MATH 3P82
REGRESSION ANALYSIS
Lecture Notes

Contents

Chapter 1 PREVIEW

Regression is a procedure which selects, from a certain class of functions, the one which best fits a given set of empirical data (usually presented as a table of x and y values with, inevitably, some random component). The 'independent' variable x is usually called the REGRESSOR (there may be one or more of these), the 'dependent' variable y is the RESPONSE variable.. The random components (called RESIDUALS) are usually assumed *normally* distributed, with the same σ and independent of each other.

The class from which the functions are selected (the MODEL) is usually one of the following types:

1. a linear function of x (i.e. $y = a + b\,x$) - simple (univariate) linear regression,
2. a linear function of x_1, x_2, ... x_k - multiple (multivariate) linear regression,
3. a polynomial function of x - polynomial regression,
4. any other type of function, with one or more parameters (e.g. $y = a\,e^{bx}$) - nonlinear regression.

The coefficients (parameters) of these models are called REGRESSION COEFFICIENTS (parameters). Our main task is going to be to find good ESTIMATORS of the regression coefficients (they should have correct expected values and variances as small as possible), to be used for *predicting* values of y when new observations are taken.

Some of the related issues are:

1. How do know (can we test) whether the relationship (between y and x) is truly linear? What if it is not (we have switch to either polynomial or nonlinear model).
2. Similarly, are the residuals truly normal and independent of each other? How do we fix the procedure if the answer is NO.
3. Even when they are normal and independent, what if their variance changes with x (here, we have to do the so called WEIGHTED REGRESSION).
4. Even when all the assumptions are properly met: In the multivariate case with many independent variables, do we really need them all to make a good prediction about y? And, if it *is* possible to reduce them (usually substantially) to a smaller subset, how do we do it (i.e. selecting the *best* five, say).?

Chapter 2 USING MAPLE

Basics

Typing an EXPRESSION (following Maple's > prompt) results in evaluating it. When the expression contains only integers (no decimal point), one gets the exact (rational) answer, as soon as at least one number in the expression is real (with a decimal point), the result is real (rounded off to 10 significant digits). The symbols $*$, / and ^ facilitate multiplication, division and exponentiation, respectively. Note that each line of your input has to end with a semicolon:

```
> 4*5-3/(5+2)+2^(-3);
```

$$\frac{1103}{56}$$

The result of any computation can be stored under a name (which you make up, rather arbitrarily), and used in any subsequent expression. Maple then remembers the value, until the end of your session, or till you deliberately replace it with a new value. Note that this (giving a name to a result) is achieved by typing the name, followed by a colon and the equal sign (a group of two symbols, representing a single operation), followed by the actual expression to be stored:

```
> a:=(3.0+4)*(2-6)+2/3-4/5;
```

$$a := -28.13333\,333$$

```
> a/7+9;
```

$$4.98095\,238$$

```
> a:=14/6;
```

$$a := \frac{7}{3};$$

```
> a/7+9;
```

$$a := \frac{28}{3};$$

(from now on, we will omit the > prompt from our examples, showing only what *we* have to type).

Maple can also handle the usual functions such as **sin, cos, tan, arcsin, arccos, arctan, exp, ln, sqrt**, etc. All angles are always measured in radians.

```
sin(3.);sqrt(8);
```

$$.14112\,00081$$

$$2\sqrt{2}$$

We can also **define** our own functions by:

```
f:=x->x^2;
```

$$f := x \to x^2$$

$f(3);$

$$9$$

where f is an arbitrary name.

Lists and Loops

Maple can store, under a single name, a whole LIST of values, thus:

$a := [3 / 2, 5, \mathbf{sqrt}(3), 7];$

$$a := [\tfrac{3}{2}, 5, \sqrt{3}, 7]$$

The individual elements of a list can be referenced by indexing (and used in computing another expression):

$a[2] * 4;$

$$20$$

One can add elements of a list by the following COMMAND (as Maple calls them):

$\mathbf{sum}('a[i]','i'= 1..4);$

$$\tfrac{27}{2} + \sqrt{3}$$

One can convert the last answer to its decimal form by:

$\mathbf{evalf}(\%);$

$$15.23205\,081$$

Note that the % symbol always refers to the previous expression.

Similarly to **sum**, one can also compute **product** of elements of a list.

To subtract say 3 from each element of the list a, redefining a correspondingly, can be achieved by:

for i **from** 1 **to** 4 **do** $a[i] := a[i] - 3$ **end do**:

Note that terminating a statement by : instead of the usual ; will prevent Maple from printing the four results computed in the process (we may not need to see them individually). Also note that, upon completion of this statement, i will have the value of 5 (any information i had contained previously will have been destroyed)!

We can easily verify that the individual elements of our a list have been updated accordingly:

$a[2];$

$$2$$

We may also create a list using the following approach:

$b := [\mathbf{seq}(2 \hat{} i, i = 1..6)];$

$$b := [2, 4, 8, 16, 32, 64];$$

Variables and Polynomials

If a symbol, such as for example x, has not been assigned a specific value, Maple considers it a variable. We may then define a to be a POLYNOMIAL in x, thus:

$a := 3 - 2 * x + 4 * x\hat{}2$;

$$a := 3 - 2\,x + 4\,x^2$$

A polynomial can be differentiated

diff(a, x);

$$-2 + 8\,x$$

integrated from, say, 0 to 3

int$(a, x = 0..3)$;

$$36$$

or plotted, for a certain range of x values

plot$(a, x = 0..3)$;

We can also evaluate it, substituting a specific number for x (there are actually two ways of doing this):

subs$(x = 3, a)$; **eval**$(a, x = 3)$;

$$33$$

$$33$$

We can also multiply two polynomials (in our example, we will multiply a by itself), but to convert to a regular polynomial form, we nee to **expand** the answer:

$a * a$; **expand**$(\%)$;

$$(3 - 2\,x + 4\,x^2)^2$$

$$9 - 12\,x + 28\,x^2 - 16\,x^3 + 16\,x^4$$

Procedures

If some specific computation (consisting, potentially, of several steps) is to be done, more than once (e.g. we would like to be able to raise each element of a list of values to a given power), we need first to design the corresponding PROCEDURE (effectively a simple computer program), for example:

$RAISETO$:= **proc**(L, N) ; **local** K, n, i ; $K := L$; $n :=$ **nops**(L) ;

for i **from** 1 **to** n **do** $K[i] := K[i]\,\hat{}\,N$ **end do** ; K **end proc** :

where $RAISETO$ is an arbitrary name of the procedure, L and N are arbitrary names of its ARGUMENTS (also called parameters), the first for the list and the second for the exponent, K, n and i are auxiliary names to be used in the actual computation (since they are **local**, they will not interfere with any such names used outside the procedure). First we copy L into K (Maple does not like it if we try to modify L directly) and find its length n (by the **nops** command). Then, we raise each element K to the power of N, and return (the last expression of the procedure) the modified list. We can organize the procedure into several lines by using Shift-Enter (to move to the next line).

We can then use the procedure as follows:

$SVFL([2, 5, 7, 1], 2)$; $SVFL([3, 8, 4], -1)$;

$[4, 25, 49, 1]$

$[\frac{1}{3}, \frac{1}{8}, \frac{1}{4}]$

Matrix Algebra

We can define a matrix by:

$a :=$ **matrix**(2, 2, [1, 2, 3, 4]) :

where 2, 2 specifies its dimensions (number of rows and columns, respectively), followed by the list of its elements (row-wise).

We can multiply two matrices (here, we multiply a by itself) by

evalm(a &∗ a) :

Note that we have to replace the usual ∗ by &∗. Similarly, we can add and subtract (using + and −), and raise a to any positive integer power (using ˆ).

We can also multiply a by a vector (of matching length), which can be entered as a list:

evalm(a &∗ [2, 5]) :

Note that reversing the order of a and [2, 5] yields a different answer.

We can also compute the **transpose** and **inverse** of a, but first we must ask Maple to make these commands available by:

with(linalg) :

We can then perform the required operation by

transpose(a) :

etc.

Similarly, to solve a set of linear equation with a being the matrix of coefficients and [2, 3] the right hand side vector, we do:

linsolve$(a,\ [2,\ 3])$:

Other useful commands:

$a :=$ **randmatrix** $(5,\ 5)$:

creates a matrix of specified dimensions with random elements,

augument$(a,\ [6,\ 2,\ 7,\ 1,\ 0])$:

attaches the list, making it an extra (last) column of a,

submatrix$(a,\ 2..4,\ 1..2)$:

reduces a to a 3 by 2 submatrix, keeping only rows 2, 3 and 4, and columns 1 and 2,

swaprow$(a,\ 2,\ 5)$:

interchanges rows 2 and 5 of a,

addrow$(a,\ 2,\ 4,\ 2/3)$:

adds row 2 multiplied by $\frac{2}{3}$ to row 4 of a.

To recall the proper syntax of a command, one can always type:

? **addrow**

to get its whole-page description, usually with examples.

Plots

Plotting a specific function (or several functions) is easy (as we have already seen):

plot($\{\sin(x), x - x\hat{}3/6\}, x = 0..Pi/2)$:

One can also plot a scattergram of individual points (it is first necessary to ask Maple to make to corresponding routine available, as follows:

with(plots) :

pointplot($[[0, 2], [1, -3], [3, 0], [4, 1], [7, -2]]$);

Note that the argument was a *list* of pairs of x-y values (each pair itself enclosed in brackets).

We can combine any two such plots (usually a scattergram of points together with a fitted polynomial) by:

$pic1 :=$ **pointplot**([**seq**($[i/5,\ \sin(i/5)], i = 1..7)]$) :

$pic2 :=$ **plot**$(\sin(x), x = 0..1.5)$:

display$(pic1, pic2)$:

Chapter 3 SIMPLE REGRESSION

The model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.1}$$

where $i = 1, 2, ..., n$, making the following assumptions:

1. The values of x are measured 'exactly', with no random error. This is usually so when we can choose them at will.

2. The ε_i are normally distributed, independent of each other (uncorrelated), having the expected value of 0 and variance equal to σ^2 (the same for each of them, regardless of the value of x_i). Note that the actual value of σ is usually not known.

The two regression coefficients are called the SLOPE AND INTERCEPT. Their actual values are also unknown, and need to be estimated using the empirical data at hand.

To find such ESTIMATORS, we use the

Maximum Likelihood Method

which is almost always the best tool for this kind of task. It guarantees to yield estimators which are ASYMPTOTICALLY UNBIASED, having the smallest possible variance. It works as follows:

1. We write down the joint probability density function of the y_i's (note that these *are* random variables).

2. Considering it a function of the parameters (β_0, β_1 and σ in this case) *only* (i.e. 'freezing' the y_i's at their *observed* values), we *maximize* it, using the usual techniques. The values of β_0, β_1 and σ to yield the maximum value of this so called LIKELIHOOD FUNCTION (usually denoted by $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}$) are the actual estimators (note that they will be functions of x_i *and* y_i).

Note that instead of maximizing the likelihood function itself, we may choose to maximize its logarithm (which must yield the same $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}$).

Least-Squares Technique

In our case, the Likelihood function is:

$$L = \frac{1}{(\sqrt{2\pi}\sigma)^n} \prod_{i=1}^{n} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]$$

and its logarithm:

$$\ln L = -\frac{n}{2}\log(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

To maximize this expression, we first differentiate it with respect to σ, and make the result equal to zero. This yields:

$$\widehat{\sigma}_m = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n}}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the values of β_0 and β_1 which minimize

$$SS \equiv \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

namely the sum of squares of the *vertical* deviations of the y_i values from the fitted straight line (this gives the technique its name).

To find $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we have to differentiate SS, separately, with respect to β_0 and β_1, and set each of the two answers to zero. This yields:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} y_i - n\,\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

and

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} x_i\, y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

or equivalently, the following so called

Normal equations

$$\begin{aligned} n\,\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i\, y_i \end{aligned}$$

They can be solved easily for β_0 and β_1 (at this point we can start calling them $\widehat{\beta}_0$ and $\widehat{\beta}_1$):

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \equiv \frac{S_{xy}}{S_{xx}}$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{3.2}$$

meaning that the regression line passes through the $(\overline{x}, \overline{y})$ point, where

$$\overline{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$\overline{y} \equiv \frac{\sum_{i=1}^{n} y_i}{n}$$

Each $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is clearly a linear combination of normally distributed random variables, their joint distribution is thus of the bivariate normal type.

```
> x := [77, 76, 75, 24, 1, 20, 2, 50, 48, 14, 66, 45, 12, 37]:
> y := [338, 313, 333, 121, 41, 95, 44, 212, 232, 68, 283, 209, 102, 159]:
> xbar := sum('x[i]','i = 1..14')/14.:
> ybar := sum('y[i]','i = 1..14')/14.:
> Sxx := sum('(x[i] - xbar)^2','i = 1..14'):
> Sxy := sum('(x[i] - xbar) * (y[i] - ybar)','i = 1..14'):
> β1 := Sxy/Sxx;
                    β1 := 3.861296955;
> β0 := ybar - β1 * xbar;
                    β0 := 31.2764689
> with(plots):
> pl1 := pointplot([seq([x[i], y[i]], i = 1..14)]):
> pl2 := plot(β0 + β1 * x, x = 0..80):
> display(pl1, pl2);
```

Statistical Properties of the three Estimators

First, we should realize that it is the y_i (not x_i) which are random, due to the ε_i term in (3.1) - both β_0 and β_1 are also fixed, albeit unknown parameters. Clearly then

$$\mathbb{E}\left(y_i - \overline{y}\right) = \beta_0 + \beta_1 x_i - \left(\beta_0 + \beta_1 \overline{x}\right) = \beta_1 \left(x_i - \overline{x}\right)$$

which implies

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot \mathbb{E}(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \beta_1$$

Similarly, since $\mathbb{E}(\overline{y}) = \beta_0 + \beta_1 \overline{x}$, we get

$$\mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0$$

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are thus UNBIASED estimators of β_0 and β_1, respectively.

To find their respective variance, we first note that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \equiv \frac{\sum_{i=1}^{n} (x_i - \overline{x})\, y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

(right?), based on which

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \mathrm{Var}(y_i)}{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^2} = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

From (3.2) we get

$$\mathrm{Var}\left(\widehat{\beta}_0\right) = \mathrm{Var}(\overline{y}) - 2\overline{x}\,\mathrm{Cov}(\overline{y},\widehat{\beta}_1) + \overline{x}^2\mathrm{Var}\left(\widehat{\beta}_1\right)$$

We already have a formula for $\mathrm{Var}\left(\widehat{\beta}_1\right)$, so now we need

$$\mathrm{Var}(\overline{y}) = \mathrm{Var}(\overline{\varepsilon}) = \frac{\sigma^2}{n}$$

and

$$\mathrm{Cov}(\overline{y},\widehat{\beta}_1) = \mathrm{Cov}\left(\frac{\sum_{i=1}^{n}\varepsilon_i}{n}, \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{S_{xx}}\right) = \frac{\sigma^2\sum_{i=1}^{n}(x_i-\overline{x})}{S_{xx}} = 0$$

(uncorrelated). Putting these together yields:

$$\mathrm{Var}\left(\widehat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

The covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is thus equals to $-\overline{x}\,\mathrm{Var}(\widehat{\beta}_1)$, and their correlation coefficient is

$$\frac{-1}{\sqrt{1+\frac{1}{n}\cdot\frac{S_{xx}}{\overline{x}^2}}}$$

Both variance formulas contain σ^2, which, in most situations, must be replaced by its ML estimator

$$\widehat{\sigma}^2_m = \frac{\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{n} \equiv \frac{SS_E}{n}$$

where the numerator defines the so called RESIDUAL (ERROR) SUM OF SQUARES. It can be rewritten in the following form (replacing $\widehat{\beta}_0$ by $\overline{y}-\widehat{\beta}_1\overline{x}$):

$$\begin{aligned} SS_E &= \sum_{i=1}^{n}(y_i-\overline{y}+\widehat{\beta}_1\overline{x}-\widehat{\beta}_1x_i)^2 = \sum_{i=1}^{n}\left[y_i-\overline{y}+\widehat{\beta}_1(\overline{x}-x_i)\right]^2 \\ &= S_{yy} - 2\widehat{\beta}_1S_{xy} + \widehat{\beta}_1^2S_{xx} = S_{yy} - 2\frac{S_{xy}}{S_{xx}}S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \\ &= S_{yy} - \frac{S_{xy}}{S_{xx}}S_{xy} = S_{yy} - \widehat{\beta}_1S_{xy} \equiv S_{yy} - \widehat{\beta}_1^2S_{xx} \end{aligned}$$

Based on (3.1) and $\overline{y} = \beta_0 + \beta_1\overline{x} + \overline{\varepsilon}$ (from now on, we have to be very careful to differentiate between β_0 and $\widehat{\beta}_0$, etc.), we get

$$\mathbb{E}(S_{yy}) = \mathbb{E}\left\{\sum_{i=1}^{n}\left[\beta_1(x_i-\overline{x}) + (\varepsilon_i-\overline{\varepsilon})\right]^2\right\} = \beta_1^2\,S_{xx} + \sigma^2(n-1)$$

(the last term was derived in MATH 2F96). Furthermore,

$$\mathbb{E}\left(\widehat{\beta}_1^2\right) = \mathrm{Var}(\widehat{\beta}_1) - \mathbb{E}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{S_{xx}} - \beta_1^2$$

Combining the two, we get

$$\mathbb{E}(SS_E) = \sigma^2(n-2)$$

Later on, we will be able to prove that $\dfrac{SS_E}{\sigma^2}$ has the χ^2 distribution with $n-2$ degrees of freedom. It is also *independent* of each $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

This means that there is a slight bias in the $\widehat{\sigma}_m^2$ estimator of σ^2 (even though the bias disappears in the $n \to \infty$ limit - such estimators are called ASYMPTOTICALLY UNBIASED). We can easily fix this by defining a new, fully unbiased

$$\widehat{\sigma}^2 = \frac{SS_E}{n-2} \equiv MS_E$$

(the so called MEAN SQUARE) to be used instead of $\widehat{\sigma}_m^2$ from now on.

All of this implies that both

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{MS_E\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right)}}$$

and

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MS_E}{S_{xx}}}} \tag{3.3}$$

have the Student `t` distribution with $n-2$ degrees of freedom. This can be used either to construct the so called CONFIDENCE INTERVAL for either β_0 or β_1, or to test any HYPOTHESIS concerning β_0 or β_1.

The corresponding Maple commands (to compute SS_E, MS_E, and the two standard errors - denominators of the last two formulas) are:

$> Syy :=$**sum**$((y[i] - ybar)\hat{}2, i = 1..14)$:
$> SSE := Syy - \beta 1\hat{}2 * Sxx$:
$> MSE := SSE/12$:
$> se1 :=$**sqrt**(MSE/Sxx):
$> se2 :=$**sqrt**$(MSE/(1/14 + xbar\hat{}2/Sxx))$:

Confidence Intervals

To construct a confidence interval for an unknown parameter, we first choose a so called CONFIDENCE LEVEL $1-\alpha$ (the usual choice is to make it equal to 95%, with $\alpha = 0.05$). This will be the probability of constructing an interval which *does* contain the true value of the parameter.

Regression coefficients

Knowing that (3.3) has the $\mathtt{t}_{n-2}$ distribution, we must then find two values (called CRITICAL) such that the probability of (3.3) falling inside the corresponding interval (between the two values) is $1-\alpha$. At the same time, we would like to have the interval as short as possible. This means that we will be choosing the critical values symmetrically around 0; the positive one will equal to $\mathtt{t}_{\frac{\alpha}{2},n-2}$, the negative one to $-\mathtt{t}_{\frac{\alpha}{2},n-2}$ (the first index now refers to the area of the remaining TAIL of the distribution) - these critical values are widely tabulated. We can also find them with the help of Maple, by:

> **with(stats)**:

> **statevalf**[**icdf,studentst**[12]](0.975);

where **icdf** stands for 'inverse cumulative density function' ('cumulative density function' being a peculiar name for 'distribution function'), and 0.975 is the value of $1-\frac{\alpha}{2}$ (leaving $\frac{\alpha}{2}$ for the tail).

The statement that (3.3) falls in the interval between the two critical values of $\mathtt{t}_{n-2}$ is equivalent (solve the corresponding equation for β_1) to saying that the value of β_1 is in the following range

$$\widehat{\beta}_1 \pm \mathtt{t}_{\frac{\alpha}{2},n-2}\sqrt{\frac{MS_E}{S_{xx}}}$$

which is our $(1-\alpha)\cdot 100\,\%$ confidence interval.

The only trouble is that, when we make that claim, we are either $100\,\%$ right or $100\,\%$ wrong, since β_1 is *not* a random variable. The probability of 'hitting' the correct value was in constructing the interval (which each of us will do differently, if we use independent samples). This is why we use the word CONFIDENCE instead of probability (we claim, with the $(1-\alpha)\cdot 100\,\%$ confidence, that the exact value of β_1 is somewhere inside the constructed interval).

Similarly, we can construct a $1-\alpha$ level-of-confidence interval for $\widehat{\beta}_0$, thus:

$$\widehat{\beta}_0 \pm \mathtt{t}_{\frac{\alpha}{2},n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}$$

Note that, since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not independent, making a joint statement about the two (with a specific level of confidence) is more complicated (one has to construct a confidence ellipse, to make it correct).

Residual variance

Constructing a $1-\alpha$ confidence interval for σ^2 is a touch more complicated. Since $\frac{SS_E}{\sigma^2}$ has the χ^2_{n-2} distribution, we must first find the corresponding two critical values. Unfortunately, the χ^2 distribution is not symmetric, so for these two we have to take $\chi^2_{\frac{\alpha}{2},n-2}$ and $\chi^2_{1-\frac{\alpha}{2},n-2}$. Clearly, the probability of a χ^2_{n-2} random variable falling between the two values equals $1-\alpha$. The resulting interval may not be the shortest of all these, but we are obviously quite close to the right solution; furthermore, the choice of how to divide α between the two tails remains simple and logical.

Solving for σ^2 yields

$$\left(\frac{SS_E}{\chi^2_{1-\frac{\alpha}{2},n-2}}, \frac{SS_E}{\chi^2_{\frac{\alpha}{2},n-2}}\right)$$

as the corresponding $(1-\alpha)\cdot 100\,\%$ confidence interval.

Maple can supply the critical values:

> **statevalf**[**icdf,chisquare**[12]](.975);

Expected-value estimator

Sometimes we want to estimate the *expected* value of y obtained with a new choice of x (let us call it x_0) which should *not* be outside the *original* range of x values (no extrapolation)! This effectively means that we want a good estimator for $\mathbb{E}(y_0) \equiv \beta_0 + \beta_1 x_0$. Not surprisingly, we use

$$\widehat{y}_0 \equiv \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \overline{y} + \widehat{\beta}_1 (x_0 - \overline{x})$$

which is clearly unbiased, normally distributed, with the variance of

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0 - \overline{x})^2$$

since $\overline{y}$ and $\widehat{\beta}_1$ are uncorrelated. This implies that

$$\frac{\widehat{y}_0 - \mathbb{E}(y_0)}{\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}}$$

must also have the t_{n-2} distribution.. It should now be quite obvious as to how to construct a confidence interval for $\mathbb{E}(y_0)$.

New y value

We should also realize that predicting an actual new value of y taken at x_0 (let us call it y_0) is a different issue, since now an (independent) error ε_0 is added to $\beta_0 + \beta_1 x_0$. For the prediction itself we still have to use the same $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ (our best prediction of ε_0 is its expected value 0), but the variance of y_0 is the variance of $\widehat{y}_0$ *plus* σ^2 (the variance of ε_0), i.e.

$$\sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0 - \overline{x})^2$$

It thus follows that

$$\frac{\widehat{y}_0 - y_0}{\sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}}$$

also has the t_{n-2} distribution.. We can then construct the corresponding $(1-\alpha)\cdot 100\,\%$ PREDICTION interval for y_0. The reason why we use another name again is that now we are combining the a priori error of a confidence interval with the usual, yet-to-happen error of taking the y_0 observation.

Hypotheses Testing

Rather than constructing a confidence interval for an unknown parameter, we may like to test a specific HYPOTHESIS concerning the parameter (such as, for example, that the exact slope is zero). The two procedures (hypotheses testing and confidence-interval construction) are computationally quite similar (even if the logic is different).

First we have to state the so called NULL HYPOTHESIS H_0, such as, for example, that $\beta_1 = 0$ (meaning that x does *not* effect y, one way or the other). This is to be tested against the ALTERNATE HYPOTHESIS H_A ($\beta_1 \neq 0$ in our case).

To perform the test, we have compute the value of a so called TEST STATISTIC T. This is usually the corresponding estimator, 'normalized' to have a simple distribution, free from unknown parameters, when H_0 is true - in our case, we would use (3.3) with $\beta_1 = 0$, i.e.

$$T \equiv \frac{\widehat{\beta}_1}{\sqrt{\frac{MS_E}{S_{xx}}}}$$

Under H_0, its distribution is $\mathtt{t}_{n-2}$, otherwise (under H_A) it has a more complicated NON-CENTRAL distribution (the non-centricity parameter equal to the actual value of β_1).

Now, based on the value of T, we have to make a decision as to whether to go with H_0 or H_A. Sure enough, if H_0 is true, the value of T must be relatively small, but how small is small? To settle that, we allow ourselves the probability of α (usually 5%) to make a so called TYPE I error (rejecting H_0 when true). Out critical values will then be the same as those of the corresponding confidence interval ($\pm\mathtt{t}_{\frac{\alpha}{2},n-2}$). We REJECT H_0 whenever the value of T enters the CRITICAL REGION (outside the interval), and don't reject (accept) H_0 otherwise. Note that the latter is a weaker statement - it is not a proof of H_0, it is more of an *inability* to disprove it! When accepting H_0, we can of course be making a TYPE II ERROR (accepting H_0 when wrong), the probability of which now depends on the actual (non-zero) value of β_1 (being, effectively, a function of these). To compute these errors, one would have to work with the non-central $\mathtt{t}_{n-2}$ distributions (we will not go into that).

Model Adequacy (Lack-of-Fit Test)

Let us summarize the assumptions on which the formulas of the previous sections are based.

The first of them (called MODEL ADEQUACY) stipulates that the relationship between x and y is linear. There are two ways of checking it out. One (rather superficial, but reasonably accurate) is to plot the resulting residuals against the x_i values, and see whether there is any systematic oscillation. The other one (more 'scientific' and quantitative) is available only when several independent y observations are taken at each x_i value. This yields and 'independent' estimate of our σ, which should be consistent with the size of the computed residuals (a precise test for doing this is the topic of this section, and will be described shortly).

The other three assumptions all relate to the ε_i's

1. being *normally* distributed,

2. having the same (*constant*) standard deviation σ,

3. being independent, i.e. *uncorrelated*.

We would usually be able to (superficially) establish their validity by scrutinizing the same e_i-x_i graph. In subsequent sections and chapters, we will also deal with the corresponding remedies, should we find any of them violated.

For the time being, we will assume that the last three assumptions hold, but we are not so sure about the straight-line relationship between x and y. We have also collected, at each x_i, several independent values of y (these will be denoted y_{ij}, where $j = 1, 2, ...n_i$).

In this case, our old residual (error) sum of squares can be partitioned into two components, thus:

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{m} n_i(\overline{y}_i - \widehat{y}_i)^2 \equiv SS_{PE} + SS_{LOF}$$

due to PURE ERROR and LACK OF FIT, respectively. Here, m is the number of distinct x_i values, and

$$\overline{y}_i \equiv \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

is the 'GROUP' MEAN of the y observations taken with x_i. Note that the overall mean (we used to call it $\overline{y}$, but now we switch - just for emphasis - to $\overline{\overline{y}}$, and call it the GRAND MEAN) can be computed by

$$\overline{\overline{y}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{m} n_i} \equiv \frac{\sum_{i=1}^{m} n_i\,\overline{y}_i}{\sum_{i=1}^{m} n_i}$$

The old formulas for computing $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (and their standard errors) remain correct, but one has to redefine

$$\begin{aligned}
\overline{x} &\equiv \frac{\sum_{i=1}^{m} n_i\,x_i}{\sum_{i=1}^{m} n_i} \\
S_{xx} &\equiv \sum_{i=1}^{m} n_i(x_i - \overline{x})^2 \\
S_{xy} &\equiv \sum_{i=1}^{m}(x_i - \overline{x})\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}}) = \sum_{i=1}^{m} n_i(x_i - \overline{x})\,\overline{y}_i
\end{aligned}$$

But the primary issue now is to verify that the model is adequate.

To construct the appropriate test, we first have to prove that, under the null hypothesis (linear model correct), $\frac{SS_{PE}}{\sigma^2}$ and $\frac{SS_{LOF}}{\sigma^2}$ are independent, and have the χ^2_{n-m} and χ^2_{m-2} distribution, respectively (where $n \equiv \sum_{i=1}^{m} n_i$, the total number of y observations).

Proof: The statements about $\frac{SS_{PE}}{\sigma^2}$ is a MATH 2F96 result. Proving that $\frac{SS_{LOF}}{\sigma^2}$ has the χ^2_{m-2} distribution is the result of the next section. Finally, since $\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ is independent of $\overline{y}_i$ (another MATH 2F96 result), and SS_{PE} is a sum of the former, and SS_{LOF} is computed based on the latter (since $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, and both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are computed using the 'group' means $\overline{y}_i$ only). □

To test the null hypothesis that the x-y relationship is linear (against all possible alternatives), we can then use the following test statistic:

$$\frac{\frac{SS_{LOF}}{m-2}}{\frac{SS_{PE}}{n-m}}$$

which (under H_0) has the $\mathsf{F}_{m-2,n-m}$ distribution. When H_0 is false, SS_{LOF} (but not SS_{PE}) will tend to be 'noticeably' larger than what could be ascribed to a purely random variation. We will then reject H_0 in favor of H_A as soon as the value of the test statistics enters the critical (right-hand tail) region of the corresponding F distribution.

```
> x := [1, 3, 6, 8.]:
> y := [[2.4, 3.2, 2.9, 3.1], [3.9, 4], [4.2], [4.1, 4.7, 5.6, 5.1, 4.9]]:
> ng := [seq(nops(y[i]), i = 1..4)]:
> n := sum(ng[i], i = 1..4):
> ybar := [seq(sum(y[i][j], j = 1..ng[i])/ng[i], i = 1..4)]:
> SSpe := sum(sum((y[i][j] - ybar[i])^2, j = 1..ng[i]), i = 1..4):
> xmean := sum(ng[i] * x[i], i = 1..4)/n:
> ymean := sum(ng[i] * ybar[i], i = 1..4)/n:
> Sxx := sum(ng[i] * (x[i] - xmean)^2, i = 1..4):
> Sxy := sum(ng[i] * (x[i] - xmean) * ybar[i], i = 1..4):
> beta1 := Sxy/Sxx:
> beta0 := ymean - xmean * beta1:
> SSlof := sum(ng[i] * (ybar[i] - beta0 - beta1 * x[i])^2, i = 1..4):
> (SSlof/2)/(SSpe/8);
                        0.9907272888
> with(stats):
> statevalf[icdf,fratio[2, 8]](0.95);
                        4.458970108
> with(plots):
> pl1 :=pointplot([seq(seq([x[i], y[i][j]], j = 1..ng[i]), i = 1..4)]):
> pl2 :=plot(beta0 + beta1 * z, z = 0.5..8.5):
> display(pl1, pl2);
```

Weighted Regression

In this section, we modify the procedure to accommodate the possibility that the variance of the error terms is not constant, but it is proportional to a given function of x, i.e.

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 \cdot g(x_i) \equiv \frac{\sigma^2}{w_i}$$

The same modification of the variance is also encountered in a different context: When, at x_i, n_i observations are taken (instead of the usual one) and the resulting *mean* of the y observations is recorded (we will still call it y_i), then (even with the constant-σ assumption for the individual observations), we have the previous situation with $w_i = n_i$. The w_i values are called WEIGHTS (observations with higher weights are to be taken that much more seriously).

It is quite obvious that maximizing the likelihood function will now require to minimize the weighted sum of squares of the residuals, namely

$$\sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

The resulting estimators of the regression coefficients are the old

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

$$\begin{aligned}
\overline{x} &\equiv \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\
\overline{y} &\equiv \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \\
S_{xx} &\equiv \sum_{i=1}^{n} w_i (x_i - \overline{x})^2 \\
S_{xy} &\equiv \sum_{i=1}^{n} w_i (x_i - \overline{x})(y_i - \overline{y})
\end{aligned}$$

One can easily show that all related formulas remain the same, except for:

$$\begin{aligned}
\mathrm{Var}(\overline{y}) &= \mathrm{Var}(\overline{\varepsilon}) = \frac{\sigma^2}{\sum_{i=1}^{n} w_i} \\
\mathrm{Var}(\widehat{\beta}_0) &= \sigma^2 \left(\frac{1}{\sum_{i=1}^{n} w_i} + \frac{\overline{x}^2}{S_{xx}} \right) \\
\mathrm{Corr}(\widehat{\beta}_0, \widehat{\beta}_1) &= \frac{-1}{\sqrt{1 + \dfrac{1}{\sum_{i=1}^{n} w_i} \cdot \dfrac{S_{xx}}{\overline{x}^2}}}
\end{aligned}$$

which require replacing n by the total weight.

Similarly, for the maximum-likelihood estimator of σ^2 we get

$$\widehat{\sigma}^2_m = \frac{\sum\limits_{i=1}^{n} w_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2}{n} = \frac{S_{yy} - \widehat{\beta}_1^2 S_{xx}}{n}$$

Since

$$\mathbb{E}(S_{yy}) = \mathbb{E}\left\{ \sum_{i=1}^{n} w_i \left[\beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})\right]^2 \right\} = \beta_1^2 \, S_{xx} + \sigma^2 (n-1)$$

remains unchanged (note that his time we did *not* replace n by the total weight) - this can be seen from

$$\mathbb{E}\left[\sum_{i=1}^{n} w_i(\varepsilon_i - \overline{\varepsilon})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} w_i(\varepsilon_i^2 - 2\varepsilon_i\overline{\varepsilon} + \overline{\varepsilon}^2)\right] =$$
$$\sum_{i=1}^{n} w_i \mathrm{Var}(\varepsilon_i) - 2\sum_{i=1}^{n} w_i^2 \frac{\mathrm{Var}(\varepsilon_i)}{\sum_{i=1}^{n} w_i} + \mathrm{Var}(\overline{\varepsilon}) \sum_{i=1}^{n} w_i = \sigma^2(n-1)$$

and so does

$$\mathbb{E}\left(\widehat{\beta}_1^2\right) = \mathrm{Var}(\widehat{\beta}_1) - \mathbb{E}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{S_{xx}} - \beta_1^2$$

we still get the same

$$\mathbb{E}(SS_E) = \sigma^2(n-2)$$

This implies that

$$\widehat{\sigma}^2 = \frac{S_{yy} - \widehat{\beta}_1^2 S_{xx}}{n-2}$$

is an unbiased estimator of σ^2. Late on, we will prove that it is still independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and has the χ^2_{n-2} distribution.

Correlation

Suppose now that *both* x and y are random, normally distributed with (bivariate) parameters μ_x, μ_y, σ_x, σ_y and ρ. We know that the *conditional* distribution of y *given* x is also (univariate) normal, with the following conditional mean and variance:

$$\mu_y + \rho\,\sigma_y \frac{x - \mu_x}{\sigma_x} \equiv \beta_0 + \beta_1\, x \tag{3.4}$$
$$\sigma_y^2\,(1 - \rho^2)$$

Our regular regression would estimates the regression coefficients by the usual $\widehat{\beta}_0$ and $\widehat{\beta}_1$. They are still the 'best' (maximum-likelihood) estimators (as we will see shortly), but their statistical properties are now substantially more complicated.

Historical comment: Note that by reversing the rôle of x and y (which is now quite legitimate - the two variables are treated as 'equals' by this model), we get the following regression line:

$$x = \mu_x + \rho\,\sigma_x \frac{y - \mu_y}{\sigma_y}$$

One can easily see that this line is inconsistent with (3.4) - it is a lot *steeper* when plotted on the same graph. Ordinary regression thus tends, in this case, to distort the true relationship between x and y, making it either more flat or more steep, depending on which variable is taken to be the 'independent' one.

Thus, for example, if x is the height of fathers and y that of sons, the regression line will have a slope less than 45 degrees, implying a false averaging trend (regression towards the mean, as it was originally called - and the name,

even though ultimately incorrect, stuck). The fallacy of this argument was discovered as soon as someone got the bright idea to fit y against x, which would then, still falsely, imply a tendency towards increasing diversity.

One can show that the ML technique would use the usual $\overline{x}$ and $\overline{y}$ to estimate μ_x and μ_y, $\sqrt{\frac{S_{xx}}{n-1}}$ and $\sqrt{\frac{S_{yy}}{n-1}}$ to estimate σ_x and σ_y, and

$$r \equiv \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \tag{3.5}$$

as an estimator of ρ (for some strange reason, they like calling the estimator r rather than the usual $\widehat{\rho}$). This relates to the fact that

$$\frac{S_{xy}}{n-1}$$

is an unbiased estimator of $\mathrm{Cov}(X,Y)$.

Proof:

$$\mathbb{E}\left\{\sum_{i=1}^{n}\left[x_i-\mu_x-(\overline{x}-\mu_x)\right]\left[y_i-\mu_y-(\overline{y}-\mu_y)\right]\right\}=$$
$$\sum_{i=1}^{n}\left[\mathrm{Cov}(X,Y)-\frac{\mathrm{Cov}(X,Y)}{n}-\frac{\mathrm{Cov}(X,Y)}{n}+\frac{\mathrm{Cov}(X,Y)}{n}\right]=$$
$$n\,\mathrm{Cov}(X,Y)\,(1-\frac{1}{n})=\mathrm{Cov}(X,Y)\,(n-1)$$

One can easily verify that these estimators agree with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the previous sections. Investigating their statistical properties now becomes a lot more difficult (mainly because of dividing by $\sqrt{S_{xx}}$, which is random). We have to use large-sample approach to derive ASYMPTOTIC formulas only (i.e. expanded in powers of $\frac{1}{n}$), something we will take up shortly.

The only *exact* result we can derive is that

$$\frac{r\sqrt{n-1}}{\sqrt{1-r^2}}=\frac{\frac{S_{xy}}{\sqrt{S_{xx}^2}}(n-2)}{\sqrt{\frac{S_{xx}S_{yy}-S_{xy}^2}{S_{xx}^2}}}=\frac{\widehat{\beta}_1}{\sqrt{\frac{MS_E}{S_{xx}}}}$$

which we know has the $\mathtt{t}_{n-2}$ distribution, *assuming* that $\beta_1=0$. We can thus use it for testing the corresponding hypothesis (the test will be effectively identical to testing H_0: $\beta_1=0$ against an alternate, using the simple model).

Squaring the r estimator yields the so called COEFFICIENT OF DETERMINATION

$$r^2=\frac{S_{yy}-S_{yy}+\frac{S_{xy}^2}{S_{xx}}}{S_{yy}}=1-\frac{SS_E}{S_{yy}}$$

which tells us how much of the original y variance has been removed by fitting the best straight line.

Large -Sample Theory

Large sample theory tells us that practically all estimators are approximately normal. Some of them of course approach normality a lot faster than others, and we will discuss a way of helping to 'speed up' this process below.

To be more precise, we assume that an estimator has the form of $f(\overline{X}, \overline{Y}, ...)$ where X, Y, ... are themselves functions of *individual* observations, and f is another function of their *sample means* (most estimators are like this), say

$$r = \frac{\overline{(x-\overline{x})(y-\overline{y})}}{\sqrt{\overline{(x-\overline{x})^2}} \cdot \sqrt{\overline{(y-\overline{y})^2}}} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}}$$

To a good approximation we can (Taylor) expand $f(\overline{X}, \overline{Y}, ...)$ around the corresponding expected values, as follows

$$f(\overline{X}, \overline{Y}, ...) \cong f(\mu_X, \mu_Y, ...) + \partial_{\overline{X}} f(...)(\overline{X} - \mu_X) + \partial_{\overline{Y}} f(...)(\overline{Y} - \mu_Y) + ...$$
$$\tfrac{1}{2}\partial^2_{\overline{X},\overline{X}} f(...)(\overline{X} - \mu_X)^2 + \tfrac{1}{2}\partial^2_{\overline{Y},\overline{Y}} f(...)(\overline{Y} - \mu_Y)^2 + \partial^2_{\overline{X},\overline{Y}} f(...)(\overline{X} - \mu_X)(\overline{Y} - \mu_Y) + ...$$

The corresponding expected value is

$$f(\mu_X, \mu_Y, ...) + \frac{\sigma_X^2}{2n}\partial^2_{\overline{X},\overline{X}} f(...) + \frac{\sigma_Y^2}{2n}\partial^2_{\overline{Y},\overline{Y}} f(...) + \frac{\sigma_X \sigma_Y \rho_{XY}}{n}\partial^2_{\overline{X},\overline{Y}} f(...) + ... \quad (3.6)$$

and the variance (based on the linear terms only):

$$\frac{\sigma_X^2}{n}\left[\partial_{\overline{X}} f(...)\right]^2 + \frac{\sigma_Y^2}{n}\left[\partial_{\overline{Y}} f(...)\right]^2 + \frac{2\sigma_X \sigma_Y \rho_{XY}}{n}\left[\partial_{\overline{X}} f(...)\right]\left[\partial_{\overline{Y}} f(...)\right] + ... \quad (3.7)$$

For example, one can show that

$$\widehat{\beta}_1 = \frac{\overline{(x-\overline{x})(y-\overline{y})}}{\overline{(x-\overline{x})^2}}$$

is *approximately* normal, with the mean of

$$\frac{\sigma_y \rho}{\sigma_x} + \frac{2\sigma_x^4 \cdot \frac{2\sigma_x \sigma_y \rho}{\sigma_x^6} + 4\sigma_x^3 \sigma_y \rho \cdot \frac{-1}{\sigma_x^4}}{2n} + ... =$$
$$\beta_1 \left[1 + ...\right]$$

(to derive this result, we borrowed some formulas of the next section). Similarly, one can compute that the corresponding variance equals

$$(\sigma_x^2 \sigma_y^2 + \sigma_x^2 \sigma_y^2 \rho^2) \cdot \frac{1}{\sigma_x^4} + 2\sigma_x^4 \cdot \frac{\sigma_x^2 \sigma_y^2 \rho^2}{\sigma_x^8} + 4\sigma_x^3 \sigma_y \rho \cdot \frac{-\sigma_x \sigma_y \rho}{\sigma_x^6} =$$
$$\frac{\sigma_y^2}{\sigma_x^2}(1 - \rho^2) + ...$$

divided by n.

We will not investigate the statistical behavior of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ any further, instead, we concentrate on the issue which is usually consider primary for this kind of model, namely constructing a

Confidence interval for the correlation coefficient

To apply (3.6) and (3.7) to r, we first realize that the three means are

$$\begin{aligned}\mathbb{E}\left[(x_i-\overline{x})(y_i-\overline{y})\right] &= (1-\frac{1}{n})\sigma_x\sigma_y\rho\\ \mathbb{E}\left[(x_i-\overline{x})^2\right] &= (1-\frac{1}{n})\sigma_x^2\\ \mathbb{E}\left[(y_i-\overline{y})^2\right] &= (1-\frac{1}{n})\sigma_y^2\end{aligned}$$

The corresponding variances (where, to our level of accuracy, we can already replace $\overline{x}$ by μ_x and $\overline{y}$ by μ_y) are easy to get from the following bivariate moment generating function

$$M(t_x,t_y)=\exp\left[\frac{\sigma_x^2t_x^2+\sigma_y^2t_y^2+2\sigma_x\sigma_y\rho t_xt_y}{2}\right]$$

They are, respectively

$$\begin{aligned}\sigma_x^2\sigma_y^2+2\sigma_x^2\sigma_y^2\rho^2-\sigma_x^2\sigma_y^2\rho^2 &= \sigma_x^2\sigma_y^2+\sigma_x^2\sigma_y^2\rho^2\\ 3\sigma_x^4-\sigma_x^4 &= 2\sigma_x^4\\ 3\sigma_y^4-\sigma_y^4 &= 2\sigma_y^4\end{aligned}$$

We will also need the three covariances, which are

$$\begin{aligned}3\sigma_x^3\sigma_y\rho-\sigma_x^3\sigma_y\rho &= 2\sigma_x^3\sigma_y\rho\\ 3\sigma_x\sigma_y^3\rho-\sigma_x\sigma_y^3\rho &= 2\sigma_x\sigma_y^3\rho\\ \sigma_x^2\sigma_y^2+2\sigma_x^2\sigma_y^2\rho^2-\sigma_x^2\sigma_y^2 &= 2\sigma_x^2\sigma_y^2\rho^2\end{aligned}$$

This means that the expected value of r equals, to a good approximation, to

$$\begin{aligned}&\rho+\frac{2\sigma_x^4\cdot\frac{3\rho}{4\sigma_x^4}+2\sigma_y^4\cdot\frac{3\rho}{4\sigma_y^4}+4\sigma_x^3\sigma_y\rho\cdot\frac{-\rho}{2\sigma_x^3\sigma_y\rho}+4\sigma_x\sigma_y^3\rho\cdot\frac{-\rho}{2\sigma_x\sigma_y^3\rho}+4\sigma_x^2\sigma_y^2\rho^2\cdot\frac{\rho}{4\sigma_x^2\sigma_y^2}}{2n}\\ =\;&\rho-\frac{1-\rho^2}{2n}+...\end{aligned}$$

Similarly, the variance of r is

$$\begin{aligned}&(\sigma_x^2\sigma_y^2+\sigma_x^2\sigma_y^2\rho^2)\left(\frac{\rho}{\sigma_x\sigma_y\rho}\right)^2+2\sigma_x^4\left(\frac{-\rho}{2\sigma_x^2}\right)^2+2\sigma_y^4\left(\frac{-\rho}{2\sigma_y^2}\right)^2+4\sigma_x^3\sigma_y\rho\left(\frac{\rho}{\sigma_x\sigma_y\rho}\right)\left(\frac{-\rho}{2\sigma_x^2}\right)\\ &+4\sigma_x\sigma_y^3\rho\left(\frac{\rho}{\sigma_x\sigma_y\rho}\right)\left(\frac{-\rho}{2\sigma_y^2}\right)+4\sigma_x^2\sigma_y^2\rho^2\left(\frac{-\rho}{2\sigma_x^2}\right)\left(\frac{-\rho}{2\sigma_y^2}\right)=1-2\rho^2+\rho^4=(1-\rho^2)^2+...\end{aligned}$$

divided by n.

Similarly, one could compute the third central moment of r and the corresponding skewness (which would turn out to be $-\frac{6\rho}{\sqrt{n}}$, i.e. fairly substantial even for relatively large samples).

One can show that integrating $\dfrac{1}{\sqrt{(1-\rho^2)^2}}$ (in terms of ρ) results in a new quantity whose variance (to this approximation) is constant (to understand the

logic, we realize that $F(r)$ has a variance given by $F'(\rho)^2 \cdot \mathrm{Var}(r)$; now try to make this a constant). The integration yields

$$\frac{1}{2} \ln \frac{1+\rho}{1-\rho} = \operatorname{arctanh} \rho$$

and, sure enough, similar analysis shows that the variance of the corresponding estimator, namely

$$z \equiv \frac{1}{2} \ln \frac{1+r}{1-r}$$

is simply $\frac{1}{n} + ...$ (carrying the computation to $\frac{1}{n^2}$ terms shows that $\frac{1}{n-3}$ is a better approximation). Its expected value is similarly

$$\frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2n} + ... \tag{3.8}$$

and the skewness is, to this approximation equal to 0. The estimator z is therefore becoming normal a lot faster (with increasing n) than r itself, and can be thus used for constructing approximate confidence intervals for ρ. This is done by adding the critical values of $\mathcal{N}(0, \frac{1}{\sqrt{n-3}})$ to z, making the two resulting limits equal to (3.8), and solving for ρ (using a calculator, we usually neglect the $\frac{\rho}{2n}$ term and use **tanh**(...); when Maple is available, we get the more accurate solutions).

Chapter 4 MULTIVARIATE (LINEAR) REGRESSION

First a little insert on

Multivariate Normal Distribution

Consider n independent, standardized, Normally distributed random variables. Their joint probability density function is clearly

$$f(z_1, z_2, ..., z_n) = (2\pi)^{-n/2} \cdot \exp\left(-\frac{\sum_{i=1}^{n} z_i^2}{2}\right) \equiv (2\pi)^{-n/2} \cdot \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right)$$

(a product of the individual pdf's). Similarly, the corresponding moment generating function is

$$\exp\left(\frac{\sum_{i=1}^{n} t_i^2}{2}\right) \equiv \exp\left(\frac{\mathbf{t}^T\mathbf{t}}{2}\right)$$

The following linear transformation of these n random variables, namely

$$\mathbf{X} = \mathbb{A}\,\mathbf{Z} + \boldsymbol{\mu}$$

where $\mathbb{A}$ is an arbitrary (regular) n by n matrix, defines a new set of n random variables having a *general* Normal distribution. The corresponding PDF is clearly

$$\frac{1}{\sqrt{(2\pi)^n}\,|\det(\mathbb{A})|} \cdot \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbb{A}^{-1})^T\mathbb{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right) \equiv$$
$$\frac{1}{\sqrt{(2\pi)^n \det(\mathbb{V})}} \cdot \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\mathbb{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right)$$

and the MGF

$$\begin{aligned}\mathbb{E}\left\{\exp\left[\mathbf{t}^T\left(\mathbb{A}\,\mathbf{Z}+\boldsymbol{\mu}\right)\right]\right\} &= \exp\left(\mathbf{t}^T\boldsymbol{\mu}\right) \cdot \exp\left(\frac{\mathbf{t}^T\mathbb{A}\mathbb{A}^T\mathbf{t}}{2}\right) \equiv \\ &\exp\left(\mathbf{t}^T\boldsymbol{\mu}\right) \cdot \exp\left(\frac{\mathbf{t}^T\mathbb{V}\,\mathbf{t}}{2}\right)\end{aligned}$$

where $\mathbb{V} \equiv \mathbb{A}\mathbb{A}^T$ is the corresponding VARIANCE-COVARIANCE MATRIX (this can be verified directly). Note that there are many different $\mathbb{A}$'s resulting in the same $\mathbb{V}$. Also note that $\mathbf{Z} = \mathbb{A}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, which further implies that

$$(\mathbf{X}-\boldsymbol{\mu})^T(\mathbb{A}^{-1})^T\mathbb{A}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X}-\boldsymbol{\mu})^T(\mathbb{A}\mathbb{A}^T)^{-1}(\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X}-\boldsymbol{\mu})^T\mathbb{V}^{-1}(\mathbf{X}-\boldsymbol{\mu})$$

has the χ^2_n distribution.

The previous formulas hold even when $\mathbb{A}$ is a matrix with fewer rows than columns.

To generate a set of normally distributed random variables having a given variance-covariance matrix $\mathbb{V}$ requires us to solve for the corresponding $\mathbb{A}$ (Maple provides us with $\mathbf{Z}$ only, when typing: **stats[random,normald]**(20)). There is infinitely many such $\mathbb{A}$ matrices, one of them (easy to construct) is lower triangular.

Partial correlation coefficient

The variance-covariance matrix can be converted into the correlation matrix, whose elements are defined by:

$$\mathbb{C}_{ij} \equiv \frac{\mathbb{V}_{ij}}{\sqrt{\mathbb{V}_{ii} \cdot \mathbb{V}_{jj}}}$$

Clearly, the main diagonal elements of $\mathbb{C}$ are all equal to 1 (the correlation of X_i with itself).

Suppose we have three normally distributed random variables with a given variance-covariance matrix. The conditional distribution of X_2 and X_3 given that $X_1 = \underline{x}_1$ has a correlation coefficient independent of the value of $\underline{x}_1$. It is called the PARTIAL CORRELATION COEFFICIENT, and denoted $\rho_{23\,|1}$. Let us find its value in terms of the ordinary correlation coefficients..

Any correlation coefficient is independent of scaling. We can thus choose the three X's to be standardized (but *not* independent), having the following tree-dimensional PDF:

$$\frac{1}{\sqrt{(2\pi)^3 \det(\mathbb{C}^{-1})}} \cdot \exp\left(-\frac{\mathbf{x}^T \mathbb{C}^{-1} \mathbf{x}}{2}\right)$$

where

$$\mathbb{C} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Since the marginal PDF of X_1 is

$$\frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x_1^2}{2}\right)$$

the conditional PDF we need is

$$\frac{1}{\sqrt{(2\pi)^2 \det(\mathbb{C}^{-1})}} \cdot \exp\left(-\frac{\mathbf{x}^T \mathbb{C}^{-1} \mathbf{x} - x_1^2}{2}\right)$$

The information about the five parameters of the corresponding bi-variate distribution is in

$$\mathbf{x}^T \mathbb{C}^{-1} \mathbf{x} - x_1^2 =$$
$$\frac{\left(\frac{x_2 - \rho_{12} x_1}{\sqrt{1-\rho_{12}^2}}\right)^2 + \left(\frac{x_3 - \rho_{13} x_1}{\sqrt{1-\rho_{13}^2}}\right)^2 - 2\,\frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}} \left(\frac{x_2 - \rho_{12} x_1}{\sqrt{1-\rho_{12}^2}}\right)\left(\frac{x_3 - \rho_{13} x_1}{\sqrt{1-\rho_{13}^2}}\right)}{1 - \left(\frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}\right)^2}$$

which, in terms of the two conditional means and standard deviations agrees with what we know from MATH 2F96. The extra parameter is our partial correlation coefficient

$$\rho_{23\,|1} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}$$

Multiple Regression - Main Results

This time, we have k independent (regressor) variables $x_1,\ x_2,...,x_k$; still only one dependent (response) variable y. The model is

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i$$

with $i = 1,\ 2, ..., n$, where the first index labels the variable, and the second the observation. It is more convenient now to switch to using the following matrix notation

$$\mathbf{y} = \mathbb{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are (column) vectors of length n, $\boldsymbol{\beta}$ is a (column) vector of length $k+1$, and $\mathbb{X}$ is a n by $k+1$ matrix of observations (with its first column having all elements equal to 1, the second column being filled by the observed values of x_1, etc.). Note that the exact values of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are, and will always remain, unknown to us (thus, they must not appear in any of our computational formulas).

Also note that your textbook calls these β's PARTIAL correlation coefficients, as opposed to a TOTAL correlation coefficient of a simple regression (ignoring all but one of the independent variables).

To minimize the sum of squares of the residuals (a SCALAR quantity), namely

$$(\mathbf{y} - \mathbb{X}\,\boldsymbol{\beta})^T(\mathbf{y} - \mathbb{X}\,\boldsymbol{\beta}) =$$
$$\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\,\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbb{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}\,\boldsymbol{\beta}$$

(note that the second and third terms are identical - why?), we differentiate it with respect to each element of $\boldsymbol{\beta}$. This yields the following vector:

$$-2\mathbb{X}^T\mathbf{y} + 2\mathbb{X}^T\mathbb{X}\,\boldsymbol{\beta}$$

Making these equal to zero provides the following maximum likelihood (least square) estimators of the regression parameters:

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} \equiv \boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}$$

The last form makes it clear that $\widehat{\boldsymbol{\beta}}$ are *unbiased* estimators of $\boldsymbol{\beta}$, normally distributed with the variance-covariance matrix of

$$\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

The 'fitted' values of $\mathbf{y}$ (let us call them $\widehat{\mathbf{y}}$), are computed by

$$\widehat{\mathbf{y}} = \mathbb{X}\,\widehat{\boldsymbol{\beta}} = \mathbb{X}\ \boldsymbol{\beta} + \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon} \equiv \mathbb{X}\ \boldsymbol{\beta} + \mathbb{H}\ \boldsymbol{\varepsilon}$$

where $\mathbb{H}$ is clearly *symmetric* and *idempotent* (i.e. $\mathbb{H}^2 = \mathbb{H}$). Note that $\mathbb{H}\mathbb{X} = \mathbb{X}$.

This means that the residuals e_i are computed by

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}$$

($\mathbb{I} - \mathbb{H}$ is also idempotent). Furthermore, the covariance (matrix) between the elements of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and those of $\mathbf{e}$ is:

$$\mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{e}^T\right] = \mathbb{E}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})\right] =$$
$$(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right](\mathbb{I} - \mathbb{H}) = \mathbb{O}$$

which means that the variables are uncorrelated and therefore *independent* (i.e. each of the regression-coefficient estimators is independent of each of the residuals – slightly counter-intuitive but correct nevertheless).

The sum of squares of the residuals, namely $\mathbf{e}^T\mathbf{e}$, is equal to

$$\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})^T(\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}$$

Divided by σ^2:

$$\frac{\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}}{\sigma^2} \equiv \mathbf{Z}^T(\mathbb{I} - \mathbb{H})\mathbf{Z}$$

where $\mathbf{Z}$ are standardized, independent and normal.

We know (from matrix theory) that any symmetric matrix (including our $\mathbb{I}-\mathbb{H}$) can be written as $\mathbb{R}^T\mathbb{D}\,\mathbb{R}$, where $\mathbb{D}$ is *diagonal* and $\mathbb{R}$ is *orthogonal* (implying $\mathbb{R}^T \equiv \mathbb{R}^{-1}$). We can then rewrite the previous expression as

$$\mathbf{Z}^T\mathbb{R}^T\mathbb{D}\,\mathbb{R}\mathbf{Z} = \widetilde{\mathbf{Z}}^T\mathbb{D}\,\widetilde{\mathbf{Z}}$$

where $\widetilde{\mathbf{Z}} \equiv \mathbb{R}\mathbf{Z}$ is still a set of standardized, independent Normal random variables (since its variance-covariance matrix equals $\mathbb{I}$). Its distribution is thus χ^2 if and only if the diagonal elements of $\mathbb{D}$ are all equal either to 0 or 1 (the number of degrees being equal to the *trace* of $\mathbb{D}$).

How can we tell whether this is true for our $\mathbb{I} - \mathbb{H}$ matrix (when expressed in the $\mathbb{R}^T\mathbb{D}\,\mathbb{R}$ form) *without* actually performing the diagonalization (a fairly tricky process). Well, such a test is not difficult to design, once we notice that $(\mathbb{I}-\mathbb{H})^2 = \mathbb{R}^T\mathbb{D}\,\mathbb{R}\mathbb{R}^T\mathbb{D}\,\mathbb{R} = \mathbb{R}^T\mathbb{D}^2\,\mathbb{R}$. Clearly, $\mathbb{D}$ has the proper form (only 0 or 1 on the main diagonal) if and only if $\mathbb{D}^2 = \mathbb{D}$, which is the same as saying that $(\mathbb{I}-\mathbb{H})^2 = \mathbb{I}-\mathbb{H}$ (which we already know is true). This then implies that the sum of squares of the residuals has χ^2 distribution. Now, how about its degrees of freedom? Well, since the trace of $\mathbb{D}$ is the same as the trace of $\mathbb{R}^T\mathbb{D}\,\mathbb{R}$ (a well known property of trace), we just have to find the trace of $\mathbb{I} - \mathbb{H}$, by

$$\mathrm{Tr}\left[\mathbb{I} - \mathbb{H}\right] = \mathrm{Tr}\left(\mathbb{I}_{n\times n}\right) - \mathrm{Tr}\left(\mathbb{H}\right) = n - \mathrm{Tr}\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right) =$$
$$n - \mathrm{Tr}\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\right) = n - \mathrm{Tr}\left(\mathbb{I}_{(k+1)\times(k+1)}\right) = n - (k+1)$$

i.e. the number of observations minus the number of regression coefficients.

The sum of squares of the residuals is usually denoted SS_E (for 'error' sum of squares, even though it is usually called RESIDUAL SUM OF SQUARES) and

computed by

$$(\mathbf{y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbb{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbb{X}\widehat{\boldsymbol{\beta}} =$$
$$= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} \equiv$$
$$\mathbf{y}^T\mathbf{y} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y}$$

We have just proved that it has the χ^2 distribution with $n-(k+1)$ degrees of freedom, and is independent of $\widehat{\boldsymbol{\beta}}$. A related definition is that of a RESIDUAL (error) MEAN SQUARE

$$MS_E \equiv \frac{SS_E}{n-(k+1)}$$

This would clearly be our unbiased estimator of σ^2.

```
> with(linalg): with(stats): with(plots):
> x1 := [2,1,8,4,7,9,6,9,2,10,6,4,8,1,5,6,7]:
> x2 := [62,8,50,87,99,67,10,74,82,75,67,74,43,92,94,1,12]:
> x3 := [539,914,221,845,566,392,796,475,310,361,383,593,614,278,750,336,262]:
> y := [334,64,502,385,537,542,222,532,450,594,484,392,392,455,473,283,344]:
> X := matrix(17,1,1.):
> X := augment(X,x1,x2,x3):
> C := evalm(inverse(transpose(X)&*X)):
> beta := evalm(C&* transpose(X)&*y);
        β := [215.2355338, 22.41975192, 3.030186331, –0.2113464404]
> e := evalm(X&*beta - y):
> MSe := sum(e['i']^2,'i'= 1..17)/13;
        MSe := 101.9978001
> for i to 4 do sqrt(C[i,i]*MSe) od;
        10.83823625
         0.9193559350
         0.07784126745
         0.01214698750
```

Various standard errors

We would thus construct a confidence interval for any one of the β coefficients, say β_j, by

$$\widehat{\beta}_j \pm \mathsf{t}_{\frac{\alpha}{2},n-k-1}\cdot\sqrt{C_{jj}\cdot MS_E}$$

where $\mathbb{C} \equiv (\mathbb{X}^T\mathbb{X})^{-1}$.

Similarly, to test a hypothesis concerning a single β_j, we would use

$$\frac{\widehat{\beta}_j - \beta_{i_0}}{\sqrt{C_{jj}\cdot MS_E}}$$

as the test statistic.

Since the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$, we know that

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbb{X}^T\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2}$$

has the χ^2_{k+1} distribution. Furthermore, since the β's are independent of the residuals,

$$\frac{\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbb{X}^T\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{k+1}}{\frac{SS_E}{n-k-1}}$$

must have the $\mathsf{F}_{k+1,n-k-1}$ distribution. This enables us to construct confidence ellipses (ellipsoids) simultaneously for all parameters or, correspondingly, perform a single test of H_0: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$.

To estimate $\mathbb{E}(y_0)$, where y_0 is the value of the response variable when we choose a brand new set of x values (let us call them $\mathbf{x}_0$), we will of course use

$$\widehat{\boldsymbol{\beta}}^T\mathbf{x}_0$$

which yields an unbiased estimator, with the variance of

$$\sigma^2\,\mathbf{x}_0^T\,(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_0$$

(recall the general formula for a variance a linear combination of random variables). To construct a corresponding confidence interval, we need to replace σ^2 by MS_E:

$$\widehat{\boldsymbol{\beta}}^T\mathbf{x}_0 \pm \mathsf{t}_{\frac{\alpha}{2},n-k-1}\cdot\sqrt{\mathbf{x}_0^T\,\mathbb{C}\,\mathbf{x}_0\cdot MS_E}$$

Predicting the actual value of y_0, one has to include the ε variance (as in the univariate case).:

$$\widehat{\boldsymbol{\beta}}^T\mathbf{x}_0 \pm \mathsf{t}_{\frac{\alpha}{2},n-k-1}\cdot\sqrt{(1+\mathbf{x}_0^T\,\mathbb{C}\,\mathbf{x}_0)\cdot MS_E}$$

Weighted-case modifications

When the variance of ε_i equals $\frac{\sigma^2}{w_i}$ or, equivalently, when the variance-covariance matrix of $\boldsymbol{\varepsilon}$ is given by

$$\sigma^2\mathbb{W}^{-1}$$

where $\mathbb{W}$ is a matrix with the w_i's on the main diagonal and 0 everywhere else, since the ε_i's remain independent (we could actually have them correlated, if that was the case).

The maximum likelihood technique now leads to minimizing the *weighted* sum of squares of the residuals, namely

$$SS_E \equiv (\mathbf{y}-\mathbb{X}\,\boldsymbol{\beta})^T\,\mathbb{W}(\mathbf{y}-\mathbb{X}\,\boldsymbol{\beta})$$

yielding

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}\mathbf{y} \equiv \boldsymbol{\beta} + (\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}\boldsymbol{\varepsilon}$$

This implies that the corresponding variance-covariance matrix is now equal to

$$(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}(\sigma^2\mathbb{W}^{-1})\mathbb{W}\mathbb{X}(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}$$

The $\mathbb{H}$ matrix is defined by

$$\mathbb{H} \equiv \mathbb{X}(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}$$

(idempotent but no longer symmetric). One can then show that $\boldsymbol{\beta}$ and $\mathbf{e}$ remain uncorrelated (thus independent) since

$$(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right](\mathbb{I}-\mathbb{H}^T)=\mathbb{O}$$

Furthermore, SS_E can be now reduced to

$$\begin{aligned}&\boldsymbol{\varepsilon}^T(\mathbb{I}-\mathbb{H})^T\,\mathbb{W}(\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon}=\\&\sigma^2\cdot\mathbf{Z}^T\,\mathbb{W}^{-1/2}(\mathbb{I}-\mathbb{H})^T\,\mathbb{W}(\mathbb{I}-\mathbb{H})\mathbb{W}^{-1/2}\mathbf{Z}\end{aligned}$$

Since

$$\mathbb{W}^{-1/2}(\mathbb{I}-\mathbb{H})^T\,\mathbb{W}(\mathbb{I}-\mathbb{H})\mathbb{W}^{-1/2}=\mathbb{I}-\mathbb{W}^{1/2}\,\mathbb{X}(\mathbb{X}^T\,\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{W}^{1/2}$$

is symmetric, idempotent, and has the trace equal to $n-(k+1)$, $\dfrac{SS_E}{\sigma^2}$ still has the $\chi^2_{n-(k+1)}$ distribution (and is independent of $\boldsymbol{\beta}$).

Redundancy Test

Having more than one independent variable, we may start wondering whether some of them (especially in combination with the rest) are redundant and can be eliminated without a loss of the model's predictive powers. In this section, we design a way of testing this. We will start with the full (UNRESTRICTED) MODEL, then select one or more independent variables which we believe can be eliminated (by setting the corresponding β equal to 0). The latter (the so called RESTRICTED or reduced MODEL) constitutes our null hypothesis. The corresponding alternate hypothesis is the usual "not so" set of alternatives, meaning that at least one of the β (i.e. not necessarily all) of the null hypothesis is nonzero.

The way to carry out the test is to first compute SS_E for both the full and restricted model. (let us call the answers SS_E^{full} and SS_E^{rest} respectively). Clearly, SS_E^{full} must be smaller that SS_E^{rest} (with more independent variables, the fit can only improve). Furthermore, one can show that SS_E^{full}/σ^2 and $(SS_E^{rest}-SS_E^{full})/\sigma^2$ are, under the assumptions of the *null* hypothesis, *independent*, χ^2 distributed, with $n-(k+1)$ and $k-\ell$ degrees of freedom respectively (where k is the number of independent variables in the full model, and ℓ tell us how many of them are left in the restricted model).

Proof: Let us recall the definition of $\mathbb{H}\equiv\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ (symmetric and idempotent). We can now compute two of these (for the full and restricted model), say $\mathbb{H}_{full}$ and $\mathbb{H}_{rest}$. Clearly, $SS_E^{full}=\mathbf{y}^T(\mathbb{I}-\mathbb{H}_{full})\,\mathbf{y}$ and $SS_E^{rest}=\mathbf{y}^T(\mathbb{I}-\mathbb{H}_{rest})\mathbf{y}$. Also,

$$\mathbb{X}_{rest}=\mathbb{X}_{full\downarrow}$$

where $\downarrow$ implies dropping the last $k-\ell$ columns. Now

$$\mathbb{X}_{full}(\mathbb{X}_{full}^T\mathbb{X}_{full})^{-1}\mathbb{X}_{full}^T\mathbb{X}_{rest}=\mathbb{X}_{full}(\mathbb{X}_{full}^T\mathbb{X}_{full})^{-1}\mathbb{X}_{full}^T\mathbb{X}_{full\downarrow}=\mathbb{X}_{full\downarrow}=\mathbb{X}_{rest}$$

since $\mathbb{A}\mathbb{B}_{\downarrow}=(\mathbb{A}\mathbb{B})_{\downarrow}$. We thus have

$$\mathbb{X}_{full}(\mathbb{X}_{full}^T\mathbb{X}_{full})^{-1}\mathbb{X}_{full}^T\mathbb{X}_{rest}(\mathbb{X}_{rest}^T\mathbb{X}_{rest})^{-1}\mathbb{X}_{rest}^T=\mathbb{X}_{rest}(\mathbb{X}_{rest}^T\mathbb{X}_{rest})^{-1}\mathbb{X}_{rest}^T$$

or

$$\mathbb{H}_{full}\,\mathbb{H}_{rest} = \mathbb{H}_{rest}$$

Taking the transpose immediately shows that, also

$$\mathbb{H}_{rest} = \mathbb{H}_{full}\,\mathbb{H}_{rest}$$

We already know why SS_E^{full}/σ^2 has the χ^2_{n-k-1} distribution: because $\mathbb{I}-\mathbb{H}_{full}$ is idempotent, with trace of $n-k-1$, *and* $(\mathbb{I}-\mathbb{H}_{full})\,\mathbf{y} = (\mathbb{I}-\mathbb{H}_{full})\,\boldsymbol{\varepsilon}$. We will now show that $(SS_E^{rest} - SS_E^{full})/\sigma^2$ has the $\chi^2_{k-\ell}$ distribution:

The null hypothesis

$$\mathbf{y} = \mathbb{X}_{rest}\,\boldsymbol{\beta}_{rest} + \boldsymbol{\varepsilon}$$

implies that

$$\begin{aligned}(\mathbb{H}_{full} - \mathbb{H}_{rest})\,\mathbf{y} = {} & (\mathbb{H}_{full}\mathbb{X}_{rest} - \mathbb{H}_{rest}\mathbb{X}_{rest})\boldsymbol{\beta}_{rest} + (\mathbb{H}_{full} - \mathbb{H}_{rest})\,\boldsymbol{\varepsilon} = \\ & (\mathbb{X}_{rest} - \mathbb{X}_{rest})\boldsymbol{\beta}_{rest} + (\mathbb{H}_{full} - \mathbb{H}_{rest})\,\boldsymbol{\varepsilon} = (\mathbb{H}_{full} - \mathbb{H}_{rest})\,\boldsymbol{\varepsilon}\end{aligned}$$

$\mathbb{H}_{full} - \mathbb{H}_{rest}$ is idempotent, as

$$(\mathbb{H}_{full} - \mathbb{H}_{rest})(\mathbb{H}_{full} - \mathbb{H}_{rest}) = \mathbb{H}_{full} - \mathbb{H}_{rest} - \mathbb{H}_{rest} + \mathbb{H}_{rest} = \mathbb{H}_{full} - \mathbb{H}_{rest}$$

and

$$\begin{aligned}&\text{Trace}(\mathbb{H}_{full} - \mathbb{H}_{rest}) = \text{Trace}(\mathbb{H}_{full}) - \text{Trace}(\mathbb{H}_{rest}) = \\ &\text{Trace}(\mathbb{I}_{full}) - \text{Trace}(\mathbb{I}_{rest}) = (k+1) - (\ell+1) = k - \ell\end{aligned}$$

Finally, we need to show that $\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H}_{full})\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^T(\mathbb{H}_{full} - \mathbb{H}_{rest})\boldsymbol{\varepsilon}$ are independent. Since the two matrices are symmetric and commute, i.e.

$$(\mathbb{I} - \mathbb{H}_{full})(\mathbb{H}_{full} - \mathbb{H}_{rest}) = (\mathbb{H}_{full} - \mathbb{H}_{rest})(\mathbb{I} - \mathbb{H}_{full}) = \mathbb{O}$$

they can be diagonalized by the *same* orthogonal transformation. This implies that SS_E^{full}/σ^2 and $(SS_E^{rest} - SS_E^{full})/\sigma^2$ can be expressed as $\widetilde{\mathbf{Z}}^T\mathbb{D}_1\,\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{Z}}^T\mathbb{D}_2\,\widetilde{\mathbf{Z}}$ respectively (using the same $\widetilde{\mathbf{Z}}$). Furthermore, since $\mathbb{D}_1$ and $\mathbb{D}_2$ remain idempotent, the issue of independence of the two quadratic forms (as they are called) is reduced to asking whether $\mathbb{D}_1\,\widetilde{\mathbf{Z}}$ and $\mathbb{D}_2\,\widetilde{\mathbf{Z}}$ are independent or not. Since their covariance matrix is $\mathbb{D}_1\mathbb{D}_2$, independence is guaranteed by $\mathbb{D}_1\mathbb{D}_2 = \mathbb{O}$. This is equivalent to $(\mathbb{I} - \mathbb{H}_{full})(\mathbb{H}_{full} - \mathbb{H}_{rest}) = \mathbb{O}$, which we already know to by true. □

Knowing all this enables us to test the null hypothesis, based on the following test statistic:

$$\frac{\dfrac{SS_E^{rest} - SS_E^{full}}{k-\ell}}{\dfrac{SS_E^{full}}{n-(k+1)}}$$

whose distribution (under the null hypothesis) is $\mathsf{F}_{k-\ell,n-k-1}$. When the null hypothesis is wrong (i.e. at least one of the independent variables we are trying to delete

is effecting the outcome of y), the numerator of the test statistic becomes unusually 'large'. The corresponding test will thus always have only one (right-hand) 'tail' (rejection region). The actual critical value (deciding how large is 'large') can be looked up in tables of the F distribution (or we can ask Maple).

At one extreme. we can try deleting *all* independent variables, to see whether any of them are relevant, at the other extreme we can test whether a specific *single* x_j can be removed from the model without effecting its predictive powers. In the latter case, our last test is equivalent to the usual (two-tail) t-test of $\beta_j = 0$.

Later on, we will tackle the issue of removing, one by one, all irrelevant independent variables from the model.

Searching for Optimal Model

Realizing that some of the independent variables may be irrelevant for y (by either being totally unrelated to y, or duplicating the information contained in the remaining x's), we would normally (especially when the original number of x's is large) like to eliminate them from our model. But that is a very tricky issue, even when we want to properly define what the 'best' simplest model should look like.

Deciding to make SS_E as small as possible will not do any good - we know that including a new x (however phoney) will always achieve some small reduction in SS_E. Trying to keep only the statistically significant x's is also quite difficult, as the significance of a specific independent variable depends (often quite strongly) on what other x's included or excluded (e.g. if we include two nearly identical x's, individually, they will appear totally insignificant, but as soon as we remove one of them, the other may be highly significant and must stay as part of the model).

We will thus take a practical approach, and learn several procedures which should get us reasonably close to selecting the 'best' subset of the independent variables to be kept in the model (the others will be simply discarded as irrelevant), even without properly defining what 'best' means. The basic two are

1. Backward elimination: Starting with the full model, we eliminate the x with the smallest $\mathtt{t} = \frac{\widehat{\beta}_j}{\sqrt{C_{jj} \cdot MS_E}}$ value, assuming this t is non-significant (using a specific α). This is repeated until all t values are significant, at which point we stop and keep all the remaining x's.

2. Forward selection: Using k models, each with a single x, we select the one with the highest t. Then we try all $k-1$ models having this x, and one of the remaining ones (again, including the most significant of these). In this manner we keep on extending the model by one x at a time, until all remaining x's prove non-significant (at some fixed level of significance - usually 5%).

Each of these two procedures can be made a bit more sophisticated by checking, after each elimination (selection), whether any of the previously eliminated (selected) independent variables have become significant (non-significant), in which case they would be included (removed) in (from) the model. The trouble is that some x may then develop a nasty habit of not being able to make up their mind, and we start running in circles by repeatedly including and excluding them. One

can take some preventive measures against that possibility (by requiring higher significance for inclusion than for kicking a variable out), but we will not go into these details. We will just mention that this modification is called STEPWISE (stagewise) elimination (selection). In this course, the procedure of choice will be backward elimination.

We will use data of our previous example, with the exception of the values of y:

$>$ $y := [145, 42, 355, 123, 261, 332, 193, 316, 184, 371, 283, 171, 270, 180, 188, 276, 319]$:
$>$ $X :=$ **matrix**$(17, 1, 1.)$:
$>$ $X :=$ **augment**$(X, x1, x2, x3)$:
$>$ $C :=$ **evalm**(**inverse**(**transpose**(X)&*X)):
$>$ $beta :=$ **evalm**$(C$&* **transpose**(X)&*$y)$;
$\beta :=$ [204.8944465, 23.65441498, 0.0250321373, –0.2022388198]
$>$ $e :=$**evalm**$(X$&*$beta - y)$:
$>$ $MSe :=$**sum**$(e$['i']^2,'i'$= 1..17)/13$;
MSe := 62.41228512
$>$ **for** i **to** 4 **do** $beta[i]/sqrt(C[i, i] * MSe)$ **od**;
24.16751153
32.89189446
0.4111008683
–21.28414937
$>$ **statevalf**[**icdf**,**studentst**[13]](0.975);
2.160368656
$>$ $X :=$ **submatrix**$(X, 1..17, [1, 2, 4])$:

In the last command, we deleted the variable with the smallest (absolute) value of t, since it is clearly nonsignificant (compared to the corresponding critical value). We then have to go back to recomputing C etc., until all remaining t values are significant.

Coefficient of Correlation (Determination)

The multiple correlation coefficient (usually called R) is computed in the manner of (3.5) between the observed ($\mathbf{y}$) and 'predicted' ($\widehat{\mathbf{y}} = \mathbb{H}\,\mathbf{y}$) values of the response variable.

First we show that $\widehat{\mathbf{y}}$ has the same mean (or, equivalently, total) as $\mathbf{y}$. This can be seen from

$$\mathbf{1}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} = \mathbf{y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{1} = \mathbf{y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}_{\downarrow} = \mathbf{y}^T\mathbb{X}_{\downarrow} = \mathbf{y}^T\,\mathbf{1}$$

where $\mathbf{1}$ is a column vector of length n with each component equal to 1, and $\mathbb{X}_{\downarrow}$ means deleting all columns of $\mathbb{X}$ but the first one (equal to $\mathbf{1}$).

If we call the corresponding mean (of $\mathbf{y}$ and $\widehat{\mathbf{y}}$) $\bar{y}$, the correlation coefficient between the two is computed by

$$R = \frac{\mathbf{y}^T\mathbb{H}\,\mathbf{y} - \frac{\bar{y}^2}{n}}{\sqrt{\left(\mathbf{y}^T\mathbb{H}^2\mathbf{y} - \frac{\bar{y}^2}{n}\right)\left(\mathbf{y}^T\mathbf{y} - \frac{\bar{y}^2}{n}\right)}}$$

Since $\mathbb{H}$ is idempotent, this equals

$$\sqrt{\frac{\mathbf{y}^T\mathbb{H}\mathbf{y} - \frac{\bar{y}^2}{n}}{\mathbf{y}^T\mathbf{y} - \frac{\bar{y}^2}{n}}}$$

R^2 defines the COEFFICIENT OF DETERMINATION, which is thus equal to

$$\frac{\mathbf{y}^T\mathbb{H}\mathbf{y} - \frac{\bar{y}^2}{n}}{\mathbf{y}^T\mathbf{y} - \frac{\bar{y}^2}{n}} = \frac{\mathbf{y}^T\mathbf{y} - \frac{\bar{y}^2}{n} - \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{H}\mathbf{y}\right)}{\mathbf{y}^T\mathbf{y} - \frac{\bar{y}^2}{n}} = \frac{S_{yy} - SS_E}{S_{yy}} \equiv \frac{SS_R}{S_{yy}}$$

where SS_R is the (due to) REGRESSION SUM OF SQUARES (a bit confusing, since SS_E is called *residual* sum of squares). It is thus best to remember the formula in the following form:

$$R^2 = 1 - \frac{SS_E}{S_{yy}}$$

It represents the proportion of the original S_{yy} removed by regression.

Polynomial Regression

This is a special case of multivariate regression, with only one independent variable x, but an x-y relationship which is clearly nonlinear (at the same time, there is no 'physical' model to rely on). All we can do in this case is to try fitting a polynomial of a sufficiently high degree (which is ultimately capable of mimicking any curve), i.e.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_k x_i^k + \varepsilon_i$$

Effectively, this is the same as having a multivariate model with $x_1 \equiv x$, $x_2 \equiv x^2$, $x_3 \equiv x^3$, etc., or, equivalently

$$\mathbb{X} \equiv \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ 1 & x_3 & x_3^2 & \cdots & x_3^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}$$

All formulas of the previous section apply unchanged. The only thing we may like to do slightly differently is our backward elimination: In each step, we will always compute the `t` value corresponding to the currently *highest* degree of x only, and reduce the polynomial correspondingly when this `t` turns out to be non-significant. We continue until we encounter a significant power of x, stopping at that point. This clearly simplifies the procedure, and appears quite reasonable in this context.

$> x := [3, 19, 24, 36, 38, 39, 43, 46, 51, 58, 61, 84, 89]:$
$> y := [151, 143, 155, 119, 116, 127, 145, 110, 112, 118, 78, 11, 5]:$
$> k := 5:$
$> X := \mathbf{matrix}(13, k):$
$>$ **for** i **to** 13 **do** **for** j **to** k **do** $X[i, j] := x[i]\hat{}(j - 1.)$ **od od**:
$> C := \mathbf{inverse}(\mathbf{transpose}(X)\&*X):$
$> beta := \mathbf{evalm}(C\&*\ \mathbf{transpose}(X)\&*y):$

```
> e := evalm(X&*beta - y):
> MSe := sum(e['i']^2,'i'= 1..13)/(13 - k):
> beta[k]/sqrt(MSe * C[k, k]);
                    -4.268191026
> statevalf[icdf,studentst[13 - k]](.975);
                     2.228138852
> k := k - 1:
> pl1 := pointplot([seq([x[i], y[i]], i = 1..13)]):
> pl2 := plot(beta[1] + beta[2] * z + beta[3] * z^2, z = 0..90):
> display(pl1, pl2);
```

We have to execute the $X :=$ **matrix**$(13, k)$ to $k := k-1$ loop until the resulting `t` value becomes significant (which, in our program, happened when we reached the quadratic coefficient).

Similarly to the simple-regression case, we should never use the resulting equation with an x outside the original (fitted) data (the so called extrapolation). This maxim becomes increasingly more imperative with higher-degree polynomials - extrapolation yields totally nonsensical answers even for relatively 'nearby' values of x.

Dummy (Indicator) Variables

Some of our independent variables may be of the 'binary' (yes or no) type. This again poses no particular problem: the yes-no (true-false, male-female, etc.) values must be translated into a numerical code (usually 0 and 1), and can be then treated as any other independent variable of the multivariate regression (and again: non of the basic formulas change). In this context, any such x is usually called an INDICATOR variable (indicating whether the subject is a male or a female).

There are other instances when we may introduce a DUMMY variable (or two) of this type on our own. For instance, we may have two sets of data (say, between the age and salary), one for the male, the other for female employees of a company. We know how to fit a straight line for each set of data, but how do we test whether the two slopes and intercepts are identical?

Assuming that the errors (ε_i) of both models have the same σ, we can pool them together, if in addition to salary (y) and age (x), we also include a 0-1 type variable (say s) which keeps track of the employee's sex. Our new (multivariate) model then reads:

$$y = \beta_0 + \beta_1 x + \beta_2 s + \beta_3 x\, s + \varepsilon$$

which means that, effectively, $x_1 \equiv x$, $x_2 \equiv s$ and $x_3 \equiv x\, s$ (the product of x and s). Using the usual multivariate regression, we can now find the best (least-square) estimates of the four regression coefficients. The results must be the same as performing, separately, two simple regressions, in the following sense:

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

(using our multivariate-fit $\widehat{\beta}$'s) will agree with the male simple regression (assuming males were coded as 0), and

$$y = (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3)\, x$$

will agree with the female simple regression. So that by itself is no big deal. But now we can easily test for identical slopes ($\beta_3 = 0$) or intercepts ($\beta_2 = 0$), by carrying out the usual multivariate procedure. Furthermore, we have a choice of performing these two tests individually or, if we like. 'collectively' (i.e. testing whether the two *straight lines* are in any way different) - this of course would have to be done by computing the full and reduced SS_E, etc. One further advantage of this approach is that we would be pooling the data and thus combining (adding) the degrees of freedom of the residual sum of squares (this always makes the corresponding test more sensitive and reliable).

Example: We will test whether two sets of x-y data can be fitted by the same straight line or not.

$> x1 := [21, 30, 35, 37, 38, 38, 44, 44, 51, 64]:$
$> x2 := [20, 36, 38, 40, 54, 56, 56, 57, 61, 62]:$
$> y1 := [22, 20, 28, 34, 40, 24, 35, 33, 44, 59]:$
$> y2 := [16, 28, 33, 26, 40, 39, 43, 41, 52, 46]:$
$> pl1 := \mathbf{pointplot}([\mathbf{seq}([x1[i], y1[i]], i = 1..10)]):$
$> pl2 := \mathbf{pointplot}([\mathbf{seq}([x2[i], y2[i]], i = 1..10)], \mathbf{color} = \mathbf{red}):$
$> \mathbf{display}(pl1, pl2);$
$> x := [op(x1), op(x2)]:$
$> y := [op(y1), op(y2)]:$
$> s := [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]:$
$> xs := [\mathbf{seq}(x[i] * s[i], i = 1..20)]:$
$> X := \mathbf{matrix}(20, 1, 1.):$
$> X := \mathbf{augment}(X, x, s, xs):$
$> beta := \mathbf{evalm}(\mathbf{inverse}(\mathbf{transpose}(X)\&*X)\&*\mathbf{transpose}(X)\&*y):$
$> e := \mathbf{evalm}(X\&*beta - y):$
$> SSeFull := \mathbf{sum}(e[i]\hat{}2, i = 1..20);$
SSeFull := 316.4550264
$> X := \mathbf{matrix}(20, 1, 1.):$
$> X := \mathbf{augment}(X, x):$
$> beta := \mathbf{evalm}(\mathbf{inverse}(\mathbf{transpose}(X)\&*X)\&* \mathbf{transpose}(X)\&*y):$
$> e :=\mathbf{evalm}(X\&*beta - y):$
$> SSeRest := \mathbf{sum}(e[i]\hat{}2, i = 1..20);$
SSeRest := 402.0404903
$> ((SSeRest - SSeFull)/2)/(SSeFull/(20 - 4));$
2.163605107
$> \mathbf{statevalf}[\mathbf{icdf}, \mathbf{fratio}[2, 6]](0.95);$
3.633723468

Since the resulting $\mathsf{F}_{2,16}$ value is nonsignificant, the two sets of data can be fitted by a single straight line.

Linear Versus Nonlinear Models

One should also realize that the basic (multi)-linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

covers many situations which at first may appear non-linear, such as, for example

$$\begin{aligned} y &= \beta_0 + \beta_1 e^{-t} + \beta_2 \ln t + \varepsilon \\ y &= \beta_0 + \beta_1 e^{-t} + \beta_2 \ln p + \varepsilon \end{aligned}$$

where t (in the first model), and t and p (in the second one) are the independent variables (all we would have to do is to take $x_1 \equiv e^{-t}$ and $x_2 \equiv \ln t$ in the first case, and $x_1 \equiv e^{-t}$ and $x_2 \equiv \ln p$ in the second case, and we are back in business. The important thing to realize is that 'linear' means linear in each of the β's, not necessarily linear in x.

A slightly more difficult situation is

$$v = a \cdot b^x$$

where v is the dependent and x the independent variable. We can transform this to a linear model by taking the logarithm of the equation

$$\ln v = \ln a + x \cdot \ln b$$

which represents a simple linear model if we take $y \equiv \ln v$, $\beta_0 \equiv \ln a$ and $\beta_1 \equiv \ln b$. The only trouble is that we have to assume the errors to be normally distributed (with the same σ) *after* the transformation (making the assumptions about errors rather complicated in the original model).

Of course, there are models which will remain essentially non-linear no matter how we transform either the independent or the dependent variable (or both), e.g.

$$y = \frac{a}{b + x}$$

We will now learn how to deal with these.

Chapter 5 NONLINEAR REGRESSION

We will assume the following model with one independent variable (the results can be easily extended to several) and k unknown parameters, which we will call b_1, b_2, ... b_k:

$$y = f(x, \mathbf{b}) + \varepsilon$$

where $f(x, \mathbf{b})$ is a specific (given) function of the independent variable and the k parameters.

Similarly to linear models, we find the 'best' estimators of the parameters by minimizing

$$\sum_{i=1}^{n} [y_i - f(x_i, \mathbf{b})]^2 \tag{5.1}$$

The trouble is that the normal equations

$$\sum_{i=1}^{n} [y_i - f(x_i, \mathbf{b})] \cdot \frac{\partial f(x_i, \mathbf{b})}{\partial b_j} = 0$$

($j = 1, 2, ...k$) are now non-linear in the unknowns, and thus fairly difficult to solve. The first two terms of the Taylor expansion (in terms of $\mathbf{b}$) of the left hand side, at some arbitrary point $\mathbf{b}_0$ (close to the exact solution), are

$$\sum_{i=1}^{n} [y_i - f(x_i, \mathbf{b}_0)] \cdot \left.\frac{\partial f(x_i, \mathbf{b})}{\partial b_j}\right|_{\mathbf{b}=\mathbf{b}_0} + \tag{5.2}$$

$$\left(\sum_{i=1}^{n} [y_i - f(x_i, \mathbf{b}_0)] \cdot \left.\frac{\partial^2 f(x_i, \mathbf{b})}{\partial b_j \partial b_\ell}\right|_{\mathbf{b}=\mathbf{b}_0} - \sum_{i=1}^{n} \left.\frac{\partial f(x_i, \mathbf{b})}{\partial b_\ell}\right|_{\mathbf{b}=\mathbf{b}_0} \cdot \left.\frac{\partial f(x_i, \mathbf{b})}{\partial b_j}\right|_{\mathbf{b}=\mathbf{b}_0} \right) (b_\ell - b_{\ell_0}) + ...$$

One can show that the first term in (big) parentheses is a lot smaller than the second term; furthermore, it would destabilize the iterative solution below. It is thus to our advantage to drop it (this also saves us computing the second derivatives). Making the previous expansion (without the offensive term) equal to zero, and solving for $\mathbf{b}$, yields

$$\mathbf{b} = \mathbf{b}_0 + (\mathbb{X}_0^T \mathbb{X}_0)^{-1} \mathbb{X}_0^T \mathbf{e}_0 \tag{5.3}$$

where

$$\mathbb{X}_0 \equiv \left[\begin{array}{cccc} \dfrac{\partial f(x_1, \mathbf{b})}{\partial b_1} & \dfrac{\partial f(x_1, \mathbf{b})}{\partial b_2} & \cdots & \dfrac{\partial f(x_1, \mathbf{b})}{\partial b_k} \\ \dfrac{\partial f(x_2, \mathbf{b})}{\partial b_1} & \dfrac{\partial f(x_2, \mathbf{b})}{\partial b_2} & \cdots & \dfrac{\partial f(x_2, \mathbf{b})}{\partial b_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f(x_n, \mathbf{b})}{\partial b_1} & \dfrac{\partial f(x_n, \mathbf{b})}{\partial b_2} & \cdots & \dfrac{\partial f(x_n, \mathbf{b})}{\partial b_k} \end{array} \right]_{\mathbf{b}=\mathbf{b}_0}$$

is the matrix of all k partial derivatives, evaluated at each value of x, and

$$\mathbf{e}_0 \equiv \left[\begin{array}{c} y_1 - f(x_1, \mathbf{b}_0) \\ y_2 - f(x_2, \mathbf{b}_0) \\ \vdots \\ y_2 - f(x_2, \mathbf{b}_0) \end{array} \right]$$

is the vector of residuals.

The standard (numerical) technique for solving them ITERATIVELY is called Levenberg-Marquardt, and it works as follows:

1. We start with some arbitrary (but reasonably sensible) INITIAL values of the unknown parameters, say $\mathbf{b}_0$. We also choose (quite arbitrarily) the first value of an iteration parameter to be $\lambda = 1$.

2. Slightly modifying (5.3), we compute a better approximation to the solution by
$$\mathbf{b}_1 = \mathbf{b}_0 + (\mathbb{X}_0^T\mathbb{X}_0 + \lambda\,\mathrm{diag}\mathbb{X}_0^T\mathbb{X}_0)^{-1}\mathbb{X}_0^T\mathbf{e}_0 \tag{5.4}$$
where 'diag' keeps the main-diagonal elements of its argument, making the rest equal to 0 (effectively, this says: multiply the diagonal elements of $\mathbb{X}_0^T\mathbb{X}_0$ by $1+\lambda$). If the sum of squares (5.1) increases, multiply λ by 10 and backtrack to $\mathbf{b}_0$, if it decreases, reduce λ by a factor of 10 and accept $\mathbf{b}_1$ as you new solution.

3. Recompute (5.4) with the new λ and, possibly, new $\mathbf{b}$, i.e.
$$\mathbf{b}_2 = \mathbf{b}_1 + (\mathbb{X}_1^T\mathbb{X}_1 + \lambda\,\mathrm{diag}\mathbb{X}_1^T\mathbb{X}_1)^{-1}\mathbb{X}_1^T\mathbf{e}_1$$
where $\mathbb{X}$ and $\mathbf{e}$ are now to be evaluated using $\mathbf{b}_1$ (assuming it was accepted in the previous step). Again, check whether this improved the value of (5.1), and accordingly accept (reject) the new $\mathbf{b}$ and adjust the value of λ.

4. Repeat these steps (iterations) until the value of (5.1) no longer decreases (within say 5 significant digits). At that point, compute $(\mathbb{X}^T\mathbb{X})^{-1}$ using the latest $\mathbf{b}$ and $\lambda = 0$.

Note that by choosing a large value of λ, the procedure will follow the direction of STEEPEST DESCENT (in a certain scale), a foolproof but inefficient way of minimizing a function. On the other hand, when λ is small (or zero), the procedure follows Newton's technique for solving nonlinear equations - fast (quadratically) converging to the exact solution *provided* we are reasonably close to it (but going crazy otherwise). So, Levenberg-Marquardt is trying to be conservative when things are not going too well, and advance rapidly when zeroing in on a nearby solution. The possibility of reaching a local (i.e. 'false') rather then global minimum is, in these cases, rather remote (normally, there is only one unique minimum); furthermore, one would readily notice it by graphing the results.

At this point, we may give (5.2) a slightly different interpretation: If we replace $\mathbf{b}_0$ by the exact (albeit unknown) values of the parameters and $\mathbf{b}$ by our least-square estimators, the e_i residuals become the actual ε_i errors, and the equation implies (using a somehow more sophisticated version of the large-sample theory):
$$\widehat{\mathbf{b}} = \mathbf{b} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon} + ...$$
indicating that $\widehat{\mathbf{b}}$ is an *asymptotically* unbiased estimator of $\mathbf{b}$, with the variance-covariance matrix of
$$(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]\,\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

The best estimator of σ^2 is, clearly

$$\frac{\sum_{i=1}^{n}\left[y_i - f(x_i, \widehat{\mathbf{b}})\right]^2}{n-k}$$

The previous formulas apply, practically without change (we just have to replace x by $\mathbf{x}$) to the case of more than one independent variable. So, the complexity of the problem (measured by the second dimension of $\mathbb{X}$) depends on the number of *parameters*, not on the number of the independent variables - for the linear model, the two numbers were closely related, but now anything can happen.

Example Assuming the following model

$$y = \frac{b_1}{b_2 + x} + \varepsilon$$

and being given the following set of observations:

x_i	82	71	98	64	77	39	86	69	22	10
y_i	.21	.41	.16	.43	.16	.49	.14	.34	.77	1.07

56	64	58	61	75	86	17	62	8
.37	.27	.29	.12	.24	.07	.64	.40	1.15

we can find the solution using the following Maple program:

```
> with(linalg):
> x := [82, 71, ....]:
> y := [.21, .41, .....]:
> f := [seq(b[1]/(b[2] + x[i]), i = 1..19)]:
> X := augment(diff(f, b[1]),diff(f, b[2])):
> b := [1, 1]:
> evalm((y - f)&*(y - f));
      4.135107228
> λ := 1 :
> bs := evalm(b):
> C := evalm(transpose(X)&*X):
> for i to 2 do C[i, i] := C[i, i] * (1 + λ) end do:
> b := evalm(b+ inverse(C)&* transpose(X)&*(y - f)):
> evalm((y - f)& * (y - f));
      7.394532277
> b := evalm(bs):
```

After several iterations (due to our selection of initial values, we first have to *increase* λ a few times), the procedure converges to $\widehat{b}_1 = 21.6 \pm 2.6$ and $\widehat{b}_2 = 10.7 \pm 2.9$. The two standard errors have been computed by an extra

```
> for i to 2 do sqrt(inverse(C)[i, i]*evalm((y - f)& * (y - f))/17) end do;
```

It is also a good idea to display the resulting fit by:

```
> with(plots):
> pl1 := pointplot([seq([x[i], y[i]], i = 1..19)]):
> pl2 := plot(b[1]/(b[2] + x), x = 6..100):
> display(pl1, pl2);
```

This can also serve, in the initial stages of the procedure, to establish 'sensible' values of b_1 and b_2, to be used as initial values of the iteration loop.

Chapter 6 ROBUST REGRESSION

In this chapter we return to discussing the simple linear model.

When there is an indication that the ε_i's are *not* normally distributed (by noticing several unusually large residuals - so called OUTLIERS), we can search for maximum-likelihood estimators of the regression parameters using a more appropriate distribution. The two most common possibilities are the LAPLACE (double exponential) and CAUCHY distribution.. Both of them (Cauchy in particular) tend to de-emphasize outliers and their influence on the resulting regression line, which is quite important when dealing with data containing the occasional crazy value. Procedures of this kind are called ROBUST (not easily influenced by outliers).

Laplace distribution

We will first assume that ε_i are distributed according to a distribution with the following PDF..

$$\frac{\exp(-\frac{|x|}{\gamma})}{2\gamma}$$

for all real values of x (the exponential distribution with its mirror reflection). Since the distribution is symmetric, the mean is equal to 0 and standard deviation is equal to $\sqrt{2}\gamma \equiv \sigma$.

The corresponding likelihood function, or better yet its logarithm, is then

$$-n\ln(2\gamma) - \frac{1}{\gamma}\sum_{i=1}^{n} |\, y_i - \beta_0 - \beta_1 x_i \,|$$

The $\frac{\partial}{\partial\gamma}$ derivative is

$$-\frac{n}{\gamma} + \frac{1}{\gamma^2}\sum_{i=1}^{n} |\, y_i - \beta_0 - \beta_1 x_i \,|$$

Making it equal to zero and solving for γ yields

$$\widehat{\gamma} = \frac{\sum_{i=1}^{n} |\, y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \,|}{n}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ represent the solution to the other two normal equations,namely

$$\begin{aligned} \sum_{i=1}^{n} \text{sign}(\, y_i - \beta_0 - \beta_1 x_i) &= 0 \\ \sum_{i=1}^{n} x_i \cdot \text{sign}(\, y_i - \beta_0 - \beta_1 x_i) &= 0 \end{aligned}$$

Solving these is a rather difficult, linear-programming problem. We will bypass this by performing the minimization of

$$\sum_{i=1}^{n} |\, y_i - \beta_0 - \beta_1 x_i \,|$$

graphically, with the help of Maple.

To find the mean and standard deviation of $\widehat{\gamma}$, we assume that n is large (large-sample theory) which allows us to replace e_i by ε_i. We thus get

$$\widehat{\gamma} \simeq \frac{\sum_{i=1}^{n} | \varepsilon_i |}{n}$$

which implies that

$$\mathbb{E}(\widehat{\gamma}) = \frac{\sum_{i=1}^{n} \mathbb{E}(| \varepsilon_i |)}{n} = \gamma + ...$$

(where the dots imply terms proportional to $\frac{1}{n}$, $\frac{1}{n^2}$, etc.), since

$$\frac{\int_{-\infty}^{\infty} |x| \, \exp(-\frac{|x|}{\gamma}) \, dx}{2\gamma} = \gamma$$

$\widehat{\gamma}$ is thus an asymptotically unbiased estimator of γ.

Similarly,

$$\mathrm{Var}(\widehat{\gamma}) \simeq \frac{\mathrm{Var}(| \varepsilon |)}{n} = \frac{\gamma^2}{n} + ...$$

(the dots now imply terms proportional to $\frac{1}{n^2}$), since

$$\frac{\int_{-\infty}^{\infty} x^2 \exp(-\frac{|x|}{\gamma}) \, dx}{2\gamma} = 2\gamma^2$$

The standard deviation of $\widehat{\gamma}$ is thus $\frac{\gamma}{\sqrt{n}} \simeq \frac{\widehat{\gamma}}{\sqrt{n}}$.

To perform the same kind of analysis for our (graphical) estimators of β_0 and β_1, we first realize that these have been obtained by minimizing the sum of a specific function (say F) of the residuals:

$$\sum_{i=1}^{n} F(e_i)$$

(in this case, F represents taking the absolute value, but we are better off by considering the general case). The two normal equations are thus

$$\begin{aligned} \sum_{i=1}^{n} F'(e_i) &= 0 \\ \sum_{i=1}^{n} x_i \, F'(e_i) &= 0 \end{aligned}$$

Expanding the left hand side with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ at the exact (albeit unknown) values β_0 and β_1 yields:

$$\begin{aligned} \sum_{i=1}^{n} F'(\varepsilon_i) - (\widehat{\beta}_0 - \beta_0) \sum_{i=1}^{n} F''(\varepsilon_i) - (\widehat{\beta}_1 - \beta_1) \sum_{i=1}^{n} x_i F''(\varepsilon_i) + ... &= 0 \\ \sum_{i=1}^{n} x_i F'(\varepsilon_i) - (\widehat{\beta}_0 - \beta_0) \sum_{i=1}^{n} x_i F''(\varepsilon_i) - (\widehat{\beta}_1 - \beta_1) \sum_{i=1}^{n} x_i^2 F''(\varepsilon_i) + ... &= 0 \end{aligned}$$

We can easily solve for

$$\left[\begin{array}{c}\widehat{\beta}_0\\ \widehat{\beta}_1\end{array}\right] = \left[\begin{array}{c}\beta_0\\ \beta_1\end{array}\right] + \left[\begin{array}{cc}\sum_{i=1}^n F''(\varepsilon_i) & \sum_{i=1}^n x_i F''(\varepsilon_i)\\ \sum_{i=1}^n x_i F''(\varepsilon_i) & \sum_{i=1}^n x_i^2 F''(\varepsilon_i)\end{array}\right]^{-1} \left[\begin{array}{c}\sum_{i=1}^n F'(\varepsilon_i)\\ \sum_{i=1}^n x_i F'(\varepsilon_i)\end{array}\right] + ... \tag{6.1}$$

The large-sample theory enables us to replace the coefficient matrix by the corresponding expected value (this is kind of tricky here since F'' relates to the Dirac function), thus:

$$\begin{aligned}\left[\begin{array}{c}\widehat{\beta}_0\\ \widehat{\beta}_1\end{array}\right] &= \left[\begin{array}{c}\beta_0\\ \beta_1\end{array}\right] + \gamma \left[\begin{array}{cc} n & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{array}\right]^{-1} \left[\begin{array}{c}\sum_{i=1}^n F'(\varepsilon_i)\\ \sum_{i=1}^n x_i F'(\varepsilon_i)\end{array}\right] + ...\\ &\equiv \left[\begin{array}{c}\beta_0\\ \beta_1\end{array}\right] + \gamma(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{F}'\end{aligned}$$

since $\mathbb{E}[F''(\varepsilon_i)] = \frac{1}{\gamma}$. Furthermore, based on $\mathbb{E}[F'(\varepsilon_i)] = 0$, we can see that the β estimators are asymptotically unbiased. Their variance-covariance matrix equals

$$\gamma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{E}[\mathbf{F}'\mathbf{F}'^T]\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \gamma^2(\mathbb{X}^T\mathbb{X})^{-1} \simeq \widehat{\gamma}^2(\mathbb{X}^T\mathbb{X})^{-1}$$

since $\mathbb{E}[\mathbf{F}'\mathbf{F}'^T] = \mathbb{I}$.

Example: In this example, we will generate our own data, using $n = 25$, $\beta_0 = 80$, $\beta_1 = -2$ and $\sigma = 10$ (the Maple program calls the regression coefficients a and b):

```
> with(linalg): with(stats): with(plots):
> x := randvector(25,entries = rand(1..35)):
> y := [seq(80 - 2 * x[i]+random[laplaced[0, 10]](1), i = 1..25)]:
> pointplot([seq([x[i], y[i]], i = 1..25)]);
> F := sum(abs(y[i] - a - b * x[i]), i = 1..25) :
> contourplot(log(F), a = 80..90, b = -2.3.. - 2.1,contours= 30);
> a := 85:  b := -2.25:
> X :=augment([seq(1, i = 1..25)], x):
> g := F/25;
                    g := 11.30304097
> for i to 2 do sqrt(inverse(transpose(X)&* X)[i, i] * g^2) end do;
                    4.585200522
                    0.2499498573
> g* sqrt(2.);  g* sqrt(2./25);
                    15.98491383
                     3.196982767
```

Our estimates for β_0 and β_1 are thus 85 ± 5 and -2.25 ± 0.25, in good agreement with the true values of 80 and -2. The σ estimate of 16.0 ± 3.2 is not that impressive (the exact value was 10), but its error is less than 2 standard errors, which we know to be quite feasible.

Cauchy Case

This time we will assume that the ε distribution is Cauchy, with the following PDF:

$$\frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + x^2}$$

Note that this distribution has *indefinite* mean (even though its median is equal to 0), and *infinite* standard deviation (σ denoting its *quartile* deviation).

The logarithm of the likelihood function is now

$$-n \ln \pi + n \ln \sigma - \sum_{i=1}^{n} \ln[\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2]$$

Differentiating with respect to σ, we get

$$\frac{n}{\sigma} - \sum_{i=1}^{n} \frac{2\sigma}{\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2}$$

Setting it to zero yields

$$\sum_{i=1}^{n} \frac{1}{1 + (\frac{e_i}{\sigma})^2} - \frac{n}{2} = 0 \tag{6.2}$$

where $e_i \equiv y_i - \beta_0 - \beta_1 x_i$.

Similarly, making the β_0 and β_1 derivatives equal to 0 results in

$$\begin{aligned} \sum_{i=1}^{n} \frac{e_i}{\sigma^2 + e_i^2} &= 0 \\ \sum_{i=1}^{n} \frac{x_i e_i}{\sigma^2 + e_i^2} &= 0 \end{aligned} \tag{6.3}$$

Maple in normally capable of solving these (nonlinear) equations for the three parameters, but we may have to help in the following way: Rather then asking it to solve (6.2) and (6.3) simultaneously, we first provide a rough estimate for σ, and solve (6.3) for β_0 and β_1 (it is always save to start with a *large* value of σ, which reduces this step to a simple regression). Using these values of β_0 and β_1, we ask Maple to solve (6.2) for σ. This cycle is repeated till convergence (σ, β_0 and β_1 no longer change). The results are of course our $\widehat{\sigma}$, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ estimates.

To investigate the statistical properties of the three estimators (now considered as random variables, not the final values), we again expand the right hand sides of the normal equations at the exact values. We should now do it with all three equations, since

$$F \equiv \ln \sigma - \ln(\sigma^2 + e_i^2)$$

is a function of all three parameters. But it is easy to see that the resulting equations decouple eventually (since the partial derivative of F with respect to σ and s_i, namely $\frac{4\sigma e_i}{(\sigma^2 + e_i^2)^2}$, has, in our approximation, a zero expected value).

So it is quite legitimate to first expand (6.2), assuming that β_0 and β_1 are fixed:

$$\sum_{i=1}^{n} \frac{\sigma^2}{\sigma^2 + \varepsilon_i^2} - \frac{n}{2} + \sum_{i=1}^{n} \frac{2\sigma\varepsilon_i^2}{(\sigma^2 + \varepsilon_i^2)^2} (\widehat{\sigma} - \sigma) + ... = 0 \tag{6.4}$$

Since

$$\mathbb{E}\left(\frac{\sigma^2}{\sigma^2+\varepsilon_i^2}-\frac{1}{2}\right)=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}\left(\frac{\sigma^2}{\sigma^2+x^2}-\frac{1}{2}\right)\frac{\sigma}{\sigma^2+x^2}\,dx=0$$

$\widehat{\sigma}$ is clearly asymptotically unbiased. Furthermore, after dividing (6.4) by n, we may replace the second coefficient by its expected value

$$\frac{1}{\pi}\int\limits_{-\infty}^{\infty}\frac{2\sigma\,x^2}{(\sigma^2+x^2)^2}\cdot\frac{\sigma}{\sigma^2+x^2}\,dx=\frac{1}{4\sigma}$$

thus:

$$\frac{\sum\limits_{i=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\varepsilon_i^2}-\frac{1}{2}\right)}{n}+\frac{1}{4\sigma}\,(\widehat{\sigma}-\sigma)+...=0$$

This means that the variance of $\widehat{\sigma}$ equals

$$\frac{16\sigma^2}{n}\mathrm{Var}\left(\frac{\sigma^2}{\sigma^2+\varepsilon_i^2}-\frac{1}{2}\right)=\frac{2\sigma^2}{n}\simeq\frac{2\widehat{\sigma}^2}{n}$$

Similarly, (6.1) now implies (since $F(\varepsilon_i)=-\ln(\sigma^2+\varepsilon_i^2)$, assuming that σ is fixed):

$$\begin{aligned}\begin{bmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{bmatrix} &= \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}+\begin{bmatrix}\sum_{i=1}^n\frac{2(\sigma^2-\varepsilon_i^2)}{(\sigma^2+\varepsilon_i^2)^2} & \sum_{i=1}^n x_i\frac{2(\sigma^2-\varepsilon_i^2)}{(\sigma^2+\varepsilon_i^2)^2}\\ \sum_{i=1}^n x_i\frac{2(\sigma^2-\varepsilon_i^2)}{(\sigma^2+\varepsilon_i^2)^2} & \sum_{i=1}^n x_i^2\frac{2(\sigma^2-\varepsilon_i^2)}{(\sigma^2+\varepsilon_i^2)^2}\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^n\frac{2\varepsilon_i}{\sigma^2+\varepsilon_i^2}\\ \sum_{i=1}^n x_i\frac{2\varepsilon_i}{\sigma^2+\varepsilon_i^2}\end{bmatrix}+...\\ &\simeq 2\sigma^2\begin{bmatrix}n & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^n\frac{2\varepsilon_i}{\sigma^2+\varepsilon_i^2}\\ \sum_{i=1}^n x_i\frac{2\varepsilon_i}{\sigma^2+\varepsilon_i^2}\end{bmatrix}+...\end{aligned}$$

which demonstrates that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are asymptotically unbiased, having the following variance-covariance matrix

$$4\sigma^4(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\,\mathbb{E}[\mathbf{F}'\mathbf{F}'^T]\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}=2\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\simeq 2\widehat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$$

Example: We will use the same parameters as in the previous example, except now the ε_i's have the Cauchy distribution with $\sigma=3$. The following Maple program does the job:

```
> with(linalg): with(stats): with(plots):
> x := randvector(25,entries = rand(1..35)):
> y := [seq(80 - 2 * x[i]+random[cauchy[0, 3]](1), i = 1..25)]:
> pointplot([seq([x[i], y[i]], i = 1..25)]);
> F1 := sum(σ^2/(σ^2 + (y[i] − a − b * x[i])^2), i = 1..25):
> F2 := sum((y[i] − a − b * x[i])/(σ^2 + (y[i] − a − b * x[i])^2), i = 1..25):
> F3 := sum((y[i] − a − b * x[i]) * x[i]/(σ^2 + (y[i] − a − b * x[i])^2), i = 1..25):
> σ := 100:
> fsolve({F2, F3}, {a = 60..100, b = −5.. − 1});
        {b = –1.979312306, a = 79.80522543}
```

```
> assign(%): σ := 'σ':
> σ := fsolve(F1 - 25/2, σ = 0..100);
                    σ := 3.351948403
> a := 'a': b := 'b':
> X := augment([seq(1, i = 1..25)], x):
> for i to 2 do sqrt(inverse(transpose(X)& * X)[i, i] * 2 * σ^2) end do;
                    1.735026661
                    0.09386891459
> σ* sqrt(2./25);
                    0.9480741785
```

This time, we are getting 79.8 ± 1.7 for β_0, -1.979 ± 0.094 for β_1 and 3.35 ± 0.95 for σ. Note that we had to iterate (about 4 times) through the **fsolve** loop to reach these values.

Chapter 7 TIME SERIES

In this chapter we study the possibility of the ε's being correlated with one another. This would normally happen when x is time, and we take a y observation (something like a stock price) every day (month, year, etc.). We will assume a simple linear (or polynomial) relationship between x and y, but the ε's are now generated by

$$\varepsilon_i = \alpha_1\varepsilon_{i-1} + \alpha_2\varepsilon_{i-2} + ... + \delta_i$$

where α_1, α_2, ... are (unknown) constants, and δ_i are *independent*, normally distributed, with the mean of zero and standard deviation of σ. This is called the AUTOREGRESSIVE model (for the ε's). We will first look in detail at the simplest case of the

Markov Model

namely

$$\varepsilon_i = \alpha_1\varepsilon_{i-1} + \delta_i \equiv \rho\varepsilon_{i-1} + \delta_i \tag{7.1}$$

A terminology note: When the ε_i's were independent, they could be seen simply as a random independent sample of size n from $\mathcal{N}(0,\sigma)$. Now, when generated in this new, rather nontrivial manner, they constitute a so called STOCHASTIC PROCESS. There are several kinds of stochastic processes; the ones with an integer index (the time scale is discrete) and continuous STATE SPACE (values of ε) are called TIME SERIES.

We will assume that this process (of generating the ε_i's) is STATIONARY, i.e. it started in a distant past (not just with our ε_1), implying that the distribution of all the ε_i's is the *same* (so is the correlation coefficient between any two *consecutive* ε_i's, etc.). The process can be stationary only when the model parameters (ρ in this case) fall in a specific range (meeting conditions of stability).

Under this assumption, we can establish that the ε_i's remain normal with the mean of zero. We can also find their common variance by taking the variance of each side of (7.1):

$$\mathrm{Var}(\varepsilon) = \rho^2\mathrm{Var}(\varepsilon) + \sigma^2$$

(note that ε_i and δ_j are uncorrelated whenever $i < j$). This implies that

$$\mathrm{Var}(\varepsilon) = \frac{\sigma^2}{1-\rho^2}$$

Note that the result is finite and positive only when $|\rho| < 1$ (this is also the stability condition).

To find the correlation coefficient between ε_{i-1} and ε_i (the so called FIRST SERIAL CORRELATION ρ_1), we multiply (7.1) by ε_{i-1} and take the expected value, getting:

$$\mathrm{Cov}(\varepsilon_{i-1},\varepsilon_i) = \rho\mathrm{Var}(\varepsilon_{i-1})$$

Dividing by $\mathrm{Var}(\varepsilon)$ yields:

$$\rho_1 = \rho$$

giving us a clear interpretation of our parameter ρ.

Similarly, to establish the k^{th} serial correlation, we multiply (7.1) by ε_{i-k} and take the expected value:

$$\mathrm{Cov}(\varepsilon_{i-k}, \varepsilon_i) = \rho \mathrm{Cov}(\varepsilon_{i-k}, \varepsilon_{i-1})$$

Dividing by the common variance yields the following recurrence formula

$$\rho_k = \rho \cdot \rho_{k-1}$$

which implies almost immediately that

$$\rho_k = \rho^k$$

This means that the variance-covariance matrix of ε_1, ε_2,....ε_n (and therefore of y_1, y_2, ... y_n) is

$$\mathbb{V} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

Luckily, this matrix has a rather simple (tri-diagonal) inverse:

$$\mathbb{V}^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \equiv \frac{1}{\sigma^2}\,\mathbb{W}$$

(check it out). Its determinant is equal to $\dfrac{\sigma^{2n}}{1-\rho^2}$.

To perform simple regression, we need to maximize the logarithm of the likelihood function, namely:

$$-\frac{n}{2}\log(2\pi) - n\log\sigma + \frac{1}{2}\log(1-\rho^2) - \frac{(\mathbf{y}-\mathbb{X}\boldsymbol{\beta})^T\mathbb{W}(\mathbf{y}-\mathbb{X}\boldsymbol{\beta})}{2\sigma^2}$$

This will yield the usual (weighted) estimators of $\boldsymbol{\beta}$ and σ^2, but now we also need to estimate ρ, based on

$$\frac{\rho}{1-\rho^2} = \frac{(\mathbf{y}-\mathbb{X}\boldsymbol{\beta})^T\mathbb{U}(\mathbf{y}-\mathbb{X}\boldsymbol{\beta})}{2\sigma^2}$$

where

$$\mathbb{U} \equiv \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & -2\rho & 1 & \cdots & 0 \\ 0 & 1 & -2\rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

or, equivalently,

$$\rho = \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sum_{i=2}^{n-1} e_i^2 + \frac{\sigma^2}{1-\rho^2}} \tag{7.2}$$

It is clear that, so solve for the maximum-likelihood estimators, one would have to iterate, e.g. start with $\rho = 0$ and do the ordinary regression, then use (7.2) to estimate ρ and come back to estimating $\boldsymbol{\beta}$ and σ^2 by the corresponding weighted regression, etc., until the estimators no longer change.

We will not derive the standard error of each of these estimators (it would be fairly tricky - we would have to use the large-n approach).

$> x := [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]:$
$> y := [126, 114, 105, 108, 95, 102, 101, 95, 83, 71, 75, 93, 102, 84, 62, 67, 63, 55, 21, 20]:$
$>$ **pointplot**([**seq**($[x[i], y[i]], i = 1..20$)]):
$> X :=$ **matrix**$(20, 2)$:
$>$ **for** i **to** 20 **do** $X[i, 1] := 1;\ X[i, 2] := x[i]$ **od**:
$> W :=$ **matrix**$(20, 20, 0)$:
$> rho := 0$:
$>$ **for** i **to** 19 **do** $W[i, i+1] := -rho;\ W[i+1, i] := -rho;\ W[i, i] := 1+rho\hat{}2$ **od**:
$> W[1, 1] := 1 : W[20, 20] := 1$:
$> beta :=$ **evalm**(**inverse**(**transpose**(X)&*W&*X)&***transpose**(X)&*W&*y);
$\beta :=$ [130.0261406, –0.9337280922]
$> e :=$ **evalm**$(y - X$&*$beta)$:
$> var :=$ **evalm**$(e$&*W&*$e)/18$;
var := 123.7606318
$> A :=$ **sum**$(e['i'] * e['i'+1], 'i' = 1..19)$:
$> B :=$ **sum**$(e['i']\hat{}2, 'i' = 2..19)$:
$> rho :=$ **fsolve**$(r = A/(B + var/(1 - r\hat{}2)), r)$;
$\rho := 0.5775786348$

Note that, to get the solution, we had to iterate (repeat the execution of the last few lines, starting with redefining the elements of W).

Yule Model

The error terms are now generated by

$$\varepsilon_i = \alpha_1 \varepsilon_{i-1} + \alpha_2 \varepsilon_{i-2} + \delta_i \tag{7.3}$$

where α_1 and α_2 are (unknown) constants, and δ_i are independent $\mathcal{N}(0, \sigma)$.

Multiplying by ε_{i-1}, taking the expected value and dividing by $\mathrm{Var}(X)$ yields

$$\rho_1 = \alpha_1 + \alpha_2 \rho_1$$

which implies that

$$\rho_1 = \frac{\alpha_1}{1 - \alpha_2} \tag{7.4}$$

Similarly, multiplying (7.3) by ε_{i-k} (where $k \geq 2$), taking the expected value and dividing by $\mathrm{Var}(X)$ results in the following recurrence formula for all the remaining serial correlation coefficients:

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2}$$

(with the understanding that $\rho_0 \equiv 1$).

Taking the variance of each side of (7.3) yields

$$\mathrm{Var}(\varepsilon) = \alpha_1^2 \mathrm{Var}(\varepsilon) + \alpha_2^2 \mathrm{Var}(\varepsilon) + 2\alpha_1\alpha_2 \mathrm{Var}(X)\,\rho_1 + \sigma^2$$

With the help of (7.4), we can now solve for

$$\mathrm{Var}(\varepsilon) = \frac{1-\alpha_2}{(1+\alpha_2)(1-\alpha_1-\alpha_2)(1+\alpha_1-\alpha_2)}\,\sigma^2$$

One can also show that the process is stable (stationary) if and only if all three factors in the denominator of the previous formula are positive.

The logarithm of the likelihood function is now:

$$-\frac{n}{2}\log(2\pi) - n\log\sigma + \log(1+\alpha_2) + \frac{1}{2}\log(1-\alpha_1-\alpha_2) + \frac{1}{2}\log(1+\alpha_1-\alpha_2) - \frac{(\mathbf{y}-\mathbb{X}\boldsymbol{\beta})^T \mathbb{W} (\mathbf{y}-\mathbb{X}\boldsymbol{\beta})}{2\sigma^2}$$

where

$$\mathbb{W} = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 & 0 & \cdots & 0 \\ -\alpha_1 & 1+\alpha_1^2 & -\alpha_1(1-\alpha_2) & -\alpha_2 & \cdots & 0 \\ -\alpha_2 & -\alpha_1(1-\alpha_2) & 1+\alpha_1^2+\alpha_2^2 & -\alpha_1(1-\alpha_2) & \cdots & 0 \\ 0 & -\alpha_2 & -\alpha_1(1-\alpha_2) & 1+\alpha_1^2+\alpha_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Setting the α_1 derivative equal to 0 yields

$$\frac{\alpha_1\sigma^2}{(1-\alpha_1-\alpha_2)(1+\alpha_1-\alpha_2)} + \alpha_1 \sum_{i=2}^{n-1} e_i^2 + \alpha_2 \sum_{i=2}^{n-2} e_i e_{i+1} = \sum_{i=1}^{n-1} e_i e_{i+1}$$

Similarly, using the α_2 derivative, we get

$$\frac{(2\alpha_2 + \alpha_1^2 - 2\alpha_2^2)\,\sigma^2}{(1+\alpha_2)\,(-1+\alpha_1+\alpha_2)\,(1+\alpha_1-\alpha_2)} + \alpha_1 \sum_{i=2}^{n-2} e_i e_{i+1} + \alpha_2 \sum_{i=3}^{n-2} e_i^2 = \sum_{i=1}^{n-2} e_i e_{i+2}$$

To a good approximation (when n is reasonably large), these can be replaced by

$$\begin{aligned} \alpha_1 \frac{\sum_{i=1}^{n} e_i^2}{n} + \alpha_2 \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{n-1} &= \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{n-1} \\ \alpha_1 \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{n-1} + \alpha_2 \frac{\sum_{i=1}^{n} e_i^2}{n} &= \frac{\sum_{i=1}^{n-2} e_i e_{i+2}}{n-2} \end{aligned}$$

$> n := 60$: $eps := vector(n)$:
$> eps[n] := 0$: $eps[n-1] := 0$:
$>$ **with(stats):with(linalg):with(plots)**:
$> eps[1] := 1.63 * eps[n] - .72 * eps[n-1] +$ **random**[**normald**[0, 1]](1):
$> eps[2] := 1.63 * eps[1] - .72 * eps[n] +$ **random**[**normald**[0, 1]](1):

$>$ **for** i **from** 3 **to** n **do** $eps[i] := 1.63 * eps[i-1] - .72 * eps[i-2] +$ **random**[**normald**[0, 1]](1) **od**:

$>$ **pointplot**([**seq**([$i, eps[i]$], $i = 1..n$)]):

$> A :=$ **sum**($eps['i'] * eps['i' + 2], 'i' = 1..n-2)/(n-2)$:

$> B :=$ **sum**($eps['i'] * eps['i' + 1], 'i' = 1..n-1)/(n-1)$:

$> C :=$ **sum**($eps['i']\hat{}2, 'i' = 1..n)/n$:

$>$ **linsolve**(**matrix**(2, 2, [C, B, B, C]), [B, A]);

[1.612192917, –0.7481655896]