Correlation and Regression

Damian Cox

August 13, 2019

Contents

1 Introduction

Correlation and Regression, at its simplest, is a statistical analysis is very often concerned with the question of how one quantity depends on another. In this section we will look at the concepts of correlation and linear regression, the formal statement of dependency of one variable on one or several explanatory variables. It is one of the most widely used and powerful tools in statistics.

2 Definitions

First we need to look at the concept of a joint distribution of two random variables, which underpins our discussion on this topic.

2.1 Joint Probability Distributions

We start with the definition of this idea.

2.1.1 Definition: Joint Probability Distribution Function

Let X and Y be two random variables. They are said to have a joint probability density function if we can write the following: Let f be a function of two real variables x and y such that $f(x, y) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \mathrm{d}y \ \mathrm{d}x = 1.$$

Then X and Y have a joint distribution if:

$$P[X \leq a, Y \leq b] = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) \mathrm{d}y \ \mathrm{d}x.$$

The function f at a point (x, y) has the interpretation that it shows how dense the probability is at that point, so that

$$f(x, y)\delta x \delta y$$

is approximately the probability that the variables X, Y will throw out values in the small rectangle around point (x, y) of size $\delta x \delta y$. We will also understand that

$$\frac{\partial^2}{\partial x \partial y} P[X \leq x, Y \leq y] = f(x, y).$$

This reflects the idea of the probability density function as the multivariate rate of change of the cumulative density function.

2.1.2 An Example of a Joint Probability Distribution Function

A point is picked at random inside a circular disc with radius a. Let R be the random variable of the distance of the point to the centre and let Θ be the random variable of the angle this line makes with a given horizontal axis. We will construct the joint probability density function for the two random variables (R, Θ). This is done by seeing that the probability of the point falling in an area A, a subset of the circle, is the ratio $A/\pi a^2$. Then the probability

$$P[R \leq r, \Theta \leq \theta]$$

is the probability that the point (R, Θ) is in a disc segment with radius r and subtending an angle of θ. The area of this disc segment is πr^2 times the ratio $\theta/2\pi$, which is then divided by πa^2 to find the required ratio of areas:

$$P[R \leq r, \Theta \leq \theta] = \frac{\theta r^2}{2\pi a^2}.$$

We can now calculate the density function by differentiating with respect to the two variables r and θ:

$$f(r, \theta) = \frac{r}{\pi a^2}, \text{ when } 0 < r < a \text{ and } 0 < \theta < 2\pi.$$

2.2 Marginal Probability Densities

We are discussing two random variables X and Y, with a joint probability density function

$$P[X \leq a, Y \leq b] = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) \mathrm{d}y \; \mathrm{d}x,$$

where f is a function of two real variables x and y such that $f(x,y) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\ \mathrm{d}x = 1.$$

The variables X and Y will have their own probability density functions; they can be written in terms of f:

$$P[X \leq a] = \lim b \to \infty P[X \leq a, Y \leq b] = \int_{-\infty}^{a}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\ \mathrm{d}x,$$

so it follows that the density function for X, denoted f_X, is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y.$$

Equivalently for Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}x.$$

2.2.1 Example: Marginal Density

A point (X,Y) is chosen at random in the disc D of radius a. Find the marginal density of the coordinate X. [It is worth noting how the coordinates of the random point are treated as two random variables.]

This problem is tackled by looking at the ratio of areas. Since any point on the circle maybe chosen with equal probability, the joint probability density function is

$$f(x,y) = \frac{1}{\pi a^2},\ \text{when}(x,y) \in D$$

and $f(x,y) = 0$ otherwise. Using the equation

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y,$$

with the limits determined by the edge of the circle:

$$f_X(x) = \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \frac{1}{\pi a^2}\mathrm{d}y = \frac{2}{\pi a^2}\sqrt{a^2-x^2}.$$

Thus we can say that the marginal distribution for X is:

$$f_X(x) = \frac{2}{\pi a^2}\sqrt{a^2 - x^2},\ \text{when}|x| < a$$

and $f_X(x) = 0$ otherwise.

2.2.2 Marginal Distributions for the Point in a Circle

For a point picked at random inside a circular disc with radius a, recall we defined R as the random variable of the distance of the point to the centre and Θ as the random variable of the angle this line makes with a given horizontal axis. The joint probability density function for the two random variables (R, Θ) was

$$P[R \leq r, \Theta \leq \theta] = \frac{\theta r^2}{2\pi a^2}.$$

The joint probability density function was

$$f(r, \theta) = \frac{r}{\pi a^2},\ \text{when } 0 < r < a \text{ and } 0 < \theta < 2\pi.$$

Then for variable R:

$$f_R(r) = \int_0^{2\pi} \frac{r}{\pi a^2} \mathrm{d}\theta = \frac{2r}{a^2},$$

for $0 \leq r \leq a$. For the variable Θ,

$$f_\Theta(\theta) = \int_0^{a} \frac{r}{\pi a^2} \mathrm{d}r = \frac{1}{2\pi}.$$

Thus the angle variable is a uniform distribution from 0 to 2π.

2.3 Independent Jointly Distributed Random Variables

Let X and Y be two random variables. The variable X may be said to be independent of Y if the probabilities for X, in other words the probability density function of X, are independent of the values of Y. It can be proved that if the two variables are independent, then

$$f(x, y) = f_X(x) f_Y(y).$$

This is an if-and-only- if relation; if we produce a joint probability mass function and it follows this form, then we conclude that the two variables are independent. This result also means that if X and Y are independent:

$$E[XY] = E[X]E[Y].$$

For an illustrative example, look at the case above of a point picked at random inside a circular disc with radius a. The joint probability density function was

$$f(r, \theta) = \frac{r}{\pi a^2}, \text{ when } 0 < r < a \text{ and } 0 < \theta < 2\pi.$$

Now compare the marginal probability densities:

$$f_R(r) = \frac{2r}{a^2}, \; f_\Theta(\theta) = \frac{1}{2\pi}.$$

Clearly $f(r, \theta) = f_R(r) f_\Theta(\theta)$ so the angle and radial distance variables are independent.

2.3.1 Example: Minimum of Two Exponential Variables

Let X and Y be two variables which follow the exponential distribution with parameters a and b respectively, so that

$$P[X < x] = 1 - e^{-ax}, P[Y < y] = 1 - e^{-by}.$$

We will produce a probability distribution for Z = min(X,Y) as follows.

Looking at the event $Z < z$, it can be seen that for the minimum to be above than a given value z, both X and Y must be above this value

z. Thus $Z < z$ is the converse of the event $X > z$ and $Y > z$. Applying the rules of probability leads to the statement:

$$P[Z < z] = 1 - P[X > z \text{ and } Y > z].$$

It follows from the distributions of X and Y that

$$P[X > z] = e^{-az}, P[Y > z] = e^{-bz}.$$

Therefore

$$P[Z < z] = 1 - P[X > z \text{ and } Y > z] = 1 - e^{-az}e^{-bz} = 1 - e^{-(a+b)z}.$$

Therefore the variable Z = min(X,Y) follows the exponential distribution with parameter $a + b$.

2.3.2 The Sum of Squares of Standard Normal Distribution

Let X and Y be two independent standard normal variables. We will produce a distribution for the quantity

$$Q = X^2 + Y^2.$$

The probability density functions for both variables, that of the standard normal variable, are:

$$f_X(x) = C\exp(-\tfrac{1}{2}x^2), f_Y(y) = C\exp(-\tfrac{1}{2}y^2), \text{ where } C = \tfrac{1}{\sqrt{2\pi}}.$$

This means that the joint probability distribution is then

$$f(x, y) = \frac{1}{2\pi}\exp[-\tfrac{1}{2}(x^2 + y^2)].$$

This means that

$$P[X^2 + Y^2 < q] = \frac{1}{2\pi}\int\int_D \exp[-\tfrac{1}{2}(x^2 + y^2)]\mathrm{d}x\mathrm{d}y.$$

The area D over which the integration is carried out is the region where $x^2 + y^2 < q$, in other words it represents a disc of radius $\sqrt{q}$. This suggests a way to carry out the integral; change to the standard polar coordinates (r, θ). The change of variable means the differential area becomes

$$\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta.$$

With this change in variables, the integral becomes

$$\int\int_D \exp[-\tfrac{1}{2}(x^2 + y^2)]\mathrm{d}x\mathrm{d}y = \int_0^{2\pi} \int_0^{\sqrt{q}} \exp[-\tfrac{1}{2}r^2]r\mathrm{d}r\mathrm{d}\theta.$$

The integration with respect to the variable θ is trivial and that for r is carried out with the substitution $u = r^2$. These are brought together for the final result:

$$P[X^2 + Y^2 < q] = 1 - e^{-\frac{1}{2}q}.$$

Thus we have shown that the sum of squares of two standard Normal variables has the exponential distribution with parameter $\frac{1}{2}$.

3 Covariance, Correlation and Regression

Following our brief look at joint probability densities, it is now possible to look at the concept of covariance and the related idea of correlation between random variables. This will then lead on to the concept of linear regression between random variables.

3.1 Covariance and Correlation

This idea naturally arises when we look at the properties of the variance of a random variable. We know that for two random variables X

and Y defined on the same sample space, the definition of expected value means:

$$E[X+Y] = E[X] + E[Y].$$

Recall that

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2.$$

Let us now look at the variance of the sum of two variables. Applying the definition above with some algebra shows that

$$Var[X+Y] = Var[X] + 2(E[XY] - E[X]E[Y]) + Var[Y].$$

Thus the variance of the sum of two variables is the sum of the two variances plus twice the quantity $E[XY] - E[X]E[Y]$. This quantity is defined as the covariance of the two variables.

3.1.1 Definition: Covariance

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

The first noteworthy thing to say about this quantity is that if two variables are independent, so that

$$E[XY] = E[X]E[Y],$$

then $Cov[X,Y] = 0$. The converse is not necessarily true except in certain cases; here is an example.

Let X be a standard normal variable and define another random variable by $Y = X^2$. These two variables are clearly not independent. The following results for these variables are proved using integration by parts:

1. For X we know that $E[X] = 0$ and $E[X^2] = 1$.

2. For Y, $E[X^2] = 1$ so $E[Y] = 1$. It is important that we know this is finite.

3. $E[X^3] = 0$.

It then follows that

$$Cov[X, Y] = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2] = 0.$$

Therefore the covariance is 0 but the variables are clearly dependent.

3.1.2 Definition: The Correlation Coefficient

Like the variance, the covariance has the disadvantage that it will be in squared units of the variables X and Y. The measure used to get around this is the correlation coefficient ρ, defined as:

$$\rho[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}.$$

This is essentially the covariance divided by the two standard deviations. It has no units and it can be proved that

$$-1 \leq \rho[X, Y] \leq 1.$$

This is done by using the properties of var[], cov[] as functional operators and then calculating the variance of the quantity Z, given by:

$$Z = \frac{Y}{Var[Y]} - \rho[X, Y]\frac{X}{Var[X]}.$$

3.2 Linear regression

Suppose that X and Y are two random variables with Y depending on X. It is required to find the coefficients α and β so that we can predict

Y from X with a linear equation $Y = \alpha + \beta X$. We mean to minimise the squared error of this prediction, in other words the quantity:

$$E[(Y - (\alpha + \beta X))^2].$$

This will be the least squares regression equation linking X and Y.

The estimates of quantities α and β produced will be denoted a and b respectively. So now values of a and b will be estimated by minimising

$$Q = E[(Y - a - bX)^2]$$

as a function of a and b. To carry this out, some algebra on Q leads to the expression:

$$Q = E[Y^2] - 2a\mu_Y - 2bE[XY] - 2ab\mu_X + b^2E[X^2] + a^2.$$

Differentiating the expression for Q with respect to a :

$$\frac{\partial Q}{\partial a} = 0 - 2\mu_Y - 2b\mu_X + 2a.$$

Setting this result to 0 gives:

$$\mu_Y = a + b\mu_X.$$

Repeating this differentiation with b gives

$$\frac{\partial Q}{\partial b} = -2E[XY] - 2a\mu_X + 2bE[X^2].$$

Now replace the expected values with the standard deviations and covariance values:

$$E[X^2] = \sigma_X^2 + \mu_X^2,\ E[XY] = cov[X,Y] + \mu_X\mu_Y = \rho_{XY}\sigma_X\sigma_Y + \mu_X\mu_Y.$$

This allows us to solve for the coefficients so that

$$b = \rho_{XY}\frac{\sigma_Y}{\sigma_X}.$$

In summary we have the equations

$$\mu_Y = a + b\mu_X \text{ and } b = \frac{\rho_{XY}}{\sigma_X \sigma_Y}.$$

These are the random variable versions of the standard least squares regression equations. When the values are put in for the expected value of the squared difference we get the residual variance:

$$E[(Y - a - bX)^2] = \sigma_Y^2 (1 - \rho_{XY}^2).$$

This is proved by subtracting the equation

$$\mu_Y = a + b\mu_X$$

from both sides of the expression for the residual

$$Y - (\alpha + \beta X).$$

3.3 The Bivariate Normal Distribution

Let ρ be a real number with

$$-1 \leq \rho \leq 1.$$

Two random variables X and Y are said to have a bivariate normal distribution if they have the following joint probability distribution:

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2} \frac{x^2 - 2\rho xy + y^2}{1 - \rho^2} \right)$$

We will see that the number ρ may be interpreted as the correlation of X and Y.

3.3.1 The Marginal Densities

For the bivariate distribution shown above, we will find the two marginal densities for X and Y. First, write

$$x^2 - 2\rho xy + y^2 = x^2 + \frac{(y - \rho x)^2}{1 - \rho^2}.$$

The joint distribution can then be written as

$$f(x,y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right).$$

The second term in this product, as a function of y, has the same form as a normal distribution density function with mean ρx and variance $1-\rho^2$. Therefore when it is integrated over y the result is 1, so that

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2).$$

This is the PDF of the standard normal distribution. By symmetry the variable Y will have the same standard normal distribution. This means we know the means and variances of X and Y, which leads to the result

$$\rho[X,Y] = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} = \rho.$$

To establish this, recall the definition of f in the following form:

$$f(x,y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right).$$

It can be quickly established, given the expected values and variance of the normal random variables X and Y, that

$$cov[X,Y] = E[XY].$$

This expected value is then

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right) \mathrm{d}y\mathrm{d}x.$$

Bring the x terms to the front to allow integration wrt y first:

$$E[XY] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right) \mathrm{d}y\mathrm{d}x.$$

The interior integral with respect to variable y is

$$\int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right) \mathrm{d}y.$$

This is now the expected value of a normally distributed random variable with mean ρx and variance $1 - \rho^2$. The answer is therefore ρx . Then:

$$E[XY] = \int_{-\infty}^{\infty} \rho x^2 \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)\mathrm{d}x = \rho E[X^2] = \rho.$$

Thus the parameter ρ is the correlation between X and Y. This result can be generalised for a general bivariate normal distribution.

The significance of the result is that we have established that the bivariate normal distribution, which has the required marginal normal distributions, has a parameter ρ which is the correlation between the two variables. We will establish some important results around this idea.

3.4 Tests for Correlation

If sample data is available, that is, a list of paired numbers (Xi, Yi) which are data values for the random variables X and Y, it may be possible to carry out a NHBST on a test value for the correlation coefficient.

3.4.1 The Sample Correlation Coefficient

Recall the definition of the sample correlation coefficient r in the following equations:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i X_i^2 - n\bar{X}^2}\sqrt{\sum_i Y_i^2 - n\bar{Y}^2}}.$$

This may be viewed as an estimate of ρ, where

$$\rho[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}} = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var[X]Var[Y]}}.$$

Û The links with the two variations of the sample value are clear. There are a few ways to conduct a test on this r as a sample value of ρ.

3.4.2 Permutation method

This first method of testing a sample value of the correlation coefficient is based on the idea of breaking the link within each pair of values X_i, Y_i, then seeing how the calculated value of r looks in this setting. This means we carry out the following steps:

1. Our Null Hypothesis is that the correlation is close to 0.
2. Choose a level of significance α.
3. For the list of data values, calculate the actual sample value of r, call it r_0.
4. Generate a permutation of the numbers 1 to n, call it p, so that the number j is mapped to $p(j)$.
5. Recalculate r for the pairs $X_i, Y_{p(i)}$. Call this r_p. Repeat with every possible permutation of the n numbers, of which there will be $n!$.
6. We note the proportion of values r_p which are further away from 0 compared to r_0, that is to say, if $r_0 > 0$, the proportion above r_0, or the proportion below for the negative r_0.
7. If this proportion is below α, we can reject the Null Hypothesis.

Essentially this procedure creates an explicit distribution for the value of r and we can compare the actual value with this distribution. It is also possible to generate a sample of values from a sample of permutations, saving computation time. Recall that for n data points, the number of permutations will be $n!$, a number that becomes very large very quickly. For example, for the case of 10 data points, the number of permutations is 3.63×10^6. Thus sampling the values of r produced by these permutations would be sufficient.

3.4.3 Bootstrap

This method is an extension of the previous idea. Here, instead of calculating a value r_p for every permutation p and identifying the implicit distribution from there, the alternative is to select, for each X_i, any one value of Y from the full list of the Y_j. This is done again from scratch each time. This may be regarded as selecting the Y_j for an X_i without replacement. Therefore we are drawing upon an even larger list of possible values of r, n^n. Typically we look at the proportion of values r_p further away from 0 than r_0, as was the case with the permutation method. This method is computationally much simpler, as it does not require the generation of a permutation.

3.4.4 The Student t Distribution

This is the most well known approach to testing a correlation. The starting point is the Null Hypothesis that the two variables X and Y follow a bivariate normal distribution with $\rho = 0$. Under this Null Hypothesis it can be shown that the quantity

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

follows the t distribution with $n - 2$ degrees of freedom. The critical values of t are used to reject the Null hypothesis or not. The alternative approach is to invert the equation mapping r to t, giving

$$r = \frac{t}{\sqrt{n-2+t^2}},$$

to generate critical values for r.

3.4.5 Fishers Transformation

The test on a correlation most widely used in practice comes from Fisher; let F be the function defined for $||x|| \leq 1$ by

$$F(x) = \tanh^{-1}(x) = \frac{1}{2}\log_e\left(\frac{1+x}{1-x}\right),$$

where

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

If r is the sample correlation coefficient from n values of two variables X and Y with a bivariate normal distribution with correlation ρ, then the quantity $F(r)$ follows the normal distribution with mean $F(\rho)$ and standard deviation $\frac{1}{\sqrt{n-3}}$, for large n. Alternatively, the quantity

$$Z = [F(r) - F(\rho)]\sqrt{n-3}$$

follows the standard normal distribution.

This method also has the advantage of producing a confidence interval for the correlation coefficient. We use the usual notation of z_β as the critical value for the standard normal distribution for probability β. For the level of significance α, the confidence interval is given by

$$F(r) - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}, F(r) - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}.$$

Thus

$$P[\rho \in [F^{-1}\left(F(r) - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right), F^{-1}\left(F(r) + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)] = 1 - \alpha.$$

Thus we have a confidence interval for the correlation coefficient.

4 Multiple Regression

We will now return and expand on the subject of linear regression, looking in detail at the case of several explanatory variables. To explain a dependent variable y in terms of a list of p independent variables x_1 to x_p, we write

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

With n data points, we have n values of y and n values of each one of the variables x_1 to x_p, so then to find the coefficients α, β_1 to β_p, we are now solving n linear equations for $p+1$ unknowns. Typically we can assume that $n > p+1$ and indeed is usually much larger.

4.1 The Least Squares Estimate

Using Y_i as the i-th value of y, let X_{ij} be the i-th value of variable x_j, so the data consists of $p+1$ lists of n numbers:

$$X_{1,1}, X_{1,2}, \ldots, X_{1,p},$$

$$X_{2,1}, X_{2,2}, \ldots, X_{2,p},$$

to

$$X_{n,1}, X_{n,2}, \ldots, X_{n,p}.$$

The regression model is now the n equations

$$Yj = \beta_1 X_{j1} + \beta_2 X_{j2} + \ldots + \beta_p X_{jp} + \alpha,$$

for $j = 1$ to n. We will write this set of equations for the coefficients in matrix form;

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} & 1 \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,p} & 1 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \\ \alpha \end{pmatrix}.$$

Set β and Y to be the column matrices

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \\ \alpha \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}$$

and set X to be the matrix

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} & 1 \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,p} & 1 \end{pmatrix}.$$

Matrix X is known as the design matrix as it contains the observed data for the independent explanatory variables.

Then the set of equations for the coefficients can be written as

$$X\beta = Y.$$

Inherent in these ideas was the model we are hoping to fit to the data, which can be written as

$$Y = X\beta + \epsilon,$$

where β is the column matrix of parameters and ϵ is a column matrix of errors, that is to say, noise or residuals.

The ordinary least squares (OLS) solution to the equation for the coefficients will be denoted as b. It is found by minimising $||\epsilon||^2$, where $||.||$ is used for the magnitude of a vector. This leads to the pseudo-inverse solution:

$$b = (X^T X)^{-1} X^T Y.$$

We will see this process soon.

This equation could be decomposed into a set of equations for the individual parameters, as was done with the simple regression case of one explanatory variable, but finding the inverse of the term $X^T X$ would be excessively complicated once there are more than two explanatory variables. The normal procedure when using these equations is to calculate $X^T Y$ and $X^T X$ and use an inverting algorithm to find $(X^T X)^{-1}$ or some other method for solving the set of $p + 1$ equations

$$(X^T X)b = X^T Y.$$

4.2 The Multiple Regression Model

The model underpinning the concept of multiple regression is described fully here.

1. Y is an n by 1 vector of identically distributed observable *response* random variables.

2. X is an n by p *design* matrix of random variables with full rank p, where each column i of X, $(X_{i,j})$ is identically distributed.

3. β is a p by 1 vector of unknown parameters and

4. ϵ is an n by 1 vector of random variables, the *random error*, where $\sigma^2 = var[\epsilon_i]$ for all i.

Then the multiple regression model is

$$Y = X\beta + \epsilon.$$

The matrix X is, as noted above, the *design* matrix, its random elements will be the values thrown up by each of the variables we are using to explain Y, the values we have for the response variable. This is effectively the data that we are trying to fit to the model. Saying it has full rank p means its columns are linearly independent; this means we have distinct data points. One of the columns of X may be all 1's, indicating an intercept. This is a slightly different way to approach this; before we assumed an intercept and talked of $p + 1$ columns.

As noted above, it is assumed that the elements of ϵ are independent and identically distributed with mean 0 and variance σ^2. We also assume that if X is random, ϵ and X are independent. The elements of ϵ are unknown.

Within this conceptual framework, set b to be the Ordinary Least Squares estimator for β, given by

$$b = (X^T X)^{-1} X^T Y.$$

This equation gives a vector of parameters b in terms of the matrices of random variables X and Y. It is therefore a random variable; it only takes on a value once the data in X and the responses in Y have been gathered.

4.2.1 Properties of the Multiple Regression Model

We will now look closer at some properties of this construction. We will prove a sequence of statements about the model which will show us why the pseudo-inverse solution is a least squares solution.

1. Recall we will typically not know the ϵ in the model

   $$Y = X\beta + \epsilon.$$

   We will estimate a vector of errors, or residuals, with the difference between Y and Xb, the predicted response:

   $$e = Y - Xb.$$

   This vector of residuals e is perpendicular to Xb. To prove this, it is firstly readily shown, by multiplication across by X^T, that $X^T e = 0$, where 0 represents the p by 1 vector of zeros. Now look at the dot product of e and Xb, which can be written as $(Xb)^T e$. This will be

   $$(Xb)^T(Y - Xb) = (Xb)^T Y - (Xb)^T Xb = b^T X^T Y - b^T X^T Xb.$$

   We know that $X^T Xb = X^T Y$, so the dot product is now

   $$b^T X^T Y - b^T X^T Xb = b^T X^T Y - b^T X^T Y.$$

   Looking at the two terms in the subtraction, each is the transpose of the other. But since both are scalars, this means that they are equal, so that $(Xb)^T e = 0$, proving the result. An important corollary of this is that the magnitude of Y can be written as

   $$||Y||^2 = ||Xb||^2 + ||e||^2.$$

2. Now we will show that the vector b is the vector of parameters which minimises the magnitude of the vector of residuals e. Define the function Q as the magnitude of $Y - Xg$ for any p by 1 vector g, so

$$Q(g) = ||Y - Xg||^2.$$

We therefore need to show that Q is minimised when $g = b$. The first step is to write

$$Q(g) = ||Y - Xg||^2 = (Y - Xg)^T(Y - Xg).$$

Then

$$Y - Xg = Y - Xb + X(b - g) = e + X(b - g).$$

Substitute this in:

$$Q(g) = (e+X(b-g))^T(e+X(b-g)) = (e^T+(b-g)^TX^T)(e+X(b-g)) =$$

$$= e^Te + e^TX(b - g) + (b - g)^TX^Te + (b - g)^TX^TX(b - g).$$

The two cross terms are 0, since they include the term X^Te or its transpose, so we can say that

$$Q(g) = e^Te + (b - g)^TX^TX(b - g).$$

For a given value of b calculated from X and Y, the term e^Te is fixed. The second term is a quadratic form giving a positive number since X^TX is a symmetric matrix. It has a clear minimum at 0 when $g = b$.

4.3 Unbiased Estimates

We will now briefly return to our study of random variables to look at the idea of bias in an estimate. Let ν be a parameter for a random variable, let D denote a sample of values for the random variable and let $H(D)$ be an estimate for this parameter calculated from the data. The estimate $H(D)$ is said to be unbiased if

$$E[H(D) - \nu] = 0.$$

We illustrate this with two examples on the mean and variance.

4.3.1 Example: Why $n - 1$?

Let us look at estimates for the mean and standard deviation. For a data set $x_1, x_2, \ldots, x_n$, the estimate for the mean is

$$\bar{X} = \frac{\sum_i X_i}{n}.$$

To determine whether or not this is an unbiased estimate of the mean μ, use the linearity of the expected value:

$$E[\bar{X} - \mu] = \frac{1}{n}\sum_i E[X_i] - \mu = 0.$$

Now look at the estimate for the variance. We will first look at the estimate

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}.$$

Some algebra will show that

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Thus the above estimate for the sample variance can be written as

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n} = \frac{\sum_i (X_i - \mu)^2}{n} - (\bar{X} - \mu)^2.$$

Therefore the expected value of this estimate is

$$E[S^2] = \frac{1}{n}E[\sum_i (X_i - \mu)^2] - E[(\bar{X} - \mu)^2].$$

The first term here is the variance itself. The second term is the variance of the mean and the central limit theorem tells us that this is $\frac{\sigma^2}{n}$, exactly if the variables X_i are normally distributed and in the limit otherwise. Therefore

$$E[S^2] = 1 - \frac{1}{n^2}\sigma^2.$$

Thus the quantity

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}$$

underestimates the variance.

As a corollary it follows that

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

is an unbiased estimate of σ^2.

4.3.2 OLS is unbiased

Let us look at the OLS estimator again and examine the question of the bias of our estimate for β. Recall the regression model is

$$Y = X\beta + \epsilon.$$

We will now prove the extremely important result that the vector $b = (X^TX)^{-1}X^TY$ is an unbiased estimator of the unknown vector parameter β: $E(b|X) = \beta$. Let ϵ be the p by 1 vector of random error. The model says $Y = X\beta + \epsilon$, so from the definition of b, it follows that

$$b = (X^TX)^{-1}X^T(X\beta+\epsilon) = (X^TX)^{-1}X^TX\beta+(X^TX)^{-1}X^T\epsilon = \beta+(X^TX)^{-1}X^T\epsilon.$$

Now the expected value is

$$E(b|X) = E(\beta + (X^TX)^{-1}X^T\epsilon|X) = \beta + (X^TX)^{-1}X^TE[\epsilon|X].$$

The vector ϵ is independent of X and each element has mean 0, so the last term is 0, completing the proof.

4.3.3 The Covariance Matrix

For a random vector u, we define its covariance matrix as

$$cov[u] = E[(u - E[u])(u - E[u])^T].$$

For an n by 1 vector u the covariance will be an n by n matrix. This definition can be written as

$$cov[u] = E[(u - E[u])(u^T - E[u]^T)].$$

In an analogous way to the scalar version, we can multiply this out to see that

$$cov[u] = E[uu^T] - E[u]E[u]^T.$$

4.3.4 The Covariance of the Least Squares Estimator

Let us look at the covariance of our estimator b. Recall ϵ is the p by 1 vector of random error in the model

$$Y = X\beta + \epsilon.$$

We already that b can be written as

$$b = \beta + (X^T X)^{-1} X^T \epsilon.$$

So then it immediately follows that

$$cov[b|X] = cov[\beta + (X^T X)^{-1} X^T \epsilon | X].$$

Noting that β is an unknown constant and using the form of the covariance matrix, this is

$$cov[b|X] = (X^T X)^{-1} X^T cov[\epsilon|X] X (X^T X)^{-1}.$$

To identify $cov[\epsilon|X]$, since it is independent of X it follows that we can write

$$cov[\epsilon|X] = cov[\epsilon] = E[\epsilon\epsilon^T] - E[\epsilon]E[\epsilon]^T.$$

The covariance matrix of a random vector with independent elements will be the variance of the elements times the identity matrix, so

$$E[\epsilon\epsilon^T] - E[\epsilon]E[\epsilon]^T = \sigma^2 I - 0,$$

where I is the identity matrix. So this means that

$$cov[b|X] = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Thus the covariance matrix of b, given X, depends on the design matrix X, the data, and the unknown variance σ^2.

We will not go into the detail here, but it can be shown that the expected value of $||e||^2$, the sum of the squares of the residuals e, in other words given X, is $\sigma^2(n-p)$. Therefore:

$$E\left[\frac{1}{n-p}\sum_i e_i^2\right] = \sigma^2.$$

It then follows that the quantity

$$\hat{S}^2 = \frac{1}{n-p}\sum_i e_i^2$$

is an unbiased estimator of σ^2. Bringing these results together, the unbiased estimate of the covariance of b is then

$$c\bar{o}v[b|X] = \left(\frac{1}{n-p}\sum_i e_i^2\right)(X^T X)^{-1}.$$

where the bar over the *cov* symbol indicates it has been estimated.

4.4 Explained Variance

We will now look at the fitted model and look at how the variance of the response variable may be deconstructed into two components which will attempt to provide more information about what underlies the model and how well it is performing. We are addressing the case when there is an intercept. Firstly, we would normally define the sample variance S_Y of the list of variables Y_i as

$$S_Y^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n}.$$

The variance of the error term e can be written similarly:

$$S_e^2 = \frac{e_i^2}{n}.$$

Define the number R by the relation

$$R^2 = 1 - \frac{S_e^2}{S_Y^2}.$$

This number R identifies the fraction of the variance that has been explained by the model

$$Y = X\beta + \epsilon$$

and the estimator

$$Y = Xb + e.$$

However, it relies on the idea that the unexplained, residual, variance of e and the remaining explained variance, that of Xb, are complementary. The definition of the variance of Xb, the predicted values of Y, will be defined with exactly the same expression as for Y or e. We will now investigate the link between the number R and the predicted vs unexplained variance.

Let u be an n by 1 vector with all entries 1, so it is identical to the column of X indicating the presence of an intercept. The variance of Y is now equal to

$$\frac{||Y - \bar{Y}u||^2}{n}.$$

Consider the following:

$$Y - \bar{Y}u = Xb - \bar{Y}u + e.$$

We know that e is perpendicular to the vector space of the columns of X and to Xb. This means that e is perpendicular to u, since the vector u is in fact one of the columns of X. Therefore we can say that

$$||Y - \bar{Y}u||^2 = ||Xb - \bar{Y}u||^2 + ||e||^2.$$

Since the means of e are 0, we can say that

$$\bar{Y} = \bar{X b} = \bar{X}b,$$

where the quantity $\bar{X}$ is a row vector of means of the columns of X. Therefore

$$||Y - \bar{Y}u||^2 = ||Xb - \bar{X}bu||^2 + ||e||^2.$$

The second term is in fact the variance of the columns of Xb, so dividing through by n we can say that

$$Var[Y] = Var[Xb] + Var[e].$$

Thus the quantity R can now be written as

$$R^2 = 1 - \frac{S_e^2}{S_Y^2} = \frac{Var[Xb]}{Var[Y]}.$$

Therefore R is the ratio of the explained standard deviation to the total standard deviation.

4.5 The F-test on the Regression Model

We will now look at a very common test applied to regression models which will yield significant information about the model. We are in the context of our regression model;

$$Y = X\beta + \epsilon,$$

with all terms as before. Finally, recall that the vector b is the ordinary least squares estimator for β, given by

$$b = (X^T X)^{-1} X^T Y.$$

Recall the F distribution linked to the chi square variable; it will be used to conduct a test on our coefficients. Let $Q_{i,k}$ be a set of independent random variables which follow the χ^2 distribution with k degrees of freedom, for $i = 1, 2$. Consider the ratio

$$\frac{Q_{1,k}/k}{Q_{1,p}/p}.$$

This variable follows the F-distribution for k, p degrees of freedom, k in the numerator and p in the denominator.

Let q be an integer such that $0 < q \leq p$. Under the hypothesis that the values of the last q elements of β are 0, let b_s be the resulting OLS estimate for β. The purpose of this test is to see whether all the coefficients are needed to explain the data; if the Null hypothesis is right and q are all zero, the data is not explained well and is unusual as measured by a statistic following the F distribution, as we will see. If the Null hypothesis is rejected, then either the model is just wrong, some of the coefficients are not zero or simply an unusual event has happened.

To see the statistic for this test, set F to be the quantity

$$F = \frac{||Xb||^2 - ||Xb_s||^2/q}{||e||^2/(n-p)}.$$

The denominator $||e||^2$ is the sum of squares of $n - p$ independent random normal variables, with mean 0, under the Null hypothesis. Therefore the ratio $||e||^2/\sigma^2$ follows the chi square distribution with $n - p$ degrees of freedom. For the Numerator, if the Null hypothesis is correct, when $||Xb_s||^2$ is subtracted from $||Xb||^2$ we are left with the sum of squares of q standard normal variables with variance σ^2. Therefore

$$\frac{||Xb||^2 - ||Xb_s||^2]}{\sigma^2}$$

follows the chi square distribution with q degrees of freedom. This means the ratio shown, with the standard deviation σ the same for both sums of squares under the Null Hypothesis, satisfies the conditions for the F distribution with q degrees of freedom in the numerator and with $n - p$ degrees of freedom in the denominator.