Applied Linear Modeling

September 10, 2019

To-do

- Exercises:
  - *pick a to-do number*
  - *write your results on the board*
  - *when finished, put your name on the to-do number and drop in Done jar*
- Create a week-2 folder on your laptop and put the following files in it (from GitHub):
  - *week-3-workshop.Rmd*
  - *water_educ_2015_who_unesco_ch8.csv*
- No new R packages for today

All the things for today

- Discussion of exercises
- Workshop
  - *Pearson's r*
  - *Spearman's rho*

What is the relationship between water access and female education?

- Access to clean water and sanitation is limited in some parts of the world, especially in countries with high rates of poverty
  - *In countries where clean water and sanitation are limited, women and girls are often responsible for the labor involved in accessing clean water for their families*
  - *This is one factor that limits educational opportunities of girls*
- The world health organization publishes data on water and sanitation access
- UNESCO publishes data on the percent of the population, males, and females who complete primary and secondary school

Importing and reviewing the data

- The data set includes water, sanitation, and education variables for 97 countries
- These data were collected in 2015 by the world health organization and UNESCO and are freely available on their websites

```
# import the water data
water.educ <- read.csv(file = "water_educ_2015_who_unesco_ch8.csv")

# examine the data
summary(object = water.educ)
```

```
##                country      med.age       perc.1dollar
##  Albania            : 1   Min.   :15.00   Min.   : 1.00
##  Antigua and Barbuda: 1   1st Qu.:22.50   1st Qu.: 1.00
##  Argentina          : 1   Median :29.70   Median : 1.65
##  Australia          : 1   Mean   :30.33   Mean   :13.63
##  Azerbaijan         : 1   3rd Qu.:39.00   3rd Qu.:17.12
##  Bahrain            : 1   Max.   :45.90   Max.   :83.80
##  (Other)            :91                   NA's   :33
##  perc.basic2015sani perc.safe2015sani perc.basic2015water
##  Min.   :  7.00     Min.   :  9.00    Min.   : 19.00
##  1st Qu.: 73.00     1st Qu.: 61.25    1st Qu.: 88.75
##  Median : 93.00     Median : 76.50    Median : 97.00
##  Mean   : 79.73     Mean   : 71.50    Mean   : 90.16
##  3rd Qu.: 99.00     3rd Qu.: 93.00    3rd Qu.:100.00
##  Max.   :100.00     Max.   :100.00    Max.   :100.00
##                     NA's   :47        NA's   :1
##  perc.safe2015water perc.in.school  female.in.school male.in.school
##  Min.   : 11.00     Min.   :33.32   Min.   :27.86    Min.   :38.66
##  1st Qu.: 73.75     1st Qu.:83.24   1st Qu.:83.70    1st Qu.:82.68
##  Median : 94.00     Median :92.02   Median :92.72    Median :91.50
##  Mean   : 83.38     Mean   :87.02   Mean   :87.06    Mean   :87.00
##  3rd Qu.: 98.00     3rd Qu.:95.81   3rd Qu.:96.61    3rd Qu.:95.57
##  Max.   :100.00     Max.   :99.44   Max.   :99.65    Max.   :99.36
##  NA's   :45
```


Codebook

- country: the name of the country
- med.age: the median age of the citizens in the country
- perc.1dollar: percentage of citizens living on $1 per day or less
- perc.basic2015sani: percentage of citizens with basic sanitation access
- perc.safe2015sani: percentage of citizens with safe sanitation access
- perc.basic2015water: percentage of citizens with basic water access
- perc.safe2015water: percentage of citizens with safe water access
- perc.in.school: percentage of school-age people in primary and secondary school
- female.in.school: percentage of female school-age people in primary and secondary school
- male.in.school: percentage of male school-age people in primary and secondary school

9/10/2019 Applied Linear Modeling (1)

file:///C:/Users/harrisj/Box/teaching/Teaching/Fall2019/week-3-materials/week-3-slides-packet.html#(1) 6/46

9/10/2019 Applied Linear Modeling (1)

file:///C:/Users/harrisj/Box/teaching/Teaching/Fall2019/week-3-materials/week-3-slides-packet.html#(1) 8/46

Descriptive statistics

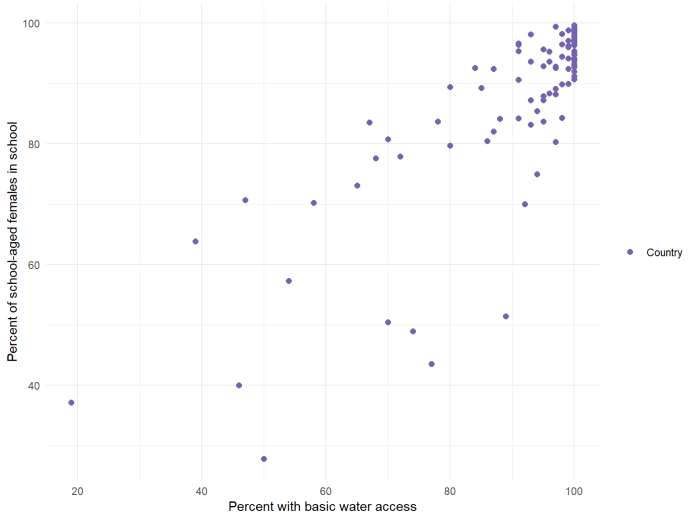
```
# open the tidyverse
library(package = "tidyverse")

# descriptive statistics for females in school and water access
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  summarize(m.f.educ = mean(x = female.in.school),
            sd.f.educ = sd(x = female.in.school),
            m.bas.water = mean(x = perc.basic2015water),
            sd.bas.water = sd(x = perc.basic2015water))
```

```
##   m.f.educ sd.f.educ m.bas.water sd.bas.water
## 1 87.01123   15.1695    90.15625     15.81693
```


Visualize the relationship

```
# explore plot of female education and water access
water.educ %>%
  ggplot(aes(y = female.in.school, x = perc.basic2015water)) +
  geom_point(aes(color = "Country"), size = 2) +
  theme_minimal() +
  labs(y = "Percent of school-aged females in school",
       x = "Percent with basic water access") +
  scale_color_manual(values = "#7463AC", name = "")
```



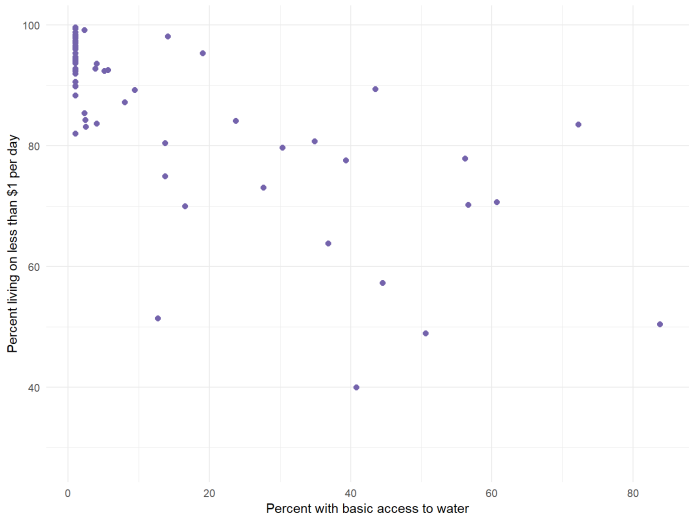
Correlation

- A correlation is a relationship between two variables
- The pattern in the graph where percent of females in school goes up when water access goes up could indicate that there is a relationship, or **correlation** between water access and percentage of females in school
- When one variable goes up as the other one goes up, the relationship could be a **positive correlation**

As the percentage of people with basic water access goes up, the percentage of females with primary or secondary education also goes up.

Poverty and water access

```
# plot of poverty and water access
water.educ %>%
  ggplot(aes(y = female.in.school, x = perc.1dollar)) +
  geom_point(size = 2, color = "#7463AC") +
  theme_minimal() +
  labs(x = "Percent with basic access to water",
       y = "Percent living on less than $1 per day")+
  scale_color_manual(values = "#7463AC", name = "")
```



Negative correlation

- When one variable goes up and the other simultaneously goes down, this is a negative correlation

As the percentage of people with basic water access goes up, the percentage of people living on less than $1 per day goes down.

Measuring correlation

- The extent to which two variables correlate is measured by a **correlation coefficient**
- The correlation coefficient has two features, strength and direction
  - *A correlation can be weak, moderate, or strong*
  - *A correlation can be positive or negative*

Computing the Pearson's r correlation between two variables

The equation for the r correlation coefficient is:

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{n-1}}{s_x s_y}$$

Where:

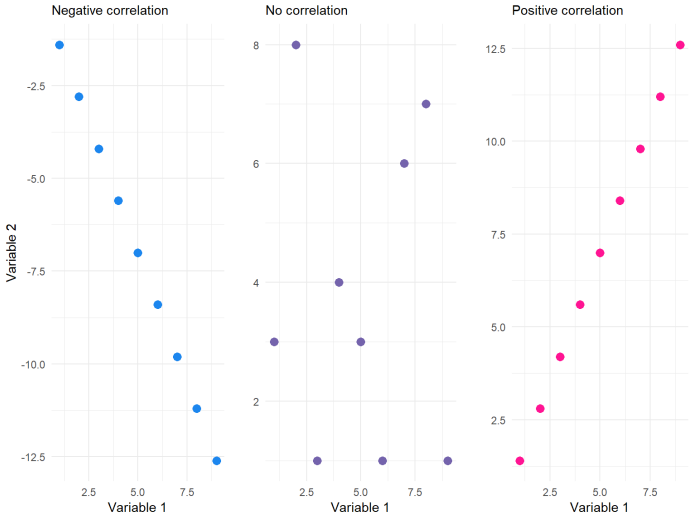
- m_x is the mean of *x* and m_y is the mean of *y*
- x_i is each observation of *x* and y_i is each observation of *y*
- *n* is the sample size
- s_x is the standard deviation of *x* and s_y is the standard deviation of *y*

Simplified:

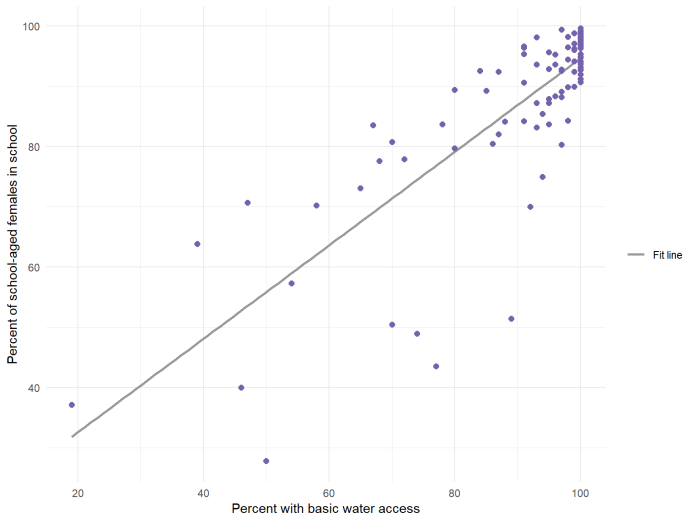
$$r_{xy} = \frac{cov_{xy}}{s_x s_y}$$

Types of correlation coefficients

- *Negative correlations* are when one variable goes up, the other goes down
- *No correlation* is when there is no discernable pattern in how two variables vary
- *Positive correlations* are when one variable goes up, the other also goes up (or when one goes down the other does too); both variables move together in the same direction



Water access and female education in countries globally



Calculating Pearson's r correlation coefficient

```
# correlation between water access and female education
water.educ %>%
  summarize(cor.females.water = cor(x = perc.basic2015water,
                                    y = female.in.school,
                                    use = "complete"))
```

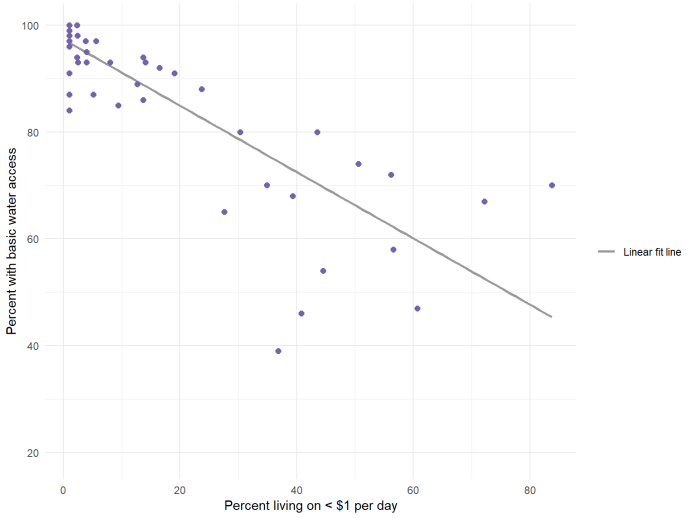
```
##   cor.females.water
## 1         0.8086651
```


Interpreting the strength of the correlation coefficient

- r = -1.0 is perfectly negative
- r = -.8 is strongly negative
- r = -.5 is moderately negative
- r = -.2 is weakly negative
- r = 0 is no relationship
- r = .2 is weakly positive
- r = .5 is moderately positive
- r = .8 is strongly positive
- r = 1.0 is perfectly positive

There is a strong positive correlation between basic water access and the percent of females in school (r = .81).

Water access and poverty in countries globally



Add correlation coefficient for poverty and water access

```
# correlations between water access, poverty, and female education
water.educ %>%
  summarize(cor.females.water = cor(x = perc.basic2015water,
                                    y = female.in.school,
                                    use = "complete"),
            cor.poverty.water = cor(x = perc.1dollar,
                                    y = perc.basic2015water,
                                    use = "complete"))
```

```
##   cor.females.water cor.poverty.water
## 1         0.8086651        -0.8320895
```

There is a moderate to strong negative correlation between percent living in poverty and percent with water access.

You try it! Graph a relationship

Copy and modify the code from above to graph the relationship between the `female.in.school` variable and basic sanitation measured by the `perc.basic2015sani` variable.

```
# relationship between female in school and sanitation
```


You try it! Compute a correlation

Add the correlation between `female.in.school` and `perc.basic2015sani` to the existing code:

```
# correlations between water access, poverty, and female education
water.educ %>%
  summarize(cor.females.water = cor(x = perc.basic2015water,
                                    y = female.in.school,
                                    use = "complete"),
            cor.poverty.water = cor(x = perc.1dollar,
                                    y = perc.basic2015water,
                                    use = "complete"))
```

```
##   cor.females.water cor.poverty.water
## 1         0.8086651        -0.8320895
```

You try it! Interpret your results

Conducting an inferential statistical test for Pearson's *r* correlation coefficient

- The correlation coefficients and plots indicated that, for this sample of countries, percentage of females in school was positively correlated with basic water access and and negatively correlated with poverty
- Does this relationship hold for all countries?
- There is an inferential statistical test that can be used to determine if the correlation coefficients in the sample likely represent similar relationships in the population

NHST Step 1: Writing the null and alternate hypotheses

H0: There is no relationship between the percent of people with water access and percent of females in school ($r = 0$)

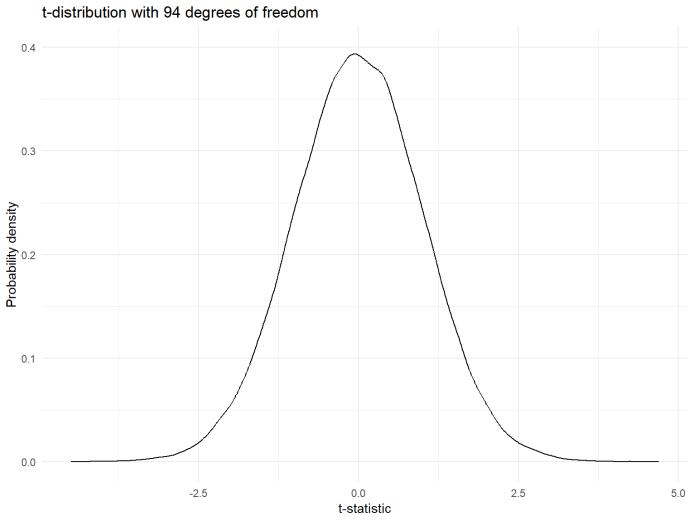
HA: There is a relationship between the percent of people with water access and percent of females in school ($r \neq 0$)

NHST Step 2: Computing the test statistic

```
# test for correlation coefficient
cor.test(x = water.educ$perc.basic2015water,
         y = water.educ$female.in.school)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water.educ$perc.basic2015water and water.educ$female.in.school
## t = 13.328, df = 94, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7258599 0.8683663
## sample estimates:
##       cor
## 0.8086651
```

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)



NHST Steps 4 & 5: Reject or retain the null hypothesis based on the p-value

- The p-value was very tiny, well under .05
- This p-value is the probability that the very strong positive relationship (r = .81) observed between percentage of females in school and percentage with basic water access would have happened if the null were true
- It is unlikely that this correlation would happen in the sample if there were no correlation in the population that this sample came from
- The output also contained confidence interval around r, so the value of r in the sample is .81 and the likely value of r in the population that this sample came from is somewhere between .73 and .87

Interpretation

The percentage of people with basic access to water is statistically significantly, positively, and very strongly correlated with the percentage of primary and secondary age females in school in a sample of countries ($r = .81$; $t(94) = 13.33$; $p < .05$). As the percentage of people living with basic access to water goes up, the percentage of females with education also goes up. While the correlation is .81 in the sample, it is likely between .73 and .87 in the population (95% CI: .73 - .87).

Checking assumptions for Pearson's r correlation analyses

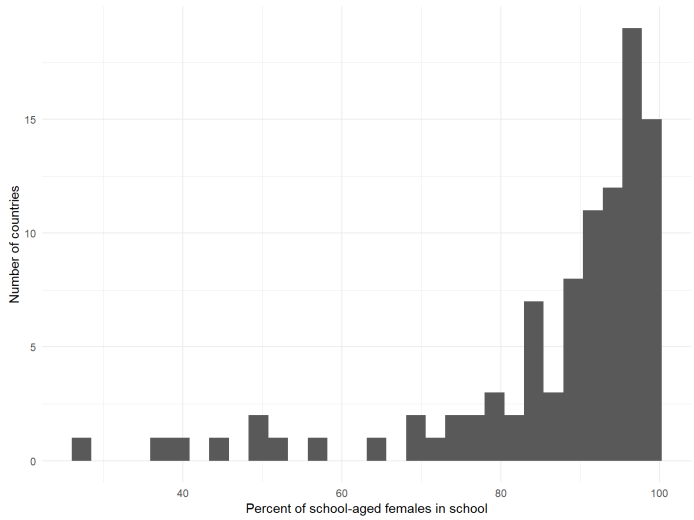
Correlation coefficients rely on four assumptions:

- Both variables are continuous
- Both variables are normally distributed
- The relationship between the two variables is *linear* (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)

Checking the normality assumption with a histogram

```
# check normality of female.in.school variable
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(x = female.in.school)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "Percent of school-aged females in school",
       y = "Number of countries")
```

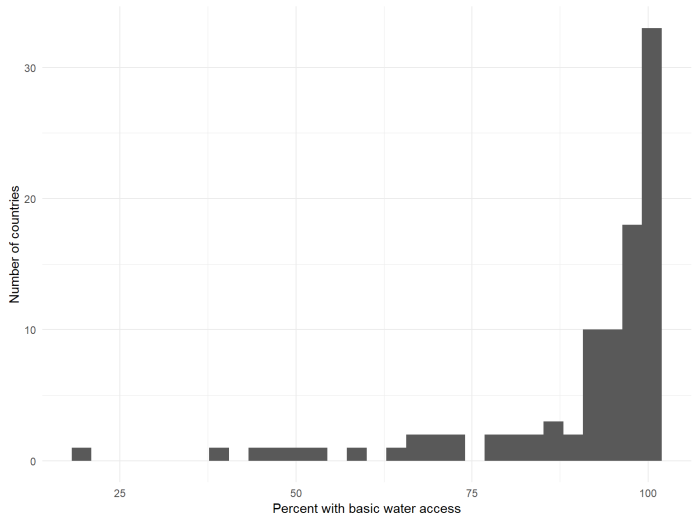

Checking normality using a histogram



Checking normality for the water access variable

```
# check normality of water access variable
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(x = perc.basic2015water)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "Percent with basic water access",
       y = "Number of countries")
```


Checking normality for the water access variable

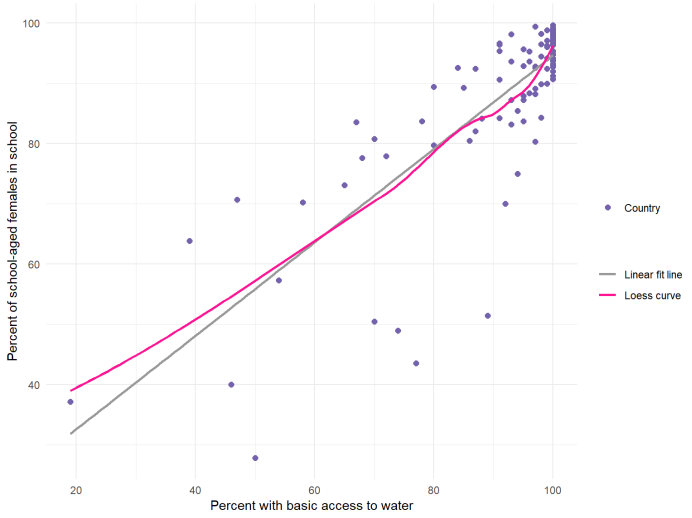


Checking the linearity assumption

- Add a Loess curve to a scatterplot to check

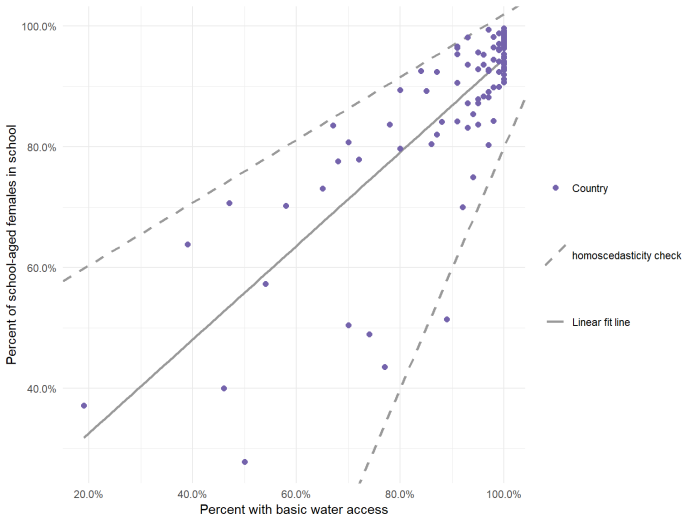
```
# female education and water graph with linear fit line and Loess curve
water.educ %>%
  ggplot(aes(y = female.in.school, x = perc.basic2015water)) +
  geom_point(aes(size = "Country"), color = "#7463AC") +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE) +
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "Percent of school-aged females in school",
       x = "Percent with basic access to water") +
  scale_color_manual(values = c("gray60", "deeppink"), name= "") +
  scale_size_manual(values = 2, name = "")
```


Checking the linearity assumption



Checking the homoscedasticity assumption

- homoscedasticity is the equal distribution of points around the line



Interpreting the assumption checking results

- Continuous variables: Met
- Normality: Not met
- Linearity: Met
- Homoscedasticity: Not met

What happens when the assumptions are not met?

- The correlation can be reported, but not generalized from the sample to the population
- Options to try when assumptions are not met:
  - *Transform the variables and try Pearson's r again*
  - *Use Spearman's rho if its assumptions are met:*
    - The variables must be at least ordinal or even closer to continuous: **Met**
    - The relationship between the two variables must be *monotonic*: **Met**

Spearman's rho

H0: There is no correlation between the percentage of females in school and the percentage of citizens with basic water access ($\rho = 0$)

HA: There is a correlation between the percentage of females in school and the percentage of citizens with basic water access ($\rho \neq 0$)

```
# spearman correlation female education and water access
spear.fem.water <- cor.test(x = water.educ$perc.basic2015water,
                            y = water.educ$female.in.school,
                            method = "spearman")
spear.fem.water
```

```
##
##  Spearman's rank correlation rho
##
## data:  water.educ$perc.basic2015water and water.educ$female.in.school
## S = 34050, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7690601
```


Interpreting rho

There was a statistically significant positive correlation between basic access to drinking water and female education ($r_s = 0.77$; $p < .001$). As the percentage of the population with basic access to water increases, so does the percentage of females in school. The data meet the monotonic relationship and variable type assumptions for Spearman's rho.

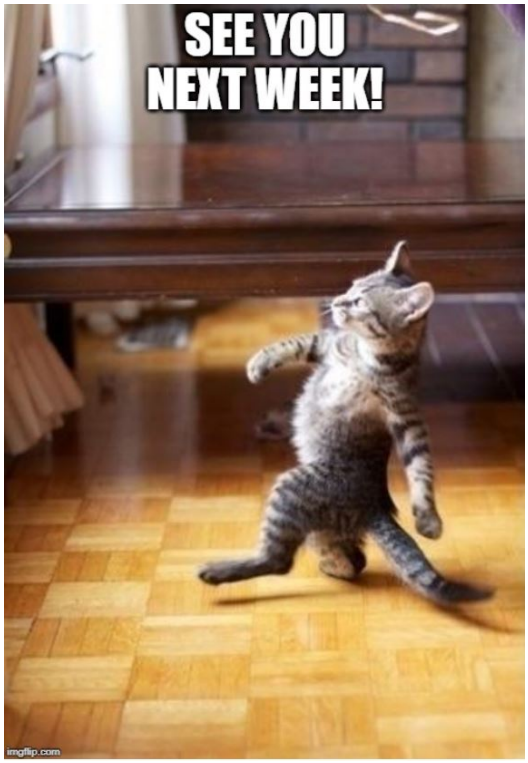
You try it! Test the correlation of sanitation with females in school

Step 1: Write the null and alternate hypotheses:

H0: There is no correlation between access to basic sanitation and percent of females in school.

HA: There is a correlation between access to basic sanitation and percent of females in school.

The End



9/10/2019 Applied Linear Modeling (1)