5.5 Multiple regression

So far, we have considered the relationship between the response variable and a *single* explanatory variable. In many cases, we wish to relate the response variable (Y, say) to several explanatory variables ($x_1, \ldots, x_k$, say – note that the subscript now indicates which of the explanatory variables we are considering, not the value taken by a single explanatory variable). This can done by straightforward generalisations of (9) and (22).

The appropriate generalisation of the linear regression model (9) is the multiple linear regression model

$$E(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{40}$$

The least-squares estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of $\alpha, \beta_1, \ldots, \beta_k$ are defined as the values which minimise

$$S(\alpha, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \left\{ y_i - (\alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}) \right\}^2,$$

where y_i is the response at the value $(x_{1i}, \ldots, x_{ki})$ of the explanatory variables $(x_1, \ldots, x_k)$. The estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ can be found by matrix algebra. (Details are given in the Honours module *Generalized Linear Models and Data Analysis.*)

The appropriate generalization of (22) is

$$Y_i \sim N(\alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent.} \tag{41}$$

If (41) holds, then the maximum likelihood estimates of $\alpha, \beta_1, \ldots, \beta_k$ are the least-squares estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

The statistic s^2, defined by

$$s^2 = \frac{1}{n - (k+1)} S\left(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k\right),$$

is an unbiased estimator of σ^2. In R, $s = \sqrt{s^2}$ is called the *residual standard error.*

Confidence intervals for α and β_l, $l = 1, ..., k$, are obtained as in the simple linear regression paradigm, although the degrees of freedom of the relevant t distribution are $n - (k+1)$. For example, for 95% confidence intervals,

$$\hat{\alpha} \pm t_{0.025, n-(k+1)} s.e.(\hat{\alpha}),$$

$$\hat{\beta}_l \pm t_{0.025, n-(k+1)} s.e.(\hat{\beta}_l),$$

where

The R command `lm` can be used for multiple linear regression. The various predictor variables are joined by `+`, as shown in the following example.

Example (Peruvian blood pressure data):

In a study of the effects of altitude on blood pressure, various measurements were made on 39 indigenous Peruvian men who had been born in the Andes and had migrated to parts of Peru at low altitude. The variables which are relevant here are

(i) `systolic` = systolic blood pressure,

(ii) `age` = age (in years),

(iii) `years` = number of years since migration to low altitude,

(iv) `weight` = weight.

The investigators believed that systolic blood pressure might be related to age, the fraction of his lifetime for which such a man had been at low altitude, and weight. Accordingly, the following R session began by calculating `fraction` as `years/age` and then regressing `systolic` on the two explanatory variables `fraction` and `weight`.

```
> fraction<-years/age
> peru<-lm(systolic~fraction+weight)
```

The summary is

```
> summary(peru)

Call:
lm(formula = systolic ~ fraction + weight)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4330  -7.3070   0.8963   5.7275  23.9819 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
fraction    -26.7672     7.2178  -3.708 0.000699 ***
weight        1.2169     0.2337   5.207 7.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 9.777 on 36 degrees of freedom
Multiple R-Squared: 0.4731, Adjusted R-squared: 0.4438 
F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06 
```

The highly significant p-values for `fraction` and `weight` indicate that there is very strong evidence against the null hypotheses that the regression coefficients of `fraction` and `weight` are zero, i.e. we can conclude that systolic blood pressure definitely depends on fraction of life lived at low altitude and on weight.

5.5.1 Polynomial regression

Multiple regression can be used to fit polynomial regression models of the form

$$Y_i \sim N(\alpha + \beta_1 x_i + \cdots + \beta_k x_i^k, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent.} \tag{42}$$

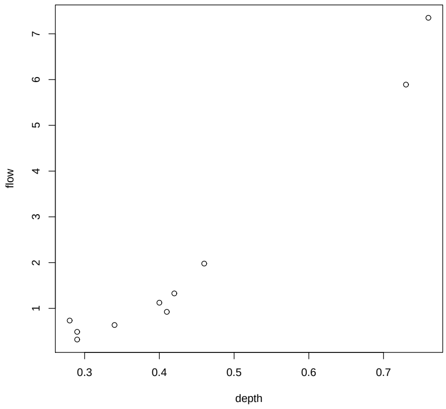
Model (42) is just the multiple regression model (40) with the powers $x, x^2, \ldots, x^k$ of x as the explanatory variables.

Example (Stream data):

Hydrologists were interested in the way in which the rate of flow in a stream varies with the depth at which the flow is measured. The following R session is a partial analysis of a data set consisting of 10 readings on flow rate (`flow`) and depth (`depth`).

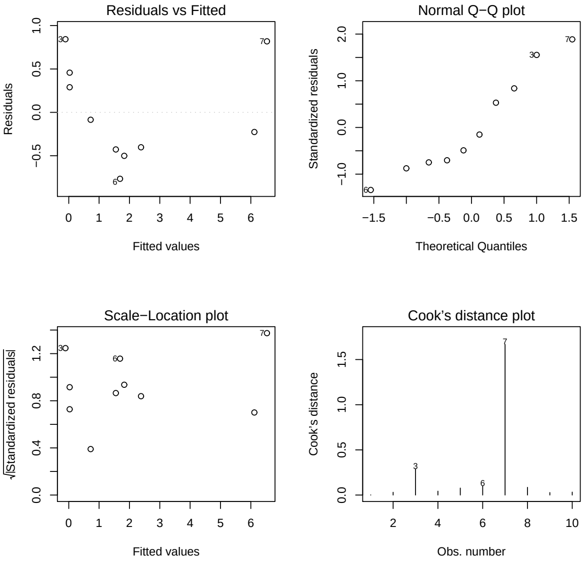
```
> plot(depth,flow,xlab="depth",ylab="flow")
```

produced the scatter plot on the next page.



This looks fairly linear, so a linear regression model of `flow` on `depth` was fitted and residual plots were produced.

```
> stream<-lm(flow~depth)
> par(mfrow=c(2,2))
> plot(stream)
```



The residual plot suggests that a quadratic of the form

$$E(Y) = \alpha + \beta_1 x + \beta_2 x^2 \tag{43}$$

(where Y denotes `flow` and x denotes `depth`) would be more appropriate than the simple linear model

$$E(Y) = \alpha + \beta x.$$

The following piece of R code fits the quadratic model (43).

```
> depthsq<-depth^2
> stream2<-lm(flow~depth+depthsq)
> summary(stream2)

Call:
lm(formula = flow ~ depth + depthsq)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.406145 -0.163666 -0.002649  0.198973  0.327658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.683      1.059   1.589   0.1561
depth        -10.861      4.517  -2.404   0.0472 *
depthsq       23.535      4.274   5.506   0.0009 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2794 on 7 degrees of freedom
Multiple R-Squared:  0.99,  Adjusted R-squared: 0.9871
F-statistic: 346.5 on 2 and 7 DF,  p-value: 1e-07
```

Note that the p-value of 0.1561 corresponding to the intercept is not significant at the 15% level, indicating that we have no reason to doubt the hypothesis $H_0 : \alpha = 0$ in (43). We might therefore choose to fit the model

$$E(Y) = \beta_1 x + \beta_2 x^2, \tag{44}$$

which can be done by

```
> stream3<-lm(flow~depth+depthsq-1)
> summary(stream3)

Call:
lm(formula = flow ~ depth + depthsq - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39946 -0.08639 -0.03278  0.14220  0.45582

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
depth    -3.7492     0.6613  -5.669 0.000471 ***
depthsq  16.9382     1.1071  15.299 3.31e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3049 on 8 degrees of freedom
Multiple R-Squared: 0.9924, Adjusted R-squared: 0.9905
F-statistic:   522 on 2 and 8 DF,  p-value: 3.343e-09
```

5.5.2 Comparison of models

A common problem in multiple regression is that of testing that certain specified regression coefficients are zero (representing the intuitive idea that the corresponding explanatory variables ‘have no effect’). For example, in the above example on the stream data, we tested the null hypothesis $H_0 : \alpha = 0$ in (43) against a 2-sided alternative.

In the above example, R calculated the p-value of 0.1561 by using a t-test. It is tempting to think that, in general, when testing that several regression coefficients are zero, we should use a sequence of t-tests, to test that each of these coefficients in turn is zero. However, this makes the calculation of p-values very complicated. It is better to use ANOVA (ANalysis Of VAriance). ANOVA is a general way of testing hypotheses which are formulated as nested linear models, using a test statistic based on the difference in the goodness-of-fit of the alternative models. ‘Nested’ means that the model corresponding to the null hypothesis is a special case of the model corresponding to the alternative hypothesis, being obtained from it by placing (linear) restrictions on the parameters.

In the context of multiple regression, we have a multiple regression model

$$Y_i \sim N(\beta_1 + \beta_2 x_{2i} + \cdots + \beta_{p_1} x_{p_1 i}, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent}, \qquad (45)$$

and we are interested in a submodel in which $p_1 - p_0$ of the regression coefficients $\beta_1, \ldots, \beta_{p_1}$ are zero. Thus we wish to test the null hypothesis

$$H_0 : \text{the specified regression coefficients are zero}$$

against the alternative hypothesis

$$H_1 : \text{there is no restriction on the specified regression coefficients.}$$

The intuitive idea is to see whether or not the *full model* (45) gives a *significantly* better fit to the data than the submodel (specified by H_0) does. Of course, the full model always fits the data a little more closely than the submodel, since it has more parameters. However, if H_0 is true then the difference in fit between the models should be quite small, while a big difference in fit would tend to suggest that H_0 is false.

The goodness-of-fit of either model to the data is measured by the *residual sum of squares* (rss), which is the sum of squares of differences between the model and data. For example for the full model, the residual sum of squares is

$$\text{rss}_1 = \sum_{i=1}^{n} \left\{ y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_{p_1} x_{p_1 i} \right) \right\}^2 . \qquad (46)$$

The residual sum of squares rss_0 for the submodel is defined similarly, but using only the least squares estimates (under H_0) of the regression coefficients used in the submodel.

Calculations similar to those used to obtain the one-sample t-test show that, under either model,

$$\frac{\text{rss}_1}{\sigma^2} \sim \chi^2_{n-p_1},$$

but only under H_0 do we have

$$\begin{aligned} \frac{\text{rss}_0}{\sigma^2} &\sim \chi^2_{n-p_0}, \\ \frac{\text{rss}_0 - \text{rss}_1}{\sigma^2} &\sim \chi^2_{p_1-p_0}, \end{aligned}$$

$$\text{rss}_0 - \text{rss}_1 \text{ is independent of } \text{rss}_1.$$

Hence

$$\frac{(\text{rss}_0 - \text{rss}_1)/(p_1 - p_0)}{\text{rss}_1/(n - p_1)} \sim F_{p_1-p_0, n-p_1} \qquad \text{under } H_0. \qquad (47)$$

If H_0 is false then this statistic will tend to be too large for consistency with $F_{p_1-p_0,n-p_1}$. This gives a test of H_0 against H_1.

Example (Stream data revisited):

Here $n = 10$. The full model is (43), i.e.

$$E(Y) = \alpha + \beta_1 x + \beta_2 x^2$$

(where Y denotes `flow` and x denotes `depth`). Thus $p_1 = 3$.

(i) If we take the null hypothesis to be

$$H_0 : \alpha = 0$$

then the submodel is (44). In this case $p_0 = 2$ and the value of the statistic (47) is $2.524921 = 1.589^2$, where 1.589 is the t-value given for `(Intercept)` in `stream2`. In the output for `stream2`, 1.589 is compared with t_7; according to (47), 1.589^2 is compared with $F_{1,7} = t_7^2$. As the p-value is 0.1561, we accept H_0.

(ii) If we take the null hypothesis to be

$$H_0 : \beta_1 = \beta_2 = 0$$

then the submodel is

$$E(Y) = \alpha,$$

so that the mean of Y does not depend on x. In this case $p_0 = 1$. The last line of `summary(stream2)` is

```
F-statistic: 346.5 on 2 and 7 DF,  p-value: 1e-07
```

meaning that the value of the statistic (47) is 346.5. This is to be compared with $F_{2,7}$. As the p-value is so small, we reject H_0.

Remark The name *Analysis of Variance* comes from the fact that (47) is based on the decomposition

$$\frac{\mathrm{rss}_0}{\sigma^2} = \frac{\mathrm{rss}_1}{\sigma^2} + \frac{\mathrm{rss}_0 - \mathrm{rss}_1}{\sigma^2}, \tag{48}$$

which splits up the (scaled) variability rss_0/σ^2 of the data about the submodel into the sum of the (scaled) variability rss_1/σ^2 of the data about the full model and the (scaled) difference $(\mathrm{rss}_0 - \mathrm{rss}_1)/\sigma^2$ between the two models. From the algebraic/geometric point of view, (48) can be regarded as an example of Pythagoras' Theorem. Under H_0, the distributions of the terms in (48) are the corresponding terms in the decomposition

$$\chi^2_{n-p_0} \sim \chi^2_{n-p_1} + \chi^2_{p_1-p_0}. \tag{49}$$