Correlation or regression, that's the question

Saskia le Cessie^{1,2}, Rolf H H Groenwold^{1,2} and Olaf M Dekkers^{1,3}

¹Department of Clinical Epidemiology, ²Department of Biomedical Data Sciences, and ³Department of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
should be addressed
to S le Cessie

Email
S.le_Cessie@lumc.nl

Abstract

There are different ways to quantify the relation between two or more continuous variables. Some researchers use correlation coefficients; others will apply regression methods such as linear regression. In this paper, we show that the choice between correlation and regression is not purely a statistical one but largely depends on the research aims. Importantly, one should always inspect the data before using either of the two methods.

*European Journal of
Endocrinology*
(2021) **184**, E15–E18

Suppose researchers measure blood glucose and cortisol levels in several individuals and aim to quantify the relationship between these two variables. This can be done by calculating a correlation coefficient or by using the regression coefficient obtained from a simple linear regression analysis. But which method to use? From a purely statistical point of view, it does not matter. If we know the correlation coefficient (unless stated otherwise, we refer to the parametric Pearson correlation coefficient), it is very easy to calculate the corresponding linear regression coefficient and vice versa. Furthermore, both methods will yield exactly the same *P*-value when testing for a relation between the two hormone measures. The choice between correlation and regression does not depend on the statistical properties of the data. Instead, it's the context of the research that matters: the research question, the aim of the study, and whether one wants to compare results between different studies.

Correlation

Let's first look at the correlation coefficient. This coefficient captures the degree to which the relationship between two variables is linear; the coefficient takes values between -1 and 1 . It is positive when higher values of, for example, glucose co-occur with higher values of cortisol, negative when higher values of glucose co-occur

with lower values of cortisol, and it will take the value 0 in the absence of a linear relation. Values of 1 or -1 indicate a perfect linear relation; all observations lie on a straight line. Figure 1 shows some examples including the correlation coefficient.

The correlation coefficient is a measure of association where the two variables play the same role and can be interchanged; the correlation between glucose and cortisol is the same as the correlation between cortisol and glucose. Within a study, it can be used to explore which variables are most strongly associated, with correlation coefficients close to 1 or -1 pointing to the strongest linear relation. However, we will later see that the correlation is less useful when used to compare results across different studies.

Regression

Another option for assessing the relationship between two continuous variables is linear regression. For example, we may explore how the mean glucose level changes as a function of cortisol and consider the linear model: $\text{mean glucose} = \alpha + \beta \times \text{cortisol}$. Note that, in this statistical approach, glucose and cortisol here have different roles, glucose is the outcome variable we want to explain, also called the dependent variable, because its mean value depends on the cortisol value, which is called the independent variable. The regression

Figure 1

Four examples of relationship between two variables and the corresponding correlation coefficient.

coefficient β indicates the mean change in glucose per unit increase in cortisol; it is the slope of the regression line. The regression coefficient β is unit dependent; if we measure cortisol in $\mu\text{g/dL}$ or $\mu\text{mol/L}$, the value of the slope of the regression line will differ (Fig. 2). Of note, the correlation coefficient and P -values are insensitive to such changes in the units. In regression, the two variables are not interchangeable. If we switch the dependent and independent variables and perform a regression analysis with cortisol as dependent variable and glucose as independent variable, we would obtain a different regression coefficient and a different regression line. Mind, however, that here the accompanying research question is how glucose influences cortisol levels. This shows that content knowledge should guide the statistical model, not the other way round. Linear regression can be expanded to adjust for confounding in a multivariable model (1).

Regression versus correlation

It does not matter whether statistical testing for an association between two variables is done by testing if the correlation coefficient differs from zero or if the slope of the regression line differs from zero; both approaches will give exactly the same test statistic and P -value. We have already discussed one important difference between correlation and regression, the fact that in correlation

Figure 2

Impact of the choice of units on the slope of regression line. In the top panel, cortisol is measured in $\mu\text{mol/dL}$ and in the bottom panel, in $\mu\text{g/L}$. The dashed line is the estimated regression line, the corresponding equation is given in the plot. The correlation coefficient (r) and P -values (P) are insensitive to such changes in the units.

both variables play the same role, while in regression, the mean of one variable is modelled as a function of the other. But there are more issues to consider:

- I. A correlation depends on the range of the variables, which limits comparisons over studies.

This is demonstrated in Fig. 3. The results are from a hypothetical study with 60 individuals, where the

relation between age and BMI is studied. In this study, age varies between 20 and 80 years and the correlation coefficient between age and BMI is 0.60. In the second plot, the results are shown when only the subgroup of individuals between age 40 and 60 is considered. The correlation is now much smaller ($r=0.26$), due to the reduction in variation of age. In contrast, the regression line remains the same (although the estimated coefficients may slightly differ because of random

variation). This phenomenon limits comparisons of correlation coefficients over studies. If one study finds a correlation of 0.7, while another study finds a correlation of 0.2 for the same two variables, this does not necessarily imply that the relation between the variables differs between studies, but could also be due to a more restricted population in the latter study. This implies that combining correlation coefficients in a meta-analysis should be done cautiously.

II. Be aware of outliers. They may greatly affect both correlations and regression models.

Both regression methods and correlations can be heavily influenced by extreme observations. Figure 4 shows an example, where the relationship between cortisol and insulin is studied. In the upper right corner, there is one very extreme observation with an insulin value of around 200 (mE/L). Both the regression analysis and the correlation indicate a positive association, the correlation coefficient is equal to 0.50 (95% CI: 0.22; 0.70, $P=0.001$). However, everything changes if we remove this one very influential point. The correlation drops to -0.15 (95% CI: -0.44 ; 0.18, $P=0.37$) and the regression line now has a negative slope. This example shows that we always have to examine our data carefully before performing the

Figure 3

Impact of age restriction on the correlation between age and BMI in a hypothetical study of this relation. First, the relation between age and BMI in the total population with age range 20–80 years; next, the subgroup of individuals between 40 and 60. In the subgroup the correlation coefficient is much smaller.

Figure 4

Impact of one influential observation on a study of the relationship between cortisol and insulin. The dashed line is the estimated regression line when the influential point at the upper right corner is included (correlation $r = 0.50$), the solid line is the obtained regression line after excluding the influential point ($r = -0.15$).

Figure 5

Example of a non-linear relation between BMI and cortisol.

The dashed line is the estimated linear regression line.

statistical analyses (2). Note that influential outliers will more often occur when variables are not normally distributed, and a logarithmic transformation may in some situations be advisable (see (3) for more details). Only if a value is biologically impossible, it should be removed from the analysis.

III. Correlations and linear regression assume a linear relation between variables.

Consider Fig. 5 where hypothetical data on BMI and cortisol levels are shown, motivated by a real study (4). There is a clear relationship between BMI and cortisol, the relationship is, however, not linear, and better described as U-shaped. The correlation coefficient in this example is close to zero ($r = -0.01$) and the regression line is very flat. The point is that if we ask the statistical program for linear regression, it will do so, even if the data do clearly not show a linear pattern. Such patterns can be missed when raw data are not inspected, for example, in a scatterplot. In the example from Fig. 5, a statistical model allowing for the U-shaped would be more appropriate (e.g. by

adding a quadratic term to the model or using more advanced non-linear modelling options).

Conclusion

The choice between correlation and regression should depend on whether one is interested in the relationship between two variables (which in case of a correlation have the same role) or whether interest is the effect of one (or more) variable(s) on another (dependent) variable. Furthermore, correlations are problematic to compare across studies, as they depend on the spread of the variables. As both the standard linear regression model and correlations assume a linear relationship between variables, and both can be heavily influenced by extreme outliers, we advise to always perform a visual inspection of the data by making scatterplots before using any of these methods.

Declaration of interest

Olaf M Dekkers is a Deputy Editor for European Journal of Endocrinology. He was not involved in the review or editorial process for this paper, on which he is listed as an author. The other authors have nothing to disclose.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

References

- 1 Groenwold RHH & Dekkers OM. Methodology for the endocrinologist: basic aspects of confounding adjustment. *European Journal of Endocrinology* 2020 **182** E5–E7. (<https://doi.org/10.1530/EJE-20-0075>)
- 2 Huebner M, le Cessie S, Schmidt C & Vach W. A contemporary conceptual framework for initial data analysis. *Observational Studies* 2018 **4** 171–192.
- 3 le Cessie S, Goeman JJ & Dekkers OM. Who is afraid of non-normal data? Choosing between parametric and non-parametric tests. *European Journal of Endocrinology* 2020 **182** E1–E3. (<https://doi.org/10.1530/EJE-19-0922>)
- 4 Schorr M, Lawson EA, Dichtel LE, Klibanski A & Miller KK. Cortisol measures across the weight spectrum. *Journal of Clinical Endocrinology and Metabolism* 2015 **100** 3313–3321. (<https://doi.org/10.1210/JC.2015-2078>)

Received 6 March 2021

Accepted 13 April 2021