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?

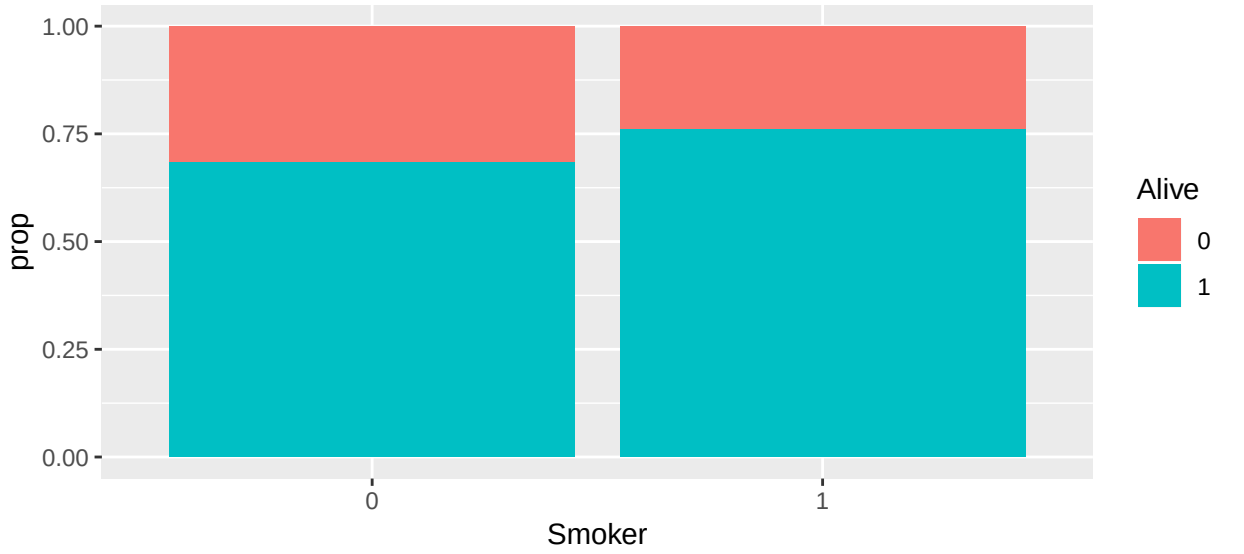
Unadjusted model

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))

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)
```

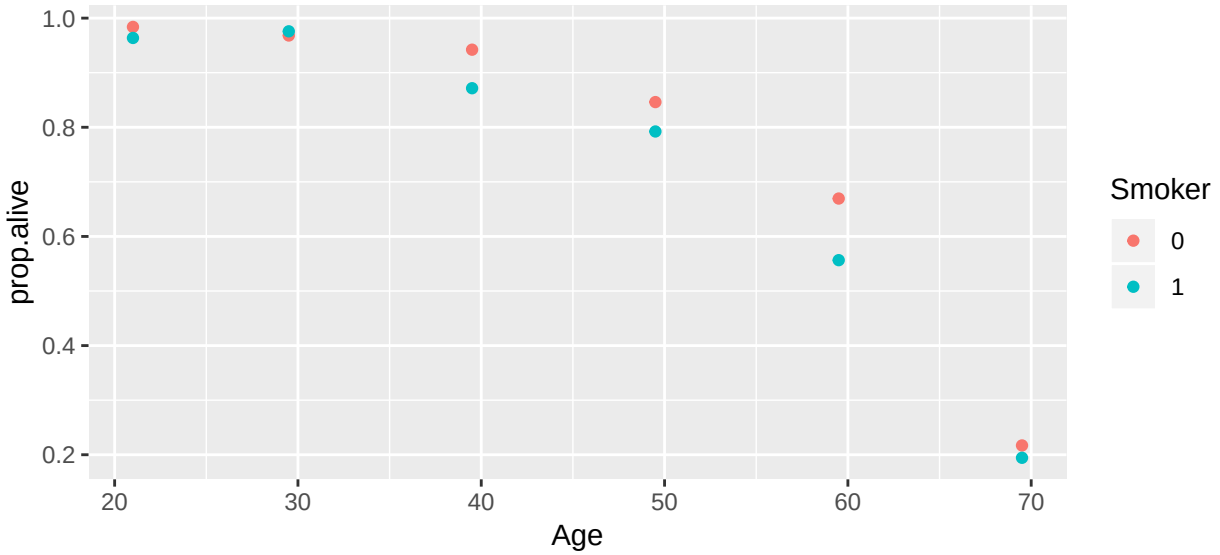
```
##
## Call:
## glm(formula = Alive ~ Smoker, family = "binomial", data = smoke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6923  -1.5216   0.7388   0.8685   0.8685
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.78052    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
```

Model adjusting for age

Describe a plot you could make to determine if there is a relationship between smoking and survival after adjusting for age.

```r
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.

```r
#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)
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?

How can we assess the fit of a logistic regression model?

Model Fit 1

```
smoke_cleaned = smoke_cleaned %>% left_join(summary2)
```

```
## Joining, by = c("Age", "Smoker")
```

```
smoke_cleaned = smoke_cleaned %>% mutate(pred.alive = predict(model.age_adjusted, smoke_cleaned, type =
cor(smoke_cleaned$prop.alive, smoke_cleaned$pred.alive)
```

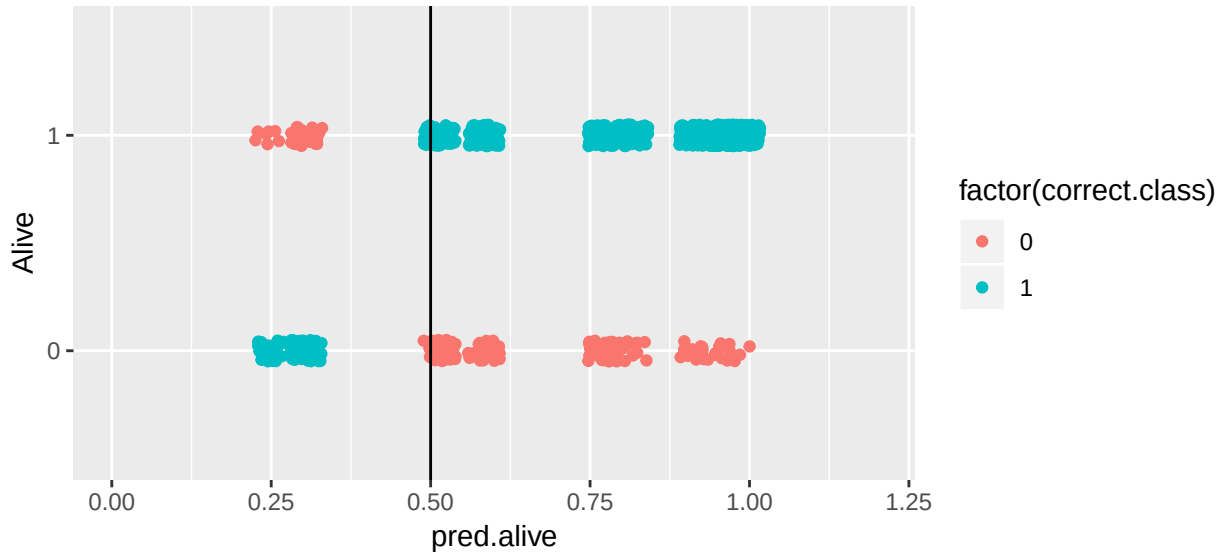
```
## [1] 0.983839
```

Model Fit 2

```
smoke_cleaned = smoke_cleaned %>%
  mutate(Alive.pred = factor(ifelse(pred.alive > 0.5,
```

```
                                          1, 0)))
smoke_cleaned = smoke_cleaned %>%
  mutate(correct.class = ifelse(Alive.pred == Alive,
                                1,0))

smoke_cleaned %>% ggplot(aes(x = pred.alive,
                             y = Alive,
                             color = factor(correct.class))) +
  geom_jitter(height = 0.05, width = 0.025) + xlim(0,1.2) +
  geom_vline(xintercept = 0.5)
```




```
library(caret)
confusionMatrix(data = smoke_cleaned$Alive.pred,
                reference = smoke_cleaned$Alive)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 130  35
##          1 162 910
##
##                Accuracy : 0.8407
##                  95% CI : (0.8191, 0.8607)
##     No Information Rate : 0.7639
##     P-Value [Acc > NIR] : 2.003e-11
##
##                   Kappa : 0.4804
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4452
##             Specificity : 0.9630
##          Pos Pred Value : 0.7879
##          Neg Pred Value : 0.8489
```

```
##              Prevalence : 0.2361
##          Detection Rate : 0.1051
##    Detection Prevalence : 0.1334
##       Balanced Accuracy : 0.7041
##
##        'Positive' Class : 0
##
```

Model with interaction

What research question(s) can we answer by fitting the model with an interaction?

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

```
##
## Call:
## glm(formula = Alive ~ Smoker * Age, family = "binomial", data = smoke_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2077   0.1082   0.2718   0.6126   1.5850
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.76239    0.62586  12.403   <2e-16 ***
## Smoker1     -1.38524    0.85556  -1.619    0.105
## Age         -0.12494    0.01050 -11.904   <2e-16 ***
## Smoker1:Age  0.01994    0.01511   1.319    0.187
## ---
## 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:  933.81  on 1233  degrees of freedom
## AIC: 941.81
##
## Number of Fisher Scoring iterations: 6
```

What would you conclude from this analysis?