Likelihood ratio tests

- Department Seminar: Dr. Mine Çetinkaya-Rundel
  September 26, 12pm - 1pm Kirby 120

  - Sign up to meet with the speaker (11 - 11:30)

- Solutions posted to Friday's class activity

Last time

Data on the RMS *Titanic* disaster. We have data on 891 passengers on the ship, with the following variables:

- `Passenger`: A unique ID number for each passenger.
- `Survived`: An indicator for whether the passenger survived (1) or perished (0) during the disaster.
- `Pclass`: Indicator for the class of the ticket held by this passengers; 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
- `Sex`: Binary Indicator for the biological sex of the passenger.
- `Age`: Age of the passenger in years; Age is fractional if the passenger was less than 1 year old.
- `Fare`: How much the ticket cost in US dollars.
- + others

$$Sex = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

Last time

Is there a relationship between passenger age and their probability of survival, after accounting for sex, passenger class, and the cost of their ticket?

β_1 = slope for female passengers

$\beta_1 + \beta_3$ = slope for males

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Age_i \cdot Sex_i + \beta_4 \log(Fare_i + 1)$$

> What hypotheses should we test to investigate this research question?

$H_0: \beta_1 = \beta_3 = 0$

H_A: at least one of $\beta_1, \beta_3 \neq 0$

Wald test

1) Rewrite the order of the terms

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1) + \beta_3 Age_i + \beta_4 Age_i \cdot Sex_i$$

2) $\beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix}$ $\quad \beta_{(2)} = \begin{bmatrix} \beta_3 \\ \beta_4 \end{bmatrix}$ $\quad H_0: \beta_{(2)} = \beta_{(2)}^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$H_A: \beta_{(2)} \neq \beta_{(2)}^0$

3) Partition variance matrix: $\mathcal{I}^{-1}(\beta) = \begin{bmatrix} \mathcal{I}^{11} & \mathcal{I}^{12} \\ \mathcal{I}^{21} & \mathcal{I}^{22} \end{bmatrix}$

$Var(\hat{\beta}_{(2)}) = \mathcal{I}^{22}$

$\mathcal{I}^{-1}(\beta) = (k+1) \times (k+1)$ matrix

Here: 5×5 matrix $\quad \mathcal{I}^{22} \in \mathbb{R}^{2 \times 2}$ (b/c $\beta_{(2)} \in \mathbb{R}^2$)

$\Rightarrow$ $\mathcal{I}^{11} \in \mathbb{R}^{3 \times 3}$

$\mathcal{I}^{12} \in \mathbb{R}^{3 \times 2}$

$\mathcal{I}^{21} \in \mathbb{R}^{2 \times 3}$

$\mathcal{I}^{22} \in \mathbb{R}^{2 \times 2}$

4) Test Statistic: $W = \left(\hat{\beta}_{(2)} - \beta^{0}_{(2)}\right)^T \left(\tilde{\Sigma}^{22}\right)^{-1} \left(\hat{\beta}_{(2)} - \beta^{0}_{(2)}\right)$

5) p-value: Under H_0, $W \sim \chi^2_q \leftarrow$ # parameters tested

Here: $W \sim \chi^2_2$

Full model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1) + \beta_3 Age_i + \beta_4 Age_i \cdot Sex_i$$

H_0: $\beta_3 = \beta_4 = 0$

$\Rightarrow$ Reduced model: $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1)$

Likelihood ratio tests

```
...
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.40695    0.44682  -3.149  0.00164 **
## Age              0.01107    0.01107   1.000  0.31730
## Sexmale         -1.27467    0.41654  -3.060  0.00221 **
## log(Fare + 1)    0.69449    0.11065   6.276 3.47e-10 ***
## Age:Sexmale     -0.03638    0.01378  -2.639  0.00831 **
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 697.21  on 709  degrees of freedom
...
```

Deviance!

> What information replaces R^2 and R^2_{adj} in the GLM output?

For logistic: AIC = 2(k+1) + deviance

Deviance

(residual deviance)

Deviance similar to SSE for linear regression

Definition: The *deviance* of a fitted model with parameter estimates $\widehat{\beta}$ is given by

(want to minimize deviance)

$$\underbrace{2\ell(\textbf{saturated model})}_{} - 2\ell(\widehat{\beta})$$

log likelihood for the fitted model

still $\sum_{i=1}^{n} \log\left(\hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}\right)$

e.g. $\sum_{i=1}^{n} \log\left(\hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}\right)$

But we estimate $\hat{p}_i$ separately for each point

$\hat{p}_i = \dfrac{e^{\hat{\beta}^T x_i}}{1+e^{\hat{\beta}^T x_i}}$

Saturated model: $\hat{p}_i = y_i$

$\Rightarrow \sum_{i=1}^{n} \log(1) = 0$

For binary logistic regression: deviance $= -2\ell(\hat{\beta})$

Residual and null deviance

```r
m1 <- glm(Survived ~ Age*Sex + log(Fare + 1),
          data = titanic, family = binomial)
summary(m1)
```

```
...
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 697.21  on 709  degrees of freedom
...
```

Residual deviance : deviance for

$\Rightarrow -2\ell(\hat{\beta}) = 697.21$

$\Rightarrow \ell(\hat{\beta}) = -\frac{697.21}{2}$

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1) + \beta_3 Age_i + \beta_4 Age_i \cdot Sex_i$

Null deviance : (residual) deviance for (intercept-only model)

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0$

$(\Rightarrow \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \bar{Y} =$ prevalence of 1s)

Linear regression : (Intuition)

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

SSTotal = SSE + SSReg

(SSModel)

$\uparrow$ null deviance

$\uparrow$ residual deviance

Drop-in-deviance:

$G = 708.04 - 697.21$

$= 10.83$

Comparing deviances

```
m1 <- glm(Survived ~ Age*Sex + log(Fare + 1),
          data = titanic, family = binomial)
summary(m1)
```

Full model

```
...
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 697.21  on 709  degrees of freedom
...
```

Reduced model

```
m2 <- glm(Survived ~ Sex + log(Fare + 1),
          data = titanic, family = binomial)
summary(m2)
```

```
...
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 708.04  on 711  degrees of freedom
...
```

Comparing deviances

Full model:

Hypotheses:

Reduced model:

Test statistic:

Comparing deviances

Full model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1) + \beta_3 Age_i + \beta_4 Age_i \cdot Sex_i$$

Reduced model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1)$$

$$G = 2\ell(\widehat{\beta}) - 2\ell(\widehat{\beta}^0)$$

> Why is G always ≥ 0?

Comparing deviances

Full model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1) + \beta_3 Age_i + \beta_4 Age_i \cdot Sex_i$$

Reduced model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Sex_i + \beta_2 \log(Fare_i + 1)$$

$$G = 2\ell(\widehat{\beta}) - 2\ell(\widehat{\beta}^0) = 10.83$$

> If the reduced model is correct, how unusual is $G = 10.83$?

Likelihood ratio test