**Name:** AADITI ROKADE **Date:** 11/20/2018

**SCORE:**

**Correlation and Regression:**

Use the GSS06 dataset and answer the following questions about the relationship between respondent’s years of education ***(‘educ’)*** and mother’s years of education ***(‘maeduc’)****.*

**Scatterplot: Respondent’s years of Education *vs* mother’s years of Education.**

1. ***What is the Pearson’s r equal to?***ANS*: 0.*4773944  
   **Interpretation**:   
   >> r = 0 indicates no linear relationship between variables  
   >> r=1 represents a perfect +ve relationship  
   >> r=-1 represents a perfect -ve relationship

***F***rom the Pearson’s r value, we can say that the there’s a moderate positive relationship between respondent’s years of education and mother’s years of education.

1. ***What is the y intercept equal to?***  *9.*3496738

**Linear Regression Output:**

**> # regression line for 'educ' and 'maeduc'**

**> educ\_model\_06 <- lm(educ ~ maeduc, data = m\_educ\_2006)**

**> # you can get just the coefficients of your model**

**> coef(educ\_model\_06)**

(Intercept) maeduc

9.3496738 0.3765627

> # you can also get a full summary

> summary(educ\_model\_06)

Call:

lm(formula = educ ~ maeduc, data = m\_educ\_2006)

Residuals:

Min 1Q Median 3Q Max

-13.3747 -1.8684 0.0019 1.8784 9.1441

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.34967 0.16166 57.84 <2e-16 \*\*\*

maeduc 0.37656 0.01352 27.86 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.776 on 2621 degrees of freedom

Multiple R-squared: 0.2284, Adjusted R-squared: 0.2281

F-statistic: 775.9 on 1 and 2621 DF, p-value: < 2.2e-16

1. ***For each additional year of education for the mother, what is the expected increase in respondent’s years of education? Are the results significant?***

Ans: for each additional year of education for the mother, the expected increase in respondent’s years of education is slope (0.3765627 ) times years plus Y intercept i.e. 9.3496738

Y= . 9.3496738 + 0.3765627 x (a is intercept and b is slope)

**Results are very significant as indicated by \*\*\* => 0 i.e. confidence level 100%**

1. ***What is the predicted value for the respondent’s years of education when mother’s years of education are equal to 14?***

**Predicted:**

**Predicted ‘educ”**

**Predicted ‘educ” years.**

Y = 9.3496738 +0.3765627 \*14 = 14.62155

Now split the database by **‘race’** and use the results to answer the same questions as above (in table from below). Use the table also to answer questions 6 -10.

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear Regression Results for Education by mother’s Education by Race** | | | |
| **Statistic** | **White** | **Black** | **Other** |
| Pearson’s r | 0.3883788 | 0.3502304 | 0.5895074 |
| y-intercept | 10.22953 | 9.97414 | 8.08603 |
| Gain per mother’s years of education | Y=10.22953 +0.31641 x | Y=9.97414 +0.27555x | Y=8.08603 +0.48469x |
| Significance level | 0 ‘\*\*\*’ | 0 ‘\*\*\*’ | 0 ‘\*\*\*’ |

1. ***For which of these three groups is the correlation strongest?*** *As indicated value of Pearson’s r* 0.3883788 (white) , 0.3502304 (black) and 0.5895074 (other) , since **OTHER** group has the highest value among the groups, the correlation is strongest for OTHER group.
2. ***Is the association positive or negative?***

A positive value of Pearson’s r suggests that the association is positive for all the groups.

1. ***For which group does mother’s years of education produce the greatest gains in respondent’s years of education?***Since the group WHITE has the highest Y-intercept, mother’s years of education produces the greatest gains in respondent’s years of education for **WHITE** group.
2. ***For whites, what percent of the variance in ‘educ’ can be explained by ‘maeduc’?***

Multiple **R-squared: 0.1508**, Adjusted R-squared: 0.1504  
  
Hence for whites, **15.08%** of variance in ‘educ’ can be explained by ‘maeduc’  
R-squared is simply square of Pearson’s r. It indicates how close the values are to the regression line.

% variance of the dependent variable is reflected in the correlation coefficient R-squared.

1. *For blacks, what percent of the variance in* ***‘educ’*** *can be explained by* ***‘maeduc’****?*

Multiple **R-squared: 0.1227**, Adjusted R-squared: 0.1201

Hence for blacks, **12.27%** of variance in ‘educ’ can be explained by ‘maeduc’  
R-squared is simply square of Pearson’s r. It indicates how close the values are to the regression line.

% variance of the dependent variable is reflected in the correlation coefficient R-squared.

Copy and paste all appropriate R output here. *Please add space as needed.*

**> m\_educ\_2006 <- gss2006 %>%**

**+ select(educ, maeduc, race) %>%**

**+ filter (!is.na(educ)) %>%**

**+ filter (!is.na(maeduc))**

Warning message:

package ‘bindrcpp’ was built under R version 3.4.4

**> ggplot(m\_educ\_2006, aes(x = maeduc, y = educ)) +**

**+ geom\_jitter(size = 1.5) +**

**+ labs(x = "Mother's Years of Education", y = "Years of Education",**

**+ title = "Relationship of Education and Mother's Education (2006)")**

**> # 1. What is the Pearson's r equal to?**

**> cor.test(m\_educ\_2006$educ, m\_educ\_2006$maeduc)**

Pearson's product-moment correlation

data: m\_educ\_2006$educ and m\_educ\_2006$maeduc

t = 27.856, df = 2621, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.447854 0.506934

sample estimates:

cor

0.4779344

**> # 2. What is the y intercept equal to?**

**> # regression line for 'educ' and 'maeduc'**

**> educ\_model\_06 <- lm(educ ~ maeduc, data = m\_educ\_2006)**

**> # you can get just the coefficients of your model**

**> coef(educ\_model\_06)**

(Intercept) maeduc

9.3496738 0.3765627

> # you can also get a full summary

> summary(educ\_model\_06)

Call:

lm(formula = educ ~ maeduc, data = m\_educ\_2006)

Residuals:

Min 1Q Median 3Q Max

-13.3747 -1.8684 0.0019 1.8784 9.1441

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.34967 0.16166 57.84 <2e-16 \*\*\*

maeduc 0.37656 0.01352 27.86 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.776 on 2621 degrees of freedom

Multiple R-squared: 0.2284, Adjusted R-squared: 0.2281

F-statistic: 775.9 on 1 and 2621 DF, p-value: < 2.2e-16

**> ggplot(m\_educ\_2006, aes(x = maeduc, y = educ)) +**

**+ geom\_jitter(size=1.5, aes(color = race, shape = race)) +**

**+ labs(x = "Mother's Years of Education", y = "Years of Education",**

**+ title = "Relationship of Education and Mother's Education (2006)")+**

**+ geom\_smooth(method = lm, color = "red")**

**> white\_06 <- m\_educ\_2006 %>%**

**+ filter(race == "white")**

**> black\_06 <- m\_educ\_2006 %>%**

**+ filter(race == "black")**

**> other\_06 <- m\_educ\_2006 %>%**

**+ filter(race == "other")**

**> # Correlation by race**

**> # from this dataset give me column $ for a variable**

**> cor.test(white\_06$educ, white\_06$maeduc) # White**

Pearson's product-moment correlation

data: white\_06$educ and white\_06$maeduc

t = 18.463, df = 1919, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.3497301 0.4257077

sample estimates:

cor

0.3883788

**> cor.test(black\_06$educ, black\_06$maeduc**) # Black

Pearson's product-moment correlation

data: black\_06$educ and black\_06$maeduc

t = 6.9853, df = 349, p-value = 1.446e-11

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2548952 0.4388224

sample estimates:

cor

0.3502304

**> cor.test(other\_06$educ, other\_06$maeduc**) # Other

Pearson's product-moment correlation

data: other\_06$educ and other\_06$maeduc

t = 13.634, df = 349, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5167135 0.6538394

sample estimates:

cor

0.5895074

**> white\_model\_06 <- lm(educ ~ maeduc, data = white\_06)**

**> # model summary**

**> summary(white\_model\_06)**

Call:

lm(formula = educ ~ maeduc, data = white\_06)

Residuals:

Min 1Q Median 3Q Max

-13.2921 -2.0265 -0.0265 1.9735 8.1377

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.22953 0.21263 48.11 <2e-16 \*\*\*

maeduc 0.31641 0.01714 18.46 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.578 on 1919 degrees of freedom

Multiple R-squared: 0.1508, Adjusted R-squared: 0.1504

F-statistic: 340.9 on 1 and 1919 DF, p-value: < 2.2e-16

> # Black

**> black\_model\_06 <- lm(educ ~ maeduc, data = black\_06)**

**> # model summary**

**> summary(black\_model\_06)**

Call:

lm(formula = educ ~ maeduc, data = black\_06)

Residuals:

Min 1Q Median 3Q Max

-11.832 -1.281 0.097 1.617 8.373

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.97414 0.45654 21.847 < 2e-16 \*\*\*

maeduc 0.27555 0.03945 6.985 1.45e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.678 on 349 degrees of freedom

Multiple R-squared: 0.1227, Adjusted R-squared: 0.1201

F-statistic: 48.79 on 1 and 349 DF, p-value: 1.446e-11

**> # Other**

**> other\_model\_06 <- lm(educ ~ maeduc, data = other\_06)**

**> # model summary**

**> summary(other\_model\_06)**

Call:

lm(formula = educ ~ maeduc, data = other\_06)

Residuals:

Min 1Q Median 3Q Max

-11.9023 -2.0860 0.0977 2.0977 9.9752

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.08603 0.34136 23.69 <2e-16 \*\*\*

maeduc 0.48469 0.03555 13.63 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.591 on 349 degrees of freedom

Multiple R-squared: 0.3475, Adjusted R-squared: 0.3456

F-statistic: 185.9 on 1 and 349 DF, p-value: < 2.2e-16

**Repeat the analysis above using the GSS2016 dataset.**

**Correlation and Regression:**

Use the GSS06 dataset and answer the following questions about the relationship between respondent’s years of education ***(‘educ’)*** and mother’s years of education ***(‘maeduc’)****.*

**Scatterplot: Respondent’s years of Education *vs* mother’s years of Education.**

1. ***What is the Pearson’s r equal to?***0.3806361

**Interpretation:**>> r = 0 indicates no linear relationship between variables  
>> r=1 represents a perfect +ve relationship  
>> r=-1 represents a perfect -ve relationship

From the Pearson’s r value, we can say that the there’s a weak-moderate positive relationship between respondent’s years of education and mother’s years of education.

1. ***What is the y intercept equal to?***

10.35119

**Linear Regression Output:**

**> # regression line for 'educ' and 'maeduc'**

**> educ\_model\_16 <- lm(educ ~ maeduc, data = m\_educ\_2016)**

**> # you can get just the coefficients of your model**

**> coef(educ\_model\_16)**

**(Intercept) maeduc**

**10.3511865 0.3014422**

**> # you can also get a full summary**

**> summary(educ\_model\_16)**

**Call:**

lm(formula = educ ~ maeduc, data = m\_educ\_2016)

**Residuals**:

Min 1Q Median 3Q Max

-12.9685 -1.9685 -0.1598 2.0315 9.6488

**Coefficients**:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.35119 0.17927 57.74 <2e-16 \*\*\*

maeduc 0.30144 0.01443 20.89 <2e-16 \*\*\*

---

**Signif. codes:** 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Residual standard error:** 2.697 on 2575 degrees of freedom

**Multiple R-squared:** 0.1449, Adjusted R-squared: 0.1446

**F-statistic:** 436.3 on 1 and 2575 DF, p-value: < 2.2e-16

1. ***For each additional year of education for the mother, what is the expected increase in respondent’s years of education? Are the results significant?***

Ans: for each additional year of education for the mother, the expected increase in respondent’s years of education is slope (0.30144) times years plus Y intercept i.e. 10.35119

Y= 10.35119+ 0.30144x (a is intercept and b is slope)

**Results are very significant as indicated by \*\*\* => 0 i.e. confidence level 100%**

1. ***What is the predicted value for the respondent’s years of education when mother’s years of education are equal to 14?***

**Predicted:**

**Predicted ‘educ”**

**Predicted ‘educ” years.**

Y= 10.35119+ 0.30144\*14 = 14.57135 years

Now split the database by **‘race’** and use the results to answer the same questions as above (in table from below). Use the table also to answer questions 6 -10.

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear Regression Results for Education by mother’s Education by Race** | | | |
| **Statistic** | **White** | **Black** | **Other** |
| Pearson’s r | 0.3709768 | 0.3026761 | 0.4553624 |
| y-intercept | 10.28725 | 11.15590 | 9.85474 |
| Gain per mother’s years of education | 0.31748 | 0.19426 | 0.33132 |
| Significance level | 0 ‘\*\*\*’ | 0 ‘\*\*\*’ | 0 ‘\*\*\*’ |

1. ***For which of these three groups is the correlation strongest?*** *As indicated value of Pearson’s r* 0.3709768 (white) , 0.3026761 (black) and 0.4553624 (other) , since **OTHER** group has the highest value among the groups, the correlation is strongest for OTHER group.
2. ***Is the association positive or negative?***A positive value of Pearson’s r suggests that the association is positive for all the groups.
3. ***For which group does mother’s years of education produce the greatest gains in respondent’s years of education?***

Since the group BLACK has the highest Y-intercept, mother’s years of education produces the greatest gains in respondent’s years of education for **BLACK** group.

***For whites, what percent of the variance in ‘educ’ can be explained by ‘maeduc’?***  
Multiple R-squared: 0.1376, Adjusted R-squared: 0.1372  
Hence for whites, **13.76%** of variance in ‘educ’ can be explained by ‘maeduc’  
R-squared is simply square of Pearson’s r. It indicates how close the values are to the regression line.

% variance of the dependent variable is reflected in the correlation coefficient R-squared.

1. ***For blacks, what percent of the variance in ‘educ’ can be explained by ‘maeduc’?***

Multiple R-squared: 0.09161, Adjusted R-squared: 0.08947

Hence for whites, **9.16%** of variance in ‘educ’ can be explained by ‘maeduc’  
R-squared is simply square of Pearson’s r. It indicates how close the values are to the regression line.

% variance of the dependent variable is reflected in the correlation coefficient R-squared.