Linear Regression

1. Introduction

The following is the standard scenario for the whole of what follows: we have measurements on a *response* (or *dependent*) variable Y, as well as corresponding values for $k \geq 1$ *explanatory* (or *independent,* or *predictor*) variables $x_1, \ldots, x_k$. In the simplest situation, considered in this chapter, $k = 1$ and we simply have pairs of observations (x_i, y_i), $i = 1, \ldots, n$ of a response variable Y measured at n different values of a single explanatory variable x.

It is usual to assume that the values of the explanatory variables are not random, but "fixed" over a range of values. Although this is often not strictly true it will be assumed throughout, as extra difficulty ensues when the x values are random variables or considered to have been measured with some error.

The aim is usually to identify the form of the model that might exist relating the Y values to those of $x_1, \ldots, x_k$, estimating the relevant parameters with optimal accuracy, testing hypotheses regarding those parameters and making predictions on the strength of the model that we decide upon. Often consideration of the type of information we need, and what we think we expect, can help in designing sensible experiments in order to collect the most useful data that funds and time will allow. This last issue will be discussed further in STAT 404.

The basic model we consider throughout is of the form

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

where $\beta_0, \ldots, \beta_k$ are unknown parameters, and ϵ indicates an error term, of zero mean and constant variance σ^2.

EXAMPLE 1.1. *From Physics you may recall Ohm's Law, which states that voltage is current times resistance, $V = IR$. Though this law is exact, any experiment performed to verify it would involve experimental error, which might reasonably be expected to have zero mean.*

The above model is termed a *linear* model, referring to the linearity in the *parameters*, not the explanatory variables. For instance, the model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

is linear, whereas

$$Y = \beta_0 + e^{-\beta_1 x_1} + \frac{\beta_2^2}{\beta_1} x_2 + \epsilon$$

is not. Such models may seem overly restrictive, but are often at least *locally* valid.

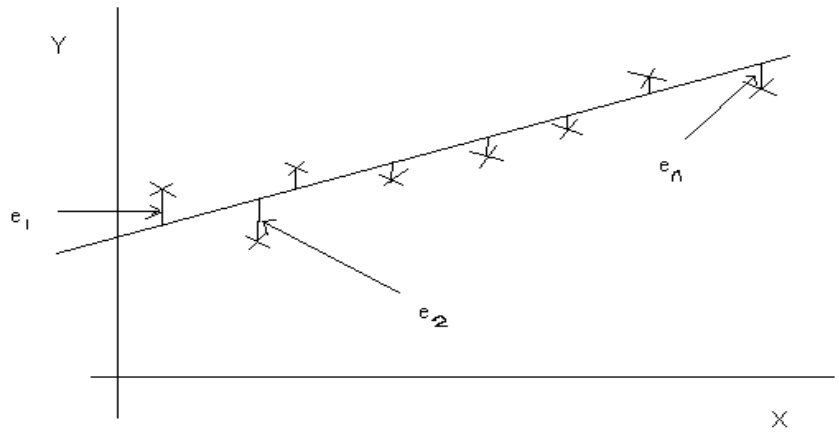
The term *regression* is often applied in the context of the material that follows.

2. Estimating the parameters

We consider in this chapter the simplest model, with just one predictor variable x, i.e.,

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

We have n pairs of observations (x_i, y_i) $i = 1, \ldots, n$ with which to verify the model and estimate the parameters β_0 and β_1. We do this via *least squares estimation*, an idea you will have met already in your introductory course. For each point (x_i, y_i) on the plot let e_i be the *vertical* distance from the point to the line fitted. Let e_i be positive if the point is above the line, and negative if below.



We choose the line which minimizes the sum of the *squares* of the errors. The sum

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

must be positive, and choosing the line to minimize this is called *least squares estimation*.

Specifically, we choose estimates to minimise the sum of vertical squared deviations from the line, that is, we minimise

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

which we do by differentiating partially in turn with respect to β_0 and β_1, setting the resulting equations to zero (for a minimum) and solving. Following these steps, we see

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \end{aligned}$$

which on equating to zero we easily derive the so-called *normal equations*

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

to be solved for (β_0, β_1) to give the estimates (b_0, b_1) say. Rearranging, we find our estimate of the slope parameter is

$$\begin{aligned} b_1 &= \frac{\sum x_i y_i - \left(\sum x_i \sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} \\ &= \frac{\sum (x_i - x)(y_i - y)}{\sum (x_i - x)^2} \qquad (1) \\ &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

say, where S_{xy} and S_{xx} are defined by the numerator and denominator of (1) respectively. It then follows that the estimate of the intercept at $x = 0$ is

$$b_0 = y - b_1 x. \qquad (2)$$

EXERCISE 2.1. *Verify the above derivations for b_0 and b_1.*

Substituting in these values, we can write our predicted value $\hat{y}$ of y at a given value of $x = x$ as

$$\hat{y} = y + b_1(x - x) \tag{3}$$

from which we see that the point $(x, \bar{y})$ must lie on the fitted line.

We call the difference between the fitted and the observed values the *residuals* of the model, and denote the ith one $e_i = y_i - \hat{y}_i$. By (3)

$$(y_i - \hat{y}_i) = (y_i - \bar{y}) - b_1 (x_i - x)$$

and summing this over $i = 1, \ldots, n$ we have

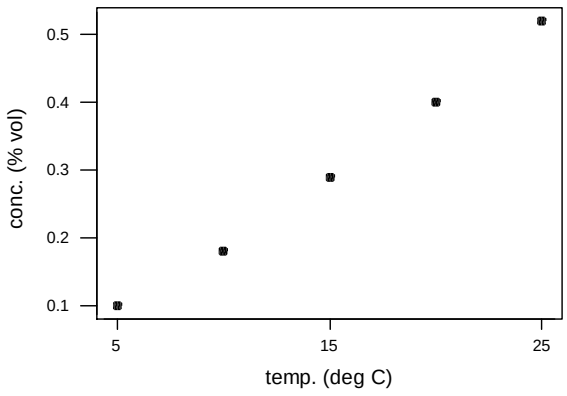
$$\begin{aligned} \sum (y_i - \hat{y}_i) &= \sum (y_i - \bar{y}) - b_1 \sum (x_i - x) \\ &= 0 \end{aligned}$$

so the sum of the residuals is zero.

EXAMPLE 2.1. *An experiment was conducted to look at the relationship between the concentration of bacteria in water, and the ambient temperature of that water. Equal volumes of water were stored at set temperatures, and the percentage volume of bacteria found in each sample was measured after a set length of time. The results are below:*

conc. (in % of vol.)	*0.10*	*0.18*	*0.29*	*0.40*	*0.52*
temp. (°C)	*5*	*10*	*15*	*20*	*25*

It is thought that a linear model might be suitable here, so the first step to establishing whether such a model is reasonable would be to plot the data.



Plot of concentration against temperature

Letting X be the temperature variable and Y the concentration, further evidence as to the suitability of a linear model will be provided by calculating r, the correlation.

x	*5*	*10*	*15*	*20*	*25*	*75*
y	*0.10*	*0.18*	*0.29*	*0.40*	*0.52*	*1.49*
x^2	*25*	*100*	*225*	*400*	*625*	*1375*
y^2	*0.010*	*0.0324*	*0.0841*	*0.160*	*0.270*	*0.557*
xy	*0.50*	*1.80*	*4.35*	*8.00*	*13.00*	*27.65*

Note that $n = 5$, $\bar{x} = 15$ and $\bar{y} = 0.298$. Via a software package (or otherwise)

$$r = 0.997.$$

This indicates nearly perfect positive correlation. Now the regression line must be calculated. Our estimate of the slope β is

$$b = \frac{cov(X, Y)}{s_X^2} = 0.0212.$$

The estimate of the intercept term is

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= -0.0200. \end{aligned}$$

The regression line for the data is therefore

$$Y = -0.0200 + 0.0212X.$$

We might use the model for predicting what level of concentration could be expected in water of a given temperature. For example, when $X = 18°C$, we predict that

$$\begin{aligned} Y &= -0.020 + 0.0212 \times 18 \\ &= 0.3724, \end{aligned}$$

so the predicted percentage volume of bacteria is 0.3724%.

Note that

(1) A *residual plot*, which plots residuals against either X or the fitted values $\hat{y}$, can be helpful in appraising a model.
(2) There should be no obvious "pattern" in the residuals.
(3) The mean of the residuals is always zero.

3. Properties of the estimators

Whatever reasonable distributional properties the Y variable may have, the following theorem holds true. Proofs are provided for completeness, but should be accessible to those who have studied MATH/STAT 302.

THEOREM 3.1. *The least squares estimates b_0 and b_1 derived above satisfy the following:*

(1) They are unbiased for β_0 and β_1 respectively.
(2) $\mathrm{Var}(b_1) = \sigma^2/S_{xx}$.
(3) $\mathrm{Var}(b_0) = \left(\sigma^2 \sum x_i^2\right) / \left(nS_{xx}\right)$.
(4) $\mathrm{Cov}(b_0, b_1) = -\left(\bar{x}\sigma^2\right)/S_{xx}$.

PROOF. (1) Since $b_1 = S_{xy}/S_{xx}$, as $\sum_1^n \bar{y}\,(x_i - \bar{x}) = 0$ we see

$$\begin{aligned} E(b_1) &= E\left(\frac{\sum y_i\,(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right) \\ &= \frac{\sum E(y_i)\,(x_i - \bar{x})}{S_{xx}} \\ &= \frac{\sum (\beta_0 + \beta_1 x_i)\,(x_i - \bar{x})}{S_{xx}} \\ &= \frac{\beta_1}{S_{xx}} \sum_1^n x_i\,(x_i - \bar{x}) \\ &= \frac{\beta_1}{S_{xx}} \sum_1^n (x_i - \bar{x})\,(x_i - \bar{x}) \\ &= \beta_1. \end{aligned}$$

Hence b_1 is unbiased for β_1. The corresponding result for b_0 is left as an exercise.

(2) Since $\mathrm{Var}(y_i) = \sigma^2$ for all i, we have

$$\begin{aligned} \mathrm{Var}(b_1) &= \frac{1}{S_{xx}^2} \sum_1^n (x_i - \bar{x})^2\,\mathrm{Var}(y_i) \\ &= \sigma^2 \frac{S_{xx}}{S_{xx}^2} \end{aligned}$$

and the result follows.

(3) This is also an exercise - you need to show that $\mathrm{Cov}(\bar{y}, b_1) = 0$.

(4) Given that y and b_1 have zero covariance, we find

$$\begin{aligned}\mathrm{Cov}(b_0, b_1) &= \mathrm{Cov}((\bar{y} - b_1\bar{x}), b_1) \\ &= -\bar{x}\mathrm{Var}(b_1) \\ &= -\frac{\bar{x}\sigma^2}{S_{xx}}.\end{aligned}$$

□

In addition, the estimators b_0 and b_1 are the *best linear unbiased estimators* of β_0 and β_1 respectively. That is to say, of all linear estimators (i.e., those of the form $a_1y_1 + \cdots + a_ny_n$, for some constants $a_1, \ldots, a_n$) which are unbiased, b_0 and b_1 have the least variance. This fact is known as the Gauss-Markov Theorem, and will not be proved here.

If in addition to the linear model, we further assume that each error term is Normally distributed with zero mean and variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$ i.i.d., then b_0 and b_1 are also Normally distributed, being linear combinations of the y_i values, with means and variances given by the above theorem. This useful property will be used later, to construct confidence intervals and tests.

4. Breaking down the sum of squares

What amount of variation in the y values can be attributed to the linear model we are fitting? We attempt to answer this question by examining some "sums of squares" which naturally arise in the model fitting. From our definition of the residuals of the model, we have

$$\begin{aligned}e_i &= y_i - \hat{y}_i \\ &= (y_i - \bar{y}) - (\hat{y}_i - \bar{y})\end{aligned}$$

and, on rewriting,

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Squaring both sides of this and summing over $i = 1, \ldots, n$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - y)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

due to the cross product term vanishing:

$$\begin{aligned}\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum b_1(x_i - \bar{x})((y_i - \bar{y}) - b_1(x_i - \bar{x})) \\ &= b_1(S_{xy} - b_1S_{xx}) \\ &= 0\end{aligned}$$

by (2). In words, we write (4) as

$$\text{SS about the mean } = \text{SS due to regression } + \text{ SS about regression}$$

where "SS" is short for "sum of squares". This is because the value $\hat{y}_i - \bar{y}$ measures the deviation of the ith predicted value from the mean, and $y_i - \hat{y}_i$ is the *residual*, namely the discrepancy between the ith observation and its fitted value. So in assessing the variation in the y values about their mean, some of the variation is due to the regression line fitted, and the remainder accounted for by the fact that all values did not lie exactly on the line. Obviously we have more confidence in a model with regression SS large relative to residual SS, and sometimes use

$$R^2 := \frac{\text{Regression SS}}{\text{Total (corrected) SS}}$$

as a measure of how the model performs, where the total (corrected) SS is just the l.h.s. of (4). A value of R^2 near unity suggest the model fits well.

Degrees of freedom. Any sum of squares has a *degree of freedom* associated with it, and this information is crucial in determining critical values for tests we will perform based on the breakdown of the variation in the response variable. The degrees of freedom (dof for short) of a SS is just the number of *independent* pieces of information required to calculate the SS, and will always be an integer. For example, the total SS, $\sum_{i=1}^{n} (y_i - \bar{y})^2$, has $n-1$ degrees of freedom, since it comprises the numbers $y_1 - \bar{y}, \ldots, y_n - \bar{y}$ only $n-1$ of which are independent since they sum to zero. Further, since $\hat{y}_i - \bar{y} = b_1 (x_i - \bar{x})$, it follows that

$$\begin{aligned} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n} b_1^2 (x_i - \bar{x})^2 \\ &= b_1^2 S_{xx} \end{aligned}$$

and so we can compute the regression SS with just one function of the y values, namely b_1. As degrees of freedom must add, we see by (4) and subtraction that the dof of the residual SS must be $n-2$, reflecting the fact that two parameters are estimated in computing it.

In general, the dof of a residual SS is the number of observations required less the number of parameters estimated.

ANOVA tables. Dividing a SS by its degrees of freedom gives the *mean square* (MS for short), and the purpose of these values will be apparent shortly. The information we obtain from breaking down the

variation in the data is often expressed in an *analysis of variance* table, or ANOVA table, as follows.

Source	DoF	SS	MS	F
Model	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSM	
Error	$n-2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE	MSM/MSE
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

The Error Mean Square (often called the residual MS) is denoted s^2, and provides an estimate of the variation about the regression line, $\sigma^2_{y|x}$ say, based on $n-2$ degrees of freedom. This unknown variance will only be σ^2 when the model fitted is true, otherwise $\sigma^2_{y|x} > \sigma^2$.

5. Normally distributed errors

The material covered so far has made no distributional assumptions about the data, and is valid for any distribution (or at least those with finite mean and variance). We now add to the basic model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, by assuming that

$$\epsilon_i \sim N(0, \sigma)$$

for each i, independently. This is often a reasonable assumption in view of the Central Limit Theorem, and can be checked by examining residuals (see earlier notes on probability plots). It is assumed valid throughout this section.

Confidence intervals for β_0 and β_1. For each parameter, confidence intervals are of the form

$$\text{estimate} \ \pm t^*_{n-2} \times se\,(\text{estimate})$$

where t^*_{n-2} is the appropriate point on the t_{n-2} distribution (often 97.5% point) and $se\,(\text{estimate})$ is the estimate of the standard deviation (e.s.d.) of the estimator, called the (estimated) standard error.

The standard deviation (s.d.) of the least squares estimate b_1 is $\sigma/S_{xx}^{1/2}$, which is estimated by replacing σ by s in situations where σ is unknown (as is most likely the case) to give the estimated standard deviation (e.s.d.). Under the assumption of Normality, two–sided $100\,(1-\alpha)\,\%$ confidence intervals for β_1 can be found by calculating

$$b_1 \pm \frac{t_{n-2}\left(1 - \frac{1}{2}\alpha\right) s}{S_{xx}^{1/2}}$$

with $t_{n-2}\left(1 - \frac{1}{2}\alpha\right)$ the $100\left(1 - \frac{1}{2}\alpha\right)\%$ ordinate of the t distribution with $n-2$ degrees of freedom. Similarly, to test the hypothesis H_0 :

$\beta_1 = \beta$ against its converse, we compare

$$\frac{b_1 - \beta}{\text{se}(b_1)} = \frac{(b_1 - \beta) S_{xx}^{1/2}}{s}$$

with the t_{n-2} distribution, and do not reject the null hypothesis if the observed value does not lie in the tails.

Corresponding results arise for inferences about β_0, with the estimated s.d. given by

$$\text{se}(b_0) = s \left(\frac{\sum x_i^2}{n S_{xx}} \right)^{1/2}.$$

Confidence intervals for mean and predicted values. If μ_y is the mean (i.e., expected) value of y the model implies

$$\mu_y = \beta_0 + \beta_1 x.$$

Confidence intervals for μ_y at $x = x_0$ are given by

$$\hat{\mu}_y \pm t_{n-2}^* \times SE\left(\hat{\mu}_y\right)$$

where $\hat{\mu}_y = b_0 + b_1 x_0$.

If we wish to predict the value of y for a given value of x, say x_0, we may compute a *prediction interval.* These are wider than the corresponding confidence intervals for $\hat{\mu}_y$, since $SE(\hat{y}) > SE\left(\hat{\mu}_y\right)$. Obviously our "best guess" of the response at a $x = x_0$ is

$$\hat{\mu}_y = \bar{y} + b_1 (x_0 - \bar{x}).$$

To assess how much confidence we should have in such an estimate, we need an idea of its sampling variance. Recalling that $\text{Cov}(\bar{y}, b_1) = 0$, we see

$$\begin{aligned} \text{Var}\left(\hat{\mu}_y\right) &= \text{Var}(\bar{y}) + (x_0 - \bar{x})^2 \text{Var}(b_1) \\ &= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \sigma^2}{S_{xx}} \end{aligned}$$

and this is estimated by replacing σ^2 by s^2 to give

$$\text{se}\left(\hat{\mu}_y\right) = s \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)^{1/2}.$$

Using this we can construct a confidence interval using the t distribution: e.g., the 95% confidence interval for the mean value of y at a given x_0 is $\hat{y}_0 \pm (t_{n-2}(0.975))\text{se}(\hat{y}_0)$.

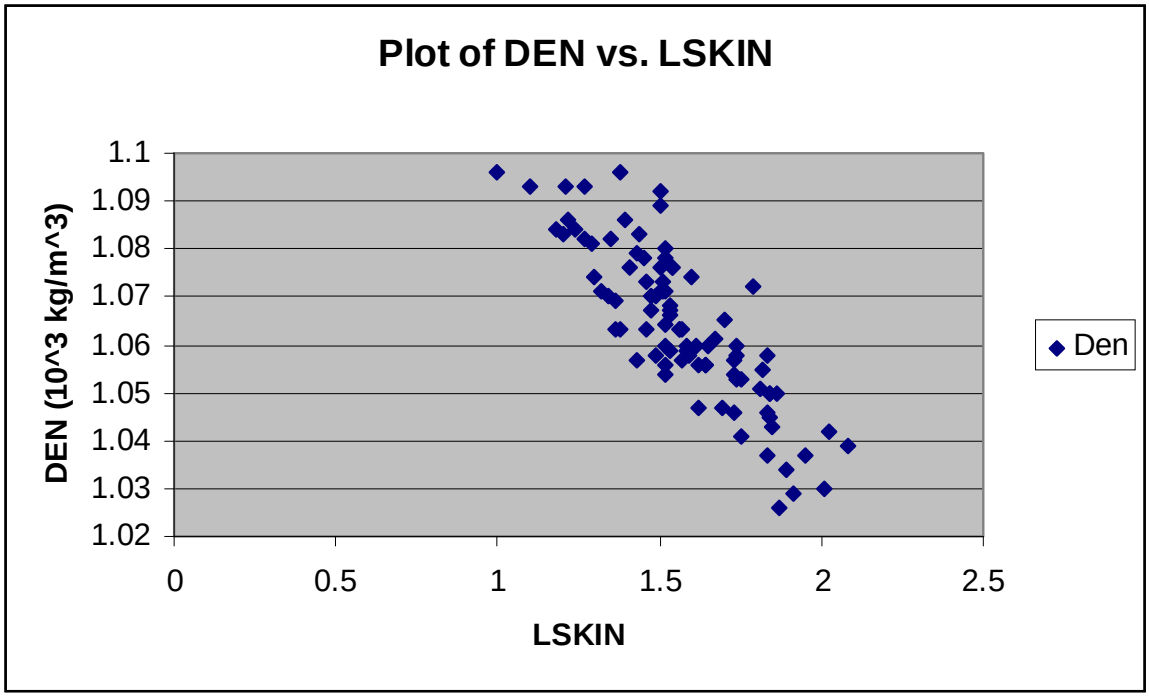
EXAMPLE 5.1. *Moore and McCabe (2003, E.g. 10.2) describe a study involving 92 college males between 20 and 29. The study aimed to see how*

$$x = \ LSKIN$$

the log of sum of four skinfold measures (biceps, triceps, subscapular and suprailiac areas), could predict the value of

$$y = \ DEN$$

body density (in 10^3 kg/m^3). A scatter plot of the data is below:



Software gives $b_0 = 1.163$, $b_1 = -0.0631$ *and so fitted line is*

$$\hat{y} = 1.163 - 0.0631x.$$

Further, we find $s = 0.00854$. *We find then for* β_0 *the estimate is 1.163 with estimated standard deviation 0.00656. A 95% confidence interval is* $(1.150, 1.176)$. *For* β_1 *corresponding interval is* $(-0.0714, -0.0549)$. *Testing* $H_0 : \beta_1 = 0$ *against* $H_a : \beta_1 \neq 0$ *compares*

$$\frac{b_1}{SE\,(b_1)} = \frac{-0.0631}{0.00414} = -15.23.$$

The predicted mean value of DEN (in $10^3 kg/m^3$) when LSKIN is 1.5 is given by

$$1.163 - 0.0631 \times 1.5 = 1.0683.$$

Now the standard error of mean here is 0.000934. This gives a 95% confidence interval of $(1.0665, 1.0702)$ *, since 97.5% point of* t_{90} *is 1.987. If a young male has LSKIN value of 1.5, can we give a 95% confidence interval for his body density, DEN? The predicted value is* 1.0683 *(in* $10^3 kg/m^3$*) as above, but a 95% prediction interval is*

$$(1.0656, 1.1001),$$

slightly wider than that given above for the mean value.

F-test for regression. A standard result in distribution theory is that if $X_1, \ldots, X_n$ are i.i.d. standard Normal, then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$, the Chi-squared distribution with n degrees of freedom. If however $X_1, \ldots, X_n$ are independent from $N(\mu_i, 1)$, then $\sum_{i=1}^n X_i^2$ has a *non-central* Chi-squared distribution, denoted

$$\sum_{i=1}^n X_i^2 \sim \chi_n^2 \left(\sum_{i=1}^n \mu_i^2 \right)$$

where $\sum_{i=1}^n \mu_i^2$ is called the *non-centrality* parameter. It can be shown that the expectation of this distribution is $n + \sum_{i=1}^n \mu_i^2$.

Consider now the regression SS, $b_1^2 S_{xx}$. We see under Normality,

$$\begin{aligned} \left(b_1 S_{xx}^{1/2}\right)^2 &\sim \left(N\left(\beta_1 S_{xx}^{1/2}, \sigma\right)\right)^2 \\ &= \left(\sigma N\left(\frac{\beta_1 S_{xx}^{1/2}}{\sigma}, 1\right)\right)^2 \\ &\sim \sigma^2 \chi_1^2 \left(\frac{\beta_1^2 S_{xx}}{\sigma^2}\right) \end{aligned}$$

and so the regression SS has a non-central Chi-squared distribution with 1 degree of freedom. Therefore

$$E(\text{Regression SS}) = \sigma^2 + \beta_1^2 S_{xx}.$$

It is possible to show that

$$\text{Residual SS} \sim \sigma^2 \chi_{n-2}^2,$$

and hence

$$E(\text{Residual SS}) = \sigma^2 (n-2)$$

from which it is apparent that s^2 is an unbiased estimator of σ^2.

From the above, we notice that if $\beta_1 = 0$ (so that the response variable does not depend linearly on the explanatory variable at all), we would expect the ratio of *mean squares* to satisfy

$$\frac{\text{Regression SS}}{\text{Residual SS}/(n-2)} \cong 1,$$

but that this ratio should be greater than unity if $\beta_1 \neq 0$. Now a ratio of independent Chi-squared variables each divided by their degrees of freedom has an F distribution, i.e.

$$\frac{\chi^2_n/n}{\chi^2_m/m} \sim F_{n,m}.$$

So to test the significance of the model fitted we compare

$$\frac{\text{Regression MS}}{\text{Residual MS}}$$

to the $F_{1,n-2}$ distribution. Since an $F_{1,n-2}$ variable is the square of a t_{n-2} variable, this test is identical to that given for $\beta_1 = 0$ earlier. When there are more terms in the model the analogous F-test does not correspond to the t–test for an individual coefficient, however tests for individual coefficients can be made by either the t or $t^2 = F$ form.

Example. The table below lists the birthweights (in g) and estimated gestational ages (in weeks) of twelve male babies.

Age	Birthweight
40	2968
38	2795
40	3163
35	2925
36	2625
37	2847
41	3292
40	3473
37	2628
38	3176
40	3421
38	2975

Fitting a linear model by least squares to these data, with birthweight as the response and age as the predictor, we find the model

$$\hat{y} = -1269 + 111.98x.$$

since

$$\sum x_i = 460, \quad \sum y_i = 36288, \quad \sum x_i^2 = 17672$$
$$\sum x_i y_i = 1395370, \quad \text{and} \quad \sum y_i^2 = 110623496.$$

Furthermore, the corresponding ANOVA table for the model is

Source	SS	Dof	MS	F
Regression	484885	1	484885	12.01
Residual	403699	10	40370	
Total (corrected)	888584	11		

Under the assumption of Normality, the F-value above tests the null hypothesis that birthweight is unrelated (at least in a linear fashion) to gestational age. Comparing with the $F_{1,10}$ distribution, we could reject this hypothesis at the 5% level.

6. Correlation

As the correlation coefficient ρ_{xy} between any two variables x and y is a measure of their collinearity, we might expect this to be closely related to the slope parameter β_1 in a suitable linear model as described above. Indeed, the *sample correlation coefficient*

$$r_{xy} := \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

is just a rescaling of the estimate of β_1, i.e.

$$b_1 = \left(\frac{S_{yy}}{S_{xx}}\right)^{1/2} r_{xy}.$$

The interpretations are slightly different: r_{xy} measures the collinearity between x and y, whilst b_1 describes the change in y when a unit change is made to x.

Further, we see that

$$\begin{aligned} r_{xy} &= (\text{sign of } b_1)\left(R^2\right)^{1/2} \\ &= R \end{aligned}$$

say, where R^2 is the multiple correlation coefficient previously defined. More generally, in cases where there are more than one linear predictor, we have

$$r_{y\hat{y}} = R.$$

7. Linear models in R

Linear models can be fitted and analysed in R Cmdr without the need for additional packages to be installed. Via Statistics → Fit models → Linear regression a linear model can be fitted with selected response and predictor variables. Graphical diagnostics can be explored via Models → Graphs → Basic diagnostic plots, the top row of plots provided being perhaps the most informative. Once a model has been

fitted, the associated ANOVA table is computed with Models → Hypothesis tests → ANOVA table

8. Exercises

(1) In the following table, x represents the heights of six fathers, whilst the y values are the corresponding heights of their eldest son, with all measurements in inches.

x	68	64	70	72	69	74
y	67	68	69	73	66	70

Fit the least squares linear model to these data. Plot the data, along with the fitted line, and comment on the adequacy of the fit.

(2) A study was made of the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected.

y	1	5	4	7	10	8	9	13	14	13	18
x	-5	-4	-3	-2	-1	0	1	2	3	4	5

Fit the least squares model for these data, and comment on the adequacy of the fit.

(3) (For background for those with MATH/STAT 302) By deriving the form of the covariance of two linear combinations of independent variables, show that $\mathrm{Cov}(\bar{y}, b_1) = 0$, where b_1 is the usual least squares estimator of the slope parameter. Hence obtain the variance of the intercept estimate b_0.

(4) Show that $R := (R^2)^{1/2}$ is the sample correlation coefficient between the response variable and the fitted values.

(5) Repeat the analysis of the birthweight data given in this chapter, using R.

(6) Recall the study on the effect of temperature on the yield of a chemical process from Q2. Construct an ANOVA table for these data and test the hypothesis that $\beta_1 = 0$ at the 5% level. Construct 95% confidence intervals for (i) β_1 and (ii) the expected value of y when $x = 3$.

(7) How well does the size of a house determine the annual tax house owners are paying? Nineteen houses are randomly selected from a city. The house size (measured in square feet of living space) and the amount of annual tax (in dollars) are

recorded for each of the 19 houses. Here are the summary statistics for the two variables:

house size : mean = 1456 sqft, standard deviation = 370 sqft
annual tax : mean = \$1707, standard deviation = \$323

The linear regression line that predicts the amount of annual tax from the house size has a slope of \$0.81 per square foot.

(a): Below is a partial ANOVA table for the regression line. Complete the table.

Source	Sum of Squares	df	Mean Square	F
Model	1610825			
Error	262428			
Total				

(b): Is there a significant linear relationship between the size of a house and the amount of annual tax charged? Carry out an appropriate test using the 1% significance level.

(c): Find the value of r^2, where r is the correlation between the size of a house and the amount of annual tax charged. Interpret this value in the context of this question.

(d): The 95% confidence interval for the population slope is found to be (0.64, 0.98) dollars per square foot. Using this information, find the standard error of the slope of the regression line.

(e): Give a point estimate for the mean annual tax paid by owners owning a 1500-square feet house.