Homework 3

Lydia Strebe

July 14, 2019

Problem 1

(a)

```
CherryPartial=read.csv("http://users.stat.umn.edu/~parky/CherryPartial.csv",
header=TRUE)

cherry_mod1=lm(time~age, data=CherryPartial)

summary(cherry_mod1)

##
## Call:
## lm(formula = time ~ age, data = CherryPartial)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2657.4  -690.9    -7.5   634.9  4766.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5581.523     40.013 139.494  < 2e-16 ***
## age            6.528      1.043   6.261 4.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1025 on 8601 degrees of freedom
## Multiple R-squared:  0.004537,   Adjusted R-squared:  0.004421
## F-statistic:  39.2 on 1 and 8601 DF,  p-value: 4.016e-10
```

The estimated slope of 6.528 means that for every year increase in age, the expected time increases by 6.528 seconds.

(b)

```
cherry_mod2=lm(time~age+state, data=CherryPartial)
```

(i) The estimated slope of 6.452 for age means that, after controlling for the effect of state, for every year increase in age, the expected time increases by 6.452 seconds.

(ii) The estimated slope of 2774.321 for stateND means that, after controlling for age, the expected time for a participant from North Dakota is 6467.806 (the intercept) plus

2774.321 seconds. In other words, it's the difference between the average time of a runner from Wyoming (the default state) and the average time of a runner from North Dakota after controlling for age.

(iii) The fitted model for a runner from MN is t^=6467.806-528.129+6.452*age or t^=5939.677+6.452*age (t^ represents estimated time).

(iv) The model time ~ age + state looks like B) 50 parallel lines, one line for each state. They are parallel because they all have the same slope: 6.452 (the effect of age), but with a different intercept.

```
cherry_mod3=lm(time~age+sex+age:sex, data=CherryPartial)

summary(cherry_mod3)

##
## Call:
## lm(formula = time ~ age + sex + age:sex, data = CherryPartial)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2697.1  -639.6   -30.5   588.8  4658.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5644.281     54.913 102.786   <2e-16 ***
## age           15.145      1.549   9.775   <2e-16 ***
## sexM        -807.682     77.269 -10.453   <2e-16 ***
## age:sexM       1.116      2.037   0.548    0.584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 956 on 8599 degrees of freedom
## Multiple R-squared:  0.1347, Adjusted R-squared:  0.1344
## F-statistic: 446.3 on 3 and 8599 DF,  p-value: < 2.2e-16
```

(i) The fitted model for a female runner is t^=5644.281+15.145*age where t^ represents estimated time.

(ii) The fitted model for a male runner is
t^=5644.281-807.682+15.145*age+1.116*age*(sexM)
or t^=4836.599+16.261*age
where t^ represents estimated time.

(e) After controlling for age and sex individually, the p-value for the interaction between age and sex is 0.584, which is not statistically significant (it is above a 0.1 significance level).

(f) H_0: Neither sex nor the interaction of age and sex are related to time (after controlling for age). I.e., $\beta_2 = \beta_3 = 0$

H_A: At least one of the regressors is related to time (after controlling for age). I.e., at least one beta is *not* equal to zero.

```
anova(cherry_mod1,cherry_mod3)

## Analysis of Variance Table
##
## Model 1: time ~ age
## Model 2: time ~ age + sex + age:sex
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   8601 9041274414
## 2   8599 7858709452  2 1182564962 646.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(g) We cannot perform a partial F-test to compare the two models from part c) and part d) because neither of the models are a subset of the other. They both contain at least one regressor that the other model does not have.

Problem 2

(a) S^2=RSS/(n-p-1)=7840/(36-3)

```
7840/(36-3)

## [1] 237.5758
```

(b) R^2=SSreg/SST=9350/17190

```
R2=9350/17190
R2

## [1] 0.5439209
```

(ii) This means that the amount of variability in breakfast cereal calories that can be explained by the model is about 54%.

(iii) R^2_{adj}=1-(1-R^2)((n-1)/(n-p-1))

```
1-(1-R2)*(35/33)

## [1] 0.5162797
```

(c) H_0: $\beta_1 = \beta_2 = 0$

H_A: At least one beta does not equal 0

F=(SSreg/p)/(RSS/33)

```
F=(9350/2)/(7840/33)
F

## [1] 19.67793

pf(q=F,df1=2,df2=33,lower.tail = FALSE)
```

```
## [1] 2.366808e-06
```

Since the p-value is very low (less than 0.05), we reject the null hypothesis. At least one of the predictors is related to calorie content.