Lecture 9: Inference with logistic regression models

Ciaran Evans

Recall: the Titanic data

Data on 891 passengers on the *Titanic*. Variables include:

- `Survived`
- `Pclass`
- `Sex`
- `Age`

$$Survived_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 Male_i + \beta_2 Age_i + \beta_3 Class2_i + \beta_4 Class3_i$$

Fitting the model in R

estimated coefficients
estimated standard error of $\hat{\beta}$
test statistic
p-value

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.777	0.401	9.416	4.682e-21
Sexmale	-2.523	0.207	-12.164	4.811e-34
Age	-0.037	0.008	-4.831	1.359e-06
Pclass2	-1.310	0.278	-4.710	2.472e-06
Pclass3	-2.581	0.281	-9.169	4.761e-20

Suppose I want to know whether there is a relation between age and the probability of survival, after accounting for passenger class and sex. What hypotheses would I test?

$H_0 : \beta_2 = 0$ (no relationship)

$H_A : \beta_2 \neq 0$ (some relationship)

Need test statistic, null distribution

Wald tests for single coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.777	0.401	9.416	4.682e-21
Sexmale	-2.523	0.207	-12.164	4.811e-34
Age	-0.037	0.008	-4.831	1.359e-06
Pclass2	-1.310	0.278	-4.710	2.472e-06
Pclass3	-2.581	0.281	-9.169	4.761e-20

$H_0: \beta_j = 0 \qquad H_A: \beta_j \neq 0$

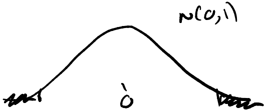
Have: $\hat{\beta}_j$, and $\hat{SE}(\hat{\beta}_j)$

Test statistic: $\dfrac{\hat{\beta}_j - 0}{\hat{SE}(\hat{\beta}_j)}$

Ex: $\dfrac{-0.037 - 0}{0.008} \approx -4.831$

Null distribution: Under H_0, $\hat{\beta}_j \approx N(\beta_j, [\hat{\mathcal{I}}^{-1}(\beta)]_{jj})$

p-value:



Another question

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.777	0.401	9.416	4.682e-21
Sexmale	-2.523	0.207	-12.164	4.811e-34
Age	-0.037	0.008	-4.831	1.359e-06
Pclass2	-1.310	0.278	-4.710	2.472e-06
Pclass3	-2.581	0.281	-9.169	4.761e-20

Suppose I want to know whether there is a relation between *passenger class* and the probability of survival, after accounting for age and sex. What hypotheses would I test?

$H_0: \beta_3 = \beta_4 = 0$ H_A: at least one of $\beta_3, \beta_4 \neq 0$

Nested models

$$Survived_i \sim Bernoulli(p_i)$$

Full model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Male_i + \beta_2 Age_i + \beta_3 Class2_i + \beta_4 Class3_i$$

Hypotheses:

$$H_0 : \beta_3 = \beta_4 = 0 \qquad H_A : \text{ at least one of } \beta_3, \beta_4 \neq 0$$

Reduced model:

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

is the full model a better fit to the data than the reduced model?

Logistic regression model performance

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
Age         -0.036985   0.007656  -4.831 1.36e-06 ***
Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

EDM: $\phi = 1$ for binomial distribution

```
(Dispersion parameter for binomial family taken to be 1)
```

related to model fit

```
    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 657.28
```

(Newton's method)

```
Number of Fisher Scoring iterations: 5
```

Logistic regression model performance

```
    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 657.28
```

- For logistic regression, deviance $= -2 \log \text{Likelihood}$
- Smaller deviances suggest a better fit to the data
- We compare nested models by comparing their deviances

· calculate deviance for full & reduced, compare values

Nested logistic regression models

```
m1 <- glm(Survived ~ as.factor(Pclass) + Sex + Age,
          family = binomial, data = titanic)

m1$deviance
```

```
## [1] 647.2831
```

```
m2 <- glm(Survived ~ Sex + Age,
          family = binomial, data = titanic)
m2$deviance
```

```
## [1] 749.9569
```

H_0: the larger model is not a better fit

Test statistic: $G = 2(\log L_{\text{full}} - \log L_{\text{reduced}})$

$= \text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}$

$G = 749.9569 - 647.2831 \approx 102$

Nested logistic regression models

Distribution: Under H_0, $G \sim \chi^2_q$

- $q =$ difference in number of parameters

e.g. $H_0: \beta_3 = \beta_4 = 0$ $\qquad q = 2$

```
m1 <- glm(Survived ~ as.factor(Pclass) + Sex + Age,
          family = binomial, data = titanic)



m2 <- glm(Survived ~ Sex + Age,
          family = binomial, data = titanic)

pchisq(m2$deviance - m1$deviance, df=2, lower.tail=F)
```

```
## [1] 5.06597e-23
```

$\approx 0 \Rightarrow$ reject H_0!