3. Regression

In this section 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 a good introduction to more sophisticated methods of linear statistical modeling.

3.1 Linear Least-square Curve Fitting

3.1.1 Independent Variable Known

Suppose that 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:

$$\hat{y} = a_o + a_1 \cdot x + \varepsilon \tag{3.1}$$

It must be noted that we assume *x* is known with precision, and that we wish to estimate *y*, based on known values of *x*. The case where both *y* and *x* contain uncertainties will be discussed later. The error ε can be minimized in the least square sense by defining an error function, *Q*, in the following way:

$$Q = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{y} - y_i\right)^2 \tag{3.2}$$

So that the error function is the sum of the squared differences between the data and our linear equation, when this is minimized by choosing the parameters a_o and a_1 in (3.1) we will have the *least-squares linear fit* to the data.

Squaring the error has several consequences.

1. The result is positive definite.

2. The minimization results in a linear problem to solve.

3. Large errors are weighted more heavily than small departures.

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 later (EOF, SVD, PCA, etc.) share these same properties of linear least squares techniques.

We want to select the constants a_O and a_1 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

$$\frac{\partial Q}{\partial a_o} = 0; \quad \frac{\partial Q}{\partial a_1} = 0 \quad \Rightarrow \text{The “normal” equations}$$

It is easy to show that these result in the following forms.

$$\frac{\partial Q}{\partial a_o} = 2a_o N + 2a_1 \sum x_i - 2\sum y_i = 0 \tag{3.3}$$

$$\frac{\partial Q}{\partial a_1} = 2a_o \sum x_i + 2a_1 \sum x_i^2 - 2\sum x_i y_i = 0 \tag{3.4}$$

These in turn can be written in the form of a matrix equation for the coefficients.

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \tag{3.5}$$

The solutions for the coefficients are:

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'\,y'}}{\overline{x'^2}}; \text{ where} (\ )' = (\ ) - \overline{(\ )} \tag{3.6}$$

Where we have used the covariance for the first time, I believe.

$$\overline{x'\,y'} \equiv \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

We see that a_1 is just the covariance of *x* with *y* divided by the variance of *x*. And:

$$a_o = \overline{y} - a_1 \overline{x} \tag{3.7}$$

As an exercise, demonstrate that the minimum value of the error functional that is obtained when the linear regression is performed is given by:

$$Q_{\min} = \overline{y'^2} - \frac{\left(\overline{x'\,y'}\right)^2}{\overline{x'^2}} = \overline{y'^2} - a_1^2\,\overline{x'^2} \tag{3.8}$$

From (3.8) we see that the minimum error, is the total variance, minus the explained part, which is related to the squared slope coefficient a_1 and the variance of the predictor.

Many other curves besides a straight line can be fitted to data using the same procedure. Some common examples are power laws and polynomials, shown in (3.9).

$$
\begin{aligned}
&y = ax^{b} \quad \Rightarrow \ln y = \ln a + b \ln x \\
&y = ae^{bx} \quad \Rightarrow \ln y = \ln a + bx \qquad (3.9) \\
&y = a_o + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n
\end{aligned}
$$

3.1.2 Both Variables 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.

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 much more elegant than the methods used by Kermack and Haldane (1950) and York (1966), so I would regard these methods now merely with historical interest. EOF/PC analysis is also easily generalized to many dimensions.

References:

Adcock, R. J., A problem in least squares, *Analyst,* **5**, 53, 1878.

Kermack, K. A. and J. B. S. Haldane, Organic Correlation and Allometry, *Biometrika,* **37,** 30-41, 1950.

Pearson, K., On lines and planes of closest fit to systems of points in space, *Phil. Mag.,* **2,** 559, 1901.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992: *Numerical Recipes*. Second Edition, Cambridge U. Press, Cambridge, UK, 963. (In the second edition, line fitting is discussed in chapter 15.)

York, D., Least-squares fitting of a straight line, *Can. J. Phys.,* **44,** 1079-1086, 1966.

3.2 Theory of Correlation

Suppose we wish to answer the question, “How ‘good’ is our least-squares fit?” We can define a measure of the ‘badness’ of the fit as the ratio of the error to the total variance:

$$\frac{\text{error}}{\text{total variance}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}-y_i\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}-y_i\right)^2}{\overline{y'^2}} \tag{3.10}$$

Here, of course, $\hat{y}$ is the value obtained from the fit. The smaller this ratio is, the better the quality of the regression.

For the case of the simple linear fit we can write the individual values in a useful form using the following sequence of steps:

$$\begin{aligned} y_i &= y_i - \hat{y} + \hat{y} \\ &= y_i^* + a_o + a_1 x_i \qquad \text{where } \mathrm{y_i^*} = y_i - \hat{y} \\ &= y_i^* + \overline{y} - a_1\overline{x} + a_1 x_i \\ y_i &= \overline{y} + a_1 x'_i + y_i^* \end{aligned} \tag{3.11}$$

This allows us to write:

$$y_i' = \left(y_i - \overline{y}\right) = a_1 x_i' + y_i^* \tag{3.12}$$

From whence we get the variance in the interesting form:

$$\overline{y'^2} = a_1^2\,\overline{x_i'^2} + \overline{y_i^{*2}} + 2a_1\,\overline{x_i' y_i^*} \tag{3.13}$$

As an exercise, you can demonstrate that the last term on the right above is identically zero, so that the variance has been decomposed into the explained part and the unexplained part. The last part is zero because we have chosen a_1 to minimize the error of a linear fit. If the covariance in the last term of (3.13) is not zero, then the linear fit is not optimum because the error is correlated with the independent variable. Dividing through by the variance of *y*, we obtain:

$$\frac{a_1^2\,\overline{x_i'^2}}{\overline{y'^2}} + \frac{\overline{y_i^{*2}}}{\overline{y'^2}} = 1 \tag{3.14}$$

In words,

$$\text{fraction of explained variance } + \text{ fraction of unexplained variance} = 1$$

But from the linear least-squares solution we know that:

$$\frac{\overline{\mathrm{x'\,y'}}}{\overline{\mathrm{x'}^2}} = a_1; \text{ so that we find}$$

$$a_1^2 \frac{\overline{x'^2}}{\overline{y'^2}} = \frac{\left(\overline{x'\,y'}\right)^2}{\overline{x'^2}\,\overline{y'^2}} = r^2 = \text{the fraction of explained variance} \qquad (3.15)$$

and where

$$r = \frac{\overline{x'\,y'}}{\sigma_x \sigma_y} = \text{the correlation coefficient; } -1 < r < 1 \qquad (3.16)$$

Note that if $\sigma_x = \sigma_y = 1$ and $\bar{x} = \bar{y} = 0$, then the linear, least-square solution is $y = r$ x.

The correlation coefficient-squared is equal to the fraction of variance explained by a linear least-squares fit between two variables.

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}; 1 - r^2 = \frac{\text{Unexplained Variance}}{\text{Total Variance}}$$

Consider the following example. 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 “explained” by sunspots is $r^2 = 0.25$. The fraction of unexplained variance is 0.75. The root-mean-square (rms) error, normalized by the total variance is thus:

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

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	**rms error**
0.98	20%
0.90	43%
0.80	60%
0.50	87%
0.30	96%

As this table illustrates, statistically significant correlations are not necessarily useful for forecasting. If you have enough data you may be able to show that a measured 0.3 correlation coefficient proves 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) 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.

The correlation coefficient *r* is often used as a measure of whether data sets are "related" or not and, as we will describe below, it can be tested for statistical significance. A number of pitfalls exist that one should avoid when using correlation coefficients for this purpose:

1. It 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.

2. It 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).

3. The statistical tests apply to independent data. Often the sample is not independent. The actual number of degrees of freedom may be much smaller than the sample size.

4. 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. These must be regarded as spurious correlations from the standpoint of seeking to find real relationships between variables that might lead to physical insight or be useful in prediction. 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? During the period between 1900-1990, the height of hemlines on dresses in the US very closely followed the Dow-Jones Average, or vice versa.

We illustrate below some of the problems of linear regression and the correlation coefficient by plotting four sets of variables, each of which has a linear fit with a correlation of 0.7 or an explained variance of 49%. You can see that in one case linear regression looks like a useful fit, but in the other cases we have a jump in the data not a linear trend, a parabola and an outlier.

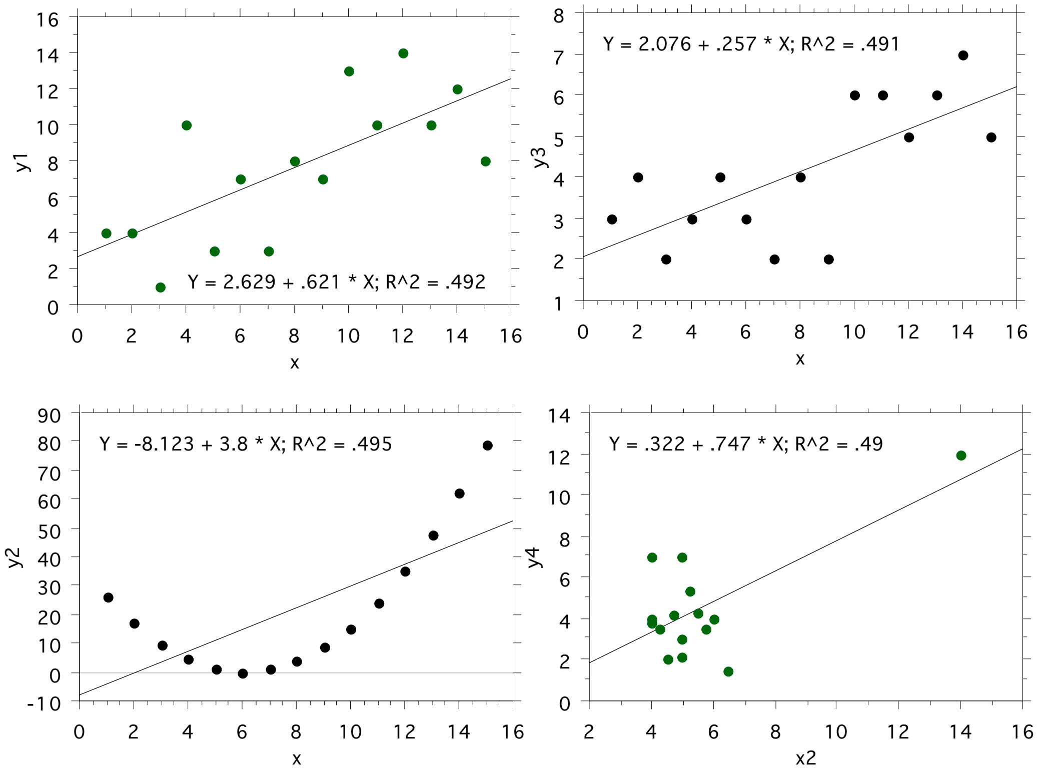

Figure of four sets of data, each with a linear correlation of 0.7 with the x-axis.

3.3 Sampling Theory of Correlation

In this section we will review the techniques for testing the statistical significance of correlation coefficients. We will suppose that 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. We will assume that we are sampling from a bivariate Normal distribution and use the Normal probability distribution.

When the true correlation coefficient is zero, the distribution is symmetric and we can make a fairly direct application of Student's t distribution. For example, suppose that we wish to test the hypothesis that $\rho = 0$, given a sample size of 18 and a sample correlation coefficient, $r = 0.32$. The steps in the process are:

1. The significance level desired is 0.05 (95%).

2. H_0: $\rho = 0$
   H_1: $\rho =$ not 0

3. The statistic used is

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

which has a Student's t distribution with $\nu = N–2$ degrees of freedom.

4. If $t > t_{0.025} = 2.12$ (for $\nu = 16$), then we reject H_0.

5. Substituting the numbers we have into the statistic, we obtain:

$$t = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35 < t_{0.025} = 2.12$$

so we cannot reject the null hypothesis that the true correlation coefficient is zero. Note that we use a two-sided test, even though the sample correlation coefficient is positive. This is because if we had followed the steps correctly we wouldn't know what the sample correlation coefficient is before we started, and have no *a priori* expectation of whether it should be positive or negative.

Fisher[1] Z-Transformation:

When the true correlation coefficient is not expected to be zero we cannot assume a symmetric, normal distribution, since $\rho \neq 0$ distributions are skewed. *Fisher's Z transformation* will convert the distribution of r into something that is normally distributed.

$$Z = \frac{1}{2}\ln\left\{\frac{1+r}{1-r}\right\};\ \mu_Z = \frac{1}{2}\ln\left\{\frac{1+\rho_o}{1-\rho_o}\right\};\ \sigma_Z = \frac{1}{\sqrt{N-3}} \qquad (3.17)$$

Here μ_Z is the expected mean of the Z statistic and σ_Z is its standard deviation. Fisher's Z transformation must be used for testing the significance of the difference between sample correlation coefficients drawn from a population acknowledged to be different from zero.

Example: Suppose $N = 21$ and $r = 0.8$. Find the 95% confidence limits on ρ.

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

If Z is normally distributed, then 95% of all values must fall within 1.96 standard deviations

[1] Sir Robert Aylmer Fisher, 1890-1962.

of Z (from two-tailed normal curve test). Therefore, 95% of the time the true mean must fall on the interval,

$$Z-1.96\,\sigma_Z<\mu_Z<Z+1.96\sigma_Z$$
$$0.6366<\mu_Z<1.5606$$

where the expression for the standard deviation of Fisher's Z has been used. These limits on the Z statistic can next be converted into limits on the true correlation.

$$\mu_Z=0.6366=\frac{1}{2}\,ln\left\{\frac{1+\rho}{1-\rho}\right\}\Rightarrow\rho=0.56$$

A handy transformation to get from μ_z to ρ is

$$\rho=\frac{\left(e^{2\mu_z}-1\right)}{\left(e^{2\mu_z}+1\right)}=\frac{\left(e^{\mu_z}-e^{-\mu_z}\right)}{\left(e^{\mu_z}+e^{-\mu_z}\right)}=\tanh(\mu_z)$$

We can state with 95% confidence that the true correlation falls on the interval, $0.56<\rho<0.92$, given that a sample of size 21 yields a sample correlation $r=0.8$.

Tests for the significance of the difference between two non-zero correlation coefficients are made by applying the Z statistic and using the fact that it 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. Then we test for the significance of the difference between the correlation coefficients by first calculating the two Z transformations,

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

From these we can calculate a z statistic, for normal probability curve,

$$z=\frac{Z_1-Z_2-\Delta_{z_1-z_2}}{\sigma_{z_1-z_2}}\text{; where }\Delta_{z_1-z_2}=\mu_{z_1}-\mu_{z_2}$$
$$\text{and }\sigma_{z_1-z_2}=\sqrt{\sigma_{z_1}^2+\sigma_{z_2}^2}=\sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}}\,.$$

3.4 Generalized Normal Equations – Multiple Regression

Suppose we have a predictand, y, that we wish to fit to a group of known predictors x_i, using the linear form:

$$y = a_o + a_1 x_1 + a_2 x_2 + ... + a_n x_n \tag{3.18}$$

In what follows we will assume that the mean has been removed from all of the variables, y and x_i. The least-squares solution for the coefficients a_i requires:

$$\begin{aligned}
a_1 \overline{x_1^2} + a_2 \overline{x_1 x_2} + a_3 \overline{x_1 x_3} + ... + a_n \overline{x_1 x_n} &= \overline{x_1 y} \\
a_1 \overline{x_1 x_2} + a_2 \overline{x_2^2} + a_3 \overline{x_2 x_3} + ... + a_n \overline{x_2 x_n} &= \overline{x_2 y} \\
a_1 \overline{x_1 x_3} + a_2 \overline{x_2 x_3} + a_3 \overline{x_3^2} + ... + a_n \overline{x_3 x_n} &= \overline{x_3 y}
\end{aligned} \tag{3.19}$$

Which can be written in matrix form:

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

or in the subscript notation:

$$\overline{x_i x_j} \, a_j = \overline{x_i y} . \tag{3.20}$$

Note that since we have removed the means of all variables, the overbarred quantities in the above expression are actually covariances.

The covariance is closely related to the variance calculation described in Chapter 1. If x and y are scalars, then the covariance Co is:

$$Co_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \quad \text{so if } \bar{x} = \bar{y} = 0 \text{, } \overline{xy} = Co_{xy}$$

We obtain the correlation by dividing the covariance by the standard deviations of both variables:

$$C_{xy} = \frac{Co_{xy}}{\sigma_x \sigma_y}$$

All 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.

The covariance matrix of x_i is on the left in (3.20) and the covariance vector of the predictors x_i with the predictand y, is on the right. If we further divide each variable by its standard deviation, so that we are working with standardized variables of mean zero and unit variance, then we can write:

$$C_{x_i x_j} a_j = C_{x_i y} \tag{3.21}$$

Where the C's are correlation matrices and vectors, respectively, on the left and right. Canned subroutines are available for doing this sort of linear modeling. The coefficients a_j are obtained by inverting the real, symmetric matrix on the left, and multiplying the inverse times the vector on the right, at least in theory. It may be quicker to use Gaussian elimination or some other numerical technique to solve for the a's. Many of these methods require that $C_{x_i x_j}$ be invertible and not singular. We will discuss below 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 $C_{x_i x_j}$ is singular.

3.4.1 Derivation of 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 (3.19) using matrix algebra. First some definitions:

Let's think of **y** and **a** as row vectors of length *N* and *m*, respectively, and the data matrix **X** as an N x m matrix, where *N* is the sample size and *m* is the number of predictors, x_m.

$$\mathbf{y} = \left[y_1\, y_2\, y_3\, y_4 \ldots . y_N \right]$$

$$\mathbf{a} = \left[a_1\, a_2\, a_3\, a_4 \ldots . a_m \right]$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & & x_{N1} \\ x_{12} & x_{22} & x_{32} & & \\ x_{13} & x_{23} & x_{33} & & \\ & & & & \\ x_{1m} & & & & x_{Nm} \end{bmatrix}$$

Now we can express our desired regression equation as,

$$\hat{\mathbf{y}} = \mathbf{aX}$$

Where we get the vector of predicted values of y, $\hat{\mathbf{y}}$, by multiplying the vector of coefficients **a** times the data matrix **X**. So far, we don’t know what the values of the coefficients, a, should be, but we are going to get these from minimizing the squared error. The error is, $Error = \mathbf{y} - \mathbf{aX}$, and in matrix notation we compute the squared error functional by taking the inner product of the error vector with itself.

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

Here superscript T indicates a transpose of the matrix, and we will use the fact that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. Now we can expand the right hand side, to get,

$$Q = \mathbf{yy}^T - \mathbf{yX}^T\mathbf{a}^T - \mathbf{aXy}^T + \mathbf{aXX}^T\mathbf{a}^T$$

The next step is to differentiate the error functional with respect to the coefficients **a**, to obtain the equation for the values of **a** that minimize the error. By experimenting with differentiating with respect to each individual coefficient, you can convince yourself that the normal rules of algebraic differentiation apply for vectors and matrices, too. Then we can write,

$$\begin{aligned}\frac{\partial Q}{\partial \mathbf{a}} &= 0 - \mathbf{yX}^T - \mathbf{Xy}^T + \mathbf{XX}^T\mathbf{a}^T + \mathbf{aXX}^T \\ &= \left(\mathbf{aXX}^T - \mathbf{yX}^T\right) + \left(\mathbf{aXX}^T - \mathbf{yX}^T\right)^T\end{aligned}$$

Note that the right hand side of the above equation 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**’s that minimizes the error, Q, by setting the right hand side to zero, or

$$\mathbf{aXX}^T = \mathbf{yX}^T \ \ or \ \mathbf{X}^T\mathbf{Xa}^T = \mathbf{Xy}^T$$

from which,

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

Looking back at (3.20) and (3.21)we can see that it is equivalent to $\mathbf{a}^T = \left(\mathbf{XX}^T\right)^{-1}\mathbf{Xy}^T$, since

$$\mathbf{X}\mathbf{X}^T = NC_{x_i x_j} \text{, and } \mathbf{X}\mathbf{y}^T = NC_{x_i y} .$$

In this way we have derived the Normal equations using matrix algebra.

3.4.2 Fourier Analysis as a Regression Problem

You may consider Fourier harmonic analysis to be a special case of a linear least-squares model. In this case the predictors are sines and cosines in some spatial dimension *z*, for example:

$$x_1 = \sin\frac{2\pi z}{L}; \quad x_2 = \cos\frac{2\pi z}{L}; \quad x_3 = \sin\frac{4\pi z}{L}; \quad x_4 = \cos\frac{4\pi z}{L}; \text{ .... } \quad (3.22)$$

If we use the regression approach, this technique will work for unevenly spaced z_i, whereas standard Fourier Transform techniques will not. Since these x_i are an orthogonal basis set, however, if the z_i are evenly spaced, the off-diagonal terms of the covariance matrix will be zero so that the a_j can be obtained algebraically, without the need for a matrix inversion. We have for evenly spaced data and orthogonal predictors:

$$a_j = \frac{\overline{x_j y}}{\overline{x_j^2}}; \quad \text{but } \overline{x_j^2} = \frac{N}{2} \text{ for all } N > 0. \text{ So that} \qquad (3.23)$$

$$a_j = \frac{2}{N}\sum_{i=1}^{N}\left\{ y_i \cdot x_j(z_i)\right\}; \quad \text{Or for example,}$$

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

This also demonstrates that Fourier analysis is optimal in a least-squares sense. 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 6.

Orthogonality:

Vectors: If we have two vectors $\boldsymbol{f}_n$ and $\boldsymbol{g}_n$ of length N, we say that they are orthogonal if their inner product is zero:

$$(\mathbf{f},\mathbf{g}) = \sum_{n=1}^{N} f_n \bullet g_n = 0$$

If N=2, these two vectors define lines that are at 90 degree angles to each other.

Continuous Functions: If we have two functions f(x) and g(x), we say that they are orthogonal on the interval 0<x<L if:

$$(f,g)=\int_0^L f(x)\,g(x)\,dx=0$$

Orthonormal Function Sets: If we have a set of functions $f_n(x)$, we say that it is an orthonormal set of functions if:

$$(f_n,f_m)=\int_0^L f_n(x)\,f_m(x)\,dx=\begin{cases}0 & if\ m\neq n\\ 1 & if\ m=n\end{cases}$$

Thus the inner product of orthogonal vectors or functions is zero, and the inner product of an orthonormal vector or function with itself is one, and with its fellows is zero.

3.4.3 Multiple Regression – How many variables to use?

Multiple regression is the regression of more than two variables. Quite often a single variable is expressed in terms of a set of other variables. For example, future stock market indices are based on present and past values of the stock market and various indices of current economic activity. In this section let's consider standardized variables whose means are zero and whose standard deviations are one.

$$x_i^* = \frac{x_i-\overline{x_i}}{\sigma_{x_i}};\quad y^* = \frac{y-\overline{y}}{\sigma_y} \qquad (3.24)$$

Here x_i is a vector of predictors and y is the predictand. Dropping the stars that indicate standardized variables, for convenience, we can then write the "normal equations" for a linear least-squares fit, following directly from (3.20) as,

$$r\left(x_i,x_j\right)a_i = r\left(x_j,y\right);\qquad \text{where } r = \text{ correlation coefficient} \qquad (3.25)$$

$r\left(x_i,x_j\right)$represents the correlation coefficient between two of the predictors, x_i and x_j. Equation (3.25) is a linear algebra problem which we could have written in subscript notation as,

$$r_{ij}a_i = r_j^y$$

where the right hand side is a correlation vector between y and the predictors x_j.

We can solve for the coefficients a_i, and find expressions for the explained and unexplained variance as we did before for the case of a single predictor. These are again related to the correlation coefficients between the variables. For simplicity, and to gain a little insight, let's consider the case of just 2 predictors. The normal equations can be expanded as,

$$\begin{aligned} r(x_1,x_1)a_1 + r(x_1,x_2)a_2 &= r(x_1,y) \\ r(x_2,x_1)a_1 + r(x_2,x_2)a_2 &= r(x_2,y) \end{aligned} \qquad (3.26)$$

or, since r_{11} = r_{22} = 1.0, and r_{12}= r_{21}, we can write,

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \qquad (3.27)$$

$$\text{So that, } a_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}; \quad a_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2} \qquad (3.28)$$

If $\hat{y}$ is a linear least-squares fit, then we can write the explained and unexplained variance,

$$\begin{gathered} \overline{y'^2} = \overline{(y_i - \hat{y})^2} + \overline{(\hat{y} - \overline{y})^2} \\ \text{Total Variance} = \text{Unexplained Variance} + \text{ Explained Variance} \end{gathered} \qquad (3.29)$$

Can you prove that (3.29) is true? You will need to use $\overline{\hat{y}} = \overline{y},\ y = \hat{y} + \varepsilon,\ and\ \overline{y\varepsilon} = 0$ to do so.

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

$$1 = \frac{Q}{\overline{y'^2}} + R^2 \qquad (3.30)$$

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

$$R^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2\, r_{1y}\, r_{2y}\, r_{12}}{1 - r_{12}^2} \qquad (3.31)$$

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. As an exercise, derive the above formula for R^2.

3.4.4 Stability of Linear Regression Equations

It turns out that in multiple regression, it is necessary to decide how many predictors to use. If too many 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. 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.

Assume that all of the correlation coefficients in the two-predictor cases are equal to 0.5, and use (3.31) to compute R^2. Then with one predictor the fraction of explained variance, R^2, is 0.25. Adding a second predictor raises R^2 to 0.33. Now suppose that r_{2y} is only 0.25, but that the other correlations retain the value of 0.5. Now the fraction of explained variance with two predictors is 0.25, the same as if only x_1 was used. r_{2y} equal to 0.25 is the minimum correlation of x_2 with *y* that is required to improve the multiple R^2 given that x_2 is correlated with x_1. This level of correlation is guaranteed, given that x_1 is correlated with *y* at the 0.5 level and x_1 and x_2 are correlated at the 0.5 level. Adding x_2 as a predictor under these circumstances does not 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.

Minimum Useful Correlation: The minimum useful correlation is the correlation of x_2 with *y* that must exist in order that adding x_2 to the regression will improve the R^2 for *y*.

$$\left|r(x_2,y)\right|_{\min useful} > \left|r(x_1,y)\cdot r(x_1,x_2)\right| \qquad (3.32)$$

We can show this by substituting $r_{2y} = r_{2y_{\min useful}} = r_{1y} \bullet r_{12}$ into the expression for the explained fraction of the variance in the two-predictor case.

$$R^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2\, r_{1y} r_{2y} r_{12}}{1 - r_{12}^2}$$

$$\begin{aligned} R^2 &= \frac{r_{1y}^2 + r_{2y}^2 - 2\, r_{1y}^2 r_{12}^2}{1 - r_{12}^2} \\ &= r_{1y}^2 \end{aligned} \qquad (3.33)$$

Thus we have proven that when r_{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_{12}, that is, the more independent the predictors are, 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, $r_{12} = 0$. Completely dependent predictors, $r_{12} = 1.0$, are useless, only one of these 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 subsequently.

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_i, 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 "explaining" more of the variance of the dependent data set. It is a good idea to use as few predictors as possible. Later we will describe how to pick the optimal set of predictors.

3.4.3 Use of Singular Value Decomposition in Multiple Regression

There are three extremely useful general decompositions of matrices (e.g. Strang, 1988). The first is the triangular factorization, LU, or lower-upper, decomposition, in which a matrix A is written as a product of lower triangular matrix *L* and an upper triangular matrix *U*.

$$A = LU \tag{3.34}$$

L has ones on its diagonal and *U* has the pivots on its diagonal.

A second important decomposition is the QR factorization. Every *m* by *n* matrix *A* with linearly independent columns can be factored into

$$A = QR \tag{3.35}$$

The columns of *Q* are orthonormal, and *R* is upper triangular and invertible. When *m=n*, all matrices are square and Q becomes orthogonal[2]..

A third decomposition, with many powerful applications, and which has therefore fostered a raging growth industry, is the singular value decomposition. It is more powerful than the others, because it places no restrictions on *A*, and its factors have very useful

[2] An orthogonal matrix is defined to be one whose columns are orthonormal (Strang, 1988)

properties.

Singular Value Decomposition: Any *n by m* matrix **A** can be factored into

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \qquad (3.36)$$

where **U** and **V** are orthogonal and Σ is diagonal. The columns of **U** (*m* by *m*) are the eigenvectors of $\mathbf{AA}^T$, and the columns of **V** (*n* by *n*) are the eigenvectors of $\mathbf{A}^T\mathbf{A}$. The *r* singular values on the diagonal of Σ (*n* by *m*) are the square roots of the nonzero eigenvalues of both $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$. .

The regression problem (3.20) can be written as a general matrix problem.

$$Ax = b \qquad (3.37)$$

Here *x* represents the vector of unknown coefficients. The formal solution is easy to write down.

$$x = A^{-1}b \qquad (3.38)$$

But in practice this formal solution is unattainable if the matrix *A* has certain properties.

(1) If the rows of A are dependent, then the equations have no solution. This happens when *b* is outside the column space of *A*. One can remedy this by projecting *b* onto the column space[3] of **A** to produce a new vector *p* and then solving the problem **Ax'=p**.

(2) If **A** has dependent columns, then the solution for *x* is not unique, and we have to add some additional criterion to pick one solution from among the possible ones.

One can force the solution to be unique in general by choosing the solution of **Ax'=p** that has minimum length (the inner product of the coefficient vector with itself is minimized). Singular value decomposition can produce this optimal solution and provide diagnostics of the subspaces of *A* along the way. But let's first do an example to make things more concrete.

Suppose we want to solve $\mathbf{Ax} = \mathbf{b}$, for **x,** where the problem is given by,

[3] The column space of a matrix is the space defined by the columns of the matrix used as coordinate vectors. If ***b*** is not contained within the column space of **A**, then there is no solution to the problem **A*x*=*b***.

$$\begin{bmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

In the column space of **A**, $\begin{bmatrix} a_1 & 0 \\ 0 & a_2 \\ 0 & 0 \end{bmatrix}$ the closest vector to **b** $= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ is $\begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix} = \mathbf{p}$.

So we solve for $\mathbf{Ax} = \mathbf{p}$. But we still have the problem that the solution for **x** is not unique.

The nullspace of **A** is $\begin{bmatrix} 0 \\ 0 \\ x_3 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_4 \end{bmatrix}$, so to get a unique solution we set these components to zero.

Now the problem has become,

$$\begin{bmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \\ 0 \end{bmatrix}$$

and the solution is obviously, $x_1 = \frac{b_1}{a_1}, \; x_2 = \frac{b_2}{a_2}$. This could have been written as the solution to the pseudoinverse problem, $\mathbf{x}^+ = \mathbf{A}^+\mathbf{p}$, or numerically,

$$\begin{bmatrix} b_1/a_1 \\ b_2/a_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/a_1 & 0 & 0 & 0 \\ 0 & 1/a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ 0 \\ 0 \end{bmatrix}$$

One aspect of singular value decomposition is that it provides a complete assessment of the subspaces of the matrix **A** of rank *r*. In particular, the columns of **U** and **V** give the orthonormal bases for all four fundamental subspaces:

The first *r* columns of **U** =	the column space of **A**
The last *m-r* columns of **U** =	the left nullspace of **A**
The first *r* columns of **V** =	the row space of **A**
The last *n-r* columns of **V** =	the nullspace of **A**

In the problem **Ax = b**, the set of all possible combinations of the columns of **A** is the subset of attainable **b**'s. This is the column space of **A**. Since **A** takes every *x* into its column space, a fundamental theorem of linear albegra, then **Ax=b** can be solved only if the **b** specified is in the column space of **A**. The nullspace is the set of all vectors **x** such that **Ax**=0. Of course, the rows and columns of a matrix define different spaces, unless the matrix is symmetric.

Let's consider a simple example of a 2x2 singular matrix **A**= $\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$, which has *m*=*n*=2, and *r*=1. The column space of **A** is all multiples of $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$, a line. The nullspace of **A** is $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$because **A** takes any vector with this direction into zero. The row space is all multiples of $\begin{bmatrix} 1 & 2 \end{bmatrix}$, which is also the column space of $\mathbf{A}^{\mathrm{T}}$. The left null space, defined by $\mathbf{A}^{\mathrm{T}}\mathbf{x} = 0$, is given by $\begin{bmatrix} -3 \\ 1 \end{bmatrix}$. Let's now look at the SVD of **A**.

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 0.316 & -0.948 \\ 0.948 & 0.316 \end{bmatrix} \begin{bmatrix} 7.07 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.447 & 0.894 \\ -0.894 & 0.447 \end{bmatrix}$$

There is one non-zero singular value (7.07) and so the rank of **A** is one. The first column of **U** is the unit vector that describes the column space of **A**, and the second column of **U** is the left nullspace, normalized, and so forth.

So as you can see, the SVD provides lots of important information about a matrix, and much of this is directly applicable to the problem **Ax=b**. First of all we need to condition **b** by projecting it onto the column space of **A**, which is given to us from SVD. Second, we need to find the shortest of all possible solutions if the solution is not unique. Any vector **x** can be split into its rowspace component $\mathbf{x'}_{\mathbf{r}}$ and its nullspace component $\mathbf{x'}_{n}$, so that $\mathbf{x'} = \mathbf{x'}_{\mathbf{r}} + \mathbf{x'}_{\mathbf{n}}$. The row component provides a solution to $\mathbf{Ax'}_{\mathbf{r}} = \mathbf{p}$, and $\mathbf{Ax'}_{\mathbf{n}} = \mathbf{0}$. Since the row space and null space vectors are orthogonal, the shortest solution vector is achieved by setting $\mathbf{x'}_{n} = 0$. All of this is handled pretty much automatically in SVD. We define the pseudoinverse as the matrix which finds the shortest solution vector $\mathbf{x}^{+}$ within the column space of **A**, $\mathbf{x}^{+} = \mathbf{A}^{+}\mathbf{b}$. This pseudoinverse is given in terms of the SVD, by

$$\mathbf{A}^{+} = \mathbf{V}\Sigma^{+}\mathbf{U}^{T} \qquad (3.39)$$

where Σ^{+} is (*n* by *m*) and is formed by putting the reciprocals of the singular values on the diagonal. It works.

Exercises:

3.1) Show that (3.8) is true.

3.2) Show that (3.29) is true.

3.3) Show that (3.30) and (3.31) are true.

3.4) A sample of 32 pairs measurements gives a correlation between temperature and precipitation of -0.5. Can a null hypothesis of zero correlation be rejected at the 95% level?

3.5) What are the 95% confidence limits on the true correlation if the sample correlation is 0.7 on the basis of 50 pairs of measurements.

3.6) What is the minimum useful correlation of *y* with *z*, if *y* is correlated with *x* at the 0.6 level, and *x* is correlated with *z* at the 0.7 level.

3.7) Use Matlab or some other canned program to find the rank and all the subspaces of the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 6 & 7 \end{bmatrix}$$

References:

Dillon, W. R., and Goldstein, M., 1984, Multivariate Analysis: Methods and Applications: Wiley, 587pp.

Draper, N. R., and Smith, H., 1966, *Applied Regression Analysis*: New York, John Wiley and Sons, 407 p.

Huff, D., 1954, *How to Lie with Statistics*: New York, W.W. Norton & Company, 142 p.

Larson, R. L. and M. L. Marx, 1986: *An Introduction to Mathematical Statistics and its Applications.* 2nd ed. Prentice-Hall, 630 pp.

Noble, B., and Daniel, J. W., 1988, *Applied Linear Algebra*: Englewood Cliffs, Prentice-Hall, 521 p.

Panofsky, H. A., and Brier, G. W., 1968, *Some Applications of Statistics to Meteorology*: University Park, Pennsylvania State University, 224 p.

Spiegel, M. R., J. Schiller, and R.A. Srinivasan, 2000, *Probability and Statistics*: Schaum's Outline Series in Mathematics: New York, McGraw Hill, 408 p. (available cheap at Amazon.)

Strang, G., 1988, *Linear Algebra and Its Applications*: San Diego, Harcourt, Brace Jovanovich, 505 p.

Wilks, D. S., 1995, *Statisical Methods in the Atmospheric Sciences*: International Geophysics Series, v. 59: San Diego, Academic Press, 467 p.