STAT 224 Autumn 2022 Homework 2

Matthew Zhao

Question 1

```
fevdata = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/fevdata.txt", h = TRUE)
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

Q1a — 6 points

```
lmm.nosmoke = lm(fev ~ age, data=subset(fevdata, sex == "Male" & smoke == "Nonsmoker"))
summary(lmm.nosmoke)
```

```
##
## Call:
## lm(formula = fev ~ age, data = subset(fevdata, sex == "Male" &
##     smoke == "Nonsmoker"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4850 -0.3506 -0.0438  0.3511  1.8230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0576     0.1147    -0.5     0.62
## age           0.2882     0.0114    25.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.556 on 308 degrees of freedom
## Multiple R-squared:  0.676,  Adjusted R-squared:  0.674
## F-statistic:  641 on 1 and 308 DF,  p-value: <2e-16
```

$\hat{\beta}_0^{mn} = -0.0576$, $s.e.(\hat{\beta}_0^{mn}) = 0.1147$

$\hat{\beta}_1^{mn} = 0.2882$, $s.e.(\hat{\beta}_1^{mn}) = 0.0114$

$\hat{\sigma}^{mn} = 0.556, n = 310$

Q1b — 8 points

$t = \frac{\hat{\beta}_j - \beta_j^0}{s.e.(\hat{\beta}_j)}$

i)

```
t = (-0.0576 - 0)/0.1147
n=654
p=1
pval = pt(t,df=n-p-1,lower.tail = F)
print(paste0("t-stat: ",signif(t,digits = 5)))
```

```
## [1] "t-stat: -0.50218"
```

```
print(paste0("df: ",n-p-1))
```

```
## [1] "df: 652"
```

```
print(paste0("P-value: ",signif(pval,digits = 5)))
```

```
## [1] "P-value: 0.69214"
```

ii)

```
t = (-0.0576 - 0.1)/0.1147
n=654
p=1
pval = 2*pt(abs(t),df=n-p-1,lower.tail = F)
print(paste0("t-stat: ",signif(t,digits = 5)))
```

```
## [1] "t-stat: -1.374"
```

```
print(paste0("df: ",n-p-1))
```

```
## [1] "df: 652"
```

```
print(paste0("P-value: ",signif(pval,digits = 5)))
```

```
## [1] "P-value: 0.16991"
```

iii)

```r
t = (0.2882 - 0.1)/0.0114
n=654
p=1
pval = pt(t,df=n-p-1)
print(paste0("t-stat: ",signif(t,digits = 5)))
```

```
## [1] "t-stat: 16.509"
```

```r
print(paste0("df: ",n-p-1))
```

```
## [1] "df: 652"
```

```r
print(paste0("P-value: ",signif(pval,digits = 5)))
```

```
## [1] "P-value: 1"
```

iv)

```r
t = (0.2882 - 0.3)/0.0114
n=654
p=1
pval = pt(t,df=n-p-1,lower.tail = F)
print(paste0("t-stat: ",signif(t,digits = 5)))
```

```
## [1] "t-stat: -1.0351"
```

```r
print(paste0("df: ",n-p-1))
```

```
## [1] "df: 652"
```

```r
print(paste0("P-value: ",signif(pval,digits = 5)))
```

```
## [1] "P-value: 0.84949"
```

Q1c — 4 points

$\hat{\beta}_j \pm t_{(n-p-1,\frac{\alpha}{2})} * s.e.(\hat{\beta}_j)$

```r
beta_hat = 0.2882
se = 0.0114
n=nrow(subset(fevdata, sex == "Male" & smoke == "Nonsmoker"))
p=1
alpha = 0.1
t = qt(alpha/2,df=n-p-1,lower.tail = F)
print(paste0("confidence interval: ",'(',
             signif(beta_hat-t*se,digits=5),',',
             signif(beta_hat+t*se,digits=5),')'))
```

```
## [1] "confidence interval: (0.26939,0.30701)"
```

We are 90% confident that β_1^{mn} is between 0.26939 and 0.30701.

Q1d — 5 points

```r
aggregate(age ~ sex + smoke, data=fevdata, mean)
```

```
##      sex     smoke    age
## 1 Female Nonsmoker  9.366
## 2   Male Nonsmoker  9.687
## 3 Female    Smoker 13.256
## 4   Male    Smoker 13.923
```

```r
aggregate(age ~ sex + smoke, data=fevdata, sd)
```

```
##      sex     smoke   age
## 1 Female Nonsmoker 2.693
## 2   Male Nonsmoker 2.778
## 3 Female    Smoker 2.245
## 4   Male    Smoker 2.465
```

$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(n-2,\frac{\alpha}{2})}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

```r
est = lmm.nosmoke$coefficients[1] + lmm.nosmoke$coefficients[2]*18
est
```

```
## (Intercept)
##        5.13
```

Estimate: $\hat{\beta}_0 + \hat{\beta}_1 x_0 = 5.13$

```
n = nrow(subset(fevdata, sex == "Male" & smoke == "Nonsmoker"))
t = qt(0.05/2,n-2,lower.tail = F)
sig_hat = 0.556
x_0 = 18
x_bar = 9.687
sd = 2.778
num = (x_0-x_bar)^2
denom = (n-1)*sd^2
int = t*sig_hat*sqrt((1/n)+(num/denom))
int
```

```
## [1] 0.1963
```

```
print(paste0('CI: (',signif(est-int,digits=5),',',signif(est+int,digits=5),')'))
```

```
## [1] "CI: (4.9339,5.3266)"
```

Q1e — 4 points

$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(n-2,\frac{\alpha}{2})}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

```
est = lmm.nosmoke$coefficients[1] + lmm.nosmoke$coefficients[2]*14
est
```

```
## (Intercept)
##       3.977
```

Estimate: $\hat{\beta}_0 + \hat{\beta}_1 x_0 = 3.977$

```
n = nrow(subset(fevdata, sex == "Male" & smoke == "Nonsmoker"))
t = qt(0.05/2,n-2,lower.tail = F)
sig_hat = 0.556
x_0 = 14
x_bar = 9.687
sd = 2.778
num = (x_0-x_bar)^2
denom = (n-1)*sd^2
int = t*sig_hat*sqrt((1/n)+(num/denom))
int
```

```
## [1] 0.1149
```

```
print(paste0('CI: (',signif(est-int,digits=5),',',signif(est+int,digits=5),')'))
```

```
## [1] "CI: (3.8625,4.0923)"
```

The interval for Q1d is wider.

Q1f — 5 points

$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(n-2,\frac{\alpha}{2})}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

```
est = lmm.nosmoke$coefficients[1] + lmm.nosmoke$coefficients[2]*14
est
```

```
## (Intercept)
##       3.977
```

Estimate: $\hat{\beta}_0 + \hat{\beta}_1 x_0 = 3.977$

```
n = nrow(subset(fevdata, sex == "Male" & smoke == "Nonsmoker"))
t = qt(0.05/2,n-2,lower.tail = F)
sig_hat = 0.556
x_0 = 14
x_bar = 9.687
sd = 2.778
num = (x_0-x_bar)^2
denom = (n-1)*sd^2
int = t*sig_hat*sqrt(1+(1/n)+(num/denom))
int
```

```
## [1] 1.1
```

```
print(paste0('CI: (',signif(est-int,digits=5),',',signif(est+int,digits=5),')'))
```

```
## [1] "CI: (2.8774,5.0775)"
```

The interval for this question is much larger since we are making a point prediction for a specific individual rather than for the average.

Q1g — 6 points

```
lmm.nosmoke.female = lm(fev ~ age, data=subset(fevdata, sex == "Female" & smoke == "Nonsmok
summary(lmm.nosmoke.female)
```

```
##
## Call:
## lm(formula = fev ~ age, data = subset(fevdata, sex == "Female" &
##     smoke == "Nonsmoker"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0984 -0.2826 -0.0135  0.2374  1.0972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67387    0.08913    7.56  5.9e-13 ***
## age          0.18209    0.00915   19.91  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.411 on 277 degrees of freedom
## Multiple R-squared:  0.589, Adjusted R-squared:  0.587
## F-statistic:  396 on 1 and 277 DF,  p-value: <2e-16
```

$\hat{\beta}_0^{fn} = 0.67387, \hat{\beta}_1^{fn} = 0.18209$

$\hat{\beta}_j \pm t_{(n-p-1,\frac{\alpha}{2})} * s.e.(\hat{\beta}_j)$

```
beta_hat = 0.18209
se = 0.00915
n=nrow(subset(fevdata, sex == "Female" & smoke == "Nonsmoker"))
p=1
alpha = 0.1
t = qt(alpha/2,df=n-p-1,lower.tail = F)
print(paste0("confidence interval: ",'(',
             signif(beta_hat-t*se,digits=5),',',
             signif(beta_hat+t*se,digits=5),')'))
```

```
## [1] "confidence interval: (0.16699,0.19719)"
```

The confidence interval for boys is (0.26939, 0.30701) while it is (0.16699,0.19719) for girls. Since there is no overlap, the lung capacity for boys most likely grows faster than for girls.

Question 2

```r
NLSY = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt", header=T)
NLSYm = subset(NLSY, Gender == "male")
```

Q2a — 6 points

```r
lm1 = lm(log(Income2005) ~ AFQT, data=NLSYm)
lm2 = lm(log(Income2005) ~ AFQT + Edu2006, data=NLSYm)
```

For the first model the coefficient of **AFQT** is `lm1$coefficients[1]` while for the second model it is `lm2$coefficients[1]`.

The coefficients are different since in the second model we also have education as a covariate. This means that the interpretation of the regression coefficient for **AFQT** is different, since in the second model it is interpreted as the change in log income from an increase of 1 percentile on the AFQT for a given level of education.

Q2b — 5 points

```r
yres = lm(log(Income2005) ~ Edu2006, data=NLSYm)$res
tres = lm(AFQT ~ Edu2006, data=NLSYm)$res
lm(yres ~ tres)$coef
```

```
## (Intercept)        tres
##  -3.023e-17   6.738e-03
```

Q2c — 2 points

```r
sst = sum((log(NLSYm$Income2005) - mean(log(NLSYm$Income2005)))^2)
ssr = sum((lm1$fitted.values - mean(log(NLSYm$Income2005)))^2)
ssr/sst
```

```
## [1] 0.1221
```

```r
sst = sum((log(NLSYm$Income2005) - mean(log(NLSYm$Income2005)))^2)
ssr = sum((lm2$fitted.values - mean(log(NLSYm$Income2005)))^2)
ssr/sst
```

```
## [1] 0.1491
```

The multiple R^2 values obtained mean that model 1 can explain 12.21% of the variation in Y and that model 2 can explain 14.91% of the variation in Y.