Chapter 4
Regression

In this chapter some aspects of linear statistical models or regression models will be reviewed. Topics covered will include linear least-squares fits of predictands to predictors, correlation coefficients, multiple regression, and statistical prediction. These are generally techniques for showing linear relationships between variables, or for modeling one variable (the predictand) in terms of others (the predictors). They are useful in exploring data and in fitting data. They are also a good introduction to more sophisticated methods of linear statistical modeling.

4.1 Ordinary linear least-squares regression

4.1.1 Independent variables are known

Suppose we have a collection of N paired data points (x_i, y_i) and that we wish to approximate the relationship between x and y with the expression:

$$\widehat{y} = a + b \cdot x + \epsilon \tag{4.1}$$

where a is called the *y-intercept* and b is the *slope of the line.* In what follows, we assume that x is known with precision, and that we wish to estimate y based on known values of x. The cases where both x and y contain uncertainties will be discussed next. The error, or residual, ϵ can be minimized in a least-squares sense by defining an error function Q in the following way:

$$Q = \frac{1}{N}\sum_{i=1}^{N}\epsilon^2 = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(bx_i + a - y_i)^2 \tag{4.2}$$

where the subscript i denotes i^{th} observation. Q is the sum of the squared differences between the data and our linear fit, and when it is minimized by choosing the parameters a and b we obtain the *least-squares linear fit* to the data.

The fact that the error is squared in the definition of Q has several important consequences.

- Q is positive definite.
- The minimization of Q (the derivative of Q) results in a linear problem to solve.
- Large errors are weighted more heavily than small errors.

The first two are very good consequences. The last can be good or bad depending on what you are trying to do. All the linear regression analysis techniques we will discuss in later chapters (EOF, SVD, PCA, etc.) share these same properties of linear least squares techniques.

We wish to select the constants a and b such that the error or risk functional Q is minimized. This is achieved in the usual way by finding the values of these constants that make the derivatives of Q with respect to them zero. Since the error is always positive and the error function has a parabolic shape, we know that these zeros must correspond to minima of the error function

$$\frac{\partial Q}{\partial a} = 0 \text{ and } \frac{\partial Q}{\partial b} = 0 \qquad \text{“The Normal Equations”} \tag{4.3}$$

It is straightforward to show that solutions to these equations results in the following

$$\frac{\partial Q}{\partial a} = 2aN + 2b\sum_{i=1}^{N} x_i - 2\sum_{i=1}^{N} y_i = 0 \tag{4.4}$$

$$\frac{\partial Q}{\partial b} = 2a\sum_{i=1}^{N} x_i + 2b\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i y_i = 0 \tag{4.5}$$

Dividing both equations by N and moving the y terms to the left-hand-side results in

$$\overline{y} = b\overline{x} + a \tag{4.6}$$

$$\overline{xy} = b\overline{x^2} + a\overline{x} \tag{4.7}$$

where $\overline{(\cdot)}$ denotes the mean across all N observations and $(\cdot)'$ will denote departures from this mean. This system of equations can also be written in matrix form,

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \tag{4.8}$$

which is especially useful when one moves to multi-linear regression with more than one independent variable.

The solutions for the regression coefficients are:

$$a = \overline{y} - b\overline{x} \tag{4.9}$$

$$b = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{4.10}$$

The term $\overline{x'y'}$ is given a special name, the *covariance* of x and y, and is defined as

$$\overline{x'y'} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \tag{4.11}$$

Thus, we see that a_1 is the covariance of x and y normalized by the variance of the independent variable x. Also, note that a, also known as the *y-intercept* is zero if the variables x and y have mean zero.

One can show that the minimum value of the error functional obtained via ordinary least-squares regression is:

$$Q_{min} = \overline{y'^2} - \frac{\overline{x'y'}^2}{\overline{x'^2}} = \overline{y'^2} - b^2\overline{x'^2} \tag{4.12}$$

Thus, we see that the minimum error is the total variance minus the explained variance, which is related to the squared slope (b) and the variance of the predictor.

4.1.2 Independent and dependent variables are uncertain

Quite often the first attempt to quantify a relationship between two experimental variables is linear regression analysis. In many cases one of the variables is a precisely known independent variable, such as time or distance, and the regression minimizes the root mean square (rms) deviation of the dependent variable from the line, assuming that the measurements contain some random error. It often happens that both variables are subject to measurement error or noise, however. In this case, to perform simple linear regression analysis one must choose which variables to define as dependent and independent. The two possible regression lines obtained by regressing y on x or x on y are the same only if the data are exactly collinear.

An alternative to simple regression is to minimize the perpendicular distance of the data points from the line in a two-dimensional space. This approach has a very long history scattered through the literature of many scientific disciplines (Adcock 1878; Pearson 1901; Kermack 1950; York 1966). The method can be elaborated to any degree desired, to take into account the different scales of the two variables in question, their uncertainty, or even the confidence one has in individual measurements (see Section 4.1.4.1.

One of the better, and more elegant, methods of doing linear fits between two variables is EOF/PC analysis, which is discussed in a later chapter of these notes. It turns out that, at least in two dimensions, doing EOF analysis minimizes the perpendicular distance from the regression line and is more elegant than the methods used by Kermack and Haldane (1950) and York (1966). EOF/PC analysis is also easily generalized to many dimensions. See Chapter 5.

4.1.3 Uncertainty estimates of ordinary least-squares regression

We want to fit a straight line to a time series of N observations y_i taken at time x_i. The linear fit is given by

$$y_i = a + bx_i + e_i, \quad i = 1, 2, ..., N \tag{4.13}$$

where e_i represents the residual error of the linear fit at each time x_i. From Chapter 4.1.1, we know that the ordinary least squares solution for parameters a and b are

$$\hat{b} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{4.14}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x} \tag{4.15}$$

and so, the errors of the fit, called the *residuals*, are

$$\hat{e}_i = y_i - (\hat{a} + \hat{b}x_i) = y_i - \hat{y}_i, \quad i = 1, 2, ..., N \tag{4.16}$$

Now, we would like to assign ranges, or confidence limits, on our estimates of a and b. We start with the unbiased estimate of the standard error variance of the residuals:

$$\hat{\sigma}_e^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{e}_i^2 = \frac{N}{N-2}(1 - r_{xy}^2)\overline{y'^2} \tag{4.17}$$

where we divide by $N - 2$ to account for the fact that two degrees of freedom were used to estimate a and b. The expression that includes the correlation coefficient, r_{xy}, follows from the derivations in Chapter 4.2.

For the time being we will assume that all of these residuals are independent of one another, but if instead they are autocorrelated, we could use a model of red noise to estimate the true number of degrees of freedom N^*, and then replace N with N^* (see Chapter 7 for a discussion of degree of freedom estimates for autocorrelated data).

From the standard error variance of the residuals, $\hat{\sigma}_e^2$, we can estimate the standard error variance of the of slope, $\hat{\sigma}_b^2$ in the following way. First,

$$\hat{\sigma}_b^2 = \frac{\hat{\sigma}_e^2}{N\sigma_x^2}, \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{4.18}$$

where we have assumed that the x_i's are precisely known. Putting the pieces together leads to

$$\hat{\sigma}_b^2 = \frac{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{4.19}$$

Looking at this equation, you can see intuitively that it is somewhat like the error in our y estimate divided by the variance in our x values - sort of like a slope of errors or variances.

Since $\frac{\widehat{b}-b}{\widehat{\sigma}_b}$ is distributed like the t-statistic with $N-2$ degrees of freedom, we can put limits on the true slope b in the following way:

$$\widehat{b} - t^{N/2}_{\alpha/2} < b < \widehat{b} + t^{N-2}_{\alpha/2}\widehat{\sigma}_b \tag{4.20}$$

where t^{N-2}_{α} is the critical value of the t-statistic for confidence level α and degrees of freedom $N-2$.

We can apply these techniques to the record of annual mean land temperature from the Goddard Institute of Space Studies (GISS) for the period 1900-2016, as shown in **Fig. 4.1**. Note that the lower limits on the trends are all positive, so we can say that the trends on the intervals are positive at 95% confidence.

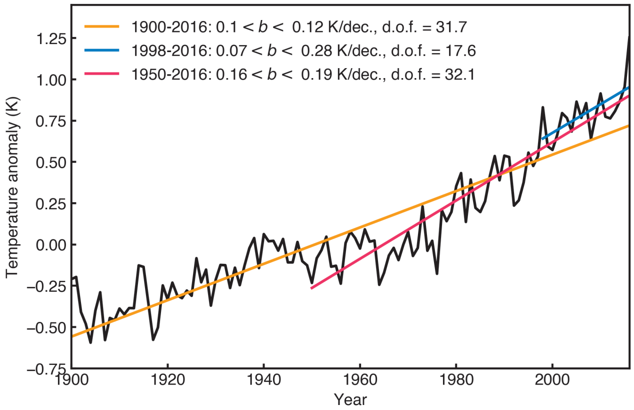

Figure 4.1 The GISS global surface temperature timeseries with linear trends (b) and ± 2.5 % uncertainties for various periods. Estimated degrees of freedom are also given. Units of the trends are Kelvin per decade.

Fig. 4.1 illustrates several aspects of linear fitting to time series. First, the result may depend sensitively on the end points of the analysis. Note that the procedure described in Section 4.1.3 assigns the shorter 1950-2017 period more degrees of freedom than the longer period from 1900-2017. This is because the longer period has an S-shape associated with the period of slow change from 1940-1980. As a result, the residuals from the linear fit for the longer period yield a large autocorrelation. The decades of the 1950's to 1970's are consistently below the line and the decades from 2000-2017 are consistently above the line. The period from 1950-2017 is better fit by a straight line and gives a larger number of degrees of freedom. Despite that the uncertainty is smaller for the longer period because the variance of the predictor is greater. The statistics support the notion that the recent trends are greater than the long-term trend at 95% significance. Starting the trend calculation at 1950 is not objective, since it was chosen by inspecting the time series, but it is true nonetheless that any starting point after 1950 or so yields the same conclusion that the recent warming is faster than the estimate for 1900-2017, unless the record is so short that the uncertainty is too great, as is the case for the 1998-2017 period.

4.1.4 Other least-squares fits

4.1.4.1 Orthogonal-least squares regression

Looking back at our derivations for ordinary least squares, it becomes apparent that the results are not symmetric for x and y. That is, it matters which variable you call x (the independent variable) and which you call y (the dependent variable). This can be seen for the example data provided in **Fig. 4.2**. If you cannot adequately justify which data should be x and which should be y, or it does not make sense to even try, *orthogonal least squares* may instead be what you want.

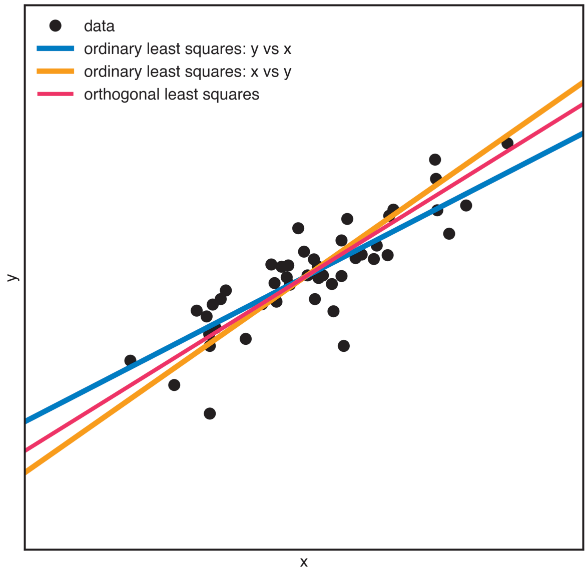

Figure 4.2 Three least-squares best fit lines calculated in different ways: (1) ordinary least squares where x is the independent variable and y is the dependent variable, (2) ordinary least squares where y is the independent variable and x is the dependent variable, and (3) using orthogonal least squares.

While ordinary linear least squares minimizes the vertical errors between the data and the best fit line (i.e. the error in y), orthogonal linear least squares minimizes the orthogonal errors, as shown in **Fig. 4.3**. It so happens that EOF analysis (to be discussed in Chapter 5) in two-dimensions provides the orthogonal least squares fit.

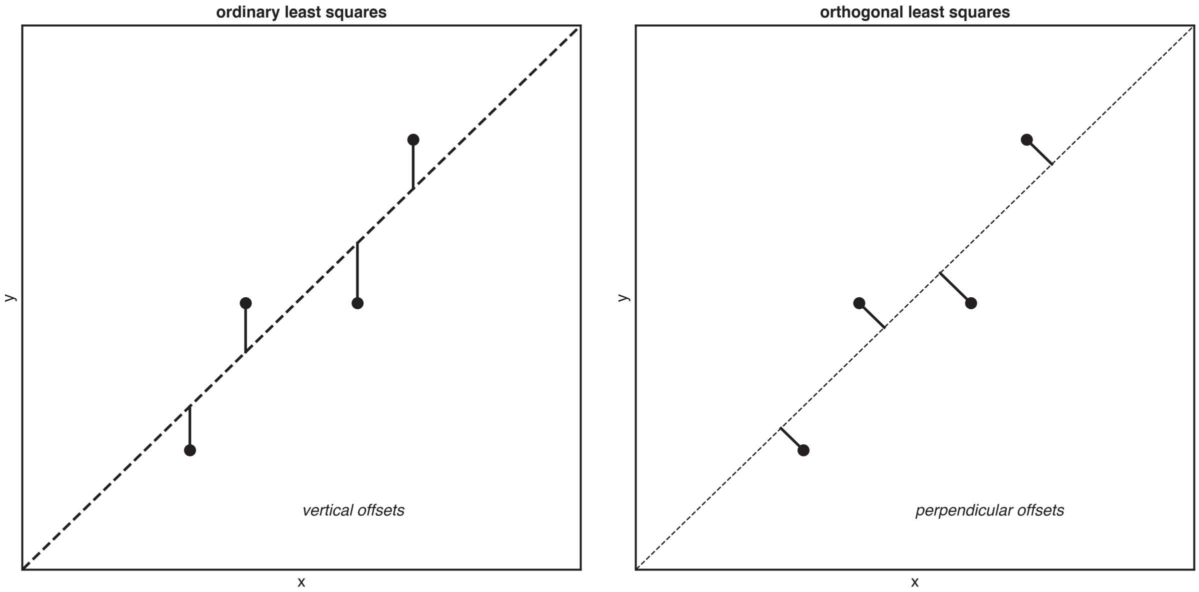

Figure 4.3 Depiction of (left) ordinary least squares regression which defines the errors based on vertical offsets and (right) orthogonal least squares regression which defines the error based on perpendicular offsets.

4.1.4.2 Power laws and polynomials

Many other curves besides a straight line can be fit to data using a similar procedure to that outlined for ordinary least-squares regression. Some common examples are power laws and polynomials such as

$$y = ax^{b} \Rightarrow \ln y = \ln a + b \ln x \tag{4.21}$$

$$y = ae^{bx} \Rightarrow \ln y = \ln a + bx \tag{4.22}$$

$$y = a_0 + a_1 x + a_x^2 + a_3 x^3 + ... + a_n x^n \tag{4.23}$$

In some cases, like that of power laws (e.g. **(4.21)**, **(4.22)**), one can use logarithms to turn the problem into a linear one, in which case, standard linear least squares methods can be used to estimate the parameters.

4.2 Correlation

4.2.1 How good is the linear fit?

How much we believe the computed regression coefficient $(\widehat{b})$ depends on the spread of the dots about the best fit line. If the dots are closely packed about the regression line, then the fit is good. The spread of the dots is given by the *correlation coefficient* r.

Here is one way to derive the correlation coefficient. By definition, the total variance of $y(t)$ is

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{4.24}$$

and by definition, the total variance of the fit of $x(t)$ to $y(t)$ (i.e. the variance of $\widehat{y}$) is

$$\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - \overline{\widehat{y}})^2 = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2 \tag{4.25}$$

where we have used the fact that

$$\overline{\widehat{y}} = a + b\overline{x} = \overline{y} \tag{4.26}$$

The percent of the total variance in y explained by the fit $\widehat{y}$ is thus given by the ratio

$$\begin{aligned}
r^2 &= \frac{\text{explained variance}}{\text{total variance}} && (4.27)\\
&= \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} && (4.28)\\
&= \frac{\sum_{i=1}^{N}(bx_i + a - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} && (4.29)\\
&= \frac{\sum_{i=1}^{N}(bx_i + \overline{y} - b\overline{x} - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} && (4.30)\\
&= \frac{\sum_{i=1}^{N}(bx_i')^2}{\sum_{i=1}^{N}(y_i'^2)} && (4.31)\\
&= \frac{(\frac{\overline{x'y'}}{\overline{x'^2}})^2\sum_{i=1}^{N}(x_i')^2}{\sum_{i=1}^{N}(y_i'^2)} && (4.32)\\
&= \frac{(\overline{x'y'})^2\sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2\sum_{i=1}^{N}(y_i'^2)} && (4.33)\\
&= \frac{(\overline{x'y'})^2\frac{1}{N}\sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2\frac{1}{N}\sum_{i=1}^{N}(y_i'^2)} && (4.34)\\
&= \frac{(\overline{x'y'})^2\cdot\overline{x'^2}}{(\overline{x'^2})^2\cdot\overline{y'^2}} && (4.35)\\
&= \frac{(\overline{x'y'})^2}{\overline{x'^2}\cdot\overline{y'^2}} && (4.36)
\end{aligned}$$

Hence,

$$r = \frac{\overline{x'y'}}{\widehat{\sigma_x}\widehat{\sigma_y}} \tag{4.37}$$

Some important points about the correlation coefficient r:

- r^2 is the fraction of variance explained by the linear least-squares fit between the two variables
- r varies between -1 and 1 and r^2 varies between 0 and 1

Note that if $\sigma_x = \sigma_y = 1$ and $\overline{x} = \overline{y}$ (that is, both x and y are standardized), then the correlation r is equal to the regression coefficient $\widehat{b}$. More generally, there is a strong relationship between the regression line and the correlation coefficient:

$$\widehat{b} = r\frac{\sigma_y}{\sigma_x} \tag{4.38}$$

Thus, the regression coefficient can be thought of as the correlation coefficient multiplied by the ratio of the standard deviations of y and x.

Worked Example 4.1.

Suppose that the correlation coefficient between sunspots and five-year mean global temperature is 0.5 ($r = 0.5$). Then the fraction of the variance of 5-year mean global temperature that is linearly explained by sunspots is $r^2 = 0.25$. That is, the fraction of unexplained variance is still 75%. The *root-mean-square error* (RMS error), normalized by the total variance is thus:

$$\left(\frac{\text{MS Error}}{\text{Total Variance}}\right)^{1/2} = \sqrt{1 - r^2} - \sqrt{0.75} = 0.87 \tag{4.39}$$

Thus, only a 13% reduction in RMS error results from a correlation coefficient of 0.5. The implications of this are further illustrated in the following table:

r	r^2	**RMS error**
0.98	.960	20.0%
0.9	.81	43.6%
0.8	.64	60.0%
0.5	.25	86.6%
0.3	.09	95.4%
0.1	.01	99.5%

In Practice.

- As **Worked Example 4.1** illustrates, statistically significant correlations are not necessarily useful for forecasting. If you have enough data you may be able to show that a measured $r = 0.3$ correlation coefficient reflects that the true correlation coefficient is different from zero at the 99% confidence level, but such a correlation, however real, is often useless for forecasting. The RMS error would be 96% of the variance. The exception to this statement about the uselessness of small correlations comes where you have a very large number of trials or chances. If you have a large volume of business (billions of dollars) spread over a large number of transactions and you shade your trades properly using the 0.3 correlation prediction, then you can actually make a lot of money...sometimes.

In Practice.

- The correlation will only show the linear relationships clearly. Nonlinear relationships may exist for which the correlation coefficient will be zero. For example, if the true relationship is parabolic, and the data are evenly sampled, the correlation coefficient would be close to zero, even though an exact parabolic relationship may exist between the two data sets.
- The correlation cannot reveal quadrature relationships (although lagged correlations often will). For example, meridional wind and geopotential are approximately uncorrelated along latitudes even though the winds are approximately geostrophic and easily approximated from the geopotential. They are in quadrature (90 degrees out of phase).
- The statistical tests (to be described next) apply to independent data. Often the sample data are not independent. The actual number of degrees of freedom may be much smaller than the sample size.
- Watch out for nonsense correlations that may occur even though the two variables have no direct relation to each other. The correlations may occur by chance or because the two variables are each related to some third variable. For example, over the past 50 years the number of books published and professional baseball games played have both increased, so that they are positively correlated. Does this mean that, if there is a players' strike, book publishing will take a nose dive?
- **Fig. 4.4** illustrates some of the problems that can arise when using linear regression and correlation coefficients to describe relationships between two data sets. This set of four examples is famously known as *Anscombe's Quartet*, as each panel has exactly the same correlation coefficient of $r = 0.82$.

4.2.2 Sampling Theory of Correlation (Pearson's correlation)

4.2.2.1 Statistical significance of correlations

The correlation, r, between two time series, $x(t)$ and $y(t)$, gives a measure of how well the two time series vary linearly with one another (or do not). But how can you tell whether the correlation you calculate is significantly different from zero? In this section we will review the techniques for testing the statistical significance of correlation coefficients.

Suppose we have N pairs of values (x_i, y_i) from which we have calculated a sample correlation coefficient r. The theoretical true value is denoted by ρ. For now, we will assume that we are sampling x and y from Normal distributions.

When the true correlation coefficient is zero, that is, when $\rho = 0$, the distribution of r is symmetric about zero and we are able to make use of the z- and t-statistic. Namely, the random variable t

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{4.40}$$

will follow the t distribution with degrees of freedom $\nu = N - 2$.

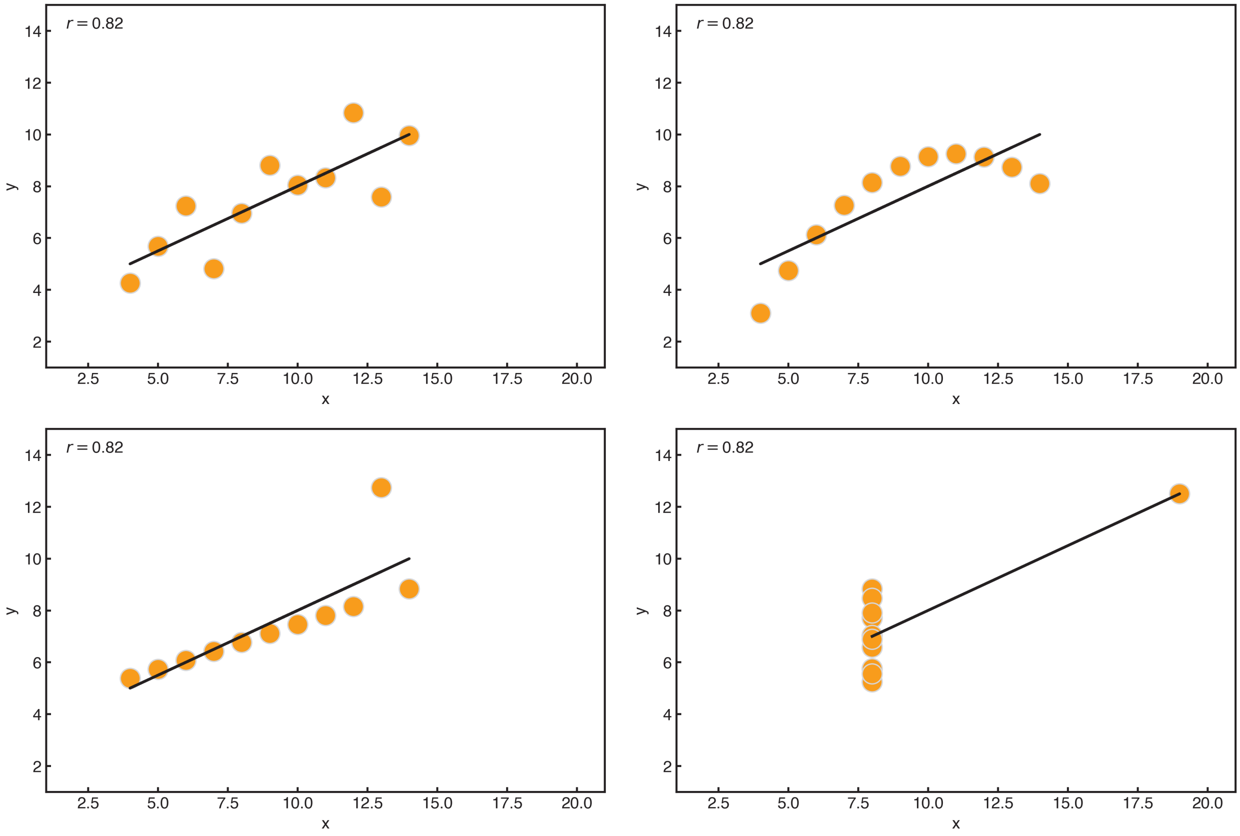

Figure 4.4 Four sets of data, known as *Anscombe's Quartet* where the correlations are all $r = 0.82$.

Worked Example 4.2.
We have two time series, each of length $N = 20$, and they are correlated at $r = 0.6$. Does this correlation exceed the 95% confidence interval under the null hypothesis that $\rho = 0$? You can assume both time series are sampled from underlying normal distributions and that the 20 observations in each data set represent 20 degrees of freedom.

....

We had no prior knowledge (before getting the samples) that the correlation would be positive or negative, so we will use a two-tailed t-test.

$t_c = 2.1$ for $\nu = N - 2 = 18$, so we want to know if the sample statistic $t > 2.1$.

$$t = \frac{0.6\sqrt{20-2}}{\sqrt{1-.6^2}} = 3.18. \tag{4.41}$$

Since $t = 3.18 > 2.1$, we can reject the null hypothesis that the true correlation is zero at 95% confidence.

In Practice.

- It turns out that the t-statistic is only applicable for $\rho = 0$ if the underlying distributions of the data are both normal, or if N is big enough that the central limit theorem applies. For well behaved distributions, a good rule of thumb is that an $N > 20$ should be sufficient for the central limit theorem to apply and the t-statistic to be appropriate for testing the null hypothesis that $\rho = 0$. Examples of the t values obtained from a range of distributions is given in **Fig. 4.5**.

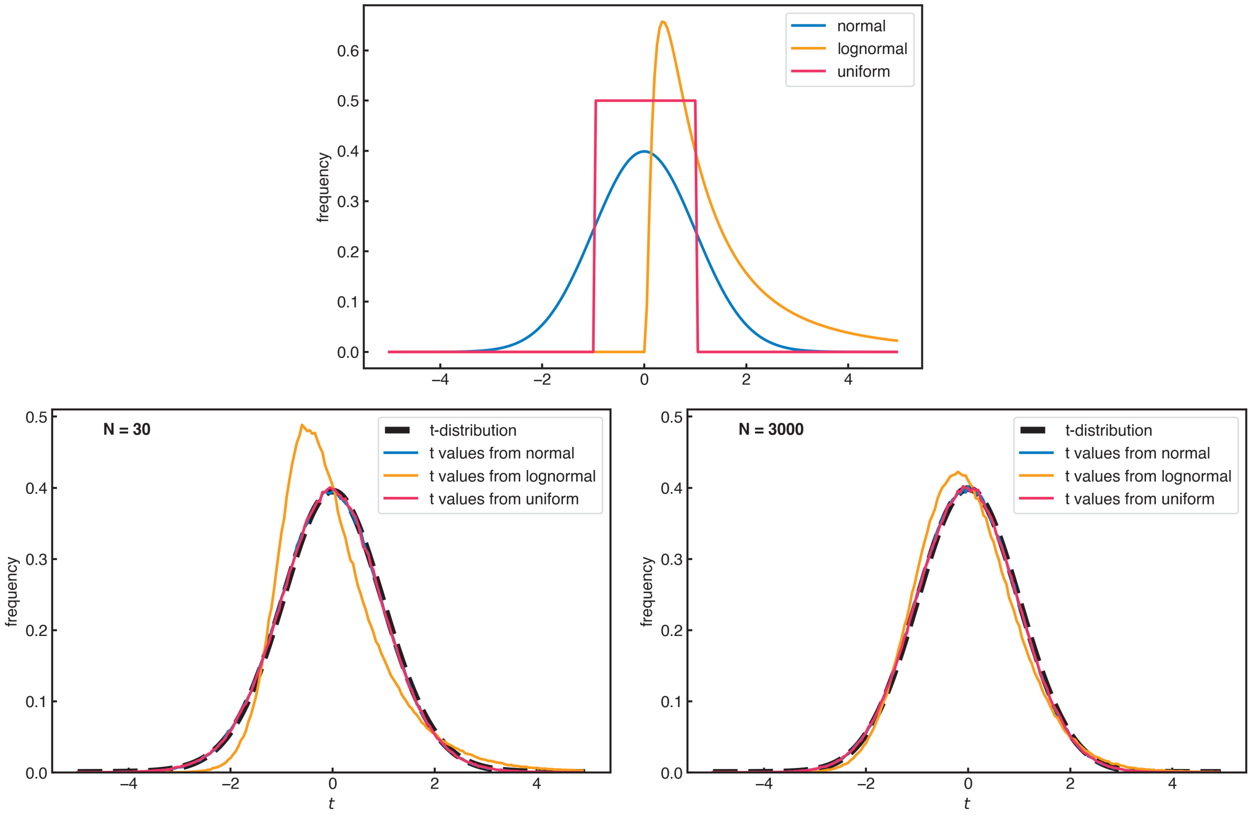

Figure 4.5 Three underlying sampling distributions and the resulting distribution of t values **(4.40)** computed from correlations obtained using $N = 30$ and $N = 3000$. The theoretical t-distribution is denoted by the black dashed line.

When the true correlation coefficient is not expected to be zero (i.e. $\rho \neq 0$), we cannot assume that the sampled correlations r will come from a symmetric, normal distribution. Instead, the distribution will be skewed due to the fact that correlations cannot exceed -1 or 1. In this instance, we must use the *Fisher-Z Transformation* to convert the distribution of r into something that is normally distributed (Z).

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{4.42}$$

The Fisher-Z statistic is then normally distributed with the following mean and standard deviation:

$$\mu_Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \quad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{4.43}$$

Thus, the confidence bounds for Z become

$$Z - t_c \sigma_Z \leqslant \mu_Z \leqslant Z + t_c \sigma_Z \tag{4.44}$$

If you have μ_Z and want the corresponding actual correlation ρ, you can use the following handy transformation

$$\rho = \frac{e^{2\mu_Z} - 1}{e^{2\mu_Z} + 1} = \frac{e^{\mu_Z} - e^{-\mu_Z}}{e^{\mu_Z} + e^{-\mu_Z}} = \tanh(\mu_Z) \tag{4.45}$$

Worked Example 4.3.
What are the 95% confidence limits on the true correlation ρ if you draw 21 samples from a normal distribution and obtain and $r = 0.8$?
....

Since we want the confidence bounds, we need to employ the Fisher-Z transformation in **(4.42)**

$$Z = \frac{1}{2}\ln\left(\frac{1+0.8}{1-0.8}\right) = 1.0986 \tag{4.46}$$

$$\sigma_Z = \frac{1}{\sqrt{21-3}} = .235 \tag{4.47}$$

Calculating $t_{0.025} = 2.1$ (using $\nu = 21 - 3$) leads to:

$$Z - 2.1\sigma_Z \leqslant \mu_Z \leqslant Z + 2.1\sigma_Z \tag{4.48}$$

$$0.61 \leqslant \mu_Z \leqslant 1.59 \tag{4.49}$$

While interesting, knowing μ_Z is not very helpful unless we convert it back to a correlation. So, plugging the bounds into **(4.45)** leads to

$$0.54 \leqslant \rho \leqslant 0.92 \tag{4.50}$$

Tests for the significance of the difference between two non-zero correlation coefficients are made by applying the Z statistic using the fact that it is normally distributed. For example, suppose we have two samples, one of size N_1 and one of size N_2, and each produce a correlation coefficient of r_1 and r_2, respectively. We test for a significant difference between these correlations by first calculating the Fisher-Z transformations for each:

$$Z_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right); \quad Z_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right) \tag{4.51}$$

From these we can calculate the typical *z*-score from

$$z = \frac{Z_1 - Z_2 - \Delta_{1,2}}{\sigma_{1,2}} \tag{4.52}$$

where

$$\Delta_{1,2} = \mu_1 - \mu_2 \tag{4.53}$$

is the transformed difference you expect (your null hypothesis). If you null hypothesis is that the true correlations of the two samples are equal (i.e. $\rho_1 = \rho_2$), then $\Delta_{1,2} = 0$. The denominator in **(4.52)** is given by

$$\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{4.54}$$

4.2.2.2 Spearman's rank correlation

Spearman's rank correlation is a nonparametric test that determines whether a set of paired data monotonically vary together, but it is not concerned with the actual amplitude of the variations, just the ranks of the values. Since this is a nonparametric test, no assumption about normality needs to be made.

The idea is very simple, the original paired data x_i and y_i get converted into ranks (position in a sorted list) X_i and Y_i and Spearman's rank correlation ρ_R is given by

$$\rho_R = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2 (Y_i - \overline{Y_i})^2}} \tag{4.55}$$

When computing the ranks, if there are duplicate values the ranks are equal to the average rank/position.

The standard error of Spearman's rank correlation ρ is given by

$$\sigma_\rho = \frac{0.6325}{(N-1)^{1/2}} \tag{4.56}$$

For significance testing on ρ_R, one can use the Fisher-Z test or the t-test (when the null hypothesis is that $\rho_R = 0$) as for the standard Pearson correlation.

There are many other nonparametric methods for calculating correlations, for example, Kendall's Tau Rank Correlation. We will not delve into these here.

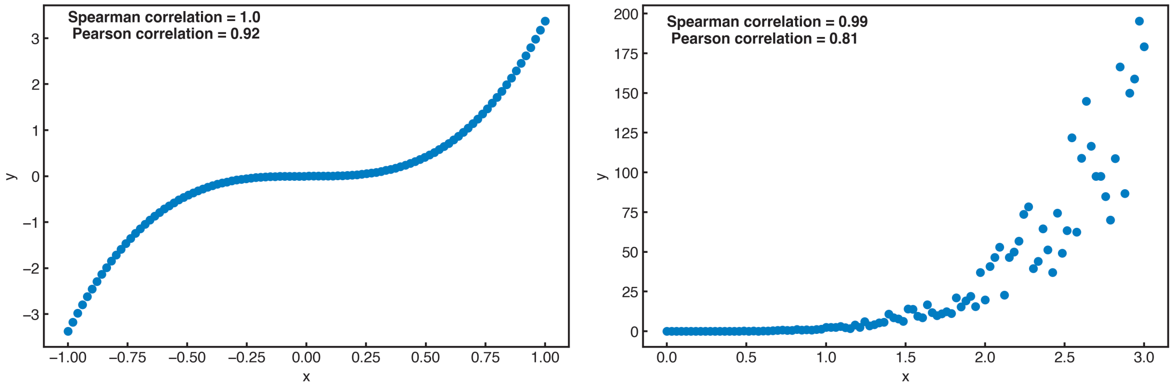

Figure 4.6 Paired data and their Spearman and Pearson correlations.

Worked Example 4.4.

Given paired data x and y, one can calculate Spearman's rank correlation using **(4.55)**. For the paired data in the table below:

Spearman rank correlation: 0.99
Pearson correlation: 0.81

x	y	rank X	rank Y
1.04	1.39	2	1
1.46	6.78	6	5
1.03	2.21	1	2
1.66	13.46	7	8
1.29	6.3	4	4
1.70	11.31	8.5	6
1.27	4.37	3	3
1.70	20.42	8.5	9
1.97	22.22	10	10
1.43	11.81	5	7

4.3 Multiple Linear Regression

4.3.1 Generalized Normal Equations

Multiple regression is the regression of more than two variables. The basic idea is that you wish to use multiple predictors x_i to predict your predictand y. That is, you wish to find the regression coefficients a_i that provide the best guess $\widehat{y}$ for your predictand y

$$\widehat{y} = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n \tag{4.57}$$

For a single predictor x, we wanted to minimize the cost function Q, defined as:

$$Q = \sum_i^N (\widehat{y}_i - y_i)^2 = \sum_i^N (a_1x_i + a_0 - y_i)^2 \tag{4.58}$$

For the multiple predictor case (predictors $x_1, x_2, x_3..., x_n$), we want to minimize

$$Q = \sum_i^N (\widehat{y}_i - y_i)^2 = \sum_i^N (a_0 + a_1x_{1,i} + a_2x_{2,i} + a_3x_{3,i} + ... + a_nx_{n,i} - y_i)^2 \tag{4.59}$$

where n is the number of predictors and N is the number of time steps. Thus, $x_{2,i}$ denotes the predictor x_2 at time step i.

For n predictors, we have $n+1$ equations derived by setting

$$\frac{\partial Q}{\partial a_i} = 0 \tag{4.60}$$

where i goes from 0 to n.

$$\begin{aligned}
\overline{y} &= a_0 + a_1\overline{x_1} + a_2\overline{x_2} + \dots + a_n\overline{x_n} && (4.61)\\
\overline{x_1y} &= a_0\overline{x_1} + a_1\overline{x_1^2} + a_2\overline{x_1x_2} + \dots + a_n\overline{x_1x_n} && (4.62)\\
\overline{x_2y} &= a_0\overline{x_2} + a_1\overline{x_2x_1} + a_2\overline{x_2^2} + \dots + a_n\overline{x_2x_n} && (4.63)\\
&\dots && (4.64)\\
\overline{x_ny} &= a_0\overline{x_n} + a_1\overline{x_nx_1} + a_2\overline{x_nx_2} + \dots + a_n\overline{x_n^2} && (4.65)
\end{aligned}$$

If we assume the mean has been removed from every variable, these simplify to n equations and n unknowns (since we now know that $a_0 = 0$ and so **(4.61)** is no longer useful).

For the jth equation:

$$\overline{x_jy} = \sum_{i=1}^{n} a_i\overline{x_jx_i} \qquad (4.66)$$

One can write this in matrix form as:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \dots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \dots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \dots \end{bmatrix} \qquad (4.67)$$

Since we have removed the means of all variables, the overbarred quantities are actually covariances. These covariances are closely related to the variance calculations from Chapter 2. If x and y are scalars, then the covariance C_{xy} is

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}), \qquad (4.68)$$

so if $\overline{x} = \overline{y} = 0$ then $\overline{xy} = C_{xy}$. The correlation between x and y is computed by dividing the covariance by the standard deviations of both variables:

$$r_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y} \qquad (4.69)$$

All of these manipulations can be done much more neatly in vector/matrix notation, and the extension to the case where y is a vector is straightforward in that context (see next section).

Using our knowledge of the covariance, we now see that if the means of our variables have been subtracted, the left-hand-side of **(4.67)** is a matrix of the covariances (termed the *covariance matrix* across all of the x_i's and the right-hand-side is a vector of the covariances between x_i and y. In this case, each horizontal line in **(4.67)** can be written in matrix notation as

$$\mathbf{C_{x_ix_j}}a_j = \mathbf{C_{x_iy}} \qquad (4.70)$$

Since the ultimate goal is to determine the a_j coefficients, one can solve for this vector by inverting the real, symmetric matrix on the left, and multiplying the inverse times the vector on the right, at least in theory.

$$\begin{aligned}
\mathbf{C_{x_ix_j}}^{-1}\mathbf{C_{x_ix_j}}a_j &= \mathbf{C_{x_ix_j}}^{-1}C_{x_iy} && (4.71)\\
a_j &= \mathbf{C_{x_ix_j}}^{-1}C_{x_iy} && (4.72)
\end{aligned}$$

However, many of the methods for computing the inverse of the covariance matrix require that $\mathbf{C_{x_ix_j}}$ j be invertible and not singular. In the following chapters we will discuss how singular value decomposition can be used to derive a very robust solution for the a_j's that is optimal even when the problem is over-determined and $\mathbf{C_{x_ix_j}}$ is singular.

In Practice.

- if each variables has been standardized (mean of 0 and standard deviation of 1), the left-hand-side of **(4.67)** is the *correlation matrix* of the x_j's, and the right-hand-side is the *correlation vector* between the x_j's and y.
- if the x_j's are time series at different locations in a data set, the covariance matrix yields information about the structures of the dominate data, and tells you something about the spatial variability of the different points
- if the predictors are linearly independent, the off diagonal elements are all 0 and the a_j's can be found algebraically

4.3.2 Derivation of the Normal Equations using Matrix Notation

Matrix notation is very powerful and compact for doing complex minimization problems and we will need to use it a lot to do more powerful methods later. As an example, then, let's derive **(4.67)** using matrix algebra. First some definitions.

Let's think of $\mathbf{y}$ and $\mathbf{a}$ as row vectors of length N and n, respectively, and the data matrix $\mathbf{X}$ as an $N \times n$ matrix, where N is the sample size and n is the number of predictors, x_j, as before.

$$\mathbf{y} = \left[y_1 \; y_2 \; y_3 \; \dots \; y_N\right] \tag{4.73}$$

$$\mathbf{a} = \left[a_1 \; a_2 \; a_3 \; \dots \; a_n\right] \tag{4.74}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{N1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{N2} \\ x_{13} & x_{23} & x_{33} & \dots & x_{N3} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{Nn} \end{bmatrix} \tag{4.75}$$

Now we can express our desired regression equation in a very compact form

$$\widehat{\mathbf{y}} = \mathbf{aX} \tag{4.76}$$

where we get the vector of predicted values of y, $\widehat{\mathbf{y}}$, by multiplying the vector of coefficients $\mathbf{a}$ by the data matrix $\mathbf{X}$.

Our goal is to determine values for the vector of coefficients $\mathbf{a}$, and we do this by minimizing the squared error of our fit (i.e. the cost function Q). In matrix notation, we compute Q by taking the inner product of the error vector with itself

$$Q = (\mathbf{y} - \mathbf{aX})(\mathbf{y} - \mathbf{aX})^T \tag{4.77}$$

Here, $(\cdot)^T$ indicates the transpose of a matrix, and we will utilize the fact that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. Expanding the right-hand-side of **(4.77)** leads to

$$Q = \mathbf{yy^T} - \mathbf{yX^Ta^T} - \mathbf{aXy^T} + \mathbf{aXX^Ta^T} \tag{4.78}$$

The next step is to differentiate Q with respect to the coefficients a_j to obtain an equation for the values of $\mathbf{a}$ that minimize the error. Doing this leads to

$$\frac{\partial Q}{\partial \mathbf{a}} = \mathbf{0} - \mathbf{yX^T} - \mathbf{Xy^T} + \mathbf{XX^Ta^T} + \mathbf{aXX^T} \tag{4.79}$$

$$= \left(\mathbf{aXX^T} - \mathbf{yX^T}\right) + \left(\mathbf{aXX^T} - \mathbf{yX^T}\right)^T \tag{4.80}$$

Note that the right hand side of (4.80) can be organized into two terms that are the transposes of each other. If a quantity is zero, then its transpose is also zero. Therefore, we can use either of the two forms above to express the minimization. We will carry along both forms in the next couple of equations, although they mean the same thing.

We obtain the optimal solution for the a_j's that minimizes the error, Q, by setting the right hand side of (4.80) equal to zero, or,

$$\mathbf{aXX^T} = \mathbf{yX^T} \text{ or } \mathbf{XX^Ta^T} = \mathbf{Xy^T} \tag{4.81}$$

from which,

$$\mathbf{a} = \mathbf{yX^T}\left(\mathbf{XX^T}\right)^{-1} \text{ or } \mathbf{a^T} = \left(\mathbf{XX^T}\right)^{-1}\mathbf{Xy^T} \tag{4.82}$$

Looking back at (4.72), we see that it is equivalent to $\mathbf{a^T} = \left(\mathbf{XX^T}\right)^{-1}\mathbf{Xy^T}$ since

$$\mathbf{XX^T} = N\mathbf{C}_{\mathbf{x_ix_j}} \text{ and } \mathbf{Xy^T} = N\mathbf{C}_{\mathbf{x_iy}} \tag{4.83}$$

Worked Example 4.5.

You may consider Fourier harmonic analysis to be a special case of a multiple linear least-squares regression model. In this case, the predictors are sines and cosines in sampling dimension z of length N. For example:

$$x_1 = \sin\frac{2\pi z}{L}; x_2 = \cos\frac{2\pi z}{L}; x_3 = \sin\frac{4\pi z}{L}; x_4 = \cos\frac{4\pi z}{L}; ... \tag{4.84}$$

If you are unfamiliar with Fourier analysis, you may want to come back to this section after studying the description of Fourier analysis in Chapter 7.

If we take a multiple linear regression approach, this technique will work for unevenly spaced z_i, whereas standard Fourier Transform techniques will not. For evenly spaced data (evenly spaced z_i) and orthogonal predictors, as is the case for these sines and cosines,

$$a_j = \frac{\overline{x_j y}}{\overline{x_j^2}}; \text{ but } \overline{x_j^2} = \frac{1}{2} \text{ for all } N > 0 \tag{4.85}$$

so that

$$a_j = \frac{2}{N}\sum_{i=1}^{N} y_i \cdot x_j(z_i) \tag{4.86}$$

for example (4.87)

$$a_1 = \frac{2}{N}\sum_{i=1}^{N} y_i \cdot \sin\left(\frac{2\pi z_i}{L}\right) \tag{4.88}$$

and these coefficients are equivalent to what is used in Fourier decomposition, demonstrating that Fourier analysis is optimal in a least-squares sense.

4.3.3 Multiple Regression - How many predictors should I use?

Multiple regression allows for one to use nearly an infinite number of predictors to predict y. However, the question is then *"how many predictors should I use?"*. To make things a bit easier, in this section we will consider standardized variables, although one should keep in mind that all equations can be rewritten without this assumption.

In the case of standardized variables, the normal equations for multiple linear-least-squares regression can be written in the following way

$$r_{x_i x_j} a_i = r_{x_j y} \tag{4.89}$$

where once again r represents the correlation. We start with the simplest case of only two predictors

$$\hat{y} = a_1 x_1 + a_2 x_2 \tag{4.90}$$

Then the normal equations can be expanded as

$$r_{x_1 x_1} a_1 + r_{x_1 x_2} a_2 = r_{x_1 y} \tag{4.91}$$

$$r_{x_2 x_1} a_1 + r_{x_2 x_2} a_2 = r_{x_2 y} \tag{4.92}$$

or in matrix notation,

$$\begin{bmatrix} r_{x_1 x_1} & r_{x_1 x_2} \\ r_{x_2 x_1} & r_{x_2 x_2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{x_1 y} \\ r_{x_2 y} \end{bmatrix} \tag{4.93}$$

But, since $r_{x_1 x_1} = r_{x_2 x_2} = 1$ and $r_{x_1 x_2} = r_{x_2 x_1}$, this can be rewritten as

$$\begin{bmatrix} 1 & r_{x_1 x_2} \\ r_{x_1 x_2} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{x_1 y} \\ r_{x_2 y} \end{bmatrix} \tag{4.94}$$

We solve for a_1 and a_2 and find that

$$a_1 = \frac{r_{x_1,y} - r_{x_1,x_2} r_{x_2,y}}{1 - r^2_{x_1,2}} \tag{4.95}$$

$$a_2 = \frac{r_{x_2,y} - r_{x_1,x_2} r_{x_1,y}}{1 - r^2_{x_1,2}} \tag{4.96}$$

If $\hat{y}$ is the best-fit, then we can write the explained and unexplained variance as

$$\overline{y^2} = \overline{(y_i - \hat{y})^2} + \overline{(\hat{y} - \bar{y})^2} \tag{4.97}$$

$$\text{Total Variance} = \text{Unexplained Variance} + \text{Explained Variance}$$

Using the fact that $\hat{y} = a_1 x_1 + a_2 x_2$ it can be shown that

$$1 = \frac{\overline{(y_i - \hat{y})^2}}{\overline{y^2}} + R^2 \tag{4.98}$$

where the fraction of explained variance R^2 is given by

$$R^2 = \frac{r^2_{x_1,y} + r^2_{x_2,y} - 2 r_{x_1,y} r_{x_2,y} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \tag{4.99}$$

In analogy with the case of simple regression, R can be defined as the multiple correlation coefficient, since its square is the fraction of explained variance.

It turns out that in multiple regression, if too many predictors are used, then the predictions associated with the regression will perform badly on independent data—worse than if fewer predictors were used in the first place. This is because using too many predictors can result in large coefficients for variables that are not actually highly correlated with the predictand. These coefficients help to fit the dependent data, but make the application to independent data unstable and potentially wildly in error. That is because you start to fit the noise, and when the noise changes the prediction is really bad. Also, sometimes these variables are better correlated with each other than they are with the predictand, which will also produce unstable predictions when used with independent data. In this case the covariance matrix you formally invert (i.e. **(4.72)**) is nearly singular.

Adding x_2 as a predictor does not always increase the explained variance. No benefit is derived from additional predictors, unless their correlation coefficient with the predictand exceeds the *minimum useful correlation* - the critical correlation required for a beneficial effect increases with the number of predictors used. Unless predictors can be found that are well correlated with the predictand and relatively uncorrelated with the other predictors, the optimum number of predictors will usually be small. The minimal useful correlation is defined as the minimum correlation between x_2 with y that will allow the addition of x_2 to improve the regression R^2. In math, this is,

$$|r_{x_2y}|_{\text{min useful}} > |r_{x_1y}r_{x_1x_2}| \tag{4.100}$$

We can show this by substituting $r_{x_2y} = r_{x_2y_{\text{min useful}}} = r_{x_1y}r_{x_1x_2}$ into the expression for the explained fraction of the variance in the two-predictor case:

$$\begin{aligned} R^2 &= \frac{r^2_{x_1,y} + r^2_{x_2,y} - 2r_{x_1,y}r_{x_2,y}r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} && (4.101) \\ &= \frac{r^2_{x_1,y} + r^2_{x_2,y} - 2r^2_{x_1,y}r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} && (4.102) \\ &= r_{x_1y} && (4.103) \end{aligned}$$

Thus, we have shown that when r_{x_2y} equals the minimum useful correlation, including the second predictor has no influence on the explained variance. What is not obvious at this point is that including such a useless predictor can actually have a detrimental effect on the performance of the prediction equation when applied to independent data, data that were not used in the original regressions. Note that the lower the value of $r_{x_1x_2}$, that is, the more independent the predictors, the better chance that both predictors will be useful, assuming that they are both correlated with the predictand. Ideally we would like completely independent predictors, i.e. $r_{x_1x_2} = 0$. Completely dependent predictors, $r_{x_1x_2} = 1$, are useless since only one of them is enough (although you can usually reduce the noise by adding them together with some judicious weighting). The desire for independent predictors is part of the motivation for empirical orthogonal functions (EOFs), which will be described in Chapter 5.

Similar, but more complicated considerations apply when deciding to use a third predictor. In general, the more predictors used, the fewer degrees of freedom are inherent in the coefficients a_j, the lower the statistical significance of the "fit" to the data points, and the less likely that the regression equations will work equally well on independent data. If predictors are added indiscriminately, you come to a point where adding predictors makes the regression work less well on independent data, even though you are accounting for more of the variance of the dependent data set. This is because you can *over fit* the data, in essence, you will use the predictors to fit the noise rather than the signal. It is a good idea to use as few predictors as possible, while still getting most of the skill you can. Later we will describe how to pick the optimal set of predictors.

Worked Example 4.6.
Say you have two predictors, x_1 and x_2 and both are correlated with the predictand y at 0.5, and are correlated with each other at 0.5, that is

$$r_{1,y} = r_{2,y} = r_{1,2} = 0.5 \tag{4.104}$$

Does adding x_2 increase your R^2 compared to a regression with x_1 alone?
....

For the first predictor only, the variance explained is

$$R_1^2 = r_{1,y}^2 = 0.25 \tag{4.105}$$

Adding a second predictor, x_2, leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.5^2 - 2 \times 0.5 \times 0.5 \times 0.5}{1 - 0.5^2} = 0.33 \tag{4.106}$$

Thus, adding a second predictor helps explain more of the variance of y.

Now let us assume that $r_{2,y} = 0.25$ and everything else remains the same. Adding the second predictor leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.25^2 - 2 \times 0.5 \times 0.25 \times 0.5}{1 - 0.5^2} = 0.25 \tag{4.107}$$

In this case, adding the second predictor does not increase the explained variance!

4.3.3.1 Adjusted R^2

In most situations, increasing the number of predictors will always increase the explained variance (R^2) because at some point the predictors will start fitting the noise rather than the signal. This will become evident when the regression model is applied to independent data - as the model will be worse than a model with fewer predictors.

How do you know when you are starting to fit the noise and thus should stop adding predictors? One tool for determining this is the *adjusted R^2*, which attempts to quantify when the additional variance explained by a new predictor is not enough to warrant its addition in the full model. The adjusted R^2 is defined as

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{4.108}$$

where p is the number of predictors (not including a constant term) and n is the sample size.

Unlike R^2, $\overline{R}^2$ only increases with the addition of a new predictor when the increase in R^2 is more than what would be expected by chance. Thus, one can use the adjusted R^2 to determine the number of predictors by plotting $\overline{R}^2$ for each additional predictor and determining when it reaches a maximum. Beyond this maximum, additional predictors will likely only degrade the fit when independent data is analyzed.