exercises_week4

Lindsey Greenhill

2/17/2021

Question 4.2

```
##       earn           height1         height2           sex
##  Min.   :     0   Min.   :4.000   Min.   : 0.000   Min.   :1.000
##  1st Qu.:  6000   1st Qu.:5.000   1st Qu.: 3.000   1st Qu.:1.000
##  Median : 16400   Median :5.000   Median : 5.000   Median :2.000
##  Mean   : 20015   Mean   :5.122   Mean   : 5.186   Mean   :1.631
##  3rd Qu.: 28000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:2.000
##  Max.   :200000   Max.   :6.000   Max.   :98.000   Max.   :2.000
##  NA's   :650      NA's   :8       NA's   :6
##       race            hisp            ed            yearbn
##  Min.   :1.000   Min.   :1.000   Min.   : 2.00   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:12.00   1st Qu.:34.00
##  Median :1.000   Median :2.000   Median :12.00   Median :50.00
##  Mean   :1.187   Mean   :1.953   Mean   :13.31   Mean   :46.98
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:15.00   3rd Qu.:60.00
##  Max.   :9.000   Max.   :9.000   Max.   :99.00   Max.   :99.00
##
##      height
##  Min.   :57.00
##  1st Qu.:64.00
##  Median :66.00
##  Mean   :66.56
##  3rd Qu.:69.00
##  Max.   :82.00
##  NA's   :8
```

Part b

```
## lm(formula = earn ~ height, data = df)
##             coef.est  coef.se
## (Intercept) -84078.32   8901.10
## height        1563.14    133.45
## ---
## n = 1379, k = 2
## residual sd = 18853.92, R-Squared = 0.09

## lm(formula = earn ~ avg_height, data = df)
##             coef.est coef.se
## (Intercept) 20014.86   507.71
## avg_height   1563.14   133.45
```

```
## ---
## n = 1379, k = 2
## residual sd = 18853.92, R-Squared = 0.09
```

fit_1 is a regression model predicting earnings from height. The intercept of the model is -84078.3. In context, this means that a person who is 0 inches tall on average makes -84078.3, however, because nobody is 0 inches tall, this interpretation doesn't really make sense. The height coefficient is 1563.1. In context, this means that as height increases by 1 inch, earnings increase by 1563.1 on average. This coefficient is statistically significant, as it is outside 2 standard errors from 0 (the t statistic is 11.7). This tells us that height matters in predicting earnings. However, the R squared of this model is .09, meaning that the model accounts for about 9% of the variability. This isn't a very high R squared value, telling us that we are leaving out important predictors of earnings.

In order to intepret the intercept from this model as average earnings for people with average height, we should transform the height variable to be centered at its mean. To do this, I created a new variable called avg_height which is equal to the height variable minus the mean of the height variable. So, a person of average height will have an avg_height value of 0. This new model (fit_2) has an intercept of 20014.86, meaning that a person with avg_height of 0, or a person with average height, makes 20014.86 on average. The avg_height coefficient is 1563.14. In context, this means that as avg_height increased by 1, or if a person is an inch taller than the average height, earnings increases by 1563.14 on average. This coefficeint is statistically significant, as it is outside 2 standard errors from 0. The R squared for this model is also .09, meaning that the model accounts for about 9% of the variability.

Part c

```r
# different model combinations

fit_3 <- lm(earn ~ avg_height + sex, data = df)
summary(fit_3)
```

```
##
## Call:
## lm(formula = earn ~ avg_height + sex, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30553 -12448  -3243   7451 171098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27025.5     1030.4  26.228  < 2e-16 ***
## avg_height     550.5      184.6   2.983  0.00291 **
## sex         -11254.6     1448.9  -7.768 1.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18460 on 1376 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1275
## F-statistic: 101.7 on 2 and 1376 DF,  p-value: < 2.2e-16
```

```r
fit_4 <- lm(earn ~ avg_height + sex + avg_height*sex, data = df)
summary(fit_4)
```

```
##
## Call:
## lm(formula = earn ~ avg_height + sex + avg_height * sex, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31209 -12591  -3172   7223 171109
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     26259.2     1248.0  21.041  < 2e-16 ***
## avg_height        772.4      275.0   2.809  0.00505 **
## sex           -10868.3     1491.6  -7.286 5.36e-13 ***
## avg_height:sex   -403.7      371.0  -1.088  0.27670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18460 on 1375 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.1277
## F-statistic: 68.22 on 3 and 1375 DF,  p-value: < 2.2e-16
```

```r
fit_5 <- lm(earn ~ sex, data = df)
summary(fit_5)
```

```
##
## Call:
## lm(formula = earn ~ sex, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28927 -12927  -2927   7380 171073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28926.9      811.9   35.63   <2e-16 ***
## sex         -14307.0     1028.6  -13.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18510 on 1377 degrees of freedom
## Multiple R-squared:  0.1232, Adjusted R-squared:  0.1225
## F-statistic: 193.4 on 1 and 1377 DF,  p-value: < 2.2e-16
```

I fit three different models for this question. The first, fit_3, is a regression model predicting earnings from avg_height and sex. The second, fit_4, is a regression model predicting earnings from avg_height, sex, and the interaction of avg_height and sex. The third, fit_5, is a regression model predicting earnings from sex. For reference, sex = 0 is male and sex = 1 is female.

fit_3 interpretation:

- The intercept is 27025.5, meaning that a person of average height and male makes 27025.5 on average.
- The avg_height coefficient is 550.5, meaning that as avg_height increases by 1, earnings increase by 550.5 on average, holding sex constant. This coefficient is statistically significant.
- The sex coefficient is -11254.6, meaning that being female decreases earnings by 11254.6 on average, holding avg_height constant. This coefficient is statistically significant.
- The R squared is .129, meaning that the model accounts for about 12.9% of the variability

fit_4 interpretation:

- The intercept is 26259.2, meaning that a person of average height and male sex makes 26259.2 on average.
- The avg_height coefficient is 772.4, meaning that as avg_height increases by 1, earnings increase by 772.4 on average, holding sex constant. This coefficient is statistically significant.
- The sex coefficient is -10868.3, meaning that being female decreases earnings by 10868.3 on average, holding avg_height constant. This coefficient is statistically significant.
- The interaction term coefficient is -403.7, meaning that a 1 inch increase in height has an additioinal 403.7 decrease on earnings if you are female. This interaction is not statistically significant.
- The R squared is .1296, meaning that the model accounts for about 12.99% of the variability

fit_5 interpretation:

- The intercept is 28926.9, meaning that a person of average height and male sex makes 28926.9 on average.
- The sex coefficient is -14307, meaning that being female decreases earnings by -14307 on average. This coefficient is statistically significant.
- The R squared is .1232, meaning that the model accounts for about 12.32% of the variability

model choice

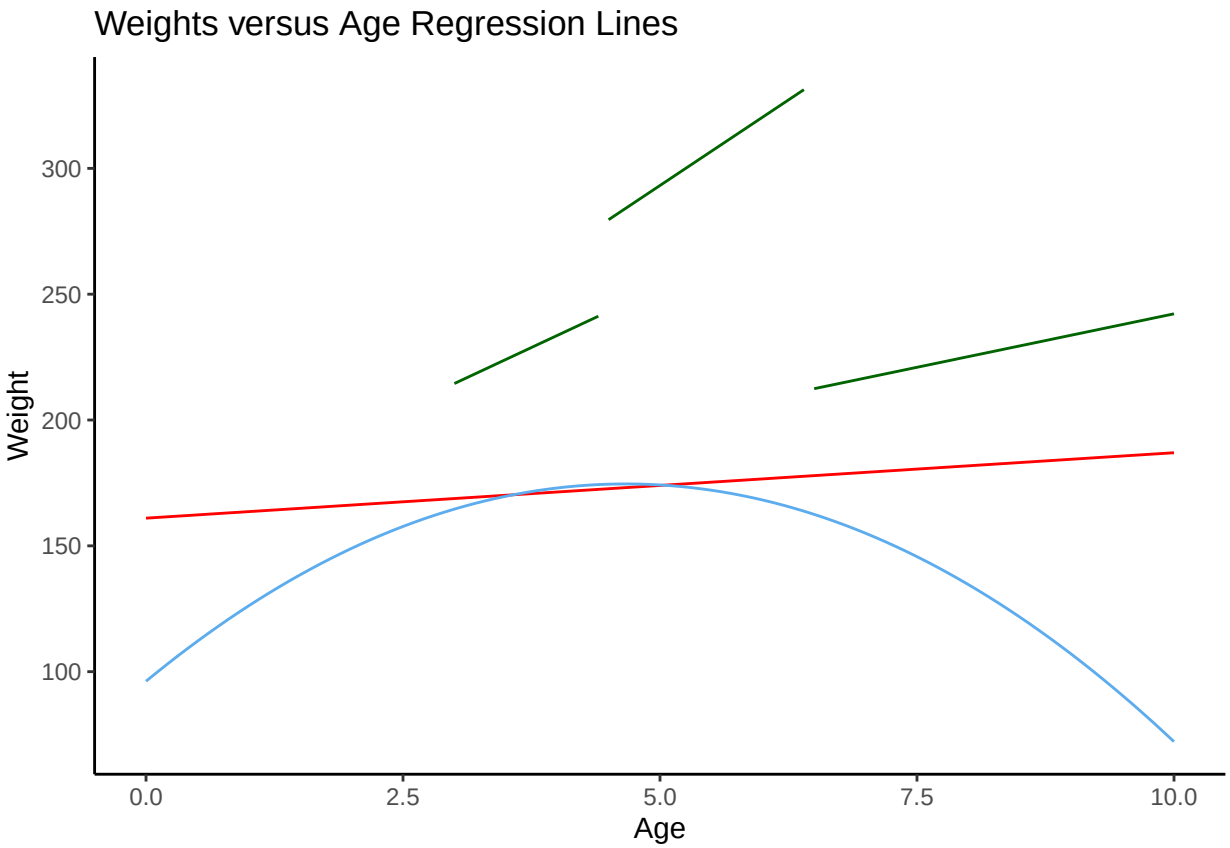
Interestingly, all of these models have very similar R squared values, so it is difficult to decide which model to use based off of that. With that being said, the models which incorporate height and sex (fit_3 and fit_4) have slightly higher R squared values, so I am inclined to prefer one of those. In between fit_3 and fit_4, it doesn't seem like the interaction adds anything substantial to the model, as the interaction coefficient is not statistically significant, so for parsimony's sake, I would go with fit_3.

Part d

See part c for coefficient interpretations.

Question 4.3

Part a


Weights versus Age Regression Lines
Weight
300
250
200
150
100
0.0
2.5
5.0
7.5
10.0
Age