exercises_week4

Lindsey Greenhill

2/17/2021

Question 4.2

```
df <- foreign::read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/earnings/heights.dta")

# there's some weird coding within the height1 and height2. seems like height 1
# is in feet and height2 is remaining inches. Height seems to be total inches.
# Also, there are two outliers in height2 where values are 98. 2 different
# options on what to do with this data. I think it would be best to delete them
# because there are no values for the earn variable for those observations
# anyways so we wouldn't use them in the regression. I'm also changing the sex
# variable to be coded as 0 or 1 because that is more the norm that I see.

df <- df %>%
  mutate(sex = sex - 1) %>%
  filter(!(is.na(earn)), ! (is.na(height)))
```

Part b

```
# the model below predicts earnings from height

fit_1 <- lm(earn ~ height, data = df)

# the coefs seem to be statistically significant according to the summary at a
# .001 level. The R squared is .09. The model accounts for 9% of the
# variability. What this tells us is that we are leaving out important variables
# that predict earnings more than height. So height seems to matter but it is a
# small fraction of the actual story.

stargazer(fit_1,
          type = "latex",
          title = "Summary Table Fit 1")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Feb 23, 2021 - 22:57:43

```
# what this does is making the variable centered at 0. They are distances from
# the mean. Also have to filter out the NA values because we don't deal with
# those until later in the book. When you have 0 height, the average outcome is
```

Table 1: Summary Table Fit 1

	Dependent variable:
	earn
height	1,563.138***
	(133.448)
Constant	$-$84,078.320***
	(8,901.098)
Observations	1,379
R^2	0.091
Adjusted R^2	0.090
Residual Std. Error	18,853.920 (df = 1377)
F Statistic	137.206*** (df = 1; 1377)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# the intercept

df <- df %>%
  mutate(avg_height = height - mean(height),
         avg_earn = earn - mean(earn))

fit_2 <- lm(earn ~ avg_height, data = df)

stargazer(fit_2,type = "latex",
          title = "Summary Table Fit 2")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Feb 23, 2021 - 22:57:43

Table 2: Summary Table Fit 2

	Dependent variable:
	earn
avg_height	1,563.138***
	(133.448)
Constant	20,014.860***
	(507.714)
Observations	1,379
R^2	0.091
Adjusted R^2	0.090
Residual Std. Error	18,853.920 (df = 1377)
F Statistic	137.206*** (df = 1; 1377)
Note:	*p<0.1; **p<0.05; ***p<0.01

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)
stargazer(fit_3,type = "latex", title = "Summary Table Fit 3")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Feb 23, 2021 - 22:57:43

Table 3: Summary Table Fit 3

	Dependent variable:
	earn
avg_height	550.545^{***}
	(184.570)
sex	$-11,254.570^{***}$
	(1,448.892)
Constant	$27,025.500^{***}$
	(1,030.387)
Observations	1,379
R^2	0.129
Adjusted R^2	0.128
Residual Std. Error	18,460.370 (df = 1376)
F Statistic	101.728^{***} (df = 2; 1376)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
fit_4 <- lm(earn ~ avg_height + sex + avg_height*sex, data = df)
stargazer(fit_4,type = "latex", title = "Summary Table Fit 4")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Feb 23, 2021 - 22:57:43

Table 4: Summary Table Fit 4

	Dependent variable:
	earn
avg_height	772.431***
	(275.023)
sex	−10,868.300***
	(1,491.646)
avg_height:sex	−403.668
	(370.951)
Constant	26,259.170***
	(1,247.989)
Observations	1,379
R^2	0.130
Adjusted R^2	0.128
Residual Std. Error	18,459.130 (df = 1375)
F Statistic	68.222*** (df = 3; 1375)
Note:	*p<0.1; **p<0.05; ***p<0.01

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


stargazer(fit_5,type = "latex", header = FALSE,
          title = "Summary Table Fit 5")
```

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:

Table 5: Summary Table Fit 5

	Dependent variable:
	earn
sex	−14,307.020***
	(1,028.646)
Constant	28,926.920***
	(811.859)
Observations	1,379
R^2	0.123
Adjusted R^2	0.123
Residual Std. Error	18,513.230 (df = 1377)
F Statistic	193.449*** (df = 1; 1377)
Note:	*p<0.1; **p<0.05; ***p<0.01

- 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

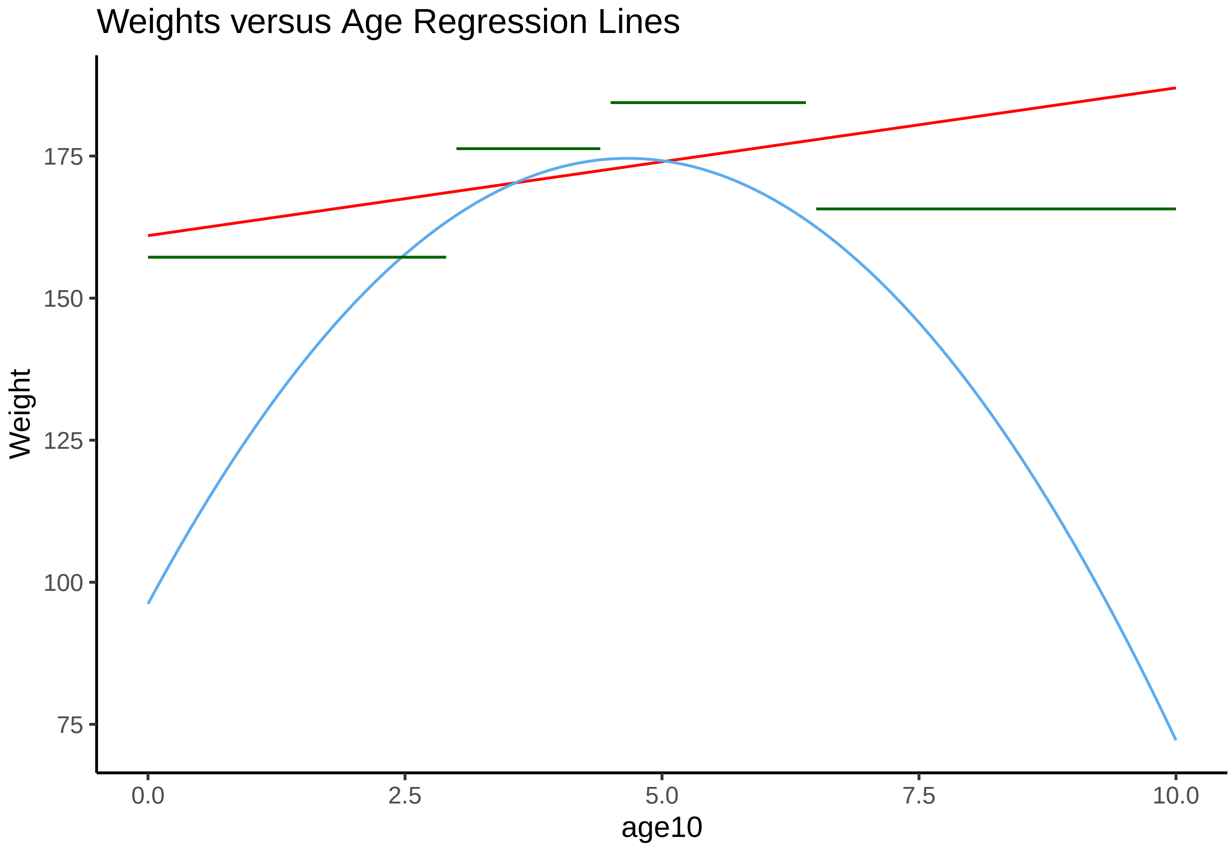
Interestingly, all of these models have very similar R squared values, so it is difficult to decide which model to use based off of that. With that being said, the models which incorporate height and sex (fit_3 and fit_4) have slightly higher R squared values, so I am inclined to prefer one of those. 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,b,c



Weights versus Age Regression Lines
Weight
175
150
125
100
75
0.0
2.5
5.0
7.5
10.0
age10