LSN 28 - Public Health Screening for Omega-3 (7.1)

MA376 - LTC J.K.Starling

Intro

(0) 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:

```
setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons_AY23-1/BlockIV_New/")
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
```

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

....

(2) 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
```

(3) What's happening here? Is this a big deal?

We have 40 obsevations that are not complete. If we had a large number of individuals with missing O3 Index values it could reduce the sample size. The conclusions we make regarding the O3 levels could be different than if we had the individuals.

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.

We will want to check to see if the distributions for the variables in the complete data set and the missing data are similar:

```
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
```

(4) So, perhaps there is some mechanism informing the missingness. What did we do above? What were the results?

....

(5) We also could look at gender proportions for missing the Omega3 index. How would we do this?

We could create a 2x2 contingency table and then conduct a chi-squared test. To do this in R, we would need to create a new column/variable with an indicator variable where 1 is missing and 0 is not missing. We would create a table with the indicator variable for rows and the sex as the columns. We would then pass the table to the prop.test() function. We see that we have a p value of 0.1077 and we fail to reject the null hypothesis and say that there is not a difference for the number of missing values by sex.

```
framing.dat <- framing.dat %>% mutate(MissingO3 = is.na(Omega3))
contrasts(framing.dat$MissingO3)
```

```
##       TRUE
## FALSE    0
## TRUE     1
```

```
MissO3.tab <- table(framing.dat$MissingO3, as.factor(framing.dat$Sex))
prop.test(MissO3.tab, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  MissO3.tab
## X-squared = 2.5874, df = 1, p-value = 0.1077
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02787101  0.29661169
## sample estimates:
##    prop 1    prop 2
## 0.5510370 0.4166667
```

Since the percentage of missing values relatively small (40 / 2495 = 1.6 %) we are probably ok just ignoring the missing data.

A Second Dataset

Let's look at another 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
```

(6) What observations can we make on this data?

It looks like we may be missing a ton here. We also note that age is decimal (becuse the observers estimated age using actual birthdates).

```
screen.full <- screen.dat %>% drop_na()
screen.mis<- screen.dat %>% filter(is.na(Sex))
c(nrow(screen.dat), nrow(screen.full), nrow(screen.mis))
```

```
## [1] 2178 1388  790
```

(7) We can compare the two subsets of data. What do we notice about these results? Why might there be missing values?

```
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
```

```
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
```

It looks like the complete data set has more observations, higher mean Omega3, higher std dev., higher mean age, and lower std. dev. age. The missing variables were actually from the way the responses were collected; participants were reqired to give birthdate and blood for Omega-3 data, but biological sex was optional.

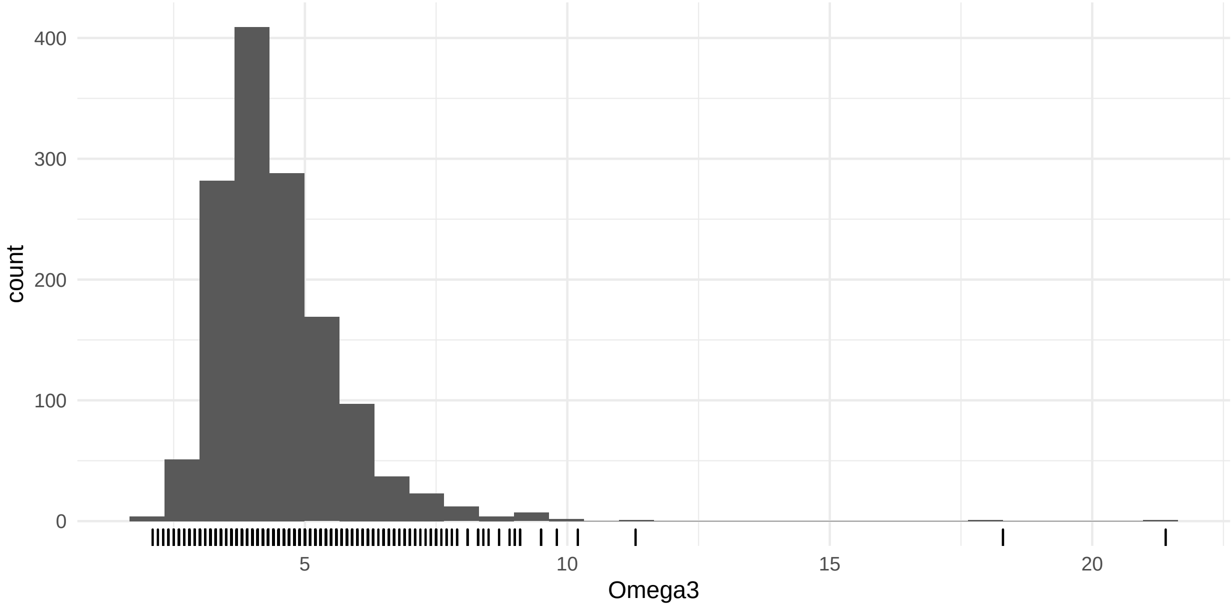
(8) Due to a high sample size, if we were to form 95% CI for Omega or Age, what could we say?

Notice that since we have a large sample size, the denominator will be large, we can say that there is a difference between the two means since the CIs do not overlap.

$$mean \pm t^* \times SE/\sqrt{n}$$

(9) We continue to explore the data. . . . What do we see? What should we do?

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

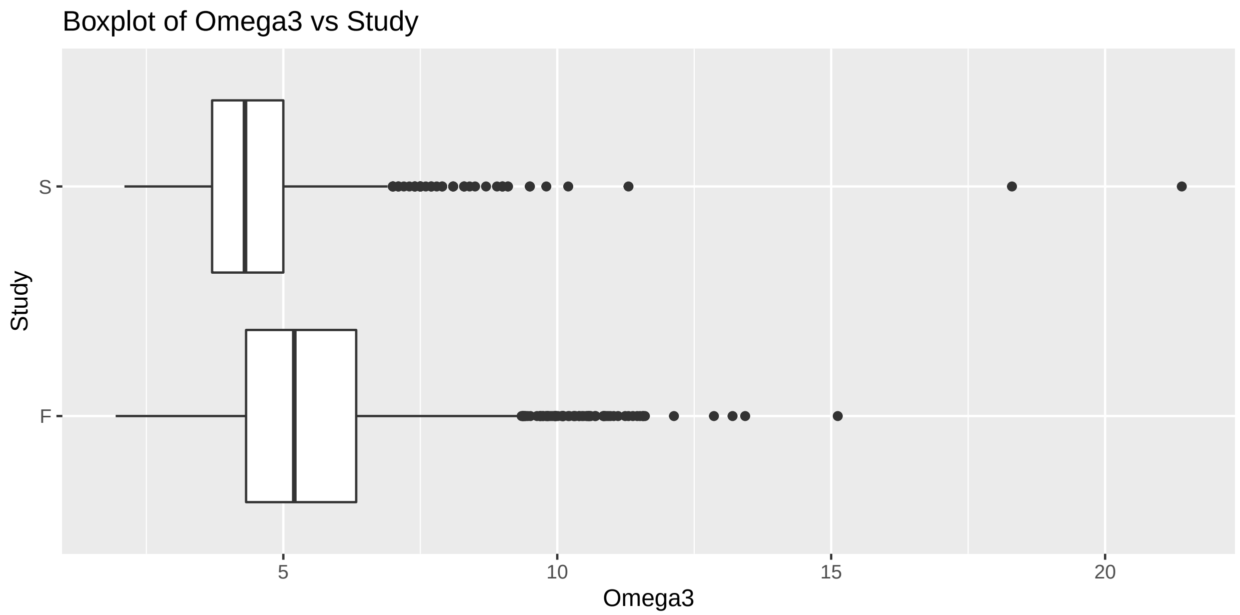


\textcolor{blue}{We see that we have two observations with extereme outliers. With only 2 outliers and over 1000 observations, these 2 individuals will probably not have much impact on our analyses, but they should still be investigated to see whether they indicate any issues with the data collection process. The first question to ask is whether the observations are clearly in error or are they contextually plausible but just unusual? For example, a negative Omega-3 Index value (biologically impossible) needs to be thought of differently than a value of 18%. A second question to ask is whether there are reasons to consider the outliers as arising from a different population. Perhaps those two observations were the only two we had for teenagers. If we removed them from the dataset, we would again restrict the population to which we are willing to generalize. Or, if these observations were the only ones taken by a particular research assistant who quit later that day, we could be justified in removing them from the dataset.}

(10) So now we want to get back to our research question. Is there a difference between the two datasets? Here's a way to combine the datasets and explore.
Is there is a difference in means? (again, why can I say this without finding a p-value?)

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

```
fulldat %>% ggplot(aes(x=Omega3, y=Study))+
  geom_boxplot() + labs(title = "Boxplot of Omega3 vs Study")
```



```
fulldat %>% group_by(Study)%>%
  summarize(omega3.mean=mean(Omega3),sd=sd(Omega3),obs=n())
```

```
## # A tibble: 2 x 4
##   Study omega3.mean    sd   obs
##   <chr>       <dbl> <dbl> <int>
## 1 F            5.46  1.66  2455
## 2 S            4.48  1.29  1388
```

Since we have a large sample size, the CIs for both populations will not overlap and we can say that the populations are different.

(11) The table for statistics for Age for each study is given below. What is the implications of this table?

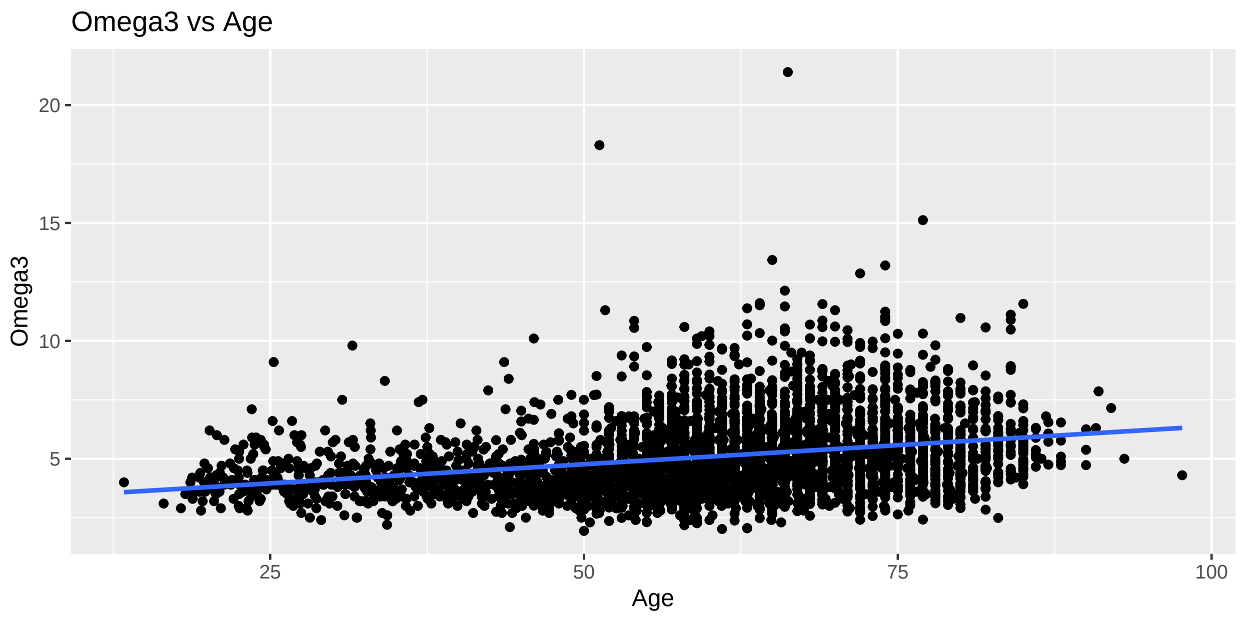
```
fulldat %>% group_by(Study)%>%
  summarize(age.mean=mean(Age),sd=sd(Age),obs=n())
```

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

Since we have a large sample size, the CIs for both populations will not overlap and we can say that the populations are different.

(12) What does the below plot suggest for our analysis?

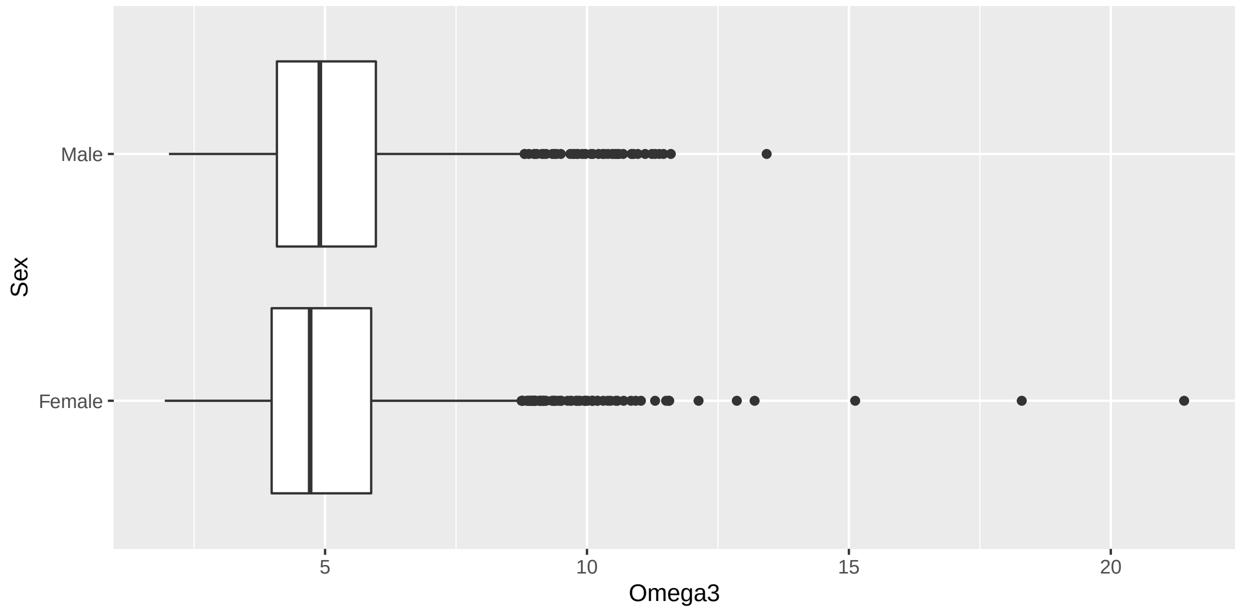
```
fulldat %>% ggplot(aes(x=Age,y=Omega3))+
  geom_point()+
  stat_smooth(method="lm",se=FALSE) + labs(title = "Omega3 vs Age")
```



There appears to be a linear relationship between omega 3 and age

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

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

(14) What models could we build?

$$(omega3)_i = \beta_0 + \beta_1(age)_i + \beta_2(study)_i + \epsilon_i$$

$$(omega3)_i = \beta_0 + \beta_1(age)_i + \beta_2(study)_i + \beta_3(sex)_i + \epsilon_i$$

(15) But we have another choice, we could put the missing `Sex` values back in to our dataset. To add the `Sex` variable in R we have to change the `<NA>` values.

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

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

(16) Now we are ready to go. Again, in typical stats classes this is where we would *start* our lesson... What's our conclusions here?

```
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
```

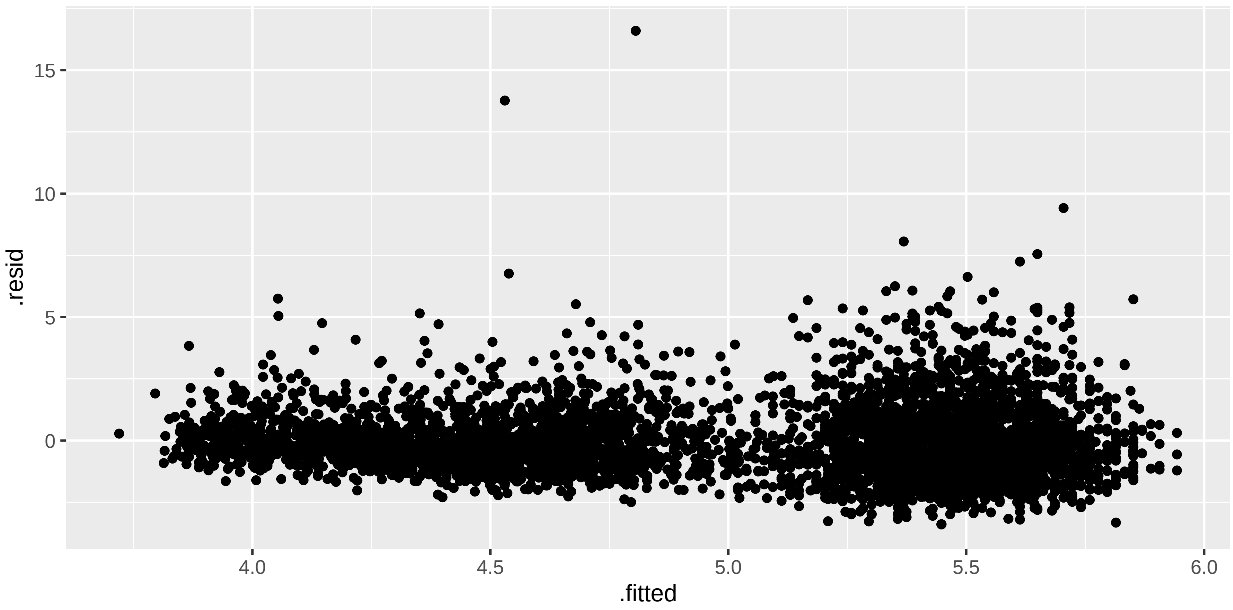
```
Anova(omega.lm,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: Omega3
##             Sum Sq   Df   F value  Pr(>F)
## (Intercept) 2812.7    1 1341.3238 < 2e-16 ***
## Age          251.4    1  119.8980 < 2e-16 ***
## Sex           11.8    2    2.8079 0.06043 .
## Study        315.4    1  150.4302 < 2e-16 ***
## Residuals   9704.7 4628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that the variable Sex is not stastically significant.

Are we done?

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



(17) One thing to do here is to repeat the analysis with the outliers removed. Do our conclusions change?

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

```
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
```

```
Anova(omega.mod.lm,type="III")
```

```
## Anova Table (Type III tests)
```

```
##
## Response: Omega3
##             Sum Sq   Df  F value  Pr(>F)
## (Intercept) 2841.4    1 1422.694 < 2e-16 ***
## Age          240.7    1  120.543 < 2e-16 ***
## Sex            9.9    2    2.478 0.08402 .
## Study        339.5    1  169.974 < 2e-16 ***
## Residuals   9238.9 4626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable Sex is still not stastically significant, and actually did not improve.