Lecture 10: Inference with logistic regression models

Recall: the Titanic data

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

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

Logistic regression model

$$\text{Survived}_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \text{Class2}_i + \beta_3 \text{Class3}_i + \beta_4 \text{Male}_i + \beta_5 \text{Age}_i$$

Fitting the model in R

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.77701265	0.401123305	9.416089	4.682044e-21
as.factor(Pclass)2	-1.30979927	0.278065527	-4.710398	2.472337e-06
as.factor(Pclass)3	-2.58062532	0.281442020	-9.169296	4.761161e-20
Sexmale	-2.52278092	0.207390924	-12.164375	4.811152e-34
Age	-0.03698527	0.007655948	-4.830919	1.359041e-06

Suppose I want to know whether there is a relation between age and the probability of survival. What hypotheses would I test?

(after accounting for Pclass & sex)

$H_0: \beta_4 = 0$

$H_A: \beta_4 \neq 0$

test stat: $\dfrac{-0.037 - 0}{0.00766}$

$\dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$

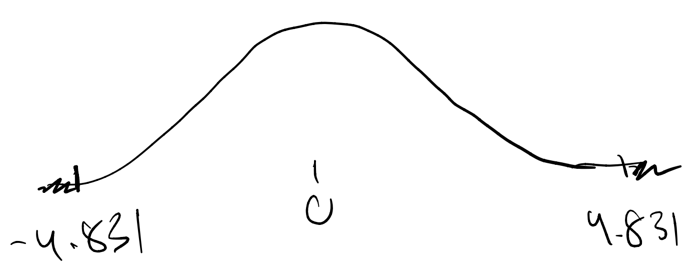
Z-test using $N(0,1)$

test statistic: -4.831

p-value: 1.4×10^{-6}

(fairly strong evidence against H_0)

$N(0,1)$



Wald tests for single coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.77701265	0.401123305	9.416089	4.682044e-21
as.factor(Pclass)2	-1.30979927	0.278065527	-4.710398	2.472337e-06
as.factor(Pclass)3	-2.58062532	0.281442020	-9.169296	4.761161e-20
Sexmale	-2.52278092	0.207390924	-12.164375	4.811152e-34
Age	-0.03698527	0.007655948	-4.830919	1.359041e-06

$\hat{\beta} \approx N(\beta, \mathcal{I}^{-1}(\beta))$ (multivariate Normal distribution)

$0.00766^2 = [\mathcal{I}^{-1}(\beta)]_{5,5}$

$\hat{\beta}_4 \approx N(\beta_4, [\mathcal{I}^{-1}(\beta)]_{5,5})$

$$\mathcal{I}^{-1}(\beta) = \begin{bmatrix} Var(\hat{\beta}_0) & & Cov(\ldots) \\ & Var(\hat{\beta}_1) & \\ Cov(\ldots) & \ddots & Var(\hat{\beta}_4) \end{bmatrix}$$

Another question

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.77701265	0.401123305	9.416089	4.682044e-21
as.factor(Pclass)2	-1.30979927	0.278065527	-4.710398	2.472337e-06
as.factor(Pclass)3	-2.58062532	0.281442020	-9.169296	4.761161e-20
Sexmale	-2.52278092	0.207390924	-12.164375	4.811152e-34
Age	-0.03698527	0.007655948	-4.830919	1.359041e-06

Suppose I want to know whether there is a relation between *passenger class* and the probability of survival. (after accounting for Sex & Age) What hypotheses would I test?

$H_0: \beta_1 = \beta_2 = 0$

$H_A:$ at least one of $\beta_1, \beta_2 \neq 0$

Recall: nested tests for linear regression

Full model : all of variables of interest

Reduced model: removes variables we test (subset of full model)

Linear regression

F-test : $$\frac{(SSE_{reduced} - SSE_{full}) / \Delta df \text{ (full vs. reduced)}}{SSE_{full} / df_{full}}$$

(under H_0)

$\sim F_{q,\, n-p}$

q = # parameters tested

n = # obs

p = # parameters in full model

Logistic regression model performance

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.777013   0.401123   9.416  < 2e-16 ***
as.factor(Pclass)2 -1.309799   0.278066  -4.710 2.47e-06 ***
as.factor(Pclass)3 -2.580625   0.281442  -9.169  < 2e-16 ***
Sexmale            -2.522781   0.207391 -12.164  < 2e-16 ***
Age                -0.036985   0.007656  -4.831 1.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
```

deviance (for logistic regression) = $-2 \log L$

Compare deviance for nested models

Nested logistic regression models

full model

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

m1$deviance
```

```
[1] 647.2831
```

reduced model

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

```
[1] 749.9569
```

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

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

(always ≥ 0)

Preview: likelihood ratio test

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

$$= 2\log\left(\frac{L_{full}}{L_{reduced}}\right)$$

Under H_0: (and other assumptions...) $\quad G \sim \chi^2_q \quad q = \#$ parameters tested

Preview: likelihood ratio test

```r
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
```

(strong evidence against H_0)