Lab 3

Alexey Arkhangelskiy

08/08/2022

(Faraway 2016) The Ohio data are from 536 children as part of a study on the effects of air pollution. Children were in the study for 4 years from ages 7 to 10. The response was whether they wheezed or not. The variables are:

`resp`: an indicator of wheeze status (1=yes, 0=no) `id`: an identifier for the child `age`: 7 yrs=-2, 8 yrs=-1, 9 yrs=0, 10 yrs=1 `smoke`: an indicator for if their mother smoked

a) Fit an appropriate generalized linear mixed model, pick an appropriate "grouping" variable, and input the other variables as fixed effects.

We will use logistic regression model since the response variable 'resp' is binary. The model will include fixed effects 'age' and 'smoke' because both might have an impact on the wheeze status. We will treat 'id' as a grouping variable (random effect) to account for the correlation between measurements from the same child across different ages. Each child (identified by id) would have a random intercept, capturing the child's inherent susceptibility to wheezing that's not captured by the fixed effects.

The model is:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \mu_\alpha + \alpha_{j[i]} + \beta_1 x_{\text{age}} + \beta_2 x_{\text{smoke}} \\ \alpha_j &\sim N(0, \sigma_\alpha^2) \end{aligned}$$

where

- $y_i = 1$ if the child wheezed, 0 otherwise
- α_j is the effect of child j.
- $x_{\text{age},i}$ is the age of the child at observation i, $x_{\text{smoke},i}$ is the smoking status of the child's mother.
- $\mu_\alpha, \beta_1, \beta_2 \in (-\infty, \infty)$ are parameters to be estimated
- $\sigma \in [0, \infty)$ is a parameter to be estimated

Load libraries

```
data(ohio,  package="faraway")

library(dplyr)
library(ggplot2)
library(lme4)
```

We fit this model using the following code

b) Using the parametric bootstrap, perform a likelihood ratio test to compare your model from a) to a random intercept-only model.

```
no_fixed_effects_model <- glmer(data=ohio,
                  formula = resp ~ 1 + (1|id),
                  family = binomial(link = "logit"))
test_stat <- 2*(as.numeric(logLik(logistic_mlm))-as.numeric(logLik(no_fixed_effects_model)))
test_stat
```

```
## [1] 8.954108
```

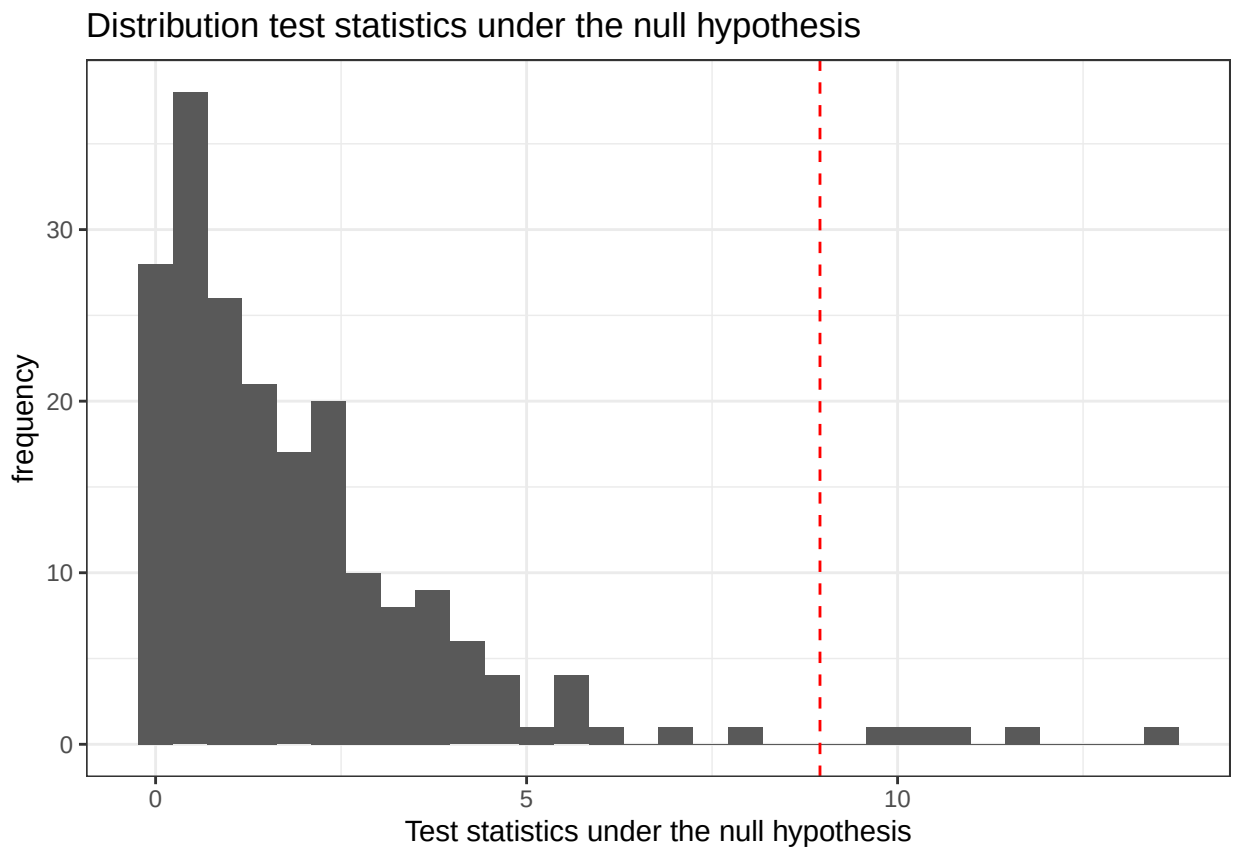
Perform parametric bootstrap:

```
n_sims=200
null_test_stats <- vector(length = n_sims)
for (i in 1:n_sims){
  y <- unlist(simulate(no_fixed_effects_model))
  bnull <- glmer(data = ohio, y ~  1 + (1|id), family = binomial, nAGQ=25)
  balt <- glmer(data = ohio, y ~  1 + smoke + age + (1|id), family =binomial, nAGQ=25)
  null_test_stats[i] <- 2*(as.numeric(logLik(balt))-as.numeric(logLik(bnull)))
}
```

c) Plot the empirical distribution of the test statistic under the null hypothesis. On your graph, indicate where your test statistic falls.

A histogram displaying the distribution of test statistics from the simulations under the null hypothesis is presented below. The test statistic from our data is marked by a red dashed line.

```
null_test_stats = data.frame(null_test_stats)
ggplot(data = null_test_stats, aes(x=null_test_stats)) +
  geom_histogram(bins = 30)+
  labs(x = "Test statistics under the null hypothesis",
       title = "Distribution test statistics under the null hypothesis",
       y = "frequency")+
  theme_bw()+
  geom_vline(xintercept = test_stat, col = "red", lty = 2)
```



d) Compute the empirical p-value. Draw a conclusion from your test.

We compute the empirical p-value using

```
emp_pval <- mean(null_test_stats$null_test_stats > test_stat)
emp_pval
```

```
## [1] 0.025
```

Only 0.5% of the bootstrapped samples produced a test statistic as extreme as (or more extreme than) your observed test_stat under the null hypothesis. This is strong evidence against the null hypothesis, suggesting that the complex model with fixed effects (logistic_mlm) fits the data significantly better than the simpler no_fixed_effects_model.

The p-values given by glmer are shown below:

```
summary(logistic_mlm)$coefficients
```

```
##               Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -3.3739541 0.27494404 -12.271421 1.289697e-34
## age         -0.1767645 0.06796623  -2.600770 9.301486e-03
## smoke        0.4147806 0.28703099   1.445072 1.484375e-01
```

Even though the coefficient for smoke is not significant by itself, the combination of age and smoke together provides a better fit than a model without them. This is reflected in the significant p-value from the LRT. Thus, the LRT indicates that including both predictors (even if one isn't individually significant) yields a better model than one with neither predictor.