Lesson 23 Ukraine Alcohol Abuse (6.2)

LTC J. K. Starling

```
library(tidyverse)
library(GGally)
library(boot)
library(broom)
```

Admin

Basic Laws

$$
\begin{aligned}
e^{\ln x} &= x & \ln(x \times y) &= \ln(x) + \ln(y); \qquad x, y > 0 \\
\frac{e^x}{e^y} &= e^{x-y} & \ln(x/y) &= \ln(x) - \ln(y); \qquad x, y > 0 \\
(e^x)^y &= e^{x \times y} & \ln(x^y) &= y \times \ln(x); \qquad x > 0
\end{aligned}
$$

Alcohol Abuse in Ukraine

Researchers were interested in rates of alcohol abuse among men and women in Ukraine, investigating questions about whether rates differ depending on variables such as sex and age, and after adjusting for such potential confounding variables, how the rates may be related to exposure to different conflicts and traumas such as living near Chernobyl. Data collection took place from February-December 2002 across all 24 of Ukraine's "states" and the republic of Crimea, using the World Mental Health-Composite International Diagnostic Interview (CIDI). The survey questions included enough information to make diagnoses of mental health disorders, including alcohol abuse, and demographic data such as sex and age. Respondents also answered the question, "Have you ever lived in the zone contaminated as a result of the Chernobyl accident?"

```
cher.dat<-read.table("cherdata.txt",
                     header = T,
                     stringsAsFactors = T)
```

(1) Using the output from the `glimpse` and `summary` commands, what is the primary response variable? What are potential explanatory variables of interest? Classify each as quantitative or categorical?

```
glimpse(cher.dat)
```

```
## Rows: 4,725
## Columns: 4
## $ live.chernobyl <fct> Yes, No, No, No, No, No, No, No, No, No, No, No, No, No~
## $ age            <int> 20, 61, 23, 21, 72, 75, 45, 40, 65, 21, 69, 71, 77, 55,~
## $ sex            <fct> Male, Female, Male, Female, Male, Female, Male, Female,~
## $ alc.post       <fct> no, no, no, no, no, no, no, no, no, yes, no, no, no, be~
```

```
summary(cher.dat)
```

```
##  live.chernobyl      age            sex         alc.post
##  No  :4350      Min.   :17.00   Female:2932   before: 210
##  Yes : 374      1st Qu.:34.00   Male  :1793   no    :4196
##  NA's:   1      Median :49.00                 yes   : 319
##                 Mean   :48.89
##                 3rd Qu.:64.00
##                 Max.   :91.00
```

response: alc.post (categorical)

explanatory: live.chernobyl (categorical/binary); age (quantitative); sex (binary)

(2) What are some observations that we may want to remove?

We would want to remove those who were alcoholics before the event (The Chernobyl explosion happened in 1986). We should also consider removing the NA from the live.chernobyl variable.

(3) If we wanted to test the association between `alc.post` and `sex` how would we do it?

We could perform analysis on the difference of the proportions between male and females and whether or not they became alcoholics. This can be accomplished with a two-sample z-test or a chi-square test. Both of these can be accomplished simultaneously using the prop.test function in R.

(4) Do the recommendations from the previous two questions. Is there an association between alcohol abuse and sex?

-Remove those who were alcoholics *before* the Chernobyl accident in 1986.

```
sub.cher <- cher.dat %>% filter(alc.post !="before") %>% droplevels()
```

- Create a 2x2 contingency table for `alc.post` and `sex`.

```
sub.cher$alc.post <- fct_rev(sub.cher$alc.post) # reverse level order
levels(sub.cher$alc.post)
```

```
## [1] "yes" "no"
```

```
cher.tab <- table(sub.cher$sex, sub.cher$alc.post)
cher.tab
```

```
##
##           yes   no
##   Female   62 2853
##   Male    257 1343
```

- Use the theory based approaches mentioned from the last section (§6.1).

```
prop.test(cher.tab, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  cher.tab
## X-squared = 305.52, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1580943 -0.1206171
## sample estimates:
##    prop 1    prop 2
## 0.0212693 0.1606250
```

```
# alternative w/o table command
# x=c(62,257)
# n=c(62+2853,257+1343)
# prop.test(x,n,correct=FALSE)
```

Yes, there appears to be an assocation between alcohol abuse and sex based on our large chi-sq value with associated small p-value.

(6) What is the odds-ratio?

The odds ratio is 8.8, meaning that males have an 8.8 times higher odds of alcohol abuse than females.

(7) What if we were interested in the association between `age` and `alc.post`? Could we repeat our analysis from above? What would we have to do?

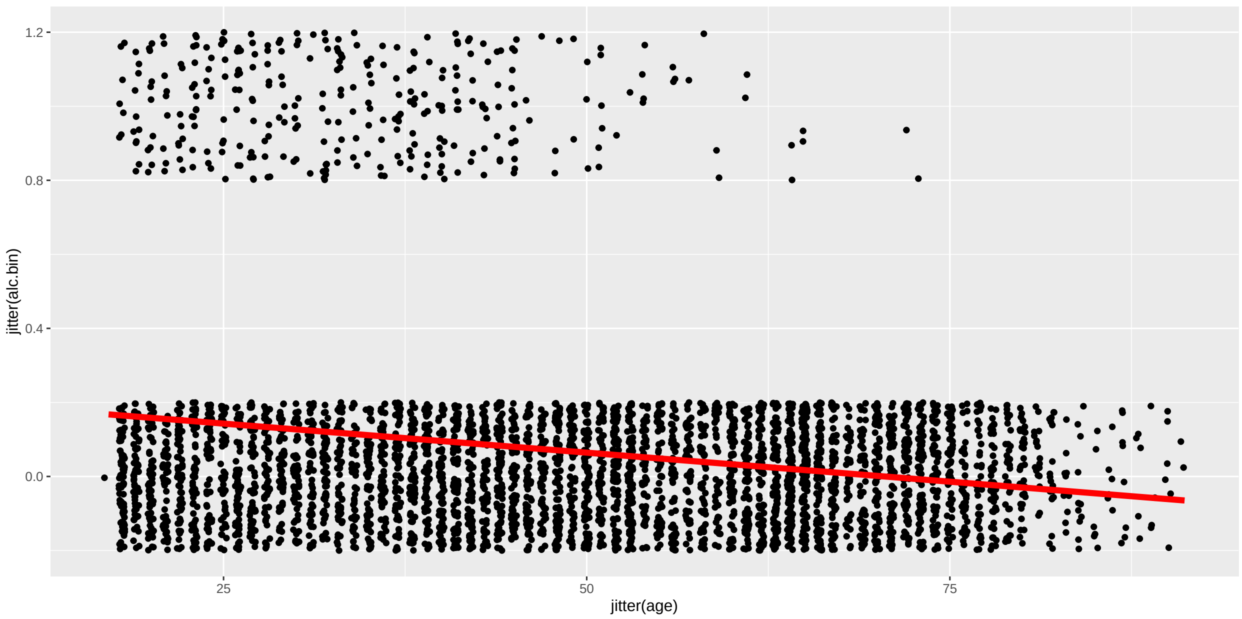
The values for sex and alc.post were both categorical/binary so our analysis was rather easy. If we use age, we would have to discretize the ages to groups of years or even to each year. This would reduce our power, particularly if we use the chi-squared distribtion since we would 'use up' many degrees of freedom with a large number of categories.

Treating age as a quantitative response variable. Median age of the nonabusers is 50 compared to 33 for the abusers. There is a noticeable shift down in the age distribution of the alcohol abusers, likely leading to a small pvalue. We could use a twosample t-test to compare the means.

(8) Perhaps we want to take a model based approach. Looking at the plot, are there issues using a linear model here?

```
sub.cher <- sub.cher %>% mutate(alc.bin = ifelse(alc.post=="yes",1,0))

sub.cher %>% ggplot(aes(x=jitter(age),y=jitter(alc.bin))) +
  geom_point()+
  stat_smooth(method="lm",se=FALSE,color="red",lwd=2)
```

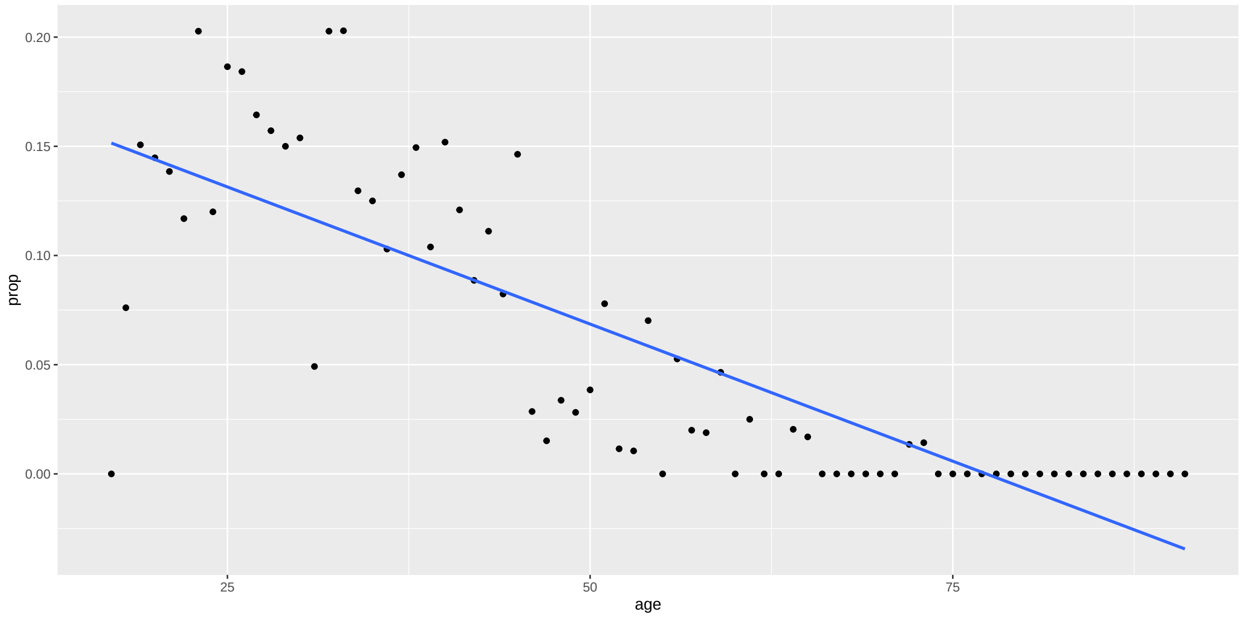


The issues are that the results aren't linearly related. We even get a negative prediction value for ages above 60-ish.

(9) Explain what is going on with the following code and figure. Does this model also have any issues?

```
grouped.cher=sub.cher %>% group_by(age)%>%summarize(prop=mean(alc.bin))

grouped.cher %>% ggplot(aes(x=age,y=prop))+geom_point()+
  stat_smooth(method="lm",se=FALSE)
```



The code is grouping the percentage of alcohol abusers by each age and then plotting the values. This model is better as far as a linear relationship is concerned, but we still will have a negative results above the 75-ish age.

(10) Either way the linear fit is concerning and not appropriate to the data. It may work better in this case, to fit a logistic regression. A logistic regression model for our data is:

$log(p/(1-p)) = beta0 + beta1 \cdot agei$

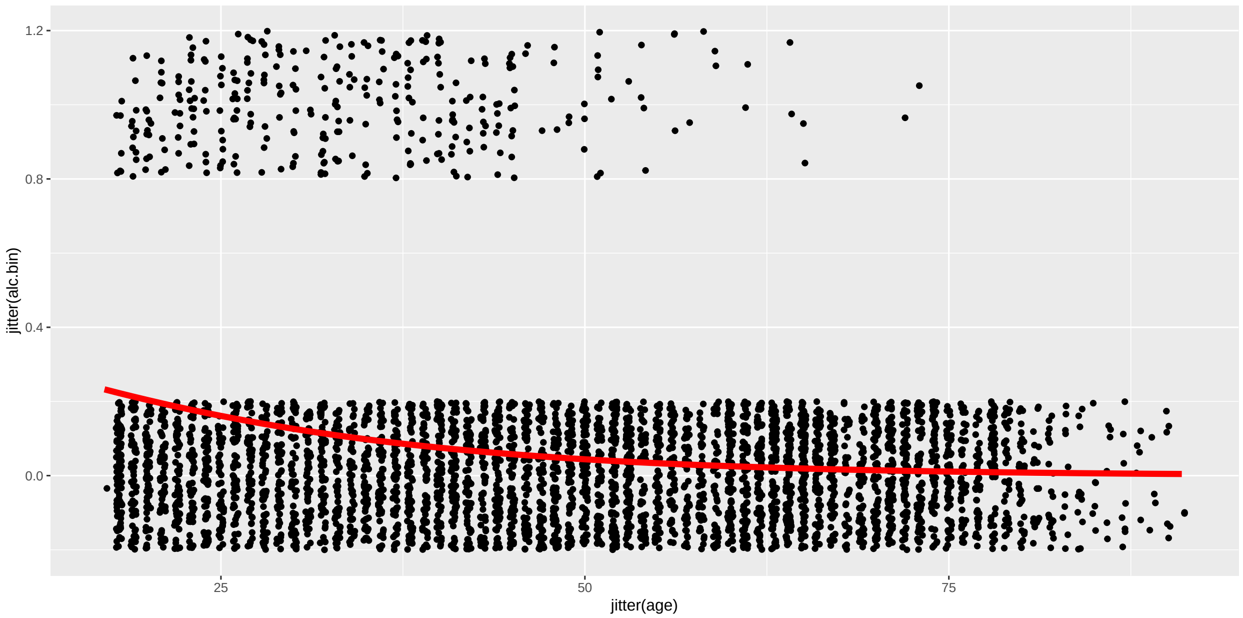
(11) The assumption, then, is we can fit a logistic curve through our data. We can do this in R by:

```
cher.glm <- glm(alc.bin ~ age, data = sub.cher, family = "binomial")
```

To see the fit we can do:

```
fitted.glm <- augment(cher.glm)

fitted.glm %>% ggplot(aes(x = jitter(age),y = jitter(alc.bin))) +
  geom_point() +
  geom_line(aes(x = age,y = inv.logit(.fitted)), lwd=2, color="red")
```



```
# ## OR
# fitted.glm %>% ggplot(aes(x = jitter(age),y = jitter(alc.bin))) +
#   geom_point() +
#   geom_line(aes(x = cher.glm$model$age,y = cher.glm$fitted.values), lwd=2, color="red")
```

Discuss the code:

The augment command adds additional columns to the data. the .fitted provides the logit modified fitted values, so you have to use the inv.logit() function to print the probabilities.

(12) R uses MLE to find the coefficients of our fitted model and can be displayed using the `coef` function. What is our fitted model? What does the intercept mean? Is there a way to make it more meaningful in our model?

```
coef(cher.glm)
```

```
## (Intercept)         age
## -0.22343898 -0.05706901
```

```
summary(cher.glm)
```

```
##
## Call:
## glm(formula = alc.bin ~ age, family = "binomial", data = sub.cher)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7277  -0.4418  -0.2839  -0.1809   2.9671
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.223439   0.157187  -1.421    0.155
## age         -0.057069   0.004153 -13.741   <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: 2305.6  on 4514  degrees of freedom
## Residual deviance: 2063.3  on 4513  degrees of freedom
## AIC: 2067.3
##
## Number of Fisher Scoring iterations: 6
```

$predictedlog - odds = log(p/(1-p)) = -0.2234 - 0.05707 \cdot agei$

We could make the explanatory variable (age - 17) so zero corresponds to the youngest person in the data set.

(13) Calculate the predicted odds of someone age 50 of being diagnosed with alcohol abuse.

What is the probability that a person aged 50 is diagnosed with alcohol abuse. (Hint: rearrange the equation)

```
myage <- 50
# predicted log-odds:
LOdds <- -0.22343898 -0.05706901 * myage
LOdds
```

```
## [1] -3.076889
```

```
# odds:
Odds <- exp(LOdds)
Odds
```

```
## [1] 0.04610244
```

```
# probability:
prob <- Odds / (1+ Odds)
prob
```

```
## [1] 0.04407067
```

(14) To interpret the slope (-0.057), let's compare the odds of someone 50 to the odds of someone 51. What if we wanted to compare the odds for two people with a 10-year age difference?

Remember that this is the odds for age = 50 and age = 51. We can compare them using the odds ratio:

$e^{-0.2234-0.05707(51)} = e^{-0.2234}e^{-0.05707(51)} = (0.80)(0.94)^{51}$

$e^{-0.2234-0.05707(50)} = e^{-0.2234}e^{-0.05707(50)} = (0.80)(0.94)^{50}$

$(0.80)(0.94)^{51}/(0.80)(0.94)^{50} = (0.94)^{51}/(0.94)^{50} = 0.94$

So, we can say that for each increase in age, we multiply the odds of success (abusing alcohol) by 0.94.

For a 10-year age diference we would have $0.94^10 = 0.54$.

(15) If we want to test gender, we could do:

```
gender.glm <- glm(alc.bin ~ sex, data = sub.cher, family = "binomial")
summary(gender.glm)
```

```
##
## Call:
## glm(formula = alc.bin ~ sex, family = "binomial", data = sub.cher)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5918  -0.5918  -0.2074  -0.2074   2.7751
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8290     0.1284  -29.83   <2e-16 ***
## sexMale       2.1754     0.1453   14.97   <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: 2305.6  on 4514  degrees of freedom
## Residual deviance: 2010.4  on 4513  degrees of freedom
## AIC: 2014.4
##
## Number of Fisher Scoring iterations: 6
```

Note that the Z value in this case is NOT the square root of the χ^2 statistic above. That's because they are testing two different things. One is testing independence of π values, the other is testing a logistic relationship between sex and alcohol.

(16) What is the predicted odds ratio for males compared to females? What sex is more likely to abuse alcohol? How does this relate to our parameters?

```r
Odds_m <- exp(-3.8290 + 2.1754)
Odds_f <- exp(-3.8290)

OR_mf <- Odds_m / Odds_f

c(Odds_m, Odds_f, OR_mf)
```

```
## [1] 0.19135977 0.02173134 8.80570670
```

```r
# check with exp of beta_male
exp(2.1754)
```

```
## [1] 8.805707
```

This corresponds to the slope 2.175, as the odds ratio between males and females is 8.8. For this indicator variable, a one unit change indicates changing from females to males. Males are more likely to have been diagnosed with alcohol abuse.

(17) What is the probability a male is diagnosed with alcoholism? What is the probability a female is diagnosed?

```r
prob_m <- Odds_m / (1 + Odds_m)
prob_f <- Odds_f / (1 + Odds_f)
c(prob_m, prob_f)
```

```
## [1] 0.16062299 0.02126913
```

(18) What is a 95% CI for β_{sex}?

$beta1 \pm 2 \times SE$

(19) So this gives us a 95% Confidence interval for the effect of sex on the log-odds. If we want a 95% CI for the multiplicative effect of gender on alcoholism we could do:

```
c(2.1754 - 2 * 0.1453, 2.1754 + 2 * 0.1453)
```

```
## [1] 1.8848 2.4660
```

```
c(exp(1.89), exp(2.46))
```

```
## [1]  6.619369 11.704812
```

So males are 6.62 to 11.7 more times likely to develop alcoholism than females, according to this analysis.