Part II

Multiple Linear Regression

Contents

1 Multiple linear regression

1.1 The model

A linear regression model that contains more than one predictor variable is called a **multiple linear regression model**. This model generalises the simple linear regression in two ways. It allows the mean function $\mathbb{E}(Y)$ to depend on more than one predictor (or regressor) variables and to have shapes other than straight lines, although it does not allow for arbitrary shapes.

1.1.1 Example

Consider the relation between the income and education of a person. It is expected that, on an average, higher level of education provides higher income. However, most people have higher income when they are older than when they are young, regardless of education. A multiple regression model that might describe this relationship is

$$\mu_{x_1,x_2} = \mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where X_1 is the level of education, X_2 is the age of a person, and $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2)$ is the mean income of a person with a specific level of education x_1 and of a specific age x_2. This is

a multiple linear regression model with two predictor variables. The model is linear because the mean is a linear function of the unknown parameters β_0, β_1, β_2.

Often it is observed that the income tends to rise less rapidly in the later earning years than in early years. To accommodate such possibility, we might extend the model to

$$\mu_{x_1,x_2} = \mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2.$$

This is how we proceed for regression modelling in a real life situation. One needs to consider the experimental condition and the phenomenon before taking the decision on how many, why and how to choose the dependent and independent variables.

1.1.2 Interactions and higher order terms

A linear regression model may take the following form:

$$\mu_{x_1,x_2} = \mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2. \tag{1.1}$$

The cross-product term, $x_1 x_2$, represents an **interaction** effect between the two predictor variables, X_1 and X_2. Interaction means that the effect produced by a change in the predictor variable on the response depends on the level of the other predictor variable(s).

A linear regression model may also take the following form:

$$\mu_{x_1} = \mathbb{E}(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3. \tag{1.2}$$

This model is also a linear regression model and is referred to as a **polynomial regression model**. Polynomial regression models contain squared and higher order terms of the predictor variables making the response surface curvilinear. It is often advised to avoid higher order term when building the models. The reasons will be explained further in the course.

1.1.3 General case

A multiple linear regression model can always be written in the form

$$\mu_{\boldsymbol{x}} = \mathbb{E}(Y|X = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

where $X = (X_1, \ldots, X_p)$ and $\boldsymbol{x} = (x_1, \ldots, x_p)$. For instance, if we let $X_3 = X_1 X_2$ in (1.1) or $X_2 = X_1^2$ and $X_3 = X_1^3$ in (1.2), we recover a multiple linear regression model with three predictor variables.

The parameters $\beta_0, \ldots, \beta_p$ are unknown parameters to be estimated. Parameter β_0 is the intercept. Each slope β_i with $i = 1, \ldots, p$ represents the expected change in the response per unit change in X_i when all of the remaining predictor variables X_j with $j \neq i$ are held constant. For this reason the parameters $\beta_0, \ldots, \beta_p$ are often called **partial regression coefficients**.

When $p = 1$, X has only one element, and we recover the simple linear regression model discussed in Part I. When $p = 2$, the mean function $\mathbb{E}(Y|X = \boldsymbol{x})$ corresponds to a plane in 3 dimensions. When $p \geq 3$, the fitted mean function is a hyperplane, the generalisation of a p-dimensional plane in a $(p+1)$-dimensional space. We cannot draw a general p-dimensional plane in our three-dimensional world.

As in the case of the simple linear regression, $\mathbb{E}(Y|X=\boldsymbol{x})$ does not adequately describe the data which show some randomness. To deal with this problem, for each observation y_i of Y we introduce a random error ε_i and write

$$y_i = \mu_i + \varepsilon_i = \mathbb{E}(Y|X=\boldsymbol{x}_i) + \varepsilon_i$$

where $\boldsymbol{x}_i$ is the level of X at which y_i was observed. The standard assumptions on the sampling distribution of the random errors ε_i are

$$\varepsilon_i \overset{ind}{\sim} N(0,\sigma^2)$$

implying that y_i is sampled from

$$Y_i = \mu_i + \varepsilon_i \overset{ind}{\sim} N(\mu_i,\sigma^2).$$

Remark 1.1. We have assumed that the predictor variables $X_1, X_2, \ldots, X_p$ are fixed (nonrandom) variables, measured without error. However, all of our results are still valid for the case where the predictors are random variables. This is certainly important, because when regression data arise from an **observational study**, some or most of the predictors will be random variables. When the data result from a **designed experiment**, it is more likely that the X_i's will be fixed variables. When the X_i's are random variables, it is only necessary that the observations on each predictor be independent and that the distribution not depend on the regression coefficients (the β_i's) or on σ^2. When testing hypotheses or constructing CIs, we will have to assume that the conditional distribution of Y given $x_1, x_2, \ldots, x_p$ be normal with mean $\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ and variance $\mathrm{Var}(Y) = \sigma^2$. □

1.2 Least squares estimation

While it is possible to estimate the parameters of more complex linear models with methods similar to those we have seen in simple linear regression, the computations become rather cumbersome very quickly. Thus, we will employ linear algebra methods to make the computations more efficient.

Suppose that $y_1, y_2, \ldots, y_n$ are responses of Y at the values $x_{11}, x_{21}, \ldots, x_{n1}$ of the variable X_1, and at the values $x_{12}, x_{22}, \ldots, x_{n2}$ of the variable X_2, and so on, as shown in Table 1.

		Predictors			
Case, i	Response, y_i	X_1	X_2	$\ldots$	X_p
1	y_1	x_{11}	x_{12}	$\ldots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{np}

Table 1: Data for multiple linear regression.

Thus each response y_i can be written as

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i$$

where ε_i is an unknown error incurred in the observation of y_i. The least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the values of $\beta_0, \beta_1, \ldots, \beta_p$ which minimize the error sum of squares,

$$SS_\varepsilon = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p)^2. \tag{1.3}$$

To find these values we need to differentiate SS_ε with respect to each $\beta_0, \beta_1, \ldots, \beta_p$:

$$\begin{aligned}
\frac{\partial SS_\varepsilon}{\partial \beta_0} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p), \\
\frac{\partial SS_\varepsilon}{\partial \beta_1} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p) x_{i1}, \\
&\vdots \qquad\qquad \vdots \\
\frac{\partial SS_\varepsilon}{\partial \beta_p} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p) x_{ip},
\end{aligned}$$

and equate

$$\left.\frac{\partial SS_\varepsilon}{\partial \beta_0}\right|_{\beta_i = \hat{\beta}_i} = \left.\frac{\partial SS_\varepsilon}{\partial \beta_1}\right|_{\beta_i = \hat{\beta}_i} = \ldots = \left.\frac{\partial SS_\varepsilon}{\partial \beta_p}\right|_{\beta_i = \hat{\beta}_i} = 0$$

for all i at the same time. This gives a system of $(p+1)$ equations in $(p+1)$ unknowns, called **normal equations**, one for each of the unknown regression coefficients. Once the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are found, one needs to perform the second derivative test to make sure they indeed minimize, not maximize the sum of squared errors.

In practice, it is more convenient to use the matrix notation of linear models. The multiple linear regression model then takes the form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The error sum of squares (1.3) can now be written as

$$SS_\varepsilon = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$$

The claim below was proven in Section 4.3 of Part I.

Claim 1.1:

The least squares estimate of $\boldsymbol{\beta}$ minimising the SS_ε is given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{1.4}$$

Example 1: Soft drink bottler

A soft drink bottler is analysing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. The analyst responsible for the study has suggested that the two most important variables affecting the delivery time (y) are the number of cases of product stocked (x_1) and the distance walked by the route driver (x_2). The analyst has collected 25 observations on delivery time, which are shown in Table 2.

y	16.68	11.50	12.03	14.88	13.75	18.11	8.00	17.83	79.24	21.50	40.33	21.00	13.50
x_1	7	3	3	4	6	7	2	7	30	5	16	10	4
x_2	560	220	340	80	150	330	110	210	1460	605	688	215	255
y	19.75	24.00	29.00	15.35	19.00	9.50	35.10	17.90	52.32	18.75	19.83	10.75	
x_1	6	9	10	6	7	3	17	10	26	9	8	4	
x_2	462	448	776	200	132	36	770	140	810	450	635	150	

Table 2: Soft drink delivery time data.

Graphics can be very useful in fitting multiple regression models. Figure 1.1 is a scatterplot matrix of the delivery time data. This is just a two-dimensional array of two-dimensional plots, where (except for the diagonal) each frame contains a scatter diagram. Thus, each plot is an attempt to shed light on the relationship between a pair of variables. This is often a better summary of the relationships than a numerical summary (such as displaying the correlation coefficients between each pair of variables) because it gives a sense of linearity or nonlinearity of the relationship and some awareness of how the individual data points are arranged over the region.

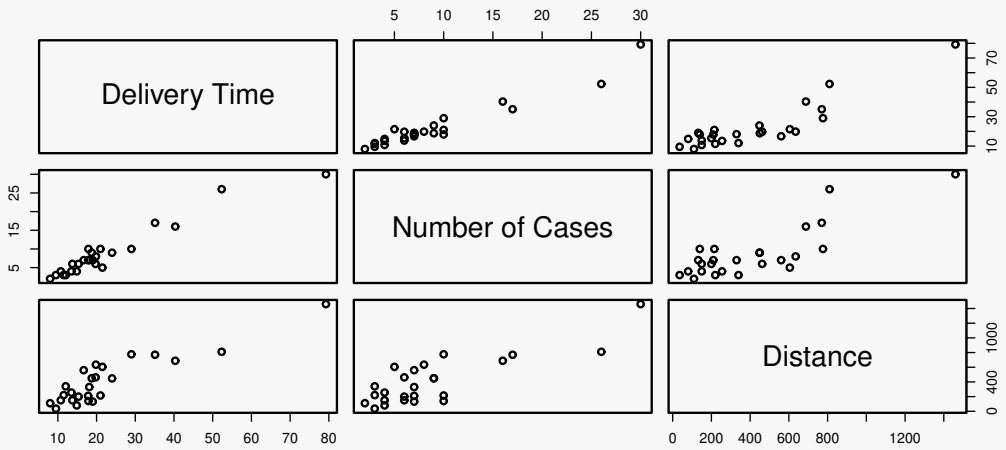

Figure 1.1: A scatterplot matrix for the delivery time data.

When there are only two regressors, sometimes a three-dimensional scatter diagram is useful in visualizing the relationship between the response and the regressors. Figure 1.2 presents this plot for the delivery time data. By spinning these plots, some software

packages permit different views of the point cloud. This view provides an indication that a multiple linear regression model may provide a reasonable fit to the data.

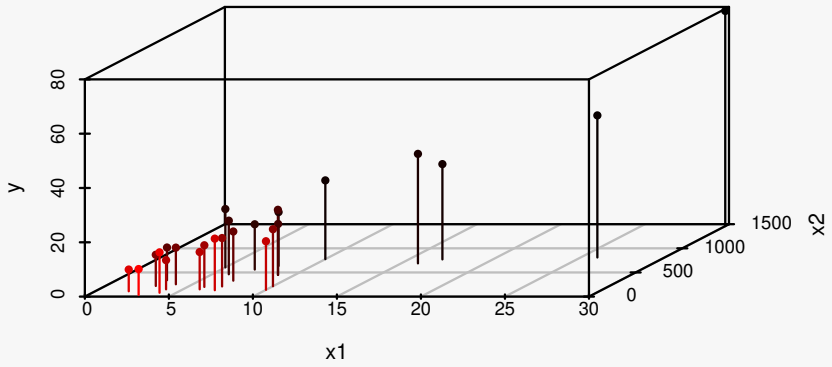

Figure 1.2: A three-dimensional scatterplot of the delivery time data.

To fit the multiple regression model we first we form the X matrix and $\boldsymbol{y}$ vector,

$$X = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix}.$$

Then the $X^T X$ matrix is

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10.232 & 133,899 & 6,725,688 \end{bmatrix}.$$

The $X^T \boldsymbol{y}$ vector is

$$X^T \boldsymbol{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}.$$

The least squares estimator $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ is then

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10.232 & 133,899 & 6,725,688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 0.11321518 & -0.00444859 & -0.00008367 \\ -0.00444859 & 0.00274378 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix} = \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}.$$

The least squares fit (with the regression coefficients reported to five decimals) is

$$\hat{y} = 2.34123 + 1.61591 x_1 + 0.01438 x_2.$$

1.3 Properties of the least squares estimator

We have made an assumption that random errors are normally distributed independent random variables. In terms of a multivariate normal distribution, this means that

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I) \tag{1.5}$$

implying

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \varepsilon \sim N_n(X\boldsymbol{\beta}, \sigma^2 I) \tag{1.6}$$

Here we recall that capital letters denote random variables, so that e.g. $\boldsymbol{y}$ is a response of $\boldsymbol{Y}$ and ε is a response of ε. For further convenience, we introduce a short-hand notation

$$C = (X^T X)^{-1}.$$

Claim 1.2:

The sampling distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C). \tag{1.7}$$

Proof. From (1.4) we know that $\hat{\boldsymbol{\beta}} = CX^T\boldsymbol{Y}$, viewed as an estimator. Hence it is a multivariate normally distributed random variable with mean

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(CX^T\boldsymbol{Y}) = CX^T\,\mathbb{E}(\boldsymbol{Y}) = CX^T\,\mathbb{E}(X\boldsymbol{\beta} + \varepsilon) = CX^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

and variance

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(CX^T\boldsymbol{Y}) = CX^T\,\mathrm{Var}(\boldsymbol{Y})\,(CX^T)^T = CX^T\sigma^2 XC^T = \sigma^2 CC^{-1}C = \sigma^2 C.$$

To end this subsection, we want to verify that (1.7) agrees with our results for the simple linear regression model. When $p = 1$, we must have

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right), \qquad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right).$$

Indeed, writing $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ in terms of its components we get $\mathbb{E}(\hat{\beta}_0) = \beta_0$ and $\mathbb{E}(\hat{\beta}_1) = \beta_1$. Showing that $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ reproduces $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$ requires a little bit more of work. At the end of Section 4.3 of Part I we showed that

$$C = (X^TX)^{-1} = \frac{1}{s_{xx}}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Comparing with

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) \end{bmatrix} = \sigma^2 C$$

we find

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2\sum_{i=1}^n x_i^2}{ns_{xx}}, \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}.$$

Variance $\mathrm{Var}(\hat{\beta}_1)$ is already in the wanted form. To put $\mathrm{Var}(\hat{\beta}_0)$ into the wanted form we note that

$$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

giving

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n s_{xx}} = \sigma^2 \frac{s_{xx} + n\bar{x}^2}{n s_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$$

thus agreeing with our earlier results.

1.4 Estimating variance of the random error term

Recall from Section 1.4 in Part I that the residual sum of squares,

$$SS_E = \sum_{i=1}^{n} e_i^2$$

measures how closely the simple linear regression model fits the data and can be used to estimate σ^2. The same is true for multiple linear regression.

Let $\boldsymbol{e}$ denote the vector of residuals,

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}},$$

where $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$ is the vector of fitted values (points on the fitted regression plane). Then the least squares estimator $\boldsymbol{E}$ of $\boldsymbol{e}$ is then given by

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}},$$

where $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is the random variable of responses and $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$ is its least squares estimator. It can be shown that[1]

$$\mathbb{E}(\boldsymbol{E}^T\boldsymbol{E}) = \sigma^2(n-p-1).$$

Therefore

$$\hat{\sigma}^2 \equiv MS_E = \frac{SS_E}{n-p-1} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p-1}$$

is an unbiased estimate of σ^2. The number

$$\nu_E = n-p-1 \tag{1.8}$$

is referred to as the **residual degree of freedom**. It is the number of linearly independent residuals e_i. We will discuss this point in more detail in Section 3.

Remark 1.2. The estimate $\hat{\sigma}^2$ is model dependent. Since σ^2 is the variance of the errors (the unexplained noise about the regression line), when comparing two models for the same data, we would usually choose a model with a smaller residual mean square. □

[1]For a prove of this statement see Appendix C.3 in (Montgomery et. al., 2012). However the techniques used in that prove go beyond the scope of the present course.

1.5 Confidence intervals

Confidence intervals (CIs) on individual regression coefficients and CIs on the mean response given specific levels of the predictors play the same important role in multiple regression that they do in simple linear regression. This section develops the one-at-a-time CIs for these cases. More general methods, leading to simultaneous CIs, will be discussed in the 3rd year module *Multivariate Statistics*.

1.5.1 Confidence interval on regression coefficients

We showed above that $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$. This means that the marginal distribution of any estimator $\hat{\beta}_j$ of β_j with $j = 0, 1, \ldots, p$ is

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$$

where c_{jj} is the jth diagonal element of the matrix $C = (X^T X)^{-1}$. Consequently,

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{n-p-1},$$

where we recall that $\hat{\sigma}^2 = MS_E = SS_E/(n-p-1)$. Hence, a $100(1-\alpha)\%$ confidence interval for the regression coefficient, β_j, is given by

$$\mathrm{CI}(\beta_j) = \left[\hat{\beta}_j - t_{\alpha/2,n-p-1} \cdot \mathrm{se}(\hat{\beta}_j),\ \hat{\beta}_j + t_{\alpha/2,n-p-1} \cdot \mathrm{se}(\hat{\beta}_j)\right] \tag{1.9}$$

where $\mathrm{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}$.

1.5.2 Confidence interval on the mean response

We now want to construct a CI on the mean response

$$\mu_0 = \mathbb{E}(Y|X = \boldsymbol{x}_0) = \boldsymbol{x}_0^T \boldsymbol{\beta}$$

at a particular level $\boldsymbol{x}_0^T = (1, x_{01}, x_{02}, \ldots, x_{0p})$ of the predictor variables $X_1, X_2, \ldots, X_p$. Replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ we obtain the least squares estimator of μ_0,

$$\hat{\mu}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + x_{01}\hat{\beta}_1 + \ldots + x_{0p}\hat{\beta}_p.$$

Claim 1.3:

The sampling distribution of $\hat{\mu}_0$ is

$$\hat{\mu}_0 \sim N(\mu_0, \sigma^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0).$$

Proof. Since $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$, we only need to compute $\mathbb{E}(\hat{\mu}_0)$ and $\mathrm{Var}(\hat{\mu}_0)$:

$$\mathbb{E}(\hat{\mu}_0) = \mathbb{E}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T\, \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T \boldsymbol{\beta}$$

and

$$\mathrm{Var}(\hat{\mu}_0) = \mathrm{Var}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T\, \mathrm{Var}(\hat{\boldsymbol{\beta}})\, \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0.$$

Therefore a $100(1-\alpha)\%$ CI on the mean response μ_0 at the level $\boldsymbol{x}_0$ is given by

$$\text{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\mu}_0)\right] \tag{1.10}$$

where $\hat{\mu}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ and $\text{se}(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0}$. This is a generalisation of the CI for the mean response in the simple linear regression.

1.5.3 Prediction interval on a new observation

The regression model can be used to predict future (new) observation y_0 of Y corresponding to a particular level of the predictor variables, say $\boldsymbol{x}_0$. The prediction interval (PI) takes into account both the error from the estimated model and the error associated with a new observation.

Claim 1.4:

The sampling distribution of $\hat{Y}_0$ is

$$\hat{Y}_0 \sim N(y_0, \sigma^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0))$$

Proof. Future observations are sampled from $\hat{Y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} + \varepsilon_0$ where $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$ and $\varepsilon_0 \sim N(0, \sigma^2)$. Thus we only need to compute the mean and variance of $\hat{Y}_0$:

$$\mathbb{E}(\hat{Y}_0) = \mathbb{E}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) + \mathbb{E}(\varepsilon_0) = \boldsymbol{x}_0^T \mathbb{E}(\hat{\boldsymbol{\beta}}) + 0 = \boldsymbol{x}_0^T \boldsymbol{\beta} = y_0$$

and

$$\text{Var}(\hat{Y}_0) = \text{Var}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) + \text{Var}(\varepsilon_0) = \boldsymbol{x}_0^T \sigma^2 C \boldsymbol{x}_0 + \sigma^2 = \sigma^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0).$$

Therefore a $100(1-\alpha)\%$ PI on a new observation y_0 at $\boldsymbol{x}_0$ is

$$\text{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{y}_0)\right] \tag{1.11}$$

where $\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ and $\text{se}(\hat{y}_0) = \sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0)}$. This is a generalisation of the PI for a new observation in the simple linear regression.

Remark 1.3. The confidence intervals that we have computed were confidence intervals for a single parameter at a time. Some problems require that several confidence or prediction intervals be constructed using the same sample data. In these cases, the analyst is usually interested in specifying a confidence coefficient that applies simultaneously to the entire set of interval estimates. A set of confidence or prediction intervals that are all true simultaneously with probability $1-\alpha$ are called **simultaneous** or **joint confidence** or **joint prediction** intervals. Their study goes beyond the scope of the present course, however a curious student may learn more about such intervals in e.g. Section 3.4.3 of (Montgomery et. al., 2012). □

Example 2: Soft drink bottler (continued)

We want to find a 95% CI on the slope β_1. We found that $\hat{\beta}_1 = 1.61591$ mins per case. The diagonal element of $C = (X^TX)^{-1}$ corresponding to β_1 is $c_{11} = 0.00274378$, and $\hat{\sigma}^2 = 10.6239$. (We enumerate rows and columns of C by $0, 1, 2, \ldots, p$.) Then

$$\hat{\beta}_1 \pm t_{0.025,25-2-1}\sqrt{\hat{\sigma}^2 c_{11}} = 1.61591 \pm 2.074\sqrt{10.6239 \times 0.00274378} = 1.61591 \pm 0.3541.$$

Hence a 95% CI on β_1 is

$$\text{CI}(\beta_1) = [1.26181,\ 1.97001].$$

Loosely speaking, there is a 95% probability that a single case carried by the delivery man contributes from 1.262 to 1.970 minutes to the delivery time.

Next, we want to find a 95% CI on the mean delivery time for an outlet requiring $x_1 = 8$ cases and where the distance walked by the delivery man is $x_2 = 275$ feet. That is,

$$\boldsymbol{x}_0^T = [1,\ 8,\ 275].$$

The estimated mean delivery time at $\boldsymbol{x}_0$ is

$$\hat{\mu}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} = [1,\ 8,\ 275]\begin{bmatrix} 2.34123 \\ 1.61591 \\ 0.01438 \end{bmatrix} = 19.22 \text{ minutes}.$$

The variance of $\hat{\mu}_0$ is estimated by

$$\begin{aligned}\hat{\sigma}^2\boldsymbol{x}_0^T C\boldsymbol{x}_0 &= 10.6239\,[1,\ 8,\ 275]\begin{bmatrix} 0.1132152 & -0.0044486 & -0.0000837 \\ -0.0044486 & 0.0027438 & -0.0000479 \\ -0.0000837 & -0.0000479 & 0.0000012 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix} \\ &= 10.6239 \times 0.05346 = 0.56794.\end{aligned}$$

giving

$$\hat{\mu}_0 \pm t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2\boldsymbol{x}_0^T C\boldsymbol{x}_0} = 19.22 \pm 2.074\sqrt{0.56794}.$$

Therefore, a 95% CI on the mean delivery time μ_0 at the level $\boldsymbol{x}_0$ is

$$\text{CI}(\mu_0) = [17.66,\ 20.78].$$

Finally, we want to find a 95% PI on the delivery time at the level $\boldsymbol{x}_0$. We have computed above that $\hat{\sigma}^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 = 0.56794$. Hence

$$\hat{y}_0 \pm t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T C\boldsymbol{x}_0)} = 19.22 \pm 2.074\sqrt{10.6239 + 0.56794} = 19.22 \pm 6.9384.$$

Therefore a 95% PI on y_0 at $\boldsymbol{x}_0$ is

$$\text{PI}(y_0) = [12.28,\ 26.16].$$

Loosely speaking, there is a 95% probability that the mean delivery time of $x_1 = 8$ cases, when the distance walked by the delivery man is $x_2 = 275$ feet, is between 17.66 and 20.78 minutes. Moreover, there is a 95% probability that a new delivery will take between 12.28 and 26.16 minutes.

2 Hypothesis testing

After fitting a multiple linear regression model and computing the parameter estimates, we have to make some decisions about the model:

- Is the model a good fit for the data?
- Do we really need all the regression variables in the model? (Generally, a model with fewer predictors and about the same "explanatory power" is better.)

There are several hypothesis tests that we can utilize to answer these questions:

- F-test for significance of regression: this test checks the significance of the **whole** regression model.
- t-test for individual regression coefficients: this test checks the significance of individual regression coefficients.
- F-test for a group of regression coefficients: this test simultaneously checks the significance of a number of regression coefficients. It can also be used to test individual coefficients.

Their results are usually reported in the coefficients and ANOVA tables that are produced as routine output in multiple regression analysis. But the tests can also be conducted "manually", if necessary. The manual computations are based on the so-called *hat matrix*.

2.1 The hat matrix

Recall that $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$. Hence the vector of fitted values can be written as

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y},$$

where

$$H = X(X^T X)^{-1} X^T \tag{2.1}$$

is called **the hat matrix**. It maps the vector of observed values into the vector of fitted values, i.e. "it puts a hat on $\boldsymbol{y}$". The hat matrix plays a central role in regression analysis, thus it is worth knowing its main properties:

(i) H is a symmetric $n \times n$ matrix,

$$H^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T = H.$$

(ii) H is an idempotent,

$$H^2 = (X(X^T X)^{-1} X^T)^2 = X(X^T X)^{-1}(X^T X)(X^T X)^{-1} X^T = H.$$

(iii) H acts as an identity operator on X,

$$HX = X(X^T X)^{-1} X^T X = X.$$

(iv) The trace of H and its rank equal $p+1$,

$$\operatorname{rank} H = \operatorname{tr} H = \operatorname{tr}(X(X^T X)^{-1} X^T) = \operatorname{tr}((X^T X)^{-1} X^T X) = \operatorname{tr} I_{p+1} = p+1.$$

(v) $H\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T H = \mathbf{1}^T$, where $\mathbf{1}$ is an n-dimensional vector of 1's.

2.2 F-test for significance of regression

The test for **significance of regression** is a test to determine if there is a **linear relationship** between the response variable Y and **any** of the predictor variables $X_1, X_2, \ldots, X_p$. This test is often thought of as an **overall** or **global test** of **model adequacy**. The appropriate hypotheses are

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0.$$

Rejection of the null hypothesis H_0 implies that at least one of the predictor variables contributes significantly to the model. (Although we don't know which one.)

The test procedure is a generalisation of the **analysis of variance** used in the simple linear regression. The total sum of squares, SS_T, is partitioned into the sum of squares due to regression, SS_R, and the residual sum of squares, SS_E, where

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Claim 2.1:

The analysis of variance identity holds true,

$$SS_T = SS_R + SS_E. \tag{2.2}$$

Proof. We first rewrite SS_R as

$$\begin{aligned} SS_R = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^T(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) &= \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}^T\mathbf{1}\bar{y} - \bar{y}\mathbf{1}^T\hat{\boldsymbol{y}} + \bar{y}^2\mathbf{1}^T\mathbf{1} \\ &= (H\boldsymbol{y})^T(H\boldsymbol{y}) - (H\boldsymbol{y})^T\mathbf{1}\bar{y} - \bar{y}\mathbf{1}^TH\boldsymbol{y} + n\bar{y}^2 \\ &= \boldsymbol{y}^TH^TH\boldsymbol{y} - \boldsymbol{y}^TH\mathbf{1}\bar{y} - \bar{y}\mathbf{1}^TH\boldsymbol{y} + n\bar{y}^2 \\ &= \boldsymbol{y}^TH\boldsymbol{y} - \boldsymbol{y}^T\mathbf{1}\bar{y} - \bar{y}\mathbf{1}^T\boldsymbol{y} + n\bar{y}^2 \\ &= \boldsymbol{y}^TH\boldsymbol{y} - n\bar{y}^2 \end{aligned}$$

since $H\mathbf{1} = \mathbf{1}$, $\mathbf{1}^TH = \mathbf{1}^T$ and $\boldsymbol{y}^T\mathbf{1} = \mathbf{1}^T\boldsymbol{y} = \sum_{i=1}^{n} y_i = n\bar{y}$. In a similar way we rewrite SS_E as

$$SS_E = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - H\boldsymbol{y})^T(\boldsymbol{y} - H\boldsymbol{y}) = \boldsymbol{y}^T(I-H)^T(I-H)\boldsymbol{y} = \boldsymbol{y}^T(I-H)\boldsymbol{y}.$$

Therefore

$$SS_R + SS_E = \boldsymbol{y}^TH\boldsymbol{y} - n\bar{y}^2 + \boldsymbol{y}^T(I-H)\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2$$

On the other hand,

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}y_i^2 - 2\sum_{i=1}^{n}y_i\bar{y} + \sum_{i=1}^{n}\bar{y}^2 = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2$$

which is the wanted result.

It can also be shown that, if the null hypothesis H_0 is true, that the sampling distributions of SS_R and SS_E are

$$\frac{SS_R}{\sigma^2} \sim \chi^2_p, \qquad \frac{SS_E}{\sigma^2} \sim \chi^2_{n-p-1},$$

and that SS_R and SS_E are independent.[2] By definition of the F-distribution, we then have that

$$F = \frac{SS_R/p}{SS_E/(n-p-1)} = \frac{MS_R}{MS_E} \sim F_{p,n-p-1}.$$

We reject the null hypothesis H_0 at a significance level α if

$$F_{cal} > F_{\alpha,p,n-p-1},$$

where F_{cal} denotes the value of the variance ratio F calculated for a given data and $F_{\alpha,p,n-p-1}$ is the percentile of the F-distribution corresponding to a cumulative probability of $(1-\alpha)$,

$$P(F > F_{\alpha,p,n-p-1}) = \alpha.$$

The calculation of the overall F-test can be summarized in the form of an Analysis of Variance (ANOVA) table shown in Table 3 below.

Source of variation	d.o.f.	SS	MS	F	P-value
Regression	$\nu_R = p$	SS_R	$MS_R = \frac{SS_R}{\nu_R}$	$F = \frac{MS_R}{MS_E}$	α
Residual	$\nu_E = n-p-1$	SS_E	$MS_E = \frac{SS_E}{\nu_E}$		
Total	$\nu_T = n-1$	SS_T			

Table 3: ANOVA table for significance of regression in multiple regression.

2.3 t-test for individual regression coefficients

Once we have determined that at least one of the predictors is important, a logical question becomes which one(s). Adding a variable to a regression model always causes the regression sum of squares, SS_R, to increase and the residual sum of squares, SS_E, to decrease.

We must decide whether the increase in SS_R is sufficient to warrant using the additional predictor in the model. So we must be careful to include only predictors that are of real value in explaining the response.

The hypotheses for testing the significance of any individual regression coefficient, such as β_j, are

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

They test if the slope associated with the j-th predictor is significantly different from zero. The test statistic for the null hypothesis is

$$T = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1},$$

[2]For complete details see e.g. Appendix C in (Montgomery et. al., 2012).

where $\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}$ and t_{n-p-1} is the Student's t distribution with $(n-p-1)$ degrees of freedom.

The null hypothesis $H_0 : \beta_j = 0$ is rejected if is the calculated value t of T is in the rejection region, that is

$$|t| > t_{\alpha/2,n-p-1}.$$

Note that this is really a **partial** or **marginal test**. That means that the test statistic (and thus the p-value of the test) depends not just on the j-th predictor, but also on **all other predictors** that are included in the model at the same time. Thus, if any predictor is added or removed from a regression model, hypothesis tests for individual slopes need to be repeated.

If the null hypothesis is rejected, we conclude that the j-th predictor has a significant influence on the response, given the other predictors in the model at the same time.

2.4 F-test for a group of predictors

Checking significance of regression variables one-by-one may sometimes be not the most effective approach in model building. We might want to simultaneously check the significance of a subset of regression coefficients instead.

We can directly determine the contribution to the regression sum of squares of a regressor, for example, X_j, given that other predictors X_i $(i \neq j)$ are included in the model by using the **extra sum of squares** method. We will use this method to directly determine the significance of a subset of regression coefficients.

Suppose our model has p predictors. We can then partition the predictors into two groups,

$$(X_1, \ldots, X_{p-r}) \quad \text{and} \quad (X_{p-r+1}, \ldots, X_p).$$

We want to simultaneously test, whether the latter group of r predictors can be removed from the model. Suppose we partition the vector of regression slopes accordingly into two parts

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

where $\boldsymbol{\beta}_1$ contains the intercept and the slopes for the first $p-r$ predictors and $\boldsymbol{\beta}_2$ contains the remaining r slopes. We want to test the hypotheses

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{vs.} \quad H_0 : \boldsymbol{\beta}_2 \neq \mathbf{0}.$$

The model may be written as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X^{red}\boldsymbol{\beta}_1 + X^{extra}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where X^{red} is the reduced $n \times (p-r+1)$ design matrix consisting of the columns of X associated with $\boldsymbol{\beta}_1$, and X^{extra} is the $n \times r$ matrix consisting of the columns of X associated with $\boldsymbol{\beta}_2$. This is called the **full model**.

The hat matrix for the full model is $H^{full} = H = X(X^TX)^{-1}X^T$, and the regression and residual sums of squares are

$$\begin{aligned} SS_R^{full} &= \boldsymbol{y}^T H \boldsymbol{y} - n\bar{y}^2, & \nu_R^{full} &= p, \\ SS_E^{full} &= \boldsymbol{y}(I-H)\boldsymbol{y}, & \nu_E^{full} &= n-p-1. \end{aligned}$$

To find the contribution of the terms in $\hat{\boldsymbol{\beta}}_2$ to the regression, we fit the model assuming that the null hypothesis $H_0 : \boldsymbol{\beta}_2 = 0$ is true. This **reduced model** is

$$\boldsymbol{y} = X^{red}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}.$$

The hat matrix for the reduced model is $H^{red} = X^{red}(X^{redT}X^{red})^{-1}X^{redT}$. The regression and residual sums of squares are

$$\begin{aligned} SS_R^{red} &= \boldsymbol{y}^T H^{red}\boldsymbol{y} - n\bar{y}^2, \qquad & \nu_R^{red} &= p - r, \\ SS_E^{red} &= \boldsymbol{y}^T (I - H^{red})\boldsymbol{y}, \qquad & \nu_E^{red} &= n - p + r - 1. \end{aligned}$$

Note that the total sum of squares is the same for both models,

$$SS_T^{full} = SS_T^{red} = SS_T = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}.$$

The regression sum of squares due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is already in the model is

$$SS_R^{extra} = SS_R^{full} - SS_R^{red} = SS_E^{red} - SS_E^{full}.$$

It has $\nu_R^{extra} = p-(p-r) = r$ degrees of freedom, and is called the **extra sum of squares due to** $\boldsymbol{\beta}_2$ because it measures the increase in the regression sum of squares that results from adding predictors $X_{p-r+1}, X_{p-r+2}, \ldots, X_p$ to a model that already contains $X_1, X_2, \ldots, X_{p-r}$. (Note that SS_R^{extra} is sometimes written as $SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - SS_R(\boldsymbol{\beta}_1)$.)

It can be shown that, assuming that the null hypothesis is true, the SS_R^{extra} is independent of SS_E^{full}, and their sampling distributions are

$$SS_R^{extra} \sim \chi_r^2, \qquad SS_E^{full} \sim \chi_{n-p-1}^2.$$

Then the test statistic for the null hypothesis is

$$F = \frac{MS_R^{extra}}{MS_E^{full}} = \frac{SS_R^{extra}/r}{SS_E^{full}/(n-p-1)} \sim F_{r,n-p-1}.$$

We reject H_0 if $F_{cal} > F_{\alpha,r,n-p-1}$, concluding that at least one of the parameters in $\boldsymbol{\beta}_2$ is not zero, and consequently at least one of the predictors $X_{p-r+1}, X_{p-r+2}, \ldots, X_p$ in X_2 contributes significantly to the regression model.

Some authors call this test a partial F-test because it measures the contribution of the predictors in X^{extra} given that the other predictors in X^{red} are already in the model. The complete computations of the partial F-test are often summarised in the form of an Analysis of Variance (ANOVA) table shown in Table 4 below.

Source of variation	d.o.f.	SS	MS	F	P-value
Residual Reduced	$\nu_E^{red} = n-p+r-1$	SS_E^{red}			
Residual Full	$\nu_E^{full} = n-p-1$	SS_E^{full}	$MS_E^{full} = \frac{SS_E^{full}}{\nu_E}$		
Extra	$\nu_R^{extra} = r$	SS_R^{extra}	$MS_R^{extra} = \frac{SS_R^{extra}}{\nu_R^{extra}}$	$F = \frac{MS_R^{extra}}{MS_E^{full}}$	α

Table 4: ANOVA table for the extra sum of squares analysis.

2.5 Coefficient of multiple determination, R^2

The coefficient of *multiple determination*, defined by

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T},$$

indicates the amount of total variability explained by the regression model. The positive square root of R^2 is called the **multiple correlation coefficient** and measures the linear association between Y and the predictor variables, $X_1, X_2, \ldots, X_p$.

The value of R^2 increases as more predictors are added to the model, even if the new predictors do not contribute significantly to the model. An increase in the value of R^2 cannot be taken as a sign to conclude that the new model is superior to the older model.

A better statistic to use is the **adjusted** R^2 statistic defined as follows:

$$R^2_{adj} = 1 - \frac{MS_E}{MS_T} = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-1)}.$$

The adjusted R^2_{adj} only increases when significant predictors are added to the model. Addition of unimportant predictors may lead to a decrease in the value of R^2_{adj}. Removal of unimportant predictors will generally lead to increase in the value of R^2_{adj}.

Example 3: Soft drink bottler (continued)

We want to test the significance of regression for the soft drink delivery time data. We have

$$SS_T = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2 = 18,310.6290 - 25 \cdot (22.384)^2 = 5784.5426.$$

We have previously found that $SS_E = 233.7317$. Hence

$$SS_R = SS_T - SS_E = 5550.8109.$$

To test $H_0 : \beta_1 = \beta_2 = 0$, we calculate the statistic

$$F_{cal} = \frac{MS_R}{MS_E} = \frac{5550.8109/2}{233.7317/(25-2-1)} = 261.2351.$$

Since $F_{0.001,2,22} = 9.612 < F_{cal}$, we conclude that delivery time is related to delivery volume and/or distance. The complete ANOVA table is shown Table 5 below.

Source of variation	d.o.f.	SS	MS	F	P-value
Regression	2	5550.8109	2775.4055	261.2351	4.44×10^{-16}
Residual	22	233.7317	10.6242		
Total	24	5784.5426			

Table 5: ANOVA table for the soft drink delivery time data.

However, this does not necessarily imply that the relationship found is an appropriate one for predicting delivery time as a function of volume and distance. Further tests of model adequacy are required.

Suppose we wish to assess the value of the predictor variable X_2 (distance) given that the predictor X_1 (cases) is in the model. The relevant hypotheses are

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0. \tag{2.3}$$

The diagonal element of $(X^T X)^{-1}$ corresponding to β_2 is $c_{22} = 0.00000123$, so the calculated value of the T statistic is

$$t_{cal} = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 c_{22}}} = \frac{0.01438}{\sqrt{10.6242 \cdot 0.00000123}} = 3.98.$$

Assuming $\alpha = 5\%$, we have $t_{crit} = t_{\alpha/2,n-2-1} = t_{0.025,22} = 2.074$. Thus $t_{cal} > t_{crit}$ and we reject the null hypothesis $H_0 : \beta_2 = 0$ with a 95% confidence level, and conclude that the predictor X_2 (distance) contributes significantly to the model given that X_1 (cases) is also in the model.

The hypotheses (2.3) can also be tested using the extra sum of squares method. The extra sum of squares due to β_2 is

$$SS_R^{extra} = SS_E^{reduced} - SS_E^{full},$$

We already know that $SS_E^{full} = 233.7317$. We only need to compute the reduced residual sum of squares, which we find to be

$$SS_R^{reduced} = \boldsymbol{y}^T (I - H^{red}) \boldsymbol{y} = 402.1349.$$

Therefore,

$$SS_R^{extra} = 402.1349 - 233.7317 = 168.4032$$

with $\nu_R^{extra} = r = 1$. This is the increase in the regression sum of squares that results from adding X_2 to a model already containing X_1. The calculated value of the test statistic is

$$F_{cal} = \frac{MS_R^{extra}}{MS_E^{full}} = \frac{168.4032/1}{10.6242} = 15.85.$$

Since $F_{0.05,1,22} = 4.30$, we reject the null hypothesis and conclude that distance (X_2) contributes significantly to the model. The complete ANOVA table is shown in Table 6 below.

Source of variation	d.o.f.	SS	MS	F	P-value
Residual Reduced	23	402.1349			
Residual Full	22	233.7317	10.6242		
Extra	1	168.4032	168.4032	15.85	6.3×10^{-4}

Table 6: ANOVA table for the extra sum of squares analysis.

3 Diagnostics and model building

3.1 Regressions diagnostics

When fitting a multiple regression model it is important to:

1. Determine whether the proposed regression model is a valid model (i.e., determine whether it provides an adequate fit to the data). This is achieved by examining regression plots involving standardised residuals and/or fitted values, and transforming response and/predictor variables, if necessary.
2. Determine which (if any) of the data points are leverage and/or outlier points, that is, points which have predictor values that have an unusually large effect on the estimated regression model and/or do not follow the pattern set by the bulk of the data, when one takes into account the given model.
3. Assess the effect of each predictor variable on the response variable, having adjusted for the effect of other predictor variables using added variable plots.
4. Assess the extent of collinearity among the predictor variables using variance inflation factors.
5. Examine whether the assumption of constant error variance is reasonable. If not, decide how to overcome this problem.
6. If the data are collected over time, examine whether the data are correlated over time. In such a case random errors are not independent and more general methods of estimation need to be used.[3]

3.1.1 Leverage points

Recall that data points which exercise considerable influence on the fitted model are called **leverage points**. Recall also that leverage measures the extent to which the fitted regression model is attracted by the given data point. We are therefore interested in the relationship between the fitted values $\hat{\boldsymbol{y}}$ and responses $\boldsymbol{y}$.

Since $\hat{\boldsymbol{y}} = H\boldsymbol{y}$, we have that

$$\hat{y}_i = (H\boldsymbol{y})_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j.$$

Since $\sum_{j=1}^{n} h_{ij} = 1$, the h_{ii} measures the extent to which the fitted value $\hat{y}_i$ is attracted by the given data point, y_i, i.e. if $h_{ii} \approx 1$, then $\hat{y}_i \approx y_i$. In other words,

A popular rule, which we shall adopt, is to classify the i-th point as a point of **high leverage** (i.e., a leverage point) in a multiple linear regression model with p predictors if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{p+1}{n}.$$

[3] See, for instance, Chapter 9 in (Sheather, 2009) or Chapter 5 in (Montgomery et. al., 2012).

3.1.2 Standardised residuals

The hat matrix H allows us to write the vector of residuals as[4]

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y}. \tag{3.1}$$

Then the least squares estimator $\boldsymbol{E}$ of $\boldsymbol{e}$ is

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I - H)\boldsymbol{Y}$$

Claim 3.1:

The sampling distribution of $\boldsymbol{E}$ is

$$\boldsymbol{E} \sim N_n(\boldsymbol{0}, \sigma^2(I - H)) \tag{3.2}$$

Proof. Since $\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$, we only need to find the mean and variance of $\boldsymbol{E}$:

$$\mathbb{E}(\boldsymbol{E}) = (I - H)\,\mathbb{E}(\boldsymbol{Y}) = (I - H)X\boldsymbol{\beta} = \boldsymbol{0}$$

since $HX = X$, and

$$\mathrm{Var}(\boldsymbol{E}) = (I - H)\,\mathrm{Var}(\boldsymbol{Y})(I - H)^T = (I - H)\sigma^2(I - H)^T = \sigma^2(I - H)$$

since $(I - H)(I - H)^T = (I - H)^2 = (I - H)$.

It follows from (3.2) that

$$\frac{E_i}{\sqrt{\sigma^2(1 - h_{ii})}} \sim N(0, 1).$$

Consequently, we introduce **standardised residuals**, r_i, by

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}.$$

We shall follow the common practice of labelling points as **outliers** in small to moderate size data sets if $|r_i| > 2$. In large data sets, we shall change this rule to $|r_i| > 4$. (Otherwise, many points will be flagged as potential outliers.) Recall, however, that a point can only be declared to be an outlier, only after we are convinced that the model under consideration is a valid one.

[4] The vector of residuals, $\boldsymbol{e}$, is orthogonal to the column space of X, that is

$$\boldsymbol{e}^T X = ((I - H)\boldsymbol{y})^T X = \boldsymbol{y}^T(I - H)X = \boldsymbol{y}^T(X - HX) = \boldsymbol{0}.$$

Since $\mathrm{rank}(X) = p + 1$, it follows from $\boldsymbol{e}^T X = \boldsymbol{0}$ that the residuals satisfy $p + 1$ linear equations, the normal equations. Indeed, transposing $\boldsymbol{y}^T(X - HX) = \boldsymbol{0}$ we obtain

$$\boldsymbol{0} = (X^T - X^T H^T)\boldsymbol{y} = X^T(I - H^T)\boldsymbol{y} = X^T(\boldsymbol{y} - H\boldsymbol{y}) = X^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}),$$

since $H^T = H$ and $H\boldsymbol{y} = \hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$. The resulting equations, $X^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, are the normal equations we found in Section 4.3 of Part I. As a consequence, there are only $\nu_E = n - p - 1$ linearly independent residuals; this number is called the residual degree of freedom, which we introduced in (1.8).

Plots of standardised residuals similar to the ones discussed in the analysis of the simple linear regression, are used to check the adequacy of a fitted multiple linear regression model. As before, standardised residuals should not show any patterns or trends when plotted against any variable. When a **valid model** has been fit, plots of standardised residuals should have the following features:

- A random scatter of points around the horizontal axis, since the mean function of the r_i is zero when a correct model has been fit;
- Constant variability as we look along the horizontal axis.

A pattern in a residual plot indicates that an incorrect model has been fit, but the pattern itself does not provide direct information on how the model is misspecified.

3.1.3 Cook's distance

The Cook's distance for a multiple linear regression model is defined by

$$D_i = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{(p+1)\hat{\sigma}^2}$$

where $\hat{\boldsymbol{y}}_{(i)}$ denotes the vector of fitted values based on the fit obtained when the ith case has been deleted from the fit. It can be shown that

$$D_i = \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}.$$

A recommended rough cut-off for noteworthy values of D_i for multiple linear regression, that we shall adopt, is $4/(n-p-1)$.

3.1.4 Added-variable plots*

We have discussed above that plots of (standardised) residuals versus predictor variables are useful in determining the effect for that predictor on the response variable. A limitation of such plots is that they may not completely show the correct or complete marginal effect of a predictor, given the other predictors in the model.

Consider a multiple linear regression model with p predictor variables,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.3}$$

Suppose we want to the introduction of an additional predictor variable Z to this model. In other words, we want to consider the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{z}\alpha + \boldsymbol{\varepsilon} \tag{3.4}$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^T$ is a vector of Z levels. We are interested in α, the regression coefficient measuring the effect of Z on $\boldsymbol{y}$, having adjusted for the effect of X on $\boldsymbol{y}$. The **added-variable plot** for predictor variable Z enables us to visually estimate α. The added-variable plot is obtained by plotting (on the vertical axis) the residuals from the model (3.3) against (on the horizontal axis) the residuals from the model

$$\boldsymbol{z} = X\boldsymbol{\delta} + \boldsymbol{\varepsilon}. \tag{3.5}$$

Notice that the residuals from model (3.3) give that part of $\boldsymbol{y}$ that is **not** predicted by X while the residuals from model (3.5) give that part of Z that is not predicted by X. Thus, the added-variable plot for predictor variable Z shows that part of $\boldsymbol{y}$ that is not predicted by X against that part of Z that is not predicted by X (i.e., the effects due to X are removed from both axes).

The vector of residuals from the model (3.3) is

$$\boldsymbol{e}_{\boldsymbol{y},X} = (I - H)\boldsymbol{y}.$$

Likewise, the vector of residuals from the model (3.5) is

$$\boldsymbol{e}_{\boldsymbol{z},X} = (I - H)\boldsymbol{z}.$$

Multiplying (3.4) by $(I - H)$ from the left hand side results in

$$(I - H)\boldsymbol{y} = (I - H)X\boldsymbol{\beta} + (I - H)\boldsymbol{z}\,\alpha + (I - H)\boldsymbol{\varepsilon}.$$

Since $HX = X$, this is equivalent to

$$\boldsymbol{e}_{\boldsymbol{y},X} = \boldsymbol{e}_{\boldsymbol{z},X}\,\alpha + \boldsymbol{e}^* \tag{3.6}$$

where $\boldsymbol{e}^* = (I - H)\boldsymbol{\varepsilon}$. Thus, α is the slope parameter in a regression of $\boldsymbol{e}_{\boldsymbol{y},X}$ (i.e., the residuals from the regression of Y on X) on $\boldsymbol{e}_{Z,X}$ (i.e., the residuals from the regression of Z on X). Let $\hat{\alpha}_{AVP}$ denote the least squares estimate of α in the model (3.6). It can be shown that $\hat{\alpha}_{AVP}$ is equal to $\hat{\alpha}$, the least squares estimate of α in the model (3.4). Furthermore, assuming that (3.4) is a valid model for the data, then the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\alpha}$. This plot will also enable the user to identify any data points which have undue influence on the least squares estimate of α.

Example 4: Italian restaurants in the New York City

The pricing data of Italians restaurants dinner menu in the NYC have been collected in order to produce a regression model to predict the price of dinner. The data includes the Price (in $) of a dinner, the average of customer rating (out of 30) on Food, Décor, and Service, and the location of the restaurant: East (1) or West (0) of the Fifth Av. (see Figure 3.1). A shortcoming of each plot in Figure 3.1 is that it looks at the effect of a given predictor on the Price ignoring the effects of the other predictors. This shortcoming is overcome by looking at added-variable plots (see Figure 3.2). The lack of statistical significance of the regression coefficient associated with the variable Service is clearly evident in the bottom left-hand plot of Figure 3.2. Thus, having adjusted for the effects of the other predictors, the variable Service adds little to the prediction of Price. Two points are identified in the top left-hand plot as having a large influence on the least squares estimate of the regression coefficient for Food. These points correspond to cases 117 and 168 and should be investigated. Case 117 corresponds to a restaurant called Veronica which has very low scores for Décor and Service, namely 6 and 14, respectively while achieving a relatively high food score of 21 given a price of $22. Case 168 corresponds to a restaurant called Gennaro, which has low scores for Décor and Service, namely 10 and 16, respectively while achieving a high food score of 24 for a relatively low price of $34.

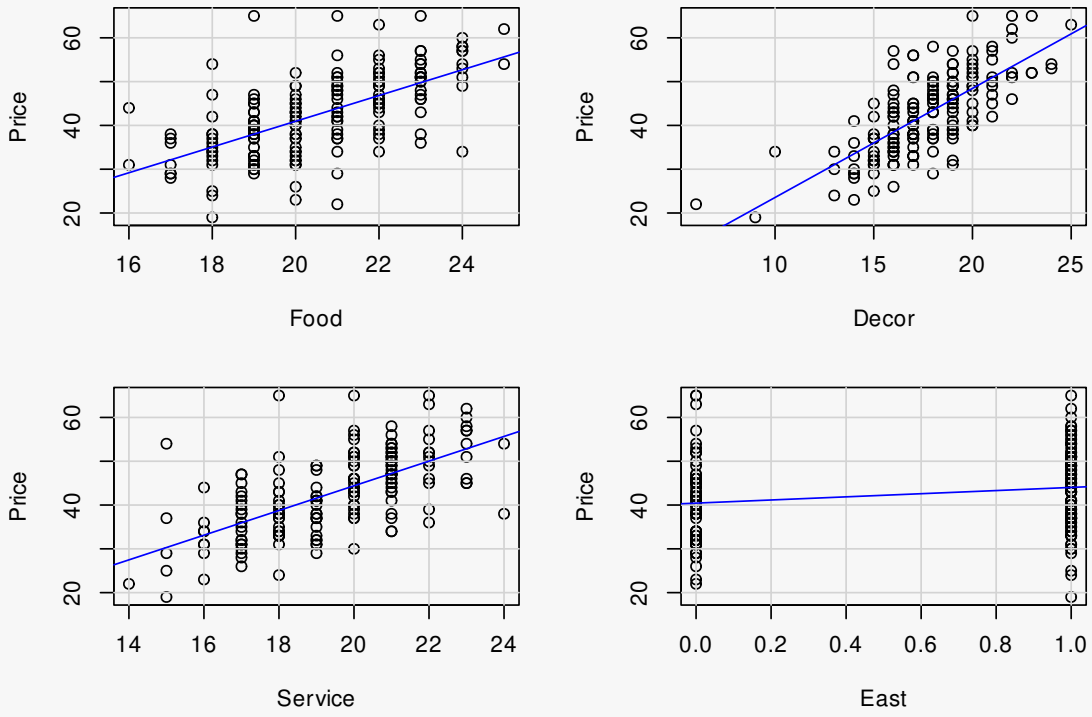

Figure 3.1: Scatter plots for the New York City restaurant data.

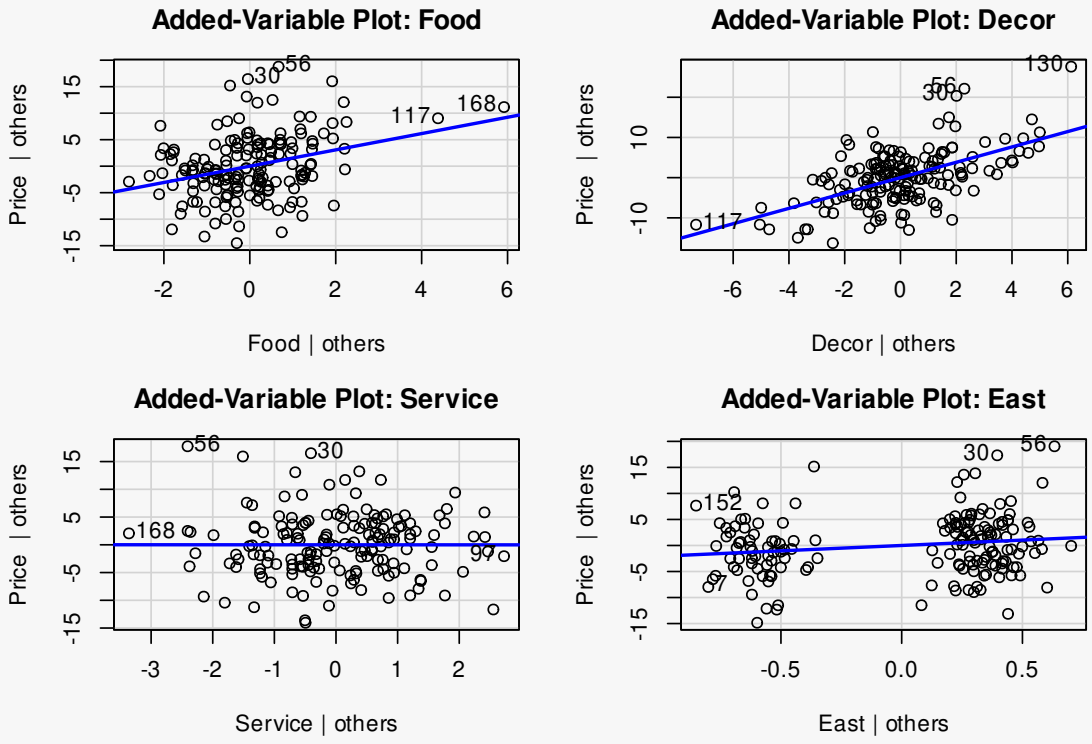

Figure 3.2: Added-variable plots for the New York City restaurant data.

3.1.5 Multicollinearity

In theory, one would like to have predictors in a multiple regression model that each have a different influence on the response and are independent from each other. In practice, the predictor variables are often correlated themselves. Multicollinearity is the prevalence of near-linear dependence among the regressors.

If one predictor were a linear combination of the other regressors, then the design matrix X would have linearly dependent columns, which would make the matrix X^TX singular (non-invertible). In practice, it would mean that the predictor that can be expressed through the other predictors **cannot** contribute any new information about the response. But, worse than that, the linear dependence of the predictors makes the estimated slopes in the regression model arbitrary.

For instance, consider a regression model in which somebody's height (in inches) is expressed as a function of arm-span (in inches). Suppose the true regression equation is

$$\mathbb{E}(Y) = 12 + 1.1X$$

Now, suppose further that when measuring the arm span, two people took independent measurements in inches (X_1) and in centimetres (X_2) of the same subjects and both variables have erroneously been included in the same linear regression model,

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

We know that $X_2 = 0.394X_1$ and thus we should have $\beta_1 + 0.394\beta_2 = 1.1$, in theory. But since this is a single equation with two unknowns, there are infinitely many possible solutions – some quite nonsensical. For instance, we could have $\beta_1 = -2.7$ and $\beta_2 = 9.645$. Of course, these slopes are not interpretable in the context of the original problem. The computer used to fit the data and to compute parameter estimates cannot distinguish between sensible and nonsensical estimates.

In general, multicollinearity is described by the correlation matrix for the p predictors in the model,

$$\mathrm{Cor}(X) = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{12} & 1 & r_{23} & \dots & r_{2p} \\ r_{13} & r_{23} & 1 & \dots & r_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1p} & r_{2p} & r_{3p} & \dots & 1 \end{bmatrix}$$

where

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{ij}}}, \qquad s_{ij} = \sum_{k=1}^{n}(x_{ki} - \bar{x}_{\bullet i})(x_{kj} - \bar{x}_{\bullet j}), \qquad \bar{x}_{\bullet i} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}\,.$$

The main diagonal elements of the inverse of the predictor correlation matrix are called the **variance inflation factors** (VIF). It can be shown that the VIF for the j-th regression coefficient can be written as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the other predictor variables. For example, if the model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ then R_1^2 is that of the model $x_{i1} = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$ and R_2^2 is that of the model $x_{i2} = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$.

Values of VIF considerably greater than 1 (usually large than 10) indicate multicollinearity problems. A few methods of dealing with multicollinearity include increasing the number of observations in a way designed to break up dependencies among predictor variables, combining the linearly dependent predictor variables into one variable, and eliminating variables from the model that are unimportant.

3.2 Factor variables

Multiple linear regression models also support the use of factor (categorical) variables. For example, gender may need to be included as a factor in a regression model. Factor variables are commonly encoded in terms of **indicator** variables. Indicator variables take on values of 0 or 1. For example, an indicator variable may be used with a value of 0 to indicate female and a value of 1 to indicate male,

$$X_1 = \begin{cases} 0 & \text{Female,} \\ 1 & \text{Male.} \end{cases}$$

In general, $n-1$ indicator variables are required to represent a factor variable with n levels. As an example, a qualitative factor having three levels, A, B, and C, may be represented as follows using two indicator variables:

$$\begin{aligned} \text{A}: &\quad X_1 = 1, \quad X_2 = 0, \\ \text{B}: &\quad X_1 = 0, \quad X_2 = 1, \\ \text{C}: &\quad X_1 = 0, \quad X_2 = 0. \end{aligned}$$

Here C is called the **reference state**. An alternative coding scheme for this example is to use a value of "−1" for all indicator variables when representing the last level of the factor:

$$\text{C}: \quad X_1 = -1, \quad X_2 = -1.$$

Indicator variables are also referred to as **dummy** or **binary** variables.

3.3 Variable selection

In this section we consider methods for choosing the "best" model from a class of multiple regression models using what are called **variable selection methods**. Interestingly, while there is little agreement on how to define the "best" model, there is general agreement in the statistics literature on the consequences of variable selection on subsequent inferential procedures, (i.e., tests and confidence intervals).

We begin by introducing some terminology. The **full model** is the following multiple regression model containing all p potential predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Throughout this chapter we shall assume that the full model is a valid regression model.

Variable selection methods aim to choose the subset of the predictors that is "best" in a given sense. In general, the more predictor variables included in a valid model the lower the bias of the predictions, but the higher the variance. Including too many predictors in a regression model is commonly called **over-fitting** while the opposite is called **under-fitting**. The two key aspects of variable selection methods are:

- Evaluating each potential subset of p predictor variables;
- Deciding on the collection of potential subsets.

We begin by considering the first aspect.

3.3.1 Coefficient of multiple determination

Recall that the coefficient of multiple determination, R^2, is defined as the proportion of the total sample variability in the response variable, Y, explained by the regression model, that is,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Adding irrelevant predictor variables to the regression equation often increases R^2. To compensate for this an adjusted coefficient of determination, R^2_{adj}, is introduced,

$$R^2_{adj} = 1 - \frac{MS_E}{MS_T} = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-1)}$$

where p is the number of predictors in the current model. It can be shown that adding predictor variables to the current model only leads to an increase in R^2_{adj} if the corresponding partial F-test statistic exceeds 1.

The usual practice is to choose the subset of the predictors that has the **highest** value of R^2_{adj}. It can be shown that this is equivalent to choosing the subset of the predictors with the **lowest** value of $\hat{\sigma}^2$.

However, choosing the subset of the predictors that has the highest value of R^2_{adj} tends towards over-fitting. For example, suppose that the maximum value is $R^2_{adj} = 0.692$ for a subset of $p = 10$ predictors, $R^2_{adj} = 0.691$ for a subset of $p = 9$ predictors and $R^2_{adj} = 0.541$ for a subset of $p = 8$ predictors. Even though R^2_{adj} increases when we go from 9 to 10 predictors there is very little improvement in fit and so the nine-predictor subset is generally preferred.

3.3.2 The Akaike's Information Criterion and Bayesian Analogue

The Akaike's Information Criterion (AIC) is a measure of "goodness" of a regression model based on maximising the log-likelihood of the model and penalising the complexity. The AIC is defined to be

$$\text{AIC} = -2\ln(L) + 2K$$

where $L = L(\hat{\beta}, \hat{\sigma}^2|Y)$ is the log-likelihood of the regression model and K is the number of estimated parameters in the model, which in our case is $K = p + 2$, since $\beta_0, \beta_1, \ldots, \beta_p$, and σ^2 are estimated in the fitted model. The measure of complexity is necessary since adding irrelevant predictor variables to the regression equation can increase the log-likelihood.

In case of the ordinary least squares the AIC is written as

$$\text{AIC} = n\ln\left(\frac{SS_E}{n}\right) + 2p.$$

This is the way R calculates the AIC. In general, the **smaller** is the AIC, the better is the model.

When the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, the AIC has a tendency for over-fitting since the penalty for model complexity is not strong enough. In such cases a corrected version of AIC should be used,

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(p+2)(p+3)}{n-p-1}.$$

The general rule is that AIC_C should be used instead of AIC unless $n/K > 40$. In particular, when n gets large, AIC_C converges to AIC, and thus for large samples there is (almost) no difference between AIC_C and AIC.

Another extension of AIC is the Bayesian information criterion defined by

$$\text{BIS} = -2\ln(L) + K\ln(n).$$

The BIC is similar to AIC except that the factor 2 in the penalty term is replaced by $\ln(n)$. When $n \geq 8$, $\ln(n) > 2$ and so the penalty term in BIC is greater than the penalty term in AIC. Thus, in these circumstances, BIC penalizes complex models more heavily than AIC, thus favouring simpler models than AIC.

For model selection purposes, there is no clear choice between AIC and BIC. BIC is asymptotically consistent as a selection criterion. What this means is that given a family of models, including the true model, the probability that BIC will select the correct model approaches one as the sample size $n \to \infty$. This is not the case for AIC, which tends to choose models which are too complex as $n \to \infty$. On the other hand, for finite samples, BIC often chooses models that are too simple, because of the heavy penalty on complexity. A popular data analysis strategy is to calculate R^2_{adj}, AIC, AIC_C, and BIC, and compare the models which minimize AIC, AIC_C, and BIC with the model that maximizes R^2_{adj}.

3.3.3 Finding the "best" model

There are two distinctly different approaches to choosing the potential subsets of predictor variables: **all possible subsets** and **stepwise methods**.

The first approach is based on considering all 2^p possible regression models and identifying the subset of the predictors that maximises a measure of fit or minimises an information criterion.

This second approach is based on examining just a sequential subset of the 2^p possible regression models. Arguably, the two most popular variations on this approach are **backward elimination** and **forward selection**.

Backward elimination starts with all potential predictor variables in the regression model. Then, at each step, it deletes the predictor variable such that the resulting model has the lowest value of an information criterion. (This amounts to deleting the predictor with the largest P-value each time.) This process is continued until all variables have been deleted from the model or the information criterion increases.

Forward selection starts with no potential predictor variables in the regression equation. Then, at each step, it adds the predictor such that the resulting model has the lowest value of an information criterion. (This amounts to adding the predictor with the smallest P-value each time.) This process is continued until all variables have been added to the model or the information criterion increases.

Backward elimination and forward selection consider at most

$$p + (p-1) + (p-2) + \ldots + 1 = p(p+1)/2$$

of the 2^p possible predictor subsets. Thus, backward elimination and forward selection do not necessarily find the model that minimizes the information criteria across all 2^p possible predictor subsets. In addition, there is no guarantee that backward elimination and forward selection will produce the same final model. However, in practice they produce the same model in many different situations.

3.4 Assessing the predictive ability of a regression model

Given that the model selection process changes the properties of the standard inferential procedures, a standard approach to assessing the predictive ability of different regression models is to evaluate their performance on a new data set (i.e., one not used in the development of the models). In practice, this is often achieved by randomly splitting the data into a **training data set** and a **test data set**, also called a **validation** data set. The training data set is used to develop a number of regression models, while the test data set is used to evaluate the performance of these models. Then the "best" model is the one having the smallest MS_E when evaluated on the test data set.

A popular strategy in assessing the predictive ability of a family of regression models is the k-fold cross-validation. The full set of data is randomly split into k subsamples. Then each model is fit k times, with each subsample being "left out" and the residual sum of squares is obtained for each model. Models are then compared by their mean residual sums of squares for the "left out" observations. Note that different splits of the data into the k folds will give different results.

These notes are a compilation of the following textbooks:

- D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th Edition, Wiley (2012). [Chapters 3–10]
- S. J. Sheather, *A Modern Approach to Regression with R*, Springer (2009). [Sections 5-7]