Simple Logistic Regression Lab - Answers

SDS 291

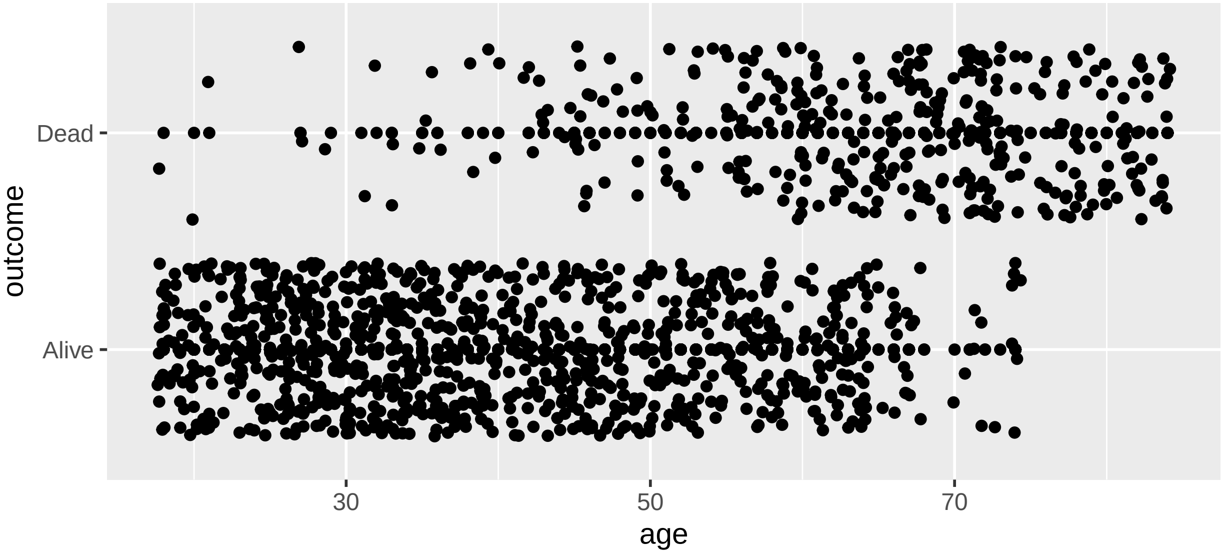
```
knitr::opts_chunk$set(echo = TRUE)
library(mosaic)
library(Stat2Data)
library(magrittr)
library(tidyverse)
data("Whickham")
Hmisc::describe(Whickham)
```

```
## Whickham
##
##  3  Variables      1314  Observations
## --------------------------------------------------------------------------------
## outcome
##        n  missing distinct
##     1314        0        2
##
## Value      Alive  Dead
## Frequency    945   369
## Proportion 0.719 0.281
## --------------------------------------------------------------------------------
## smoker
##        n  missing distinct
##     1314        0        2
##
## Value         No   Yes
## Frequency    732   582
## Proportion 0.557 0.443
## --------------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1314        0       67        1    46.92    20.06       21       25
##      .25      .50      .75      .90      .95
##       32       46       61       71       77
##
## lowest : 18 19 20 21 22, highest: 80 81 82 83 84
## --------------------------------------------------------------------------------
```

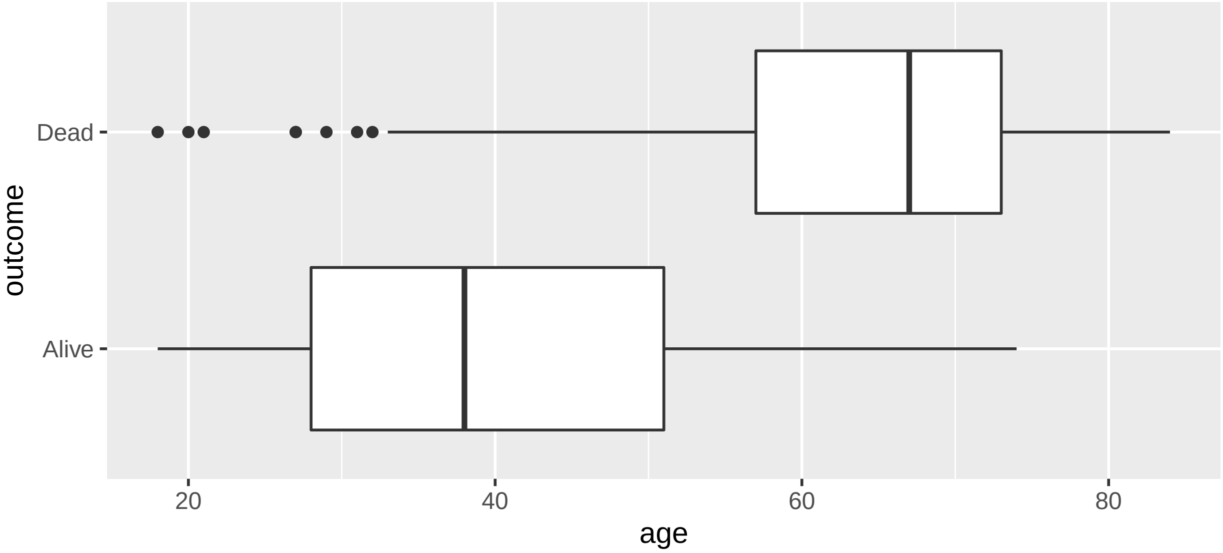
Age and Outcome

#1. Visualize the relationship between age and the outcome

```
qplot(x=outcome,y=age,data=Whickham)+geom_jitter()+coord_flip()
```



```
qplot(x=outcome,y=age,data=Whickham,geom="boxplot")+coord_flip()
```



We see positive relationship between age and death. In the boxplots, the median age (the center, vertical line in the middle of the box) of those who were dead 20 years later is higher/older than the women who were alive 20 years later.

The jittered plot also helps illustrate the age distribution – there are more younger people in this sample, as seen by the greater density of points at younger ages.

A note about the code: `R` only knows how to build a boxplot where the categorical (or binary) variable is on the x-axis. So you need the `coord_flip()` component to switch the response variable back on the y-axis.

##2. Fit a logistic model to test the relationship between outcome and age (with code below).

```
Whickham$outcome<-relevel(Whickham$outcome, ref="Alive")
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
## -1.8953  -0.5538  -0.2293   0.4277   3.2296
##
## 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
```

A note about the code: Logistic regression assumes that the response variable is 0/1 (1 reflecting the level of the outcome that you want to consider as your outcome. Like in linear regression when the quantitative variable is already ordered numerically and R knows which is higher, you sometimes need to tell `R` what you want the levels of your variable to be. Especially if the variables are already labeled as factor variables (like Alive/Dead).

##3. 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.

$log(odds)$ = -7.403 + 0.122·age

The direction is positive, as reflected by the positive β_1 - as age increases, so do the log odds of death. For every 1 additional year of age, the log(odds) of being dead 20 years later increases, on average, 0.122. The magnitude of the association seems small, though it's really difficult to tell on the log(odds) scale since it's not a very intuitive set of units.

##4. Based on this model, what is the probability that a 60 year old was dead?

There are (at least) three ways to calculate this answer. For the two that use `R` to do the calculation, you have to define a value for age - here, it's 60 - so that it knows by what to multiply the $\hat{\beta}_1$ coefficient.

Use the predict function

```
newdata = data.frame(age=60)
predict(m0, newdata, type="response")
```

```
##         1
## 0.4771564
```

Program the math yourself

```r
age<-60
logodds<-coef(m0)[1]+ (coef(m0)[2]*age)
m0_data<-as.data.frame(cbind(age,logodds))
m0_data <- m0_data %>%
  mutate(odds=exp(logodds),
        prob = odds/(1+odds))
m0_data
```

```
##   age     logodds     odds      prob
## 1  60 -0.09143792 0.912618 0.4771564
```

Now you have a dataframe with a column for age and three columns for the log(odds), the odds, and the probability.

Do the math manually

log(odds)= -7.403 + 0.122·60

$log(odds_{60})$ = -0.091

$odds_{60} = e^{log(odds_{60})}$= 0.913

$\hat{\pi}_{60} = \frac{odds_{60}}{1+odds_{30}}$= 1.004

A note about the code: I'm "cheating" here, because I'm actually having `R` do the math on the back end rather than pasting these numbers into RMarkdown. If you want to get into the bells and whistles of RMarkdown and reproducible documents, take a look at the RMD file. You ***absolutely*** *do not need to do this in your own work* – this is meant more for people who already feel really comfortable with RMarkdown and want to grow their skills.

##5. What is the odds ratio for the association between age and dying?

```r
exp(coef(m0))
```

```
##  (Intercept)          age
## 0.0006093452 1.1295975993
```

Interpretation: Each additional year of age is associated with a 1.13 times higher odds of being dead 20 years later. In other words, of two individuals who were one-year apart in age, the older individual has 13% higher odds of being dead 20 years in the future than the younger individual.

You could also calculate the odds for someone a year older (say, 61) and compare the odds.

$odds_{61} = e^{\hat{\beta}_0+\hat{\beta}_1 \cdot 61}$= 1.031

$OR = \frac{odds_{61}}{odds_{60}}$= $\frac{1.031}{0.913}$ =1.13

Outcome and Smoker: 2-by-2 table

Construct a two-way table to explore whether smoking is associated with the outcome (being dead).

A note about the code (below): We're reordering the levels of the explanatory and response variables to get a certain order/organization of the 2-by-2 table. For our purposes in this class, I'll try to demonstrate only this arrangement to simplify things. However, you may get other arrangements – in your homework, group projects, etc. – so it's good to think about why we're doing what we're doing so you can calculate odds and odds ratios regardless of the table arrangement. What we want is:

Response (column) / Explanatory (row)	Response: Yes	Response: No
Explanatory (Yes)	**a**	**b**
Explanatory (No)	**c**	**d**

So the OR of the explanatory variable having the response you want is: $\frac{a/b}{c/d}$ or $\frac{\text{success/failures of explanatory(yes)}}{\text{success/failures of explanatory(no)}}$. Since the OR is the difference in the odds of the outcome/response for a 1-unit-difference in X, here a one-unit difference is going from explanatory(yes) to explanatory(no).

```
Whickham$smoker<-relevel(Whickham$smoker, "Yes", "No")
Whickham$outcome<-relevel(Whickham$outcome, "Dead","Alive")
```

##6.Calculate the the proportion of smokers who were dead and of non-smokers who were dead 20 years later.

```
tally(~ smoker + outcome, margins=FALSE, data=Whickham)
```

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

Proportion of smokers who were dead: 139/(139+443) = 0.239 Proportion of non-smokers who were dead: 230/(230+502) = 0.314

So we know that there were fewer smokers among those who died than there were, proportionally, than those who were alive 20 years later. But what we might be interested in is to compare smokers to non-smokers.

We can think of these values as *conditional* proportions – the proportion of people with the outcome conditional on them being a smoker.

You can also get these values from the the CrossTable function from the gmodels package.

```
gmodels::CrossTable(Whickham$smoker, Whickham$outcome,
prop.r=TRUE, prop.c=FALSE, prop.chisq=FALSE, prop.t=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1314
##
##
##                 | Whickham$outcome
## Whickham$smoker |      Dead |     Alive | Row Total |
## ----------------|-----------|-----------|-----------|
##             Yes |       139 |       443 |       582 |
##                 |     0.239 |     0.761 |     0.443 |
## ----------------|-----------|-----------|-----------|
##              No |       230 |       502 |       732 |
##                 |     0.314 |     0.686 |     0.557 |
## ----------------|-----------|-----------|-----------|
##    Column Total |       369 |       945 |      1314 |
## ----------------|-----------|-----------|-----------|
##
##
```

##7.Calculate the Odds Ratio using the numbers of people in the relevant table cells (a-d).

```
#OR<-(a/b)/(c/d)
OR<-(139/443)/(230/502)
OR
```

```
## [1] 0.6848366
```

Interpretation The odds of being dead 20 years later for a smoker was 0.685 times lower than non-smokers. Thus, smokers had lower odds of being dead 20 years later than non-smokers.

Outcome and Smoker: modeling

First we have to be sure that we are going to calculate the same OR that we did above – smokers' odds of dying compared to non-smokers'.

##8.Fit a logistic model of the relationship between smoking status and outcome.

Note about code: We are going to re-set the reference groups for both the explanatory and response. Why? Remember that in order to make the table above arranged as it was, we wanted the order to be Dead and then Alive (from left to right). So now `R` thinks that Dead is the reference and Alive is the level of the response variable that we're interested in. Same for smokers – it has smokers as the reference. So unless you run the `relevel()` code below, your regression model would be estimating the log(odds) of being Alive for non-smokers compared to smokers. Since we're interested in the relationship of smokers and death, we have to reset the reference groups.

```
Whickham$outcome<-relevel(Whickham$outcome, ref="Alive")
Whickham$smoker<-relevel(Whickham$smoker, ref="No")
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
## -0.8685  -0.8685  -0.7388   1.5216   1.6923
##
## 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
```

##9.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.

$log(odds)$ = -0.781 + -0.379·smoker

We see that the association is negative (the $\hat{\beta}_1$ coefficient is negative). Again, it's hard to intererpret the magnitude on the log(odds) scale, so we probably want an odds ratio to make more sense of this relationship.

##10.Using the two-way table above, show that you can calculate the coefficient for smoking status from your regression model.

```
exp( cbind( OR=coef(m1), confint(m1) ) )
```

```
## Waiting for profiling to be done...

##                    OR     2.5 %    97.5 %
## (Intercept) 0.4581673 0.3913417 0.5347849
## smokerYes   0.6848366 0.5345661 0.8750872
```

We see that the odds ratio of the odds of death for smokers compared to the odds of death for non-smokers is the same from the model as it was before from the 2-by-2 table.

We also have a 95% confidence interval here, which you can get by exponentiating the `confint()` function that we used in linear regression for the confidence intervals.

Interpretation We are 95% confidence that the true odds of death for smokers, compared to non-smokers, are, on average, between 0.535 and 0.875 in the population. Since this 95% CI does not include the null odds ratio (OR=1), we have evidence to suggest that the smokers' odds of death is statistically different than non-smokers'.

##11.Based on your model, what's the probability that a smoker was dead?

```
newdata = data.frame(smoker="Yes")
predict(m1, newdata, type="response")
```

```
##         1
## 0.2388316
```

```
smoker<-1
logodds<-coef(m1)[1]+ (coef(m1)[2]*smoker)
m1_data<-as.data.frame(cbind(smoker,logodds))
m1_data <- m1_data %>%
  mutate(odds=exp(logodds),
        prob = odds/(1+odds))
m1_data
```

```
##   smoker   logodds      odds      prob
## 1      1 -1.159096 0.3137698 0.2388316
```

log(odds)= -0.781 + -0.379·1

$log(odds_{\text{smoker}})$ = -1.159

$odds_{smoker} = e^{log(odds_{\text{smoker}})}$= 0.314

$\hat{\pi}_{smoker} = \frac{odds_{smoker}}{1+odds_{smoker}}$= -1.972

(If you recall, this probability is also the conditional proportion we saw before – of smokers, what proportion had died.)

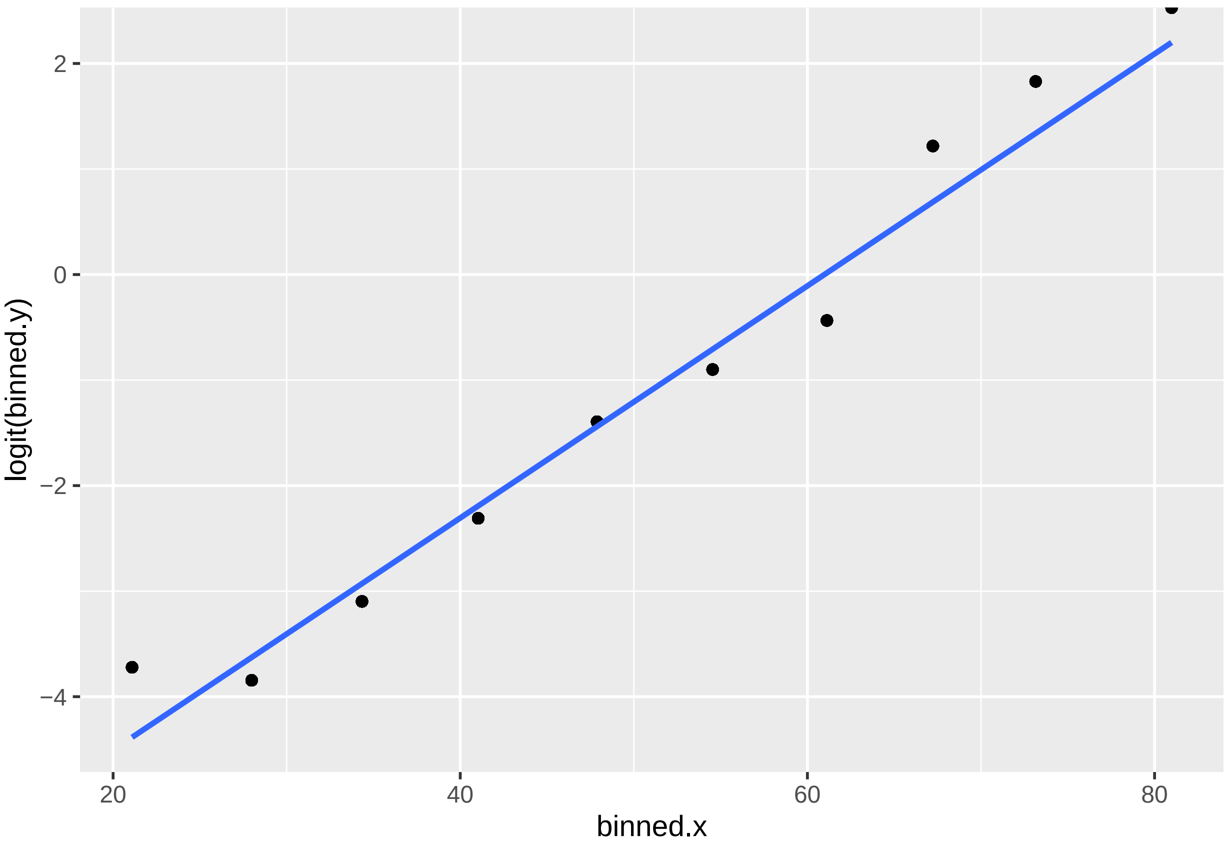
##12.What are the regression assumptions? Evaluate whether they are met.

[Hint: for the question above, check the textbook in Chapter 9.3-9.4 for what the assumptions are. See the code below for one of the assumptions (linearity, p.473). The book suggest ~5 bins, but we're using 10]

```
#let's make `outcome` a numeric variable between 0 and 1 so that the proportion makes sense
#and decide what the 10 age groups should be by "cutting" the data into 10 evenly sized groups.
Whickham <- Whickham %>%
  mutate(isDead=as.numeric(if_else(outcome=="Dead",1,0)),
         ageGroup = cut(age, breaks=10))

# Then generate the proportion of those alive for each age group (so you have something to plot on the
# And the average age for each age group (so you have something to plot on the x-axis)
Whickham_bin <- Whickham %>%
  group_by(ageGroup) %>%
  mutate(binned.y=mean(isDead), binned.x=mean(age))

#is this roughly linear on the log-odds scale?
qplot(y=logit(binned.y), x=binned.x, data=Whickham_bin) + geom_smooth(method="lm", se=FALSE, fullrange=
```


logit(binned.y)
binned.x