Simulation and parametric bootstrap

Reminders:

- HW 2 due Friday (noon)
- Study session Thursday, 7-8pm Kirby 120

HW 2, Q2 : Titanic data

Response: Survival

Explanatory variables: Pclass, Sex, Age

Pclass: 1, 2, 3

Sex: male (1), female (0)

Age: numeric variable

Issue: want Pclass to be a categorical variable in our model!

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Age}_i + \beta_3 \text{SecondClass}_i + \beta_4 \text{ThirdClass}_i$$

Sex_i: $=1$ male, $=0$ female (indicator variable)

SecondClass_i: $=1$ if Pclass $=2$, $=0$ else

ThirdClass_i: $=1$ if Pclass $=3$, $=0$ else

(indicator variables)

β_3: difference in log odds between first & second class passengers, holding Sex & Age fixed

e^{β_3}: change in odds

β_4: change in log odds between first & third class passengers

glm(Survived ~ Sex + Age + as.factor(Pclass), ...)

forces Pclass to be categorical

Recap: Steps in hypothesis testing

$$Admit_i \sim Bernoulli(\pi_i) \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{GRE}_i$$

- Specify hypotheses

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$

- Calculate a test statistic

$$G = \text{deviance for reduced model - deviance for full model}$$

- Calculate a p-value

Recap: calculating p-values

> What *is* a p-value?

p-value: How unusual the data is *if* H_0 is true. e.g.,

random variables (varies from sample to sample)

$$P(G \geq 13.92 | H_0)$$

$\Rightarrow$ G has a distribution

> How do we calculate a p-value?

Compare the observed test statistic to the null distribution

(distribution of G if H_0 is true)

> How do we get the null distribution?

Plan

Want: distribution of G if H_0 is true

Plan:

1) Simulate data for which H_0 is true
2) Calculate G on simulated data
3) Repeat many times to approximate null distribution

Exploring the null distribution with simulation

- Want to know how G (drop in deviance) behaves if H_0 is true
- So, need data for which we *know* H_0 is true!

see HW 1

```
x <- runif(1000, 0, 5)
p <- exp(-3 + 0*x)/(1 + exp(-3 + 0*x))
y <- rbinom(1000, 1, p)

m1 <- glm(y ~ x, family = binomial)
summary(m1)
```

$Y_i \sim \text{Bernoulli}(\pi_i)$

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -3 + 0X_i$

H_0 is $\beta_1 = 0$

```
...
##
##     Null deviance: 367.04  on 999  degrees of freedom
## Residual deviance: 367.00  on 998  degrees of freedom
...
```

$G = 367.04 - 367.00 = 0.04$

Exploring the null distribution with simulation

- Simulating one set of data only gives us one statistic under H_0
- We need to simulate many datasets to explore the full distribution

> How can we efficiently create *many* simulated datasets?

for loop!

Idea:
Do the following many times:
- Simulate data
- calculate test statistic
- Store result

pseudo-code

```
for (i in ....){
    - simulate
    - test statistic
    - store result
}
```

Exploring the null distribution with simulation

- Simulating one set of data only gives us one statistic under H_0
- We need to simulate many datasets to explore the full distribution

store results of simulation

```
null_statistics <- c()
nsim <- 500
for(i in 1:nsim){
  x <- runif(1000, 0, 5)
  p <- exp(-3 + 0*x)/(1 + exp(-3 + 0*x))
  y <- rbinom(1000, 1, p)
  m1 <- glm(y ~ x, family = binomial)
  null_statistics[i] <- m1$null.deviance -
    m1$deviance
}
```

repeat 500 times

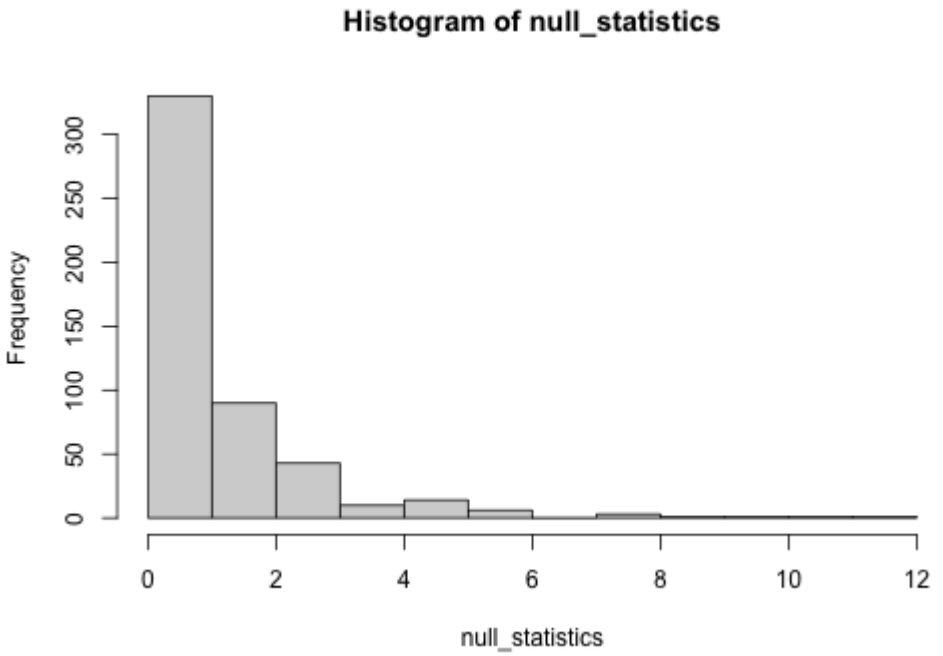
simulate data under H_0

store result

test statistic

Exploring the null distribution with simulation

```
hist(null_statistics)
```



1
13.92

Are there any issues with this approach to approximating the null distribution?

The problem with simulation

Goal: Want to know how unusual the observed test statistic ($G = 13.92$) for the grad admissions data is if H_0 is true (there is no relationship between GRE score and admission)

Potential problem: Our simulated data looks nothing like the grad admissions data

Strategy:

1) Fit the reduced model on the grad admissions data, pretend the reduced model is correct (i.e. H_0 is true)

repeat many times!

2) Simulate data from reduced model

3) Calculate a test statistic from simulated data

Step 1

Step 1: Pretend the reduced model is correct:

$$Admit_i \sim Bernoulli(\pi_i)$$

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

> How do I fit this model in R?

Step 1

Step 1: Pretend the reduced model is correct:

$$Admit_i \sim Bernoulli(\pi_i)$$

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

intercept only model 1

```
m1 <- glm(admit ~ 1, family = binomial,
          data = grad_app)
summary(m1)
```

```
...
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7653     0.1074  -7.125 1.04e-12 ***
## ---
...
```

$\hat{\beta}_0 = -0.77$

Step 2

Step 2: Simulate from the reduced model

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -0.77$$

> What should I change in my code to simulate from this reduced model?

```
x <- runif(1000, 0, 5)
p <- exp(-3 + 0*x)/(1 + exp(-3 + 0*x))
y <- rbinom(1000, 1, p)
```

Step 2

Step 2: Simulate from the reduced model

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -0.77$$

```
x <- grad_app$gre
p <- exp(-0.77 + 0*x)/(1 + exp(-0.77 + 0*x))
y <- rbinom(length(x), 1, p)
```

Step 3

Step 3: Calculate a test statistic for the simulated data

```
x <- grad_app$gre
p <- exp(-0.77 + 0*x)/(1 + exp(-0.77 + 0*x))
y <- rbinom(length(x), 1, p)

m <- glm(y ~ x, family = binomial)
m$null.deviance - m$deviance
```

```
## [1] 0.01527109
```

> In the observed data, $G = 13.92$. In this simulated data, $G = 0.015$. Does the observed test statistic seem unusual, if H_0 were true?

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_7.html

Class activity

```
null_statistics <- c()
nsim <- 500
for(i in 1:nsim){
  x <- runif(1000, 0, 5)
  p <- exp(-3 + 0*x)/(1 + exp(-3 + 0*x))
  y <- rbinom(1000, 1, p)
  m1 <- glm(y ~ x, family = binomial)

  null_statistics[i] <- m1$null.deviance -
    m1$deviance
}
```

> How do I adapt this code to simulate many times from the reduced model?

Class activity

output

```
null_statistics <- c()
nsim <- 500
for(i in 1:nsim){
  x <- grad_app$gre
  p <- exp(-0.77 + 0*x)/(1 + exp(-0.77 + 0*x))
  y <- rbinom(length(x), 1, p)
  m1 <- glm(y ~ x, family = binomial)

  null_statistics[i] <- m1$null.deviance -
    m1$deviance
}
```

sequence

1, 2, 3, ..., 500

for loop:

output: what we want to calculate

sequence: what we iterate over

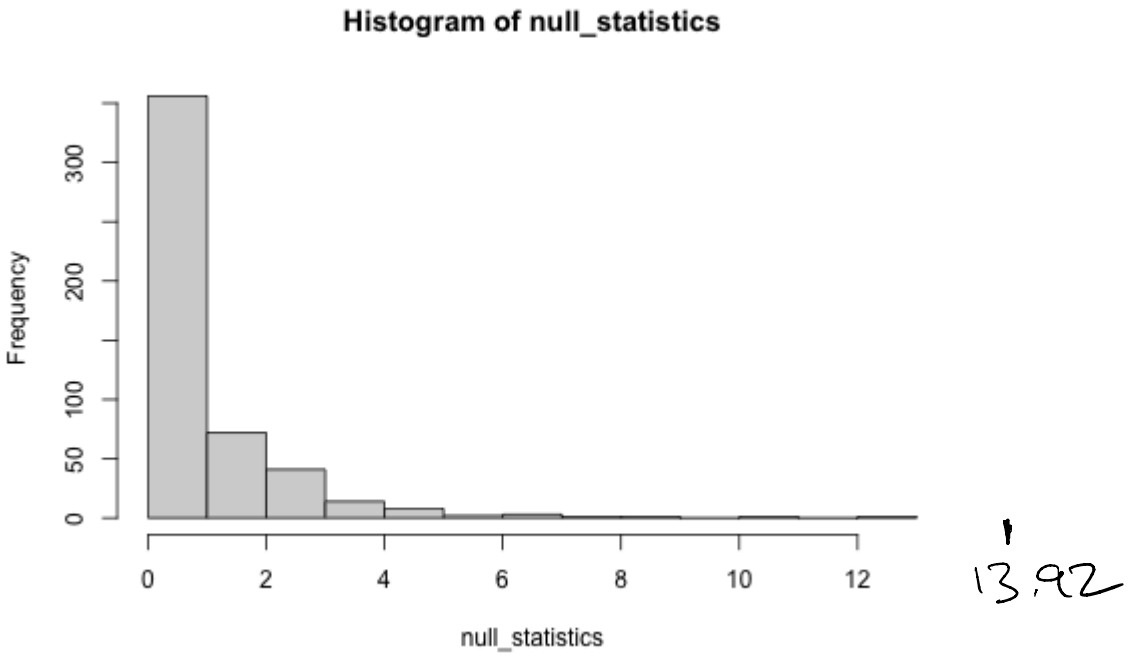
body: calculation for each iteration

Class activity

$P(G \geq 13.92 \mid H_0)$

$p \approx$ fraction of simulations in which $G \geq 13.92$

```
hist(null_statistics)
```



Compared to the test statistics simulated from the reduced model, how unusual is the observed test statistic of $G = 13.92$?

Approximating a p-value

p-value: How unusual the data is *if* H_0 is true. e.g.,

$$P(G \geq 13.92 | H_0)$$

We can approximate this probability using the test statistics simulated from the reduced model:

```
mean(null_statistics >= 13.92)
```

```
## [1] 0
```

p-value ≈ 0

Parametric bootstrapping