Lesson 23 Ukraine Alcohol Abuse (6.2)

LTC J. K. Starling

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

Admin

Basic Laws

$$\begin{aligned} e^{\ln x} &= & \ln(e^x) &= \\ e^x \times e^s &= & \ln(x \times y) &= \\ \frac{e^x}{e^y} &= & \ln(x/y) &= \\ (e^x)^y &= & \ln(x^y) &= \end{aligned}$$

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                              
```

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

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

(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
```

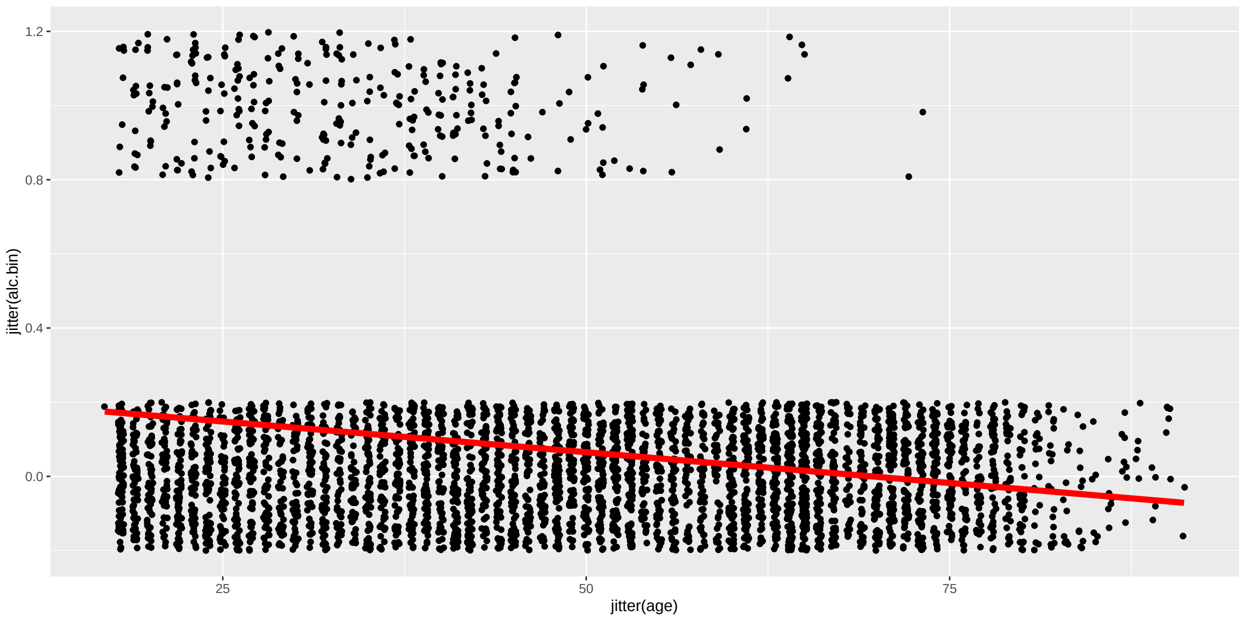
```
# alternative w/o table command
# x=c(62,257)
# n=c(62+2853,257+1343)
# prop.test(x,n,correct=FALSE)
```

(6) What is the odds-ratio?

(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?

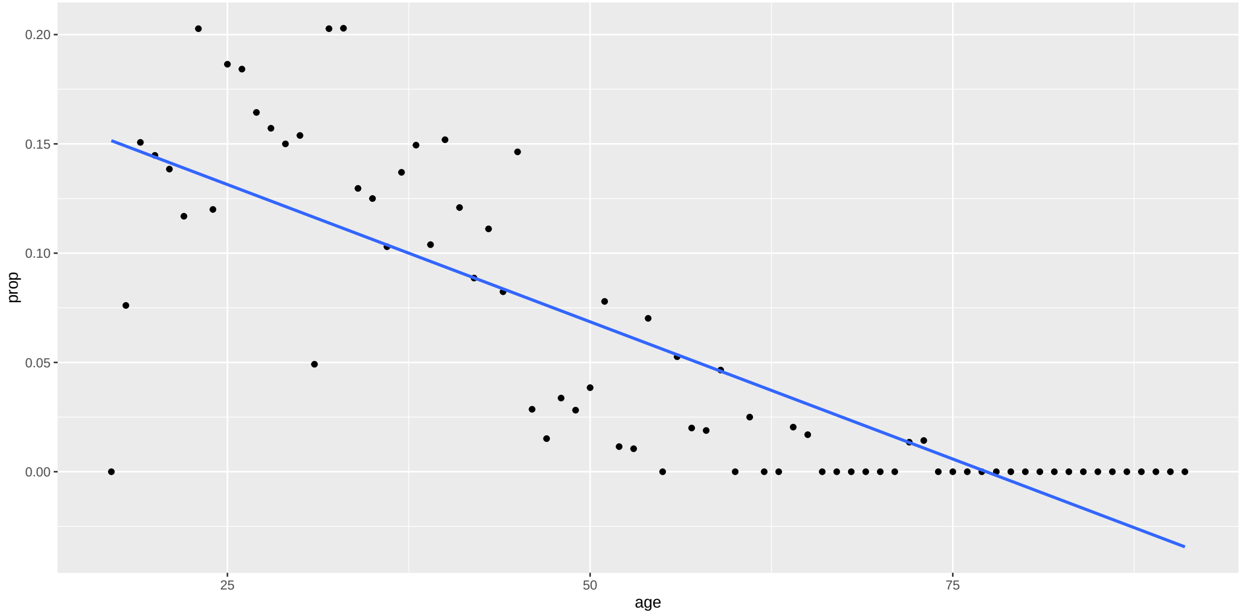
(8) Perhaps we want to take a model based approach. Looking at the plot, are there issues using a linear model here?

```
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)
```



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

```r
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)
```



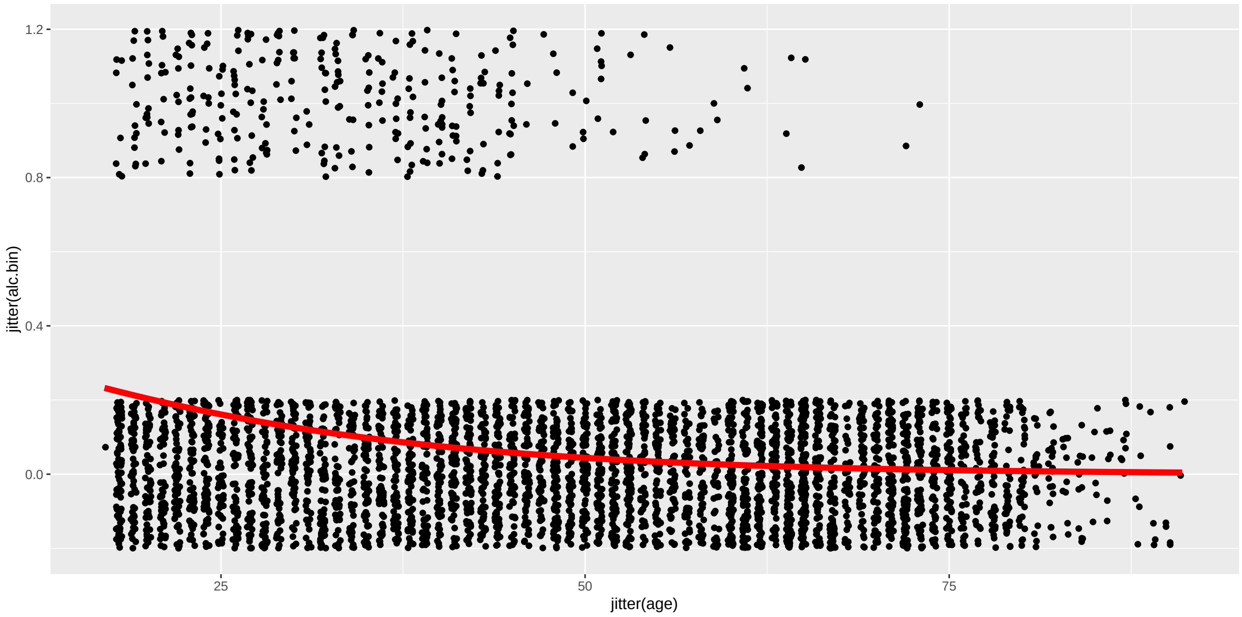
(10) Either way the linear fit is concerning and not appropriate to the data. It may work better in this case, to fit a logistic regression. A logistic regression model for our data is:

(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:

(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
```

(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)

(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?

(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?

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

(18) What is a 95% CI for β_{sex}?

(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.