# Logistic regression

## Inference

### Prof. Maria Tackett

[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 examine the relationship between various health characteristics and the risk of having heart disease in the next 10 years.

**high_risk**: 1 = High risk, 0 = Not high risk

**age**: Age at exam time (in years)

**education**: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College

**currentSmoker**: 0 = nonsmoker; 1 = smoker

# Modeling risk of coronary heart disease

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -5.385 + 0.073 \text{ age} - 0.242 \text{ ed2} - 0.235 \text{ ed3} - 0.020 \text{ ed4}$$

# Hypothesis test for $\beta_j$

**Hypotheses**: $H_0 : \beta_j = 0$ vs $H_a : \beta_j \neq 0$

**Test Statistic**:

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

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

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

# Confidence interval for $\beta_j$

We can calculate the **C% confidence interval** for $\beta_j$ as the following:

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

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

# Confidence interval for $\beta_j$

We can calculate the **C% confidence interval** for $\beta_j$ as the following:

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

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

This is an interval for the change in the log-odds for every one unit increase in $x_j$.

# Interpretation in terms of the odds

The change in **odds** for every one unit increase in $x_j$.

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

# Interpretation in terms of the odds

The change in **odds** for every one unit increase in $x_j$.

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

**Interpretation**: 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.

# Model

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

### Hypotheses

$$H_0 : \beta_1 = 0 \ \text{ vs } \ H_a : \beta_1 \neq 0$$

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|------|---------|-----------|-----------|---------|----------|-----------|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

**Test statistic**

$$z = \frac{0.0733 - 0}{0.00547} = 13.4$$

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

**P-value**

$$P(|Z| > |13.4|) \approx 0$$

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|------|---------|-----------|-----------|---------|----------|-----------|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

```
2 * pnorm(13.4,lower.tail = FALSE)
```

```
## [1] 6.046315e-41
```

# Let's look at the coefficient for **age**

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

**Conclusion**: The p-value is very small, so we reject $H_0$. The data provide sufficient evidence that age is a statistically significant predictor of whether someone is high risk of having heart disease, *after accounting for education*.

STA 210

# Comparing models

# Log likelihood

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

- Measure of how well the model fits the data

- Higher values of $\log L$ are better

- **Deviance** $= -2 \log L$

  - $-2 \log L$ follows a $\chi^2$ distribution with $n - p - 1$ degrees of freedom

# 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

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

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

# Drop-in-deviance test

**Hypotheses**:

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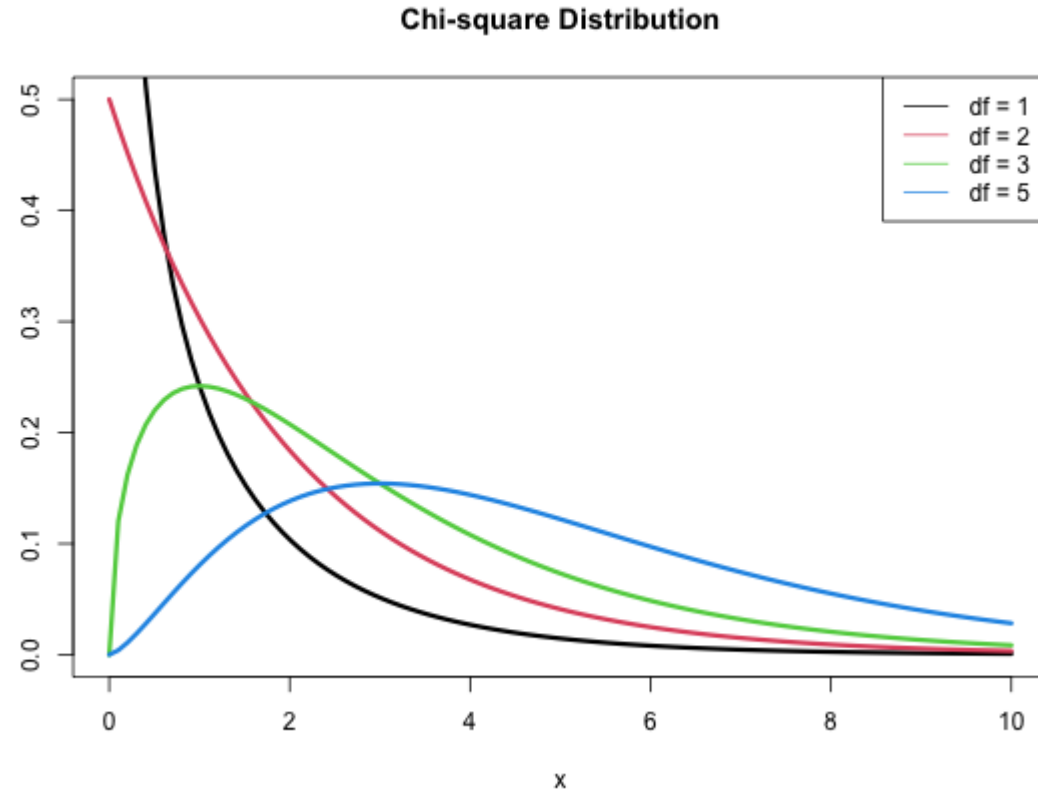$$H_a : \text{ at least } 1 \ \beta_j \text{ is not } 0$$

**Test Statistic**:

$$G = (-2 \log L_{reduced}) - (-2 \log L_{full})$$

**P-value**: $P(\chi^2 > G)$,

calculated using a $\chi^2$ distribution with degrees of freedom equal to the difference in the number of parameters in the full and reduced models

STA 210

# $\chi^2$ distribution



Chi-square Distribution

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | -5.385 | 0.308 | -17.507 | 0.000 | -5.995 | -4.788 |
| age | 0.073 | 0.005 | 13.385 | 0.000 | 0.063 | 0.084 |
| education2 | -0.242 | 0.112 | -2.162 | 0.031 | -0.463 | -0.024 |
| education3 | -0.235 | 0.134 | -1.761 | 0.078 | -0.501 | 0.023 |
| education4 | -0.020 | 0.148 | -0.136 | 0.892 | -0.317 | 0.266 |

Should we add **currentSmoker** to this model?

# Should we add **currentSmoker** to the model?

```
model_reduced <- glm(TenYearCHD ~ age + education,
              data = heart, family = binomial)
```

```
model_full <- glm(TenYearCHD ~ age + education +
                      currentSmoker,
              data = heart, family = binomial)
```

# Should we add **currentSmoker** to the model?

```
# Calculate deviance for each model
(dev_reduced <- glance(model_reduced)$deviance)
```

```
## [1] 3300.135
```

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

```
## [1] 3279.359
```

# Should we add **currentSmoker** to the model?

```r
# Calculate deviance for each model
(dev_reduced <- glance(model_reduced)$deviance)
```

```
## [1] 3300.135
```

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

```
## [1] 3279.359
```

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

```
## [1] 20.77589
```

# Should we add **currentSmoker** to the model?

```r
# p-value
#1 = number of new model terms in model 2
pchisq(test_stat, 1, lower.tail = FALSE)
```

```
## [1] 5.162887e-06
```

# Should we add **currentSmoker** to the model?

```
# p-value
#1 = number of new model terms in model 2
pchisq(test_stat, 1, lower.tail = FALSE)
```

```
## [1] 5.162887e-06
```

**Conclusion**: The p-value is very small, so we reject $H_0$. The data provide sufficient evidence that the coefficient of **currentSmoker** is not equal to 0. Therefore, we should add it to the model.

# 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_reduced, model_full, test = "Chisq") %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   Resid..Df Resid..Dev    df Deviance     p.value
##       <dbl>      <dbl> <dbl>    <dbl>       <dbl>
## 1      4130      3300.    NA       NA  NA
## 2      4129      3279.     1     20.8  0.00000516
```

# Model selection

Use AIC or BIC for model selection

$$AIC = -2 * \log L - n \log(n) + 2(p + 1)$$

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

# AIC from `glance` function

Let's look at the AIC for the model that includes **age**, **education**, and **currentSmoker**

```
glance(model_reduced)$AIC
```

```
## [1] 3310.135
```

# AIC from `glance` function

Let's look at the AIC for the model that includes **age**, **education**, and **currentSmoker**

```
glance(model_reduced)$AIC
```

```
## [1] 3310.135
```

Calculating AIC

```
- 2 * glance(model_reduced)$logLik + 2 * (4 + 1)
```

```
## [1] 3310.135
```

# Comparing the models using AIC

Let's compare the full and reduced models using AIC.

```
glance(model_reduced)$AIC
```

```
## [1] 3310.135
```

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

```
## [1] 3291.359
```

Based on AIC, which model would you choose?

# Comparing the models using BIC

Let's compare the full and reduced models using BIC

```
glance(model_reduced)$BIC
```

```
## [1] 3341.772
```

```
glance(model_full)$BIC
```

```
## [1] 3329.323
```

Based on BIC, which model would you choose?