STAT 212: Principles of Statistics II

Lecture Notes: Chapter 1 (Part C)

Multiple Linear Regression

Patricia Ning

Dept. of Statistics

Texas A&M University

Multiple Regression

In multiple regression, we have one dependent variable/response/outcome Y and more than one independent variable/predictors/covariates.

Examples

(1) Study involving preschool and elementary school children.

y = developmental score, which is an average of scores for many characteristics such as verbal ability, graphical perception, communication ability, etc.

x_1 = age of child

x_2 = no. of older siblings

(2) Medical study

y = systolic blood pressure of an individual

x_1 = age

x_2 = weight

x_3 = cholesterol level

Our regression models will have the form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

As usual $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$, *which doesn't depend on* $x_1, \ldots, x_k$, and $\epsilon \sim N(0, \sigma^2)$.

The expected value of Y given $x_1, \ldots, x_k$ is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Again, all the $E(\cdot)$'s and Var$(\cdot)$'s above denote *conditional* expectation and variance, respectively, given the predictor values $x_1, \ldots, x_k$.

This model is more flexible than it might first appear, since *a given x_i might be a function of other independent variables.*

Two distinct situations are possible:

I. $k =$ number of separate independent variables in the model. In other words, $k + 1$ measurements are obtained from each experimental unit, one measurement being y.

II. Number of separate independent variables is p, which is less than k.

Examples of II:

Suppose $p = 2$ with independent variables x_1 and x_2. Two possible models are as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \qquad (M1)$$

Here $k = 3$ with $x_3 = x_1 x_2$ (also known as the *interaction effect* between x_1 and x_2).

Or we could have

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon, \qquad (M2)$$

in which case $k = 4$ with $x_3 = x_1^2$ and $x_4 = x_2^2$ (i.e. the quadratic effects of x_1 and x_2).

The function

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is a *plane*, which has no curvature. **If** one suspects that the *surface of averages* $E(Y|x_1, x_2)$ has *curvature*, then the models $M1$ and $M2$ might be more appropriate to use.

Read the discussion on pg. 423-425 (textbook).

Obtaining the least squares estimates:

Given the model on pg. 82N, we want to find least squares estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_k$. This works in exactly the same way as in polynomial regression.

Data comes in the form:

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \quad i = 1, \ldots, n.$$

Define

$$f(b_0, b_1, \ldots, b_k) =$$

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})]^2.$$

Choose $b_0, b_1, \ldots, b_k$ to minimize f.

$$\frac{\partial f(b_0,\ldots,b_k)}{\partial b_j} = 0 \quad \text{if and only if}$$

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})] x_{ij} = 0,$$

for $j = 0, 1, \ldots k$, and with $x_{i0} = 1$ for each i.

As before, we can write the solution to this set of *linear equations* in matrix form.

Let $\boldsymbol{X}$ be the $n \times (k+1)$ matrix with $(j+1)$st column equal to:

$$[x_{1j}\ x_{2j}\ \cdots\ x_{nj}]^T, \quad \text{for each } j = 0, 1, \ldots, k.$$

$$\text{Let} \quad \boldsymbol{b} = [b_0\ b_1\ \cdots\ b_k]^T,$$

$$\text{and} \quad \boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^T.$$

The normal equations are:

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y},$$

which have the solution:

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

The vector $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0 \; \widehat{\beta}_1 \; \cdots \; \widehat{\beta}_k]^T$ of least squares estimates is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

In an `R` session suppose that the response and independent variables are called `y` and `x1`, `x2`, `x3`. Then, we could obtain the least squares estimates using the command:

```
summary(lm(y~x1+x2+x3)).
```

Example 8: *Regression analysis of mesquite tree data (dataset available in course website)*

The data in this example concern the prediction of total production of photosynthetic biomass (leaves) of mesquite trees by using certain easily measured aspects of the plant as opposed to actual harvesting of the mesquite.

Data on 20 mesquite trees from the same geographic location are given on page 91N.

The variables are:

z = leafwt = total weight (in grams) of photosynthetic material derived from the actual harvesting of mesquite

u_1 = diam1 = canopy diameter (in meters) measured along the longest axis of the tree parallel to the ground

u_2 = diam2 = canopy diameter (in meters) measured along the shortest axis of the tree parallel to the ground

u_3 = totht = total height (in meters) of the mesquite tree

u_4 = canht = canopy height (in meters) of the mesquite tree

It's desired to be able to predict leafwt using the other measurements.

It's more natural to consider a *multiplicative* model rather than a linear one since leaf weight should be nearly proportional to canopy volume, and canopy volume should be nearly proportional to the product of canopy dimensions.

The (multiplicative) model we will consider is:

$$Z = \beta_0^* u_1^{\beta_1} u_2^{\beta_2} u_3^{\beta_3} u_4^{\beta_4} \epsilon'.$$

To analyze the data by means of a linear model, we can take logarithms, which leads to:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$, where

$$\beta_0 = \log(\beta_0^*), \quad \epsilon = \log(\epsilon'),$$

$$y = \log(z), \qquad x_1 = \log(u_1), \quad x_2 = \log(u_2),$$

$$x_3 = \log(u_3), \quad x_4 = \log(u_4).$$

As usual, we'll assume that $\epsilon \sim N(0, \sigma^2)$.

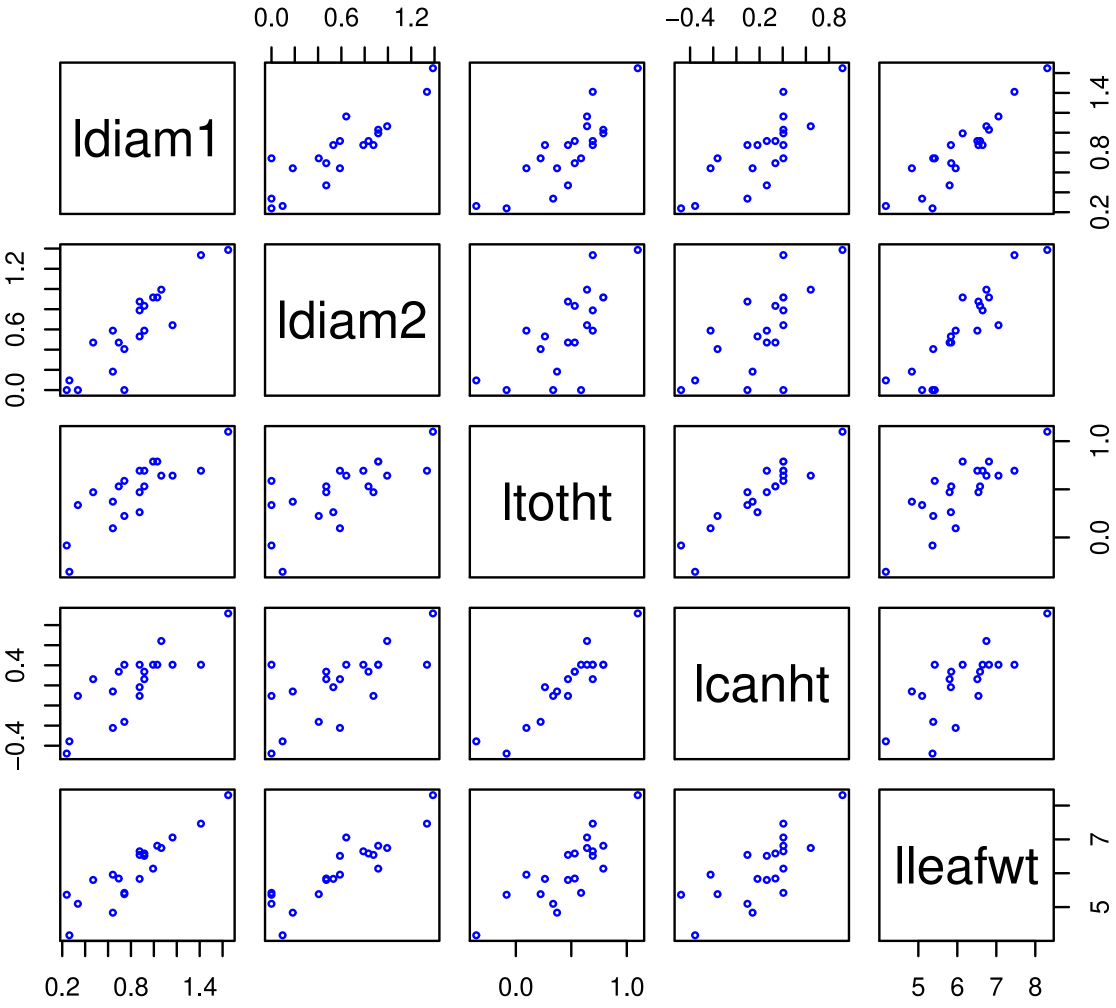
We'll be able to find a prediction interval for $y = \log(z)$ using the linear model above.

If we're 95% sure that y is between, say, y_1 and y_2, then we're 95% sure that z is between e^{y_1} and e^{y_2}. ☺ (See Supplementary Notes).

Mesquite tree data

diam1	diam2	totht	canht	leafwt
2.50	2.3	1.70	1.40	723.0
5.20	4.0	3.00	2.50	4052.0
2.00	1.6	1.70	1.40	345.0
1.60	1.6	1.60	1.30	330.9
1.40	1.0	1.40	1.10	163.5
3.20	1.9	1.90	1.50	1160.0
1.90	1.8	1.10	0.80	386.6
2.40	2.4	1.60	1.10	693.5
2.50	1.8	2.00	1.30	674.4
2.10	1.5	1.25	0.85	217.5
2.40	2.2	2.00	1.50	771.3
2.40	1.7	1.30	1.20	341.7
1.90	1.2	1.45	1.15	125.7
2.70	2.5	2.20	1.50	462.5
1.30	1.1	0.70	0.70	64.5
2.90	2.7	1.90	1.90	850.6
2.10	1.0	1.80	1.50	226.0
4.10	3.8	2.00	1.50	1745.1
2.80	2.5	2.20	1.50	908.0
1.27	1.0	0.92	0.62	213.5

Scatterplot matrix



The horizontal axis for a column of plots corresponds to the variable named in that column. Likewise the vertical axis for each row corresponds to the variable named in that row.

The least squares estimates found in R are as follows:

$$\widehat{\beta}_0 = 4.3043 \qquad \widehat{\beta}_1 = 0.9579$$

$$\widehat{\beta}_2 = 1.0194 \qquad \widehat{\beta}_3 = 1.1650$$

$$\widehat{\beta}_4 = -0.6040$$

The coefficient of x_4 seems to be a little troubling. What does it imply?

Variance estimation (estimation of σ^2):

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik},$$

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

$$\widehat{\sigma}^2 = \frac{SSE}{n-k-1} = MSE.$$

Coefficient of determination:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

$$SST = SSR + SSE.$$

$$R^2 = \frac{SSR}{SST},$$

the fraction of total variation in the y_i's explained by the model being considered.

Inference: Confidence intervals and tests for β_i

Done exactly as in polynomial regression. The formulas are exactly the same.

$$\text{Var}(\hat{\beta}_i) = \sigma^2 c_{i+1},$$

with c_j the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Confidence intervals for $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ (i.e. the conditional mean of Y) and prediction intervals for Y at $(x_1, \ldots, x_k)$ are also the same.

$$\widehat{\mu}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$$

$$\boldsymbol{x}^T = [1 \;\; x_1 \;\; x_2 \;\; \cdots \;\; x_k]$$

Confidence interval for $E(Y)$ at $\boldsymbol{X} = \boldsymbol{x}$:

$$\widehat{\mu}(\boldsymbol{x}) \pm t_{n-k-1;\alpha/2} \widehat{\sigma} \sqrt{\boldsymbol{x}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}}$$

Prediction interval for Y at $\boldsymbol{X} = \boldsymbol{x}$:

$$\widehat{\mu}(\boldsymbol{x}) \pm t_{n-k-1;\alpha/2} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}}$$

The F-test (or ‘model utility’ test in textbook):

An interesting hypothesis to test is:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

If our model is correct, the last hypothesis says that Y *is unrelated with* $x_1, \ldots, x_k$.

A test statistic for this hypothesis is:

$$F = \frac{SSR/k}{MSE} = \frac{R^2/k}{(1-R^2)/(n-k-1)}.$$

If $\epsilon_1, \ldots, \epsilon_n$ are normally distributed, then under H_0, F has the so-called F-distribution with degrees of freedom k and $n-k-1$, i.e. $F_{k,n-k-1}$.

Rejection region: Reject H_0 and conclude that at least one $\beta_j \neq 0$ (for *some* $j = 0, 1, 2, \ldots, k$)

$$\text{if } F \geq F_{k,n-k-1;\alpha}, \quad \text{where}$$

$F_{k,n-k-1;\alpha}$ is the $(1-\alpha)^{th}$ quantile of $F_{k,n-k-1}$.

Alternatively, you may compute the p-value: $P(F > F_{obs})$ where $F \sim F_{k,n-k-1}$ and F_{obs} is the *observed value* of the test statistic F above.

You can find this value (or get a good approximation) using the F distribution tables or via `R`.

Reduction method of testing:

Suppose, for example, that $k = 4$ and we want to test the hypothesis:

$$H_0 : \beta_3 = 0,\ \beta_4 = 0.$$

This hypothesis says that *x_3 and x_4 are not needed in the same model that contains x_1 and x_2.*

This hypothesis would be interesting in a case where x_1 and x_2 were known to be related to Y, but it's not clear whether x_3 and x_4 are related to Y or not.

$$y = \text{measure of product quality}$$

$$x_1 = \text{experience of production line workers}$$

$$x_2 = \text{purity measure of raw materials}$$

The variables x_1 and x_2 are well-known to affect production quality. How about

$$\begin{aligned} x_3 &= \text{average height of production} \\ &\quad\ \text{line workers,} \quad \text{or} \\ x_4 &= \text{temperature of factory at} \\ &\quad\ \text{production time?} \end{aligned}$$

You might be tempted to test $H_0 : \beta_3 = 0,\ \beta_4 = 0$ by carrying out separate tests of $H_{03} : \beta_3 = 0$ and $H_{04} : \beta_4 = 0$. But *this is not a good idea. Doing so can lead to incorrect conclusions*.

Instead, you should use the *reduction method.*

Formal details of the reduction method:

Without loss of generality, suppose we want to test the last l coefficients to be 0 for a model with k coefficients originally (i.e. k predictors):

$$H_0 : \beta_{k-\ell+1} = \beta_{k-\ell+2} = \cdots = \beta_k = 0.$$

Estimate $\beta_0, \beta_1, \ldots, \beta_k$ in the *full* model containing all k independent variables. Let SSE_f be the SSE for this model.

Estimate $\beta_0, \beta_1, \ldots, \beta_{k-\ell}$ in the *reduced* model that contains only $x_1, \ldots, x_{k-\ell}$. Let SSE_r be the SSE for this model.

The test statistic is:

$$F = \frac{(SSE_r - SSE_f)/\ell}{SSE_f/(n-k-1)}.$$

Reject H_0 if and only if $F \geq F_{\ell,n-k-1;\alpha}$. Can also compute the p-value accordingly.

The F statistic is necessarily positive. This is because SSE_r *must be* at least as large as SSE_f. *Why?*

Example 8 (continued): Inference for mesquite tree data

In an R session, I defined `y`, `x1`, `x2`, `x3` and `x4` to be the natural logarithms of z, u_1, u_2, u_3 and u_4, respectively.

Then I used the command

```
fit=lm(y~x1+x2+x3+x4).
```

Test of model utility

In the last line of the output from `summary(fit)`, we see that the F-statistic for the test of model utility is 28.6 with a P-value of $7.307 \cdot 10^{-7}$. So, we conclude that the model is useful, in the sense that at least one of the four independent variables has a nonzero coefficient.

Test for a single regression coefficient

Suppose we want to test $H_0 : \beta_2 = 0$, which says "$x_2 =$ log(canopy diameter along shorter axis) is not needed in the same model with the other three independent variables."

According to the output from `summary(fit)`, the P-value for the test of this hypothesis is 0.0353. Therefore, we *reject H_0 and conclude that we should have x_2 in the model.*

Coefficient of determination

Note that the R^2 value is pretty high: 0.884. So, the model is explaining about 88% of the variation in log(leaf weight). That's good!

Estimation of mean response and prediction

Consider trees with the following dimensions:

$$u_1 = 2.5 \quad u_2 = 2.0 \quad u_3 = 1.7 \quad u_4 = 0.92$$

Estimate the average log(leaf weight) of such trees by means of a 95% confidence interval. Using the commands

```
X=data.frame(x1=log(2.5),x2=log(2),
  x3=log(1.7),x4=log(0.92))
predict(fit,X,interval="conf")
```

the interval is

$$(5.98, 7.13).$$

Guess the leaf weight of a single tree with the dimensions above. Using

```
predict(fit,X,interval="predict")
```

we're 95% sure that the log(leaf weight) is in

$$(5.58, 7.53),$$

and so we're 95% sure that the leaf weight is between $e^{5.58} = 265.1$ grams and $e^{7.53} = 1863.1$ grams.

Testing that a subset of variables is not needed

For these data, it makes sense that some of the independent variables are highly correlated with each other.

For example, it wouldn't be surprising that tall trees tend to have tall canopies. This in fact is the case. Look at the third row and fourth column of the scatterplot matrix.

This suggests that maybe we don't need all four independent variables in the model: *the information about leaf weight conveyed by tree height may be contained in canopy height.*

Likewise, we may need only one of the two canopy diameters. Let's test the following hypothesis then:

$$H_0 : \beta_1 = 0, \beta_3 = 0.$$

This hypothesis says we don't need log(diameter 1) and log(total height) in the same model with log(diameter 2) and log(canopy height).

Use the *reduction method* to test the hypothesis. The command

```
anova(fit)
```

gives us the SSE for the *full* model:

$$SSE_f = 2.0368.$$

Now we fit a *reduced* model that has in it only the two independent variables `x2` (log-diameter 2) and `x4` (log-canopy height). The SSE for this model is

$$SSE_r = 2.8261.$$

The F-statistic is

$$F = \frac{(2.8261 - 2.0368)/2}{2.0368/15} = 2.906.$$

From the F table or R we find $F_{2,15;0.05} = 3.68$. So, we would not reject H_0. There isn't sufficient evidence to conclude that we need `x1` and `x3` in the model.

Note that R^2 for the reduced model is 0.84. This is only a bit smaller than the R^2 of 0.88 for the full model.

In other words, *the reduced model explains almost as much of the variance in leaf weight as does the full model*. For reasons of simplicity, we might then want to go with the simpler model.

To predict leaf weight, we'd only have to make two measurements from a tree as opposed to four.

Model selection

In multiple regression we're faced with the problem of choosing a good subset of independent variables. This is especially important *when the researcher has measured all independent variables but the kitchen sink.* ☺

There are many tools for selecting a good subset of variables. We'll talk about three: R^2, AIC and BIC.

Two basic principles:

- If the models being compared all have the same number of independent variables, just choose the model with the highest R^2 value.

- If the models have different numbers of independent variables, compare models using either AIC or BIC. *Smaller* values of AIC or BIC indicate better models.

When you don't have more than 5 or 6 independent variables, a reasonable thing to do is to fit all possible models and pick out the best candidates using AIC or BIC.

The number of possible models when you have k variables is

$$\binom{k}{0} + \binom{k}{1} + \cdots + \binom{k}{k} = 2^k.$$

The command `leaps` in the R package `leaps` will fit all possible models and give you the R^2 value for each one. You will find instructions on how to install the package and use `leaps` under R Codes and Instructions at eCampus.

The command `leaps` lists all models with the same number of variables consecutively, and in order from largest to smallest R^2 value.

According to `leaps`, the best models for the mesquite data are as follows:

No. vars.	Best model	R^2	AIC	BIC
1	`ldiam1`	0.818	26.06	29.04
2	`ldiam2, ltotht`	0.865	22.17	26.16
3	`ldiam1, ldiam2, ltotht`	0.878	22.04	27.02
4	`ldiam1, ldiam2, lcanht, ltotht`	0.884	23.07	29.05

So, *AIC prefers the three variable model with* `ldiam1, ldiam2, ltotht` and *BIC prefers the two variable model with* `ldiam2, ltotht`.

Plotting residuals

As in straight line and polynomial regression, residuals are valuable tools for helping us determine if the model assumptions are met (at least approximately).

If the model is correct, the residuals

$$e_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n,$$

should behave like a random sample from a normal distribution.

Standardized residuals:

$$e_i^* = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}}, \quad i = 1, \ldots, n.$$

These should behave like a random sample from a *standard* normal distribution.

Methods of plotting:

Plot	Purpose
e_i^* vs. i	Check independence of errors.
e_i^* vs. $\widehat{y}_i$	Check constant variance.
e_i^* vs. an independent variable	Check for nonlinear relationship.
Normal probability plot or kernel density estimate	Check normality

Influential observations

Sometimes a single observation

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$$

can have a big influence on the estimates of $\beta_0, \beta_1, \ldots, \beta_k$ and/or σ^2.

Such an observation is called an *outlier*. There are two basic kinds of outliers:

1. $\boldsymbol{x}_i$ is in the "middle" of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ but y_i is unusually large or small.

2. $\boldsymbol{x}_i$ is far away from the rest of the $\boldsymbol{x}_j$s.

The two kinds of outliers have different effects. *The first kind doesn't have much effect on the $\widehat{\beta}_i$s, but can inflate $\widehat{\sigma}^2$.*

The second kind can have a big effect on the $\widehat{\beta}_i$s.

In multiple regression spotting outliers can be difficult because with more than two independent variables we can't plot y against the vector of independent variables.

A useful diagnostic is *Cook's D*. This measures the distance between $\widehat{\boldsymbol{\beta}}$ and an estimate of $\boldsymbol{\beta}$ calculated *after deleting the ith observation from the data set*.

This means there is one Cook's D value for each of the n cases in the data set. If a D value is "big," then the corresponding observation is influential, i.e., an outlier.

How large is "large?" Cook's D values less than 1 are not usually indicative of an outlier.

Values of D larger than 1.5 suggest the possibility of an outlier. If $D > 1.5$, fit the model without the suspicious case and look at how much the estimates of the β_is change.

In R suppose you have put regression output from `lm` into an object called `fit`. To obtain Cook's D value for all the data, use the command

```
output=cooks.distance(fit).
```

You could then plot the Cook's D values as follows:

```
plot(output,ylab=‘‘Cook’s D’’).
```

For more details on Cook's D values and their computation, see Supplementary Notes.