Chapter 15

Multiple Linear Regression

by Chris Wild and George Seber[1]
The University of Auckland

Modern computer packages have made the "doing" of multiple regression, in the sense of fitting specified models, comparatively trivial. The two aspects that require the development of sophistication are in understanding what the program output is telling you about the data and the process of building models to use for analysis. The latter, in particular, is more of an art than a science and challenges even experienced professional statisticians. The most important thing to develop in a first exposure to multiple regression is an understanding of the basic model and of the types of information that successful regression analyses provide. This is our primary focus. We also discuss model criticism and illustrate the processes involved in model building.

15.1 Using Several Explanatory Variables

15.1.1 Why Use Several Predictors?

In problem 10 of Review Exercise 12, we used regression to predict the asking prices of Renault-5 cars from their ages. There are, however, other things besides age that affect the value of a car and thus the asking price. We should be able to do a better job of predicting prices if, in addition to age, we can also use information on such things as distance traveled (mileage), extras fitted, number of previous owners, and general condition of the body and interior.

In simple regression in Chapter 12, we were using a single explanatory variable X to explain the behavior of our response variable Y. In doing this, we were really just dipping our toes in the water. In most real situations we can come up with quite a long list of variables that are likely to affect the behavior of the response we are interested in. It seems intuitively compelling that we should be able to make better predictions, or come to a better understanding of mechanisms, using a whole set of explanatory variables than we could from any one of them alone. Put more simply, the more we know the better we should be able to do.

In real investigations where regression is used, the most critical phase of the study is to come up with a list of characteristics that could plausibly be useful in predicting or explaining the response of interest. These lists come from combining the knowledge and intuitions of the investigators with the results of activities such as brainstorming with people who are knowledgable about the area under study and reading reports on previous studies. Sometimes it is comparatively obvious how we should measure a characteristic (e.g., age). Sometimes it is extremely difficult. For example, how can we capture the notion of "the general condition of the interior of a car" as some sort of measurement? All of this results in a list of variables that are both plausible explanatory variables for the response and are things that we will actually be able to measure in practice. The study then proceeds to gathering data on these variables and the response variable for a properly chosen set of individuals. There are a number of types of question we often try to answer using such studies.

- Which of the variables in our list really are important predictors and which seem to carry little information about the response?

We may want to know, for example, what are the most important risk factors for adult-onset diabetes? What are the best predictors of company takeover? What factors best predict the way in which a company will report its performance in its accounts?

- How can we use the data we have collected to build a prediction equation that combines information from all of our explanatory variables and gives predictions which are close to the observed responses?

More generally, we need ways of building a good *model* for the response we are likely to get at any given combination of values of the explanatory variables.

- How large an effect of does each of our explanatory variables have on the response?

To what extent does a high-fat diet increase the risk of diabetes? How big a drop in car price can we expect over 5 years? What is the effect of maternal smoking on the birth weights of babies once we have adjusted for the effects of other important predictors such as the sizes of the parents?

QUIZ ON SECTION 15.1.1

1. Why do we want to consider the effects of several explanatory variables rather than just a single variable?
2. How do we arrive at a list of candidate explanatory variables for inclusion in a study?
3. What types of question are studies that use regression generally trying to answer?

EXERCISES FOR SECTION 15.1.1

1. A data set we sometimes use in class relates the sizes of insurance claims for fire damage of houses to the distance of the house from the fire station. What other characteristics of the house do you expect would be predictive of the size of a claim?

2. Another data set we use was collected to evaluate the effects of an advertising campaign on attendance figures for a zoo. The advertisements were broadcast for periods of about a week separated by longer rest periods where the ads were not broadcast. Was the advertising effective? You might think that question would be easy to answer; we just look at attendances shortly after the ads have been run and see if they are larger than usual. Unfortunately, it is not this easy. Zoo attendances go up and down all of the time. There are, however, patterns in this variation. What other factors can you think of that might influence attendance at the zoo? (In looking for an advertising effect, we have to find ways to allow, or adjust, for these other variables.)

3. Think about how you would measure some of the characteristics that you have listed in questions 1. and 2.

15.1.2 Displaying the data

In practice, good investigations start with a question and a great deal of up-front work goes into such things as coming up with sets of useful explanatory variables, solving

measurement problems, and devising good experimental, sampling or other observational designs. For the rest of the Chapter, we will assume this job and the ensuing data collection has been done well. If it hasn't we are already in trouble. With data analysis, it really is a case of garbage in, garbage out. So from now on we are in the business of trying to make sense of the data that we have collected.

Example 15.1.1 *Can We Predict the Quality of Bordeaux Wine Vintages?*

The quality of Bordeaux wine vintages vary from year to year. Some years such as 1982, 1983 and 1985 produced outstanding vintages, while others like 1984 produced wine of lesser quality. The data in Table 15.2.1 were collected over a period of 27 years. The price is an index obtained by taking the wholesale price fetched by mature wines, adjusting it for inflation, and scaling it so that the value for 1961 is 100. We shall use this index as a measure of the quality of the wine. Interest centers on how climatic conditions affect the quality of the resulting wine. If possible, we would like to be able use climate information for a vintage to predict its quality before the wines go on to the market. Three climatic variables are given, namely, TEMP (average temperature in $^{\circ}C$ over the growing season), H.RAIN (rainfall over harvest time in mm), and W.RAIN (winter rainfall in mm).

TABLE 15.1.1 Bordeaux Wine Data

year	price	temp	h.rain	w.rain	year	price	temp	h.rain	w.rain
1952	37	17.1	160	600	1968	11	16.2	292	610
1953	63	16.7	80	690	1969	12	16.5	244	575
1955	45	17.1	130	502	1970	40	16.7	89	622
1957	22	16.1	110	420	1971	27	16.8	112	551
1958	18	16.4	187	582	1972	10	15.0	158	536
1959	66	17.5	187	485	1973	16	17.1	123	376
1960	14	16.4	290	763	1974	11	16.3	184	574
1961	100	17.3	38	830	1975	30	16.9	171	572
1962	33	16.3	52	697	1976	25	17.6	247	418
1963	17	15.7	155	608	1977	11	15.6	87	821
1964	31	17.3	96	402	1978	27	15.8	51	763
1965	11	15.4	267	602	1979	21	16.2	122	717
1966	47	16.5	86	819	1980	14	16.0	74	578
1967	19	16.2	118	714					

Source: Ashenfelter, O., Ashmore, D., and Lalonde, R. (1995)
"Bordeaux wine vintage quality and weather". *Chance*, **8** (4),xxx–xx.

How can we display this data to see what is going on? The most obvious approach is to look at a whole lot of scatter plots. Statistical packages give us a very convenient way of doing this. The result is most often called a *scatter plot matrix* (as in Minitab), but Splus and R call it a *pairs plot*. Figure 15.1 is a scatter plot matrix for the Bordeaux wine data. It simply consists of a page of scatter plots with every variable plotted against every other variable. To read the plots we need to understand how the axes are labeled. We should mentally drag the names in the central boxes horizontally or vertically onto the axes of the individual plots. For example, every plot in the second column of plots has PRICE as its x-axis. Every plot in the first row of plots has YEAR as its y-axis.

If we are interested in how well the other variables predict PRICE, we will want PRICE on the vertical axis and so will be looking at the second row of plots. All of the relationships with PRICE seem fairly weak, but PRICE appears to decrease over time (i.e., as YEAR increases), increase with with greater average temperatures over the growing season (TEMP), decrease as we get more harvest rain (H.RAIN) and perhaps increase slightly with more winter rain (W.RAIN). (Any relationships between the climate variables themselves are even weaker.) We note also that the 1961 vintage, which has PRICE=100, stands out as a possible outlier in many of the plots.

Unfortunately, although sets of scatter plots like these can give us useful ideas about what *might* be going on in the data, we cannot put too much trust in what we see in them for reasons that will soon become apparent.

Example 15.1.2 *Individual scatter plots can be misleading*

Consider the artificial data given in Table 15.1.2.

Table 15.1.2 Artificial data with two X-variables

Obs. :	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X_1 :	1.0	0.4	0.7	0.5	1.0	0.8	0.6	0.4	0.3	0.5	0.6	0.7	0.5	0.2
X_2 :	0.1	0.7	0.4	0.9	0.3	0	0.9	0.3	0.6	0.3	0.4	0.8	1.0	0.6
Y :	1.9	1.9	1.9	1.6	1.7	2.2	1.5	2.3	2.1	2.2	2.0	1.5	1.5	2.2

Figure 15.1.2(a) plots Y versus X_1 and Fig. 15.1.2(b) plots Y versus X_2. From these graphs it appears that X_1 is unrelated to Y and X_2 is only weakly related to Y **but** X_1, X_2 and Y happen to be related exactly! In fact, for every observation,

$$Y = 3 - X_1 - X_2$$

exactly. If you don't believe it check and see! The individual scatter plots give us no inkling that such a strong relationship exists. ■

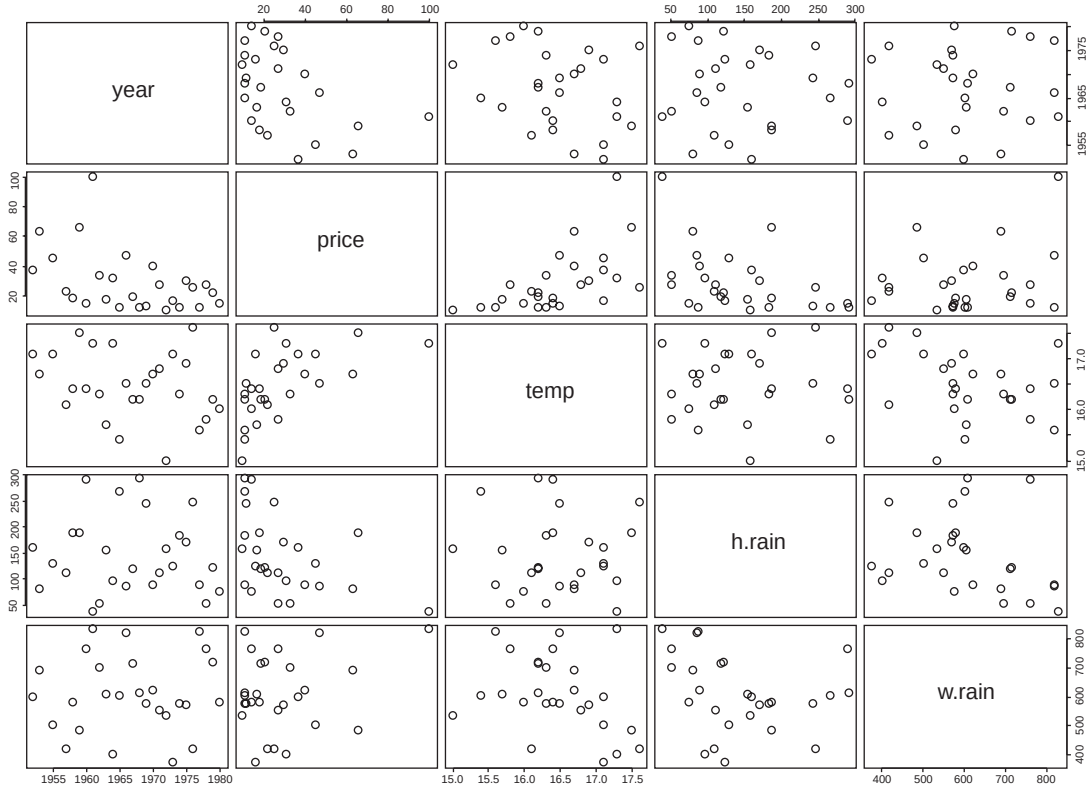


FIGURE 15.1.1 Scatter plot matrix for the Bordeaux wine data.

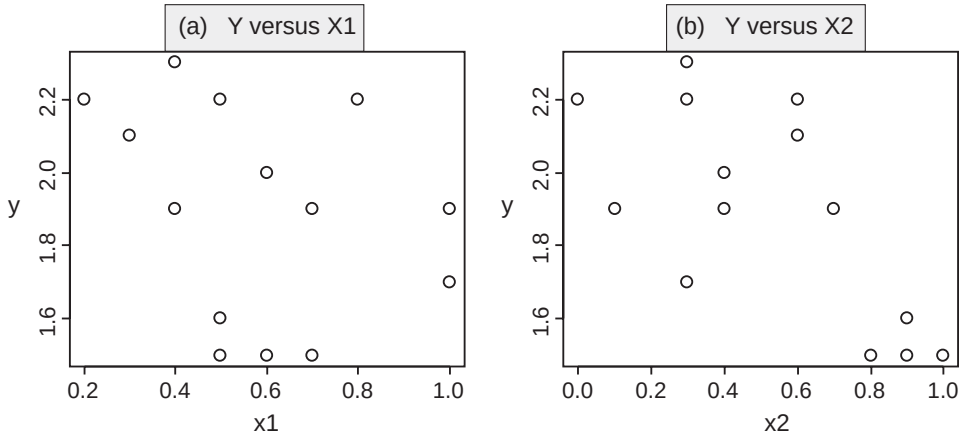


FIGURE 15.1.2 Plotting Y against each X-variable.

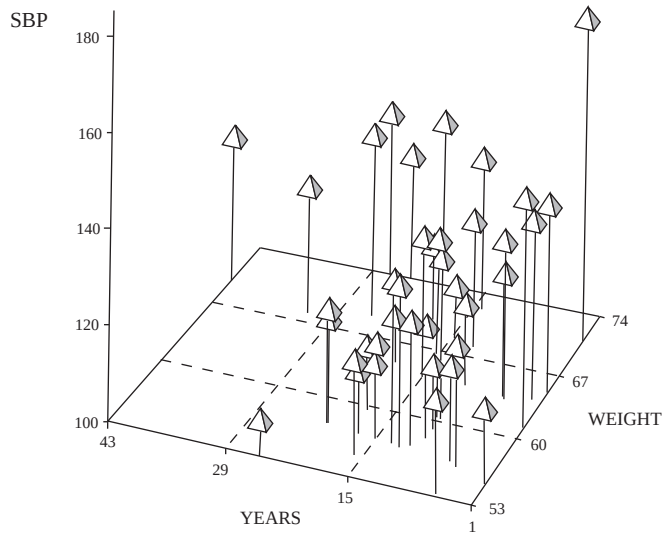


FIGURE 15.1.3 3-dimensional scatter plot for the Peru Indians data.

In the example above we saw a situation where the scatter plots failed to show that an important relationship exists. In addition, an individual scatter plot can show a relationship with an X-variable when none exists, and it can even get the direction of a relationship (positive or negative) wrong. You should be particularly suspicious of scatter plots of Y versus individual X-variables when the scatter plot matrix shows strong relationships between some of the X-variables. These same problems carry over to fitted regression models. The regression coefficient from a simple regression of Y on a single X-variable can be spuriously significant, spuriously insignificant, and even get its sign wrong.

Since considering the explanatory variables one at a time can be misleading, what we need are ways of displaying data, or fitting models, that consider all of the variables together. One option, when we have two X-variables, is a *3-dimensional scatter plot* as in Fig. 15.1.3.

Example 15.1.3 *Long Term Effect of Altitude on Blood Pressure*

Anthropologists have been interested in determining what are the long-term effects, if any, of altitude on human blood pressure. One suggested way of doing was to sample a population of Peruvian Indians who were native to the high Andes mountains but had since migrated to lower climes. The subjects were males over 21 years old who were born at high altitudes and had parents also born at high altitudes. Previous research

suggested that migration of this kind might cause higher blood pressure at first, but over time the blood pressure would decrease. The variables measured were as follows.

- AGE, the subject's age.
- YEARS, the number of years since migration.
- WEIGHT, the subject's weight in kilograms.
- HEIGHT, the subject's height in mm.
- BP, the subject's systolic blood pressure.

The data for 39 men are given in Table 15.3.1.

TABLE 15.3.1 Peruvian Indian Data

age	years	weight	height	BP	age	years	weight	height	BP
21	1	71	1629	170	38	18	59.5	1513	114
22	6	56.5	1569	120	38	11	61	1653	136
24	5	56	1561	125	38	11	57	1566	126
24	1	61	1619	148	39	21	57.5	1580	124
25	1	65	1566	140	39	24	74	1647	128
27	19	62	1639	106	39	14	72	1620	134
28	5	53	1494	120	41	25	62.5	1637	112
28	25	53	1568	108	41	32	68	1528	128
31	6	65	1540	124	41	5	63.4	1647	134
32	13	57	1530	134	42	12	68	1605	128
33	13	66.5	1622	116	43	25	69	1625	140
33	10	59.1	1486	114	43	26	73	1615	138
34	15	64	1578	130	43	10	64	1640	118
35	18	69.5	1645	118	44	19	65	1610	110
35	2	64	1648	138	44	18	71	1572	142
36	12	56.5	1521	134	45	10	60.2	1534	134
36	15	57	1547	120	47	1	55	1536	116
37	16	55	1505	120	50	43	70	1630	132
37	17	57	1473	114	54	40	87	1542	152
38	10	58	1538	124					

Source: Ryan, T. A., Jr., Joiner, B. L. and Ryan, B.F. (1976).
Minitab: Student Handbook Duxbury Press: North Scituate, Mass.

Figure 15.1.3 gives a perspective drawing of a 3-dimensional scatter plot. Blood pressure (BP) is plotted vertically above a horizontal plane whose edges are the YEARS axis and WEIGHT axis. We can get some idea of the relationship between BP and YEARS and WEIGHT from the plot. For example, the blood pressures seem larger at the back of the picture (large weights) than at the front, and there may be an increase from left to right (as YEARS gets *smaller*). But then again we may be being mislead by perspective effects in the picture and the single very large value at the back right. The ideal would be a 3-dimensional physical construction that we could walk around and view from different angles. Many statistical packages allow us to mimic this on a 2-dimensional screen by letting us rotate 3-dimensional scatter plots in arbitrary ways, a procedure which is usually called *spinning*. The movement involved in spinning gives us the feeling of depth and relative position. But we are still limited to three dimensions (i.e., Y plus two explanatory variables), and three dimensional plots still have the same sorts of inadequacies when there are many more variables to be considered as the two-dimensional scatter plots do.

There are other promising graphical tools, like *coplots*, but these will not be considered in this chapter. Instead we will go on to look at fitting models as a means of understanding multivariable regression data.

QUIZ ON SECTION 15.1.2

1. What is a scatter plot matrix?

2. Can we trust what we see in a Y versus X_j plot when there are several important explanatory variables for Y? Justify your answer.

3. Can we trust the results of a regression of Y on X_j when there are several important explanatory variables for Y?

4. What device do computer programs use to give the effect, on a 2-dimensional screen, of walking around a 3-dimensional scatter plot and looking at it from different angles.

5. Can we trust what we see in a 3-dimensional Y versus X_1 and X_2 plot when there are several more explanatory variables which are very predictive of Y?

EXERCISE FOR SECTION 15.1.2

Obtain a scatter plot matrix (pairs plot) for the Peruvian Indians data and discuss what you see in the individual scatter plots.

15.2 The multiple linear regression model

15.2.1 Introducing the model

The simplest possible model for use with a single explanatory variable X is the simple linear model discussed in Section 12.4. Here, the ith observation has its Y-value related to its X-value through

$$Y_i = \beta_0 + \beta_1 x_i + U_i,$$

where U_i is a random error with underlying true mean $\mu_U = 0$ and standard deviation $\sigma_U = \sigma$. When making formal statistical inferences from data, we further assumed that the errors were independent and Normally distributed. As we saw in Section 12.4, this model generates patterns of constant scatter about a linear trend. This very simple model fits many data sets very well, but other data sets force us to reach for more complicated models that allow for a greater variety of trend shapes. In Example 12.4.4, we saw how including a quadratic, or x^2 term, in our model to form

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + U_i,$$

lets us allow for a gently curved trend. Furthermore, the quadratic model enables us to test whether the trend really is curved by testing $H_0 : \beta_2 = 0$, as we did in Example 12.4.4.

Now that we have more than one explanatory variable, we have to expand our notation somewhat. Let $X_1, X_2, \ldots, X_k$ denote our explanatory variables (AGE, YEARS, WEIGHT and HEIGHT in Example 15.1.3). Let x_{ij} be the observed value of X_j for the ith individual. For example, in Example 15.1.3, the second X-variable is YEARS and the 3rd person has $x_{3\,2} = 5$.

When we have two explanatory variables, X_1 and X_2, the simplest model we can construct is of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + U_i.$$

This model generates a pattern in which the Y-values are randomly scattered about a plane in three dimensional space as depicted in Fig. 15.2.1. Think of the X_1 and X_2 axes as being laid out on a table top. The Y-axis is vertical, perpendicular to the table top. We have tried, in Fig.15.2.1, to portray the regression plane as a sheet of glass. The black balls are elevated above the glass and the little stick joining each ball vertically to the plane shows how far the observation is elevated above the plane. The lengths of these sticks are the sizes of the errors. We also see through the glass plane to balls which fall below the plane and appear as grey. They too have vertical sticks giving the sizes

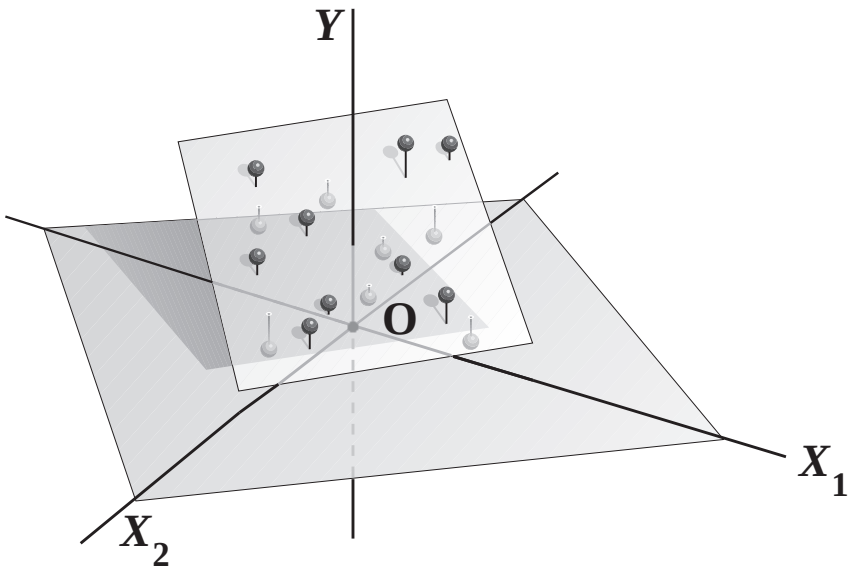


FIGURE 15.2.1 Data points in 3 dimensions scattered about a plane.

of their errors. A model of the sort shown in Fig. 15.2.1 is obviously a very reasonable candidate model for the data we see in Fig. 15.1.3. The above two X-variable model also extends in a very natural way to an arbitrary number of variables, k, as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + U_i,$$

although we can no longer represent the model diagrammatically in 3-dimensional space.

Extensions

The trend patterns in our data may be considerably more complicated than the plane depicted in Fig. 15.2.1. With a single X-variable, adding squared and even higher-level polynomial terms allows us to cater for increasingly complicated trend shapes. Similarly, with two or more X-variables, adding squared terms in one or more of the variables (e.g. YEAR^2) and cross product terms (e.g. YEAR×WEIGHT) allows us to fit surfaces to our data which respectively bend and twist. These can all be catered for within the multiple linear regression model. We can set $X_1 = \text{AGE}$, $X_2 = \text{AGE}^2$, $X_3 = \text{WEIGHT}$, $X_4 = \text{AGE} \times \text{WEIGHT}$, $X_5 = \text{YEAR}$, and so on. Interpretation of the coefficients of variables which have polynomial and/or crossproduct terms in a model is, however, beyond the scope of this chapter.

Regression Analysis

The regression equation is
BP = 90.2 - 0.161 age - 0.538 years + 1.50 weight - 0.0276 height

Predictor	Coef	SE Coef	T	P
Constant	90.18	52.48	1.72	0.095
age	-0.1613	0.2795	-0.58	0.568
years	-0.5380	0.2195	-2.45	0.020
weight	1.4987	0.3173	4.72	0.000
height	-0.02761	0.03674	-0.75	0.457

[Note: Prior to release 13, Minitab labeled the standard error column "StDev".]

FIGURE 15.2.2 Minitab regression output for the Peruvian Indian data.

15.2.2 Model Fitting

Any set of values we choose for the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, will produce a predicted value for the Y-variable for the ith individual, namely

$$\widehat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik},$$

which can be compared with the value that we actually observe, namely y_i. We want to choose coefficient values that give predictions which are close, on average, to observed data. As in Section 12.3.1 we choose the set of coefficient values which gives the smallest value of the sum of squared prediction errors, $\sum(y_i - \widehat{y}_i)^2$. In other words we are again fitting our model to the data using least squares. The coefficients that define the least squares surface are denoted $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ and are called *the least squares estimates*. The (least squares) *fitted regression surface*, which we will use in an attempt to describe the trend in the data, is given by

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_k x_k.$$

Example 15.2.1 *A Least Squares Fit to the Peruvian Indian Data*

Figure 15.2.2 gives the results of fitting a multiple regression model to the Peruvian Indian data from Example 15.1.3 using BP as the response variable and AGE, YEARS, WEIGHT and HEIGHT as explanatory variables. We see, for example, that the least squares estimate of the intercept is $\widehat{\beta}_0 = 90.18$, the least squares estimate of the coefficient of AGE is $\widehat{\beta}_{AGE} = -0.1613$, and so on. The resulting equation for predicting blood pressures is

$$\widehat{BP} = 90.18 - 0.1613 \times \text{AGE} - 0.5380 \times \text{YEARS} + 1.4987 \times \text{WEIGHT} - 0.0276 \times \text{HEIGHT}.$$

The first individual in the file ($i = 1$) has AGE=21, YEARS=1, WEIGHT=71, and HEIGHT=1629 giving a predicted (or *fitted*) value for the blood pressure of 147.7 (our $\widehat{y}_1$) in comparison with the observed blood pressure for this individual of 170 (our y_1). The *residual*, or estimated error is $\widehat{u}_1 = y_1 - \widehat{y}_1 = 170 - 147.7 \approx 22$. In spatial terms (cf. Fig.15.2.1), this observation falls 22 units above the fitted least squares regression surface.

As the chapter proceeds, we will learn to answer questions such as:

- Does our model actually fit the data?
- What can we infer about the population or process that gave rise to this data?
- Which variables really do affect Y?

Our first task is to better understand the model itself.

15.2.3 Interpreting Coefficients

We use these models in situations where the data appears to be randomly scattered about the fitted regression surface. We are thus thinking in terms of a Y-value on the regression surface at $X_1 = x_1, \ldots, X_k = x_k$ being the average of the Y-values we would expect to see if we were able take repeated observations at this set of X's. For example, the first Peruvian Indian in Table 15.1.3 has AGE=21, YEARS=1, WEIGHT=71, HEIGHT=1629 and $\widehat{y} = \widehat{BP} \approx 148$ for this set of X-values. What this means to us is that if we collected data on a much larger sample from this population of Indians and obtained quite a few Indians with AGE=21, YEARS=1, WEIGHT=71, HEIGHT=1629, then we estimate that the average of those Indians' blood pressures would about 148.

The least squares coefficients $\widehat{\beta}_j$ give us estimates of how the average value of Y changes as the values of the X-variables change. We think of the least squares estimate of the coefficient of X_j, $\widehat{\beta}_j$, being an estimate of some unknown true coefficient β_j. But because the least squares coefficients, and resulting $\widehat{y}$ values, are just estimates from data, they are therefore subject to sampling variation (just as in Section 12.4.1). Therefore, in due course, we will have to reach once more for our trusty tools of confidence intervals and significance tests to cope with the uncertainty this sampling variation causes. In the meantime, however, we want to concentrate on learning to interpret the coefficients, without the added complication of considering sampling variation.

Our model is that the response for an individual with $X_1 = x_1, X_2 = x_2, \ldots X_k = x_k$ is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + U,$$

where $\mathrm{E}(U) = 0$. We have:

$$\text{1st individual}: \quad E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \ldots + \beta_k x_k.$$

Consider now another individual whose X-values are identical to those of the first except that X_j has increased by w units to $X_j = x_j + w$.

$$\text{2nd individual}: \quad E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j (x_j + w) + \ldots + \beta_k x_k.$$

Subtracting the first expression from the second, we see that the difference between the two values of $\mathrm{E}(Y)$ is $\beta_j w$. In particular, for a 1-unit increase in X_j ($w = 1$), $\mathrm{E}(Y)$ changes by β_j. When $X_1 = 0, \ldots, X_k = 0$, we find that $\mathrm{E}(Y) = \beta_0$.

Interpreting the Coefficients

$\boldsymbol{\beta_0}$, the **constant, or intercept term**,
is the expected Y-value when every $X_j = 0$.

Provided the values of all X-variables except X_j remain unchanged,

- $\boldsymbol{\beta_j}$, the **coefficient of** X_j is *the change* in $\mathrm{E}(Y)$ associated *with each 1-unit increase in* X_j.
- If X_j increases by w-units, then $\mathrm{E}(Y)$ changes by $\beta_j w$ units.

[Changing the value of one variable with all the rest remaining unchanged does not always make sense.]

Example 15.2.1 cont. *Interpreting the Coefficients*

The estimated coefficients are given in Fig. 15.2.2. The estimated intercept is 90.2. But what does an intercept relate to here? It relates to the average BP value when all of the X-variables are set to zero, i.e., and an Indian with AGE = 0, YEARS = 0, WEIGHT = 0 and HEIGHT = 0. This is not a meaningful combination of values so we shall ignore the intercept.

The other coefficients can be interpreted as the effects of a 1-unit change in the X-variable under consideration, all other variables being held constant. Thus, to interpret the coefficient of AGE (-0.16), we think in terms of Indians who are identical on all variables except their age and estimate that average blood pressure will go down ($\widehat{\beta}_{AGE}$ is negative) by 0.16 units for every additional year of age. Similarly, since $\widehat{\beta}_{YEARS} = -0.53$, the model estimates that, with all other factors remaining unchanged, blood pressure goes down by 0.53 units, on average, for every additional year since migration down from

the mountains. Similarly it estimates that, with all other variables remaining constant, average blood pressure goes up by 1.48 units for every additional kilogram in weight and down by 0.028 units for every additional mm of height. For every additional 1 cm = 10×1 mm in height, average blood pressure is estimated to go down by $0.028 \times 10 = 0.28$ units ($= \widehat{\beta}_{HEIGHT} \times 10$).

15.2.4 Using centered and standardized variables

By *centering a variable*, say AGE, we mean that we construct a new variable AGE.C, say, defined by

$$\text{AGE.C} = \text{AGE} - \text{mean(AGE)}.$$

If you replace AGE by AGE.C in our model, you will find that only the intercept changes. (Try it and see!) All other coefficients (and their standard errors P-values, etc.) remain unchanged. When we replace all variables in the model in Fig. 15.2.2 by their centered equivalents, we find that the coefficients of AGE.C, YEARS.C, WEIGHT.C and HEIGHT.C are identical to those for their uncentered equivalents in Fig. 15.2.2, but that the intercept has changed to 127.4. Not only has the intercept changed, it is now a meaningful quantity. When we set our new X-variables to zero, i.e., set AGE.C=0, YEARS.C=0, WEIGHT.C=0 and HEIGHT.C=0, we are equivalently setting the original variables equal to their average values. The intercept now relates to average BP at the average value of AGE, the average value of YEARS, and so on.

> If we center all of our X-variables before putting them into the model, the intercept relates to the average value of Y when each of the original X-variables is set equal to its average value.

Example 15.2.2 *A Least Squares Fit to the Bordeaux Wine Data*

Figure 15.2.3 gives the results of fitting a multiple regression model to the Bordeaux wine data from Example 15.1.1 using PRICE as the response variable and TEMP (average temperature in $^\circ C$ over the growing season), H.RAIN (rainfall over harvest time in mm), and W.RAIN (winter rainfall in mm) as the explanatory variables. Here, again, the intercept is not meaningful. It relates to average PRICE in a year when all the X-variables are set to zero, i.e., TEMP=0, H.RAIN=0 and W.RAIN=0, and we never see such years. Interpreting the coefficients, with all other variables being held constant, the model estimates that average PRICE goes up by 22 units for every additional $^\circ C$ in

```
lm(formula = price ~ temp + h.rain + w.rain)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -365.45179   77.63849  -4.707 9.66e-05 ***
temp          22.50086    4.28502   5.251 2.51e-05 ***
h.rain        -0.09296    0.03746  -2.481   0.0208 *
w.rain         0.06103    0.02247   2.717   0.0123 *
```

FIGURE 15.2.3 Some regression output from R for the Bordeaux wine data.

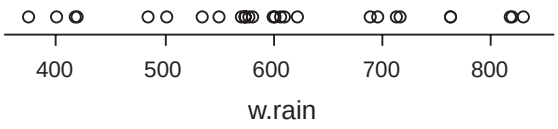


FIGURE 15.2.4 Dot plot of W.RAIN.

average growing temperature, goes down by 0.09 units for every additional 1 mm in rain over harvest time, and goes up by 0.06 units for every additional 1 mm of rain over the winter.

It looks, on the face of it, as though the rain variables are having almost no effect on PRICE at all because their coefficients are so small. But the coefficient of W.RAIN is telling us about the effect of a 1-unit increase in W.RAIN which means only 1 mm more rain over a whole winter. It's not surprising the effect is small! Figure 15.2.4 plots W.RAIN and we see a range of more than 400 mm in winter rain values.

We could decide that 100 mm is a more useful increase in W.RAIN to be looking at than 1 mm. Our coefficient tell us that for every additional 100 mm winter rain, average PRICE is estimated to go up by $\widehat{\beta}_{W.RAIN} \times 100 = 6$ units. ■

From the baove discussion we see that we cannot judge the size of the effect of a variable from the size of its regression coefficient in isolation. The size of $\widehat{\beta}_j$ depends on the units that X_j is measured in. If you convert the units of W.RAIN from mm to cm, so that the numerical values are divided by 10, then the coefficient becomes 10 times larger. If you convert to lots of 100 mm, the coefficient becomes 100 times larger. When interpreting the effect of a variable, we want to be quantifying the effect of a change of a sensible size. We could do this by looking at the extent of the variation in that variable, as we did with W.RAIN in Example 15.2.2. Next we will look at a more automated approach.

The effect of a standardized X-variable

By the expression *standardizing* a variable, say W.RAIN, we mean that we construct a new variable W.RAIN.S, say, defined by

$$\text{W.RAIN.S} = \frac{\text{W.RAIN} - \text{mean(W.RAIN)}}{\text{sd(W.RAIN).}}$$

The new variable W.RAIN.S has mean 0 and standard deviation 1. A 1-unit increase in W.RAIN.S corresponds to a 1 standard-deviation increase in W.RAIN. Thus, if we use W.RAIN.S in the model, the coefficient of W.RAIN.S tells us about the change in Y associated with every 1-sd increase in W.RAIN. Also we will always get the same coefficient regardless of the units we use to measure W.RAIN.
[Note that because standardized variables have mean zero, they are also centered.]

Effect of a standardized variable

The coefficient of a standardized X-variable is the change in average Y-value associated with a *one standard deviation* (1-sd) increase in the original X-variable.

Fig. 15.2.3(a) gives the regression output for the Bordeaux wine data with all explanatory variables having been standardized. The coefficients are now much more comparable in size. The fitted model estimates that a 1-sd increase in TEMP increases PRICE, on average, by about 15 units, whereas a 1-sd increase in H.RAIN decreases average PRICE by about 8 units, and a 1-sd increase in W.RAIN increases average PRICE by about 7 units. The effect of a 1-sd increase in temperature seems about twice as large as the effects of comparable increases in the rainfall variables; these latter two effects being very similar in size (but opposite in direction).

Notes:

(a) Bordeau wine data: Using standardised explanatory variables to model PRICE

```
Coefficients:
             Estimate  Std.Error t-value p-value
(Intercept)   28.815      2.565  11.233 8.17e-11
temp.S        14.673      2.794   5.251 2.51e-05
h.rain.S      -6.792      2.737  -2.481   0.0208
w.rain.S       7.875      2.899   2.717   0.0123
```

(b) Peru Indians data: Using standardised explanatory variables to model BP

```
             Estimate  Std.Error t-value  p-value
(Intercept)  127.410      1.669  76.337  < 2e-16
age.S         -1.240      2.148  -0.577   0.5676
years.S       -5.431      2.216  -2.451   0.0195
weight.S      10.639      2.252   4.724 3.91e-05
height.S      -1.455      1.936  -0.752   0.4575
```

FIGURE 15.2.3 Regression output with standardized X-variables (and unstandardized Y).

1. Standardizing X-variables puts different variables measured on different scales on to the same footing and thus makes their effects much more comparable.

2. Standardizing is a good idea during data exploration but, when reporting results to nonstatistical people, we recommend reporting the effects of changes in terms of the original variables as these are easier to understand.

3. If all of the X-variables are standardized, as they are in Fig. 15.2.3, then the magnitudes (ignoring signs) of the regression coefficients are in the same rank order (largest to smallest) as the t-values, and in the reverse order to the P-values.

4. If all of the X-variables are standardized, the intercept is the average Y-value when all X's are set equal to their average values.

Standardized variables and partial correlation

If we standardize both Y and X_j, then the resulting coefficient can then be interpreted as estimating the *partial correlation* between Y and X_j after adjusting for the effects of the other variables. We can loosely think of this in terms of the correlation between Y-values and X_j-values for individuals who are identical on all other variables.

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.0000	0.1224	4.12e-14	1.0000
temp.S	0.7002	0.1333	5.251	2.51e-05
h.rain.S	-0.3241	0.1306	-2.481	0.0208
w.rain.S	0.3758	0.1383	2.717	0.0123

FIGURE 15.2.4 Fitting the regression model to the Bordeaux wine data with Y = PRICE and the three X-variables all standardized.

We refitted the model we have been using for the Bordeaux wine data after standardizing all of the variables, including PRICE. The results are given in Fig. 15.2.4. We get the following estimated partial correlations: 0.70 between PRICE and TEMP; -0.32 between PRICE and H.RAIN; and 0.38 between PRICE and W.RAIN.

QUIZ ON SECTION 15.2

1. To what quantity does the intercept (or constant term) in a multiple linear regression relate?
2. What quantity does the coefficient of X_j tell us about?
3. What is meant by "centering a variable"?
4. What quantity does the intercept relate to if all of the fitted X-variables are centered?
5. What is meant by "standardizing a variable"?
6. If X_{jS} is the standardized version of X_j and we regress Y on a set of X-variables including X_{jS}, what quantity does the coefficient of X_{jS} tell us about?
7. What quantity does the intercept relate to if all of the fitted X-variables are standardized?
8. Let X_{jS} be the standardized version of X_j and Y_S be the standardized version of Y. If we regress Y_S on a set of X-variables which includes X_{jS} rather than X_j, what quantity does the coefficient of X_{jS} tell us about?

EXERCISES FOR SECTION 15.2

1. This exercise is built around the Bordeaux wine data and assumes that the model fits the data adequately.

   (a) Fit a linear model to the Bordeaux wine data with response variable PRICE and explanatory variables TEMP, H.RAIN and W.RAIN and check that you obtain the results in Fig. 15.2.3 (expect small computer-package-specific differences in labeling).

(b) Replace TEMP in your regression by the centered version of TEMP. What about the output changes? What stays the same?

(c) Replace all of the explanatory variables in your regression in (a) by their centered versions. Compare this output with that from both (a) and (b). What about the output changes? What stays the same? Also compare the intercept with the mean value of PRICE. What do you notice?

(d) Replace TEMP in your regression in (a) by the standardized version of TEMP. What about the output changes? What stays the same?

(e) Replace all of the explanatory variables in your regression in (a) by their standardized versions. Compare this output with that from both (a) and (d). What about the output changes? What stays the same? Also compare the intercept with the intercept in (c) and the mean value of PRICE. What do you notice?

(f) Replace both PRICE and TEMP in your regression in (a) by their standardized versions but leave all other variables unchanged. What about the output changes? What stays the same? Compare also with the output in Fig. 15.2.4.

2. This exercise is built around the Peruvian Indians data and assumes that the model fits the data adequately.

(a) You may be interested in whether the effects you have seen in problem 1. are special to this data set. They are not, but you may wish check and to see using this second data set.

(b) Obtain an estimate of the effect on average blood pressure of a 500 g (or 0.5 kg) increase in weight (all other variables being held constant).

(c) Give an estimate of the effect on average blood pressure of a 1 standard-deviation increase in weight (all other variables being held constant).

(d) Give an estimate of the partial correlation between blood pressure and weight (all other variables being held constant).

3. Read about the Fuel Usage data in Appendix A15.1.

(a) Fit a linear regression model to these data with LITERS as the response variable and DIST and MO.JAN as the explanatory variables, and interpret the coefficients.

(b) Estimate the change in liters consumed, on average, associated with an 100-kilometers increase in the number of kilometers traveled since the last fill-up.

(c) Find the partial correlation between LITERS and DIST and try to explain what it means.

(d) Repeat (a) for a model in which KM.LTR is the response and DIST and MO.JAN are the explanatory variables.

15.3 Inference

15.3.1 Inferences About Coefficients

There is a great deal of the basic regression output that we have not yet discussed. We refer to the standard error, t-statistic and P-value columns. What are all these things there to tell us about? Here, we are just revisiting Section 12.4. The only thing that is really new is that our output panel has more lines. Here is a quick recap.

Our least squares estimates of the coefficients are not the truth; they are just estimates from data. We know that, as such, they are subject to sampling variation, and this makes us unsure about what the true values really are. The *standard error* for the estimated coefficient of X_j is an estimate of its variability in repeated sampling. We also want to employ our usual tools of confidence intervals for the unknown true values of the coefficients and significance tests for values we hypothesize for these unknown true values.

Under the standard multiple linear-regression model used by the programs,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + U_i, \quad \text{where } U_i \sim \text{Normal}(\mu_{U_i} = 0, \sigma_{U_i} = \sigma),$$

for $i = 1, \ldots, n$ independently. Equivalently, given all of the values $x_{i\,j}$ the individuals have for their X-variables,

$$Y_i \sim \text{Normal}(\mu_{Y_i}, \sigma_{Y_i} = \sigma), \quad \text{where } \mu_{Y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik},$$

for $i = 1, \ldots, n$ independently. When these assumptions are true, it can be shown that the sampling distribution of the t-statistics $t_0 = \dfrac{\widehat{\beta}_j - \beta_j}{\text{se}(\widehat{\beta}_j)}$ is Student($df = n - k - 1$). The resulting **confidence intervals** for the unknown true value of β_j are of the form

$$\text{estimate} \ \pm \ t\,\text{std errors} \ = \ \widehat{\beta}_j \pm t\,\text{se}(\widehat{\beta}_{\text{j}})$$

and the **t-test statistic** for testing a hypothesized value for the unknown true β_j, i.e. testing $H_0 : \beta_j = c$ is

$$t_0 \;=\; \frac{\text{estimate} - \text{hypothesized value}}{\text{std error}} \;=\; \frac{\widehat{\beta}_j - c}{\text{se}(\widehat{\beta}_\text{j})}.$$

The "no effect" hypothesis

One of the big questions we said we needed to be able to answer was, "Which of our X variables really do affect Y?" The t-statistic and the P-value on the line of standard output corresponding to variable X_j relate to a test of $H_0 : \beta_j = 0$, and thus $t_0 =$ estimate/standard error. The printed P-value is typically a 2-sided P-value.

When $\beta_j = 0$, the model for Y no longer depends in any way on X_j. Thus, when our linear model fits, testing for $\beta_j = 0$ is equivalent to testing a hypothesis that says X_j **has no effect** *on Y over and above that already captured by the other variables.* Or equivalently, we could interpret it as X_j has *no predictive value* for Y over and above that already captured by the other variables. The reason why we have stressed this idea of "over and above that already captured in the other variables" is that if we take out one of the other X-variables out of the model or put a new one in, the relationships between these other variables and X_j can cause the whole picture to change. X_j may then become, or may then cease to be, a significant predictor of Y. There are other ways that people use to express this "over and above" idea. We may talk, for instance, of X_j having no effect on Y once we have *adjusted for*, or *allowed for*, or *controlled for* the effects of the other variables in the model, or *partialled out the effects* of the other variables in the model.

When the P-value corresponding to X_j is small (significant), we do indeed have evidence that X_j is related to Y and we can see the direction of the relationship by looking at the the sign of $\widehat{\beta}_j$. If the P-value is large (nonsignificant), it is plausible that there is no real relationship because the distance between the estimate $\widehat{\beta}_j$ and zero can be explained simply in terms of sampling variation.[2] If the P-value is large, we will use the expression, "we have no evidence of a relationship. . .".[3]

Example 15.3.1 *Interpreting the Test Results for the Peruvian Indian Data*

[2]If you received a dollar from every writer of a research report that mistakenly claimed that their nonsignificant P-value demonstrated that no relationship existed, you would soon become quite rich.

[3]Of course the "we have no evidence ..." is referring just to the data set we are currently looking at. There may well be evidence in other data sets.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	90.1774	52.4751	1.7185	0.0948	-16.4647	196.8195
age	-0.1613	0.2795	-0.5772	0.5676	-0.7293	0.4066
years	-0.5380	0.2195	-2.4511	0.0195	-0.9841	-0.0919
weight	1.4987	0.3173	4.7238	0.0000	0.8539	2.1434
height	-0.0276	0.0367	-0.7516	0.4575	-0.1023	0.0470

FIGURE 15.3.1 Excel regression output for the Peruvian Indian data.

Figure 15.3.1 gives Excel output for the multiple regression model for BP in the Peruvian Indian data using AGE, YEARS, WEIGHT and HEIGHT as explanatory variables.

The AGE and HEIGHT effects are nonsignificant. Having adjusted for the effects of the other variables in the model, there is no evidence of an AGE effect or a HEIGHT effect on blood pressure. There is very strong evidence of a weight effect (P-value is less than 10^{-4} and the effect is positive (BP goes up with WEIGHT). There is also evidence of a YEARS effect (P-value ≈ 0.02). The direction of the effect is negative, i.e., average BP goes down as the number of YEARS the person has been down from the mountains increases.[4]

To estimate the sizes of these effects, taking sampling variation into account, we need confidence intervals for the true values of the coefficients. From Fig. 15.3.1, we see that a 95% confidence interval for the true value of β_{WEIGHT} is given by approximately [0.85, 2.14]. The coefficient tells tells about the change in average value of $Y =$ BP associated with a 1-unit increase in WEIGHT (all other variables remaining fixed). We see that with 95% confidence, when WEIGHT goes up by 1 kg, average blood pressure goes up by somewhere between 0.85 and 2.14 units.

What is new here? Previously, when we interpreted the estimated coefficient, we quoted a point estimate for the effect of WEIGHT on BP without giving any idea of uncertainty. Now we are quoting a range of values to allow for the uncertainty due to sampling variation. All of our statements to follow are made with 95% confidence, and all concern the effect of a change in the stated variable when all other variables are held fixed. We will not keep repeating these qualifications.

The confidence interval for β_{YEARS} is given approximately by $[-0.98, -0.09]$. Thus, with 95% confidence, average blood pressure goes down by somewhere between 0.09 and

[4]Since this was the effect the investigators were expecting to find (see the introductory paragraph of Example 15.1.3), we would be justified in quoting a 1-sided P-value of $0.02/2 = 0.01$.

0.98 units for every additional year since migrating down from the mountains. For every additional 10-year period[5], average BP goes down by somewhere between 0.9 and 9.8 units. The confidence intervals also show that whilst we have no evidence of an AGE or HEIGHT effect, we cannot claim that they have no effect. The confidence interval for β_{AGE}, for example, tells us that a 1-year increase in AGE is associated with anywhere between an 0.73 unit decrease and an 0.41 unit increase in average BP. For 10 years, this ranges between a 7.3 unit decrease and a 4.1 unit increase. We are simply rather ignorant about any effect that AGE might have.

Excel's output, given in Fig. 15.3.1, is unusual in that it supplies these intervals automatically. Most packages make you work a little to get them.

Confidence Intervals

Our confidence interval for the true value of a coefficient is of the form

$$\text{estimate} \ \pm \ t \, \text{std errors}.$$

For linear model with k X-variables and an intercept, $\quad df = n - k - 1.$

[Note that when we have a single X-variable, $k = 1$ and $df = n - 2$ as in Section 12.4.2.]

Our data set contains data on $n = 39$ men, and we have fitted $k = 4$ variables so the degrees of freedom are $df = 39 - 4 - 1 = 34$. When $df = 34$, the t-multiplier for a 95% confidence interval is given by $t = 2.032$. Thus the 95% confidence interval for β_{AGE} is given by $-0.1613 \pm 2.032 \times 0.2795$. This gives the interval for the AGE coefficient given in Fig. 15.3.1 to within rounding error. You may wish to check that you can get the other values.

15.3.2 The Test for No Regression

Figure 15.3.2(a) shows a little more of the standard regression output produced by Minitab. Most packages produce something very like this. The equivalent portion of the output from R and Splus, given in Fig. 15.3.2(b), is a little more abbreviated. In the Minitab output we get an "analysis of variance" panel that looks only a little different from what we saw in 1-way analysis of variance. Something is obviously being tested. What is it?

[5]The values of YEARS in Table 15.1.3 vary between 1 and 43 years.

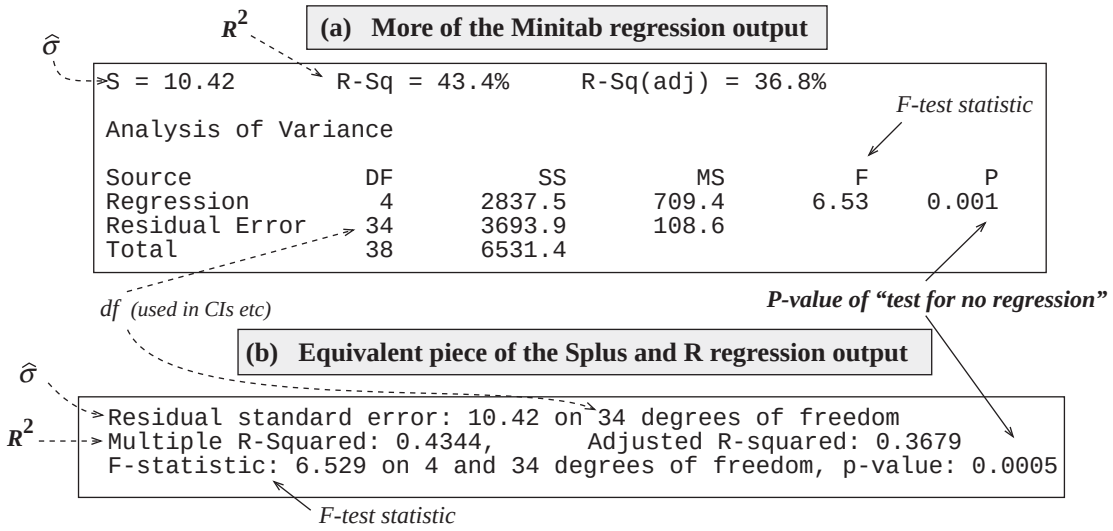


FIGURE 15.3.2 Regression output containing the test for no regression.

The test results relate to a null hypothesis that says that the true values of all of the regression coefficients except the intercept are zero:

$$\text{No regression:} \quad H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0.$$

What this means is that none of the X-variables have any predictive value for Y. The test result is almost always significant (as it is in Fig. 15.3.2) as it is extremely rare for an investigator's intuition to be so bad as to come up with a completely dud set of predictors. The test is more often called the *test for regression*. We show evidence for the existence of some regression by being able to reject the null hypothesis of no regression.

Several other aspects of the output have been labeled in Fig. 15.3.2, in particular, see $\widehat{\sigma}$ which is the estimate of the standard deviation of the errors,[6] and *df* $(= n - k - 1)$ for use in confidence interval construction. Another quantity that has been labeled is R^2, which is the subject of the next subsection.

QUIZ ON SECTION 15.3

1. Why do we need to test hypotheses and calculate confidence intervals for regression coefficients?
2. What do the standard errors of the coefficients tell us about?

[6]Note that $\widehat{\sigma}$ is the square root of the MS, or mean square, entry for "Residual Error" (e.g. in Fig. 15.3.2(a), $\sqrt{108.6} = 10.42$).

3. In what two forms can the standard multiple linear regression model be written?
4. What is the formula for the t-test statistic printed on the regression output?
5. What hypothesis does it test? Mathematically? Verbally or intuitively?
6. If you took one of the other X-variables out of the model and refitted, would you expect the P-value for the coefficient of X_j to be different or remain the same?
7. If you added another X-variable to the model and refitted, would you expect the P-value for the coefficient of X_j to be different or remain the same?
8. How do we interpret a significant P-value here?
9. How do we interpret a nonsignificant P-value?
10. Write down the formula used for the degrees of freedom to be used for t-tests and intervals, and define the terms in that formula.

The following questions concern Section 15.3.2

11. What hypothesis is tested by what is commonly known as the "test for regression"? Try to give both a technical answer and an intuitive verbal answer.
12. What test statistic is used?
13. How could we interpret (i) a significant "test for regression" result? (ii) a non-significant result?

EXERCISES FOR SECTION 15.3

1. Apply the ideas learned in Section 15.3.1 to the output from the model we have been using for the Bordeaux wine data. (Assume that this model fits the data.)

   (a) What do we learn from the P-values and the signs of the coefficients?

   (b) Obtain and interpret confidence intervals for the coefficients. For the rain variables, interpret both in terms of a 1 mm change and a 100 mm change.

2. Read about the Fuel Usage data in Appendix A15.1.

   (a) Fit a linear regression model to these data with LITERS as the response variable and DIST and MO.JAN as the explanatory variables. What do we learn from the P-values and the signs of the coefficients?

   (b) Obtain and interpret confidence intervals for the coefficients. For the variable DIST, interpret both in terms of a 1 km change and a 100 km change.

   (c) Repeat (a) and (b) for a model in which KM.LTR is the response and DIST and MO.JAN are the explanatory variables.

3. Read about the Coursework Data in Appendix A15.2.

   (a) Fit a linear regression model to these data with EXAM as the response variable and TEST1, TEST2 and ASSIGN as the explanatory variables.

   (b) What do we learn from the P-values and the signs of the coefficients?

   (c) Now refit with EXAM as the response variable and only ASSIGN as explanatory? What do we learn from the P-values and the signs of the coefficients?

   (d) Can you give any plausible explanation for the difference in the conclusions reached between (b) and (c)? (Read carefully about the nature of the assignment marks).

15.4 Does Our Model Fit the Data?

Unfortunately, the phrase "the fit of a model" is often used in two quite different senses. One sense concerns *the predictive or explanatory ability* of the model ("Are the predicted values very close to the observed y-values?"). The other sense concerns *whether the assumptions* of the multiple linear regression model are satisfied.

It is possible for a model to give good point predictions[7] but clearly violate model assumptions. A linear fit to the computer timings data in Fig. 12.4.9 (where the trend is slightly curved) gives quite good point predictions within the range of the data. At 0.935 the correlation coefficient is very high. A linear model fitted to the rain gauge data in Fig. 12.4.8 gives even better point predictions. This time the trend is linear but there is a strong departure from the constant spread assumption (see Fig. 12.4.2(d)). What goes wrong in these situations is that statistical tests, confidence intervals and prediction intervals cease to have the statistical behavior we expect of them. For example, calculated "95% confidence intervals" no longer have 95% coverage properties.

It is also possible – in fact, quite common – for a model to have low predictive ability and yet satisfy all of the model assumptions. In other words, the assumptions about the form of the trend and the statistical behavior of the errors (as seen through the residuals) are satisfied, but because σ is large, there is a great deal of scatter. Consequently, the trend does not predict individual data points very well.

Section 15.4.1 discusses R^2, a measure of the predictive power of a model. Section 15.4.2 discusses the checking of model assumptions.

[7] By *point predictions* we are distinguishing between single number predictions and prediction intervals (Section 12.4.3).

15.4.1 Measuring Predictive or Explanatory Power

As discussed in Section 12.5.2, when we have a single X-variable, the correlation coefficient r of the data points (x_i, y_i) is a measure of how closely the set of data points came to falling on a straight line. When r is $+1$ or -1, the points fall exactly on the line without any scatter (Fig. 12.5.2). Thus, y-values can be predicted without error from their x-values and the equation of the line. When $r = 0$ there is no linear relationship and a line will not provide useful predictions of y-values from x-values. It can be shown mathematically that r is also the correlation coefficient of the pairs $(\widehat{y}_i, y_i)$. This alternative description is very useful as it automatically generalizes to models where there is more than one explanatory variable.

Suppose now that we have a linear model with k explanatory variables. We define R to be the correlation coefficient of the pairs $(\widehat{y}_i, y_i)$. As already indicated, $R = r$ when $k = 1$. Unless we are using a pathologically ill-fitting model, R will be positive. Now R^2, generally known as the *coefficient of determination*, routinely appears on regression output (see both Figs 15.3.2(a) and (b)). When multiplied by 100% it tells us about the percentage of the variation in the observed y-values that can explained, via the model, by the variation in the values of the X-variables.

Explanatory Power of the Model

- R is the correlation between the observed and predicted y-values.

The coefficient of Determination, R^2

- is the *proportion* of the variation in the y-values that has been explained by the fitted model;
- is usually re-expressed in terms of "percent variation explained".

For the model fitted to the Peruvian Indians data in Fig. 15.3.2, $R^2 = 0.434$ (or 43.4%). This tells us that about 44% of the variation in blood-pressure values can be explained, using the fitted model, in terms of the variation present in the AGE, YEARS, WEIGHT and HEIGHT measurements. The balance of the variation, namely 56% remains, as yet, unexplained.

If we are dealing with an ill-fitting model, we may be able to increase the percent variation in BP explained by AGE, YEARS, WEIGHT and HEIGHT a little by building a model that fits better (see Section 15.4.3 for ideas on model building). Otherwise, we conclude that the variables we have are not capable of telling the whole story. The

only way to do better job of explaining the behavior of Y = BP will be to find some new (additional) X-variables, or combinations of the original variables, that are good predictors.

Models cannot provide very precise predictions unless R^2 is close to 100% (at least above 90%). Otherwise there is too much unexplained variation (scatter) left in the data, and our prediction intervals have to become quite wide to accommodate it.

The coefficient of Determination

For the remainder of this subsection, we shall delve more deeply into the background and interpretation of R^2. The material is rather technical and can be omitted on a first reading.

The theoretical table corresponding to Fig. 15.3.2 is given by

TABLE 15.1.1 Analysis-of-Variance Table for Regression

Source	SS	df	Mean SS	F -statistic	P-value
Regression	$\sum(\widehat{y}_i - \overline{y})^2$	3	s^2_{reg}	$f_0 = \frac{s^2_{reg}}{s^2}$	$\text{pr}(F \geq f_0)$
Residual	$\sum(y_i - \widehat{y}_i)^2$	$n-k$	s^2		
Total	$\sum(y_i - \overline{y})^2$	$n-1$			

Here, $y_i - \overline{y}$ is the (signed) distance between y_i and the center of the y's. We can rewrite $y_i - \overline{y}$ (by adding and subtracting $\widehat{y}_i$) as

$$\begin{array}{rccccc} & y_i - \overline{y} & = & (\widehat{y}_i - \overline{y}) & + & (y_i - \widehat{y}_i) \\ \text{or} & \text{deviation of observation} & = & \text{deviation of trend value} & + & \text{residual .} \end{array}$$

It can be shown algebraically that this additive property goes through for the sums of squares as well. Thus,

$$\begin{array}{rccccc} & \sum(y_i - \overline{y})^2 & = & \sum(\widehat{y}_i - \overline{y})^2 & + & \sum(y_i - \widehat{y}_i)^2 \\ \text{or} & \text{Total.SS} & = & \text{Regression.SS} & + & \text{Residual.SS,} \end{array}$$

where SS is an abbreviation for *Sum of Squares*. Now,

- Total.SS is a measure of the variation in the y_i's.

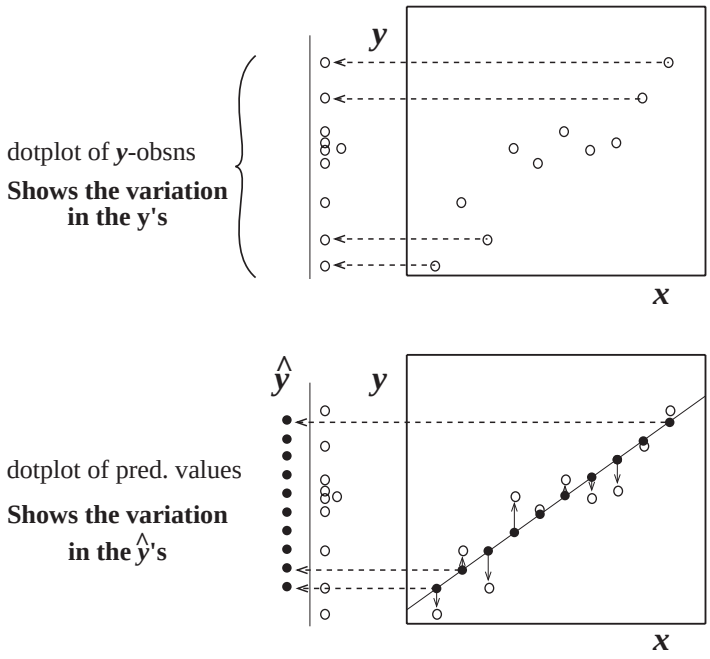


FIGURE 15.4.1 Variation in the y-values and in the $\widehat{y}$-values.

- Regression.SS is the corresponding measure of the variation in the predicted values.[8]
- Residual.SS is the corresponding measure of the variation in the residuals.

The residual sum of squares represents the unexplained component of the variation. We try to fit models that capture all of the trend. Having done that, we think of what is left over (the scatter) as random noise. We can find no patterns in it. We cannot "explain" it. The regression sum of squares represents the component of the variation that is "explained by the model" in the sense depicted in Fig 15.4.1 (with a single X-variable). Regression.SS $= \sum(\widehat{y}_i - \overline{y})^2$ is the variation in the $\widehat{y}$'s. It is the amount of variation we would still have even if the model fitted perfectly, i.e., even if there was no residual scatter. It results from the variation in the X-variable(s). It can be shown mathematically that

$$R^2 = \frac{\text{Regression.SS}}{\text{Total.SS}}.$$

Thus R^2 is the proportion of the total variation in the observed y's that has been "explained by the model" – it is the *proportion of variation explained*.

[8] It transpires that $\overline{y}$ is also the mean of the $\widehat{y}_i$s and that the residuals have mean 0.

We know that R is the correlation between the observed and predicted y-values. This tells us that $0 \leq R^2 \leq 1$, if R^2 is close to 0 then there is little correlation between the observed and predicted values, and if R^2 is close to 1 they are highly correlated. Furthermore,

$$R^2 = \frac{\text{Regression.SS}}{\text{Total.SS}} = \frac{\text{Total.SS} - \text{Residual.SS}}{\text{Total.SS}} = 1 - \frac{\text{Residual.SS}}{\text{Total.SS}}.$$

Thus, an R^2 of close to 1 corresponds to a set of residuals, or estimated errors, that are very small relative to the variation in the y-values. This indicates that the model will be good for prediction. An R^2 close to 0 corresponds to a set of residuals that are almost as variable as the y-values. In other words, the scatter about the fitted surface is almost as large as the original variation in y-values. When R^2 is close to 0 we are essentially doing no better than predicting every one of the y_i's using their sample mean $\overline{y}$. The X-variables are not giving useful predictive information.

Adjusted R-squared

Right next to R^2 on the output panels in Fig. 15.3.2, you will find `R-sq (adj)` in (a) or `Adjusted R-squared` in (b). The value is a little smaller than R^2. Why the adjustment? It can be shown mathematically that, when you add another X-variable to a model, R^2 will become smaller. This happens even if the additional variable has essentially no predictive value. The adjustment allows for the number of variables fitted in the model, and it enables us to make fair comparisons of the predictive abilities of models with differing numbers of X-variables.

EXERCISE showing that r is also the correlation coefficient between y and $\widehat{y}$. (***PUT IN AN EXERCISE from Chance Encounters?***)

15.4.2 Model Checking

We will now consider evaluating whether our model fits the data in the sense that the assumptions appear to be satisfied. The main tools for doing this, namely residual plots, have already been described and used in Section 12.4.4, and we suggest that you now take the time to re-read Section 12.4.4.

The standard multiple linear regression model is given by,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + U_i, \quad \text{where } U_i \sim \text{Normal}(\mu_{U_i} = 0, \sigma_{U_i} = \sigma),$$

for $i = 1, \ldots, n$ independently. All of the discussion in Section 12.4.4 concerning the implicit assumptions about the error behavior and the way we use the residuals $\widehat{u}_i = y_i - \widehat{y}_i$ to check the assumptions about the straight line model also apply to the multiple linear regression model exactly as before. The only real change from Section 12.4.4 is that our residuals have been captured from a model with a more complex trend and that we now have more than one X-variable to plot our residuals against.

The basic principle behind model building and residual plotting for these and also for more complicated models is as follows. We try to build a model which captures and models all of the trends, or patterns, in our data. The residuals represent what is left over. If we have successfully modeled all patterns, the residuals should look random and contain no patterns. If we find pronounced patterns in our residuals, this means that there are patterns in the data that our modeling has not as-yet captured. We will therefore need to enlarge or change our model to capture these unmodeled patterns.

The standard residual plots we look at are

- Residuals versus fitted (or predicted) values ($\widehat{u}$ versus $\widehat{y}$)
- Residuals versus each of the X-variables in turn

Any appreciable trends seen in these plots tell us that our trend model is not working. A curve in a plot of residuals versus X_j suggests we try adding an X_j^2 term. We can try to address problems in the residuals versus fitted values plot that do not respond to this by checking whether other available X-variables should have been included in the model, adding cross-product terms (e.g., $X_2 \times X_4$), or by using transformations of Y and/or the X-variables (see Chapter 16). Fans in the residuals are addressed by using a transformation of Y (Chapter 16) or by weighted least squares (which is outside the scope of the present account).

If the data has been collected in time order, we also plot

- Residuals versus time (or observation number if time is not available)
- Residuals versus *lagged* (or previous) residuals to look for serial correlation
  [i.e., plot the points $(\widehat{u}_1, \widehat{u}_2)$, $(\widehat{u}_2, \widehat{u}_3)$, …, $(\widehat{u}_{n-1}, \widehat{u}_n)$]

The standard test for serial correlation is called the Durbin-Watson test.

It is only when we have fixed all of the other problems in the model that we investigate Normality of the residuals with:

- Normal probability plots (also known as Normal Q-Q plots) of the residuals

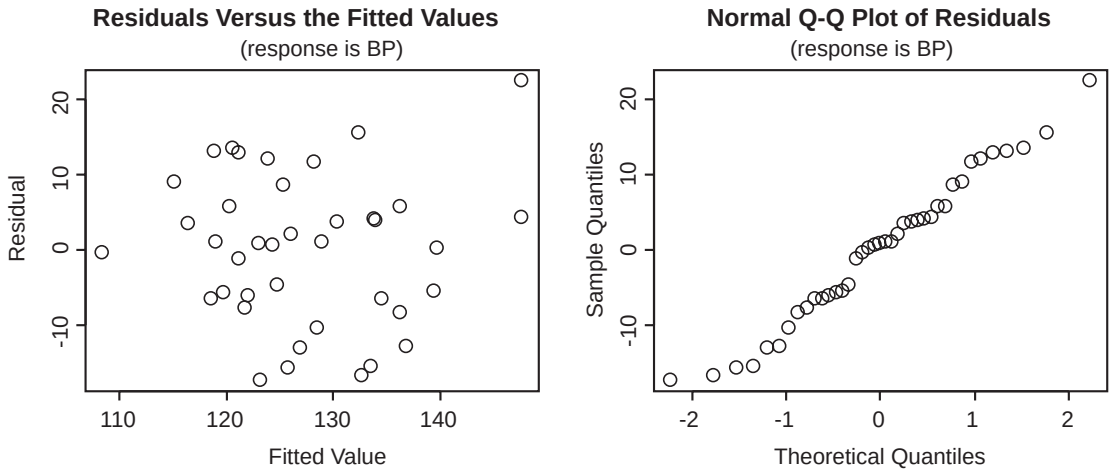


FIGURE 15.4.2 Two residual plots for the Bordeaux wine data.

and formal tests such as the Wilk-Shapiro W-test and the Weisberg-Bingham test.

Example 15.4.1 *Model Checking for the Peruvians Indian Data*

We performed residual plots to check the model fitted to the Peruvian Indians data in which BP is the response variable and AGE, YEARS, WEIGHT and HEIGHT were explanatory variables. [This model produced the output in Figs 15.3.1 and 15.3.2(a).] We found no serious problems with any of the residual plots. They all looked sufficiently like "horizontal patternless bands." We thought we were perhaps seeing a slight hint of a fan in the RESIDUALS versus HEIGHT plot, but repeatedly plotting normally distributed random numbers against HEIGHT convinced us that this could easily have just been random. Figure 15.4.2 gives the residuals versus fitted or predicted) values plot and the Normal Q-Q plot. The latter looks sufficiently linear and formal tests, using both the W-test and the Weisberg-Bingham test, detected no evidence against the Normality assumption.

*** SLIGHT PROBLEM JUST NOTICED, A BIG AGE*YEARS INTERACTION***

Example 15.4.2 *Model Checking for the Bordeaux Wine Data*

We have performed residual plots to check the model fitted to the Bordeaux data in which PRICE is the response variable and TEMP, H.RAIN and W.RAIN were explanatory variables. [This model produced the output in Fig. 15.2.3.] Figure 15.4.3(a) shows the residuals versus fitted values plot. We see a curved trend which is very pronounced,

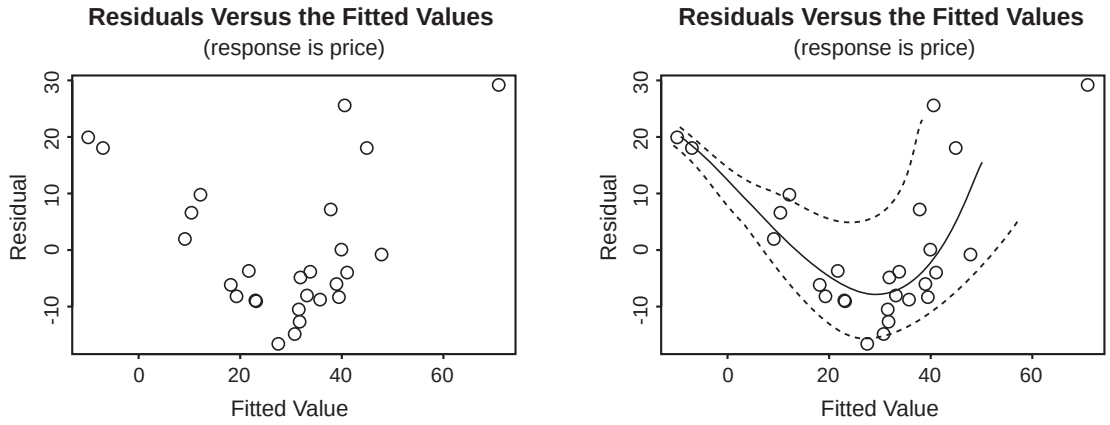


FIGURE 15.4.3 Two more residual plots for the Bordeaux wine data.

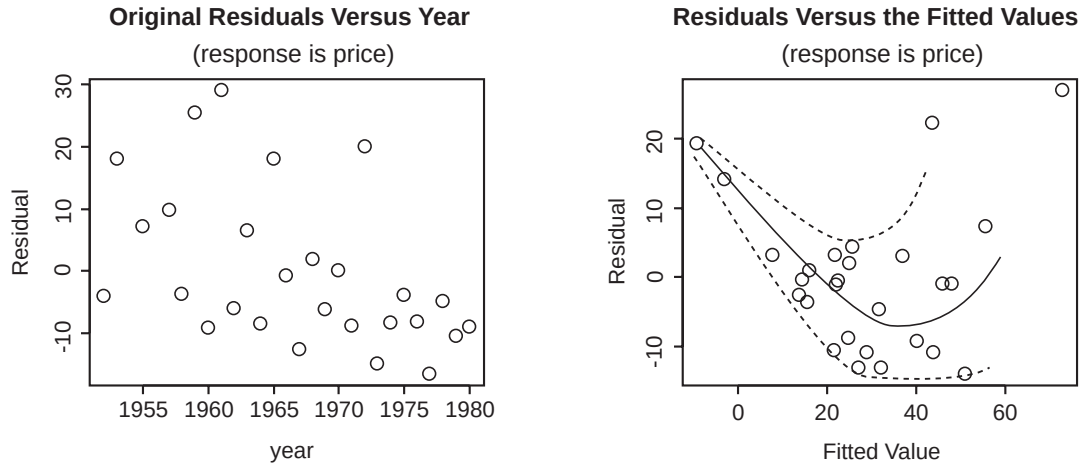
and also a fan. These have been emphasized in Fig. 15.4.3(b). The model is clearly deficient. We would put no trust in any inferences drawn from this model. We will build a model that does fit this data as an extended Case Study in Section 15.4.3.

15.4.3 A CASE STUDY: Building a model for the wine data

This subsection illustrates the process of model building. Parts of the Case Study will involve greater levels of sophistication than could reasonably be expected from a beginner in multiple regression. We begin where Example 15.4.2 left off.

When we tried plotting residuals versus the other X-variables, we saw a suggestion of a curve in the RESIDUALS versus TEMP and nothing suspicious in the RESIDUALS versus H.RAIN and RESIDUALS versus W.RAIN plots. The curve in the TEMP plot was too slight to account for the pronounced curve in Fig. 15.4.3. Nevertheless, we tried adding a TEMP^2 term to the model and refitting. The residuals versus fitted values plot was virtually unchanged from Fig. 15.4.3.

There is one more variable, YEAR, that we have not used. You may recall that we saw a trend in the PRICE versus YEAR scatter plot in Fig. 15.1.1. We then plotted the residuals from the original fit versus YEAR (see Fig. 15.4.4(a)). There is a clear trend downwards. This suggests that we should add YEAR to the model. Once YEAR had been added to the model, the trend in the residuals versus YEAR plot (not shown) has been removed, but the residuals versus fitted values plot (shown in Fig. 15.4.4(b)) still shows a strong curved trend and the fan is now even more pronounced than it was in



(a) Before YEAR has been added to the model **(b)** After YEAR has been added to the model

FIGURE 15.4.4 Adding YEAR to the regression model.

Fig. 15.4.3(b). We have included the only other variable we have, and we have tried to allow for curved trends by adding squared terms. None of it helped appreciably. Adding cross-product terms (e.g., TEMP*W.RAIN) doesn't help either. Where do we go from here?

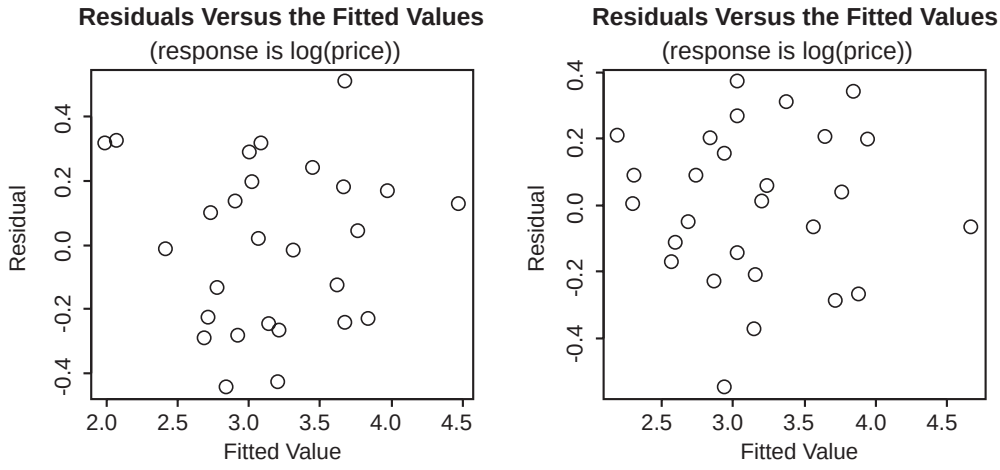
Experience has shown us that when we have a fan in the residuals versus fitted values plot with larger fitted values corresponding to greater scatter, and particularly when it is combined with a curved trend like that in Fig. 15.4.4(b), we usually do much better building a model for[9] $\log(Y)$ than we do building a model for Y itself. Thus we decided to try using log(PRICE) as our response variable, and TEMP, H.RAIN, W.RAIN and YEAR as explanatory variables. Our model is now

$$\log(\text{PRICE}) = \beta_0 + \beta_{\text{TEMP}}\text{TEMP} + \beta_{\text{H.RAIN}}\text{H.RAIN} + \beta_{\text{W.RAIN}}\text{W.RAIN} + \beta_{\text{YEAR}}\text{YEAR} + U,$$

where U is our random error term.

The residuals versus fitted values plot from the log(PRICE) model (Fig. 15.4.5(a)) is enormously better than any we have seen for this data before. We are still not quite happy with it, however. Discounting the two points in the top left hand corner, we see an upward drift. (We could well be getting overly picky here.) There is a slight curve in the residuals versus YEAR plot, so we tried adding YEAR^2. The resulting residuals

[9]In our account, as in most packages "log" is used to refer to natural logarithms ($\log_e$) which often correspond to the ln button on a calculator.



(a) Before YEAR^2 has been added **(b)** After YEAR^2 has been added

FIGURE 15.4.5 Modeling log(PRICE).

versus fitted values plot in Fig. 15.4.5(b) looks better and the plots of residuals versus each of the X-variables in the model (not shown) look good. The P-value for the squared term is less than 5% so we will leave it in for the time being.

We are fairly happy with the residual analysis of this model so far. This is, however, data collected over time so we have some more things things to check out. YEAR is our time variable and we have already checked the residual plot for YEAR. But we have yet to check for *serial correlation* (i.e. time-order correlation between the residuals). Fig. 15.4.6(a) shows the plot of residuals versus lagged residuals. There is a hint of a downward drift, but the standard formal test for serial correlation, namely the *Durbin-Watson test*, gives a non-significant result.[10] Now that everything else has checked out, we check for Normality of the errors. The Normal Q-Q plot of the residuals, given in Fig. 15.4.6(b), is very close to linear. We have finally built a model for which all of the assumptions appear to be satisfied.

The process we have gone through above is the statistical modeling cycle depicted in Fig. 15.4.7. As we knew no theory to guide our original choice of model, we used the simple model containing the variables of interest as a starting point and progressively added features, each time reacting to the inadequacies we saw in the current model in

[10]If we had found significant serial correlation, we would then have set about enlarging our model to incorporate a description of the dependence pattern in the errors using a time series model for the errors. We would have begun by trying the simplest such model, the *autoregressive model* of order 1, or AR(1).

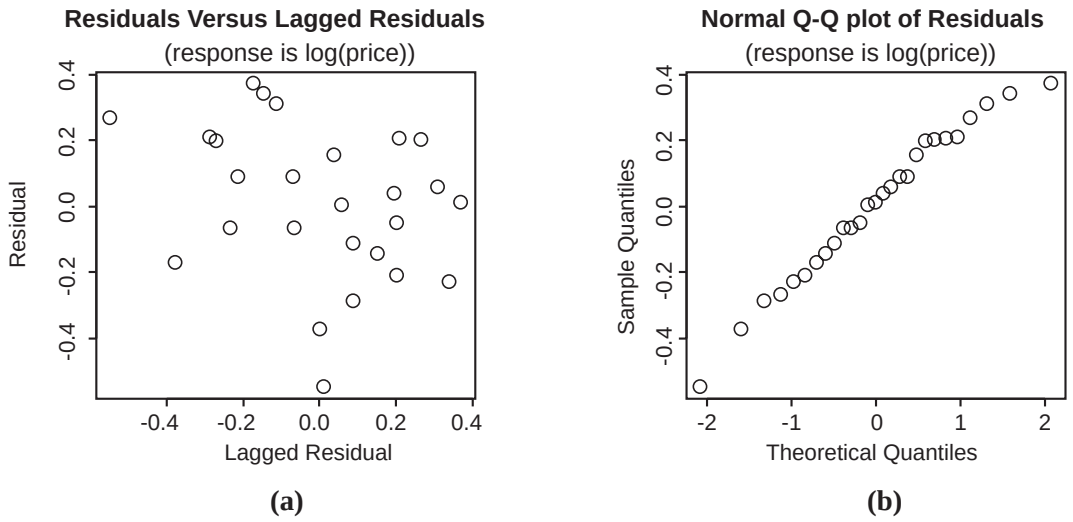


FIGURE 15.4.6 Two residual plots for the "final" Bordeaux wine data model.

the residual plots. We have eventually come up with a model that fits well. Now it is time to use it to make inferences about the data.

Basic computer output from the model is given in Fig. 15.4.8(a). There are complications in interpreting the coefficients of this model as the coefficients tell us about changes in log(PRICE). We are more interested in what happens to PRICE. In Chapter 16, which discusses the use of transformations in data analysis, we will learn how to interpret the coefficients of a linear model for $\log(Y)$ in terms of multiplicative (rather than additive) effects, and we will revisit this analysis. In the meantime, we note that if $\log(Y)$ increases (resp. decreases) with an X-variable, then so does Y. For example, when all other variables held fixed, we have strong evidence that log(PRICE), and hence PRICE, goes down with increasing rainfall at harvest time and goes up with increasing rainfall in the winter. Also wines grown under the same climatic conditions are getting cheaper (in inflation-adjusted currency) over time, and we have some evidence that the relationship with temperature is curved. If you plot the $-7.175\,\mathrm{TEMP} + 0.2381\,\mathrm{TEMP}^2$ over the range of the data, you will find that the estimated temperature effect is increasing with TEMP.

The analysis above used data on Bordeaux vintages from 1952 to 1980. Figure 15.4.8(b) gives the climate-variable values for the next twelve years, i.e., from 1981 to 1992. Figure 15.4.8(c) gives the results of using the model in (a) to predict log(PRICES) for these new vintages. The highest point predictions are 4.99 for 1989 and 5.18 for 1990, so

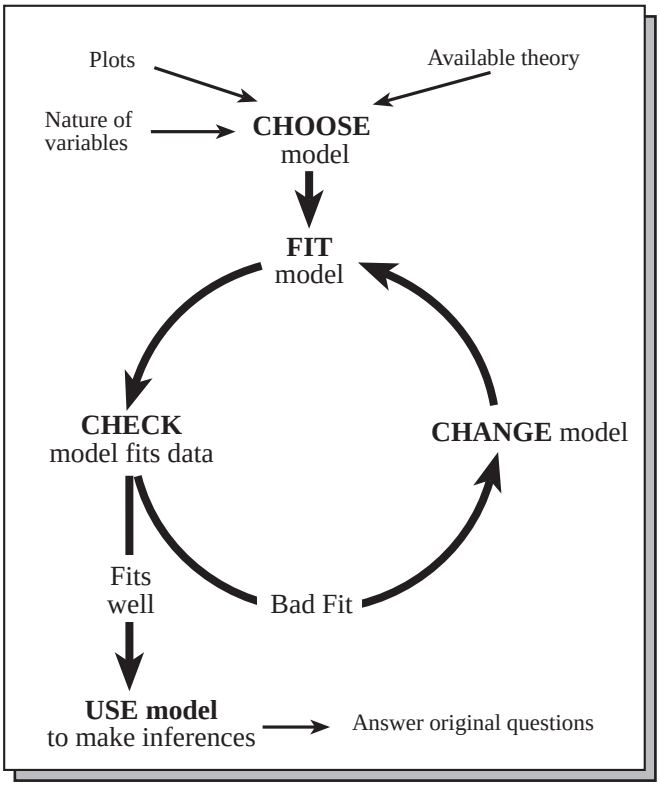


FIGURE 15.4.7 The statistical modeling cycle.

(a) Basic Minitab output

Regression Analysis

```
The regression equation is
log.price = 106 - 7.17 temp + 0.238 temp.sq - 0.00381 h.rain
            + 0.00153 w.rain - 0.0253 year

Predictor        Coef     SE Coef          T        P
Constant       105.95       32.03       3.31    0.003
temp           -7.175       3.486      -2.06    0.052
temp.sq        0.2381      0.1065       2.24    0.036
h.rain     -0.0038060   0.0007285      -5.22    0.000
w.rain      0.0015263   0.0004602       3.32    0.003
year        -0.025264    0.006447      -3.92    0.001

S = 0.2580      R-Sq = 86.3%     R-Sq(adj) = 83.1%
```

(b) Input data for the prediction

temp.p	temp.p.sq	h.rain.p	w.rain.p	year.p
17.0	289.00	111	535	1981
17.4	302.76	162	712	1982
17.4	302.76	119	845	1983
16.5	272.25	119	591	1984
16.8	282.24	38	744	1985
16.3	265.69	171	563	1986
17.0	289.00	115	452	1987
17.1	292.41	59	808	1988
18.6	345.96	82	443	1989
18.7	349.69	80	468	1990
17.7	313.29	183	570	1991
17.9	320.41	342	543	1992

(c) Results of prediction

Minitab output from the prediction

year.p	Fit	StDev Fit	95.0% CI	95.0% PI	
1981	3.1399	0.1248	(2.8804, 3.3994)	(2.5439, 3.7359)	
1982	3.5972	0.1884	(3.2054, 3.9889)	(2.9328, 4.2615)	
1983	3.9386	0.2255	(3.4695, 4.4076)	(3.2259, 4.6512)	X
1984	2.7182	0.1324	(2.4429, 2.9935)	(2.1151, 3.3212)	
1985	3.4610	0.1604	(3.1275, 3.7945)	(2.8292, 4.0928)	
1986	2.3000	0.1453	(1.9978, 2.6021)	(1.6842, 2.9158)	
1987	2.8464	0.1664	(2.5003, 3.1925)	(2.2079, 3.4849)	
1988	3.6721	0.2011	(3.2539, 4.0903)	(2.9919, 4.3524)	
1989	4.9908	0.5236	(3.9018, 6.0798)	(3.7768, 6.2048)	XX
1990	5.1820	0.5746	(3.9870, 6.3770)	(3.8720, 6.4919)	XX
1991	3.4280	0.2551	(2.8974, 3.9585)	(2.6734, 4.1825)	X
1992	3.0167	0.3376	(2.3146, 3.7189)	(2.1331, 3.9004)	XX

our annotation

```
X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

FIGURE 15.4.8 Using our "final" model.

these look like the most promising wines. If we want to allow for uncertainty, it is the "95.0% PI" prediction intervals we should be looking at since we are interested in actual log(PRICE) values, not averages (see Section 12.4.3 for discussion of the two types of interval). We notice that the prediction intervals are very wide even though, with $R^2 \approx 86\%$ our fitted model explained approximately 86% of the variation in log(PRICE). We only get precise prediction intervals when R^2 is very large indeed. Of course we are really not interested in predicting log(PRICE) at all. What we would really like to predict is the PRICE. Fortunately, this is easy. We turn a prediction for log(PRICE) into a prediction for PRICE by exponentiating. For example, our point prediction for PRICE in 1981 is 23.1, since $\exp(3.1399) = 23.1$. We treat the interval endpoints in the same way. Our prediction interval for the PRICE of the 1981 vintage extends from $\exp(2.5439) = 12.7$ to $\exp(3.7359) = 41.9$.

After thought

When this data was introduced in Example 15.1.1, PRICE was intended to be a measure of the quality of the wine. It was, in fact, the inflation adjusted average price fetched for the vintage. But we have seen that, having adjusted for the effects of the climate variables, the average PRICE has been going down over time. Does this mean that the quality has been going down? Not at all. Consider what has been happening with computers. The inflation-adjusted prices of comparable pieces of computing equipment have dropped rapidly over time so that the computers we buy now are much cheaper than machines with inferior capabilities were in the past. Inflation adjustment is an overall average adjustment for a whole"basket of goods" which ignores the fact that different types of goods and services become more or less expensive at different rates. Increasing competition from premium wines from regions such as California and Australiasia could have been a factor. It is quite likely that premium wines have been becoming cheaper in relative terms. Only wine experts could hope to begin to answer the question as to whether Bordeaux vintages have been deteriorating. What we believe is that inflation adjusted PRICE is perhaps deficient as a measure for comparing the qualities of wines over time. Something like log(PRICE) $-0.025\times$YEAR might be a better measure for the period 1952 to 1980.

15.4.4 Other Model Diagnostics

Over recent years, the development of suitable diagnostics for various statistical models has become somewhat of an industry, and rightly so as it is an important topic. In fact whole books have been written on just regression diagnostics. We have focused, in

this chapter, on the residual plots as a diagnostic tool, which will be adequate for many regression models. However, there are two other aspects that we want to mention briefly just so that you will be familiar with the terms if you come across them. These are *influence* and *multicollinearity*.

Influence

An influential point is a data point which has a profound effect on the fitting of the model, that is, if we shift the point slightly there can be a substantial change in the values of the regression coefficients for the fitted model. The model is sensitive to the position of the point. This concept of influence was introduced briefly for a straight line model in Section 12.4.4 under the heading of "Outliers in X", and you should consult that paragraph and look at Fig. 12.4.13. There the data point (in 2-dimensions) has a large X-value, and the point may be in keeping with the trend determined by the remaining data points or it may act as an outlier in that it is well away from the trend. Thus an influential point comes in two flavors; it may blend nicely with the rest of the data and confirm the apparent trend already there, or it may clash with the apparent trend and suggest a very different picture. Methods of detecting and dealing with such points lie outside the scope of this book. However, some regression packages automatically highlight so-called *high-leverage points*. These points are essentially outliers in the X-space and have the potential to be influential. We typically reanalyse our data with those points omitted and see how much our conclusions would change. If the conclusions do change considerably we have to think very carefully about the scope and purpose of our analysis and what we should now believe.

Multicollinearity

To introduce the notion of "multicollinearity" suppose we measure two variables X_1 and X_2 for a regression model with n data points. Unbeknown to us $X_2 = 2X_1 + 3$. If we fit both X_1 and X_2 as explanatory variables we run into two problems, computational and interpretational. Firstly, the model can't be fitted as the computer can't handle this so-called *ill-conditioning*. Secondly, it is clear that once one of the variables is in the model, the other variable is redundant and provides no new information about the response Y. Clearly we must choose either X_1 or X_2, but not both; and it does not matter which one we choose. Because of the exact linear relationship $2X_1 - X_2 + 2 = 0$, the n pairs of values for X_1 and X_2 lie on a straight line (i.e., are "collinear"), and the correlation coefficient for the pairs will have a correlation coefficient of $r = 1$. In practice, however, the linear relationship may only be approximate so that the correlation is not 1 but something close to 1 such as r=0.95. Both variables can now be fitted, but we still have the problem of not needing both in the model. With just two regressors we

can plot X_1 versus X_2 and thus see what is going on. However, with a larger number of explanatory variables we need other methods for detecting the presence of a linear relationship like $X_1 + 3X_2 + 2X_3 \approx 4$ or even several such relationships. Because the X variables satisfy (or, more usually, approximately satisfy) a linear relationship, we say that the variables are (approximately) collinear, and we refer to the problem as the problem of "multicollinearity". Fitting and interpreting models with multicollinearity is tricky, and the reader is referred to more advanced texts for details.

QUIZ ON SECTION 15.4

(Intro. paragraphs of Section 15.4)

1. In what two very distinct senses do people talk about "the fit of a model"?
2. Why could a model fit well in one of these senses and fail to fit in the other?

(Section 15.4.1)

3. What measure do we use of the explanatory, or predictive power of a model?
4. When we have a single X-variable, the correlation coefficient r measures the correlation between X and Y. What other correlation does it measure?
5. What is R the correlation between?
6. What name is commonly given to R^2?
7. How, intuitively, do we interpret R^2?
8. When will a model give us reasonably precise predictions?
9. If our model assumptions are satisfied, but R^2 is still small, what does this tell us? What avenues remain for improving the situation?

(Sections 12.4.4 and 15.4.2)

10. What assumptions about the errors are made by the linear model?
11. Which assumptions are critical for all types of inference?
12. What types of inference are relatively robust against departures from the Normality assumption?
13. Four types of residual plot were described. What were they and what can we learn from each?
14. What is an "outlier in X" and why do we have to be on the lookout for such observations?
15. What is the basic principle behind residual plotting in the context of model building?
16. What "fix" is suggested by a curved trend in a residuals versus X_j plot.
17. What are some other possible remedies for curved trends in a residuals versus fitted values plot?
18. How can a fan in a residuals versus fitted values plot be addressed?
19. What is "serial correlation" and when do we look for it?

20. What residual plot can be used to diagnose serial correlation?
21. What is the usual test used to test for serial correlation?
22. Is non-Normality of residuals the first or last problem we should worry about?

(Section 15.4.3)

23. What is *serial correlation* and when do we check for it?
24. How do we check for it graphically? What is the name of the usual test for serial correlation in regression models?
25. Describe the statistical modeling cycle.
26. What three considerations feed into our choice of a first model to fit to the data?
27. What is one useful source of ideas about how we might change a model that does not fit?
28. How can we convert a prediction for $\log(Y)$ to a prediction for Y?
29. How can we convert a prediction *interval* for $\log(Y)$ to a prediction interval for Y?

(Section 15.4.4)

30. What is an influential point?
31. What should we do if our computer package identifies a point as being a high-leverage point?
32. What is multicollinearity and why is it a problem?

EXERCISES FOR SECTION 15.4

1. xxx
2. xxx
3. xxx

*****EXERCISES

There is one Indian who is very overweight for his height. Is he the influential point? You may have found him in your pairs plot

15.5 Using Grouping Factors as Explanatory Variables

15.5.1 Including a Binary Variable

A *binary variable* is one that can only take 2 values. It is used to code for membership of one of 2 possible groups, for example male/female, employed/unemployed, alive/dead, or bankrupt/not bankrupt. Whereas once numeric codes were always used, it is now common to use the group names themselves.

We will look at a special, and very useful, way of coding whether an individual does or does not belong to a group of interest.

Use of dummy, or indicator variables

$$\text{Suppose that we code} \quad X_F = \begin{cases} 1, & \text{if the person is female;} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, males have $X_F = 0$ and females have $X_F = 1$. Now suppose that we blindly include this variable in a multiple regression model. Under a linear model, an individual with $X_1 = x_1, \ldots, \ldots, X_k = x_k$ and $X_F = x_F$ has

$$E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_F x_F$$

What does the coefficient of X_F represent? A female has

$$\text{Female}: \quad E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \ + \beta_F \times 1$$

An otherwise identical male has

$$\text{Male} \quad : \quad E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \ + \beta_F \times 0.$$

The difference is β_F. Thus, the coefficient of X_F is the difference in average Y-value between females and otherwise identical males (identical in the sense of having the same values for the other X-variables).

The *indicator* or *dummy variable* for having characteristic A is of the form

$$\text{CHAR.A} = \begin{cases} 1, & \text{if the individual has characteristic A;} \\ 0, & \text{if the individual does not have characteristic A.} \end{cases}$$

The coefficient of CHAR.A is the difference in average Y-value between individuals with A and individuals without A who are the same on all other X-variables.

(a) Response = EXAM, Explanatory = TEST1, TEST2, ASSIGN and **IS.FEMALE**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.60350    0.81609   5.641 2.29e-08 ***
test1        0.85168    0.08400  10.139  < 2e-16 ***
test2        1.48335    0.08452  17.549  < 2e-16 ***
assign       0.02590    0.04061   0.638   0.5238
is.female   -0.59948    0.30127  -1.990   0.0469 *
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

(b) Response = EXAM, Explanatory = **IS.FEMALE** (alone)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.920635   0.290716  82.282   <2e-16 ***
is.female    0.007272   0.413756   0.018    0.986

            95 % C.I.lower    95 % C.I.upper
(Intercept)       23.35005          24.49122
is.female         -0.80481           0.81935
```

FIGURE 15.5.1 Using IS.FEMALE as an explanatory variable.

Alternatively, we can think of the coefficient for CHAR.A as giving the difference in average Y-values between those with and without characteristic A, once we have *adjusted for* (or allowed for, or controlled for, or partialled out) the effects of the other variables in the model. Of course, we can only apply these interpretations to data if the model fits!

Example 15.5.1 *Looking For a Sex Difference in Exam Marks*

Figure 15.5.1(a) gives the results of fitting a linear model to predict EXAM marks of the 1995 Auckland class from TEST1 marks, TEST2 marks ASSIGN (assignment) marks and IS.FEMALE, where IS.FEMALE is an indicator variable for being female. The P-value for IS.FEMALE coefficient is just less than 5% and we calculate the 95% confidence interval for the true value to be given by (-1.19, -0.01). We do seem to have some evidence that females are not doing as well as males, scoring somewhere between 0 and 1.2 marks less on average, *than a male with the same test and assignment scores.*

If we do not adjust for coursework marks as in Fig. 15.5.1 (b), which uses IS.FEMALE as the only explanatory variable, there is no significant gender difference.

Notes:

1. One plausible (to us!) explanation for the change in significance is as follows. It could be that women are working harder on average than men. (If you check, you will find that females do have significantly higher assignment scores than males.) If we are comparing two people who have achieved the same scores, it is plausible that the person who has worked less hard for those scores has a greater natural talent in the subject and this may tend to show in the exam. Well, that's one idea – can you think of some other possible explanations?

2. It is generally true when we fit a regression using a indicator (or dummy) variable as the sole explanatory variable, the P-value and confidence interval for the effect of that variable is identical to the P-value and confidence interval we get from a pooled 2-sample t-test for a difference in mean Y-values between the two groups.

   In this case, you can verify that the P-value and confidence interval from a pooled 2-sample t-test for a difference in average exam scores between females and males is the same as the P-value and confidence interval for the IS.FEMALE effect in Fig. 15.5.1(b). (Can you see why this is?)

EXERCISE FOR SECTION 15.5.1

Fit a linear regression model to the Fuel Usage Data in Appendix A15.1. using KM.LTR as the response variable and DIST, MO.JAN, and LONG as explanatory.

1. What can you conclude from the P-value and the sign of the coefficient for the LONG effect?

2. Obtain and interpret the confidence interval for the true coefficient of LONG.

15.5.2 Including a Factor

A *(grouping) factor*, or *categorical variable*, is a variable that specifies which of a set of groups each individual belongs to. In the SAS package, a factor is called a *class variable*. The set of groups are called the *levels* of the factor. In coursework data set of Example 15.5.1 and Appendix A15.1, the variable SEX is a factor with two levels "male" and "female". DEGREE is a factor with 3 levels ("BA", "BCom" and "BSc") corresponding to the 3 different types of degree students are enrolled for.[11] We are going to want to see the extent to which DEGREE helps explain the differences we see in exam performance.

[11]Likewise DEGREE.CODE is a factor with 3 (numeric) levels 1, 2 and 3 which is being used as numeric codes for"BA", "BCom" and "BSc" respectively.

If we want to use a factor with more than 2 levels as an explanatory variable in a linear model we do so by including in the model a dummy variable for every level of the factor except one. What effect does this have?

Suppose that our factor has J levels. Let G_1 be a dummy variable for being in group 1, G_2 be a dummy variable for being in group 2, and so on. Suppose we omit the dummy variable for the first group. Our linear model is

$$E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \quad + \beta_{G_2} G_2 + \beta_{G_3} G_3 + \ldots + \beta_{G_J} G_J$$

Considering just individuals who have $X_1 = x_1, \ldots, X_k = x_k$, we have the following.

An individual in group 1 is not in any of the other groups and thus has $G_2 = G_3 = \ldots = G_J = 0$. Substituting these values into the model, we get

$$\text{Group 1 individuals:} \quad \mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

An individual in group 2 has $G_2 = 1$ and every other $G_j = 0$, so

$$\text{Group 2 individuals:} \quad \mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{G_2}.$$

By subtracting the Group 1 line from the Group 2 line, we see that β_{G_2} is the difference in average Y-value between Group 2 individuals and Group 1 individuals (who are the same on all other X-variables).

An individual in group 3 has $G_3 = 1$ and every other $G_j = 0$, so

$$\text{Group 3 individuals:} \quad \mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{G_3}.$$

Subtracting the Group 1 line from the Group 3 line, we see that β_{G_3} is the difference in average Y-value between Group 3 individuals and Group 1 individuals. And so it goes on for each of the other groups. Thus Group 1 is being treated as a *baseline* group to which all other groups are being compared and β_{G_j} measures the difference between group j and group 1, with all other variables being held constant.

Using a (grouping) factor as an explanatory variable

- Include a dummy variable for every level of the factor except one.
- The omitted group beomes the baseline group to which all other groups are compared.
- The **coefficient for group** j is *the difference* in average Y-values *between group* j *and the baseline group*

  (for individuals who are the same on all of the other X-variables.)
- The **intercept** is the average Y-value for members of the baseline group with every other $X_j = 0$.

Alternatively, we can think in terms of the coefficient for group j as giving the difference in average Y-values between group j and the baseline group once we have *adjusted for* (or allowed for, or controlled for, or partialled out) the effects of the other variables in the model. Of course, we can only apply these interpretations to data if the model fits.

Example 15.5.2 *Looking for Differences by Degree in Exam Marks*

This is a continuation of Example 15.5.1 using the data set described in Appendix A15.1. The people sitting the examinations in Example 15.5.1 were each enrolled for one of 3 different degrees; a BA (Arts), a BCom (business), or a BSc (science). The variable DEGREE tells which degree each student is enrolled for. We want to see if there is a difference in average performance between students enrolled for different degrees.

Since DEGREE is a factor with 3 levels, we include in our model dummy variables for two of those levels. Output from fitting 2 models is given in Fig. 15.5.2. By default, Splus/R have put in dummies for the levels BCom and BSc and omitted the level BA. Thus, the BA group is the baseline group to which the other two groups will be compared. The coefficient in the `degreeBCom` row relates to the difference between the BCom group and the BA group, and the coefficient in the `degreeBSc` row relates to the difference between BSc and BA.

EXAM is the response variable for both models. In Fig. 15.5.2(a) we are using DEGREE as the sole explanatory variable. In Fig. 15.5.2(b), we are using TEST1,

(a) Using DEGREE alone

Regression output / *CIs added*

```
Coefficients:
             Estimate Std.Error t value Pr(>|t|)   95% C.I.lower  C.I.upper
(Intercept)  22.1158    0.6214  35.588  < 2e-16     20.89608       23.33550
degreeBCom    1.3683    0.7182   1.905 0.057090     -0.04133        2.77795
degreeBSc     2.4096    0.6787   3.550 0.000405      1.07752        3.74162

Residual standard error: 6.057 on 868 degrees of freedom
Multiple R-Squared: 0.01679,      Adjusted R-squared: 0.01452
F-statistic: 7.409 on 2 and 868 degrees of freedom,   p-value: 0.0006447
```

Same as from 1-way ANOVA when the factor is the only explanatory variable

(b) Using DEGREE and the other variables

Regression output / *CIs added*

```
Coefficients:
             Estimate Std.Error t-value Pr(>|t|)   95% C.I.lower  C.I.upper
(Intercept)  3.20947   0.95250   3.370 0.000786      1.33998        5.07897
test1        0.85684   0.08822   9.712  < 2e-16      0.68368        1.03000
test2        1.47111   0.08587  17.132  < 2e-16      1.30258        1.63964
assign       0.03613   0.04148   0.871 0.383994     -0.04528        0.11753
sex          0.56348   0.30151   1.869 0.061980     -0.02830        1.15526
degreeBCom   0.98902   0.51258   1.929 0.053998     -0.01703        1.99508
degreeBSc    0.62368   0.49416   1.262 0.207254     -0.34622        1.59359

Residual standard error: 4.287 on 864 degrees of freedom
Multiple R-Squared: 0.5098,       Adjusted R-squared: 0.5064
F-statistic: 149.7 on 6 and 864 degrees of freedom,   p-value:     0
```

The "no regression" hypothesis now refers to none of the variables, including test1, etc. having any effect

(c) An additional ANOVA table

```
> anova(fit)
Analysis of Variance Table

Response: exam
           Df  Sum Sq Mean Sq  F value  Pr(>F)
test1       1  9873.0  9873.0 537.2233 < 2e-16
test2       1  6493.2  6493.2 353.3161 < 2e-16
assign      1     0.8     0.8   0.0446 0.83279
sex         1    72.9    72.9   3.9678 0.04669
degree      2    70.6    35.3   1.9219 0.14695
Residuals 864 15878.4    18.4
```

FIGURE 15.5.2 Using DEGREE as an explanatory variable.

TEST2, ASSIGN, and SEX as well as DEGREE as explanatory variables. The former helps us answer the question, "Are there differences in average exam scores between students enrolled in different degrees?" The latter helps us answer the question, "Are there any differences in average exam scores between students enrolled in different degrees *once we have adjusted for the effects of gender and performance during the year*?" These are quite different questions.

When we fit a factor as the sole explanatory variable, we are really doing a 1-way analysis of variance, and the F-test results for the test for no regression are identical to the F-test results from a 1-way analysis of variance. The regression intercept estimate is the sample mean for the baseline (omitted) group, and the regression coefficient estimate for group j is the difference between the group j sample mean and the baseline sample mean.

We see from the F-test in Fig. 15.5.2(a) that we have very strong evidence of a difference in mean test scores between the three different degree groups (P-value $\approx .0006$). The P-value of 0.057 on the `degreeBCom` line tells us that we have some evidence of a difference between the BCom group and the BA group, with the 95% confidence interval for the true difference putting the BCom mean as being somewhere between 0.4 marks smaller than the BA mean and 2.8 marks larger. Reading the `degreeBSc` line, we have very strong evidence of a (positive) difference between the BSc mean and the BA mean. With 95% confidence the true BSc mean is bigger than the BA mean by between about 1 mark and 3.7 marks. The regression output does not give us a comparison between the BSc and BCom means. To get test results and confidence intervals for this comparison, we would have to refit the model changing the omitted level from BA to BCom (or BSc but not both).

We now turn to the output in Fig. 15.5.2(b). In this case, we have fitted other variables as well as DEGREE. The "no regression" hypothesis no longer addresses just the effect of DEGREE but rather "no effect by any of the variables." Many packages will give an analysis of variance table similar to that in Fig. 15.5.2(c), which breaks up the regression sum of squares by variable. Here we see the results of an F-test for a null hypothesis of no differences between DEGREE groups *once we have adjusted for the effects of* gender and performance in on-course assessment. Since P-value ≈ 0.15, there is no evidence of a difference after adjustment. Thus the differences in exam performance by DEGREE that we saw in Fig. 15.5.2(a) appear to be explained by gender differences and differences in performance in on-course assessment. Confidence intervals are also given in Fig. 15.5.2(b), and they can be interpreted as for (a) with the proviso that we are now no longer looking at differences between the DEGREE groups as a whole, but

looking at differences after adjustment. This is the same as looking at differences in average exam score by DEGREE group, restricted to people with the same gender and on-course performance scores.

Notes:

1. The models we have fitted assume that the (adjusted) effects of a difference in factor levels are the same regardless of the values of the other variables. For example, we are assuming that the true DEGREE differences among people who did well in on-course assessment are the same as the DEGREE differences among people who did poorly in on-course assessment. This so-called *no interactions* assumption can be tested by checking for the significance of cross products between the factor-level dummy variables and other variables. This is beyond the scope of the present account.

2. Some packages (e.g. Minitab and Excel) cannot automatically produce the results of the overall F-test of no differences between factor levels after adjustment. We will now show you how to do this "by hand".

The F-test by hand

The following F-test can be used to test any hypothesis that corresponds to collapsing a larger regression model (the *full model*) to a smaller model[12] (the *collapsed model*). The idea behind the test statistic is as follows. The residual sum of squares (Residual.SS) is a measure of how well a model fits the data.[13] A model with more variables will have a smaller Residual.SS as it has more flexibility to "get close" to the data points. The test is based on whether the Residual.SS for the full model is sufficiently smaller than that for the collapsed model to be convincing evidence that the additional variables are necessary.

Step 1: Fit the full model.

Let $\text{Residual.SS}_{Full}$, $\text{Residual.MS}_{Full}$ and $\text{Residual.df}_{Full}$ be, respectively, the *sum* of squares, the *mean* square, and the degrees of freedom *from the residual line* of the analysis of variance table for the full model.

Step 2: Fit the collapsed model.

Let $\text{Residual.SS}_{Collapsed}$ be the residual *sum* of squares for the collapsed model.

[12] i.e., one with fewer terms.

[13] In the sense of having predicted values ($\widehat{y}$) close to the observed y-values.

Step 3: Form the F-test statistic

$$f_0 = \frac{\text{Residual.SS}_{Collapsed} - \text{Residual.SS}_{Full}}{d \times \text{Residual.MS}_{Full}},$$

$$\text{where } d = \text{Residual.df}_{Collapsed} - \text{Residual.df}_{Full}$$

Step 4: Obtain the P-value as $\text{pr}(F \geq f_0)$,
where $F \sim F(df_1 = d, df_2 = \text{Residual.df}_{Full})$.

This process is depicted in Fig. 15.5.3 using Excel output.

Example 15.5.3 *The F-test for Differences by Degree Computed by Hand*

We wish to perform the (adjusted) F-test for degree shown in Fig. 15.5.2 "by hand". In Step 1, we fit the full model with TEST1, TEST2, ASSIGN, SEX and DEGREE as explanatory variables and read off the appropriate terms from the output. In Step 2, we the collapse the above model by omitting DEGREE (because our hypothesis is that there is no DEGREE effect). Thus the collapsed model includes only includes TEST1, TEST2, ASSIGN, and SEX. The fitted models from Steps 1 and 2 and the calculation in Step 3 are presented in Fig. 15.5.3 from which we obtain $f_0 = 1.92$. The P-value (Step 4) is given by $\text{pr}(F \geq 1.92)$, where $F \sim F(df_1 = 2, df_2 = 864)$ which gives us P-value $= 0.147$.

EXERCISES FOR SECTION 15.5.2

1. xxx
2. xxx
3. xxx

QUIZ ON SECTION 15.5

1. What is a binary variable?
2. What is an indicator, or dummy, variable?
3. If we include an indicator variable in a regression model, what does its coefficient measure?
4. How do we include a (grouping) factor as an explanatory variable in a regression model?
5. Which level of the factor becomes the baseline group for comparisons?

The Full Model

ANOVA

	df	SS	MS	F	Significance F
Regression	6	16510.6	2751.760	149.73	4.4E-130
Residual	864	15878.4	18.378		
Total	870	32389.0			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.7730	0.9188	4.11	4.4E-05	1.9697	5.5763
assign	0.0361	0.0415	0.87	3.8E-01	-0.0453	0.1175
test1	0.8568	0.0882	9.71	1	0.6837	1.0300
test2	1.4711	0.0859	17.13	7	1.3026	1.6396
is.female	-0.5635	0.3015	-1.87	2	-1.1553	0.0283
degreeBCom	0.9890	0.5126	1.93	5.4E-02	-0.0170	1.9951
degreeBSc	0.6237	0.4942	1.26	2.1E-01	-0.3462	1.5936

$$d = \textbf{Residual.df}_{\textbf{Collapsed}} - \textbf{Residual.df}_{\textbf{Full}} = 866 - 864 = 2$$

The Collapsed Model

ANOVA

	df	SS	MS	F	Significance F
Regression	4	16439.9	4109.979	223.16	1.34E-131
Residual	866	15949.1	18.417		
Total	870	32389.0			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4.6035	0.8161	5.64	2.3E-08	3.0017	6.2053
assign	0.0259	0.0406	0.64	5.2E-01	-0.0538	0.1056
test1	0.8517	0.0840	10.14	6.7E-23	0.6868	1.0166
test2	1.4833	0.0845	17.55	3.2E-59	1.3175	1.6492
is.female	-0.5995	0.3013	-1.99	4.7E-02	-1.1908	-0.0082

$$f_0 = \frac{\textbf{Residual.SS}_{\textbf{Collapsed}} - \textbf{Residual.SS}_{\textbf{Full}}}{d \times \textbf{Residual.MS}_{\textbf{Full}}} = \frac{15949.1 - 15878.4}{2 \times 18.378} = 1.92$$

$Residual.SS_{Full}$

$Residual.MS_{Full}$

$Residual.df_{Full}$

$Residual.df_{Collapsed}$

$Residual.SS_{Collapsed}$

FIGURE 15.5.3 Doing the *F*-test "by hand". (Annotated Excel Ouput.)

6. What does the coefficient of a dummy variable that has been included (say, the dummy for group j) measure?
7. What does the intercept term in the model represent?
8. We have learned about a new F-test in Section 15.5.2. What is the purpose of this test? How do we interpret a small P-value from this test? How do we interpret a large P-value from this test?
9. What major assumption are we making in interpeting the coefficients of dummy variables in the way we have learned in Section 15.5?

APPENDIX for Chapter 15: Data Sets

A15.1: Fuel Usage Data

File name: fuel.dat
Description

These data originated from a car's journey-log kept by a brother of University of Waterloo statistics professor Jack Robinson over a 12 month period. The car was being used in Canada and the US. Each time he filled his car with gas (petrol), he recorded (raw data) his odometer reading (number of kilometers the car had traveled to date; the date; `liters`, number of liters of fuel put into the car; and some comments about servicing. etc. From these the variables below were obtained.

Variables:

`odo`	Odometer reading at fill-up
`dist`	distance traveled in kilometers since last fill up
`liters`	Liters of petrol put into tank on a given fill-up
`km.ltr`	fuel consumption in kilometers traveled per liter
`days`	days since last fill up
`long`	for long trip – (prior fill up less than 2 days earlier). (1 = yes, 0 = no)
`month`	month of the year with values 1 = Jan, 2 = Feb, etc.
`mo.jan`	months away from January (the coldest month in the year) e.g., December and February both have `mo.jan` = 1 November and March both have `mo.jan` = 2, etc.

A15.2: Coursework Data

File name: coursework.dat
Description

These data consist of the examination and coursework scores for the students doing second year Data Analysis at Auckland in 1995. At that time, the course ran through the whole year and students completed 10 Assignments and 2 tests during the year. There is a great deal of help available for assignments so that anyone who is prepared to put the work in can get good assignment marks. Our main interest will be in `Exam` and the other variables will be treated as explanatory.

Variables:

`exam`	Mark in the exam, out of 40.
`assign`	Adjusted overall mark for assignments, out of 25
`test1`	Adjusted mark in test 1, out of 12.5
`test2`	Adjusted mark in test 2, out of 12.5
`sex`	Gender of student, with levels: male, female
`is.female`	Gender coded 1 = female, 0 = male
`degree`	Degree student is enrolled in, with levels: BA, BSc and BCom
`degree.code`	Degree coded as 1 = BA, 2 = BCom and 3 = BSc