i^2$

- To test the hypotheses

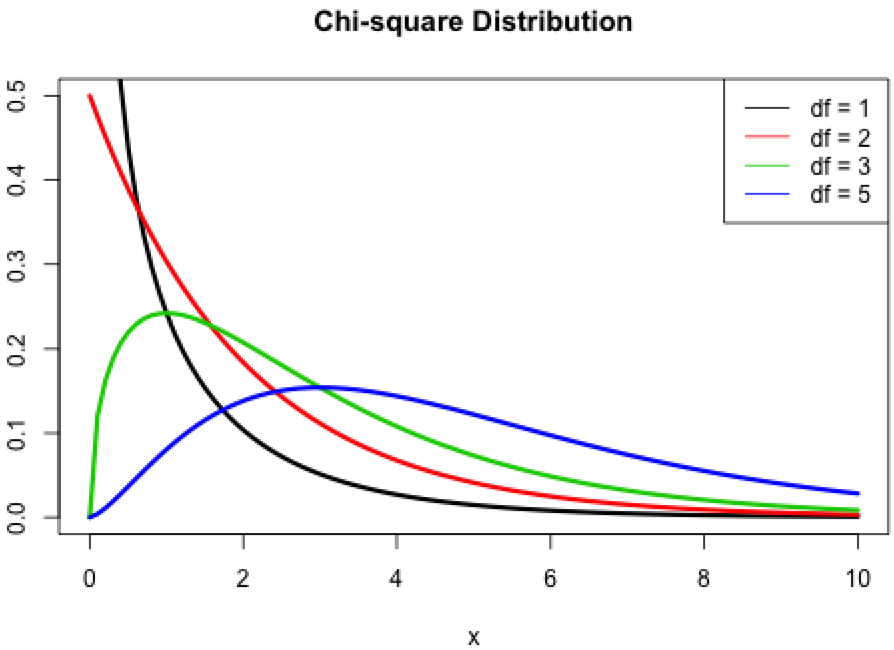
$$
\begin{aligned}
H_0 &: \beta_{q+1} = \cdots = \beta_p = 0 \\
H_a &: \text{ at least one } \beta_j \text{ is not} 0
\end{aligned}
$$

the **drop-in-deviance statistic** is $G^2_{reduced} - G^2_{full}$

The p-value for the test is calculated using a Chi-square distribution (χ^2) with degrees of freedm equal to the difference in the number of parameters in the full and reduced models

STA 210

χ^2 distribution



Calculate p-value for the drop-in-deviance test as $P(\chi^2 > \text{test statistic})$

Should we add **`education`** to the model?

```
model_red <- glm(TenYearCHD ~ ageCent + currentSmoker + totChol,
              data = heart_data, family = binomial)
model_full <- glm(TenYearCHD ~ ageCent + currentSmoker + totChol +
               education,
              data = heart_data, family = binomial)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.617	0.277	-9.448	0.000	-3.161	-2.074
ageCent	0.078	0.006	12.701	0.000	0.066	0.091
currentSmoker1	0.449	0.099	4.540	0.000	0.255	0.643
totChol	0.003	0.001	2.503	0.012	0.001	0.005
education2	-0.266	0.119	-2.233	0.026	-0.502	-0.034
education3	-0.319	0.144	-2.215	0.027	-0.607	-0.041
education4	-0.116	0.160	-0.725	0.468	-0.436	0.192

STA 210

Should we add **`education`** to the model?

```
# Deviances
(dev_red <- glance(model_red)$deviance)
```

```
## [1] 2894.989
```

```
(dev_full <- glance(model_full)$deviance)
```

```
## [1] 2887.206
```

```
# Drop-in-deviance test statistic
(test_stat <- dev_red - dev_full)
```

```
## [1] 7.783615
```

Should we add **`education`** to the model?

```
# p-value
1 - pchisq(test_stat, 3) #3 = number of new model terms in model2
```

```
## [1] 0.05070196
```

What is your conclusion for the test?

Drop-in-Deviance test in R

- We can use the **anova** function to conduct this test
  - Add **test = "Chisq"** to conduct the drop-in-deviance test

```
anova(model_red, model_full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: TenYearCHD ~ ageCent + currentSmoker + totChol
## Model 2: TenYearCHD ~ ageCent + currentSmoker + totChol + education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3654     2895.0
## 2      3651     2887.2  3   7.7836   0.0507 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

STA 210

Model selection

- Use AIC or BIC for model selection

$$AIC = -2 * L - {\color{red}n\log(n)} + 2(p + 1)$$

$$BIC = -2 * L - {\color{red}n\log(n)} + log(n) \times (p + 1)$$

where $L = \sum_{i=1}^{n}[y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i)]$

STA 210

AIC from `glance` function

Let's look at the AIC for the model that includes `ageCent`, `currentSmoker`, and `totCholCent`

```
glance(model_red)$AIC
```

```
## [1] 2902.989
```

Calculating AIC

```
(L <- glance(model_red)$logLik)
```

```
## [1] -1447.495
```

```
- 2 * L + 2 * (3 + 1)
```

```
## [1] 2902.989
```

Should we add education to the model?

Recall:

- Reduced Model includes `AgeCent`, `currentSmoker`, `totCholCent`
- Full Model includes `AgeCent`, `currentSmoker`, `totCholCent`, `education` (categorical)

```
glance(model_red)$AIC
```

```
## [1] 2902.989
```

```
glance(model_full)$AIC
```

```
## [1] 2901.206
```

Based on the AIC, which model would you choose?

What remaining questions do you have about logistic regression?