Intro to Logistic Regression

Kevin Cummiskey

November 18, 2019

```
library(tidyverse)
```

Section 6.2 Introduction to 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.

Calculate the odds ratio and logs odds ratio of being alive comparing smokers to nonsmokers.

Perform the chi-square test to determine if there is a significant association between smoking and survival. State the appropriate hypotheses and report the chi-square test statistic and p-value.

Does your result above mean smoking raises the probability of being alive?

Useful log rules for this lesson

- $e^{\ln r} = r$
- $\ln(e^r) = r$
- $e^r \times e^s = e^{r+s}$
- $\ln\left(\frac{r}{s}\right) = \ln(r) - \ln(s)$

Introduction to Logistic Regression

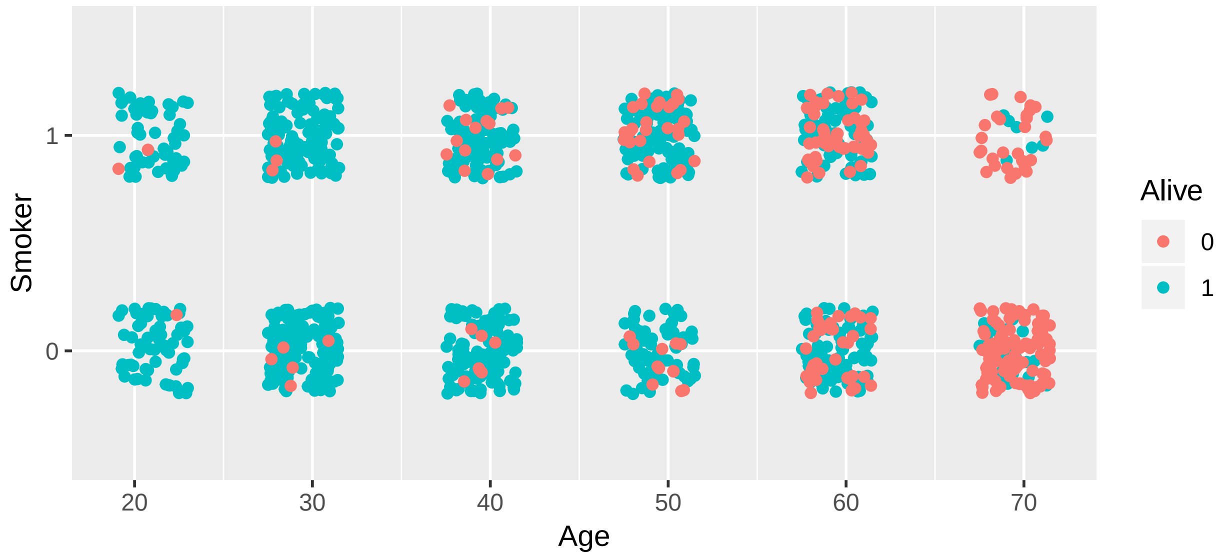
```
smoke = read.table(file = "smoke.csv", header = T, sep = ",")
smoke$Alive = factor(smoke$Alive)
smoke$Smoker = factor(smoke$Smoker)
smoke %>% group_by(Smoker,Alive) %>% count() %>% spread(key = "Smoker",value = "n")
```

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

```
## # A tibble: 2 x 3
## # Groups:   Alive [2]
```

```
##   Alive   `0`   `1`
##   <fct> <int> <int>
## 1 0       166   126
## 2 1       502   443
```

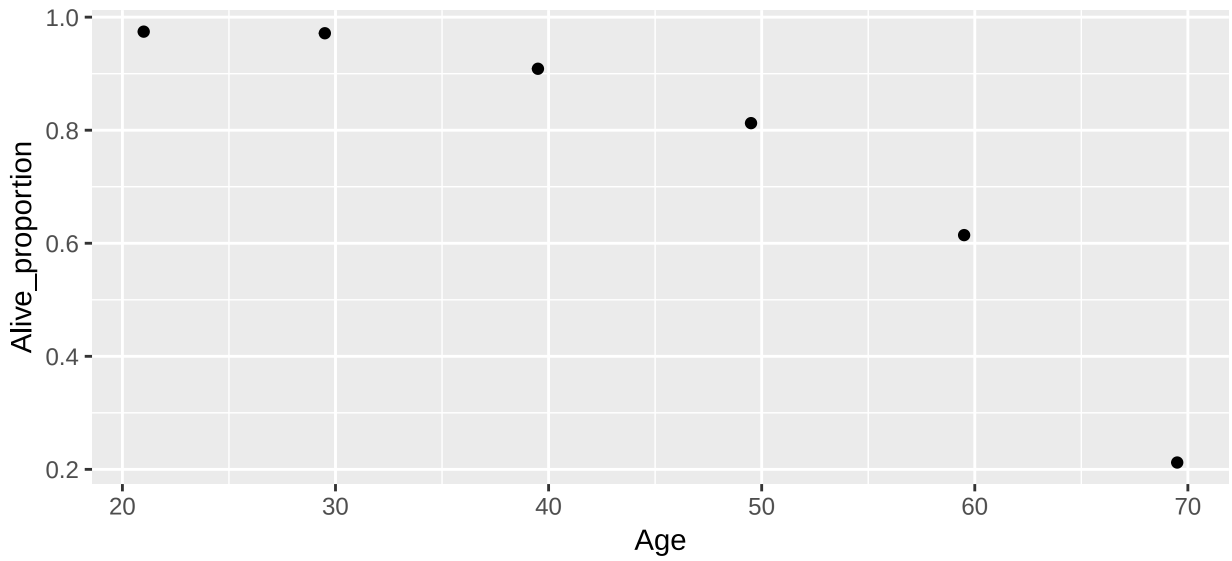
```
smoke %>% ggplot(aes(x = Age, y = Smoker, color = Alive)) + geom_jitter(width = 2, height = 0.2)
```



Based on the plot, is age a confounding variable of the association between smoking and survival? Explain.

To demonstrate logistic regression, let's look at the relationship between survival and age.

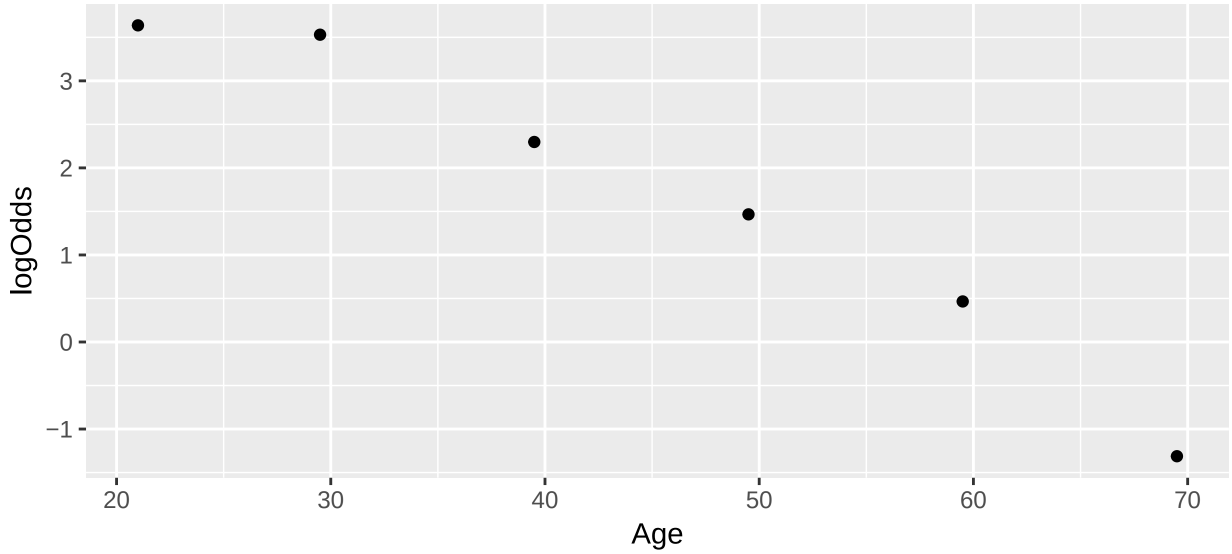
```
#Let's look at a plot of age vs proportion alive
summary = smoke %>% count(Age, Alive) %>%
  group_by(Age) %>%
  mutate(Alive_proportion = n/sum(n),
         total = sum(n)) %>%
  filter(Alive == 1) %>% select(-c(Alive,n))

summary %>%
  ggplot(aes(x = Age, y = Alive_proportion)) +
  geom_point()
```



Explain two reasons a linear model for age and alive proportion is not appropriate?

Let's try a log odds (or *logit*) transformation.

```
summary = summary %>%
  mutate(logOdds = log(Alive_proportion/(1-Alive_proportion)))
summary %>% ggplot(aes(x = Age, y = logOdds)) +
  geom_point()
```



Is a linear model on the log odds appropriate? How does the logit transformation ensure probabilities are between 0 and 1?

Let's fit the following model.

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

How do we interpret β_0 and β_1? Would we expect β_1 to be positive or negative? Why is there no ϵ_i on this model?

```r
# In practice, we would just do this:
model_age = glm(Alive ~ Age, data = smoke, family = "binomial")
summary(model_age)
```

```
##
## Call:
## glm(formula = Alive ~ Age, family = "binomial", data = smoke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0327   0.1422   0.2295   0.6782   1.5837
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.972626   0.410228    17.0   <2e-16 ***
## Age         -0.113535   0.007232   -15.7   <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: 1352.0  on 1236  degrees of freedom
## Residual deviance:  938.4  on 1235  degrees of freedom
## AIC: 942.4
##
## Number of Fisher Scoring iterations: 5
```

What would we conclude from these results?