> ##### R Lab 7: Correlation and Regression -- 2006 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//RtmpdD6J2d/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 2006 Data

> gss2006 <- read\_csv("GSS\_2006.csv")

Parsed with column specification:

cols(

.default = col\_character(),

prestg10 = col\_integer(),

prestg105plus = col\_integer(),

sppres10 = col\_integer(),

sppres105plus = col\_integer(),

papres10 = col\_integer(),

papres105plus = col\_integer(),

mapres10 = col\_integer(),

mapres105plus = col\_integer(),

sei10 = col\_double(),

spsei10 = col\_double(),

pasei10 = col\_double(),

masei10 = col\_double(),

sei10educ = col\_double(),

spsei10educ = col\_double(),

pasei10educ = col\_double(),

masei10educ = col\_double(),

sei10inc = col\_double(),

spsei10inc = col\_double(),

pasei10inc = col\_double(),

masei10inc = col\_double()

# ... with 110 more columns

)

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

Warning: 7 parsing failures.

row # A tibble: 5 x 5 col row col expected actual file expected <int> <chr> <chr> <chr> <chr> actual 1 1722 physhlth an integer DONT KNOW 'GSS\_2006.csv' file 2 2958 prozfor1 an integer DONT KNOW 'GSS\_2006.csv' row 3 3164 adults no trailing characters " or more" 'GSS\_2006.csv' col 4 3170 sphrs2 no trailing characters + hrs 'GSS\_2006.csv' expected 5 3246 physhlth an integer DONT KNOW 'GSS\_2006.csv'

... ................................. ... ................................................................. ........ ....................................... [... truncated]

Warning message:

In rbind(names(probs), probs\_f) :

number of columns of result is not a multiple of vector length (arg 1)

**> 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

> # Scatter plots by race

>

> # White

**> ggplot(white\_06, 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 (2006)") +**

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

**> # Black**

**> ggplot(black\_06, 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 (2006)") +**

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

**> # Other**

**> ggplot(other\_06, 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 (2006)") +**

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