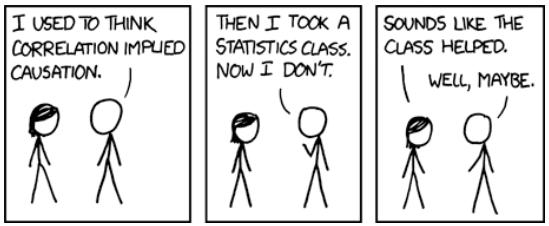
Biostats Week 8: Bivariate for two continuous variables

Week 8 schedule

- t-test review
- correlation workshop
- sign up for book club topic



Correlation workshop outline

The remaining two-variable situation is when there are two continuous variables. Here is the outline for learning about *correlation* analyses:

1. Using graphs to make a prediction
2. Computing and interpreting Pearson's r correlation coefficient
3. Inference and correlation coefficients
4. Assumption checking for correlation analyses
5. Spearman's r correlation coefficient
6. Partial correlations

The clean water problem

- The lack of access to clean water and sanitation worldwide impacts people living in poverty and poor women and girls in particular
- Specifically, women and girls tend to be responsible for collecting water for their families, often walking long distances in unsafe areas and carrying heavy loads
- Lack of access to sanitation facilities also puts women and girls at greater risk for harassment and assault and keeps teenage girls out of school in many parts of the world

The data

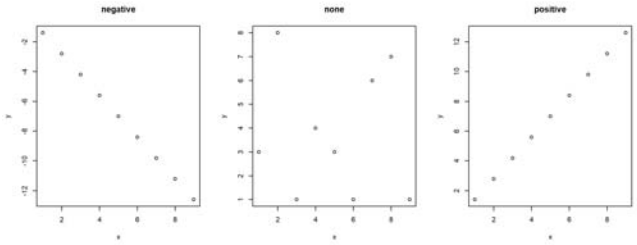
The World Health Organization data tracks water access and sanitation globally. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) tracks education rates by sex globally. These two data sources are merged to create a data set that is saved here: https://tinyurl.com/y7k5uqq9

The observations are countries and the variables are:

- country: the name of the country
- med.age: the median age of the population in the country
- perc.1dollar: percentage of the population living on $1 per day or less
- perc.basic2015sani: percentage of the population with basic sanitation access
- perc.safe2015sani: percentage of the population with safe sanitation access
- perc.basic2015water: percentage of the population with basic water access
- perc.safe2015water: percentage of the population 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

Types of correlations

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



Correlation strength

Correlations can range from -1 to 1.

The size of the correlation (regardless of whether it is negative or positive) is its strength. While there is no firm cutoff, most resources will characterize correlation coefficients as:

- very weak 0 to .2
- weak .2 to .39
- moderate .4 to .59
- strong .6 to .79
- very strong .8 to 1

Pearson's r correlation coefficient calculation

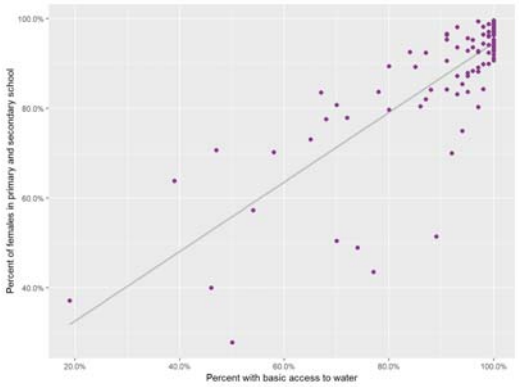
The most commonly used correlation coefficient is Pearson's r:

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

Where:

- i is an individual observation
- x_i is the value of x for each observation
- $\bar{x}$ is the mean of x
- y_i is the value of y for each observation
- $\bar{y}$ is the mean of y
- n is the sample size
- s_x and s_y are the standard deviations of x and y

Basic water access and percent of females in school



Pearson's r for female education and water

A correlation coefficient can quantify the relationship:

```
# compute the correlation coefficient
cor(waterData$female.in.school, waterData$perc.basic2015water, use='complete')
```

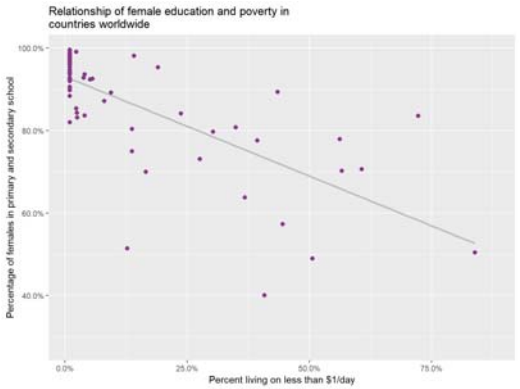
```
## [1] 0.8086651
```

The correlation of 0.81 is positive and very strong.

The interpretation:

The percent of females in school is very strongly positively correlated with the percent of citizens with basic access to drinking water (r = 0.81). As basic access to water goes up, the percent of school-age females in school also increases.

What is the correlation here?



Computing the correlation coefficient

```
# compute the correlation coefficient
cor(waterData$female.in.school, waterData$perc.1dollar, use='complete')
```

```
## [1] -0.7144238
```

Interpretation:

The graph and correlation coefficient show a strong negative relationship between poverty and females in school (r = -0.71). That is, as poverty goes up, female education goes down.

Check your understanding

Interpret the correlation between female education and safe water access:

```
# compute the correlation coefficient
cor(waterData$female.in.school, waterData$perc.safe2015water, use = 'complete')
```

```
## [1] 0.7606593
```


Inference and correlation coefficients

The correlation coefficients and plots indicated that, for this sample of countries:

- female education was positively correlated with basic water access and safe water access
- female education was negatively correlated with poverty

These are the relationships in **some countries** (our sample), but do these relationships hold in **all countries** (the population)?

A statistical test can be used to determine if the sample comes from a population where there is a relationship between the two variables of interest.

The null and alternate hypotheses

For basic water access and percent of females in school:

- H0: There is no correlation between basic water access and female education (r = 0)
- HA: There is a correlation between basic water access and female education (r does not equal 0)

For poverty and percent of females in school:

- H0: There is no correlation between poverty and female education (r = 0)
- HA: There is a correlation between poverty and female education (r does not equal 0)

The statistical test

The null hypothesis can be tested by using *a variation on the one-sample t-test*, compare the correlation coefficient (instead of the mean) to a hypothesized value of zero. The formula for a one-sample t-test is:

$$t = \frac{\bar{x} - \mu}{se_{\bar{x}}}$$

The standard error for a correlation coefficient is computed:

$$se_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Substituting r and the se_r into the one-sample t-statistic formula and simplifying, we get:

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

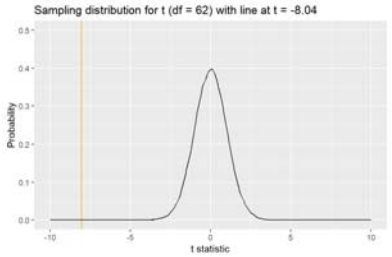
Using the formula

Use of this formula requires r and n

- r = -0.7144238
- n = 64

$$t = \frac{-0.7144238\sqrt{64-2}}{\sqrt{1-(-0.7144238)^2}} = -8.0395$$

t-statistic and sampling distribution



- The area under the curve of the sampling distribution contains all possible values of the t-statistic for samples with 62 d.f. that *came from a population where r = 0*
- The probability of getting a t-statistic of -8.04 (or larger) for a sample from *a population where r = 0* is the shaded area under the curve to the left of the orange line
- This is a very tiny probability of getting a sample where the t-statistic is this big (or bigger) *if r = 0 in the population*
- So, *r is probably not 0 in the population that this sample came from*

What is the actual probability or p-value?

Output from the `cor.test` function:

```
# test for correlation coefficient
corFemPov <- cor.test(waterData$female.in.school, waterData$perc.1dollar)
corFemPov
```

```
##
##  Pearson's product-moment correlation
##
## data:  waterData$female.in.school and waterData$perc.1dollar
## t = -8.0395, df = 62, p-value = 3.379e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8167942 -0.5684392
## sample estimates:
##       cor
## -0.7144238
```

- The correlation is r = -0.71
- The t-statistic is t = -8.04
- The degrees of freedom are d.f. = 62
- The p-value is 0.00000000003379057 or p < .001
- The 95% confidence interval for r is 95% CI: -0.57 to -0.82

What about the null hypothesis?

With a p-value this small, it is unlikely that the sample came from a population where r = 0, so…

- REJECT THE NULL HYPOTHESIS, in favor of the alternate hypothesis that there is a correlation between poverty and female education (r is not equal to 0).
- The correlation is *statistically significant*, that is, it is statistically unlikely that it came from a population where r = 0.
- Interpret the findings:

The percent of people living on $1 per day or less is statistically significantly strongly negatively correlated with the percent of primary and secondary age females in school in a country (r = -0.71; t(62) = -8.04; p < .001). The sample likely came from a population where the correlation between poverty and female education was between -0.57 and -0.82 (95% CI: -0.57 to -0.82). As the percent of people living on $1 per day or less goes up, the percent of school-age females in school goes down.

Check your understanding

Write the null and alternate hypotheses to test whether there is a correlation between female education and basic water access. Examine the output from a correlation test and interpret the results.

- H0:
- HA:

```
##
##  Pearson's product-moment correlation
##
## data:  waterData$female.in.school and waterData$perc.basic2015water
## 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
```

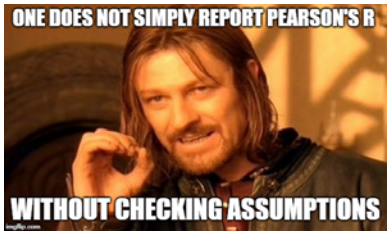
Reject or fail to reject the null hypothesis?

Write the interpretation:

Assumption checking for Pearson's r

There are several assumptions for Pearson's r:

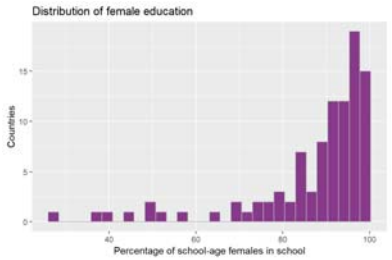
- Both variables are continuous
- Both variables are normally distributed
- The relationship between the two variables is *linear* (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)



Check assumption 1: Both variables are continuous

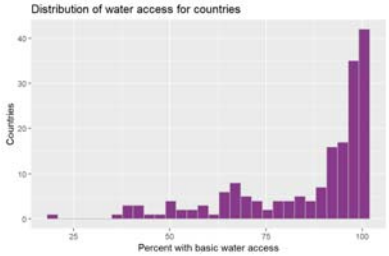
Assumption 1: For the correlation between females in school and basic water access, both variables are continuous. The first assumption is MET.

Check assumption 2: Both variables are normally distributed



- The data do not appear to be normally distributed.
- The distribution is very *left* or *negatively* skewed, where there are values that create a longer tail to the left of the histogram.
- The second assumption is NOT MET for this variable.

Check assumption 2: Both variables are normally distributed

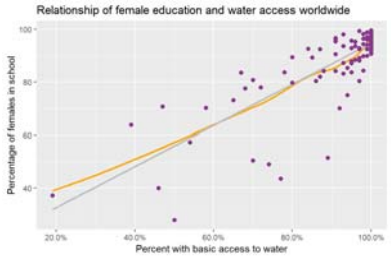


- The histogram shows a distribution that is extremely *left skewed*.
- The assumption is NOT MET for this variable.

Check assumption 3: Linearity

The linearity assumption is met if a scatterplot of the two variables shows a relationship *that falls along a line*.

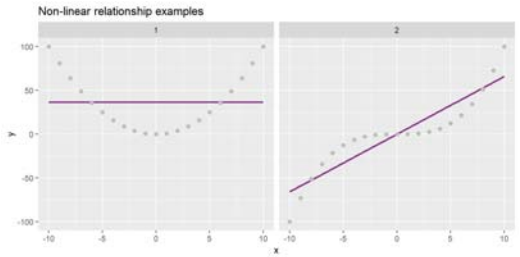
The straight line in the plot appears to represent the relationship well, so this assumption may be met. If it is difficult to tell, a *Loess curve* can be added to double-check. A Loess curve shows the actual relationship between the two variables without constraining the line to be straight (see orange line):



In this case, an orange Loess curve shows some minor deviation from linear at the lower percentages, but overall the relationship seems close to linear. This assumption is MET.

Wait, what does a non-linear relationship look like?

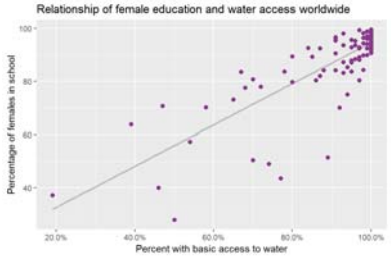
A non-linear relationship might look like one of these graphs where the relationships that do not fall along a straight line:



Both of these plots show relationships between x and y, but the relationships are not linear. They fall along curves instead of along a straight line.

Assumption 4: Homoscedasticity

The final assumption is the equal distribution of points around the line, which is often called the assumption of *homoscedasticity*. In the plot below, the points seem closer to the line on the far right and then are a little more spread out around the line on the left.



Checking homoscedasticity statistically

Although the difference in spread from the left to the right is not dramatic until the values are over 90% access to water, it might be worth using a statistical test to check whether the difference in spread from one end to the other is statistically significant.

A Breusch-Pagan test examines the **null hypothesis that the variance is constant**.

```
##
##  studentized Breusch-Pagan test
##
## data:  waterData$female.in.school ~ waterData$perc.basic2015water
## BP = 12.368, df = 1, p-value = 0.0004368
```

- The Breusch-Pagan test statistic has a low p-value associated with it (BP = 12.37; p = 0.0004), indicating that the null hypothesis would be rejected.
- The null hypothesis was that the variance is constant, so this means that the assumption of constant variance is not met.
- This assumption is NOT MET.

Assumption checking conclusion

- In all, the Pearson's r analysis for female education and water access met two of the four assumptions.
- It failed the assumptions of normally distributed variables and homoscedasticity.
- **While the results of the correlation analysis can be reported, since it failed two assumptions, the results should not be generalized beyond the sample.**
- Rewrite your conclusion:

In the sample, the percent of people with basic water access is very strongly positively correlated with the percent of primary and secondary age females in school in a country (r = 0.81). As the percent of people living on $1 per day or less goes up, the percent of school-age females in school goes down.

What now? Spearman's rho correlation coefficient

One option for when assumptions are unmet for a correlation analysis is the Spearman's rank correlation coefficient, rho.

- Spearman's rho is computed by ranking each value for each variable from lowest to highest and then computing the extent to which the two variable ranks are the same.
- So, for a Spearman's rho of female education and water access, the values of female education would be ranked from lowest to highest and the values of water access would be ranked from lowest to highest.
- Then, once the ranks are assigned, the correlation coefficient is computed:

$$rho = \frac{6\sum d^2}{n(n^2-1)}$$

Where:

- d is the difference between the ranks of the two variables
- n is the number of observations

Computing Spearman's rho

H0: There is no correlation between female education and water access (rho = 0) HA: There is a correlation between female education and water access (rho is not 0)

```
# spearman correlation female education and water access
spearCorrFemWater <- cor.test(waterData$female.in.school,
                              waterData$perc.basic2015water,
                              method = "spearman")
spearCorrFemWater
```

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

Statistical significance of rho

Instead of a t-statistic and p-value, the Spearman test reports the S test statistic, which is computed:

$$S = (n^3 - n)\frac{1 - r_{ranked}}{6}$$

Where r_{ranked} is the Pearson's r for ranks of the values of the two variables. Weirdly, the S is not used to find the p-value (or really for anything), instead the p-value is determined by computing a t-statistic using a Pearson's r for the ranked values and n-2 degrees of freedom:

$$t = r_{ranked}\sqrt{\frac{n-2}{1-r^2}}$$

If you wanted to examine how S and t are related, you could use algebra to isolate the common r_{ranked} and solve for S or t, but for now, we will just use the t-statistic to determine the significance of rho.

Interpretating the Spearman's rho

There is a statistically significant strong positive correlation between basic access to drinking water and female education (rho = 0.77; p < .001). As the percent of the population with basic access to water increases, so does the percent of school-age females in school.

Assumptions for Spearman's rho

The assumptions for the Spearman's rho correlation are:

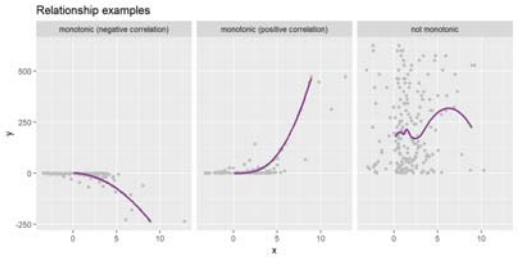
- The variables are at least ordinal on this scale:
  - *Nominal: Categories that don't have a logical order (e.g., religion, marital status)*
  - *Ordinal: Categories that have a logical order (e.g., highest level of school completed)*
  - *Interval: Can only take specific values along a continuum (e.g., number of cars in a parking lot, number of people in class)*
  - *Ratio: Can take any value along a continuum (e.g., height, weight)*
- the relationship between the two variables must be **monotonic**

Assumption 1: Variable types

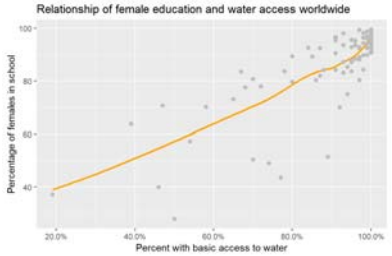
The female education and water variables are both *ratio*. This assumption is met.

Assumption 2: Monotonic relationship

A **monotonic** relationship is a relationship where one variable goes up as the other variable goes up, or one variable goes down while the other goes up. That is, the relationship does not have to follow a straight line, it can curve as long as it is always heading in the same direction. Here are a couple of examples to demonstrate:



Assumption 2: Testing the assumption



The line suggests that the relationship between female education and water access meets the monotonic assumption since the values of female education consistently go up as the values of access to water go up. The relationship does not change direction. This assumption IS MET.

Spearman's rho conclusion

Given the analyses meet both assumptions, report and interpret the Spearman correlation coefficient.

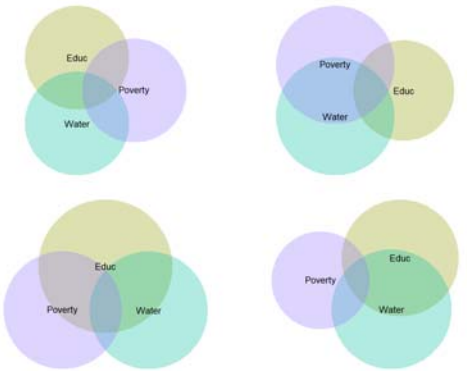
The interpretation:

There is a statistically significant positive correlation between basic access to drinking water and female education ($rho = 0.77$; $p < .001$). As the percent of the population with basic access to water increases, so does the mean years of education for female citizens. The data meet the assumptions for Spearman's rho.

Partial correlations

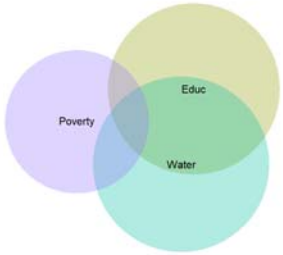
What if female education and water access are both related to poverty? What if this relationship is influencing how they are related to each other?

There is a method called *partial correlation* for examining how multiple variables overlap. Examine a few Venn Diagrams with different patterns of shared variance to clarify the idea:



Shared variance

- The different amounts of overlap representing the amounts of shared variance among the variables.
- For example, we might be interested in how much overlap there is between female education and water access after accounting for poverty.
- In the Venn Diagrams, this would be the overlap between yellow and green only.



Calculating partial correlations

The output shows the partial correlation between each pair of variables, while accounting for, or *controlling* for, the third variable. Partial correlation can be computed using Pearson's r or Spearman's rho depending on which assumptions are met. These are Spearman's rho given that the variables did not meet the Pearson's assumptions.

```
##                   perc.1dollar perc.basic2015water female.in.school
## perc.1dollar         1.0000000          -0.5977239       -0.2841782
## perc.basic2015water -0.5977239           1.0000000        0.4305931
## female.in.school    -0.2841782           0.4305931        1.0000000
```

Interpret one of these:

Controlling for poverty, the partial correlation between female education and basic water access is moderate (rho = 0.43).

Partial correlation significance

```
##                   perc.1dollar perc.basic2015water female.in.school
## perc.1dollar      0.000000e+00       2.312820e-07     0.0239966731
## perc.basic2015water 2.312820e-07       0.000000e+00     0.0004272781
## female.in.school    2.399667e-02       4.272781e-04     0.0000000000
```

The p-values indicate statistical significance for several of the correlation coefficients, including the correlation between females in school and basic water access.

Interpret a partial correlation:

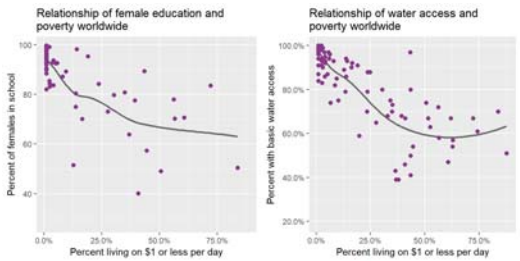
The partial correlation between percent of females in school and the percentage of citizens who have basic water access was moderate, positive, and statistically significant (r_s = 0.43; t = 3.73; p =.0004). Even after poverty is accounted for, increased basic water access is related to increased percent of females in school.

Assumption checking

The assumptions that applied to the two variables for a Pearson's r or Spearman's rho correlation would apply to all three variables for the corresponding partial correlation.

So, in this case, each variable would need to be at least interval and each pair of variables would need to have a monotonic relationship.

- Assumption 1: Variable type
  - *The variables are ratio, so this assumption IS MET.*
- Assumption 2: Monotonic relationship



Assumption checking results

In this case the analyses appear to meet the monotonic assumption for the poverty variable and the female education variable but not for the poverty variable and the basic water access variable.

The results can still be reported, but without meeting the assumptions for the statistical test, interpreting the statistical significance is a problem.

In situations like this there are a few possible strategies including: (1) interpreting the results for the sample only, and (2) recoding one of the variables to be categorical and using a different type of analysis.

Interpretation

Re-write the interpretation given the assumptions were not met:

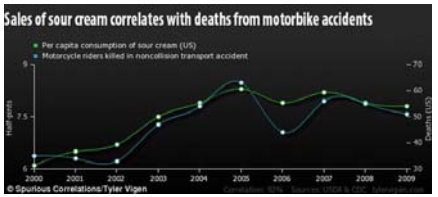
The partial Spearman's rho correlation between mean years of education for females and the percentage of citizens who have basic water access was moderate and positive ([] = 0.43. Even after poverty is accounted for, an increased basic water access was related to increased years of education for female citizens in this sample of countries.

Other options…proceed with caution

- In addition to the two strategies above, many people transform variables by taking the square root, log, or inverse of the variable values.
- These transformations often work to help analysts meet a normality assumption or other assumptions for a particular type of analysis.
- Although transformation of variables allows meeting of assumptions, the interpretation of results suffers.
- For example, instead of reporting that the percentage of females in school is positively correlated with poverty rate, an inverse transformation of poverty rate would result in the interpretation that the percentage of females in school is positively correlated with the inverse of the poverty rate.
- For this reason (and a few others), do your best to avoid data transformations whenever possible.

And finally…

Just because two variables are statistically significantly very strongly correlated does not mean that one causes the other.



The end

The challenge is on GitHub