

Example6__3

Kevin Cummiskey

November 25, 2019

Section 6.3 Multiple Logistic Regression

Review

In the article “The incidence of thyroid disorders in the community: A twenty-year follow-up of the Wickham survey” by Vanderpump et al., 443 of the 582 smokers and 502 of the 732 nonsmokers were still alive at the 20 year follow-up.

Let Y_i be 1 if subject i survives 20 years and 0 if subject i does not survive such that $Y_i \sim \text{Bernoulli}(\pi_i)$. Consider the following model:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i$$

where x_i equals 1 if subject i is a smoker and 0 if subject i is not a smoker.

Interpret the following quantities and provide estimates based on the data.

- β_0
- $\exp(\beta_0)$
- $\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
- β_1
- $\exp(\beta_1)$
- $\beta_0 + \beta_1$
- $\exp(\beta_0 + \beta_1)$
- $\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$

Multiple Logistic Regression

Why is it important to adjust for age in this analysis?

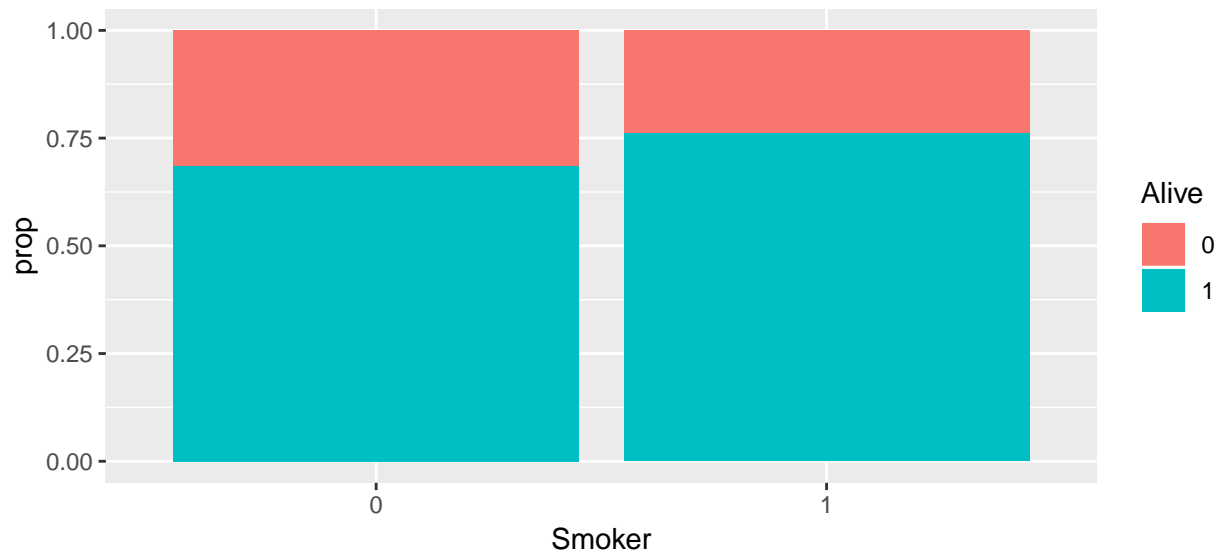
Here is a plot of the unadjusted association between smoking and 20-year survival.

```
smoke = read.table(file = "smoke.csv", header = T, sep = ",")
smoke$Alive = factor(smoke$Alive)
smoke$Smoker = factor(smoke$Smoker)

summary = smoke %>% group_by(Smoker, Alive) %>%
  count() %>% group_by(Smoker) %>% mutate(prop = n/sum(n))
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
summary %>% ggplot(aes(x = Smoker, y = prop, fill = Alive)) +
  geom_col()
```



Let's fit the logistic regression model in the review. Confirm the model produces the same odds ratio as the sample statistics.

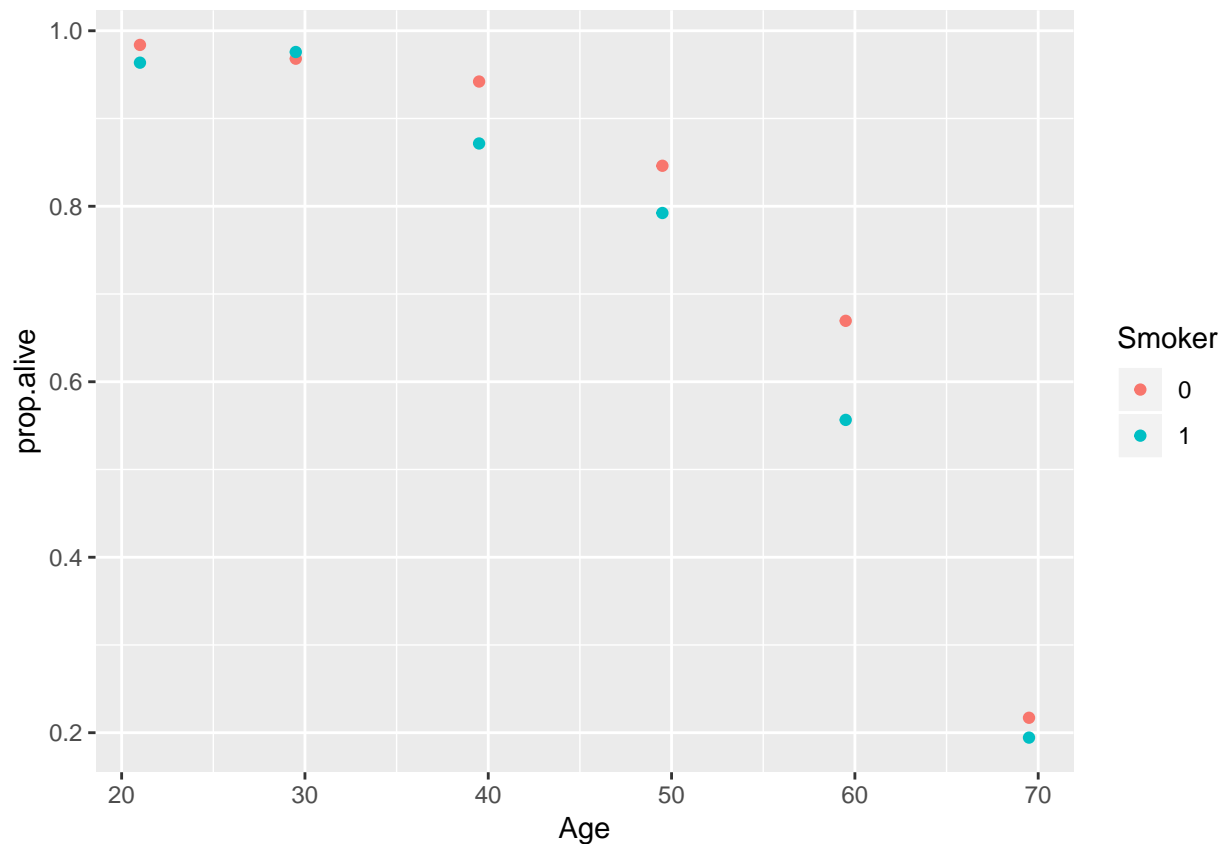
```
model.crude = glm(Alive ~ Smoker, data = smoke, family = "binomial")
summary(model.crude)
```

```
##
## Call:
## glm(formula = Alive ~ Smoker, family = "binomial", data = smoke)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6923  -1.5216   0.7388   0.8685   0.8685
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.78052    0.07962   9.803  < 2e-16 ***
## Smoker1      0.37858    0.12566   3.013  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1560.3  on 1313  degrees of freedom
## Residual deviance: 1551.1  on 1312  degrees of freedom
## AIC: 1555.1
##
## Number of Fisher Scoring iterations: 4
```

Let's look at the survival of smokers versus nonsmokers with the same levels of age.

```
summary2 = smoke %>% group_by(Age, Smoker, Alive) %>%
  count() %>% group_by(Age, Smoker) %>% mutate(prop.alive = n/sum(n)) %>%
  filter(Alive == 1) %>% select(Age, Smoker, prop.alive)
```

```
summary2 %>% ggplot(aes(x = Age, y = prop.alive, color = Smoker)) +  
  geom_point()
```



Does it appear smokers or nonsmokers have a better odds of surviving for 20 years?

Let's fit the following model where $Y_i \sim \text{Bernoulli}(\pi_i)$.

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$

where $x_{1,i}$ is smoker (1) or nonsmoker (0) and $x_{2,i}$ is the subject's age at the start of the study.

```
#remove older folks - no survivors 75 years or older.
```

```
smoke_cleaned = smoke %>% filter(Age < 70)
```

```
model.age_adjusted = glm(Alive ~ Smoker + Age, data = smoke_cleaned, family = "binomial")  
summary(model.age_adjusted)
```

```
##
```

```
## Call:
```

```
## glm(formula = Alive ~ Smoker + Age, family = "binomial", data = smoke_cleaned)
```

```
##
```

```
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -3.1026  0.1277  0.2396  0.6351  1.6675
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.2411      0.4502  16.085 <2e-16 ***
## Smoker1      -0.2823      0.1677  -1.683  0.0923 .
## Age          -0.1160      0.0075 -15.467 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1352.02  on 1236  degrees of freedom
## Residual deviance:  935.54  on 1234  degrees of freedom
## AIC: 941.54
##
## Number of Fisher Scoring iterations: 6
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.5.2
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##      recode
## The following object is masked from 'package:purrr':
##
##      some
```

```
Anova(model.age_adjusted, type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: Alive
##           LR Chisq Df Pr(>Chisq)
## Smoker      2.86  1   0.09101 .
## Age        415.22  1   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does the adjusted odds ratio compare to the unadjusted?