Regression & AR1

Contents

1.1 Linear Regression

In linear regression, the goal is to determine

- the linear fit of X and Y (the regression coefficient)
- the robustness of the fit (the correlation coefficient)

Regression is simple but powerful, however, this also makes it easily misused. In a sense, the entire class (EOFs; Fourier analysis) is all based on regression.

How do we find the slope and y-intercept of the line that best fits the observed data? Let's assume $x(t)$ and $y(t)$ are time series sampled at N time steps (so that each point represents a time step).

First, we have to define what "best-fit" means. For now, we will use the conventional definition which means that we want to reduce the sum of the squared errors of y.

Using the method of least squares:

$$\hat{y}(t) = a_1 x(t) + a_0 \tag{1}$$

where

- $\hat{y(t)}$ denotes the estimate of $y(t)$ based on the linear relationship with $x(t)$
- a_1 denotes the slope, a.k.a. the regression coefficient
- a_0 denotes the y-intercept

Define the error of the fit as the sum of squares of the $y(estimate) - y(actual)$:

$$Q = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{N} (a_1 x_i + a_0 - y_i)^2 \tag{2}$$

where the subscript i denotes the time step.

The error is squared so that

- the error is positive definite (don't want positive and negative errors canceling out)
- the minimization of Q (the derivative of Q) is a linear problem

Note: the square causes larger errors to be more heavily weighted

We now follow steps from our college Calculus I course and find the a_1 and a_0 that minimize Q (sometimes called the cost function):

$$\frac{dQ}{da_0} = 0 \tag{3}$$

$$0 = 2\sum_{i=1}^{N}(a_1 x_i + a_0 - y_i) \tag{4}$$

$$0 = a_1\sum_{i=1}^{N} x_i + a_0 N - \sum_{i=1}^{N} y_i \tag{5}$$

$$\frac{dQ}{da_1} = 0 \tag{6}$$

$$0 = 2\sum_{i=1}^{N}(a_1 x + a_0 - y_i)x_i \tag{7}$$

$$0 = a_1\sum_{i=1}^{N} x_i^2 + a_0\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i \tag{8}$$

Divide through by N and move the y terms to the left hand side, where overbars denote the mean and primes denote departures from the mean:

$$\overline{y} = a_1\overline{x} + a_0 \tag{9}$$

$$\overline{xy} = a_1\overline{x^2} + a_0\overline{x} \tag{10}$$

Two equations, two unknowns.

The solutions are:

$$a_1 = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} \tag{11}$$

Note that,

$$\overline{xy} = \overline{x}\cdot\overline{y} + \overline{x'y'} \tag{12}$$

and

$$\overline{x^2} = \overline{x}^2 + \overline{x'^2} \tag{13}$$

Hence,

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{14}$$

$$a_0 = \overline{y} - a_1\overline{x} \tag{15}$$

a_1

- slope of the line
- regression coefficient
- equal to the covariance of x and y divided by the variance of x

a_0

- y-intercept
- note that if the means of the time series are 0 (they are anomalies), then $a_0 = 0$.

One can put confidence limits on the slope a_1 in a manner of ways. For example, a jackknife approach can be used to determine the sensitivity to removing a single point. Alternatively, the standard error σ_{a_1} (standard deviation) of the slope is given by

$$\sigma_{a_1}^2 = \frac{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{16}$$

where we assume that x is known exactly. 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.

Then, the confidence bounds on the slope follow the normal distribution, and the confidence interval for the true slope b is given by

$$a_1 - t_\alpha^{N-2} \cdot \sigma_{a_1} < b < a_1 + t_\alpha^{N-2} \cdot \sigma_{a_1} \tag{17}$$

The $N-2$ comes from the fact that two degrees of freedom were used to estimate a_1 and a_0.

1.1.1 How good is the fit?

How much we "believe" the regression coefficient (a_1) 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

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{18}$$

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

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \hat{\overline{y}})^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2 \tag{19}$$

where we have used the fact that

$$\hat{\overline{y}} = a_1\overline{x} + a_0 = \overline{y} \tag{20}$$

The percent of the total variance in $y(t)$ explained by the fit $\hat{y}(t)$ is thus given by the ratio

$$\begin{aligned}
r^2 &= \frac{\text{explained variance}}{\text{total variance}} && (21)\\
&= \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} && (22)\\
&= \frac{\sum_{i=1}^{N}(a_1x_i + a_0 - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} && (23)\\
&= \frac{\sum_{i=1}^{N}(a_1x_i + \overline{y} - a_1\overline{x} - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} && (24)\\
&= \frac{\sum_{i=1}^{N}(a_1x_i')^2}{\sum_{i=1}^{N}(y_i'^2)} && (25)\\
&= \frac{(\frac{\overline{x'y'}}{\overline{x'^2}})^2\sum_{i=1}^{N}(x_i')^2}{\sum_{i=1}^{N}(y_i'^2)} && (26)\\
&= \frac{(\overline{x'y'})^2\sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2\sum_{i=1}^{N}(y_i'^2)} && (27)\\
&= \frac{(\overline{x'y'})^2\frac{1}{N}\sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2\frac{1}{N}\sum_{i=1}^{N}(y_i'^2)} && (28)\\
&= \frac{(\overline{x'y'})^2\cdot\overline{x'^2}}{(\overline{x'^2})^2\cdot\overline{y'^2}} && (29)\\
&= \frac{(\overline{x'y'})^2}{\overline{x'^2}\cdot\overline{y'^2}} && (30)
\end{aligned}$$

Hence,

$$r = \frac{\overline{x'y'}}{\sigma_x\sigma_y} \tag{31}$$

Recall that $\sigma_x = (\overline{x'^2})^{1/2}$.

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

The correlation coefficient r

- $r = \sqrt{r^2} = \left(\frac{\overline{x'y'}}{\sigma_x\sigma_y}\right)$

– r varies between -1 and 1

Relationships between r and r^2:

$$r = 0.99 \Rightarrow r^2 = 0.98 \tag{32}$$
$$r = 0.90 \Rightarrow r^2 = 0.81 \tag{33}$$
$$r = 0.70 \Rightarrow r^2 = 0.49 \tag{34}$$
$$r = 0.50 \Rightarrow r^2 = 0.25 \tag{35}$$
$$r = 0.25 \Rightarrow r^2 = 0.06 \tag{36}$$

1.1.2 Relationship between the slope of the regression line and the correlation coefficient:

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{37}$$

so, it follows that

$$a_1 = r\frac{\sigma_y}{\sigma_x} \tag{38}$$

– the regression coefficient (slope) of $y(t)$ on $x(t)$ can be thought of as the correlation coefficient multiplied by the ratio of the standard deviations of y and x

– regression coefficients give you information about the correlation coefficient and the relative amplitudes of variations of y and x

– in the special case where x and y are standardized, the correlation coefficient and the regression coefficient are equal

1.1.3 General comments on linear regression

– only works for linear relationships

– does not reveal relationships that are lagged or out of phase

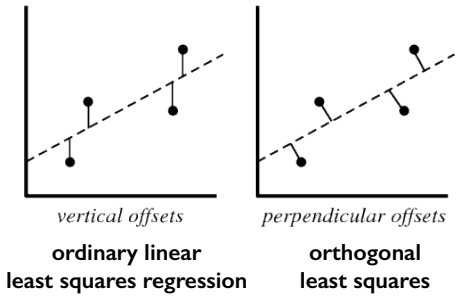
– need to be careful about estimating the true sample size (more on this later)

– correlation DO NOT reveal cause and effect

– flipping the x and y will not give you the same results, so very important to physically justify your choice of x and y!

1.1.4 Orthogonal-Least Squares Regression

If you can't justify which of your data is dependent and which is independent, or it doesn't make sense to do so, orthogonal least squares may actually be what you want. In this case, you minimize the orthogonal distances, rather than the vertical distances:



It just so happens that in 2-dimensions, EOF analysis (to be discussed later) will give you the orthogonal least squares fits. So, in a few weeks, you will be capable of calculating this too.

Example: NOTEBOOK LSQ_OLS.IPYNB

1.1.5 Filtering with linear regression

Consider the decomposition of a variable y into a fraction that is linearly congruent with x and the fraction uncorrelated with x:

$$y(t) = y(t)_{\mathrm{fitted}} + y(t)_{\mathrm{residual}} \tag{39}$$

The fit of $y(t)$ from $x(t)$ is simply:

$$y(t)_{\mathrm{fitted}} = a_1 x(t) + a_0 \tag{40}$$

where

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}} = r\frac{\sigma_y}{\sigma_x} \tag{41}$$

If the means of $y(t)$ and $x(t)$ are zero, then,

$$y(t)_{\mathrm{fitted}} = a_1 x(t) \tag{42}$$

hence,

$$y(t)_{\mathrm{residual}} = y(t) - \frac{\overline{x'y'}}{\overline{x'^2}} \cdot x(t) = y(t) - r\frac{\sigma_y}{\sigma_x} \cdot x(t) \tag{43}$$

- $y(t)_{\mathrm{fitted}}$ represents the LSQ fit of $x(t)$ to $y(t)$
- by construction, $y(t)_{\mathrm{residual}}$ is uncorrelated with $x(t)$

- the fraction of variance of $y(t)$ explained by $x(t)$ is r^2 (ratio of variance of $y(t)_{\text{fitted}}$ to $y(t)$)
- the fraction of variance of $y(t)$ that is not explained by $x(t)$ is $1 - r^2$

1.2 Theory of correlation (Pearson's correlation)

1.2.1 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 together (or don't).

$-1 \leqslant r \leqslant 1$, with numbers closer to 1 implying that the time series vary linearly with one another.

Now, we will discuss techniques for testing the statistical significance of correlations. We will denote the sample correlation as r and the theoretical true value as ρ.

If $\rho = 0$, we can use a form of the z-statistic and t-statistic:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{44}$$

Worked Example: EXAMPLE 2.1

2.1 Testing the hypothesis that $\rho = 0$

We have two time series, each of length 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.

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 $t > 2.1$.

$$t = \frac{0.6\sqrt{20-2}}{\sqrt{1-.6^2}} = 3.18. \tag{45}$$

Since $t > 2.1$, we can reject the null hypothesis.

2.2 Testing the hypothesis that ρ is not equal to 0

What are the confidence limits on the true correlation if you drew 21 samples from a normal distribution and obtained $r = 0.8$?

$$Z = \frac{1}{2}\ln\left(\frac{1+0.8}{1-0.8}\right) = 1.0986 \tag{46}$$

$$\sigma_Z = \frac{1}{\sqrt{21-3}} = .235 \tag{47}$$

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

$$Z - 2.1\sigma_Z \leqslant \mu_Z \leqslant Z + 2.1\sigma_Z \tag{48}$$

$$0.61 \leqslant \mu_Z \leqslant 1.59 \tag{49}$$

We still need to get this back into correlation form, so plugging into the equation for ρ gives

$$0.54 \leqslant \rho \leqslant 0.92 \tag{50}$$

It turns out, the above statistic only works if the underlying distributions are normal, or if N is big, the central limit theorem applies. From what I have found, $N > 20$ or so should be sufficient for the central limit theorem to apply here. However, we can test this of course with our computers.

Example: NOTEBOOK TESTING_NORMALITY_OF_CORRELATIONS.IPYNB

If $\rho \neq 0$, we must use a test called the Fisher-Z Transformation. When the true correlation is not zero, the underlying distribution is not symmetric, and so we cannot use the normal distribution (t/z-test). However, the Fisher-Z Transformation "transforms" the distribution of r into something that is normally distributed.

$$\text{Fisher-Z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{51}$$

Turns out, the Fisher-Z statistic is normally distributed with a mean and standard deviation of:

$$\mu_Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{52}$$

$$\sigma_Z = \frac{1}{\sqrt{N-3}} \tag{53}$$

Thus, the confidence bounds for Z become:

$$Z - t_c\sigma_Z \leqslant \mu_Z \leqslant Z + t_c\sigma_Z \tag{54}$$

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

$$\rho = \frac{e^{2\mu_Z}-1}{e^{2\mu_Z}+1} = \tanh(\mu_Z) \tag{55}$$

As a reminder,

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{1-e^{-2x}}{1+e^{-2x}} \tag{56}$$

$$\tanh ix = i\tan x \tag{57}$$

Note, to calculate a confidence interval for correlations, you need to use the Fisher-Z transformation.

Worked Example: EXAMPLE 2.2

1.2.2 Comparing two non-zero sample correlations

If we want to test the difference between two correlations that are non-zero, we can once again use the Fisher transformation for each and use the fact that Z is normally distributed. Suppose we have two samples of size N_1 and N_2 which give correlation coefficients of r_1 and r_2. We test for a significant difference between these correlations by first calculating the Fisher-Z transformation for each:

$$Z_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) \tag{58}$$

$$Z_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right) \tag{59}$$

$$\tag{60}$$

and then calculating our normal z-score:

$$z = \frac{Z_1 - Z_2 - \Delta_{1,2}}{\sigma_{1,2}} \tag{61}$$

where

$$\Delta_{1,2} = \mu_1 - \mu_2 \tag{62}$$

is the *transformed* difference you expect (your null hypothesis). If your null hypothesis is that the true correlations of the two samples are equal ($\rho_1 = \rho_2$), then

$$\Delta_{1,2} = \mu_1 - \mu_2 = 0 \tag{63}$$

$\sigma_{1,2}$ is given in the following way

$$\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{64}$$

1.2.3 Spearman's rank correlation

Spearman's rank correlations is a nonparametric test that tests whether a set of paired data monotonically vary together (when one goes up, the other goes down), but it is not concerned with how much it goes up or down. Since this is a nonparametric test, no assumption about normality needs to be made.

The idea is very simple, your original paired data x_i and y_i get converted into ranks X_i and Y_i and

$$\rho = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2 (Y_i - \overline{Y_i})^2}} \tag{65}$$

When there are duplicate values, the ranks are equal to the average position.

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

$$\sigma_\rho = \frac{0.6325}{(N-1)^{1/2}} \tag{66}$$

To determine significance, you can use the Fisher-Z test and the t-test (for a null hypothesis that $\rho = 0$) as for the Pearson correlation.

Example: SEE SLIDES 08_CORRELATION.PDF

Note: A second nonparametric method is called Kendall's Tau Rank Correlation. We won't go into this here.

1.3 Autocorrelation & Estimating the # of Independent Samples

Thus far, we have assumed that our time series (and data sets) are all stationary and have no intrinsic memory. Now, we will discuss these assumptions, and how to determine the true number of degrees of freedom in an autocorrelated data set.

1.3.1 Stationarity

Stationarity implies that the statistics of a time series (its mean and higher-order moments) are independent of time, i.e. unchanging in time. In general, we will assume that this is the case. Note that this means one should *remove any trend* in the data before performing the analysis. The trend can be removed in the method previously discussed using linear regression.

1.3.2 Autocorrelation

The autocovariance function ($\gamma(t)$) is the covariance of a time series with itself at another time, as measured by a time lag (or lead) τ. For a time series $x(t)$, it is defined as

$$\gamma(\tau) = \frac{1}{(t_N - \tau) - t_1} \sum_{t=t_1}^{t_N-\tau} [x'(t) \cdot x'(t+\tau)] \tag{67}$$

where t_1 and t_N are the starting and end points of the time series, respectively, and the prime denotes departures from the long-term mean.

Example: DRAW OUT EXAMPLE OF HOW THIS WORKS ON THE BOARD.

Note that for a continuous time series with time positions $k = 1, 2, 3...N$:

$$\gamma(\tau) = \overline{x'(t)x'(t+\tau)} \tag{68}$$

and for $\tau = 0$, the autocovariance is $\gamma(0) = \overline{x'^2}$ = variance.

The more commonly used *autocorrelation* $\rho(\tau)$ is just $\gamma(\tau)$ normalized by $\gamma(0)$. It is simply the correlation of a time series with itself at another time.

Notes on the autocorrelation:

- γ is symmetric about $\tau = 0$
- $-1 \leqslant \rho(\tau) \leqslant 1$
- $\rho(0) = 1$
- if the time series is not periodic, $\rho(\tau) \to 0$ as $\tau \to \infty$

Example: SEE SLIDES 08_CORRELATION.PDF

1.3.3 The first order autoregressive model

Also referred to as a "first order Markov process" or "red noise".

red noise process: "today is like yesterday + noise"

A red noise time series is defined mathematically as:

$$x(t) = a \cdot x(t - \Delta t) + b \cdot \epsilon(t) \tag{69}$$

where

- x is a standardized variable
- Δt is the time interval between data points (and is assumed to be a constant here)
- a lies between 0 and 1 and measures the memory of the previous state
- $(t - \Delta t)$ is the day before day t
- $\epsilon(t)$ is a random variable drawn from the standard normal distribution and represents noise in the system

To determine a: multiply the l.h.s. and r.h.s. by $x(t - \Delta t)$ and take the time average

$$\overline{x(t)x(t - \Delta t)} = a \cdot \overline{x(t - \Delta t) \cdot x(t - \Delta t)} + b \cdot \overline{\epsilon(t)x(t - \Delta t)} \tag{70}$$

- since x is standardized (variance of 1), the first term of the r.h.s. is $a \cdot 1$
- since $\epsilon(t)$ is random in time, assuming your time series is long enough, the last term on the r.h.s is 0.

Thus, for a standardized x

$$a = \overline{x(t)x(t - \Delta t)} = \gamma(\tau = 1) = \gamma(1) \tag{71}$$

That is, a is the autocovariance at lag Δt, or, one time step ahead. Since x is standardized, this is also the autocorrelation at lag Δt, so,

$$a = \rho(\Delta t) = \rho(1) \tag{72}$$

for a standardized time series.

What about b - the magnitude of the noise? Since $x(t)$ and $\epsilon(t)$ both have unit variance, one can square both sides of the red-noise equation and then take the average to solve for b:

$$\overline{x^2(t)} = \overline{a^2 x^2(t-\Delta t)} + \overline{b^2 \cdot \epsilon^2(t)} \tag{73}$$
$$1 = a^2 \cdot 1 + b^2 \cdot 1 \tag{74}$$
$$b = \sqrt{1-a^2} \tag{75}$$

One can then use the equation for a red noise process to predict the value of x at a later time. For example, two time steps into the future:

$$x(t) = a \cdot x(t-\Delta t) + b \cdot \epsilon(t) \tag{76}$$
$$x(t+\Delta t) = a \cdot x(t) + b \cdot \epsilon(t) \tag{77}$$
$$x(t+2\Delta t) = a \cdot x(t+\Delta t) + b \cdot \epsilon(t) \tag{78}$$

Multiply both sides by $x(t)$ and time average:

$$\overline{x(t) \cdot x(t+2\Delta t)} = \overline{x(t) a \cdot x(t+\Delta t)} + \overline{x(t) b \epsilon(t)} \tag{79}$$
$$\overline{x(t) \cdot x(t+2\Delta t)} = \overline{a \cdot x(t) \cdot x(t+\Delta t)} + 0 \tag{80}$$
$$\tag{81}$$

Therefore,

$$\rho(2\Delta t) = a\rho(\Delta t) = \rho^2(\Delta t) \tag{82}$$

since $a = \rho(\Delta t)$. More generally,

$$\rho(n\Delta t) = \rho^n(\Delta t). \tag{83}$$

The function that has this property is the exponential: $e^{(nx)} = (e^x)^n$. So, it turns out that the autocorrelation for a red-noise time series is an exponential:

$$\rho(n\Delta t) = e^{(-n\Delta t)/T_e} \tag{84}$$

where T_e is the e-folding time of the autocorrelation function (more on this in a second). In other words, the autocorrelation function of red noise decays exponentially for increasing lag $\tau = n\Delta t$.

The e-folding time-scale is the time it takes for the autocorrelation to drop to $1/e = 0.368$ of the original value (1), and can be computed as

$$T_e = -\frac{\Delta t}{\ln(a)} \tag{85}$$

So, if $\Delta t = 1$, and $\rho(\Delta t) = \rho(1) = a = 0.6$, then the e-folding time of the autocorrelation function is $T_e = 2$.

1.3.4 White Noise

Whereas red noise is defined as

$$x(t) = a \cdot x(t - \Delta t) + b \cdot \epsilon(t) \tag{86}$$

White noise is the special case where $\rho(\tau > 0) = 0$, so $a = 0$.

White noise has equal power at all frequencies and has zero autocorrelation (no memory of the previous time steps). In geophysical applications, white noise is generally assumed normally distributed.

Example: NOTEBOOK CORRELATION_WITH_MEMORY_EXAMPLES.IPYNB

1.3.5 Independent samples and effective sample size

Recall that the sample size N greatly impacts the variance of the sample means.

Persistence in a data set leads to an *overestimation* of the sample size, because each data point is not independent of those around it.

Consider the t-statistic:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{N-1}}} \tag{87}$$

If we assume no persistence in a red-noise time series, the sample mean standard deviation $\hat{s} = \frac{s}{\sqrt{N-1}}$ will be an underestimate and thus the t statistic will be over-estimated.

The most convenient way to deal with persistence in your time series is to introduce an *effective sample size*, N^*.

$N^* \leqslant N$ and can be substituted into the original formulas in place of N.

The estimation of N^* is generally approached assuming that the data follows a first-order autoregressive process (red noise). In this case, N^* can be estimated using the approximation:

$$\frac{N^*}{N} \cong \frac{1 - \rho(\Delta t)}{1 + \rho(\Delta t)} \text{ (see Wilks page 127)} \tag{88}$$

where N is the number of data points in the time series, Δt is the time step and $\rho(\Delta t)$ is the autocorrelation at lag 1.

- for white noise, if $\rho(1) = 0$, $N^* = N$
- as $\rho(1)$ increases, N^* decreases

The above approximation is nearly identical to the discrete version of the effective sample size proposed by Leith (Journal of Applied Meteorology, p. 1066, 1973), given by

$$N^* \cong \frac{N\Delta t}{2T_e} = \frac{\text{total length of record}}{\text{two times the e-folding time of the autocorrelation}} \tag{89}$$

where T_e is the e-folding time of the autocorrelation function of the time series. The factor of 2 is included because any given point in a red noise time series can be predicted by points both before and after that point.

As T_e (the "redness" of the time series) increases, we get fewer degrees of freedom from each observation.

Note that the above can be re-written as:

$$\frac{N^*}{N} \cong \frac{\Delta t}{2T_e} = \frac{\Delta t}{-2\frac{\Delta t}{\ln a}} = \frac{\ln a}{-2} \tag{90}$$

As an extreme example, say I have a time series $x(t)$, and I tell you that $x(1) = 2$ and that all other values of x are given by $x(1)$. It wouldn't make sense for the sample mean calculation to have your degrees of freedom equal to N, since really, all you need to know is $x(1)$ and then you know all of the other values.

Using the Leith formula, one can compute N^* as a function of the lag-1 autocorrelation of a time series:

$\rho(\Delta t)$	< 0.16	0.3	0.5	0.7	0.9
N^*/N	1	0.6	0.35	0.18	0.053

Finally, Bretherton et al. (Journal of Climate, pg. 1990, 1999) take a less conservative approximation and have suggested using:

$$\frac{N^*}{N} \cong \frac{1 - \rho^2(\Delta t)}{1 + \rho^2(\Delta t)} \tag{91}$$

This formula yields almost 2 times more degrees of freedom than the Leith approximation. This approximation may be used when one is analyzing variance or higher-order moments. Otherwise, if one is interested in the mean, $\frac{N^*}{N} \cong \frac{1-\rho(\Delta t)}{1+\rho(\Delta t)}$ should be used.

Example: SHOW PLOT OF COMPARISON OF LEITH AND BRETHERTON

Example: DISCUSS OTHER METHODS OF DEALING WITH AUTOCORRELATION

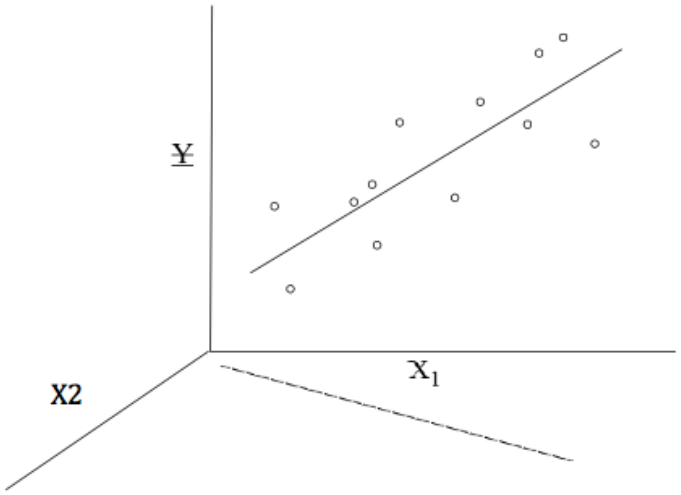
1.4 Multiple regression (multi-linear regression)

Basic idea: Generalize the derivation of the regression coefficient to multiple linear predictors.

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n \tag{92}$$

Note that now the fit is in multiple phase space.

Consider the two predictor case:



What does it mean if X_1 and X_2 are at right angles?

- they are orthogonal predictors (the inner product is 0)
- they give you independent information
- if X_1 and X_2 cover all possible combinations for the space, they are said to "form a basis"

If X_1 and X_2 are not orthogonal

- they are not independent
- they repeat information, (are redundant)

Note that if only X_1 was used to predict Y, one could still get a best-fit-line, but all information in the X_2 direction would be ignored.

The usefulness of independent predictors will later motivate EOF analysis.

Generalized normal equations

For a single predictor x, we want to minimize the cost function Q which is:

$$Q = \sum_{i}^{N} (\hat{y}_i - y_i)^2 = \sum_{i}^{N} (a_1 x_i + a_0 - y_i)^2 \tag{93}$$

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

$$Q = \sum_{i}^{N} (\hat{y}_i - y_i)^2 = \sum_{i}^{N} (a_0 + a_1 x_{1,i} + a_2 x_{2,i} + a_3 x_{3,i} + ... + a_n x_{n,i} - y_i)^2 \tag{94}$$

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

$$\frac{\partial Q}{\partial a_i} = 0 \tag{95}$$

where i goes from 0 to n.

$$\begin{aligned}
\overline{y} &= a_0 + a_1\overline{x_1} + a_2\overline{x_2} + ... + a_n\overline{x_n} && (96)\\
\overline{x_1 y} &= a_0\overline{x_1} + a_1\overline{x_1^2} + a_2\overline{x_1 x_2} + ... + a_n\overline{x_1 x_n} && (97)\\
\overline{x_2 y} &= a_0\overline{x_2} + a_1\overline{x_2 x_1} + a_2\overline{x_2^2} + ... + a_n\overline{x_2 x_n} && (98)\\
&... && (99)\\
\overline{x_n y} &= a_0\overline{x_n} + a_1\overline{x_n x_1} + a_2\overline{x_n x_2} + ... + a_n\overline{x_n^2} && (100)
\end{aligned}$$

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 (96) is no longer useful).

For the jth equation:

$$\overline{x_j y} = \sum_{i=1}^{n} a_i \overline{x_j x_i} \tag{101}$$

One can write this in matrix form as:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1 x_2} & \overline{x_1 x_3} & ... \\ \overline{x_2 x_1} & \overline{x_2^2} & \overline{x_2 x_3} & ... \\ \overline{x_3 x_1} & \overline{x_3 x_2} & \overline{x_3^2} & ... \\ ... & & & \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ... \end{bmatrix} = \begin{bmatrix} \overline{x_1 y} \\ \overline{x_2 y} \\ \overline{x_3 y} \\ ... \end{bmatrix}$$

This can also be written as

$$\mathbf{C_{x_i x_j}} a_j = C_{x_i y} \tag{102}$$

1. since $\overline{x} = 0$ for all j, the l.h.s. is the *covariance matrix* of the predictors. The diagonal elements are the *variances* of the predictors and the off-diagonal elements are the *covariances* between predictors.

2. the r.h.s. is the covariance vector of the predictors (x_j) and the predictand (y)

3. if each variable has been standardized (mean of 0, standard deviation of 1), the l.h.s. is the *correlation matrix* of x_j, and the r.h.s. is the *correlation vector*

4. if x are time series at different locations in a data set, the covariance matrix yields information about the structures of the data, and tells you something about the spatial variability of the different points

5. if the predictors are linearly independent, the off diagonal elements are all 0 and the a_j's can be found algebraically

6. otherwise, the a_j's are found as the inverse of the covariance matrix $\times$ the r.h.s. vector (solve for a but using matrix algebra). There are a variety of techniques for inverting the covariance matrix.

$$\mathbf{C_{x_i x_j}}^{-1}\mathbf{C_{x_i x_j}}a_j = \mathbf{C_{x_i x_j}}^{-1}C_{x_i y} \tag{103}$$

$$a_j = \mathbf{C_{x_i x_j}}^{-1}C_{x_i y} \tag{104}$$

1.4.1 Multiple regression - how many variables should I use?

To make life a bit easier, let's assume that our predictors x_j and predictand y have all be standardized. Then, the normal equations for multiple linear-least-squares regression can be written in the following way:

$$r(x_i, x_j)a_i = r(x_j, y) \tag{105}$$

where r represents the correlations/covariances. Let's take the simple case where we only have two predictors:

$$\hat{y} = a_1 x_1 + a_2 x_2 \tag{106}$$

Then,

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1 x_2} \\ \overline{x_2 x_1} & \overline{x_2^2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \overline{x_1 y} \\ \overline{x_2 y} \end{bmatrix}$$

can be re-written as

$$\begin{bmatrix} 1 & r_{1,2} \\ r_{1,2} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{1,y} \\ r_{2,y} \end{bmatrix}$$

since $r_{1,1} = r_{2,2} = 1.0$ and $r_{1,2} = r_{2,1}$.

We solve for a_1 and a_2 and find that

$$a_1 = \frac{r_{1,y} - r_{1,2}r_{2,y}}{1 - r_{1,2}^2} \tag{107}$$

$$a_2 = \frac{r_{2,y} - r_{1,2}r_{1,y}}{1 - r_{1,2}^2} \tag{108}$$

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

$$\overline{y^2} = \overline{(y - \hat{y})^2} + \overline{(\hat{y} - \overline{y})^2} \tag{109}$$

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

(but don't forget that $\overline{y} = 0$). Using the fact that $\hat{y} = a_1 x_1 + a_2 x_2$ it can be shown that

$$1 = \frac{\overline{(y - \hat{y})^2}}{\overline{y^2}} + R^2 \tag{110}$$

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

$$R^2 = \frac{r_{1,y}^2 + r_{2,y}^2 - 2 r_{1,y} r_{2,y} r_{1,2}}{1 - r_{1,2}^2} \tag{111}$$

To demonstrate this, let's do an example: 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

$$r_{1,y} = r_{2,y} = r_{1,2} = 0.5 \tag{112}$$

For the first predictor only, the variance explained is

$$R_1^2 = r_{1,y}^2 = 0.25 \tag{113}$$

Adding a second predictor, x_2, leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.5^2 - 2 \times 0.5 \times 0.5 \times 0.5}{1 - 0.5^2} = 0.33 \tag{114}$$

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

However, now lets assume that $r_{2,y} = 0.25$ and everything else remains the same. Then, adding the second predictor leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.25^2 - 2 \times 0.5 \times 0.25 \times 0.5}{1 - 0.5^2} = 0.25 \tag{115}$$

Thus, adding the second predictor did not increase the explained variance!

To add an additional predictor, one must be sure that it *increases* the explained variance. There is a *minimum useful correlation* correlation between the new predictor and the predictand in order to make adding it useful.
Minimum Useful Correlation

$$|r(x_2, y)|_{\text{min useful}} > |r(x_1, y) \cdot r(x_1, x_2)| \tag{116}$$

You will see that in our example above

$$|r(x_2, y)|_{\text{min useful}} = 0.5 \times 0.5 = 0.25 \tag{117}$$

In other words, $r_{2,y}$ must be greater than 0.25 in order to add anything to the fit. Similar (but more complex) steps can be followed for adding a 3rd predictor.

If you think about it, it would be ideal for $r_{1,2} = 0$, in which case, you have two completely independent predictors. On the flip side, $r_{1,2} = 1.0$ is completely useless, since x_2 provides you with no additional information than you already got from x_1.

What's more, adding additional predictors can actually be detrimental overall when applying the fit to independent data. This is because you can *over fit* the data, in essence, using the predictors to fit the noise, rather than the signal only. It is a good idea to use as few predictors as possible - and test the fit on independent data after the regression coefficients have been determined. We will discuss more about how to pick optimal predictors in the future.

Example: ADJUSTED R^2: HTTPS://EN.WIKIPEDIA.ORG/WIKI/COEFFICIENT_OF_DETERMINATION#ADJUSTED_R2

Example: NOTEBOOK MINIMUM_CORR_FOR_ADDED_VALUE.IPYNB

1.4.2 Granger causality

Granger-causality is a statistical approach for determining whether one time series, x, is useful in predicting another time series y. While it has a seemingly fancy name, it is really just two multi-linear regressions. The name implies that the method can determine whether x *causes* y, however, it is important to note that it cannot do this! In addition, Granger causality cannot determine whether there is a third driver causing the other two, nor does it account for instantaneous relationships (i.e. lag zero).

With these caveats out of the way, let's see what Granger causality can do. If we are interested in predicting $y(t)$, the idea is to first see how much of y_t can be predicted using only lagged values of y itself. That is, step one is to perform the following multi-linear regression

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_k y_{t-k} \tag{118}$$

$$= a_0 + \sum_{\tau=1}^{k} a_\tau y_{t-\tau} \tag{119}$$

where the a_is are the regression coefficients.

Step two is to ask "can I obtain additional, *unique* information about y_t using lagged values of x?" That is, does adding information from x provide me with additional information about y beyond what y already contains itself? To answer this, we once again perform multi-linear regression, only now we add lagged values of x

$$y_t = b_0 + b_1 y_{t-1} + ... + b_k y_{t-k} + c_p x_{t-p} + ... + c_k x_{t-k} \tag{120}$$

$$= b_0 + \sum_{\tau=1}^{k} b_\tau y_{t-\tau} + \sum_{\tau=p}^{k} c_\tau x_{t-\tau} \tag{121}$$

In our notation, $\mathtt{p}$ is the smallest lag considered and $\mathtt{k}$ is the largest, such that $\mathtt{p} \geqslant 1$ and $\mathtt{k} \geqslant 1$.

The final step is to now see if adding lagged values of $\mathtt{x}$ provided additional predictive power of $\mathtt{y}_\mathtt{t}$. To do this, we require that two conditions are met

1. There exists at least one significant $\mathtt{c}$ according to a t-test.
2. The addition of the $\mathtt{c}$ terms collectively add power to the regression according to an F-test.

The second condition involves comparing the amount of variance explained in (119) with the amount of variance explained in (121), and as we have learned, we can do this with an F-test.

If you find that lagged values of $\mathtt{x}$ do indeed provide unique information about $\mathtt{y}_\mathtt{t}$ and that this increase in variance explained is significant, then it is said that "$\mathtt{x}$ *Granger-causes* $\mathtt{y}$".

To apply this approach, you will need to choose how many lags to include, that is, what is $\mathtt{p}$ and $\mathtt{k}$. You may choose these based on physical understanding of your system, or using more sophisticated methods of determining how many lags to use, such as the Akaike information criterion.

Example: SEE SLIDES 08A_GRANGER_CAUSALITY.PDF