Lsn 28

Clark

Review

You are interested in whether or not the Red Sox are more likely to win at Fenway Park. You fit the following logistic regression model for the 162 games in the 2018 season:

- Response Variable: whether or not the Red Sox won
- Explanatory Variables: Field (Home/Away), Opponent

```r
model = glm(Result ~ Field + Opp, data = redsox, family = "binomial")

summary(model)
```

```
##
## Call:
## glm(formula = Result ~ Field + Opp, family = "binomial", data = redsox)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0229  -1.1214   0.5393   0.9496   1.5890
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.399e+00  1.111e+00   1.260    0.208
## FieldHome    4.559e-01  3.708e-01   1.229    0.219
## OppBAL       5.299e-02  1.268e+00   0.042    0.967
## OppCHW      -1.885e+00  1.341e+00  -1.405    0.160
## OppCLE      -1.951e+00  1.343e+00  -1.453    0.146
## OppDET      -9.249e-01  1.403e+00  -0.659    0.510
## OppHOU      -1.885e+00  1.341e+00  -1.405    0.160
## OppKCR       1.475e-15  1.555e+00   0.000    1.000
## OppLAA       1.595e+01  1.603e+03   0.010    0.992
## OppMIA       1.595e+01  1.964e+03   0.008    0.994
## OppMIN      -1.368e+00  1.342e+00  -1.019    0.308
## OppNYM      -1.162e+00  1.660e+00  -0.700    0.484
## OppNYY      -1.532e+00  1.193e+00  -1.284    0.199
## OppOAK      -2.329e+00  1.403e+00  -1.659    0.097 .
## OppPHI      -1.627e+00  1.491e+00  -1.091    0.275
## OppSEA      -1.303e+00  1.341e+00  -0.971    0.331
## OppTBR      -1.292e+00  1.195e+00  -1.082    0.279
## OppTEX       2.156e-01  1.544e+00   0.140    0.889
## OppTOR      -3.020e-01  1.236e+00  -0.244    0.807
## OppWSN       1.617e+01  2.284e+03   0.007    0.994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 206.23  on 161  degrees of freedom
## Residual deviance: 175.76  on 142  degrees of freedom
## AIC: 215.76
```

```
##
## Number of Fisher Scoring iterations: 16
```

Is there evidence of an association between winning and field after adjusting for opponent? Explain.

In terms of this scenario, discuss why you would adjust for opponent.

Using the model, calculate the estimated probability of the Red Sox beating the Yankees (NYY) at home.

Is the probability you calculated in the last question equal to the observed percent of home games the Red Sox won against the Yankees in the data? In general, when will predicted probabilities equal observed probabilities?

Example 7.1

Credit for the below exercise belongs to LTC Nick Clark.

We tend to do students a disservice in statistics courses because the data is given to you perfectly clean and accurate. This is, of course, not how life really works.

For example, researchers were interested in how Omega-3 index values compare between the Framingham Heart study and different subsets of the U.S. population, so they collected data in seven cities. That data look like:

```
framing.dat<-read.csv("FraminghamOmega3.csv")
head(framing.dat)
```

```
##      O3I    Sex Age Omega3...
## 1 0.0470 Female  56      4.70
## 2 0.0464 Female  66      4.64
## 3 0.0413   Male  75      4.13
## 4 0.0760   Male  68      7.60
## 5 0.0717   Male  75      7.17
## 6 0.0405 Female  82      4.05
```

The first little bit of cleaning I'd do is change the name `Omega3...` for convenience

```
framing.dat <- framing.dat %>%
  mutate(Omega3=Omega3...)%>%
  select(-`Omega3...`)
```

While this isn't necessary, it does help others read my code. The next thing we might do is to determine if we have complete data. If we don't and we just start calculating statistics we might do:

```
mean(framing.dat$Omega3)
```

```
## [1] NA
```

And we see we have problems. Why is R doing this?

To see how many observations have NAs in them we can do the following:

```
complete<-complete.cases(framing.dat)
sum(complete)
```

```
## [1] 2455
```

```
nrow(framing.dat)
```

```
## [1] 2495
```

What's happening here?

Is this a big deal?

To examine which cases are missing data we can do:

```
fram.complete <- framing.dat %>% drop_na()
```

Then we do:

```
fram.mis<- framing.dat %>% filter(is.na(Omega3))
```

Now we've made two datasets, one of the missing data and one that is not missing. The reason we are doing this is we want to explore what sort of missingness we have. The best case scenario is our data are **Missing Completely at Random**.

As an aside this is different than what our book is saying.

But if our data are MCAR, then the values we did measure should be relatively consistent, as a quick check:

```
t.test(fram.complete$Age,fram.mis$Age)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  fram.complete$Age and fram.mis$Age
## t = 3.4518, df = 36.201, p-value = 0.001433
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.021639 7.778701
## sample estimates:
## mean of x mean of y
##  66.48350  61.58333
```

So, perhaps there is some mechanism informing the missingness. We also could look at gender proportions. How would we do this?

In this case, we are probably ok just ignoring the missing data.

Let's look at the second dataset

```
screen.dat<-read.csv("ScreeningsOmega3.csv")
screen.dat <- screen.dat %>%
  mutate(Omega3=Omega3...)%>%
  select(-Omega3...)
head(screen.dat)
```

```
##     Age   O3I  Sex Omega3
## 1 13.35 0.040 Male    4.0
## 2 14.16 0.057 <NA>    5.7
## 3 15.14 0.029 <NA>    2.9
## 4 15.24 0.034 <NA>    3.4
## 5 15.77 0.047 <NA>    4.7
## 6 16.14 0.031 <NA>    3.1
```

It looks like we are missing a ton here.

```
screen.full <- screen.dat %>% drop_na()
nrow(screen.dat)
```

```
## [1] 2178
```

```
nrow(screen.full)
```

```
## [1] 1388
```

To explore this fuller we could summarize it:

```
screen.mis<- screen.dat %>% filter(is.na(Sex))
screen.mis %>% summarize(count=n(),mean.Omeg=mean(Omega3),sd.Omeg=sd(Omega3),mean.age=mean(Age),sd.age=
```

```
##   count mean.Omeg  sd.Omeg mean.age   sd.age
## 1   790  4.357089 1.050074 44.78682 17.02019
```

Which we compare to:

```
screen.full %>% summarize(count=n(),mean.Omeg=mean(Omega3),sd.Omeg=sd(Omega3),mean.age=mean(Age),sd.age=
```

```
##   count mean.Omeg  sd.Omeg mean.age   sd.age
## 1  1388  4.479611 1.291694 49.84399 15.10445
```

Due to a high sample size, if we were to form 95% CI for Omega or Age, what could we say?

So, yeah, different. For now, let's just deal with the complete cases.

We continue to explore the data:

```
screen.full%>%ggplot(aes(x=Omega3))+
  geom_histogram()+
  geom_rug(sides="b")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

What do we see? What should we do?

So now we want to get back to our research question. Is there a difference between the two datasets? Here's how I would combine the datasets:

```
screen.full<- screen.full %>% mutate(Study="S")
fram.complete <- fram.complete %>% mutate(Study="F")
fulldat <- bind_rows(screen.full,fram.complete)
```

Now we can explore

```
fulldat %>% ggplot(aes(x=Study,y=Omega3))+
  geom_boxplot()
```

```
fulldat %>% group_by(Study)%>%
  summarize(mean=mean(Omega3),sd=sd(Omega3),obs=n())
```

```
## # A tibble: 2 x 4
##   Study  mean    sd   obs
##   <chr> <dbl> <dbl> <int>
## 1 F      5.46  1.66  2455
## 2 S      4.48  1.29  1388
```

So, there is a difference in means (again, why can I say this without finding a p-value?)

But, we also have:

```r
fulldat %>% group_by(Study)%>%
  summarize(mean=mean(Age),sd=sd(Age),obs=n())
```

```
## # A tibble: 2 x 4
##   Study  mean    sd   obs
##   <chr> <dbl> <dbl> <int>
## 1 F      66.5  9.10  2455
## 2 S      49.8 15.1   1388
```

What does the below plot suggest for our analysis?

```r
fulldat %>% ggplot(aes(x=Age,y=Omega3))+
  geom_point()+
  stat_smooth(method="lm",se=FALSE)
```

So we are building out our sources of variation diagram mentally. What else could be impacting Omega 3?

```r
fulldat %>% ggplot(aes(x=Sex,y=Omega3))+
  geom_boxplot()
```

So, now we could look at a statistical model to account for the fact that we might need to adjust for Sex and Age if we want to note the true differences in the studies.

The model could be:

But we have another choice, we could put the missing `Sex` values back in to our dataset

```r
screen.dat<- screen.dat %>% mutate(Study="S")
fulldat.mod <- bind_rows(screen.dat,fram.complete)
```

Then we can build the following model:

To fit this in R we have to change the `<NA>` values. (Note we have to be a touch careful with this) As an aside, learn SQL commands and R data structures!

```r
fulldat.mod <- fulldat.mod %>%
  mutate(Sex=as.character(Sex))%>%
  replace_na(list(Sex="Missing"))
```

Now we are ready to go. Again, in typical stats classes this is where we would *start* our lesson…

```r
omega.lm <- lm(Omega3~Age+Sex+Study,data=fulldat.mod)
summary(omega.lm)
```

```
##
## Call:
## lm(formula = Omega3 ~ Age + Sex + Study, data = fulldat.mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3978 -0.9154 -0.2194  0.6556 16.5946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.292864   0.117214  36.624   <2e-16 ***
## Age          0.018332   0.001674  10.950   <2e-16 ***
## SexMale     -0.115910   0.049853  -2.325   0.0201 *
## SexMissing  -0.055035   0.065987  -0.834   0.4043
## StudyS      -0.701759   0.057216 -12.265   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.448 on 4628 degrees of freedom
## Multiple R-squared:  0.1324, Adjusted R-squared:  0.1317
## F-statistic: 176.6 on 4 and 4628 DF,  p-value: < 2.2e-16
```

What's our conclusions here?

Are we done?

```r
omega.lm %>% ggplot(aes(x=.fitted,y=.resid))+
  geom_point()
```

One thing to do here is to repeat the analysis with the outliers removed

```r
fulldat.removed <- fulldat.mod %>% filter(omega.lm$residuals<10)
```

```r
omega.mod.lm <- lm(Omega3~Age+Sex+Study,data=fulldat.removed)
summary(omega.mod.lm)
```

```
##
## Call:
## lm(formula = Omega3 ~ Age + Sex + Study, data = fulldat.removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3956 -0.9132 -0.2116  0.6654  9.4232
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.315273   0.114407  37.719   <2e-16 ***
## Age          0.017942   0.001634  10.979   <2e-16 ***
## SexMale     -0.108054   0.048655  -2.221   0.0264 *  
## SexMissing  -0.033384   0.064414  -0.518   0.6043    
## StudyS      -0.728346   0.055866 -13.037   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.413 on 4626 degrees of freedom
## Multiple R-squared:  0.1397, Adjusted R-squared:  0.1389 
## F-statistic: 187.7 on 4 and 4626 DF,  p-value: < 2.2e-16
```

Do our conclusions change?