Lsn 26

Clark

Admin

Basic Laws - KNOW these

Researchers were interested in rates of alcohol abuse among men and women in Ukraine, investigationg questiona about whether rates differ depending on variables such as sex and age, and after adjustin g for such potential confoudning variables, how the rates may be related to explsure to different conflicts and traumas such as living near Chernobyl.

```
cher.dat<-read.table("http://www.isi-stats.com/isi2/data/cherdata.txt",header=T)
```

In total we have:

```
cher.dat %>% summarize(count=n())
```

```
##   count
## 1  4725
```

Observations. We first remove those that had alcohol abuse *before* Chernobyl

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

```
##   count
## 1  4515
```

If we wanted to test the association betwen `alc.post` and `sex` how would we do it?

```
x=c(62,257)
n=c(62+2853,257+1343)
prop.test(x,n,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## 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
```

From here, what is the odds-ratio?

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?

Perhaps we want to take a model based approach. Looking at the plot:

```
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)
```

Are there issues using a linear model here? We might also do?

```
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)
```

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:

The assumption, then, is we can fit a logistic curve through our data. In R we can do:

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

To see the fit we can do:

```
library(broom)
fitted.glm<-augment(cher.glm)
fitted.glm %>% ggplot(aes(x=age,y=alc.bin))+geom_point()+
  geom_line(aes(x=age,y=inv.logit(.fitted)),lwd=2,color="red")
```

Let's talk about this code.....

The coefficients of our fitted model are found through MLE and are:

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

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

So our fitted model is:

The predicted odds of someone age 50 of being diagnosed with alcohol abuse is:

To interpret the slope, let's compare the odds of someone 50 to the odds of someone 51

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.

What is the predicted odds ratio for males compared to females? How does this relate to our parameters?

What is the probability a male is diagnosed with alcoholism? What is the probability a female is diagnosed?

What is a 95% CI for β_{sex}?

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:

```
exp(1.89)
```

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

```
exp(2.46)
```

```
## [1] 11.70481
```

So males are 6.67 to 11.7 more times likely to develop alcoholism than females, according to this analysis.