# Lsn 25 - AY23

Clark

## Admin

Does smoking impact 20 year survival after adjusting for age at the initial interview?

Our sources of variation diagram is:

One way to explore our data is from a mosaic plot (which is similar to a segmented bar chart). This can be implimented via the `ggmosaic` library in R

```
library(ggmosaic)
dat<-read.csv("smoke.csv")
dat.mod <- dat %>% mutate(age.fac=as.factor(Age),smoke.fac=as.factor(Smoker))
dat.mod %>% ggplot()+
  geom_mosaic(aes(x=product(smoke.fac,age.fac),fill=smoke.fac))
```

From here, what concerns might we have in any analysis of the data?

We can find the empirical odds ratios by:

```
group.odds=dat %>% group_by(Age,Smoker)%>%summarize(odds=mean(Alive)/(1-mean(Alive)))
```

```
## `summarise()` has grouped output by 'Age'. You can override using the `.groups`
## argument.
```

```
odds.ratios=group.odds %>% group_by(Age)%>%summarize(ratio=odds[Smoker==1]/odds[Smoker==0])
odds.ratios
```

```
## # A tibble: 6 x 2
##     Age ratio
##   <dbl> <dbl>
## 1  21   0.434
## 2  29.5 1.33
```

```
## 3  39.5 0.417
## 4  49.5 0.694
## 5  59.5 0.620
## 6  69.5 0.871
```

From here, what does it look like is happening?

However, if we take the full data set we have:

```
smoke.glm<-glm(Alive~Smoker,data=dat,family="binomial")
coef(smoke.glm)
```

```
## (Intercept)      Smoker
##   1.1066123   0.1506755
```

So the odds ratio between Smokers and non-Smokers is:

This is an example of *Simpson's Paradox* and is due to the confounding in our data. To account for this, we need to adjust for age in our model.

Our logistic regression model becomes:

To fit this we do:

```
smokeage.glm<-glm(Alive~Smoker+Age,data=dat,family="binomial")
summary(smokeage.glm)
```

```
##
## Call:
## glm(formula = Alive ~ Smoker + Age, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1026   0.1277   0.2396   0.6351   1.6675
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.2411     0.4502  16.085   <2e-16 ***
## Smoker       -0.2823     0.1677  -1.683   0.0923 .
## Age          -0.1160     0.0075 -15.467   <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: 1352.02  on 1236  degrees of freedom
## Residual deviance:  935.54  on 1234  degrees of freedom
## AIC: 941.54
##
## Number of Fisher Scoring iterations: 6
```

The fit can be seen by:

```
library(broom)
library(boot)
fit.dat<-augment(smokeage.glm)%>%
group_by(Age,Smoker)%>%summarize(fit=inv.logit(mean(.fitted)))
samp.props=dat %>% group_by(Age,Smoker)%>%summarize(ps=mean(Alive))
fit.dat %>%ggplot(aes(x=Age,y=fit,group=Smoker,color=as.factor(Smoker)))+
  geom_line()+geom_point(aes(x=Age,y=ps,color=as.factor(Smoker)),data=samp.props)
```

While perhaps this is difficult to see, one of the assumptions of the model is that the odds ratio between smokers and non-smokers of the same age is:

From our empirical data perhaps this isn't the case. To correct this we can add an interaction term and our model becomes:

From here, we see that for a given age, the odds ratio between smoker and non-smoker is:

This is fit by:

```
full.glm<-glm(Alive~Smoker*Age,data=dat,family="binomial")
summary(full.glm)
```

```
##
```

```
## Call:
## glm(formula = Alive ~ Smoker * Age, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2077   0.1082   0.2718   0.6126   1.5850
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.76239    0.62586  12.403   <2e-16 ***
## Smoker      -1.38524    0.85556  -1.619    0.105
## Age         -0.12494    0.01050 -11.904   <2e-16 ***
## Smoker:Age   0.01994    0.01511   1.319    0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1352.02  on 1236  degrees of freedom
## Residual deviance:  933.81  on 1233  degrees of freedom
## AIC: 941.81
##
## Number of Fisher Scoring iterations: 6
```

The new curves are:

```
fit.dat<-augment(full.glm)%>%
group_by(Age,Smoker)%>%summarize(fit=inv.logit(mean(.fitted)))
```

```
## `summarise()` has grouped output by 'Age'. You can override using the `.groups`
## argument.
```

```
fit.dat %>%ggplot(aes(x=Age,y=fit,group=Smoker,color=as.factor(Smoker)))+
  geom_line()+geom_point(aes(x=Age,y=ps,color=as.factor(Smoker)),data=samp.props)
```

So is it helpful to have an interaction term here? To address this, we need something similar to ANOVA. However our assumptions for ANOVA cannot, in any way, be satisfied. Commonly, this is done through computing Deviance of a model. Deviance of a model is related to maximum likelihood estimation. For logistic regression, we compute the log-likelihood of the null model and compare it to the log-likelihood of our model, which has a $\chi^2$ distribution with $p$ degrees of freedom where $p$ is the number of terms in our model (in this case 3)

This is similar to the ANOVA test we do at the beginning to say is our model better than the null model. We can get more specific by doing:

```
library(car)
Anova(full.glm,type="III") #Warning here...
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: Alive
##            LR Chisq Df Pr(>Chisq)
```

```
## Smoker        2.635  1     0.1045
## Age         275.822  1     <2e-16 ***
## Smoker:Age    1.738  1     0.1874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Just as in linear regression we need to check performance. It doesn't really make sense to look at $R^2$ here (why?), so what is commonly done is to form what is called a **Confusion matrix**

```
library(caret)
pdata<- predict(full.glm,type="response")
cm<-confusionMatrix(data=as.factor(pdata>0.5),reference = as.factor(dat$Alive==1))
cm$table
```

```
##            Reference
## Prediction FALSE TRUE
##      FALSE   130   35
##      TRUE    162  910
```

There's a ton here to unpack and we won't have time to go into it (take MA478!)

However, one thing to note is if we didn't include an interaction term we would have:

```
pdata<- predict(smokeage.glm,type="response")
cm<-confusionMatrix(data=as.factor(pdata>0.5),reference = as.factor(dat$Alive==1))
cm$table
```

```
##            Reference
## Prediction FALSE TRUE
##      FALSE   130   35
##      TRUE    162  910
```

Which shows we're not gaining a ton by including an interaction here. However if we use the first model of only smoking we have:

```
pdata<- predict(smoke.glm,type="response")
min(pdata)
```

```
## [1] 0.751497
```

```
max(pdata)
```

```
## [1] 0.7785589
```

Which, says that under this model we would always predict someone was alive!

Let's interpret our results here: