**> ##### R Lab 7: Correlation and Regression -- 2016 GSS Data ####**

**> install.packages("tidyverse")**

Error in install.packages : Updating loaded packages

Restarting R session...

**> install.packages("tidyverse")**

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/tidyverse\_1.2.1.tgz'

Content type 'application/x-gzip' length 77756 bytes (75 KB)

==================================================

downloaded 75 KB

The downloaded binary packages are in

/var/folders/79/tx9tjz8j0hl99904bkw5dy740000gn/T//Rtmp5JrvrI/downloaded\_packages

> library(tidyverse)

── Attaching packages ─────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

✔ ggplot2 3.0.0 ✔ purrr 0.2.5

✔ tibble 1.4.2 ✔ dplyr 0.7.6

✔ tidyr 0.8.1 ✔ stringr 1.3.1

✔ readr 1.1.1 ✔ forcats 0.3.0

── Conflicts ────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

Warning messages:

1: package ‘tidyverse’ was built under R version 3.4.2

2: package ‘ggplot2’ was built under R version 3.4.4

3: package ‘tibble’ was built under R version 3.4.3

4: package ‘tidyr’ was built under R version 3.4.4

5: package ‘purrr’ was built under R version 3.4.4

6: package ‘dplyr’ was built under R version 3.4.4

7: package ‘stringr’ was built under R version 3.4.4

8: package ‘forcats’ was built under R version 3.4.3

> # Importing GSS 2016 Data

> gss2016 <- read\_csv("GSS\_2016.csv")

Parsed with column specification:

cols(

.default = col\_character(),

agekdbrn = col\_integer(),

babies = col\_integer(),

bigbang = col\_logical(),

bigbang1 = col\_logical(),

bigbang2 = col\_logical(),

boyorgrl = col\_logical(),

cohort = col\_integer(),

cohrs2 = col\_integer(),

condrift = col\_logical(),

coninc = col\_double(),

conrinc = col\_double(),

copres10 = col\_integer(),

copres105plus = col\_integer(),

cosei10 = col\_double(),

cosei10educ = col\_double(),

cosei10inc = col\_double(),

dateintv = col\_integer(),

earnrs = col\_integer(),

educ = col\_integer(),

electron = col\_logical()

# ... with 81 more columns

)

See spec(...) for full column specifications.

|=====================================================================================| 100% 17 MB

Warning: 2 parsing failures.

row # A tibble: 2 x 5 col row col expected actual file expected <int> <chr> <chr> <chr> <chr> actual 1 1210 old1 no trailing characters " or older" 'GSS\_2016.csv' file 2 2760 earnrs an integer eight or more 'GSS\_2016.csv'

**> m\_educ\_2016 <- gss2016 %>%**

**+ 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\_2016, 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 (2016)")**

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

**> cor.test(m\_educ\_2016$educ, m\_educ\_2016$maeduc)**

Pearson's product-moment correlation

data: m\_educ\_2016$educ and m\_educ\_2016$maeduc

t = 20.887, df = 2575, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.3471254 0.4131760

sample estimates:

cor

0.3806361

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

**> # 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

**> ggplot(m\_educ\_2016, 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 (2016)")+**

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

**> white\_16 <- m\_educ\_2016 %>%**

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

**> black\_16 <- m\_educ\_2016 %>%**

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

**> other\_16 <- m\_educ\_2016 %>%**

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

**> # Correlation by race**

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

**> cor.test(white\_16$educ, white\_16$maeduc) # White**

Pearson's product-moment correlation

data: white\_16$educ and white\_16$maeduc

t = 17.427, df = 1903, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.3315906 0.4090720

sample estimates:

cor

0.3709768

**> cor.test(black\_16$educ, black\_16$maeduc) # Black**

Pearson's product-moment correlation

data: black\_16$educ and black\_16$maeduc

t = 6.5392, df = 424, p-value = 1.782e-10

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2138155 0.3865688

sample estimates:

cor

0.3026761

**> cor.test(other\_16$educ, other\_16$maeduc) # Other**

Pearson's product-moment correlation

data: other\_16$educ and other\_16$maeduc

t = 7.9894, df = 244, p-value = 5.377e-14

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.3502363 0.5491591

sample estimates:

cor

0.4553624

**> # White**

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

**> # model summary**

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

Call:

lm(formula = educ ~ maeduc, data = white\_16)

Residuals:

Min 1Q Median 3Q Max

-13.097 -2.097 -0.097 1.903 7.713

Coefficients:

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

(Intercept) 10.28725 0.22885 44.95 <2e-16 \*\*\*

maeduc 0.31748 0.01822 17.43 <2e-16 \*\*\*

---

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

Residual standard error: 2.671 on 1903 degrees of freedom

Multiple R-squared: 0.1376, Adjusted R-squared: 0.1372

F-statistic: 303.7 on 1 and 1903 DF, p-value: < 2.2e-16

**> # Black**

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

**> # model summary**

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

Call:

lm(formula = educ ~ maeduc, data = black\_16)

Residuals:

Min 1Q Median 3Q Max

-7.4871 -1.4871 -0.4871 1.6371 7.8728

Coefficients:

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

(Intercept) 11.15590 0.37082 30.084 < 2e-16 \*\*\*

maeduc 0.19426 0.02971 6.539 1.78e-10 \*\*\*

---

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

Residual standard error: 2.362 on 424 degrees of freedom

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

F-statistic: 42.76 on 1 and 424 DF, p-value: 1.782e-10

**> # Other**

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

**> # model summary**

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

Call:

lm(formula = educ ~ maeduc, data = other\_16)

Residuals:

Min 1Q Median 3Q Max

-10.8367 -1.8306 0.1663 2.1694 10.1453

Coefficients:

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

(Intercept) 9.85474 0.46296 21.286 < 2e-16 \*\*\*

maeduc 0.33132 0.04147 7.989 5.38e-14 \*\*\*

---

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

Residual standard error: 3.249 on 244 degrees of freedom

Multiple R-squared: 0.2074, Adjusted R-squared: 0.2041

F-statistic: 63.83 on 1 and 244 DF, p-value: 5.377e-14

**> ggplot(white\_16, aes(x = maeduc, y = educ)) +**

**+ geom\_jitter(size=1.5, color = "royal blue") +**

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

**+ title = "White - Education and Mother's Education (2016)") +**

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

**> ggplot(black\_16, aes(x = maeduc, y = educ)) +**

**+ geom\_jitter(size=1.5, color = "light coral") +**

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

**+ title = "Black - Education and Mother's Education (2016)") +**

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

**> ggplot(other\_16, aes(x = maeduc, y = educ)) +**

**+ geom\_jitter(size=1.5, color = "green3") +**

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

**+ title = "Other - Education and Mother's Education (2016)") +**

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