STAT 512 Homework 5

Kathleen Wendt

03/24/2020

Part 1: Heart data

For this problem use the data described in Ott and Longnecker Example 12.22 (p 664 in the 7th edition). The data are available from Canvas as "CKheart.csv". Read the description of the data in the book. You can use the output in the book to check your own R calculations.

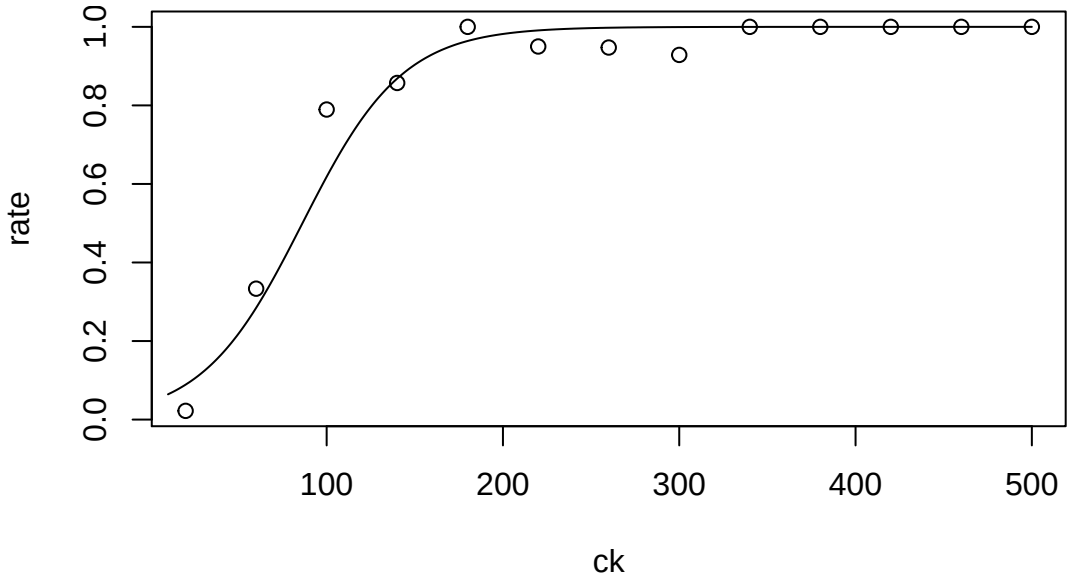
Question A: Logistic regression

Use `glm()` to fit a logistic regression model that estimates the probability of a heart attack as a function of CK value. Include the Coefficients table in your assignment.

term	estimate	std.error	statistic	p.value
(Intercept)	-3.0283596	0.3669773	-8.252171	0
ck	0.0351044	0.0040811	8.601712	0

Question B: Plot

Construct a plot of the data with the fitted logistic regression curve overlaid. Include the plot in your assignment.



Question C: Odds ratio of CK

Give an estimate of the odds ratio corresponding to CK and an approximate 95% confidence interval.

The odds ratio estimate is 1.036. The corresponding confidence interval is (1.028, 1.045).

Question D: CK odds

Give a one-sentence description of the odds of heart attack among those with a given level of CK, compared to the odds of a heart attack among those with a level of CK ten points higher. (4 pts)

A 10-point increase in CK is associated with 1.4x increased odds of a heart attack.

Question E: Psuedo R-squared

Calculate McFadden's Pseudo R^2 for the model.

McFadden's Pseudo R^2 is 0.856.

Question F: LD-90

Give an estimate of the CK level at which doctors would be 90% sure that a subject has had a heart attack.

At a CK level of 148.86, doctors would be 90% confident that the subject had a heart attack.

Part 2: Birth weight

An observational study was done to investigate risk factors associated with low infant birth weight. Data from 189 (singleton) pregnancies were collected at Baystate Medical Center, Springfield, MA during 1986. The response variable was low (1 if birth weight was less than 2.5 kg, 0 otherwise). The predictor variables included: `age` (mother's age in years), `mwt` (mother's weight in pounds prior to pregnancy), `race` (mother's race, 1= white, 2=black, 3=other) and smoke (1=mother smoked during pregnancy, 0 otherwise). The data is available from Canvas as "birthweight.csv".

Important note: Be sure to define race and smoke as factors!

Question A: Race

To examine the relationship between low vs race: calculate the proportion of births resulting in low birth=weight for each race category and present the p-value from a chi-square test. (4 pts)

Based on the Chi-square test, there is no difference in the proportions of low vs. normal birth weight based on mother race, $p = 0.08189 > \alpha = 0.05$.

Question B: Smoking

To examine the relationship between low vs smoke: calculate the proportion of births resulting in low birth-weight for each smoke category and present the p-value from a chi-square test. (4 pts)

Based on the Chi-square test, there is a difference in the proportions of low vs. normal birth weight based on mother smoking, $p = 0.03958 < \alpha = 0.05$.

Question C: Logistic regression

Run a logistic regression with smoke as the only predictor variable. Calculate the emmeans using `type = "response"` for each smoke group (copy/paste the results to your assignment). *Note: these should match your simple proportions from part B.* (4 pts)

```
##  smoke  prob     SE  df asymp.LCL asymp.UCL
##  0     0.252 0.0405 Inf     0.181     0.339
##  1     0.405 0.0571 Inf     0.300     0.520
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

Question D: AIC model selection

Now consider all 4 predictors (age, mwt, race, smoke). Using best subsets selection with AIC criteria, which variables are included in the final model? Include the Coefficients table and Type3 Anova table in your assignment. (4 pts)

NOTE: Use the selected model from the previous question for all further questions!

Based on AIC best subsets selection, the "best" model includes maternal weight, race, and smoking as predictors of low birth weight.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1092208	0.8821088	-0.1238179	0.9014595
mwt	-0.0132595	0.0063102	-2.1012808	0.0356163
race2	1.2900945	0.5108750	2.5252644	0.0115611
race3	0.9705149	0.4122349	2.3542766	0.0185588
smoke1	1.0600059	0.3783229	2.8018552	0.0050810

term	statistic	df	p.value
mwt	4.960053	1	0.0259394
race	9.325993	2	0.0094381
smoke	8.244433	1	0.0040877

Question E: Smokers vs. non-smokers

Based on the model selected above, give the estimated odds ratio and corresponding 95% CI for Smokers vs Non-Smokers (smoke 1 vs 0).

The estimated odds ratio for low birth weight in maternal smokers vs. non-smokers is 2.886. The corresponding confidence interval is (1.395, 6.198).

Question: F: New smoke emmeans

Calculate the emmeans using `type = "response"` for each smoke group (copy/paste the results to your assignment). *Note that these values are different from what you found in part C because of the additional variables included in the model.*

```
## $emmeans
##  smoke  prob     SE  df asymp.LCL asymp.UCL
##  0     0.254 0.0467 Inf     0.174     0.356
##  1     0.496 0.0710 Inf     0.360     0.632
```

```
##
## Results are averaged over the levels of: race
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
##
## $contrasts
##  contrast odds.ratio    SE  df z.ratio p.value
##  0 / 1         0.346 0.131 Inf -2.802  0.0051
##
## Results are averaged over the levels of: race
## Tests are performed on the log odds ratio scale
```

Question G: Compare by maternal race

Run Tukey adjusted pairwise comparisons for race. Discuss your findings. (4 pts)

White mothers (Level 1) have significantly lower odds of having an infant with a low birth weight, compared to black mothers (Level 2), $p = 0.031$, and compared to mothers of other races (Level 3), $p = 0.049$.

Question H: Hoslem Test

Give the p-value corresponding to the Hosmer-Lemeshow test. Use `hoslem.test()` from the `ResourceSelection` package with `g = 10` groups. Based on this test, is there evidence of lack of fit?

Based on the Hosmer-Lemeshow test, $p = 0.4997$, we fail to reject the null hypothesis; there is no evidence for lack of fit.

Appendix

```
# load packages
library(tidyverse)
library(janitor)
library(broom)
library(kableExtra)
library(MASS)
library(emmeans)
library(MuMIn)
library(car)
library(ResourceSelection)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# 1. read and prepare heart data
heart_data <- readr::read_csv("data/CKheart.csv") %>%
  janitor::clean_names()
# 1a. build logistic regression model of heart attack probability
heart_logreg <- glm(cbind(with_ha, without_ha) ~ ck,
                    family = binomial(link = "logit"),
                    data = heart_data)
kableExtra::kable(broom::tidy(heart_logreg))
# 1b. create rate variable for heart data
heart_data <- heart_data %>%
  dplyr::mutate(rate = with_ha/(with_ha + without_ha))
# 1b. create new data sequence
heart_seq <- seq(10, 500, 1)
# 1b. predict new data
p_hat <- predict(heart_logreg, list(ck = heart_seq), type = "response")
# 1b. plot data
plot(rate ~ ck, data = heart_data)
lines(p_hat ~ heart_seq)
# 1c. heart ck odds ratio
exp(heart_logreg$coef)
# 1c. heart ck ci
exp(confint(heart_logreg))
# 1d. exp ck coeff x10
exp(0.0351*10)
# 1e. calculate pseudo r2
heart_null <- glm(cbind(with_ha, without_ha) ~ 1,
                  family = binomial(link = "logit"),
                  data = heart_data)
1 - logLik(heart_logreg)/logLik(heart_null)
# 1f. create probability sequence for LD
heart_prob <- seq(0.1, 0.9, 0.05)
# 1f. calculate LDs
MASS::dose.p(heart_logreg, cf = 1:2, p = heart_prob)
# 2. read and prepare birth weight data
```

```r
birth_data <- readr::read_csv("data/birthweight.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(low = as.factor(low),
                race = as.factor(race),
                smoke = as.factor(smoke))
# 2a. build maternal race/low birth weight table
race_weight <- table(birth_data$race, birth_data$low)
# 2a. calculate proportions
prop.table(race_weight, 1)
# 2b. conduct chi square test
chisq.test(race_weight)
# 2a. build maternal smoking/low birth weight table
smoke_weight <- table(birth_data$smoke, birth_data$low)
# 2a. calculate proportions
prop.table(smoke_weight, 1)
# 2b. conduct chi square test
chisq.test(smoke_weight)
# 2c. build logistic regression with smoke
birth_logreg <- glm(low ~ smoke,
                    family = binomial(link = "logit"),
                    data = birth_data)
# 2c. extract emmeans based on smoke log reg
emmeans::emmeans(birth_logreg, ~ smoke, type = "response")
# 2d. build full model with all possible predictors
birth_full <- glm(low ~ .,
                  family = binomial(link = "logit"),
                  data = birth_data)
# 2d. extract aic for all possible models
options(na.action = "na.fail")
MuMIn::dredge(birth_full, rank = "AIC")
# 2d. build final model based on aic
birth_final <- glm(low ~ mwt + race + smoke,
                   family = binomial(link = "logit"),
                   data = birth_data)
# 2d. show coefficients table
kableExtra::kable(broom::tidy(birth_final))
# 2d. show Anova type 3 table
kableExtra::kable(broom::tidy(car::Anova(birth_final, type = 3)))
# 2e. estimated OR for smokers vs. non
exp(birth_final$coef)
# 2e. CI for OR smokers vs. non
exp(confint(birth_final))
# 2f. extract emmeans based on final low birth weight log reg
emmeans::emmeans(birth_final, pairwise ~ smoke, type = "response")
# 2g. Tukey-adjusted pairwise comparisons for race
emmeans::emmeans(birth_final, pairwise ~ race, type = "response")
# 2h. hosmer-lemeshow test for goodness of fit
ResourceSelection::hoslem.test(birth_final$y, fitted(birth_final), g = 10)
```