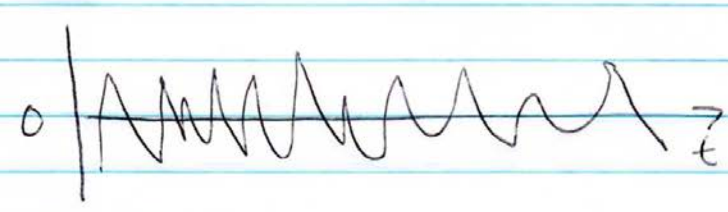
How to account for autocorrelation in evaluating statistical significance?

→ Problem: Data may not be independent in time. *This persistance reduces the degrees of freedom in the data.

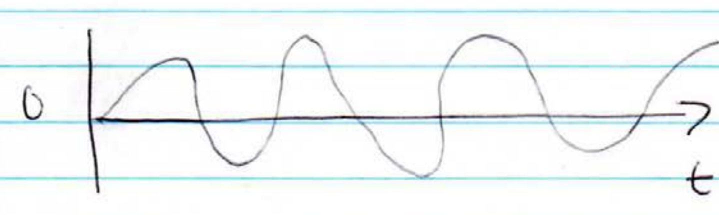
Example mentioned before:
ENSO time series (of whatever index) monthly for 50 years. Is the true sample size = 50 years × 12 $\frac{\text{months}}{\text{yr}}$ = 600 months?

→ **Answer**: NO because sea surface temperatures in the tropical Pacific are fairly persistent from month to month.

The more autocorrelated data is, the "smoother" it will appear in time:



Little correlation in time
"White"



More correlation in time
"Red"

Various ways to account for the autocorrelation
Simplest way is just to use lag - 1
autocorrelation.

Lag - 1 Autocorrelation

$\rightarrow$ time

Original Data: $a_1 \quad a_2 \quad a_3 \quad a_4 \ldots a_n$

Lagged Data: $\quad a_1 \quad a_2 \quad a_3 \ldots a_{n-1} \quad a_n$

Calculating the correlation of the following data series: (a_i, a_{i-1})
where i = time index

This correlation is called lag - 1 autocorrelation and is denoted by ρ_1.

Use ρ_1 to get a modified sample size:

Effective sample size $\rightarrow$ $$\boxed{n' \simeq n \frac{1-\rho_1}{1+\rho_1}}$$

n = original # of samples

n' = # of samples accounting for persistence.

where $n' < n$

The effective sample size can also be used to inflate the variance accounting for autocorrelation

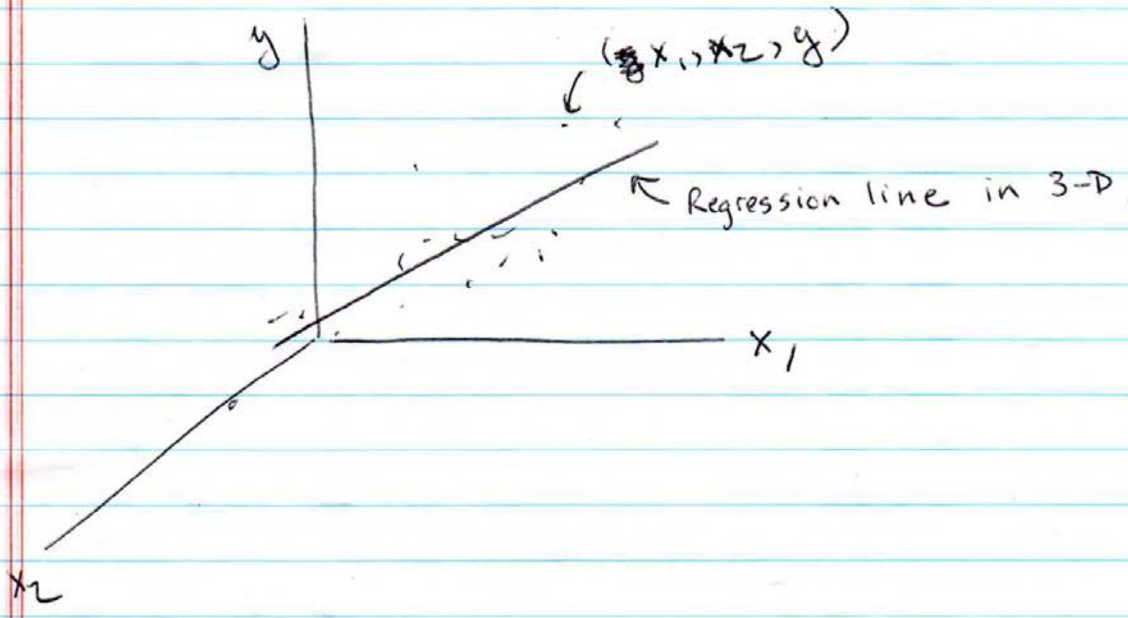
Adjusted variance $\rightarrow$ $$\frac{s^2}{n'} = \frac{s^2}{n}\left(\frac{1+\rho_1}{1-\rho_0}\right)$$

Multiple Regression

Idea: Generalize the derivation of the linear regression coefficient to multiple predictors:

$$\hat{y} = a_0 + a_1 x_i + a_2 x_2 + \dots a_N x_N$$

Fit now is in a multiple phase space.



If x_1 and x_2 are orthogonal predictors (i.e. they are not related) they give independent information regarding y.

Recall the expressions we got for the case of one predictor from our discussion of regression before

By least-squares error minimization, get:

$$\bar{y} = a_0 + a_1 \bar{x}$$ → Minimization wrt a_0

$$\overline{xy} = a_0 \bar{x} + a_1 \overline{x^2}$$ → Minimization wrt a_1

Now generalized to multiple predictors:

$$\bar{y} = a_0 + a_1 \bar{x}_1 + a_2 \bar{x}_2 + \dots + a_n \bar{x}_n$$ → Minimization with respt. to a_0.

Get multiple equations for minimization wrt a_1

For first predictor (x_1)

$$\overline{x_1 y} = a_0 \bar{x}_1 + a_1 \overline{x_1^2} + a_2 \overline{x_2 x_1} + \dots + a_n \overline{x_n x_1}$$

For n^{th} predictor (x_n)

$$\overline{x_n y} = a_0 \bar{x}_n + a_1 \overline{x_1 x_n} + a_2 \overline{x_2 x_n} + \dots a_n \overline{x_n^2}$$

If we assume normalized variables (i.e. mean = 0 and std. dev. = 1), then the equation reduces to:

$$\overline{(x_i x_j)}\, a_i = \overline{x_i y}$$

Covariance of predictors (2-D) MATRIX

Coefficients (1-D) VECTOR

Covariance of predictors and predictands (1-D) VECTOR

In matrix notation:

$$\begin{array}{cc} & j \rightarrow \\ i \downarrow & \begin{bmatrix} \overline{x_1^2} & \overline{x_1 x_2} & \overline{x_1 x_3} & \cdots & \overline{x_1 x_n} \\ \overline{x_2 x_1} & \overline{x_2^2} & \overline{x_2 x_3} & \cdots & \overline{x_2 x_n} \\ \overline{x_3 x_1} & \overline{x_3 x_2} & \overline{x_3^2} & \cdots & \overline{x_3 x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ \overline{x_n x_1} & \overline{x_n x_2} & \overline{x_n x_3} & \cdots & \overline{x_n^2} \end{bmatrix} \end{array} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \overline{x_1 y} \\ \overline{x_2 y} \\ \overline{x_3 y} \\ \vdots \\ \overline{x_n y} \end{bmatrix}$$

LHS: Covariance matrix of predictors

RHS: Covariance vector of predictors and predictands.

* If x's are standardized, LHS becomes the correlation matrix.

If the predictors are orthogonal, then the cross correlation terms (i.e. $\overline{x_1 x_2}$, etc.) are $= 0$.

Then the correlation coefficients are:

$$a_1 = \frac{\overline{x_1 y}}{\overline{x_1^2}} \; ; \; a_2 = \frac{\overline{x_2 y}}{\overline{x_2^2}} \quad etc\ldots$$

Otherwise must invert the covariance matrix to solve..

<u>Case of two predictors</u>

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2$$ Regression equation

If the variance terms $=1$ (i.e. $\overline{x_1^2} = 1$ and $\overline{x_2^2} = 1$), then a_1 and a_2 are:

$$a_1 = \frac{\overline{x_1 y} - (\overline{x_2 y})(\overline{x_1 x_2})}{1 - (\overline{x_1 x_2})^2}$$

$$a_2 = \frac{\overline{x_2 y} - (\overline{x_1 y})(\overline{x_1 x_2})}{1 - (\overline{x_1 x_2})^2}$$

Can get this using matrix multiplication..
(possible HW problem)

If the variables (predictors + predictands) have been normalized; can express this in terms of correlation coefficients (r)

$$a_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}$$

$$a_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}$$

Calculate the explained variance and unexplained variance, then, in a similar way as with one predictor.

Additional notes

- Mutiple Linear regression can be used as a data filter - and often is.

- Same caveats as with linear regression / correlation mentioned earlier.

- Can't just keep adding potential predictors blindly without evaluating how much they improve exp. variance.