STATS310.3 Simple linear regression

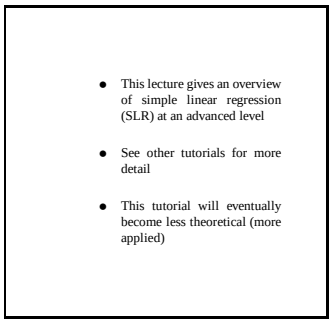
Gunnar Stefansson

June 18, 2012

Contents

1 Overview of simple linear regression

1.1 Background



This tutorial gives an introduction to simple linear regression. It assumes familiarity with elementary probability theory and random variables (see e.g. tutorials in Stats on the tutor-web).

This tutorial is rather theoretical for an "applied" course in simple linear regression. The text should therefore be considered a placeholder.

Optional theoretical background:

Recall that a random variable X is a real-valued (measurable) function, the expected value of a random variable is denoted $E[X]$ and the variance is $V[X]$.

The probability distribution of a random variable may be described by its cumulative distribution function,

$$F(x) := P[X \leq x]$$

or, when this is differentiable, by the density function, $f = F'$.

These notions extend to multivariate cases. In particular, a function f is a density function for a vector $mathbfY = (Y_1, \ldots, Y_n)'$ of random variables if $f(\mathbf{y} \geq 0$ for all $\mathbf{y} \in \mathbb{R}^n$ and

$$P[Y = 1 \leq y_1, \ldots Y_n \leq y_n] = \int_{t_n=-\infty}^{y_n} \ldots \int_{t_1=-\infty}^{y_1} f(y_1, \ldots, t_n) dt_1 \ldots dt_n$$

A collection of random variables is independent if the joint density is the product of the individual ones.

Densities may have unknown parameters. These may be estimated using e.g. least squares or maximum likelihood.

The **Maximum Likelihood Estimate, MLE** is the value of a parameter which maximizes the likelihood function.

It is fairly easy to see that the MLE for μ for the normal distribution is given by the mean of the y-values.

The density function describing a Gaussian random variable (a variable having the normal distribution) is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad y \in \mathbf{R}^n$$

.

For a set of n measurements from independent and identically distributed (i.i.d.) normal distributions,

the joint density is given by.

$$\begin{aligned} f(y_1,\ldots,y_n) &= \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_i \frac{(y_i-\mu)^2}{2\sigma^2}} \quad y_1,\ldots,y_n \in \mathbf{R} \end{aligned}$$

For given values of the parameters this function can be used to describe how probable certain outcomes are.

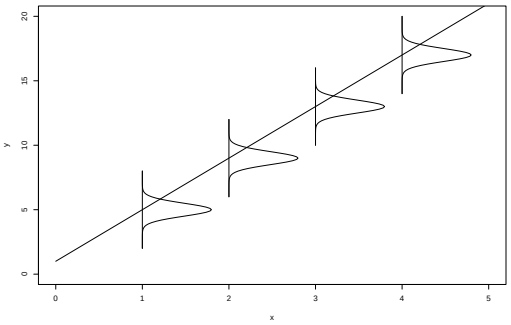
Once the experiment has been conducted and the data have been observed, this function can be evaluated for different values of the parameters. In this context, the function is called a **likelihood function**:

$$L(\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_i \frac{(y_i-\mu)^2}{2\sigma^2}} \quad \mu \in \mathbf{R}, \sigma \in \mathbf{R}_+$$

1.2 Informal regression

Have data as (x,y)-pairs
Scatterplot indicates relationship
Want to "fit a line" through the data
Evaluate the fit

1.3 Formal regression



Fixed numbers, x_i
Random variables: $Y_i \sim n(\alpha+\beta x_i, \sigma^2)$
or: $Y_i = \alpha+\beta x_i+\varepsilon_i$
$\varepsilon_i \sim n(0,\sigma^2)$ independent and identically distributed (i.i.d.)
The data:

$$y_i = \alpha+\beta x_i + e_i$$

Regression analysis is concerned with finding and estimating relationships between sets of measurements. Usually it is assumed that there is a desire to predict one set of measurements from another set.

Assume therefore that there are fixed numbers, x_i, such that the measurements are outcomes of random variables which are of the form

$$Y_i \sim n(\alpha+\beta x_i, \sigma^2)$$

or

$$Y_i = \alpha+\beta x_i+\varepsilon_i, \quad i = 1,\ldots,n$$

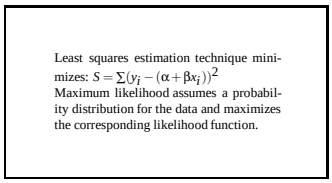
where $\varepsilon_i \sim n(0,\sigma^2)$ are independent and identically distributed (i.i.d.) and α, β are some fixed (but usually unknown) numbers.

The data are therefore of the form:

$$y_i = \alpha + \beta x_i + e_i.$$

The *x*-values are commonly called independent variables whereas the *y*-values, which depend on these, are termed dependent variables.

1.4 Estimation methods



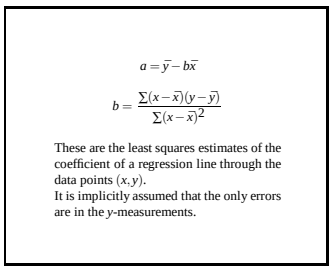
The method of least squares estimates α and β by minimizing the sum of squares

$$S = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

which is viewed as a function of of α and β.

Maximum likelihood assumes a probability distribution for the data and maximizes the corresponding likelihood function. In the regression setting it is commonly assumed that the data come from a Gaussian distribution. In this case the parameter estimates become the same as from least squares. In addition an estimate of the variance can be obtained as a part of the procedure.

1.5 The point estimates of a and b



It is quite simple to differentiate the function *S* with respect to α and β in order to find the values which minimize this sum of squares.

This gives minimum values at a point estimate. Those numerical values are denoted by *a* and *b*:

$$a = \bar{y} - b\bar{x}$$

and

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

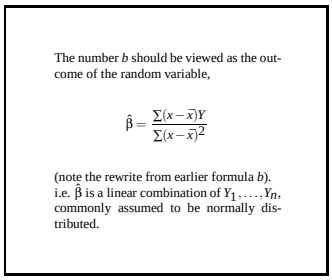
These are the least squares estimates of the coefficients of a regression line through the data points (x_i, y_i), $i = 1, \ldots, n$.

It must be kept in mind that it is assumed throughout that the x-vales are fixed numbers and any measurement errors are only present in the y-values.

Also note that though this computed regression line is commonly called the best line through the data, this is only in the context of least squares. There are other possible definitions of the quality of a line, particularly when one assumes that the Y_i-variables have a different distribution from the Gaussian.

To iterate: α and β are the unknown correct parameters and a, b are numerical estimates of these values, based on a specific data set.

1.6 The estimator and the estimate



A datum (y_i) is considered the realization of a random variable (Y_i).

It follows that the number b is actually a realization (outcome) of the random variable

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2},$$

or

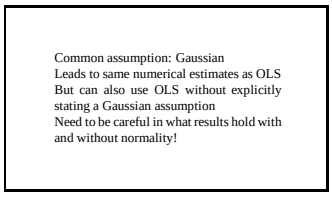
$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2},$$

where we have slightly rewritten the original formula for b.

It is therefore seen that $\hat{\beta}$ is a linear combination of $Y_1, \ldots, Y_n$, which are most commonly assumed to be independently normally distributed.

It is useful to distinguish between the numerical outcome, b, the unknown number being estimated, β, and the estimator, $\hat{\beta}$, though this distinction does become tedious at later stages.

1.7 Assumptions



If the random variables are assumed to come from independent normal (Gaussian) distributions, then the density of Y_i can be written down and the joint p.d.f. of all the variables is the product of the individual p.d.f.'s.

When the joint density is viewed as a function of the parameters, for fixed values of the data, it is termed a **likelihood function**.

Maximum likelihood estimation is undertaken by maximizing the likelihood function, and the resulting estimators are termed **maximum likelihood estimators**.

Simple manipulation shows that OLS provides numerically the same estimates as maximum likelihood in the case of normal distributions.

The interpretation is not the same, however, as an assumption of normality will provide more detailed results concerning the distribution of the estimators.

1.8 On expected values and variances

From elementary statistics course it is assumed known that if Y is a random variable, then the expected value, $\mu = E[Y]$ is (when it exists) given by

$$E[Y] = \int y f(y) dy$$

if Y has a continuous density (f), or by

$$E[Y] = \sum_y y p(y)$$

if Y has a discrete distibution with probability mass function p.

The variance, $\sigma^2 = V[Y]$ of the random variable Y with expected value μ is given by

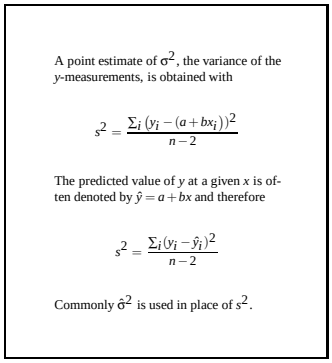
$$V[Y] = E\left[(Y-\mu)^2\right]$$

(when this exists).

In particular, for a random variable Y with an expected value of the form $E[Y] = \alpha + \beta x$, the variance is given by

$$V[Y] = E\left[(Y-(\alpha+\beta x))^2\right].$$

1.9 Estimating dispersion


A point estimate of σ^2, the variance of the y-measurements, is obtained with

$$s^2 = \frac{\sum_i (y_i - (a+bx_i))^2}{n-2}$$

The predicted value of y at a given x is often denoted by $\hat{y} = a + bx$ and therefore

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

Commonly $\hat{\sigma}^2$ is used in place of s^2.

A point estimate of σ^2 is obtained with

$$s^2 = \frac{\sum_i (y_i - (a+bx_i))^2}{n-2}.$$

We will commonly use the notation

$$SSE = \sum_i (y_i - (a+bx_i))^2$$

and

$$MSE = SSE/(n-2)$$

so that $s^2 = MSE$.

This will be used to derive an estimator for the variance of $\hat{\beta}$, denoted s_b^2 (or $\sigma_{\hat{\beta}}^2$).

To be accurate, one should always differentiate between an unknown quantity being estimated (e.g. $\alpha,\ \beta$), the estimator which is a random variable (e.g. $\hat{\alpha},\ \hat{\beta}$) and the estimate itself which is numerical outcome (e.g. $a,\ b$).

If we consider the corresponding random variables, we have

$$\hat{\sigma}^2 = \frac{\sum_i \left(Y_i - (\hat{\alpha} + \hat{\beta} x_i)\right)^2}{n-2}.$$

where the random variable $\hat{\sigma}^2$ is an **estimator** for σ^2, the variance of the y-measurements (or random variables, Y, rather).

As is done here, the notation s is commonly used for the numerical outcome of the random variable $\hat{\sigma}$ and s^2 or $MSE = SSE/(n-2)$ for the numerical outcome of $\hat{\sigma}^2$.

Similarly, when estimating the accuracy of the slope s_b is used to denote the numerical outcome of $\hat{\sigma}_{\hat{\beta}}$. In these cases, the notation can become quite cumbersome and these distinctions are therefore commonly omitted.

For a given value of the x-variable, the predicted value of y is denoted by

$$\hat{y}$$

so that

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

and therefore the numerical estimate of variance can be denoted by:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}.$$

It should be noted that although this is one natural point estimate of σ^2, it is by no means the only one. An alternative estimator is the MLE, given by

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}.$$

As will be seen later, the MLE is biased and the division by $n-2$ in place of n is what is needed to make an unbiased estimator.

1.10 Correlation and explained variation

Recall the the correlation coeffficient r is always between -1 and 1.
Write $SSE = \sum(y-\hat{y})^2$ (sum of squared errors, i.e. error after regression), and $SSTOT = \sum(y-\bar{y})^2$ (total sum of squares, i.e. before regression)
Definition: The explained variation is

$$R^2 = 1 - \frac{SSE}{SSTOT}$$

Note:

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \ldots = r^2$$

Recall the correlation r, which is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

is always between -1 and 1. The correlation is a useful concept, but one must note that r has no simple and direct interpretation other than the very vague "measures how close the x and y data are to being on a straight line".

Consider therefore the sum of squared errors, i.e. deviations from the straight line:

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

It is natural to compare this sum of squared errors to the sum of squares which is obtained if no relationship is assumed between y and x. This latter, total, sum of squares is denoted $SSTOT$ and computed with:

$$SSTOT = \sum_i (y_i - \bar{y}_i)^2.$$

Note that SSE is the variation which is still unexplained after a linear relationship has been assumed, but $SSTOT$ is the variation to begin with, i.e. the total variation in the y-data. It is now reasonable to define the proportional variation which remains unexplained, $SSE/SSTOT$ and hence the explained variation is $1 - SSE/SSTOT$.

Definition: The explained variation is

$$R^2 = 1 - \frac{SSE}{SSTOT}$$

It must be noted that this is the same concept as before since

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = \ldots = r^2.$$

We thus see that although r has no simple direct interpretations, R^2 has a natural interpretation and is therefore considerably more useful.

1.11 Output from regression software

1.12 Overview and vocabulary

Phrase	Synonym or abbreviation	Explanation
Regression	Simple linear regression, SLR	The act of fitting a regression line through data
Standard error		The standard deviation of a summary statistic
Regression analysis		The act of fitting and analysing a regression line through data
Least squares estimation	LS or OLS	The use of minimum sums of squares to estimate parameters
Likelihood function		The joint pdf of observations, viewed as a function of parameters
Maximum likelihood estimation		Estimating parameters by maximizing the likelihood function
Maximum likelihood estimator	MLE	The estimator resulting from maxizing the likelihood function
SSTOT		Total sum of squares
SSE		Sum of squared errors

2 Matrix representation of simple linear regression

2.1 Purpose of matrix representation

It is easy to set up matrices which describe the simple linear regression model. Solving this using matrix algebra gives an alternative representation of the estimators.

It turns out that a number of results are simpler to obtain using geometry and matrix algebra, as opposed to calculus. The matrix version of the regression problem also give a powerful tool for deriving estimates of variances and covariances of estimators.

2.2 Matrix form of simple linear regression

$\mathbf{y} \in R^n$ = vector of measurements

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

the "$\mathbf{X}$-matrix"
min $\sum(y_i - (\alpha + \beta x_i))^2$ is equivalent to finding

$$\beta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

to mininmize $||\mathbf{y} - \mathbf{X}\beta||^2$
Number notation: $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$

Denote the vector of measurements by $\mathbf{y} \in R^n$ and let

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

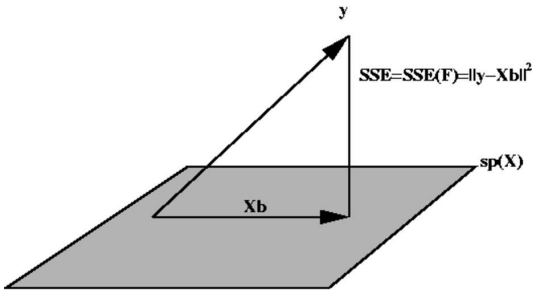
be the "**X**-matrix" so the regression problem of minimizing $\sum(y_i - (\alpha + \beta x_i))^2$ is equivalent to finding

$$\beta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

such that $||\mathbf{y} - \mathbf{X}\beta||^2$ is minimized.

When working with numerical outcomes the notation $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ is common. Here, e_i, $i = 1, \ldots, n$ are the numerical deviations from the model.

2.3 Prediction as linear projection



Viewing a linear model geometrically provides considerable new insight into the regression problem,

The basic model is of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{X}$ is a matrix of with dimensions $n \times p$ (in the present case $p = 2$). The best predictor of $\mathbf{y}$ within this model... (i.e. in the sense of closest in norm) is obtained using an orthogonal projection of $\mathbf{y}$ onto the plane spanned by the column vectors of the matrix $\mathbf{X}$. This orthogonal projection is denoted by $\hat{\mathbf{y}}$. Now, as the orthogonal project is contained in $sp\{\mathbf{X}\}$, it must be some linear combination of the column vectors of $\mathbf{X}$ and hence one can write $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ for some choice of $\hat{\beta}$.

Note: If V is a subspace of $\mathbb{R}^n$, then the orthogonal projection of a vector $\mathbf{u}$ onto V is the vector $\mathbf{v}$ which has the properties that $\mathbf{v} \in V$ and $\mathbf{u} - \mathbf{v} \in V^{\perp}$ where $V^{\perp}$ is the collection (vector space) of vectors which are orthogonal to all vectors in V.

The orthogonal projecction of $\mathbf{y}$ onto $span\{\mathbf{X}\}$, the span of the column vectors of $\mathbf{X}$ is thus the vector $\hat{\mathbf{y}} \in span\{\mathbf{X}\}$ such that $\mathbf{y} - \hat{\mathbf{y}} \in span\{\mathbf{X}\}^{\perp}$.

Since $\hat{\mathbf{y}} \in \mathbf{X}$, it follows that $\hat{\mathbf{y}}$ is a linear combination of the column vectors of $\mathbf{X}$, so we can write $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ for some $\hat{\beta}$.

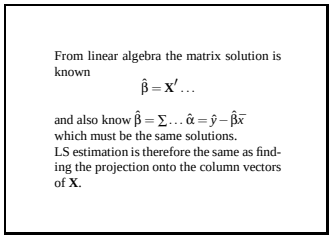
If $\hat{\mathbf{y}}$ is to be a projection onto $span\{\mathbf{X}\}$, the residual, $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$, is in $span\{\mathbf{X}\}^{\perp}$. Since $\hat{\mathbf{e}}$ is orthogonal to all vectors in $span\{\mathbf{X}\}$, it is also orthogonal to the column vectors of $\mathbf{X}$.

It follows that $\mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = 0$, i.e.

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

which are the **normal equations**.

2.4 Geometric solution to the simple linear regression problem



From linear algebra, the solution to the projection problem is known:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

This solution is fairly easy to obtain either through projection consideration, where we split $\mathbf{y}$ into orthogal components $\hat{\mathbf{y}} + \mathbf{e}$ or through differentiation of the norm, $||\mathbf{y} - \mathbf{X}\beta||^2$. In either case it becomes obvious that the column vectors of the $\mathbf{X}$-matrix must be orthogonal to the residual vector, $\mathbf{y} - \mathbf{X}\beta$. It follows from this that the solution must satisfy the **normal equations** $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$ which again implies the above solution if the matrix inverse exists. The condition of the existence of the inverse is common in regression problems but will be omitted for more complex linear models.

From earlier results, it is also known that

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

which must therefore give the same solutions.

It is seen that least squares estimation is numerically equivalent to obtaining the orthogonal projection of the data vector onto the space spanned by the column vectors of $\mathbf{X}$.

Example:

The simplest linear model is $y_i = \mu + e_i$, i.e. each measurement is a deviation from a common mean.

This model in matrix notation becomes: $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{1}$$

Matrix algebra quickly gives $\mathbf{X}'\mathbf{X} = [n]$, which is a 1×1 matrix with element n, (or simply the number n).

The inverse of this matrix is of course $1/n$. It is also easy to see that $\mathbf{X}'\mathbf{y} = \sum y_i$ and therefore $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \bar{y}$, which is the same estimate as before.

The sum of squared deviations becomes

$$SSE = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 = \sum(y - \hat{y})^2 = \sum(y - \bar{y})^2$$

and the estimate of variance is

$$MSE = SSE/(n-p) = \frac{\sum(y-\bar{y})^2}{n-1}$$

as before.

2.5 Overview and vocabulary

Phrase	Synonym or abbreviation	Explanation
MSE		Mean squared error

3 Distributions of linear projections of vectors of random variables*

3.1 Linear combinations of independent random variables

c a column vector
Y a vector of independent random variables
Same σ, expected values may differ, $E[\mathbf{Y}] = \mu$
Then

$$E\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mu$$

$$V\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mathbf{c}\sigma^2$$

Suppose **c** a column vector and **Y** a vector of independent random variables with a common variance, σ^2, but possibly different expected values. Then the mean and variance of the linear combination, $\mathbf{c}'\mathbf{Y}$, are given by

$$E\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mu$$
$$V\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mathbf{c}\sigma^2$$

These results are trivial to ascertain since the components, Y_i, are independent and hence e.g.

$$\begin{aligned} V\left[\mathbf{c}'\mathbf{Y}\right] &= V\left[\sum_i c_i Y_i\right] \\ &= \sum_i c_i^2 V\left[Y_i\right] \\ &= \mathbf{c}'\mathbf{c}\sigma^2 \end{aligned}$$

3.2 Covariance between linear combinations of independent random variables

a, **b** column vectors
Y a vector of independent random variables
Same σ, expected values may differ, $E[\mathbf{Y}] = \mu$
Then

$$Cov\left[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}\right] = \mathbf{a}'\mathbf{b}\sigma^2$$

Suppose **a**, **b** are column vectors and **Y** a vector of independent random variables with a common variance, σ^2, but possibly different expected values. Then the covariance between the linear combinations, $\mathbf{a}'\mathbf{Y}$ and $\mathbf{b}'\mathbf{Y}$, is given by

$$Cov\left[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}\right] = \mathbf{a}'\mathbf{b}\sigma^2$$

This follows from looking at the linear combinations as sums of components and noting that the covariance is a sum of all possible combinations, all of which are zero except where the same Y_i-combinations appear:

$$\begin{aligned}
Cov\left[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}\right] &= Cov\left[\sum_i a_i Y_i, \sum_j b_j Y_j\right] \\
&= \sum_{i,j} Cov\left[a_i Y_i, b_j Y_j\right] \\
&= \sum_{i,j} a_i b_j Cov\left[Y_i, Y_j\right] \\
&= \sum_i a_i b_i Cov\left[Y_i, Y_i\right] + \sum_{i,j:i\neq j} a_i b_j Cov\left[Y_i, Y_j\right] \\
&= \sum_i a_i b_i V\left[Y_i\right] \\
&= \mathbf{a}'\mathbf{b}\sigma^2
\end{aligned}$$

This result indicates that if the projection vectors, **a** and **b** are orthogonal, then the covariance remains zero. Note also that strictly, independence of the original variables is not required, but only zero covariance which is not the same condition in the general case.

In the case of two Gaussian random variables, it is, however, true that they have zero covariance if and only if they are independent. This can be seen by observing the bivariate Gaussian density function which neatly factors if and only if the covariance is zero.

3.3 Linear projections of independent random variables

A an $n \times n$ matrix
Y a vector of n independent random variables, mean μ, $V[Y_i] = \sigma^2$.
Then

$$E[\mathbf{AY}] = \mu$$

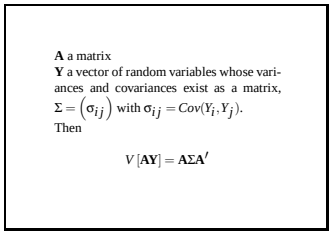
$$V[\mathbf{AY}] = \mathbf{A}\mathbf{A}'\sigma^2$$

Let $\mathbf{A}$ be a $q \times n$ matrix and $\mathbf{Y}$ an n-vector of independent random variables with common variance but possibly different expected values, then

$$E[\mathbf{AY}] = \mathbf{A}\mu$$
$$V[\mathbf{AY}] = \mathbf{AA}'\sigma^2$$

This can be derived either by considering the componentwise composition of $\mathbf{AY}$ or by writing A as a collection of row vectors and using the earlier results.

3.4 Linear transformations of dependent random variables



Let $\mathbf{A}$ be an $n \times n$ matrix and $\mathbf{Y}$ a vector of random variables whose variances and covariances exist as a matrix, $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = Cov(Y_i, Y_j)$.

This general situation occurs in regression analysis when measurements arrive in such a fashion that they can not be assumed to be independent. Several such examples certainly exist and the theory therefore needs to be properly developed.

This is also an important result when studying distributional properties of estimators, which are typically already linear combinations of original variables and hence no longer independent.

The first step is to derive the variance of projections of such variables. As before, this can be done by studying components or by looking at vector-wise linear combinations.

We obtain

$$V[\mathbf{AY}] = \mathbf{A\Sigma A}'$$

4 The expected value and variance of the estimators in simple linear regression

4.1 Expected value of the slope estimator

The estimator for the slope is unbiased:

$$\hat{\beta} = \sum_i \frac{(x_i - \bar{x})Y_i}{\sum_j (x_j - \bar{x})^2}$$

$$\Rightarrow E\hat{\beta} = \sum_i \frac{(x_i - \bar{x})E[Y_i]}{\sum_j (x_j - \bar{x})^2}$$

$$= \sum_i \frac{(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_j (x_j - \bar{x})^2} = \ldots = \beta$$

Note: This result only depends on the mean structure of Y_i, not the p.d.f. or even the variance.

Some properties of the various estimators need to be derived. In particular it is important to derive the expected value and variance of both the estimators of slope and intercept. These will be derived using only the assumptions needed, but for latter statistical inference more assumptions will be added in order to derive the probability distributions of these estimators.

The expected value is easy to compute based on writing the slope estimator as a linear combination of the dependent variables. First write

$$\hat{\beta} = \sum_i \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} Y_i$$

so that

$$\begin{aligned} E\hat{\beta} &= \sum_i \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} E[Y_i] \\ &= \sum_i \frac{(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_j (x_j - \bar{x})^2} \\ &= \frac{\sum_i (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_i (x_i - \bar{x})^2} \\ &= \beta \end{aligned}$$

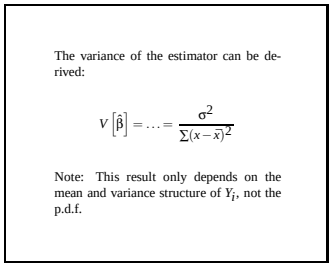
where we have used the facts that the sum of deviations $x_i - \bar{x}$ is zero and the corollary that the sum of $(x_i - \bar{x})x_i$ is the same as the sum of squares, $(x_i - \bar{x})^2$

It follows that the slope estimator, $\hat{\beta}$, is unbiased.

These results only use the assumption on the mean of the Y_i-variables. Thus, they also hold even if the variables are not Gaussian.

Also one should make a note that the slope estimator is also unbiased even if the variance structure is wrong. In particular, the slope is still unbiased whether the measurements all have the same variances or not.

4.2 Variance of the slope estimator



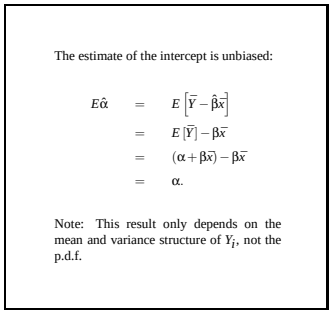
Slightly more work yields the variance of the estimator:

$$\begin{aligned} V\left[\hat{\beta}\right] &= V\left[\sum_i \frac{(x_i-\bar{x})Y_i}{\sum_j (x_j-\bar{x})^2}\right] \\ &= \sum_i \frac{(x_i-\bar{x})^2}{\left(\sum_j (x_j-\bar{x})^2\right)^2} V[Y_i] \\ &= \frac{\sigma^2}{\sum(x-\bar{x})^2} \end{aligned}$$

Here, we have used elementary facts concerning the variance operator, $V[aU+V] = a^2V[U]+V[V]$ if U and V are independent, extended in an obvious fashion to a linear combination of the independent random variables, Y_i.

As for the slope estimator, these results only use the assumption on the mean, variance and independence of the Y_i-variables. Thus, they also hold even if the variables are not Gaussian.

4.3 Expected value of the intercept estimator



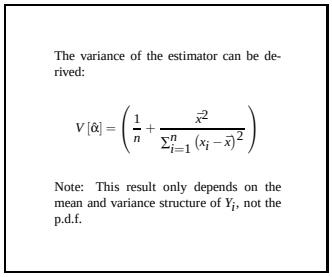
The expected value of the intercept estimate can be derived in a number of ways, the obvious being an attack on the basic equation:

$$\begin{aligned} E\hat{\alpha} &= E\left[\bar{Y}-\hat{\beta}\bar{x}\right] \\ &= E[\bar{Y}]-\beta\bar{x} \\ &= (\alpha+\beta\bar{x})-\beta\bar{x} \\ &= \alpha. \end{aligned}$$

We have thus shown that the estimator is unbiased.

These results only use the assumption on the mean, variance and independence of the Y_i-variables. Thus, they also hold even if the variables are not Gaussian.

4.4 Variance of intercept estimator



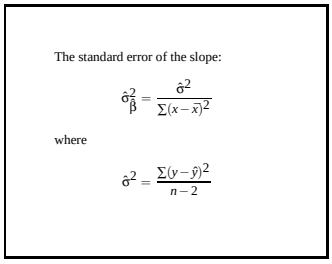
The variance of the estimator can be derived by rearranging terms in the formula for $\hat{\alpha}$ so that it is written as a linear combination of the Y_i-variables.

$$\hat{\alpha} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)Y_i$$

$$V[\hat{\alpha}] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

As for the mean, these results only use the assumption on the mean, variance and independence of the Y_i-variables. Thus, they also hold even if the variables are not Gaussian.

4.5 Estimating slope accuracy



When computing variances and standard deviations of derived quantities it is customary to call these standard errors to distinguish from from standard deviations in the meaning of simple deviations from a common mean.

The estimated standard error of the slope is usually denoted by

$$\hat{\sigma}^2_{\hat{\beta}}.$$

The natural estimator of the slope variance

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum(x-\bar{x})^2},$$

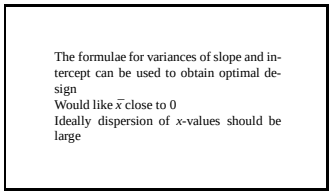
given the earlier estimator of the variance of the y-values. What remains, however, is to develop distributional properties of the estimator, since this is crucial for inference (drawing formal statistical conclusions about the true slope).

where

$$\hat{\sigma}^2 = \frac{\sum(y-\hat{y})^2}{n-2}$$

is the customary estimator of the variance of the y-values around the regression line.

4.6 Experimental design issues

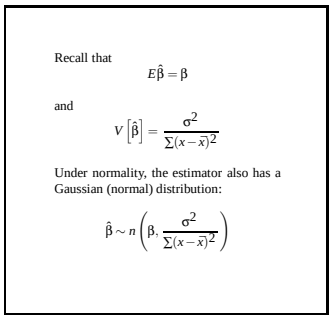


The formulae for slope and intercept can be used to obtain optimal design for a given sample size.

In the regression setting, the location of the x-values may be at the discretion of the experimenter. The experimenter may then choose to allocate these values so as to obtain minimum variance in the estimators of slope and intercept.

5 Distribution of estimators in SLR

5.1 Marginal distribution of estimator of slope



In addition to the mean and variance, the distributions of the estimators need to be derived.

Recall that the mean and variance of the slope estimator are given by

$$E\hat{\beta} = \beta$$

and

$$V\left[\hat{\beta}\right] = \frac{\sigma^2}{\sum(x-\bar{x})^2}$$

Since the estimator is a linear combination of the Y_i-variables, which are Gaussian, $\hat{\beta}$ is also normally distributed and we obtain:

$$\hat{\beta} \sim n\left(\beta, \frac{\sigma^2}{\sum(x-\bar{x})^2}\right)$$

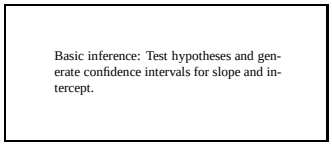
These results will form the basis for testing hypotheses and computing confidence intervals for β.

5.2 Marginal distribution of estimator of intercept

Exercise: Derive the marginal pdf of $\hat{\alpha}$.

6 Inference in SLR

6.1 Elements of inference in simple linear regression



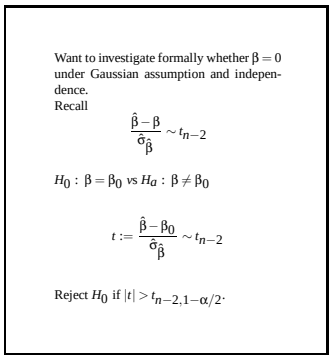
Earlier results on the estimators can be used to make inference on the true slope and intercept.

The first question raised is whether there is any relationship between the x and y measurements, i.e. whether the slope is zero. This can be phrased as a general hypothesis test for the slope.

Although hypothesis tests are important, they give no information if the hypothesis can not be rejected and hence confidence intervals tend to be more informative in general.

Both hypothesis tests and confidence intervals can be derived for the intercept as well as the slope, although inference for the intercept tends not to be as commonly used.

6.2 Testing hypotheses concerning the slope



As the distribution of $\hat{\beta}$ is known one can derive statistical tests of hypotheses concerning β. In general one would like to test hypotheses of the form $H_0 : \beta = \beta_0$.

A particularly common hypothesis is one of whether there is any relationship, i.e. $H_0 : \beta = 0$.

In order to test the general hypothesis concerning the slope, $H_0 : \beta = \beta_0$ *vs* $H_a : \beta \neq 0$ it should be noted that

$$t := \frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-2}$$

in general.

Note: This statement concerning the t-distribution tacitly assumed that the two quantities in the ratio are statistically independent, and that the χ^2-distribution applies to the denominator. This statement is easier to prove as a whole in the general case using matrix algebra.

It follows that if H_0 is correct, then

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

should come from a t-distribution with $n-2$ degrees of freedom.

H_0 will be rejected if the computed t-value is too extreme, i.e. if $|t| > t_{n-2,1-\alpha/2}$.

Example: Suppose we have a few measurements, (x,y), to be used in a regression analysis.

	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
	1	1.0	-2	4	-5	25	10	2.6		0.36
	2	5	-1	1	-1	1	1	3.8		1.44
	3	6	0	0	0	0	0	6.0		0
	4	7	1	1	1	1	1	8.2		1.44
	5	11	2	4	5	25	10	10.4		0.3
$\sum$	15	30	0	10	0	52	22			3.60
	$\bar{x}=3$	$\bar{y}=6$								

$$\hat{\beta} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{22}{10} = 2.2$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 6 - 2.2 \times 3 = -0.6$$

$$\hat{\sigma}^2 = \frac{\sum(y-\hat{y})^2}{n-2} = \frac{3.60}{3} = 1.2$$
$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum(x-\bar{x})^2} = \frac{1.2}{10} = 0.12$$

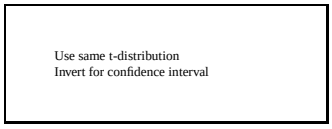
95% Confidence interval for β:

$$\hat{\beta} \pm t_{n-2.0975}\hat{\sigma}_{\hat{\beta}} = 2.2 \pm 3.182 \cdot \sqrt{0.12}$$

Testing $H_0: \ \beta = 2$ vs $Ha: \ \beta \neq 2$

$$\frac{\hat{\beta}-2}{\hat{\sigma}_{\beta^2}} = \frac{\hat{\beta}-2}{\sqrt{0.12}} = \ldots$$

6.3 Confidence interval for the slope



Given the t-distribution of the ratio,

$$t := \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-2},$$

it is easy to write down a probability statement, basically stating that there is probability $1-\alpha$ that t is between $\pm t_{n-2,1-\alpha/2}$.

These inequalities can be inverted to obtain a probability statement of two random endpoints incorporating the true value of β.

Given subsequent data, i.e. a realisation of the random variables, the data is used to compute the actual interval and a statement is made that the unknown parameter lies within the interval.

6.4 Inference for the intercept

Same procedure as for the slope, but replacing β and $\hat{\beta}$ with α and $\hat{\alpha}$ gives a confidence statement for the intercept.

6.5 Overview and vocabulary

Phrase	Synonym or abbreviation	Explanation
Statistical inference		The act of drawing a formal conclusion based on data

7 Covariance between estimators and inference*

7.1 Covariance between estimates of slope and intercept

Need to derive $Cov(\hat{\alpha}\&\hat{\beta})$ for general purposes
Can use this for inference b (for $\hat{Y}$ etc-not line -2.6 waits!) but it is easier to rewrite $\hat{Y}$ as linear combination.

It is in general useful to consider not only the variances of the estimators, $\hat{\alpha}$ and $\hat{\beta}$, but also the covariance between these estimators.

Take the simple model with $Y_i = \alpha + \beta x_i + \varepsilon_i$ and $\varepsilon_i \sim n(0, \sigma^2)$, i.i.d., so that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

The simplest way to obtain $Cov(\hat{\alpha}, \hat{\beta})$ is by rewriting the two formulae in terms of linear combinations of the Y_i-variables:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right) Y_i$$

from which the covariance follows easily since we have $Cov(Y_i, Y_k) = 0$ if $i \neq k$ and $Cov(Y_i, Y_i) = \sigma^2$.

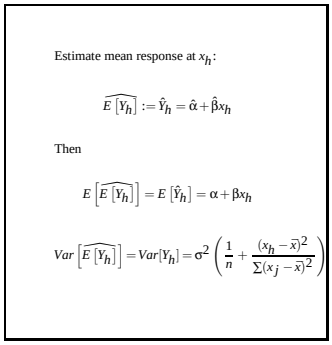
Completion of the above derivation requires the equality

$$Cov(aU + bW, cU + dW) = acCov(U, U) + bdCov(W, W) = acV[U] + bdV[W]$$

for independent random variables U and W. This equality can be derived by expanding the above left hand side using the defining expression for the covariance between two random variables,

$$Cov(S, T) = E[(S - \mu_S)(T - \mu_T)].$$

7.2 Estimating a point on the regression line



The mean (or, more accurately, expected) response at a new x-value, x_h is most naturally done by using the corresponding point on the estimated regression line:

$$\widehat{E[Y_h]} = \hat{\alpha} + \hat{\beta} x_h$$

Given $Cov(\hat{\alpha}, \hat{\beta})$ it is now possible to compute the variance of this estimator directly from the above formula.

$$Var\left[\widehat{E[Y_h]}\right] = Var\left[\hat{\alpha} + \hat{\beta} x_h\right] = \ldots$$

Alternatively, the same variance can be obtained by rewriting the formula for $\widehat{E[Y_h]}$ as a single linear combination of the Y_i-variables:

$$\widehat{E[Y_h]} = \hat{\alpha} + \hat{\beta} x_h = (\bar{Y} - \hat{\beta}\bar{x}) + \hat{\beta} x_h = \bar{Y} + (x_h - \bar{x})\hat{\beta} = \frac{1}{n}\sum Y_i + (x_h - \bar{x})\frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$$

i.e.

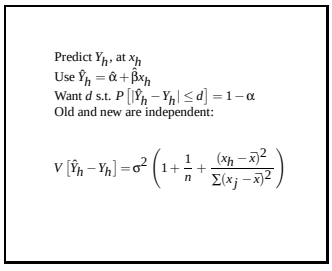
$$\widehat{E[Y_h]} = \sum\left(\frac{1}{n} + (x_h - \bar{x})\frac{(x_i - \bar{x})}{\sum(x_j - \bar{x})^2}\right) Y_i$$

From this the variance of the estimator follows easily after noting that cross-product terms cancel:

$$Var\left[\widehat{E[Y_h]}\right] = Var[Y_h] = \sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

This gives the basis for inference for $\widehat{E[Y_h]}$.

7.3 Predicting a new observation



Predicting the response, Y_h, at a new x-value, x_h is most naturally done by using the corresponding point on the regression line:

$$\hat{Y}_h = \hat{\alpha} + \hat{\beta} x_h$$

Notice, however, that as a prediction for a future point, this includes two sources of variation or error, first due to the measurement errors in the original data through variation in the parameter estimates and secondly through the future measurement errors at this point.

This will give prediction intervals for future measurements. Notice that these are not quite the same as confidence intervals, since a prediction interval makes a statement about the outcome of a random variable and this is a probability statement. A confidence statement is a statement about an unknown number and is therefore a different concept.

In the current setting we will want to be able to make a statement of the form

$$P\left[|\hat{Y}_h - Y_h| \leq d\right] = 1 - \alpha.$$

and the "d" must be chosen so as to fulfill the statement.

Can use independence of new and old obs so

$$V\left[\hat{Y}_h - Y_h\right] = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

This will give t_{n-2}-distributions for prediction intervals. Note that there is an important probabilistic difference between prediction intervals and confidence intervals.

As we now have a probability distribution:

$$\frac{Y_h - \hat{Y}_h}{s_{pred}} \sim t_{n-2}$$

we can also make diagnostic inference on whether a particular new observation is likely to be produced by the same mechanism as the earlier observations.

7.4 Predicting mean of several new observation

For mean of m new get

$$V\left[\bar{Y}_h - Y_h\right] = \sigma^2\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

For the average of several, m, new observations, at the same value, x_h of the x-variable, we get a slightly different variance estimate.

$$V\left[\bar{Y}_h - Y_h\right] = \sigma^2\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

8 Statistical packages

8.1 The R statistical package

R is freely available on the Internet.
Students can pick this up and install on their home computers.

Students are expected to obtain and install R. R is freely available and can be downloaded from the Internet.

Although R is free, it is very extensive. It is designed for easy extensibility and the emphasis is on easy graphical display and model searches.

The built-in help system is well-designed and is recommended for all users.

Books on S and Splus generally apply to R as well.

Typical commands in R include

- scan. Reads data, one element at a time.
- read.table. Reads tabular data.
- print. Outputs data to screen or file.
- lm. Fits linear model.
- summary. Summarizes output, e.g. from linear model.
- plot. General plotting function.
- rnorm. Generates normal (pseudo-)random deviates.

Example: A typical R example. The following sequence inputs matrix data in columns x, y and z, from the file "test.dat" into R and subsequently prints the data and does a simple linear regression.

The commands also plot a few examples of randomly generated data.

```
dat<-read.table("test.dat",col.names=c("x","y","z"))
print(dat)
summary(lm(y~x,data=dat))
x<-1:100
y<-2+0.5*x+rnorm(100)*5*x
plot(x,y)
plot(dat$x,dat$y)
```

Note that dat becomes a data frame, which is a bit like a matrix, but the columns have names and can be referred to as dat$x etc.

8.2 Linear statistical models with R

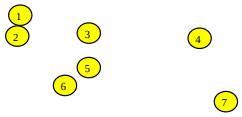
Suppose that within R a user has two columns of data, "x" and "y", which come in pairs and there is a need to fit a straight line through the data points.

Having plotted the data, this is followed by specifying the model, which should be of the form $y = \alpha + \beta x$. The model notation in R for this simple linear model is

$$y \sim x.$$

The tilde character ($\sim$) indicates that the left-hand side is a dependent variable and the model is on the right-hand side. On the right hand side it is implicitly assumed that there will be an intercept (α in the mathematical model) and therefore there is only a need to list the "dependent" variable(s), in this case only x.

To fit the actual model the "lm" function is used (lm being short for "linear model"):



Mathematical model:

$$y = \alpha + \beta x + \varepsilon$$

R definition:

$$y \sim x$$

lm(y x)
Storing the output fm<-lm(y x).

Figure 1: Example output from a simple linear model fit of the form y=a+bx. Items (1)-(2) are the estimates of a and b respectively. The estimate of the standard error of b is given by (3). The P-value for testing whether the true (underlying) value of b is zero is in (4). Items (5)-(7) give the MSE, R-squared and P-value for the entire model, respectively.

```
lm(y~x)
```

In order to process the model results, the fitted model is stored under some name, e.g. “fm”:

```
fm<-lm(y~x).
```

Example: Suppose the data are given by

```
x    1    2    3    4    5    6
y   -7   -6    0    0   -2    6
```

A simple linear model can be fitted to the data and the results output using:

```
> summary(lm(y~x))
```

The results are shown in the figure.

Note: The output from the various lm-related programs is quite detailed and although a statistics course can be designed around the interpretation of the results, some basic knowledge is essential.

Consider the output given in the figure.

Example: Consider a data set with a dependent variable *y*, an independent variable *x* and a factor, f:

```
   x  f         y
1  1  A  6.367151
2  2  A 10.783743
3  3  A 11.528125
4  4  A 15.564471
5  5  A 18.509431
6  1  B  4.608247
7  2  B  6.849981
8  3  B 12.301949
9  4  B 14.251640
10 5  B 16.483796
11 1  C  6.293174
12 2  C  7.905664
13 3  C 10.640212
14 4  C 15.881404
15 5  C 16.679703
```

If this data set is read in using read.table, the f-column will automatically become a factor and can be used directly in a model such as

```
lm(y~f+x)
```

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8277 -0.9488 -0.1151  0.7969  2.1061

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7466     0.6992   3.928  0.00173 **
x             2.9656     0.2108  14.066 3.04e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.155 on 13 degrees of freedom
Multiple R-Squared: 0.9383,     Adjusted R-squared: 0.9336
F-statistic: 197.9 on 1 and 13 DF,  p-value: 3.043e-09


> fm<-lm(y~f+x)
> drop1(fm,test="F")
Single term deletions

Model:
y ~ ff + x
       Df Sum of Sq     RSS     AIC  F value     Pr(F)
<none>               10.317   2.386
 f      2     7.018  17.335   6.170   3.7414    0.0576 .
x       1   263.837 274.153  49.585 281.3080 3.499e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

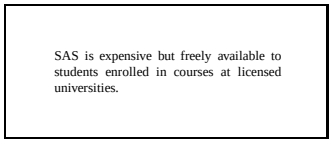
Use the resid function to extract residuals, then plot these and standardize to test for normality etc.

Use anova(fm1,fm2) to compare two models.

Having obtained the model, the coefficients can now be obtained, summary statistics of the model can be listed and the analysis of variance corresponding to the model is obtained:

```
fm<-lm(y~x).
summary(fm)         # General summary of model fit
anova(fm)           # Additional variation explained by each effect
drop1(fm)           # Marginal test of each effect in a model
coef(fm)            # Extract coefficients of fitted model
resid(fm)           # Extract residual
fitted(fm)          # Extract fitted values
```

8.3 The SAS statistical package



Students at licenced universtities can obtain and install SAS. SAS is an expensive package, but it is possibly the most extensive statistical package available. Under a university license it is available to enrolled students.

SAS is best known for classical statistical analyses such as linear models or univariate analyses where this package excels. This program is extremely well tested and runs on most computer platforms.

Detailed instructions on using SAS are available on http://www.tutor-web.net in various tutorials under Statistics.

Example: A typical SAS example. The following sequence inputs data in columns x, y and z, from the file “F:
test.dat” into SAS and subsequently prints the data, computes means and does a simple linear regression.

```
libname mystore 'F:\';
data mystore.mysasset;
  infile 'F:\test.dat';
  input x y z;
proc print;
proc means;
proc glm;
  model z=x;
```

The libname causes the data to be stored between SAS runs, so the data step can be omitted in the next run, by using instead the libname statement alone and referring explicitly to mystore.mysasset.