Quantitative Understanding in Biology Short Course
Session 6
Correlation and Regression

Jason Banfelder
Luce Skrabanek

April 15th, 2025

1 Correlation

Linear correlation and linear regression are often confused, mostly because some bits of the math are similar. However, they are fundamentally different techniques. We'll begin this section of the course with a brief assessment of linear correlation, and then spend a good deal of time on linear and non-linear regression.

If you have a set of pairs of values (call them x and y for the purposes of this discussion), you may ask if they are correlated. Let's spend a moment clarifying what this actually means. First, the values must come in pairs (e.g., from a paired study). It makes no sense to ask about correlation between two univariate distributions.

Also, the two variables must both be observations or outcomes for the correlation question to make sense. The underlying statistical model for correlation assumes that both x and y are normally distributed; if you have systematically varied x and have corresponding values for y, you cannot ask the correlation question (you can, however, perform a regression analysis). Another way of thinking about this is that in a correlation model, there isn't an independent and a dependent variable; both are equal and treated symmetrically. If you don't feel comfortable swapping x and y, you probably shouldn't be doing a correlation analysis.

The standard method for ascertaining correlation is to compute the so-called Pearson correlation coefficient. This method assumes a linear correlation between x and y. You could have very well correlated data, but if the relationship is not linear the Pearson method will underestimate the degree of correlation, often significantly. Therefore, it is always a good idea to plot your data first. If you see a non-linear but monotonic relationship between x and y you may want to use the Spearman correlation; this is a non-parametric method. Another option would be to transform your data so that the relationship becomes linear.

In the Pearson method, the key quantity that is computed is the correlation coefficient, usually written as r. The formula for r is:

$$r = \frac{1}{n} \sum \left[\frac{(x_i - \overline{x})}{\mathrm{SD}_x} \cdot \frac{(y_i - \overline{y})}{\mathrm{SD}_y} \right] \tag{1}$$

The correlation coefficient ranges from -1 to 1. A value of zero means that there is no correlation between x and y. A value of 1 means there is perfect correlation between them: when x goes up, y goes up in a perfectly linear fashion. A value of -1 is a perfect anti-correlation: when x goes up, y goes down in an exactly linear manner.

Note that x and y can be of different units of measure. In the formula, each value is standardized by subtracting the average and dividing by the SD. This means that we are looking at how far each value is from the mean in units of SDs. You can get a rough feeling for why this equation works. Whenever both x and y are above or below their means, you get a positive contribution to r; when one is above and one is below you get a negative contribution. If the data are uncorrelated, these effects will tend to cancel each other out and the overall r will tend toward zero.

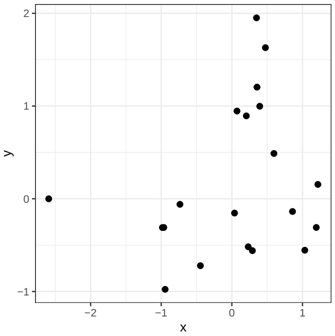
A frequently reported quantity is r^2. For a linear correlation, this quantity can be shown to be the fraction of the variance of one variable that is explained by the other variable (the relationship is symmetric). If you compute a Spearman correlation (which is based on ranks), r^2 does not have this interpretation. Note that for correlation, we do not compute or plot a 'best fit line'; that is regression!

Many people take their data, compute r^2, and, if it is far from zero, report that a correlation is found, and are happy. This is a somewhat naïve approach. Now that we have a framework for statistical thinking, we should be asking ourselves if there is a way to ascertain the statistical significance of our computed r or r^2. In fact there is; we can formulate a null hypothesis that there is no correlation in the underlying distributions (they are completely independent), and then compute the probability of observing an r value as large or larger in magnitude as the one we actually observed just by chance. This p-value will be a function of the number of pairs of observations we have, as well as of the values themselves. Similarly, we can compute a CI for r. If the p-value is less than your pre-established cutoff (or equivalently, if your CI does not include zero), then you may conclude that there is a statistically significant correlation between your two sets of observations.

To compute the correlation between two sets of (paired) measurements in Prism, begin by entering the data into an "XY" data sheet. Then click analyze, and choose "Correlation"; Prism will dutifully report the relevant r, r^2, and p-value. When using the default options, a 95% CI for r will be reported. If you're interested in the 95% CI of r^2, you can square each limit of the CI independently, but be careful to take due account of changes in sign. For example, if the 95% CI of r is -0.3 − 0.5, then the 95% CI of r^2 is 0 − 0.25.

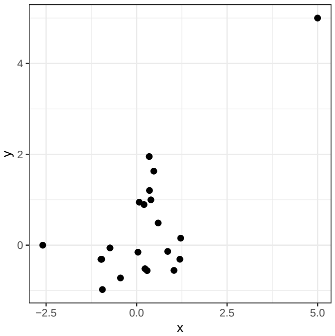
In many cases, you may see weak r^2s reported in the literature, but no p-value or CI. If you wish, you can compute a p-value yourself just by knowing n (the number of pairs) and r; see a text if you need to do this.

A important point about linear correlation is that it is sensitive to outliers. Let's explore this with an example. We begin by considering 20 uncorrelated datapoints, which are plotted here, and presented numerically in an appendix and available in the supplementary file `uncorr.txt` for you to import into Prism.



The correlation coefficient, r, for these data is 0.24, with a CI of -0.23 − 0.62. The fact that the CI includes zero indicates that it is plausible that there really is no correlation between x and y, which is consistent with the high (non-significant) p-value.

Now, let's add one outlier, say the point $x = 5, y = 5$. The plot now looks like this...



...and the correlation coefficent is now 0.7, with a CI of 0.39 − 0.87. We also have a significant p-value. Recall that one of the assumptions of the correlation test is that both x and y are normally distributed. Do you think that this applies in this case?

2 Introduction to Modeling

Regression, or curve fitting, is a much richer framework than correlation. There are several reasons why we may want to perform a regression analysis:

1. Artistic: We want to present our data with a smooth curve passing near the points. We don't have any quantitative concerns; we just want the figure to "look good".
2. Predictive: We want to use the data we've collected to predict new values for an outcome, given measured values for an explanatory variable. This may be, for example, a standardization curve for an instrument or assay. We'd like our predictions to be as consistent with our data as possible, and we don't care too much about the math that generates our predicted values (although simpler equations would be preferred for practical purposes).
3. Modeling: We want to test a hypothesized mechanistic model of a system against data, or we wish to use a mechanistic model we believe in to predict new data. In this case, we do care about the mathematical structure of our model, as it is derived from (or informs) our mechanistic model.

As basic scientists trying to figure out how the world works, we will focus on the third technique.

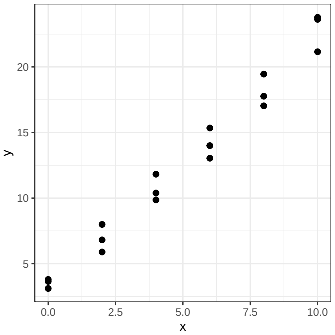
In accordance with Occam's Razor, all else being equal, we prefer simpler models to more complex ones. In mathematical terms, we prefer:

- Fewer explanatory variables
- Fewer parameters (model coefficients, etc.)
- Linear over non-linear relationships
- Monotonic vs. non-monotonic relationships
- Fewer interactions among explanatory variables

Of course, there is always a natural tension, because more complex models tend to fit our data better. There is always a tradeoff between model complexity and accuracy, and scientific judgment is often necessary to resolve this inherent conflict. As we shall see, there are some objective techniques that can help.

3 Simple Linear Regression

We'll start with the mechanics of a simple linear regression; you have probably done this before. Say we have our pairs of values, shown in the plot below, and we wish to fit a line to them. These data are shown in the appendix, and found in the supplementary file named `linear.txt`.



In this case, simple inspection makes it clear that a linear model would be appropriate. It is also worth noting the structure of the data: there are three datapoints for each time value.

To perform a linear regression on this data in Prism, begin again by importing the data into an XY datasheet. Then click Analyze, and choose linear regression, and Prism will derive a linear model that best fits the data, and the resultant line will be added to your plot of the datapoints.

We see that the best-fit line has a slope of 1.909 and an intercept of 3.147. Prism also reports the 95% CI of the slope and intercept of this regression line.

Furthermore, we get an r^2 of 0.978, which is the same value you would get if you performed a correlation analysis on the same dataset. This is one of the reasons why correlation and regression are often confused. For regression, though, life is not as simple as just looking at r^2.

When you perform a basic regression (linear or otherwise), the model parameters are chosen to minimize the sum of the squares of the residuals. A residual is the difference between a predicted and an observed value. This breaks the symmetry in the mathematics between x and y. You will get different lines if you regress $y = f(x)$ and $x = g(y)$.

Note that in the example above, there were three datapoints for each x value. Prism gives you the option to enter data like this with three y columns. While this is fine, if you do so, when performing a linear regression, the default option in Prism will be to regress to the **mean** y-value for each x-value. We recommend that you don't do this! By taking averages before regression, you're throwing away important information about the variance of the underlying data. While you might get tighter CIs on the estimated parameters, this is likely to be an overestimation of the certainty of the parameter estimates. By keeping all of your data in one column, you can't make this choice.

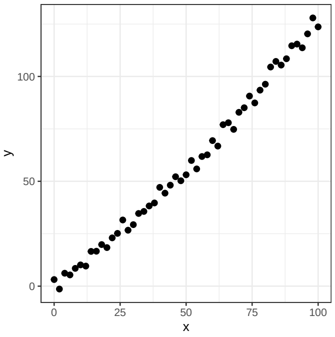
We have used the term linear a few times without formally defining what that means. We've taken for granted here that a linear regression fits data to a model of the form $y = mx + b$, and for present purposes, this is what we'll take "linear" to mean. However, you should be aware that in other contexts, the word "linear" implies something slightly different; for instance, linear algebra is the study of systems that follow the form $\mathbf{y} = \mathbf{A} \cdot \mathbf{x}$; this is a matrix equation whose one-dimensional analog is $y = ax$; there is no intercept term. A transformation that includes such an intercept is called an "affine" transformation. Converting inches to centimeters is a linear transformation, whereas converting temperatures from Fahrenheit to Kelvin is an affine transformation.

4 Plotting and Interpreting Residuals

If you worked through the previous examples above in Prism, you may have noticed that Prism gives you the option of preparing a plot of residuals for each regression.

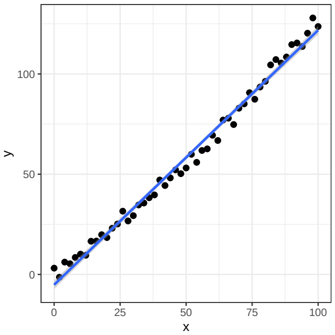
You should always choose this option and inspect the results, as the following will demonstrate.

Here is another dataset that we can fit some sample models to:

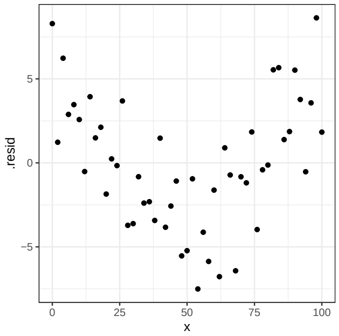


The data can be imported into Prism from the provided supplemental file, `residuals.txt`. If you perform a linear regression on this data in Prism, you should obtain a best-fit line with a slope of 1.270 and an intercept of -5.137. The 95% CI of both parameters of the line are also quite narrow; the CI of the slope is 1.233 − 1.307. The CI of the intercept is -7.283 − -2.990, which is small relative to the range of the y-values over the whole dataset. Additionally, the r^2 value looks very strong at 0.99.

While all of these metrics seem to indicate a good fit, there is a subtle but fundamental problem with the model. It may not be apparent from the default plot that Prism makes of the raw data and the best fit line, but with some manipulation of the scale and careful examination, you can observe that the model tends to systematically underpredict the data at extreme values of x, and systematically overpredict the data at intermediate values.



Subtle patterns like this can often be made more readily apparent by plotting the residuals (i.e., the differences between the observed and predicted values for each point) as a function of the x value.

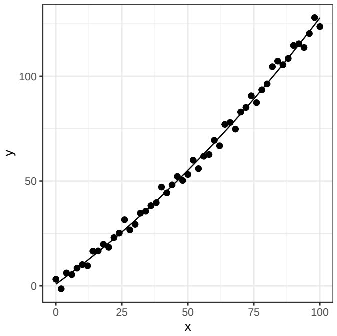


In a good model, the residuals should be normally distributed, and show no systematic patterns. Here was can clearly see that this is not the case; there is something else going on that the model is not capturing. Thus, despite the good numeric metrics, we would reject this simple linear model.

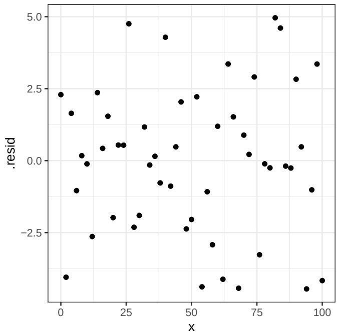
When Prism make plots of residuals, it plots them as a function of the independent variable, x. A more common practice is to plot residuals as a function of the fitted value of y. This is more generalizable, in that you can easily prepare such a plot, even when there are multiple explanatory variables in your model; i.e., when you're fitting to the model $z = f(x, y)$.

5 Non-Linear Regression

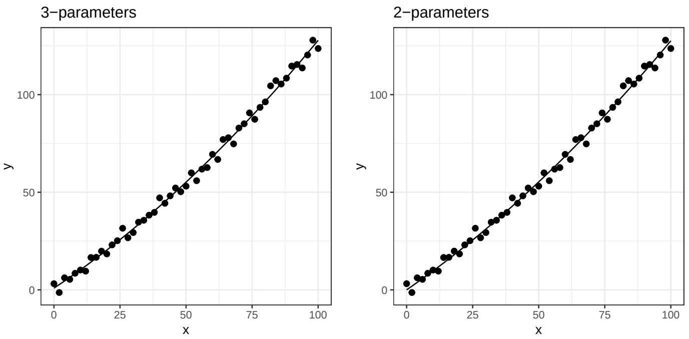
Continuing the example above, if we are perceptive (or lucky), we might try adding a quadratic term to our model. In Prism, choose to analyze the same data, but select Non-Linear Regression, and use a second order polynomial model. Because we want to be able to inspect the distribution of residuals, be sure to ask for such a plot (you'll find the option on the Diagnostics tab). Also, make sure that Prism computes the 95% CIs of the three model parameters. Prism will then fit to the equation $y = B_0 + B_1 \cdot x + B_2 \cdot x^2$.



If you inspect the plot of residuals from this model, you'll observe that there's no longer an obvious systematic pattern to the residuals.



If you're following along in Prism, you might notice that in our revised model, the 95% CI of the intercept term included zero. Since we prefer models with fewer parameters, and it seems that the intercept parameter isn't needed, we might try a third model that doesn't include this parameter: $y = B_1 \cdot x + B_2 \cdot x^2$. In Prism, this can be achieved by fitting to the three-parameter quadratic model, but then fixing $B_0 = 0$ on the Constrain tab. In this case, the two parameter quadratic model predicts a curve that is nearly identical to that predicted by the three parameter model.



A systematic comparison of our alternative models is probably in order at this point:

	Model formula	SSQ[1]
model 1	$y = B_0 + B_1 \cdot x$	733.84
model 2	$y = B_0 + B_1 \cdot x + B_2 \cdot x^2$	321.20
model 3	$y = B_1 \cdot x + B_2 \cdot x^2$	325.73

Here we can see that model 2 (the full, three-parameter model) has the smallest sum-of-squares of residuals. This should not be surprising, since both model 1 and model 3 are special cases of model 2. Since curve fitting boils down to adjusting the model parameters to minimize this SSQ, it is not possible for model 2 to fit worse than model 1 or model 3.

Given all this, which model do you think best represents the data? We've eliminated model 1 due to the systematic variable in residuals. But should we use model 2, which fits the points slightly better than model 3, but at the expense of more model complexity (three parameters instead of two)? We'll look more deeply into this question in our next session.

[1] Use non-linear regression to a straight line to get Prism to report SSQs for linear models.

6 Appendix

6.1 Uncorrelated data.

	x	y
1	0.2321	-0.5180
2	0.4751	1.6298
3	-0.9633	-0.3096
4	0.0728	0.9463
5	0.2049	0.8940
6	0.5956	0.4886
7	-0.4451	-0.7217
8	1.2178	0.1547
9	-2.5927	-0.0001
10	0.3947	0.9978
11	0.3490	1.9515
12	1.0328	-0.5550
13	0.8589	-0.1373
14	-0.9446	-0.9768
15	1.1953	-0.3095
16	0.2907	-0.5595
17	0.0380	-0.1542
18	-0.7352	-0.0600
19	0.3560	1.2042
20	-0.9831	-0.3106

6.2 Linear data.

	x	y
1	0.000	3.796
2	2.000	5.891
3	4.000	10.387
4	6.000	13.038
5	8.000	17.766
6	10.000	23.625
7	0.000	3.111
8	2.000	6.813
9	4.000	11.819
10	6.000	13.998
11	8.000	17.032
12	10.000	23.778
13	0.000	3.648
14	2.000	7.993
15	4.000	9.860
16	6.000	15.343
17	8.000	19.453
18	10.000	21.154

6.3 Residual exercise data.

	x	y
1	0.000	3.153
2	2.000	-1.370
3	4.000	6.172
4	6.000	5.365
5	8.000	8.487
6	10.000	10.139
7	12.000	9.581
8	14.000	16.576
9	16.000	16.664
10	18.000	19.836
11	20.000	18.397
12	22.000	23.032
13	24.000	25.170
14	26.000	31.559
15	28.000	26.692
16	30.000	29.333
17	32.000	34.666
18	34.000	35.634
19	36.000	38.256
20	38.000	39.679
21	40.000	47.117
22	42.000	44.352
23	44.000	48.154
24	46.000	52.181
25	48.000	50.265
26	50.000	53.115
27	52.000	59.932
28	54.000	55.913
29	56.000	61.830
30	58.000	62.631
31	60.000	69.416
32	62.000	66.804
33	64.000	77.011
34	66.000	77.934
35	68.000	74.770
36	70.000	82.906
37	72.000	85.084
38	74.000	90.652
39	76.000	87.382
40	78.000	93.477
41	80.000	96.301
42	82.000	104.508
43	84.000	107.176
44	86.000	105.433
45	88.000	108.449
46	90.000	114.646
47	92.000	115.438
48	94.000	113.674
49	96.000	120.315
50	98.000	127.915
51	100.000	123.651