Correlation Analysis in R, Part 1: Basic Theory

Petr Baranovskiy @ www.dataenthusiast.ca

2021-01-01 21:30:00

Introduction

There are probably tutorials and posts on all aspects of correlation analysis, including on how to do it in R. So why more?

When I was learning statistics, I was surprised by how few learning materials I personally found to be clear and accessible. This might be just me, but I suspect I am not the only one who feels this way. Also, everyone's brain works differently, and different people would prefer different explanations. So I hope that this will be useful for people like myself – social scientists and economists – who may need a simpler and more hands-on approach.

These series are based on my notes and summaries of what I personally consider some the best textbooks and articles on basic stats, combined with the R code to illustrate the concepts and to give practical examples. Likely there are people out there whose cognitive processes are similar to mine, and who will hopefully find this series useful.

Why correlation analysis specifically, you might ask? Understanding correlation is the basis for most other statistical models; in fact, these things are directly related, as you will see just a few paragraphs further down in this post.

Although this series will go beyond the basic explanation of what a correlation coefficient is and will thus include several posts, it is not intended to be a comprehensive source on the subject. Some topics won't be included because I do not know much about them, or (for example, Bayesian correlations) because I am not planning to include them. Since I will be focusing on performing correlation analysis in R, I won't be addressing basic statistical concepts such as variance, standard deviation, etc.

I was inspired to write these series by The Feynman Technique of Learning.

Correlation Coefficient

A variable that is related to another variable is a *covariate*. Covariates share some of their variance (hence *co*-variance). But variance depends on the scale of measurement, and thus is sensitive to changes in the scale. Therefore, when you measure covariance, it is not a very useful number in a sense that you can't say whether the variables share a lot or only a little bit of their variance by simply looking at the number.[1]

To overcome the problem of covariance being dependent on the measurement scale, we need a unit of measurement into which any scale of measurement can be converted. This unit is the standard deviation (SD). Converting units of measurement into SD units is known as *standardization*. In this case we are converting covariance, and it is done by simply dividing the covariance by the product of standard deviations of the covariates.

[1] Variance is measured in the units that are the *squares* of the units of the variable itself, e.g. m^2, kg^2, $hours^2$, etc. Naturally, the resulting number per se doesn't explain much. m^2 may *seem* clear (although in this case it isn't), but what is hour squared?

This gives us the ***Pearson product-moment correlation coefficient***, more often referred to as the ***Pearson correlation coefficient***, or simply the ***Pearson's*** r:

$$r = \frac{COVxy}{SDx * SDy}$$

So what does the correlation coefficient do in practical terms? Suppose, you have collected your data and now have a scatterplot in front of you. As such, you don't know much about how the variables are related in your data yet. So you'll have to start by somehow describing and summarizing your data.

The logical first step would be to find $mean_x$ and $mean_y$,[2] and then mark a point where they intersect – the *point of averages*. After we have found the point of averages, we can measure the spread of the data points using SDx and SDy,[3] which let us know how spread out the data is horizontally and vertically. But if we knew only the point of averages and the standard deviations of our variables, we still wouldn't know if and how strongly the variables are associated. This is exactly what the correlation coefficient tells us!

The ***correlation coefficient*** *is the measure of linear association between variables.* "If there is a strong association between two variables, then knowing one helps a lot in predicting the other. But when there is a weak association, information about one doesn't help much in guessing the other" (Freedman, Pisani, and Purves 1998, 121).

What does the *"measure of linear association"* mean? It simply means that if the relationship between the variables can be graphically summarized with a straight line, and the correlation coefficient measures how much the data points are clustered around the line. Further in this series, I will be using the letter r interchangeably with the term "correlation coefficient", unless I am speaking about a specific type of coefficient such as the Kendall's tau τ or the Spearman's rho ρ.

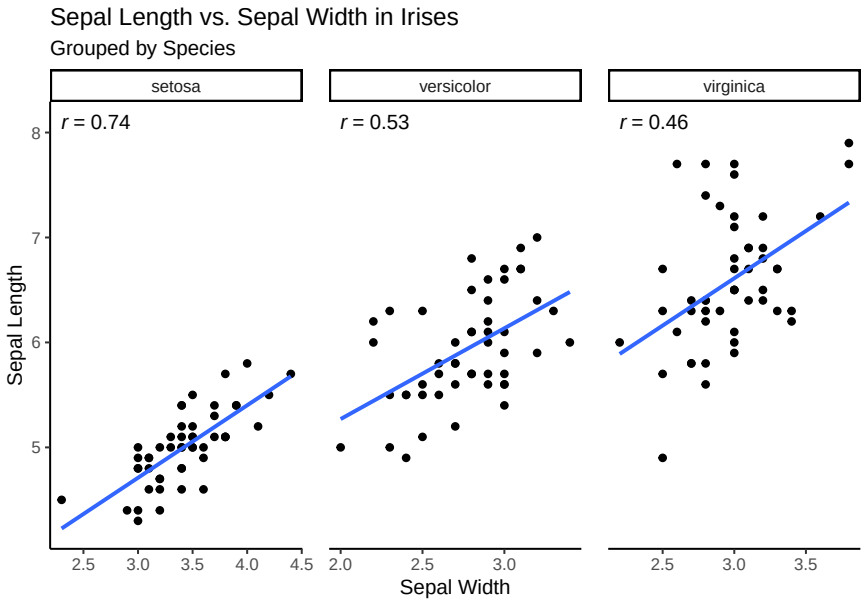
Let's illustrate these concepts using a scatterplot:

```
# load {tidyverse} for convenience
library(tidyverse)

ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length)) +
  facet_wrap(~ Species, scales = "free_x") +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE) +
  ggpubr::stat_cor(aes(label = tolower(..r.label..)), label.y = 8.1) +
  theme_classic() +
  theme(panel.spacing = unit(1, "lines")) +
  labs(x = "Sepal Width",
       y = "Sepal Length",
       title = "Sepal Length vs. Sepal Width in Irises",
       subtitle = "Grouped by Species")
```

[2] Or $\bar{X}$ and $\bar{Y}$ in proper mathematical notation, which I prefer to avoid because it makes things unnecessarily harder for beginners.

[3] A fancy math way of referring to (sample) standard deviation would be s_x and s_y .



The closer r is to 1 or -1, the tighter are the data points grouped around the line and the stronger is the association between the variables; the closer r is to 0, the looser is the grouping and the weaker is the association. If the coefficient is positive, the variables are positively associated (as x deviates from the mean, y deviates in the same direction), and the line goes up. If it is negative, they are negatively associated (as x deviates from the mean, y deviates in the opposite direction), and the line goes down.

Some important things to remember (Freedman, Pisani, and Purves 1998, 126, 128, 144–45):

- $r = .80$ does not mean that 80% of the points are tightly clustered around the line, nor does it indicate twice as much linearity as $r = .40$.
- $r = -.80$ indicates the same amount of clustering as $r = .80$.
- The appearance of a scatter diagram depends on r and on the SD: if the SDs are smaller, the points look more tightly clustered, same if the r is larger; scatterplots with the same r can look tightly or loosely clustered, depending on their SDs. Thus, r measures clustering not in absolute terms but relative to the SD.

Correlation Coefficient Features

Correlation coefficient has some interesting features that stem from the fact that it is a pure number, i.e. does not depend on the units in which the variables are measured. I will illustrate these using the `iris` dataset, but first I will make my own version `iris_l` (`l` for "local") with shorter variable names, so that the code is more concise and readable:

```
# shorten var names for convenience
iris_l <- select(iris, sl = Sepal.Length, sw = Sepal.Width)
```

- $CORxy = CORyx$:

```
cor(iris_l$sl, iris_l$sw) == cor(iris_l$sw, iris_l$sl)
```

```
## [1] TRUE
```

- Adding or subtracting the same number to *one or both* variables doesn't change r:

```
iris_l <- iris_l %>%
  mutate(sl_plus = sl + 7.25, sw_plus = sw + 7.25) %>%
  mutate(sl_minus = sl - 7.25, sw_minus = sw - 7.25)
```

```
cor(iris_l$sl, iris_l$sw)
cor(iris_l$sl_plus, iris_l$sw_plus)
cor(iris_l$sl_minus, iris_l$sw_minus)
cor(iris_l$sl, iris_l$sw_plus)
cor(iris_l$sl_minus, iris_l$sw)
```

r stays the same: -0.1175698

- Multiplying *one or both* variables by the same *positive* number doesn't change r:

```
iris_l <- mutate(iris_l, sl_mpos = sl * 3.5, sw_mpos  = sw * 3.5)
```

```
cor(iris_l$sl, iris_l$sw_mpos)
```

```
## [1] -0.1175698
```

```
cor(iris_l$sl_mpos, iris_l$sw_mpos)
```

```
## [1] -0.1175698
```

- Multiplying *both* variables by the same *negative* number doesn't change r; multiplying *only one* variable by a *negative* number changes the direction (sign) but not the absolute value of r:

```
iris_l <- mutate(iris_l, sl_mneg = sl * -3.5, sw_mneg  = sw * -3.5)
```

```
cor(iris_l$sl_mneg, iris_l$sw_mneg)
```

```
## [1] -0.1175698
```

```
cor(iris_l$sl, iris_l$sw_mneg)
```

```
## [1] 0.1175698
```

Keep these features in mind when converting units in which your variables are measured, e.g. if you are converting temperature from Fahrenheit to Celsius.

Also, here are some important (and hopefully, unnecessary) reminders:

$$-1 \leq r \leq 1$$

$$correlation \neq causation$$

$$correlation \neq no\ causation$$

Correlation Coefficient and the Regression Method

Although the meaning of the phrase "measure of association between variables" is hopefully now clear, the concept may still seem a bit abstract. How can it help me in my analysis? At least for a linear association, the answer is quite straightforward: *r predicts by how much y will change upon the change in x – in standard deviation units, not in the original units of measurement.* Fortunately, since we know the SD, we can then easily convert SD units into the original units in our model.

This is known as ***the regression method***, which can be formulated as follows: *associated with each change of one SD in x, there is a change of r * SD in y, on average* (Freedman, Pisani, and Purves 1998, 160).

Let's illustrate how the regression method works using my favorite dataset about penguins (Gorman, Williams, and Fraser 2014), because penguins are cool. First, let's install the package with data:

```
# install palmerpenguins
install.packages("palmerpenguins")
```

Then, let's get data for one penguin species:

```
# load data
library(palmerpenguins)

# select gentoo penguins
gentoo <- penguins %>%
  filter(species == "Gentoo") %>%
  select(c(2, 3, 6)) %>% # keep only relevant data
  drop_na()
```

Then, let's explore how bill length correlates with body mass in Gentoo penguins and calculate SDs for both variables. We'll be needing the results of these calculations, so I am saving them as separate data objects. Let's also take a look at the mean values of these variables:

```
gentoo_r <- cor(gentoo$bill_length_mm, gentoo$body_mass_g)
sdx <- sd(gentoo$bill_length_mm)
sdy <- sd(gentoo$body_mass_g)
```

```
gentoo_r
```

```
## [1] 0.6691662
```

```
sdx
```

```
## [1] 3.081857
```

```
sdy
```

```
## [1] 504.1162
```

```
mean(gentoo$bill_length_mm)
```

```
## [1] 47.50488
```

```
mean(gentoo$body_mass_g)
```

```
## [1] 5076.016
```

In this example $r = 0.67$ (approx.). Remember that r shows by how much y changes *in SD units* when x changes by 1 SD unit. Since we know the SDs for bill length and body mass of Gentoo penguins, we can predict the body mass of a Gentoo penguin using the regression method:

First, let's find out how much body mass will change when bill length changes by ± 1 SD.

$$0.67 * SD(body\ mass) = 0.67 * 504 = 337.68$$

Thus, when bill length changes by ± 3.08 mm, i.e. by 1 $SD(bill length)$, body mass changes by ± 337.68 grams in the same direction (because r is positive). This means that a Gentoo penguin with a bill length of 50.59 mm (approx. 1 SD above the mean) will have the *predicted* body mass of:

$$mean(body\ mass) + r * SD(body\ mass) = 5076 + 337.68 = 5413.68\ grams$$

I should stress that this is the ***predicted value***. One can also think about it as the *most likely value of y at the corresponding value of x*. Actual values, of course, vary. Note also how SD units got converted into actual measurement units when we did our calculations.

Correlation Coefficient and Linear Regression (and Penguins)

Correlation should not be confused with regression, since r is standardized, and the regression equation is unit-specific. But does r matter for the regression equation? You bet! It is the key component of the formula for the *slope of the regression line*:[4]

$$r * \frac{SDy}{SDx}$$

And the regression equation itself is: $\hat{y} = intercept + x * slope$,[5] or:

$$\hat{y} = intercept + x * (r * \frac{SDy}{SDx})$$

The basic concepts should be more or less clear by now, so let's translate them into actual calculations. Using the regression equation (of which our correlation coefficient `gentoo_r` is an important part), let us predict the body mass of three Gentoo penguins who have bills 45 mm, 50 mm, and 55 mm long, respectively.

First, let's find the intercept:

```
gentoo_lm <- lm(body_mass_g ~ bill_length_mm, data = gentoo)
intercept <- as.numeric(gentoo_lm$coefficients[1])
```

And then paste the values (bill lengths, the correlation coefficient `gentoo_r`, and standard deviations `sdx` and `sdy`) into the regression equation. This will produce the predicted body mass values in grams:

[4] The slope of the regression line is also known as the ***regression coefficient***, which should not be confused with the *correlation* coefficient. The ***regression line*** is the line that estimates the average value for y corresponding to each value of x (Freedman, Pisani, and Purves 1998, 160).

[5] The ***intercept*** is simply the predicted value of y when $x = 0$. Keep in mind that the intercept is a *theoretical* value that may or may not be practically meaningful (e.g. you might have an intercept of -1010 grams, which makes no practical sense). It is called "intercept" because at $x = 0$ the regression line intersects with the Y axis of a plot. Also, did you note this little hat on top of $\hat{y}$? It is simply a mathematical way of saying that this is the *predicted* value of y.

```r
x <- c(45, 50, 55)
intercept + x * (gentoo_r * sdy/sdx)
```

```
## [1] 4801.834 5349.130 5896.426
```

Note that this is not how you run a predictive linear model in R. Above I just wrote the regression equation as close to its textbook form as possible using the R code. In practice, you'd be using the `predict()` function, which is far more convenient:

```r
x <- data.frame(bill_length_mm = c(45, 50, 55))
predict(gentoo_lm, x)
```

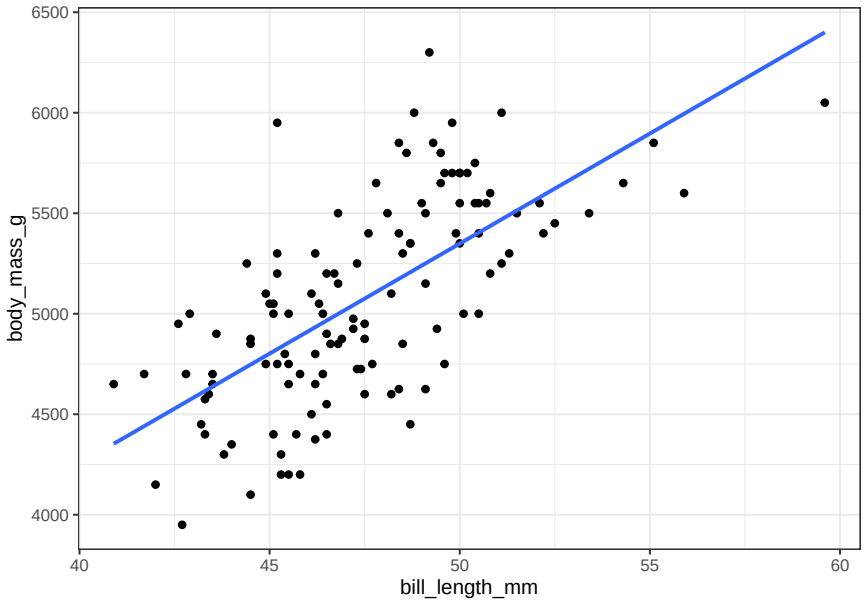
```
##        1        2        3
## 4801.834 5349.130 5896.426
```

Keep in mind that `predict()` requires a specific syntax not only inside `predict()`, but also inside `lm()`:

- inside `lm()`, use the `data` argument to refer to the dataset, do not subset with `$` (i.e. do *not* write the formula as `gentoo$body_mass_g ~ gentoo$bill_length_mm`),
- predictor variable should be passed as a *dataframe*, not vector, and
- predictor variable should have the same name as it had in the model (in our example, the predictor is named `bill_length_mm` inside both the `lm()` function and in the `x` dataframe).

We can also visually check the output of our regression equation and of the `predict()` function against a plot:

```r
gentoo %>%
  ggplot(aes(x = bill_length_mm, y = body_mass_g)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE) +
  theme_bw()
```



You can play around with assigning different values to `x` and re-running the calculations to predict weights of different penguins. Of course, always keep in mind that the values produced by a statistical model are just that – *predicted* values – and may or may not make sense. For example, if you set bill length to 1 meter, you'll get an enormous penguin that weighs 109.3 kg, which should not be possible without advanced genetic engineering. Or if you set it to a negative 50 mm, you'll get a penguin with negative mass, which might exist somewhere in the realms of theoretical physics but certainly not in the real world. Always use common sense when building your model.

Statistical Significance and Confidence Intervals of the Correlation Coefficient

First, a few quotes and definitions (emphasis mine):

> The observed ***significance level*** is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null (Freedman, Pisani, and Purves 1998, 481).

> The ***P-value*** of a test is the chance of getting a big test statistic – assuming the null hypothesis to be right. *P is not the chance of the null hypothesis being right* (Freedman, Pisani, and Purves 1998, 482).

> Scientists test hypotheses using probabilities. In the case of a correlation coefficient … if we find that the observed coefficient was very unlikely to happen if there was no effect [correlation] in the population, then we can gain confidence that the relationship that we have observed is statistically meaningful (Field, Miles, and Field 2012, 210).

I will not be providing a more in-depth breakdown of the concept of statistical significance here, as there are some great explanations in textbooks (including the ones listed in the Bibliography section of this post) and online (for example, here and here). I particularly recommend chapters 26 "Tests of Significance" and 29 "A Closer Look at Tests of Significance" in Freedman, Pisani, and Purves (1998), where statistical significance is explained very simply and clearly. Most importantly, they address common misunderstandings and misuses of significance testing and p-values, accompanied by detailed examples.

Briefly defining a confidence interval (CI) is much harder, as this deceptively simple topic can be *very* easily misinterpreted. I thus strongly recommend to read at least chapter 21 "The Accuracy of Percentages" in Freedman, Pisani, and Purves (1998), as well as an open-access article by Sim and Reid (1999). *After* you have read these, take a look at this simulation for a nice visual aid. Here I will only provide the formal and informal definitions, and a short but important quote:

> Over infinite repeated sampling, and in the absence of selection, information, and confounding bias, the α-level confidence interval will include the true value in α% of the samples for which it is calculated (Naimi and Whitcomb 2020).

> If we were to draw repeated samples from a population and calculate a 95% CI for the mean of each of these samples, the population mean would lie within 95% of these CIs. Thus, in respect of a particular 95% CI, we can be 95% confident that this interval is, of all such possible intervals, an interval that includes the population mean rather than an interval that does not include the population mean. *It does not … express the probability that the interval in question contains the population mean, as this [probability] must be either 0% or 100%* (Sim and Reid 1999).[6]

[6] When thinking about this, replace the word "mean" with the words "statistic" and "parameter" (for the sample and the population, respectively), since CIs can be calculated for pretty much any statistic, not just the mean.

> The chances are in the sampling procedure, not in the parameter (Freedman, Pisani, and Purves 1998, 384).

So, let's illustrate the correlation coefficient's CI and p-value using the `rstatix::cor_tes()` function. Same can be done with R's default `stats::cor.test()`, but I prefer `rstatix` because it returns output as a dataframe instead of a list, as well as for other reasons to be addressed in detail in the next post in this series.

```
install.packages("rstatix")
```

First, let's test the *two-directional* hypothesis that body mass is correlated with bill length in Gentoo penguins. It is called two-directional because our test includes two possibilities: that a higher bill length is correlated with a *higher* body mass, and that a higher bill length is correlated with a *lower* body mass. The null-hypothesis (H_0) is that they are uncorrelated, i.e. that bill length and body mass change independently of each other.

```
# Two-directional test
rstatix::cor_test(gentoo, bill_length_mm, body_mass_g)
```

```
## # A tibble: 1 x 8
##   var1           var2          cor statistic        p conf.low conf.high method
##   <chr>          <chr>       <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 bill_length_mm body_mass_g  0.67      9.91 2.68e-17    0.558     0.757 Pearson
```

As we see, `bill_length_mm` and `body_mass_g` are positively correlated: $r = 0.67$, and the results are highly statistically significant due to an extremely low p-value: 2.68e-17, which allows us to reject the null hypothesis. In other words, we can say that bill length and body mass are likely correlated not just in our sample, but in the whole population of Gentoo penguins. The output also gives us the lower and higher limits of the 95% CI (default) for the correlation coefficient. You can set a lower or a higher confidence level for the CI with the `conf.level` argument. For example:

```
# 99% CI - see how CI changes
rstatix::cor_test(gentoo,
                  bill_length_mm, body_mass_g,
                  conf.level = 0.99)
```

```
## # A tibble: 1 x 8
##   var1           var2          cor statistic        p conf.low conf.high method
##   <chr>          <chr>       <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 bill_length_mm body_mass_g  0.67      9.91 2.68e-17    0.518     0.780 Pearson
```

Try assigning a lower confidence level (e.g. 90%) and see how it affects the CI.

Let's now test if higher bill length correlates with a higher body mass (this would be a common sense assumption to make):

```
# Directional test: greater
rstatix::cor_test(gentoo,
                  bill_length_mm, body_mass_g,
                  alternative = "greater")
```

```
## # A tibble: 1 x 8
##   var1           var2          cor statistic        p conf.low conf.high method
##   <chr>          <chr>       <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 bill_length_mm body_mass_g  0.67      9.91 1.34e-17    0.578         1 Pearson
```

Based on the outcome of this test, we can say that a *higher* bill length is likely correlated with a *higher* body mass not just in our sample, but in the whole population of Gentoo penguins.

Finally, let's test if a higher bill length correlates with a *lower* body mass, i.e. if it could be that the longer the beak, the smaller the penguin. Admittedly, this does not sound like a particularly plausible idea, but sometimes our research can produce unexpected results – this is when it can be most fun:

```
# Directional test: less
rstatix::cor_test(gentoo,
                  bill_length_mm, body_mass_g,
                  alternative = "less")
```

```
## # A tibble: 1 x 8
##   var1           var2          cor statistic     p conf.low conf.high method
##   <chr>          <chr>       <dbl>     <dbl> <dbl>    <dbl>     <dbl> <chr>
## 1 bill_length_mm body_mass_g  0.67      9.91     1       -1     0.744 Pearson
```

This time no surprises – this particular result was not significant at all at $p = 1$, which is the highest a p-value can ever be. This means that the data is fully consistent with H_0.[7]

Note how in all our tests r remained the same, but p-value and CI changed depending on the hypothesis we were testing.

R-squared

The ***coefficient of determination***, R^2 is a measure of the amount of variability in one variable that is shared by the other variable (Field, Miles, and Field 2012, 222).[8] "When we want to know if two variables are related to each other, we … want to be able to … explain some of the variance in the scores on one variable based on our knowledge of the scores on a second variable"(Urdan 2011, 87). In other words, when variables are correlated, they share a certain proportion[9] of their variance, which is known as the ***explained variance*** or ***shared variance***. R^2 is a very valuable measure, as it tells us the proportion (or the percentage if we multiply R^2 by 100) of variance in one variable that is shared by the other variable. Calculating R^2 is very simple:

$$R^2 = r^2$$

If the concept of shared variance is still not entirely clear, take a look at this visualization, where R^2 is explained graphically as a Venn diagram.

Note that the term *coefficient of determination* may be somewhat misleading. Correlation by itself does not signify causation, so there is no reason why it would magically become causation when you square the correlation coefficient. Sometimes people refer to R^2 as "the variance in one variable explained by the other", but we should remember that this does not imply causality. Also, this is why I think that the term *shared variance* is preferable to the *explained variance*.

[7] The null hypothesis in this case would be that as bill length *increases*, body mass does not *decrease*. And indeed, it does not. On the contrary, body mass goes up, as we see from the previous tests.

[8] Pearson's product-moment correlation coefficient is denoted by both r and R. Typically, the upper-case R is used in the context of regression, because it represents the multiple correlation coefficient; however, for some reason when we square r, an upper case R is used (Field, Miles, and Field 2012, 209). I think this is done because academic types are evil and enjoy confusing people :)

[9] Which can also be expressed as a percentage.

Finally, let's calculate R^2 in R using bill length and body mass of Gentoo penguins as our covariates:

```
# expressed as a proportion:
cor(gentoo$bill_length_mm, gentoo$body_mass_g)^2
```

```
## [1] 0.4477834
```

```
# expressed as a percentage:
cor(gentoo$bill_length_mm, gentoo$body_mass_g)^2 * 100
```

```
## [1] 44.77834
```

This concludes the basic theory and definitions behind correlation analysis. In the next post, I will focus on performing and reporting correlation analysis in R.

This post is also available as a PDF.

Bibliography

Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. First edit. London, Thousand Oaks, New Delhi, Singapore: SAGE Publications.

Freedman, David, Robert Pisani, and Roger Purves. 1998. *Statistics*. Third edit. New York, London: W.W. Norton & Company.

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. 2014. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis)." *PLoS ONE* 9 (3). https://doi.org/10.1371/journal.pone.0090081.

Naimi, Ashley I, and Brian W Whitcomb. 2020. "Can Confidence Intervals Be Interpreted?" *American Journal of Epidemiology* 189 (7): 631–33. https://doi.org/10.1093/aje/kwaa004.

Sim, Julius, and Norma Reid. 1999. "Statistical inference by confidence intervals: Issues of interpretation and utilization." *Physical Therapy* 79 (2): 186–95. https://doi.org/10.1093/ptj/79.2.186.

Urdan, Timothy C. 2011. *Statistics in Plain English*. New York: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780203851173.