

Multiple Logistic Regression

SDS 291

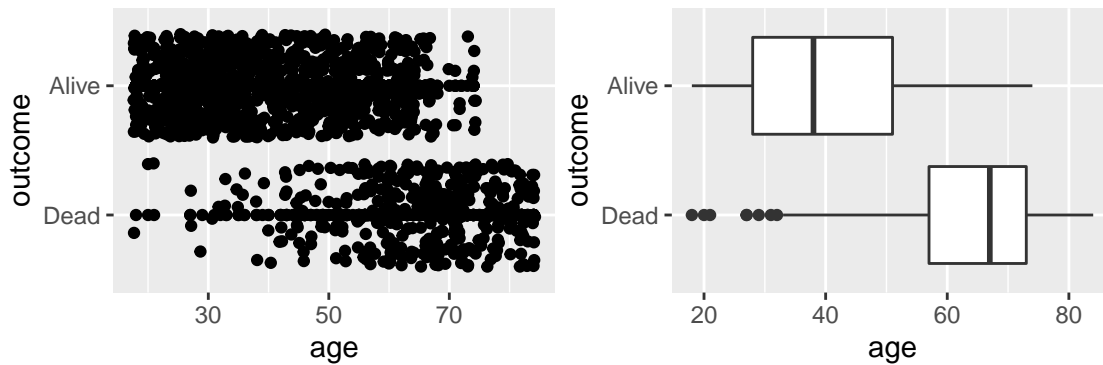
April 15, 2020

We're going to work with the `Whickham` data contains observations about women, and whether they were alive 20 years after their initial observation (`outcome` is a 2 level factor variable - Alive/Dead). You can learn more about these data from the `mosaicData` help feature if you'd like.

Specifically, we're interested in: the association of age, smoking status (smoker), and 20-year survival (outcome: alive (success), dead (reference / failure)). Bring in the relevant packages and the data (below, from the `mosaic` package).

```
library(mosaic)
library(tidyverse)
data("Whickham")
Whickham$outcome<-relevel(Whickham$outcome, ref="Dead")
```

Age and Outcome



```
m0<-glm(outcome~age, data=Whickham, family=binomial)
summary(m0)
```

```
##
## Call:
## glm(formula = outcome ~ age, family = binomial, data = Whickham)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2296  -0.4277   0.2293   0.5538   1.8953
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.403126   0.403522  18.35  <2e-16 ***
## age        -0.121861   0.006941 -17.56  <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: 1560.32  on 1313  degrees of freedom
## Residual deviance:  946.51  on 1312  degrees of freedom
## AIC: 950.51
##
## Number of Fisher Scoring iterations: 6
```

1. Write the fitted equation and interpret the results (in these units) in light of the question. Be sure to comment on the magnitude and direction of the association.
2. Based on this model, what is the probability that a 60 year old was alive 20 years after the initial survey?

Smoking Status and Outcome (Alive)

```
Whickham$smoker<-factor(Whickham$smoker, levels=c("Yes", "No"))
Whickham$outcome<-factor(Whickham$outcome, levels=c("Alive", "Dead"))
tally(~ smoker + outcome, margins=FALSE, data=Whickham)
```

```
##           outcome
## smoker Alive Dead
##   Yes   443  139
##   No   502  230
```

3. Calculate the Odds Ratio of non-smokers being alive in 20 years compared to smokers from the table above.

```
Whickham$smoker<-relevel(Whickham$smoker, ref= "No")
Whickham$outcome<-relevel(Whickham$outcome, ref= "Dead")
m1<-glm(outcome~smoker, data=Whickham, family=binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = outcome ~ smoker, family = binomial, data = Whickham)
##
## 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 ***
## smokerYes    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
```

4. Show that you can calculate the coefficient for smoking status from your regression model as you did in #3.

5. Based on your model, what's the probability that a smoker was alive 20 years later?

6. Based on what you know about the risk of death for age and smoking status, do these results make sense? Explain your answer.

Multiple Logistic Regression

```
m2<-glm(outcome~age+smoker, data=Whickham, family=binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = outcome ~ age + smoker, family = binomial, data = Whickham)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2795  -0.4381   0.2228   0.5458   1.9581
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.599221   0.441231  17.223  <2e-16 ***
## age         -0.123683   0.007177 -17.233  <2e-16 ***
## smokerYes    -0.204699   0.168422  -1.215    0.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1560.32  on 1313  degrees of freedom
## Residual deviance:  945.02  on 1311  degrees of freedom
## AIC: 951.02
##
## Number of Fisher Scoring iterations: 6
```

7. What is the odds ratio for smokers compared to non-smokers in this model? Interpret in a sentence in the context of this real-world problem.

8. What is the probability of a 60 year old non-smoker being alive 20 years later?

9. What is the probability of a 40 year old smoker being alive 20 years later?

10. What does this model help us to understand about our simple logistic regression estimates above?

Optional - Interaction Term

11. What would an interaction term between age and smoking status do in this model? How would an interaction term affect the OR for age?

12. How do the coefficients in the interaction model relate to the separate models for Age for smokers and non-smokers (below)?

Interaction

```
m3<-glm(outcome~age+smoker+age*smoker, data=Whickham, family=binomial)
summary(m3)

##
## Call:
## glm(formula = outcome ~ age + smoker + age * smoker, family = binomial,
##      data = Whickham)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3983  -0.4256   0.2163   0.5598   1.9283
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.169231   0.606600  13.467  <2e-16 ***
## age          -0.133231   0.009953 -13.386  <2e-16 ***
## smokerYes     -1.457843   0.837232  -1.741   0.0816 .
## age:smokerYes  0.022235   0.014495   1.534   0.1250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1560.32  on 1313  degrees of freedom
## Residual deviance:  942.68  on 1310  degrees of freedom
## AIC: 950.68
##
## Number of Fisher Scoring iterations: 6
```

Smokers

```
Whickham_smoker<- Whickham %>% filter(smoker=="Yes")
m3_smoker<-glm(outcome~age, data=Whickham_smoker, family=binomial)
summary(m3_smoker)

##
## Call:
## glm(formula = outcome ~ age, family = binomial, data = Whickham_smoker)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0009   0.1337   0.3044   0.6362   1.8457
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.71139   0.57702  11.63  <2e-16 ***
## age          -0.11100   0.01054 -10.53  <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: 639.89  on 581  degrees of freedom
## Residual deviance: 453.72  on 580  degrees of freedom
## AIC: 457.72
##
## Number of Fisher Scoring iterations: 5
```

Non-Smokers

```
Whickham_nonsmoker<- Whickham %>% filter(smoker=="No")
m3_nonsmoker<-glm(outcome~age, data=Whickham_nonsmoker, family=binomial)
summary(m3_nonsmoker)
```

```
##
## Call:
## glm(formula = outcome ~ age, family = binomial, data = Whickham_nonsmoker)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3983  -0.4609   0.1532   0.4357   1.9283
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.169231   0.606600   13.47  <2e-16 ***
## age         -0.133231   0.009953  -13.39  <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: 911.23  on 731  degrees of freedom
## Residual deviance: 488.96  on 730  degrees of freedom
## AIC: 492.96
##
## Number of Fisher Scoring iterations: 6
```


13. Is this model with the interaction term a better fit than the model with Age alone (Model 0 above), or than the model with just Smoking alone (Model 1)?

```
library(lmtest)
lrtest(m3,m0)
```

```
## Likelihood ratio test
##
## Model 1: outcome ~ age + smoker + age * smoker
## Model 2: outcome ~ age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    4 -471.34
## 2    2 -473.25 -2  3.8255    0.1477
```

```
lrtest(m3,m1)
```

```
## Likelihood ratio test
##
## Model 1: outcome ~ age + smoker + age * smoker
## Model 2: outcome ~ smoker
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    4 -471.34
## 2    2 -775.56 -2 608.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- a. What are the null and alternative hypotheses for each of these tests?

- b. What is the test statistic and p-value for each and what does that mean about the test?

- c. What do these tests tell you about the relationships between age, smoking, and survival over 20 years in this cohort of women from Wickham?