Chapter 13

Multiple Linear Regression

Chapter Objectives

- State and interpret the multiple regression model.
- Obtain and interpret estimates of model coefficients.
- Obtain and interpret fitted values and residuals associated with a fitted multiple regression model.
- Interpret sums of squares, degrees of freedom, and mean squares.
- Interpret the R^2 associated with a fitted multiple regression model and use it to assess how well the model fits the data.
- Carry out t tests for the coefficients in a multiple regression model.
- Obtain t confidence intervals for the coefficients in a multiple regression model.
- Carry out a regression model F test.
- Decide whether the t tests (and F test) associated with a multiple regression analysis are appropriate for a given set of data.
- Recognize multicollinearity and describe the main consequences of performing multiple regression in the presence of multicollinearity.
- Use the adjusted R^2 and the Akaike Information Criterion (AIC) to compare suitabilities of models containing different sets of explanatory variables.

Key Takeaways

- A multiple regression analysis is used to estimate the equation of a linear relationship between a response variable and multiple numerical explanatory variables. Non-linear patterns in data can be transformed to linear ones prior to conducting the analysis.
- A t test for a coefficient is a test for whether there's a linear relationship between the response variable and a particular explanatory variable, holding the other explanatory variables constant.
- The model F test is a test for whether there's a linear relationship between the response variable and at least one of the explanatory variables.
- Both the t and F tests require either that the response variable is normally distributed or the sample size is large. A log transformation can make a right-skewed response variable more normal prior to conducting a t or F test.
- A multiple regression model describes variation in a response variable in terms of multiple numerical explanatory variables. It contains two parts: one representing non-random variation due to the linear relationship to the explanatory variables and another representing random variation (random error).
- Sums of squares in multiple regression are statistics that measure variation in the observed values of a response variable due to the linear relationship to the explanatory variables and due to random error.
- Mean squares are another way to measure variation. They're obtained by dividing sums of squares

by their degrees of freedom. The degrees of freedom associated with a sum of squares is determined by how many of its squared deviations are "free to vary." The values of two mean squares are directly comparable, but the values of two sums of squares aren't necessarily comparable.

- The R^2 is a statistic that measures how well a fitted multiple regression model fits the data. Expressed as a percent, its value is interpreted as the percent of variation in the response variable that's explained by variation in the explanatory variables. A larger R^2 value indicates a better fitting model. The R^2 value always increases as more explanatory variables are added to the model.
- The regression model F test statistic is a ratio of two mean squares. Its numerator measures variation that's due to the explanatory variables and its denominator variation that's due to random error.
- The t test statistic for a coefficient indicates how many standard errors away from zero the estimate of the coefficient lies.
- Multicollinearity refers to correlations among explanatory variables. The main consequence of performing multiple regression in the presence of multicollinearity is that it can lead to unreliable estimates of model coefficients.
- The adjusted R^2 is a statistic that measures how well a multiple regression model fits the data, adjusting for how many explanatory variables are contained in the model. A larger adjusted R^2 value indicates a better fitting model, but the value doesn't always increase as more explanatory variables are added to the model.
- Either the the adjusted R^2 or the AIC statistic can be used to compare suitabilities of two or more multiple regression models that differ by the sets of explanatory variables they contain.
- Automated (e.g. stepwise) variable selection procedures can assist with the task of deciding which explanatory variables to include in a model.

13.1 Introduction

In environmental studies, it's often the case that a *single* explanatory variable doesn't adequately explain the variation in the response variable, and instead *multiple* explanatory variables are needed. In this chapter, we look at methods for analyzing data consisting of a response variable and multiple *numerical* explanatory variables. We saw how to analyze data for which there were two categorical explanatory variables (factors) in Chapter 11.

13.1.1 Uses for Multiple Linear Regression

Some reasons for including more than one predictor in a regression model include:

- The model that includes the additional predictor may end up explaining substantially more of the variation in Y, thereby producing more precise predictions of Y values.
- Including multiple predictors in the model allows us to study the effect of one predictor on Y while *controlling* for the effects of the other predictors.
- The model that includes more than one predictor can be used to investigate how Y responds to *simultaneous* changes in the predictors.

The following example illustrates the first two of these reasons. Example 13.5 illustrates the third.

Example 13.1: Uses for Multiple Regression

The scatterplot below shows, for the metropolitan areas of 28 U.S. cities, the water consumption (log of millions of liters/day) for commercial, industrial, and residential uses versus the median income (standardized as a z-score), which serves as a measure of the city's wealth [4].

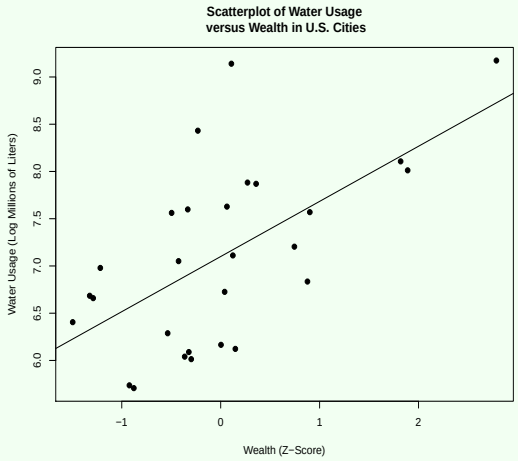


Figure 13.1: Scatterplot of water usage (log of millions of liters/day) versus wealth (z-score of the city's median income) for 28 U.S. metropolitan areas.

The data are shown below. Included are the cities' populations (in 2000).

Water Usage for U.S. Metropolitan Areas

City	Water Usage	Population	Wealth
New York	9.17	21,286,485	2.787
Los Angeles	9.14	16,373,645	0.108
Chicago	8.43	9,157,540	-0.231
DC/Baltimore	8.11	6,484,212	1.819
San Francisco	8.01	6,262,629	1.890
Detroit/Ann Arbor	7.88	5,456,428	0.271
Dallas	7.87	5,221,801	0.358
Atlanta	7.57	4,265,642	0.901
Seattle	7.60	3,554,760	-0.333
Miami	7.56	3,876,380	-0.496
Phoenix	7.63	3,251,876	0.063
Minneapolis	6.84	2,968,806	0.877
Denver	7.20	2,581,506	0.745
Pittsburgh	6.68	2,516,960	-1.324
St. Louis	6.73	2,225,418	0.040
Portland/Salem	7.05	2,265,223	-0.426
San Antonio	6.66	1,592,383	-1.289
Salt Lake City	6.98	1,333,914	-1.217
Las Vegas	7.11	1,563,282	0.123
Providence	6.17	1,497,564	0.003
Jacksonville	6.01	1,100,491	-0.298
Dayton/Springfield	6.09	950,558	-0.322
Albany/Schenectady	6.12	875,583	0.148
Albuquerque	6.29	712,738	-0.536
Omaha	6.04	716,998	-0.364
Little Rock	5.71	583,845	-0.878
Stockton	5.74	563,598	-0.923
Mobile	6.41	540,258	-1.496

The equation of the fitted regression line shown in the scatterplot is

$$\hat{Y} \;=\; 7.10 + 0.584X,$$

Thus we estimate that the log of a city's water usage will be about 0.584 units higher, on average, for every increase of one *standard unit* in its median income. The 28 cities are very different in size though, ranging from rather small (Mobile, AL, pop. 540,258) to very large (New York, NY, pop. 21,286,485). There are two main reasons why we may want to include population as a predictor

along with wealth in a model describing water consumption.

First, notice in Fig. 13.1 that there's substantial variation away from the regression line, so the prediction of a city's water usage based on its wealth alone won't be very reliable. The deviations away from the line are due to the fact that wealth isn't the only factor that determines how much water a city uses. Among the other factors is the size of the city: larger cities use more water. Clearly, incorporating information about a city's size would make the prediction of its water usage more reliable. Example 13.10 will show that the excess variation away from the line is reduced substantially when city size is included in the statistical model.

The second reason for including population in the model is to *control* for its effect on water usage while investigating the effect of wealth on water usage. It turns out that wealthier cities tend to be larger. A direct comparison of the water usage for a high-income city to that of a low-income one isn't valid if the high-income city is much larger. A better approach, if the goal is to find out if wealth impacts water usage, would be to compare the water usages of high- and low-income cities that are the same size, that is, while holding size constant. By including population as a predictor in the model, we're able to do just that – investigate the relationship between water usage and wealth while holding city size constant. This will be illustrated further in Examples 13.8 and 13.19.

In the last example, because wealthier cities tend to be larger, the effect of wealth on a city's water usage is ***confounded*** with the effect of city size (see Chapter 2 for a discussion of confounding). Thus *one way to control for the effects of confounding variables is to include them in the statistical model.* We'll return to this point in Section 13.18.

13.1.2 Notation

In this chapter, we'll use *multiple regression* to investigate the relationship between a response variable $\boldsymbol{Y}$ and two or more predictor variables $\boldsymbol{X_1, X_2, \ldots, X_p}$, with $\boldsymbol{p}$ denoting the number of predictor variables.

Example 13.2: Multiple Regression Notation

Efficient design of municipal waste incinerators requires knowing the energy content of the waste that's to be incinerated. To investigate the relationship between the energy content of waste and the composition of the waste, the following variables were measured on each of 30 waste specimens [6]:

Y = Energy content (kcal/kg)

X_1 = Percent plastics by weight
X_2 = Percent paper by weight
X_3 = Percent garbage by weight
X_4 = Percent moisture by weight

The data are below.

Municipal Waste Composition

Waste Specimen	Energy Content	Plastics	Paper	Garbage	Water
1	947	18.69	15.65	45.01	58.21
2	1407	19.43	23.51	39.69	46.31
3	1452	19.24	24.23	43.16	46.63
4	1553	22.64	22.20	35.76	45.85
5	989	16.54	23.56	41.20	55.14
6	1162	21.44	23.65	35.56	54.24
7	1466	19.53	24.45	40.18	47.20
8	1656	23.97	19.39	44.11	43.82
9	1254	21.45	23.84	35.41	51.01
10	1336	20.34	26.50	34.21	49.06
11	1097	17.03	23.46	32.45	53.23
12	1266	21.03	26.99	38.19	51.78
13	1401	20.49	19.87	41.35	46.69
14	1223	20.45	23.03	43.59	53.57
15	1216	18.81	22.62	42.20	52.98
16	1334	18.28	21.87	41.50	47.44
17	1155	21.41	20.47	41.20	54.68
18	1453	25.11	22.59	37.02	48.74
19	1278	21.04	26.27	38.66	53.22
20	1153	17.99	28.22	44.18	53.37
21	1225	18.73	29.39	34.77	51.06
22	1237	18.49	26.58	37.55	50.66
23	1327	22.08	24.88	37.07	50.72
24	1229	14.28	26.27	35.80	48.24
25	1205	17.74	23.61	37.36	49.92
26	1221	20.54	26.58	35.40	53.58
27	1138	18.25	13.77	51.32	51.38
28	1295	19.09	25.62	39.54	50.13
29	1391	21.25	20.63	40.72	48.67
30	1372	21.62	22.71	36.22	48.19

We'll explore these data graphically in Example 13.3, and in Example 13.4 we'll summarize the relationships between the variables by their correlations. An objective of the study was to use the data to develop a model for predicting energy content from the $p = 4$ predictor variables. In Example 13.7 we'll fit the *multiple regression model* to the data, and in Example 13.9 we'll assess how well it fits.

Note that for each of the waste specimens in the last example, all four predictors were measured along with the response. Data for which more than one variable is measured on each of $\boldsymbol{n}$ individuals are referred to as ***multivariate*** data. We store the data (for use with statistical software) in columns as below.

Observation	Y	X_1	X_2	$\cdots$	X_p
1	Y_1	X_{11}	X_{21}	$\cdots$	X_{p1}
2	Y_2	X_{12}	X_{22}	$\cdots$	X_{p2}
3	Y_3	X_{13}	X_{23}	$\cdots$	X_{p3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_{1n}	X_{2n}	$\cdots$	X_{pn}

As shown above, we use the notation

$\boldsymbol{p}$ = The number of predictor (explanatory) variables $X_1, X_2, \ldots, X_p$.
$\boldsymbol{n}$ = The number of individuals upon which $X_1, X_2, \ldots, X_p$ and Y are measured, or sample size.
$\boldsymbol{Y_i}$ = The value of the response variable for the ith individual.

$\boldsymbol{X_{1i}, X_{2i}, \ldots, X_{pi}}$ = The values of the p predictor variables $X_1, X_2, \ldots, X_p$ for the ith individual.

For a single individual, the set of response and predictor variable values is referred to as a ***multivariate observation***. In the municipal waste example, there are $n = 30$ multivariate observations, each represented by a row in the data table.

13.1.3 Summarizing and Graphing Multivariate Data

The first step in analyzing a multivariate data set is to *explore* the data using graphics and summary statistics. Conventional ***univariate*** methods (means, standard deviations, histograms, boxplots, etc.) can be performed separately on each variable, but ***multivariate*** methods, which involve more than one variable at a time, are needed to explore the *interrelationships* among the variables.

Scatterplot Matrices

One of the most useful graphical displays for multivariate numerical data is a ***scatterplot matrix***, which consists of scatterplots of the variables in a multivariate data set taken, two at a time, arranged in rows and columns of an array.

Example 13.3: Scatterplot Matrices

The scatterplot matrix for the municipal waste data of Example 13.2 is shown below.

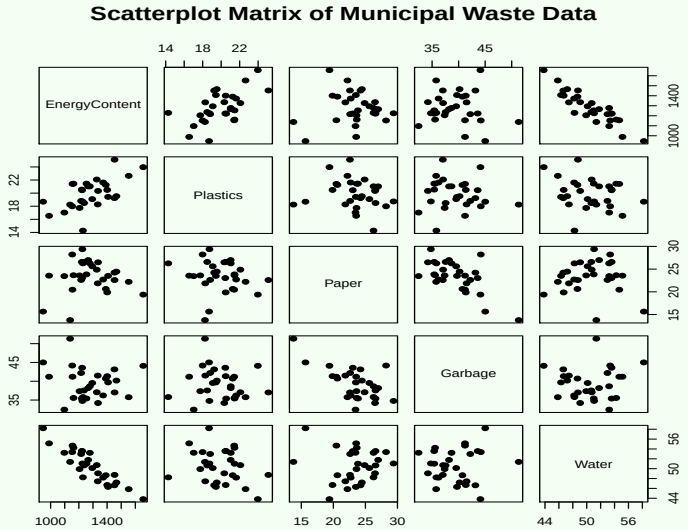


Figure 13.2: Scatterplot matrix of the variables in the municipal waste data set.

The labels along the diagonal refer to the axes in the scatterplots. For example, the y-axis of the scatterplots in the entire first row is energy content, and the x-axis for plots in the entire fifth column is water.

Each of the plots below the diagonal shows the same two variables as one of the plots above it, but with the axes transposed. Thus, for example, the plot in the third row, second column shows the same two variables (plastics and paper) as the one in the third column, second row, but their axes are transposed.

Correlation Matrices

A ***correlation matrix*** is just an array (table) showing, in rows and columns, the correlations corresponding to the plots in the scatterplot matrix.

Example 13.4: Correlation Matrices

Here's the correlation matrix for the municipal waste data.

	Energy Content	Plastics	Paper	Garbage	Water
Energy Content	1.00	0.59	0.04	-0.09	-0.90
Plastics	0.59	1.00	-0.15	-0.09	-0.26
Paper	0.04	-0.15	1.00	-0.63	-0.01
Garbage	-0.09	-0.09	-0.63	1.00	0.07
Water	-0.90	-0.26	-0.01	0.07	1.00

The table entries are correlations, with row and column labels indicating the variables. For example, the rightmost entry in the top row, $r = -0.90$, is the correlation between energy content and percent water, and corresponds to the scatterplot in the same position of the scatterplot matrix of Example 13.3.

The entries along the diagonal are correlations between variables and themselves, and hence equal to 1.0. The matrix is symmetric about its diagonal, meaning the value in the ith row, jth column is the same as the one in the jth row, ith column, because the correlation between, say, energy content and water is the same as that between water and energy content.

Three-Dimensional Scatterplots

A useful way to plot values of a response variable versus *two* predictors is in a ***three-dimensional scatterplot***, where each of the two horizontal axes in a three-dimensional coordinate system correspond to a predictor and the vertical axis corresponds to the response. Unlike a scatterplot matrix, which only shows how the response variable is related to the predictors *one at a time*, a scatterplot matrix shows how it's related to them *simultaneously*.

Example 13.5: Three-Dimensional Scatterplot

Streams and lakes contain dissolved oxygen that supports fish and other aquatic life. But organic pollutants consume dissolved oxygen when they chemically degrade via oxidation.

The chemical oxygen demand (COD) of a water supply is the amount of oxygen that would be needed to chemically degrade the organic compounds contained in the water. It's used an indirect measure of organic pollution that includes chemicals, petroleum, and solvents.

In the study of pollutants in highway runoff in the Pear River Delta, South China described in Example 8.7 of Chapter 8, COD (mg/L) was measured in runoff for each of $n = 18$ rainfall events [2]. Also recorded were the rain depth (mm) and the length of the antecedent dry period (days). The data are below.

Chemical Oxygen Demand in Highway Runoff

Rainfall Event	Rainfall Date	COD	Rain Depth	Antecedent Dry Period
1	9/25/05	296	4.1	0.83
2	2/26/06	256	9.0	7.83
3	3/22/06	518	6.7	0.43
4	4/06/06	451	3.6	6.38
5	4/23/06	469	2.9	10.57
6	5/17/06	323	3.4	6.66
7	6/15/06	161	9.6	0.73
8	7/15/06	336	28.6	4.87
9	7/16/06	119	19.6	0.16
10	7/25/06	379	10.9	6.22
11	7/26/06	75	27.2	0.98
12	4/06/06	177	18.5	6.35
13	4/23/06	295	17.6	10.57
14	5/02/06	93	8.5	0.68
15	5/06/06	29	26.8	1.17
16	5/10/06	46	20.3	3.81
17	5/31/06	37	16.3	0.38
18	6/12/06	77	1.3	2.54

We'll consider COD to be the response variable and investigate its relationship to the two predictors, rain depth and antecedent dry period. Thus

Y = COD

X_1 = Rain depth
X_2 = Antecedent dry period

The scatterplot matrix of the data is shown below.

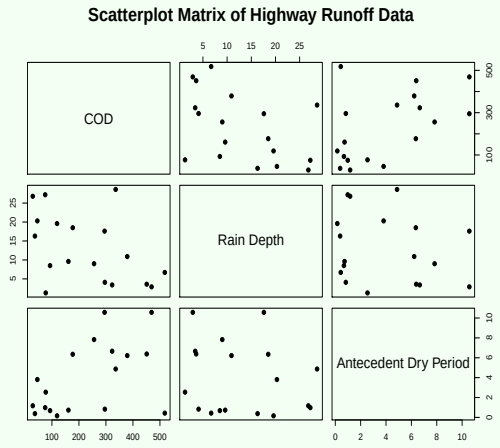


Figure 13.3: Scatterplot matrix of the variables in the highway runoff data set.

From the scatterplot matrix, we see that COD has a slight *negative* relationship to rain depth and a slight *positive* relationship to antecedent dry period. We also see that there's no relationship

between the two predictors, rain depth and antecedent dry period.

The scatterplot matrix shows the relationships between COD and the predictors *one predictor at a time*. The three-dimensional scatterplot below shows it's relationship to them *simultaneously*.

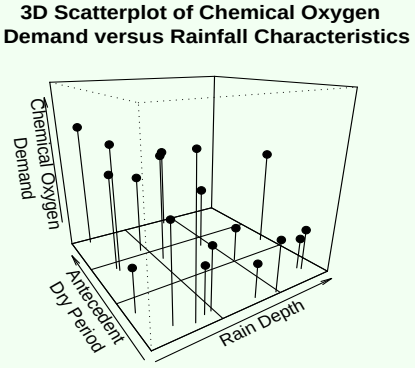


Figure 13.4: Three-dimensional scatterplot of the chemical oxygen demand in highway runoff versus rain depth and antecedent dry period.

We see from the plot that the COD tends to be highest when the rain depth is small and the antecedent dry period long. We can fit a plane to the data using the method of *least squares* described later in this chapter. This so-called *fitted regression plane* is shown in the plot below.

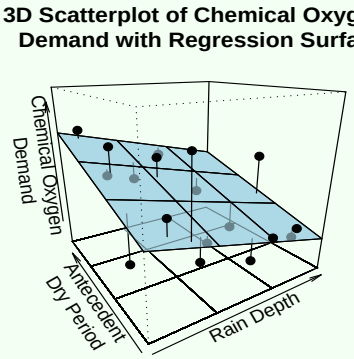


Figure 13.5: Regression plane fitted to the data on COD, rain depth, and antecedent dry period.

With the regression plane fitted to the data, the nature of the relationship between COD and the two predictors becomes clearer. The upward tilt of the plane in the direction of increasing antecedent dry period indicates that, for a fixed rain depth, COD tends to increase as the antecedent dry period increases. Likewise, the downward tilt in the direction of increasing rain depth indicates that, for a fixed antecedent dry period, COD tends to decrease as the rain depth increases. In Example 13.6 we'll see the actual equation for the regression plane, and we'll see how to use it to quantify *how much* the COD changes for given changes in rain depth and antecedent dry period.

13.2 The Multiple Linear Regression Model

As for simple linear regression (Chapter 12), when a response variable exhibits a relationship to predictor variables, there's usually an underlying natural process driving that observed pattern, but in the context of multiple regression, the underlying process can involve more than one predictor variable. For example, the observed relationship of COD to rain depth and antecedent dry period is driven by the dilution of COD as rain depth increases and the buildup of organic matter on surfaces during the antecedent dry period. As for simple linear regression, a common goal of multivariate studies is to draw inferences about the underlying process that's driving the pattern in the data, and to achieve this goal we'll employ a *statistical model* that describes the data via the underlying process and then *estimate* unknown constants in the model (*parameters*) using the data.

The model, called the ***multiple linear regression model***, is an extension of the simple linear regression model that allows for more than one predictor variable, and it captures the "overall pattern" in the data while also allowing for "deviations" away from that "overall pattern."

We'll first look at the model for the case in which there are only two predictors and later (Section 13.2.2) look at the more general case for which there may be more than two.

13.2.1 Multiple Linear Regression Model with Two Predictors

The Equation of a Plane

In a three-dimensional coordinate system, points are represented by the values of the three coordinates, X_1, X_2, and Y. A plane such as the one shown in Figure 13.5 can be represented by a function $Y = f(X_1, X_2)$ having the form

$$f(X_1, X_2) \;=\; a + b\,X_1 + c\,X_2, \tag{13.1}$$

where a, b, and c are constants that determine, respectively, the height at which the plane intersects the y-axis, the tilt of the plane in the direction of X_1, and tilt in the direction of X_2. As an example, the plane whose equation is

$$Y \;=\; 40 + 2.3\,X_1 + 1.4\,X_2 \tag{13.2}$$

is shown in Fig. 13.6. The height Y of this plane over any point on the "floor" of the coordinate system is obtained by plugging that point's X_1 and X_2 values into the right side of (13.2). For example, the height above the point $X_1 = 10$, $X_2 = 5$ is

$$\begin{aligned} Y &= 40 + 2.3\,(10) + 1.4\,(5) \\ &= 70, \end{aligned}$$

which is depicted by the dashed vertical line in Fig. 13.6.
By plugging in $X_1 = 0$ and $X_2 = 0$, we see that the plane intercepts the y-axis at $Y = 40$. Notice that the plane tilts upward in the direction of X_1 and upward in the direction of X_2, but the tilt is steeper in the X_1 direction. In fact, the gradient (slope) along any line in the X_1 direction (holding X_2 constant) is 2.3, the coefficient of X_1 in the equation of the plane, whereas the gradient in the X_2 direction is only 1.4, the coefficient of X_2. To see why, suppose we allow X_1 to vary but we fix the value of X_2 at, say, $X_2 = 5$. Then Y takes values on the line depicted in Fig. 13.7, and the equation of this line (after replacing X_2 by 5 in (13.2)) is

$$\begin{aligned} Y &= 40 + 2.3\,X_1 + 1.4\,(5) \\ &= 47 + 2.3\,X_1 \end{aligned}$$

which is seen to have slope 2.3. A similar argument shows that the gradient of the plane in the X_2 direction is 1.4.

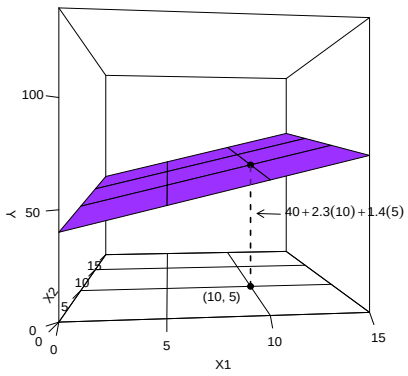


Figure 13.6: The plane whose equation is $Y = 40 + 2.3\,X_1 + 1.4\,X_2$.

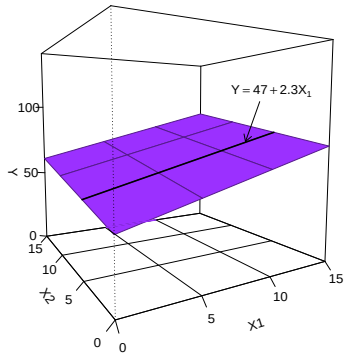


Figure 13.7: The plane whose equation is $Y = 40 + 2.3\,X_1 + 1.4\,X_2$ and a line on the plane in the X_1 direction (with X_2 fixed at 5).

The Multiple Regression Model with Two Predictors

For multivariate data with two predictors, the ***multiple linear regression model*** consists of a plane corresponding to the underlying natural process driving the "overall pattern" in the data and an error term corresponding to a "deviation" away from that "overall pattern."

Multiple Regression Model with Two Predictors:

$$Y_i \;=\; \beta_0 + \beta_1\,X_{1i} + \beta_2\,X_{2i} + \epsilon_i, \tag{13.3}$$

where

- Y_i is the observed value of the response variable for the ith individual ($i = 1, 2, \ldots, n$).
- X_{1i} is the observed value of the first predictor variable (X_1) for the ith individual.
- X_{2i} is the observed value of the second predictor variable (X_2) for the ith

individual.

$\boldsymbol{\beta_0}$ is the y-intercept of the ***true regression plane***.
$\boldsymbol{\beta_1}$ is the ***coefficient*** for X_1 in the ***true regression plane***.
$\boldsymbol{\beta_2}$ is the ***coefficient*** for X_2 in the ***true regression plane***.
ϵ_i is a random error term following a N(0, σ) distribution, and the ϵ_i's are independent of each other.

The model relates the response variable Y to the two predictors X_1 and X_2 by way of the so-called ***true regression plane***, $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$. The (unknown) ***parameters*** of the model are the ***coefficients*** β_0, β_1, and β_2, representing the y-intercept, gradient in the X_1 direction, and gradient in the X_2 direction, respectively, of the true regression plane and σ, the error distribution's standard deviation. In practice, their values will be estimated from the data.

As for simple linear regression, the random error ϵ represents the deviation of Y above or below the plane resulting from the net effect of all *other* factors, *besides* X_1 and X_2, as well as measurement error. A COD measurement in highway runoff will deviate above or below the plane due to factors such as traffic volume prior to the rain storm, types of vehicles using the highway, their conditions, and so on.

The standard deviation σ represents the size of a typical error. In the model, its value doesn't depend on X_1 or X_2, so the amount of variation of Y above or below the plane is assumed to be the same regardless of the values of the predictors. The model is depicted graphically in the figure below.

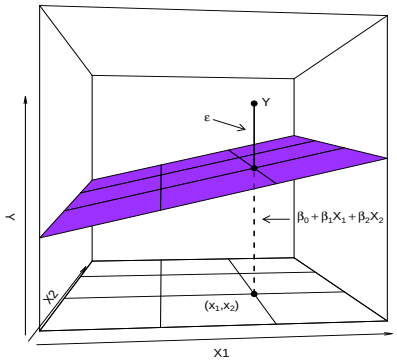


Figure 13.8: Graphical depiction of the multiple regression model with two predictors.

Comments:

- As for simple linear regression, the multiple regression model can be thought of as describing separate, distinct populations of Y values, one population for each combination of values of X_1 and X_2, where the population means all lie on the plane $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. See Fig. 12.11 in Chapter 12. In the study of COD in highway runoff, each population would correspond to rainstorm events of a given rain depth and antecedent dry period. The true regression plane $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ is sometimes called the ***true mean response plane***.

- The aforementioned Y populations are assumed to be *normal* and to all have the same standard deviation σ.

- No assumptions are made about the X_1 and X_2 variables. Their values don't even have to be randomly selected – they can be hand-picked, as would be the case if they were values of two explanatory variables in a designed experiment.

- The assumption of normality (and independence) of the ϵ's is needed for testing hypotheses about β_0, β_1, and β_2 and (constructing confidence intervals for them).

Interpretation of the Model Coefficients When There Are Two Predictors

When there are only two predictors in the model, the coefficient β_1 is the change in Y associated with a one-unit increase in X_1 *while holding* X_2 *constant*. Likewise, β_2 is the change in Y associated with a one-unit increase in X_2 *while holding* X_1 *constant*. The signs of the coefficients indicate whether Y tends to increase or decrease as the predictor increases.

The condition that *the other predictor is held constant* is especially relevant to investigating the influence of one predictor on the response while ***controlling*** (or "adjusting") for the effect of the other one. For instance, in Example 13.1, we saw that wealthier cities tend to use more water but also tend to be larger, so the higher water usage could be due to the larger population sizes. To investigate the relationship between the wealth and water usage while *controlling* for the population size, we need to include population size in the model along with wealth. In this case, the coefficient for wealth can be interpreted as the increase in water usage for each one-unit increase in wealth *while population size remains constant*. We'll revisit this in Example 13.8.

In practice, of course, the (unknown) model coefficients will have to be estimated from the data. We'll see how they're estimated in Section 13.3.

13.2.2 Multiple Linear Regression Model with p Predictors

For multivariate data with more than two predictors, we can no longer use a plane in our statistical model to represent the "overall pattern" in the data. But in the same way that we generalized from a line in two dimensions to a plane in three, we can generalize from a plane in three dimensions to a so-called *hyperplane* in more than three. The ***multiple linear regression model*** consists of a *hyperplane* corresponding to the underlying natural process driving the "overall pattern" in the data and an error term corresponding to a "deviation" away from that "overall pattern."

Multiple Linear Regression Model with p Predictors:

$$Y_i \;=\; \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i, \tag{13.4}$$

where

- Y_i is the observed value of the response variable for the ith individual ($i = 1, 2, \ldots, n$).
- $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the observed values of the p predictors for the ith individual.
- $\boldsymbol{\beta_0}$ is the y-intercept of the ***true regression mean response***.
- $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_p}$ are the ***coefficients*** for $X_1, X_2, \ldots, X_p$ in the ***true regression mean response***.
- ϵ_i is a random error term following a N(0, σ) distribution, and the ϵ_i's are independent of each other.

Although we can no longer visualize the model using a plane, the p-predictor version of the model has several features in common with the two-predictor version. The response variable Y is related to the p predictors via the so-called ***true regression mean response***, $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$. The (unknown) ***parameters*** of the model are the ***coefficients*** $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, representing the y-intercept, gradient in the X_1 direction, gradient in the X_2 direction, and so on, and σ, the error distribution's stan-

dard deviation. In practice, the parameters will be estimated from the data.

Comments:

- As for the two-predictor case, the multiple regression model can be thought of as describing separate, distinct populations of Y values, one for each combination of values of $X_1, X_2, \ldots, X_p$, where the population means all have the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.
- The aforementioned Y populations are assumed to be *normal* and to all have the same standard deviation σ.
- No assumptions are made about the variables $X_1, X_2, \ldots, X_p$. Their values don't even have to be randomly selected – they can be hand-picked, as would be the case if they were values of p explanatory variables in a designed experiment.
- The assumption of normality (and independence) of the ϵ's is needed for testing hypotheses about $\beta_0, \beta_1, \ldots, \beta_p$ and (constructing confidence intervals for them).

Interpretation of the Model Coefficients When There Are p Predictors

When there are more than two predictors, the model coefficients are no longer the y-intercept and slopes of a tilted plane. But other than, that they're interpreted just as they would be for the two-predictor case. In particular, for each $k = 1, 2, \ldots, p$, the coefficient β_k is the change in Y associated with a one-unit increase in X_k *while holding all of the other predictor variables constant.* The signs of the coefficients indicate whether Y tends to increase or decrease as the predictor increases.

As for the two-predictors case, the condition that *the other predictor variables are held constant* in the interpretation of the β_k's means we can investigate the influence of X_k on Y while simultaneously ***controlling*** (or "adjusting") for the effects of the other $p - 1$ predictors.

13.3 Least Squares Estimation of the Model Coefficients

We estimate the (unknown) model coefficients $\beta_0, \beta_1, \ldots, \beta_p$ from the data using the ***method of least squares*** introduced in Chapter 12. In the two-predictor case, this means finding a plane for which the *vertical* deviations of the points away from it in a three-dimensional scatterplot are small. The plane in Fig. 13.5 was fitted to the COD data using least squares. When there are more than two predictors, the idea is the same – we want the deviations away from the *hyperplane* to be small.

More formally, the ***least squares estimates*** of the (unknown) model coefficients, denoted $\boldsymbol{b_0, b_1, \ldots, b_p}$, give the smallest possible value for the sum of squared deviations

$$\sum_{i=1}^{n} [Y_i \; - \; (b_0 + b_1 \, X_{1i} + \cdots + b_p \, X_{pi})]^2 \, ,$$

which can be thought of as measuring how "close" the ***fitted multiple regression model***

$$\hat{Y} \;\; = \;\; b_0 + b_1 \, X_1 + \ldots + b_p \, X_p \tag{13.5}$$

comes to the observed Y values. The estimated coefficients are computed entirely from the data and are readily obtainable using statistical software.

13.4 Using the Fitted Regression Model

The fitted regression model (13.5) serves as the *estimate* of the true mean response $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. The symbol $\hat{Y}$ is used instead of Y to remind us that it's the equation of the *fitted* regression model, not the theoretical regression model (13.4). We can use the fitted model to predict values of the response Y and to draw conclusions about the relationship between the response and the predictors.

Using the Fitted Model to Predict Y: We can predict the response of an individual based on that individual's set of values of the predictor variables by plugging those values into the fitted regression equation.

Interpretation of the Estimated Coefficients: The estimated coefficients $b_1, b_2, \ldots, b_p$ have the same interpretation as the true (unknown) coefficients $\beta_1, \beta_2, \ldots, \beta_p$ except that they're *estimates*. Thus for each $k = 1, 2, \ldots, p$, the coefficient b_k is the *estimated* change in Y associated with a one-unit increase in X_k *while holding all of the other predictor variables constant*, that is while *controlling* for them. The signs of the coefficients indicate whether Y tends to increase or decrease as the predictor increases.

Examples 13.6 and 13.7 illustrate prediction and the interpretation of the estimated coefficients. Example 13.8 shows how we use multiple regression to *control* for the effect of one variable while investigating the effect of another.

Example 13.6: Using the Fitted Multiple Regression Model

For the highway runoff data of Example 13.5, the response variable is chemical oxygen demand (COD) and there are $p = 2$ predictors, rain depth and antecedent dry period. The theoretical regression model for the data is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where Y = COD, and X_1 = rain depth and X_2 = antecedent dry period. The model was fitted to the data using statistical software, resulting in estimated coefficients

$$b_0 = 241, \qquad b_1 = -6.87, \qquad \text{and} \qquad b_2 = 19.9,$$

and thus the equation of the fitted regression model is

$$\hat{Y} = 241 - 6.87\, X_1 + 19.9\, X_2.$$

This is the equation of the plane shown in Fig. 13.5. Based on its coefficients, we estimate that for each one-mm increase in rain depth (X_1), the COD *decreases* by 6.87 μg/L, on average, and for each additional antecedent dry day (X_2), the COD *increases* by 19.9 μg/L.

Based on the fitted model, if the rain depth is 10 mm and the antecedent dry period three days, we'd predict the COD concentration to be

$$\begin{aligned} \hat{Y} &= 241 - 6.87\,(10) + 19.9\,(3) \\ &= 232 \end{aligned}$$

μg/L. This is the height of the plane in Fig. 13.5 above the point (10, 3) on the "floor" of the coordinate system.

In the next example there are more than two predictors, so the fitted model is no longer a plane.

Example 13.7: Using the Fitted Multiple Regression Model

For the municipal waste data of Example 13.2, the theoretical regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon\,,$$

where

Y = Energy content (kcal/kg)

X_1 = Percent plastics by weight
X_2 = Percent paper by weight
X_3 = Percent garbage by weight
X_4 = Percent moisture by weight

The model was fitted to the data using statistical software, and the resulting estimated model coefficients are

b_0	=	2245	b_3	=	4.30
b_1	=	28.9	b_4	=	-37.4
b_2	=	7.64			

Thus the fitted model is

$$\hat{Y} = 2245 + 28.9\,X_1 + 7.64\,X_2 + 4.30\,X_3 - 37.4\,X_4.$$

Intuitively, because the energy content of waste refers to its potential to incinerate, it makes sense that b_1 is positive whereas b_4 is negative, since waste with a higher plastic content is more combustible, but waste with higher moisture content is less combustible. The values of b_1 and b_4 quantify *how much* the energy content changes as the plastic and moisture contents increase. For each one-percent increase in plastic, the energy increases by 28.9 kcal/kg, and for each one-percent increase in moisture, the energy *decreases* by 37.4 kcal/kg. By contrast, the values of b_2 and b_3 are relatively close to zero, indicating that paper and garbage don't have as much of an effect on energy as plastic and moisture do.

The predicted energy content for a waste consisting of 20% plastics, 20% paper, 35% garbage, and 45% moisture is

$$\begin{aligned}\hat{Y} &= 2245 + 28.9(20) + 7.64(20) + 4.30(35) - 37.4(45)\\ &= 1443.3\end{aligned}$$

kcal/kg.

The next example illustrates how multiple regression can be used to control for the effect of one variable while investigating the effect of another.

Example 13.8: Controlling for the Effect of a Variable

In Example 13.1 we saw that wealthier cities tend to use more water, but also tend to be larger, so the higher water usage may be due to their larger population size.

To determine the impact that wealth has on a city's water usage, while *controlling* for the size of the city, we can simply include city size in the regression model along with wealth, and then look at the estimated coefficient for wealth.

The theoretical regression model for the data is

$$Y \;=\; \beta_0 + \beta_1\, X_1 + \beta_2\, X_2 + \epsilon\,,$$

where

Y = Water usage (log millions of liters/day)

X_1 = Wealth (z-score of the city's median income)
X_2 = City size (population in millions)

The estimated model coefficients (from statistical software) are

$$b_0 \;=\; 6.48, \qquad b_1 \;=\; 0.11, \qquad \text{and} \qquad b_2 \;=\; 0.16,$$

so the fitted model is

$$\hat{Y} \;=\; 6.48 \,+\, 0.11\, X_1 \,+\, 0.16\, X_2.$$

A graph the fitted model as a plane in three-dimensional space is shown on the left below. Also shown, on the right, is a scatterplot matrix of the data.

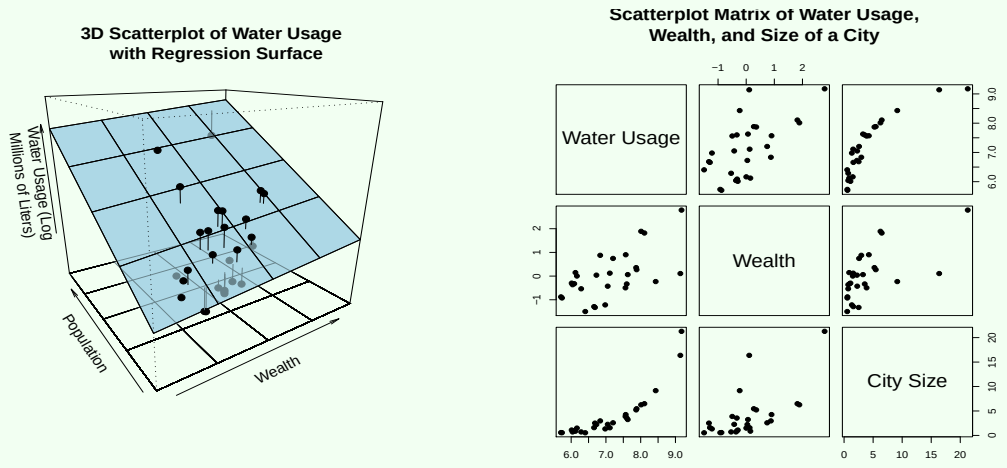


Figure 13.9: Regression plane fitted to data on water usage versus wealth and city size for 28 U.S. cities (left); scatterplot matrix the data (right).

The estimated coefficient for wealth indicates that a city's (log) water usage increases by 0.11 units, on average, for each one-unit increase in its wealth (z-score) *while holding city size constant*, that is, while *controlling* for the size of the city. Thus two same-sized cities whose wealths differ by one unit would be expected to differ in water use about 0.11 units.

By contrast, if we *don't* control for city size, and instead just fit a *simple* linear regression model with wealth as the (one) predictor, its estimated coefficient (from Example 13.1) is $b_1 = 0.58$, indicating for each one-unit increase in wealth, water usage increases by 0.58 units, more than five times as much as when we control for size!

13.5 Fitted Values and Residuals

The fitted multiple regression model provides an estimate of the true regression mean response. For each of the n individuals in a multivariate data set, we define the individual's ***fitted value*** (also called ***predicted values***), denoted $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ as follows.

Fitted Value: For the ith individual in the data set,

$$\hat{Y}_i \;=\; b_0 + b_1\, X_{1i} + b_2\, X_{2i} + \ldots + b_p\, X_{pi},$$

where $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the values of the p predictor variables for that individual.

The fitted values are the values we'd predict for Y by plugging each observed set of predictor values $X_{1i}, X_{2i}, \ldots, X_{pi}$ into the fitted regression model equation. There will be n fitted values, one for each individual in the data set. For the data on COD in highway runoff in Fig. 13.5, they're the points lying on the fitted regression plane from which the vertical deviations emanate.

It turns out that the average of the fitted values is equal to the average of the Y_i's.

Fact 13.1 The mean of the fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ from a fitted multiple regression model is equal to the mean of the observed responses $Y_1, Y_2, \ldots, Y_n$, that is,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \bar{Y}.$$

We'll also be interested in evaluating the random "deviations" away from the overall pattern, that is, values of the random error term ϵ in the model. A ***residual***, denoted $\boldsymbol{e_i}$, is defined as the difference between the ith individual's observed Y value and the fitted value for that individual.

Residual: For the ith individual in the data set,

$$e_i \;=\; Y_i - \hat{Y}_i,$$

where Y_i is the observed response for the ith individual and $\hat{Y}_i$ is the fitted value.

For the COD data, the residuals are the vertical line segments in Fig. 13.5. A residual will be positive if Y_i lies above the plane, and negative if it lies below the plane. There will be a total of n residuals, one for

each individual in the data set.

In the two-predictor case, the fact that the points in a three-dimensional scatterplot don't all lie on a plane tells us that other factors, besides the predictors $X_1, X_2, \ldots, X_p$, affect the value of Y. A residual is the net effect of these other factors. *The residual e_i approximates the random error term ϵ_i* in the model (13.4). In Section 13.10 we'll use the residuals to estimate the standard deviation of the $N(0, \sigma)$ error distribution, and in Section 13.16 we'll use them to check the normality assumption.

It can be shown that the residuals always sum to zero.

Fact 13.2 The residuals in a multiple regression analysis sum to zero, that is,

$$\sum_{i=1}^{n} e_i \;=\; 0.$$

13.6 Two Sources of Variation in Y

A response variable Y will be affected by numerous factors, not just the predictor variables that were chosen to be included in the regression model. In a multiple regression analysis, there will be two sources of variation in the responses:

1. Variation due to differences from one individual to the next in the set of values of the predictors $X_1, X_2, \ldots, X_p$.
2. Variation due to differences from one individual to the next in the values of all other factors (besides $X_1, X_2, \ldots, X_p$).

These correspond, respectively, to the nonrandom "overall pattern" in the data and the random "deviations" away from that pattern, that is, errors. In the study of the COD in highway runoff, the two sources of variation are, on the one hand, the two predictors rain depth and antecedent dry period, and on the other, all the other factors that affect COD (such as traffic volume, types of vehicles, their conditions, and so on).

13.7 Sums of Squares

Introduction

We'll measure the contributions of the two sources of variation in the response variable using sums of squares, as we did in the context of simple linear regression. We'll use sums of squares to:

1. Assess *how well* the multiple regression model fits the data.
2. Estimate the standard deviation σ of the error distribution.
3. Test a hypothesis to decide if *any* of the p predictor variables $X_1, X_2, \ldots, X_p$ explains some of the variation in the responses.

Variation Due to Error

Variation in Y due to random error, that is, due to all other factors besides $X_1, X_2, \ldots, X_p$, is measured by the ***error sum of squares***, denoted **SSE** and defined as follows.

Error Sum of Squares:

$$\text{SSE} \;=\; \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \;=\; \sum_{i=1}^{n} e_i^2\,.$$

The error sum of squares is just the ***sum of squared residuals***. The SSE will be large when the variation in Y due to random *error* is large.

Variation Due to the X_k's

Variation in Y that's due to differences from one individual to the next in the set of values of $X_1, X_2, \ldots, X_p$ is measured by the ***regression sum of squares***, denoted **SSR** and defined as follows.

Regression Sum of Squares:

$$\text{SSR} \;=\; \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.$$

The regression sum of squares is the sum of squared deviations of the fitted values away from the mean $\bar{Y}$ of the responses, so it reflects variation in the $\hat{Y}_i$'s. SSR reflects variation in the fitted values. When there are just two predictors, the fitted values lie on the fitted plane, and in this case SSR will be large when the plane has a *steep tilt*. In general, SSR will be large when the variation in Y due to the predictors $X_1, X_2, \ldots, X_p$ is large.

13.8 ANOVA-Like Partition of the Total Variation in Y

Introduction

It turns out, as will be seen, that in a multiple regression analysis, the two types of variation in Y, that due to the nonrandom relationship to the predictors and that due to random error, account for *all* of the variation in the responses.

Total Variation

The *total variation* in the response values is measured by the ***total sum of squares***, as usual, again denoted **SSTo** and defined as follows.

Total Sum of Squares:

$$\text{SSTo} \;=\; \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Because SSTo measures total variation in the responses, it will be large if either the predictors or the random error contribute a large degree of variation to Y. Thus SSTo reflects both variation due to $X_1, X_2, \ldots, X_p$ and due to error.

Partition of the Total Variation

It can be shown that the sums of squares in a multiple regression analysis satisfy the following ***ANOVA-like partition*** of the total variation in the responses.

Fact 13.3 The sums of squares defined above satisfy the following relation.

$$\text{SSTo} = \text{SSR} + \text{SSE}.$$

This decomposes the variation in the responses into components corresponding to the two sources:

$$\text{Total Variation} = \text{Variation Due to } X_1, X_2, \ldots, X_p + \text{Variation Due to Error}$$

13.9 Degrees of Freedom

Associated with each sum of squares is a corresponding degrees of freedom.

Degrees of Freedom: For multiple regression, the degrees of freedom are:

$$\begin{aligned} df \text{ for SSTo} &= n-1 \\ df \text{ for SSR} &= p \\ df \text{ for SSE} &= n-(p+1) \end{aligned}$$

There are $n-1$ degrees of freedom for SSTo because only $n-1$ of the deviations used to compute it are "free to vary". For SSR, the degrees of freedom is the number of predictors in the model, p, because (it can be shown) only p of the deviations used to compute SSR are "free to vary." The degrees of freedom for SSE is the sample size n minus the number of parameters $\beta_0, \beta_1, \ldots, \beta_p$ (including the intercept), $p+1$, because, it turns out, that's how many of the deviations used to compute SSE are "free to vary."

As was the case for ANOVA and simple linear regression, the degrees of freedom are additive.

Fact 13.4 The degrees of freedom given above satisfy the following relation.

$$df \text{ for SSTo} = df \text{ for SSR} + df \text{ for SSE}.$$

13.10 Mean Squares

Recall that a ***mean square*** is a sum of squares divided by its degrees of freedom. The two mean squares, the ***mean square for regression***, or **MSR**, and ***mean squared error***, or **MSE**, will be used later to test for a relationship between Y and $X_1, X_2, \ldots, X_p$.

Mean Squares: For multiple regression, the mean square for regression and mean squared error are

$$\begin{aligned} \text{MSR} &= \frac{\text{SSR}}{p} \\ \text{MSE} &= \frac{\text{SSE}}{n-(p+1)}. \end{aligned}$$

Estimating σ

The MSE is the "average" (using $n-(p+1)$) *squared* residual, so its square root measures the size of a typical residual. Thus, because the residuals are approximations of the random errors ϵ in the multiple re-

gression model, we use $\sqrt{\text{MSE}}$ as an estimator of the standard deviation σ of the N(0, σ) error distribution.

Estimator of $\boldsymbol{\sigma}$: In a multiple regression analysis, the estimator of σ, denoted $\hat{\boldsymbol{\sigma}}$, is

$$\hat{\sigma} \;=\; \sqrt{\text{MSE}}.$$

Example 13.14 illustrates the use of the square root of the MSE to estimate σ.

13.11 Assessing the Fit of the Regression Model

When reporting the results of a multiple regression analysis, it's difficult to show in a graph how well the model fits the data, especially when there are more than two predictors. For this reason, we often include in the report a statistic that measures how well the model fits the data.

In this section, we'll look at two such statistics:

1. The mean squared error, MSE, or its square root.
2. The coefficient of multiple determination, denoted R^2.

These statistics were also used in the context of simple linear regression in Chapter 12. In Section 13.19 we'll look at two other measures of model fit, the *adjusted* R^2 and the *Akaike information criterion*, that are used primarily for deciding which of several candidate predictor variables to include in a model.

13.11.1 The MSE as a Measure of Fit of the Model

When a model fits a data set well, the deviations away from the fitted model, or residuals, will be small. Because the mean squared error is the size of an "average" *squared* residual, and its square root represents the size of a *typical* residual, either one of these statistics can be used to assess how well the model fits. A *smaller* value of the MSE (or $\sqrt{\text{MSE}}$) indicates that the model fits the data *better*.

13.11.2 The Coefficient Of Multiple Determination R^2

A drawback to using the mean squared error (or its square root) to assess the fit of a model is that its value depends on the units of measure of the response variable. For example, changing Y from centimeters to inches will change the value of the MSE.

In a multiple regression analysis, we can avoid this problem by instead measuring the fit of the model using the ***coefficient of multiple determination***, denoted $\boldsymbol{R^2}$, defined as follows.

Coefficient of Multiple Determination:

$$R^2 \;=\; \frac{\text{SSR}}{\text{SSTo}} \;=\; 1 \,-\, \frac{\text{SSE}}{\text{SSTo}}\,. \tag{13.6}$$

Because SSR measures variation in the responses due to differences among individuals in their predictor values $X_1, X_2, \ldots, X_p$, and SSTo measures *total* variation in the responses, we can think of R^2 as

$$R^2 \;=\; \frac{\text{Variation in } Y \text{ Due to } X_1, X_2, \ldots, X_p}{\text{Total Variation in } Y}.$$

In other words, R^2 is the *proportion* of variation in the response variable that can be explained by differences among individuals in their predictor values $X_1, X_2, \ldots, X_p$.

Properties and Interpretation of R^2: The following properties of R^2 provide insight into its interpretation.

1. The value of R^2 will always be between zero and one (because it's a proportion).
2. R^2 tells us *how well* the regression model fits the data:
   - An R^2 value near zero means the model *doesn't* fit very well (because only a small fraction of the Y variation is explained by $X_1, X_2, \ldots, X_p$).
   - An R^2 value near one means the model fits the data very well (because a large fraction of the Y variation is explained by $X_1, X_2, \ldots, X_p$).

Example 13.9: The Coefficient of Multiple Determination R^2

For the municipal waste data of Example 13.2, with variables

Y = Energy content (kcal/kg)

X_1 = Percent plastics by weight
X_2 = Percent paper by weight
X_3 = Percent garbage by weight
X_4 = Percent moisture by weight

The fitted regression model (from Example 13.7) is

$$\hat{Y} = 2245 + 28.9\,X_1 + 7.64\,X_2 + 4.30\,X_3 - 37.4\,X_4.$$

The sums of squares (obtained using statistical software) are

$$\text{SSTo} = 689,710, \quad \text{SSR} = 664,931, \quad \text{and} \quad \text{SSE} = 24,779.$$

The value of R^2, using (13.6), is

$$R^2 = 1 - \frac{24,779}{689,710} = 0.964,$$

indicating that the model fits the data very well. In fact, it indicates that 96.4% of the variation in the energy content of waste is explained by its plastic, paper, garbage, and moisture contents, and only 3.6% by all other factors combined. Apparently the combustibility (energy) of waste is determined almost entirely by the four predictor variables used in the model.

Improving Model Fit by Adding Predictors

Recall that the errors ϵ in a regression model (and the residuals after fitting the model to the data) are due the net effects on Y of all other factors *besides* the predictor variables $X_1, X_2, \ldots, X_p$ that are in the model.

By adding another variable, X_{p+1} say, as a predictor to the model, X_{p+1} is no longer lumped in with all the other factors that contribute to the sizes of the errors (and residuals), so the errors (residuals) become smaller. Likewise, by removing a predictor, X_p say, from a model, X_p is now lumped in with all the other factors that contribute to the errors (residuals), which as a result become larger. This is stated formally in the following fact.

Fact 13.5 Removing a predictor variable from a regression model always results in a *larger* SSE and a *smaller* R^2. Adding a predictor variable to the model always results in a *smaller* SSE and a *larger* R^2.

Another way to interpret this is that the model that includes the additional predictor will explain more of the variation in Y, that is, will fit the data better. The following example illustrates.

Example 13.10: Improving Model Fit by Adding Predictors

For the data on water usage in U.S. cities given in Example 13.1, the fitted *simple* linear regression model, with Y = Water usage and X_1 = Wealth, is

$$\hat{Y} = 7.10 + 0.58X_1 .$$

It's plotted in the scatterplot of Fig. 13.1. The square root of the MSE and R^2 for this model, obtained using statistical software, are

$$\sqrt{\text{MSE}} = 0.778 \qquad \text{and} \qquad R^2 = 0.369,$$

so a typical residual is size 0.778 and only 36.9% of the variation in water usage is explained by wealth alone.

If we refit the model, but this time with both X_1 = Wealth and X_2 = City Size as predictors, the fitted model becomes

$$\hat{Y} = 6.48 + 0.11\,X_1 + 0.16\,X_2,$$

and the square root of the MSE and R^2 become

$$\sqrt{\text{MSE}} = 0.502 \qquad \text{and} \qquad R^2 = 0.747.$$

Now, with both wealth *and* city size in the model, a typical residual size is smaller, 0.502, and 74.7% of the variation in water usage is explained by these two predictors, much more than was explained by wealth alone.

13.12 *t* Tests for the Regression Model Coefficients

Each of the parameters $\beta_1, \beta_2, \ldots, \beta_p$ in the regression model (13.4) represents the change in Y, on average, associated with a one-unit increase in the corresponding predictor (while holding the other predictors constant). When a particular coefficient β_k is zero, Y has no relationship to the corresponding predictor X_k. We'll be interested, therefore, in testing the hypotheses

$$H_0 : \beta_k = 0 \tag{13.7}$$
$$H_a : \beta_k \neq 0. \tag{13.8}$$

for each $k = 1, 2, \ldots, p$. The null hypothesis says there's *no relationship* between Y and X_k, and the alternative says *there is a relationship*. Occasionally we'll also be interested in testing the corresponding hypotheses about the intercept β_0,

$$H_0 : \beta_0 = 0$$
$$H_a : \beta_0 \neq 0,$$

but this is usually of less interest.

Sampling Distribution of b_k

Because the estimate b_k of the true (unknown) coefficient β_k is computed from the data, and the response values $Y_1, Y_2, \ldots, Y_n$ vary from one sample to the next, b_k is a random variable that varies from sample to sample.

The hypothesis test for β_k will be based on how different b_k is from zero, so to carry out the test, we'll need to be able to distinguish between a difference from zero that's due to chance and one that's due to more than just chance. For this, we'll need the sampling distribution of b_k.

Fact 13.6 Suppose we have multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and follow a $\mathrm{N}(0, \sigma)$ distribution.

Then for each $k = 0, 1, \ldots, p$, b_k follows a *normal* distribution with mean β_k and standard error denoted $\boldsymbol{\sigma_{b_k}}$, which is to say,

$$b_k \sim \mathrm{N}(\beta_k,\ \sigma_{b_k}).$$

An expression for σ_{b_k} in terms of the error standard deviation σ and the observed values of the predictors $X_1, X_2, \ldots, X_p$ can be found in many books on multiple regression, including [5].

It follows that if we standardize b_k, the resulting random variable Z follows a standard normal distribution, that is,

$$Z = \frac{b_k - \beta_k}{\sigma_{b_k}} \sim \mathrm{N}(0,\ 1).$$

t Test Statistics for Model Coefficients

In practice, we have to estimate the standard error σ_{b_k} of b_k. We'll denote the ***estimated standard error*** by $\boldsymbol{S_{b_k}}$. Details about how S_{b_k} is computed can be found in books on multiple regression such as [5]. In practice, we obtain its value using statistical software.

According to the next fact, when we standardize b_k using the *estimated* standard error, the resulting standardized variable follows a t distribution. This fact will be used to develop the t test procedures for the coefficients.

Fact 13.7 Suppose we have multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and follow a $\mathrm{N}(0, \sigma)$ distribution.

Then for each $k = 0, 1, \ldots, p$,

$$\frac{b_k - \beta_k}{S_{b_k}} \sim t(n - (p+1)),$$

the t distribution with $n - (p+1)$ degrees of freedom.

The ***t test statistic for a coefficient***, denoted $\boldsymbol{t}$, is obtained by replacing β_k in Fact 13.7 by it's null hypothesized value zero.

t Test Statistic for a Coefficient:

$$t = \frac{b_k - 0}{S_{b_k}}. \tag{13.9}$$

Because b_k is an estimator of the true coefficient β_k, if H_0 was true, and β_k equal to zero, we'd expect b_k to be close to zero, in which case t would be close to zero too. But if H_a was true, we'd expect b_k to differ from zero in the direction specified by H_a, in which case t would differ from zero in that direction too. Therefore,

1. *Large positive* values of t provide evidence in favor of $H_a : \beta_k > 0$.
2. *Large negative* values of t provide evidence in favor of $H_a : \beta_k < 0$.
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \beta_k \neq 0$.

Furthermore, t measures (approximately) how many standard errors the estimate b_k is away from zero, and in what direction (positive or negative). To decide if an observed value of t provides statistically significant evidence against the null hypothesis, we'll need its sampling distribution under H_0, which, from Fact 13.7, is the following.

Sampling Distribution of t Under H_0: Suppose we have multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and follow a $\mathrm{N}(0, \sigma)$ distribution.

Then when

$$H_0 : \beta_k \ = \ 0$$

is true,

$$t \ \sim \ t(n - (p + 1)).$$

The t Test Procedure for a Model Coefficient

P-values and critical values (for the rejection region approach) for the t test for a coefficient are obtained from the tails of the $t(n - (p + 1))$ distribution, as summarized below.

t Test for β_k

Assumptions: Data consist of multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and either they follow a $N(0, \sigma)$ distribution or n is large.

Null hypothesis: $H_0 : \beta_k = 0$ (for any one of $k = 0, 1, \ldots, p$)

Test statistic value: $t = \frac{b_k}{S_{b_k}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t distribution with $n-(p+1)$ d.f.:	Rejection region = t values such that:*
$H_a : \beta_k > 0$	to the right of t	$t > t_{\alpha,n-(p+1)}$
$H_a : \beta_k < 0$	to the left of t	$t < -t_{\alpha,n-(p+1)}$
$H_a : \beta_k \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-(p+1)}$ or $t < -t_{\alpha/2,n-(p+1)}$

* $t_{\alpha,n-(p+1)}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-(p+1)$ d.f.

Note: Statistical software packages always report the results of the *two-sided* test of

$$\begin{aligned} H_0 : \beta_k &= 0 \\ H_a : \beta_k &\neq 0 \end{aligned}$$

when they perform regression analyses. To carry out a *one-sided* test, when the observed t value differs from zero in the direction specified by H_a, we divide the reported p-value by two.

Carrying out the t Test for a Coefficient

When a multiple regression analysis is carried out using statistical software, the results of the tests for coefficients are summarized in a table of the form below.

Predictor	Estimated Coefficent	Standard Error	t	P-value
Intercept	b_0	S_{b_0}	$t = b_0/S_{b_0}$	p
X_1	b_1	S_{b_1}	$t = b_1/S_{b_1}$	p
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_p	b_p	S_{b_p}	$t = b_p/S_{b_p}$	p

Example 13.11: t Test for a Coefficient

For the highway runoff data of Example 13.5, with COD as the response variable and rain depth and antecedent dry period as predictors, statistical software reports the following output:

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	240.65	71.26	3.377	0.0042
Rain Depth	-6.87	3.59	-1.913	0.0751
Antecedent Dry Period	19.93	9.08	2.195	0.0443

From the output, the equation of the fitted model is

$$\hat{Y} = 241 - 6.87\,X_1 + 19.9\,X_2.$$

To test whether the relationship of COD to rain depth is statistically significant, the hypotheses are

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

The value of the test statistic (13.9) is $t = -1.913$ and the p-value is 0.0751. Using a level of significance $\alpha = 0.05$, we fail reject H_0 and conclude that the relationship isn't statistically significant.

To test whether COD is related to the antecedent dry period, the hypotheses are

$$\begin{aligned} H_0 : \beta_2 &= 0 \\ H_a : \beta_2 &\neq 0 \end{aligned}$$

and the observed test statistic value is $t = 2.195$ with p-value 0.0443. Using $\alpha = 0.05$, we reject H_0 and conclude that the relationship between COD and antecedent dry period is statistically significant. Furthermore, because the coefficient is positive, we can conclude that longer antecedent dry periods are associated with higher COD concentrations in runoff.

Example 13.12: t Test for a Coefficient

For the municipal waste data of Example 13.2, with energy content as the response and percent plastics, paper, garbage, and water as the predictors, statistical software reports the following:

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	2245.1	177.9	12.62	0.000
Plastics	28.922	2.823	10.24	0.000
Paper	7.643	2.314	3.30	0.003
Garbage	4.297	1.916	2.24	0.034
Water	-37.356	1.834	-20.37	0.000

Thus the equation of the fitted model is

$$\hat{Y} = 2245 + 28.9X_1 + 7.64X_2 + 4.30X_3 - 37.4X_4\,.$$

Using a level of significance $\alpha = 0.05$, the p-values indicate that all four predictors have a statistically significant relationship to energy content. The first three (plastics, paper, and garbage) have positive coefficients, meaning that higher amounts of these variables result in higher energy contents. The fourth (water) has a negative coefficient, so higher amounts of water are associated with lower energy contents.

13.13 t Confidence Intervals for the Regression Model Coefficients

A **$100(1-\alpha)\%$ *confidence interval for* β_k**, the true (unknown) coefficient for X_k in the regression model, is given, for each $k = 0, 1, \ldots, p$, by

Confidence Interval for a Model Coefficient:

$$b_k \pm t_{\alpha/2,n-(p+1)} S_{b_k},$$

where b_k is the least squares estimate of β_k, $t_{\alpha/2,n-(p+1)}$ is the $100(1-\alpha/2)$th percentile of the t distribution with $n-(p+1)$ degrees of freedom, and S_{b_k} is the estimated standard error of b_k (obtained using software).

We can be $100(1-\alpha)\%$ confident that the true (unknown) coefficient β_k will be contained in this confidence interval somewhere.

Example 13.13: t Confidence Interval for β_k

For the study of COD in $n = 18$ highway runoff specimens, Example 13.11 gives the regression analysis output from statistical software. From the output, the estimated coefficient for antecedent dry period and its (estimated) standard error are

$$b_2 = 19.93 \qquad \text{and} \qquad S_{b_2} = 9.08.$$

From a t distribution table, using $n-(p+1) = 15$ degrees of freedom, the critical value for a 95% confidence interval for β_2 is $t_{\alpha/2,n-(p+1)} = 2.131$. Thus the 95% confidence interval is

$$\begin{aligned} 19.93 \pm 2.131(9.08) &= 19.93 \pm 19.35 \\ &= (0.58,\ 39.28). \end{aligned}$$

We can be 95% confident that β_2, the true change in COD, on average, for each one-day increase in the antecedent dry period, is between 0.58 and 39.28 mg/L.

13.14 Regression Model F Test

The t tests of Section 13.12 are used to test *one at a time* whether each predictor is related to the response variable. It's sometime useful to test *all at once* whether *any* of the predictors are related to the response. We do this using the ***regression model F test***. The null and alternative hypotheses are

$$\begin{aligned} &H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ &H_a\text{: Not all } \beta_k\text{'s equal 0} \end{aligned} \tag{13.10}$$

The null hypothesis says that *none* of the predictors $X_1, X_2, \ldots, X_p$ are related to Y. In other words, it says that the *entire model* is of no use for explaining variation in Y. The alternative says that *at least one* of the predictors *is* related to Y, but doesn't specify which ones. In other words, it says that the model is useful for explaining Y variation.

The ***overall regression model F test statistic*** is

***F* Test Statistic for the Regression Model**:

$$F = \frac{\text{MSR}}{\text{MSE}}. \tag{13.11}$$

The numerator measures variation in Y due to the predictor variables $X_1, X_2, \ldots, X_p$ and will be large when one or more of the predictors is related to Y. Thus F will be large when the variation in Y due to $X_1, X_2, \ldots, X_p$ is large relative to the variation due to random error. It follows that

Large values of F provide evidence *against* H_0 in favor of H_a.

To decide if an observed value of F is large enough to provide statistically significant evidence against the null hypothesis, we'll need its sampling distribution under H_0.

Sampling Distribution of *F* Under H_0: Suppose we have multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and follow a $\text{N}(0, \sigma)$ distribution. Then when

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

is true,

$$F \sim F(p,\, n - (p+1)),$$

the F distribution with numerator degrees of freedom p and denominator degrees of freedom $n - (p+1)$.

P-values (and critical values for the rejection region approach) are obtained from the *right* tail of the $F(p, n-(p+1))$ distribution.

The F test procedure is summarized below.

Regression Model *F* Test for $\beta_1, \beta_2, \ldots, \beta_p$

Assumptions: Data consist of multivariate observations described by the multiple regression model (13.4), where the ϵ_i's are independent and either they follow a $\text{N}(0, \sigma)$ distribution or n is large.

Null hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

Test statistic value: $F = \dfrac{\text{MSR}}{\text{MSE}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or F is in rejection region.

Alternative hypothesis	P-value = area under F-distribution with p and $n-(p+1)$ d.f.:	Rejection region = F values such that:*
H_a : Not all β_k's equal 0	to the right of F	$F \geq F_{\alpha,p,n-(p+1)}$

* $F_{\alpha,p,n-(p+1)}$ is the $100(1-\alpha)$th percentile of the F distribution with p and $n-(p+1)$ d.f.

13.15 The Multiple Regression ANOVA Table

The degrees of freedom, sums of squares, mean squares, observed F test statistic value, and p-value from a multiple regression analysis are usually summarized in ***multiple regression ANOVA table*** having the form shown below.

Source	DF	SS	MS	F	P-value
Regression	p	SSR	$\text{MSR} = \text{SSR}/p$	$F = \text{MSR}/\text{MSE}$	p
Error	$n-(p+1)$	SSE	$\text{MSE} = \text{SSE}/(n-(p+1))$		
Total	$n-1$	SSTo			

Example 13.14: Regression ANOVA Table and Model *F* Test

For the highway runoff data of Example 13.5, with COD as the response and rain depth and antecedent dry period as predictors, the multiple regression ANOVA table produced by statistical software is below.

Source	DF	SS	MS	F	P-value
Regression	2	179,843	89,922	5.228	0.0189
Error	15	258,006	17,200		
Total	17	437,849			

For the regression model F test of

$$H_0\text{: } \beta_1 = \beta_2 = 0$$
$$H_a\text{: } \beta_1 \text{ and } \beta_2 \text{ don't both equal 0,}$$

the observed test statistic value is $F = 5.228$, and the p-value is 0.0189. Using a level of significance $\alpha = 0.05$, we reject H_0 and conclude that at least one of the two predictors, rain depth and antecedent dry period, is related to COD.

Also, the error sum of squares and mean squared error are $\text{SSE} = 258,006$ and $\text{MSE} = 17,200$, respectively, so a typical deviation of a COD concentration above or below the fitted regression plane in Fig. 13.5 is $\sqrt{\text{MSE}} = \sqrt{17,200} = 131.1$ mg/L, and this is our estimate of the standard deviation σ of the $\text{N}(0,\sigma)$ error distribution in the regression model.

13.16 Using Residuals to Check the *t* and *F* Test Assumptions

The t tests for the coefficients of the regression model and the F test for the overall regression model rely on three assumptions:

1. The errors ϵ_i in the regression model follow a normal distribution.
2. The the standard deviation σ of the error distribution doesn't change with the values of the predictor variables.
3. The responses Y_i are independent of each other, or equivalently, the errors ϵ_i are independent.

The third assumption (independence) is usually addressed in the study design by separating observations sufficiently in space and time. The other assumptions (normality and common σ) are checked via plots of the residuals.

Checking the Normality Assumption

To check the normality assumption, we look at a normal probability plot or a histogram of the residuals.

Example 13.15: Checking Assumptions

For the data on COD in highway runoff (Example 13.5), a normal probability plot and histogram of the residuals, obtained using statistical software after fitting the model, are below.

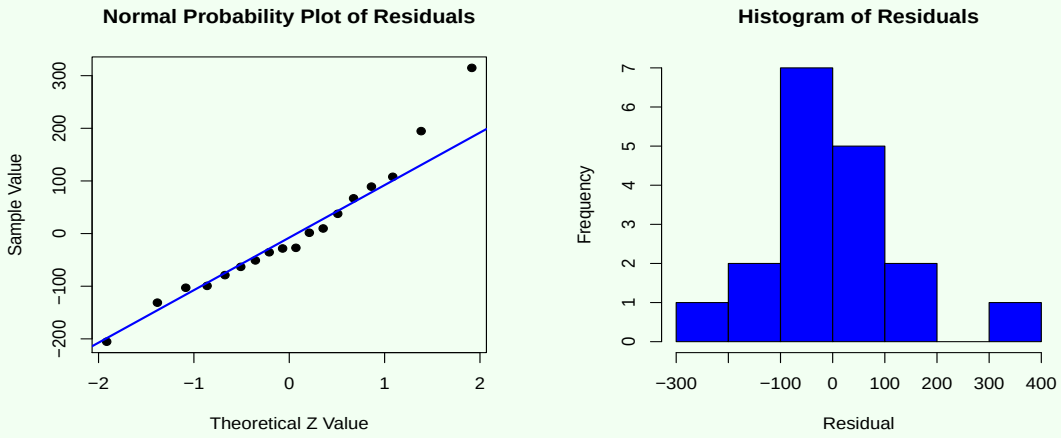


Figure 13.10: Normal probability plot (left) and histogram (right) of the residuals for the data on COD in highway runoff.

Both plots indicate that the assumption of normality of the errors appears to be met.

Checking the Constant σ Assumption

There are a few ways to check the assumption that the error standard deviation σ doesn't change with the values of $X_1, X_2, \ldots, X_p$.

1. **Plot the residuals versus each of the predictor variables $X_1, X_2, \ldots, X_p$**: We can look at plots of the residuals versus the values of each predictor variable X_k, with a horizontal line at $y = 0$. The amount of vertical spread above and below the line should be roughly the same from left to right, and in particular, it shouldn't increase (or decrease) as X_k increases.
2. **Plot the residuals versus the fitted values**: We can look at a plot of the residuals versus the fitted values, with a horizontal line at $y = 0$. The amount of vertical spread above and below the line should be roughly the same from left to right, and in particular, it shouldn't increase with the fitted value.

The idea behind the second plot is that it's not uncommon for the *variation* in the values of a response variable to be larger when their *mean* is larger. Monthly elevations of the Nile River, for example, vary more than those of the little creek that runs through the park in your neighborhood. The weights of adult blue whales vary more than the weights of a given anchovy species. Because the fitted values are the estimated mean responses for given values of the predictors, if the variation in the responses increases with the mean, it would show up in the plot as an increasing amount of vertical spread from left to right. See Fig. 13.11

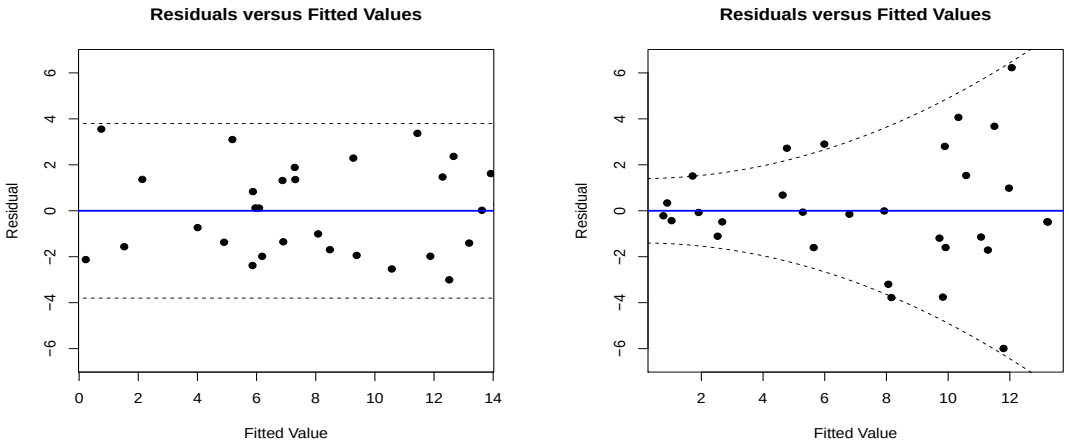


Figure 13.11: Plots of residuals versus fitted values for which the constant standard deviation assumption is met (left) and not met (right). Dashed lines are merely to enhance the appearances of the plots.

Example 13.16: Checking Assumptions

For the data on COD in highway runoff, a plot of the residuals versus the fitted values is below.

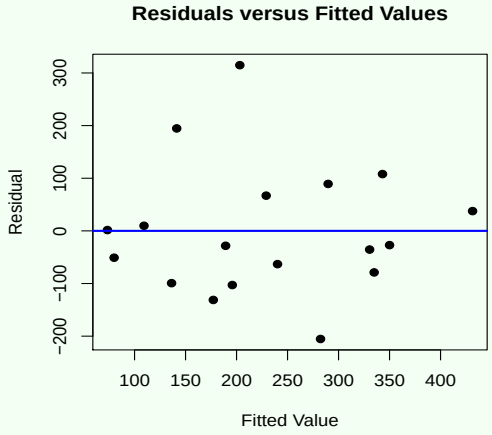


Figure 13.12: Plot of the residuals versus the fitted values for the data on COD in highway runoff.

The amount of vertical spread of the points away from the horizontal line is roughly constant as we move from left to right in the plot, and in particular the spread isn't systematically changing. Therefore the assumption that the error standard deviation σ is constant (doesn't depend on the values of the predictors) appears to be met. Thus, since normality was confirmed in Example 13.15, the t and F test results of Examples 13.11 and 13.14 are valid.

13.17 Dealing With a Non-Constant Standard Deviation: Transformations

When the constant standard deviation assumption isn't met, that is, when σ changes with the values of one or more of the predictors $X_1, X_2, \ldots, X_p$, it's sometimes possible to stabilize it by transforming the Y observations The most commonly used transformation for this purpose is the log transformation of the Y_i's, but other transformations in the Ladder of Powers might also be used.

13.18 Multicollinearity and Its Consequences

Introduction

When two or more of the predictor variables in a multiple regression analysis are correlated with each other, we say that there's ***multicollinearity*** among them.

When two variables are correlated, the value of one provides information about the value of the other. A snake's length, for example, provides information about how much the snake weighs. The age of a tree hints at how large its diameter might be. This implies that when two correlated predictors are both included in a multiple regression model, there's some redundancy in terms of the information they provide about the Y variable. If the predictors are highly correlated, it may be pointless to include both of them in the model – one or the other might do just about as well.

One way to decide if just one or the other of two predictors does about as well both in terms of explaining Y variation is to look at the change in R^2 after adding the second predictor to a model that already includes the first. We know that the R^2 will go up (Section 13.11.2), but if it doesn't go up by very much, that tells us that the second predictor doesn't contribute very much *additional* information toward explaining Y variation, and therefore can be left out of the model.

Example 13.17: Multicollinearity

Incineration is one method for disposing of household and industrial waste. But incineration emits pollutants such as dioxin, which can lead to human health problems such as damage to the immune system.

In an experiment to determine whether adjustments to the incineration process can reduce dioxin emissions, waste was incinerated under 18 different operating conditions corresponding to settings on the incinerator, and the dioxin at the outlet was measured under each condition [3].

The response variable was

Y = Dioxin at the outlet (ng-TE/Nm^3)

and the variables that determined the operating conditions were

X_1 = Furnace bed temperature (°C)
X_2 = Furnace top temperature (°C)
X_3 = Oxygen (O_2, %)
X_4 = Secondary/primary air ratio
X_5 = Total air supply (Nm^3/min)

The resulting data and a scatterplot matrix of it are below.

Dioxin in Incinerator Emissions

Experimental Run	Dioxin	Furnace Bed Temperature	Furnace Top Temperature	O_2	Secondary/Primary Air Ratio	Total Air Supply
1	13.70	618	845	11.3	1.48	497
2	29.53	609	783	12.1	1.63	525
3	14.98	597	746	13.2	1.65	530
4	18.15	589	858	12.5	1.79	535
5	8.89	583	876	12.5	1.77	530
6	6.61	604	874	12.7	1.76	525
7	20.12	612	858	12.5	1.47	493
8	7.08	603	864	11.9	2.04	560
9	11.49	601	866	11.6	2.31	604
10	19.34	603	890	12.1	1.78	509
11	19.4	603	890	12.1	1.78	509
12	5.20	601	915	12.8	1.64	485
13	11.00	605	840	12.1	1.98	546
14	14.00	603	861	11.8	1.82	521
15	25.78	600	930	10.8	1.23	446
16	26.38	600	930	10.8	1.23	446
17	30.41	593	961	11.0	1.50	462
18	31.27	596	948	11.5	1.39	442

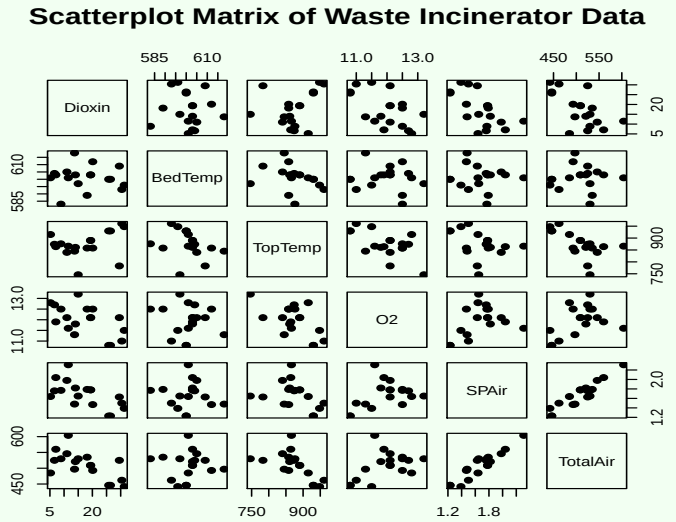


Figure 13.13: Scatterplot matrix of the variables in the waste incinerator data set.

Two of the predictors, total air supply and secondary/primary air ratio, stand out as being highly correlated. Primary air refers to air supplied from below the burning waste, secondary air to that supplied from above it, and total air to the sum of the primary and secondary air supplies.

Because they're correlated, we'd expect that if we add secondary/primary ratio to a regression model that *already* includes total air (along with the other three predictors), it won't contribute very much toward the model's ability to explain dioxin emissions variation.

In fact, when we add secondary/primary air ratio to the model, the R^2 only increases from 0.605

to 0.620, so the percentage of variation in dioxin emissions that's explained by the model only increases from 60.5% to 62.0%.

If our goal is to develop a simple model for dioxin emissions, we could leave secondary/primary ratio out of the model.

In the last example, because secondary/primary ratio and total air were correlated, we found that it wasn't useful to add secondary/primary ratio to a model that already includes total air. If we'd done it the other way around, that is, added total air to the model that already included secondary/primary ratio, we'd have found that it isn't useful to add total air to the model. This raises the question, *which* of the two predictors should be included in the model? We'll see in Section 13.19 some methods for dealing with this ambiguity.

Some Consequences of Multicollinearity

The ambiguity as to which predictors to include in a model and which ones to leave is just one of several consequences of multicollinearity that include the following.

Consequences of Multicollinearity: Multicollinearity among predictors can lead to the following:

- Estimates of model coefficients may differ depending on which other predictors are included in the model. See Examples 13.18 and 13.19.
- Results of t tests for coefficients may differ depending on which other predictors are included in the model. See Examples 13.18 and 13.19.
- The standard errors of coefficient estimates may be *very large*. See Example 13.18.
- The regression model F test may indicate that *at least one* predictor is related to Y even though the t tests don't indicate that any of them are. See Example 13.20.

We'll look at these more closely. The first two somewhat annoying ones imply that our conclusions about the relationship between a given predictor and the response can be different depending on which other predictors are included in the model. The third means our coefficient estimates can be unreliable. The next example illustrates.

Example 13.18: Consequences of Multicollinearity

Continuing from the previous example on waste incinerator dioxin emissions, when we fit the model *without* secondary/primary ratio included we get the following coefficient estimates and their standard errors and t test results.

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	391.230	169.291	2.311	0.0379
Bed Temp	-0.248	0.208	-1.191	0.2551
Top Temp	-0.082	0.043	-1.891	0.0812
O2	-7.085	2.858	-2.479	0.0277
Total Air	**-0.135**	**0.045**	**-2.996**	**0.0103**

The coefficient for total air indicates that the dioxin emissions decrease by an estimated 0.135 units

for each one-unit increase in total air (holding the other predictors constant), with a standard error of 0.045, and this observed relationship is statistically significant (t = -2.996, p-value = 0.0103).

If we fit the model again, but this time *with* secondary/primary ratio included in the model, we get the following.

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	479.251	215.401	2.225	0.0460
Bed Temp	-0.269	0.215	-1.252	0.2344
Top Temp	-0.118	0.068	-1.720	0.1111
O2	-7.806	3.102	-2.516	0.0271
Total Air	**-0.262**	**0.191**	**-1.370**	**0.1957**
SP Ratio	17.322	25.292	0.685	0.5064

Now the estimated decrease in dioxin emissions for each one-unit increase in total air (holding the other predictors constant) is 0.262 units, with a much larger standard error (0.191). As a result of the large standard error, the t test statistic is closer to zero (t = -1.370) and the observed relationship is no longer statistically significant (p-value = 0.1957).

The reason why estimated coefficients can change upon adding another predictor to the model is that those coefficients indicate the change in the response variable associated with a one-unit increase in the predictor *while controlling for the other predictors* that are in the model. For variables *not* in the model, no such control is imposed. What this means is that the coefficient for a predictor in the model reflects the change in the response associated with a one-unit change in the predictor as well as associated *simultaneous* changes in all other variables not included in the model.

The reason why a t test statistic can change when another predictor is added to the model is that both its numerator (the coefficient) and denominator (the standard error) can change.

The next example illustrates just *how much* our conclusions about a predictor can change when we add another one to the model and reminds us of the importance of *controlling* for the effects of confounding variables by including them in the model.

Example 13.19: Controlling for the Effects of Confounding Variables

Consider again the study of the relationship between a city's water usage and its wealth described in Example 13.1. Recall that wealthier cities tend to use more water, but also tend to be larger, so the effects of wealth and city size on water usage are *confounded*.

In Example 13.8, we saw that the coefficient for wealth changes when we control for city size (by adding city size to the model). We'll see now that the hypothesis test results change too.

When we fit a model with wealth as the only predictor, ignoring city size, we get:

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	7.10	0.147	48.266	0.0000
Wealth	0.58	0.150	3.896	0.0006

and we conclude that wealthier cities use statistically significantly more water (b_1 = 0.58, p-value = 0.0006).

But if we control for city size (by including it in the model), we get:

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	6.48	0.139	46.507	0.0000
Wealth	0.11	0.124	0.872	0.3910
City Size	0.16	0.026	6.118	0.0000

and wealth is no longer significant ($b_1 = 0.11$, p-value $= 0.3910$).

We can conclude that the perceived effect of wealth on a city's water usage in the first analysis is due at least in part to the fact that wealthier cities tend to be larger.

Another consequence of multicollinearity that we occasionally encounter is a regression model F test indicating that at least one predictor is related to Y even though the t tests don't indicate that any of them are. These results don't necessarily contradict each other. They tell us that when the predictors in the model are allowed to vary *together*, they explain variation in Y, but when only one is allowed to vary at a time, *while the others are held constant*, it's either unrelated to Y or its relationship is too weak to detect.

This is most easily understood when there are only two predictors as in the next example.

Example 13.20: Consequences of Multicollinearity

Continuing with the study on waste incinerator dioxin emissions (Example 13.18), fitting a model with just the two predictors total air and secondary/primary ratio gives:

Predictor	Estimated Coefficient	Standard Error	t	P-value
Intercept	69.236	31.671	2.186	0.0451
Total Air	-0.072	0.112	-0.645	0.5289
SP Ratio	-8.894	17.345	-0.513	0.6156

and neither of the individual predictors is statistically significant. This says that if the value of one of the two predictors is changed but not the other, the effect on dioxin emissions isn't discernible.

It turns out, though, that the regression model F test statistic is $F = 5.258$ and the p-value is 0.0186, which tells us that when total air and secondary/primary ratio *both* change, there's a discernible effect on dioxin emissions.

The three-dimensional scatterplot and fitted regression plane are shown below.

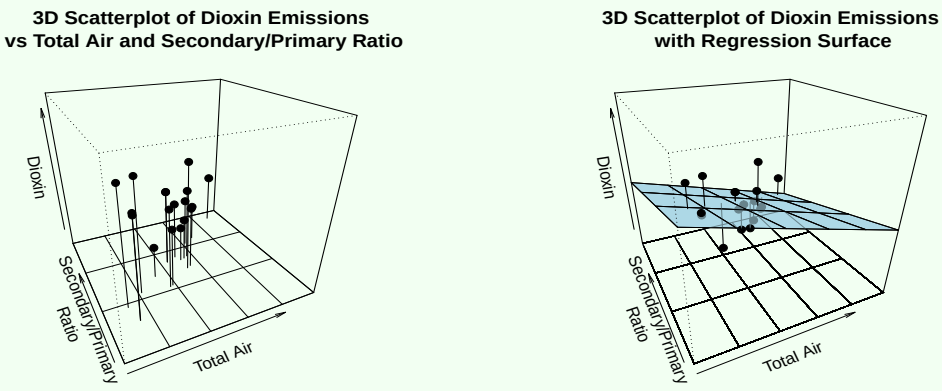


Figure 13.14: Three-dimensional scatterplot of dioxin emissions versus total air and secondary/primary ratio (left); regression plane fitted to the data (right).

The plane's steepest gradient (tilt) is in the direction in which total air *and* secondary/primary ratio *both* increase, which is consistent with the F test result. Its gradients in the directions in which one predictor increases but not the other are less steep, which helps explain the t test results.

Recall that when the standard error of an estimated coefficient is large, the estimate is considered to be imprecise. To see why the standard errors can be large in the presence of multicollinearity, consider the extreme (and rare) case in which there are just two predictors that are *perfectly* correlated. Thus there's a straight-line relationship between X_1 and X_2 as seen in the three-dimensional scatterplot on the left side of Fig. 13.15. In this case, infinitely many regression planes, two of which are shown in the right plot of Fig. 13.15, would fit the data equally well, and the fitted plane would be entirely free to "wobble."

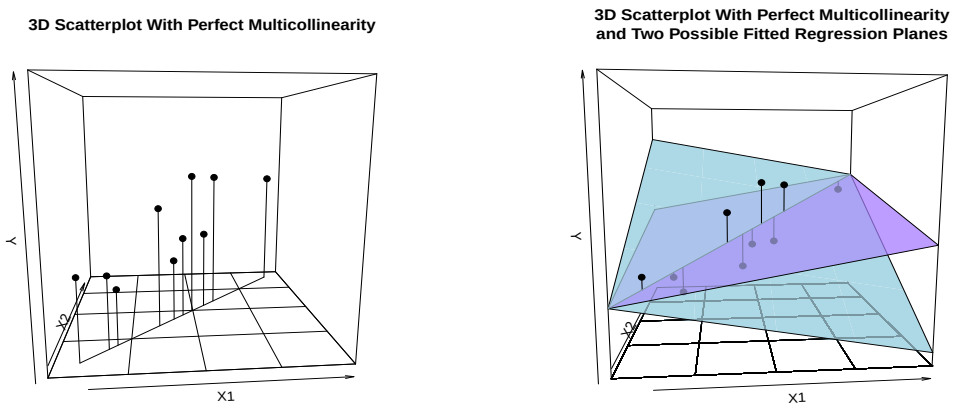


Figure 13.15: Three-dimensional scatterplot of data with perfect multicollinearity (correlation) between the two predictors (left); two possible regression planes, both fitted to data using the method of least squares (right).

In practice, it's rare for two predictors to be *perfectly* correlated, so there won't be infinitely many planes that fit equally well. But if the correlation is strong, there will be infinitely many that all fit the data *almost* as well, and as the Y values vary from one sample to the next, the fitted regression plane will "wobble." This variability from sample to sample in the tilt of the fitted plane results in large standard errors of the estimated coefficients.

Assessing Whether or Not Multicollinearity is Present

Scatterplot matrices and the aforementioned consequences of multicollinearity can be used to *assess* whether multicollinearity is present.

Assessing Whether Multicollinearity is Present: The following are indications that multicollinearity among predictors exists:

- Predictors appear related to each other in the scatterplot matrix.
- Estimates of model coefficients differ depending on which other predictors are included in the model.
- Results of t tests for coefficients differ depending on which other predictors are included in the model.
- The standard errors of coefficient estimates are unexpectedly *large*.
- The regression model F test indicates that *at least one* predictor is related to Y even though the t tests don't indicate that any of them are.

If we need a statistic that measures the severity of the multicollinearity in a given set of data, we can use the so-called ***variance inflation factor***. Details can be found in [5].

Remedying the Problems Associated with Multicollinearity

To remedy the problems associated with multicollinearity, we simply leave one or more of the correlated predictors out of the model. Often this can be done without substantially sacrificing how well the model fits the data. See Example 13.17. The next section describes ways of deciding *which* predictors should be omitted from the model.

13.19 Variable Selection

13.19.1 Introduction

It's not uncommon for a data set to contain as many as 20 or more variables that could potentially be used as predictors in a regression model. We've seen (Section 13.11.2) that including more predictors in a model results in a model that explains more of the variation in the response variable – the R^2 goes up. But models that contain too many predictors can be difficult to interpret and cumbersome to use, so we usually prefer to limit the number of predictors to just a handful that we think are important and that produce a model that fits the data well. This means there are two goals when deciding which predictors to include in a model.

Two Goals of Selecting Predictors for a Model: When deciding which predictors to include in a regression model and which ones to leave out, there are two goals:

- The model should fit the data well and explain a large proportion of the Y variation.
- The model should contain only a small number of predictors (for example, 2 - 5).

The two goals aren't compatible with each other – to satisfy the first one, we could add more predictors to the model, but that would violate the second one.

Deciding which variables to include in a model can be a challenge. To help us, a few so-called ***variable selection procedures*** have been developed. Here are three of them:

1. All-subsets variable selection
2. Best subsets variable selection
3. Stepwise variable selection (three versions):
   - Forward selection
   - Backward elimination
   - Bidirectional search

These procedures are entirely automated, and are carried out by statistical software. The choice of which one to use in a given situation will depend largely on how many variables are available to choose from. If there are fewer than, say, five or six, the all-subsets procedure should be used. If there are about six to 20, best subsets should be used. If there more than about 20, one of the stepwise procedures should be used.

13.19.2 Variable Selection Criteria

Variable selection procedures employ a statistic called a ***variable selection criterion*** to decide which of two or more models is most appropriate for a given set of data. We'll look at two commonly used variable selection criteria:

1. The adjusted R^2, denoted R^2_{adj}.
2. Akaike information criterion, or AIC.

The Adjusted R^2

Because the usual R^2 goes up every time we add another predictor to a model, it wouldn't make sense to use R^2 to compare the suitabilities of two models for a given set of data if one of the models contains more predictors than the other and one of our goals is to limit the number of predictors in the model.

What we need is a statistic whose value reflects *both* goals of selecting predictors – a good fitting model *and* a limited number of predictors. One such statistic is the ***adjusted* R^2**, denoted R^2_{adj} and defined as follows.

Adjusted R^2:

$$R^2_{\text{adj}} \;=\; 1 - \frac{\text{SSE}/(n-(p+1))}{\text{SSTo}/(n-1)} \;=\; 1 - \frac{\text{MSE}}{S_y^2}.$$

where S_y^2 is the sample variance of $Y_1, Y_2, \ldots, Y_n$.

Comparing R^2_{adj} to the (unadjusted) R^2 in (13.6), we see that the sums of squares in R^2_{adj} are divided by their degrees of freedom, whereas in (13.6) they aren't. Two things happen to R^2_{adj} when we add a predictor to a model:

- The SSE decreases, reflecting the improved fit of the model. This tends to make R^2_{adj} bigger.
- The degrees of freedom for SSE, $n-(p+1)$, decreases (because p goes up by one). This tends to make R^2_{adj} smaller.

The result is that the value of R^2_{adj} can go up *or* down when we add another predictor to the model. If adding a predictor results in a *larger* R^2_{adj}, we take it to mean that the gain in model fit is worth the trade-off of having the extra predictor in the model. But if adding the predictor results in a *smaller* R^2_{adj}, it tells us that the gain in model fit isn't enough to justify putting the additional predictor in the model.

Using R^2_{adj} to Select Predictors for a Regression Model:

To decide which of two or more models is most appropriate for a given set of data, we choose the one that has a higher R^2_{adj} value.

Note that unlike the (unadjusted) R^2, R^2_{adj} isn't a proportion, and it's value doesn't necessarily lie between zero and one. In fact, it's possible for R^2_{adj} to take negative values.

Example 13.21: The Adjusted R^2

Consider the two models fitted to the water usage data, one of which contains only one predictor, wealth, and the other two predictors, wealth and city size, given in Example 13.10.

For the model containing just wealth, the adjusted R^2 (from statistical software) is

$$R^2_{\text{adj}} \;=\; 0.344.$$

For the model with both city size and wealth, the adjusted R^2 is

$$R^2_{\text{adj}} \;=\; 0.727.$$

Based on the R^2_{adj} values, we should include city size in the model along with wealth.

Akaike Information Criterion

The ***Akaike information criterion***, denoted **AIC**, is another commonly used criterion for variable selection. It's defined as follows.

Akaike Information Criterion AIC:

$$\text{AIC} \;=\; n\log(\text{SSE}) \;-\; n\log(n) \;+\; 2p\,.$$

For AIC:

1. The second term $n\log(n)$ is a *constant* the sense that it doesn't depend on which predictors, nor how many, are in the model.
2. The first term $n\log(\text{SSE})$ will be *small* if the model *fits the data well*.
3. The last term $2p$ will be *small* if the *number of predictors in the model, p, is small*. It acts as a penalty for including too many predictors in the model.

Thus a model that fits the data well *and* contains only a limited number of predictors will have a *small* AIC value.

Using AIC to Select Predictors for a Regression Model:

To decide which of two or more models is most appropriate for a given set of data, we choose the one that has a lower AIC value.

13.19.3 All-Subsets Variable Selection

If only a small number of variables are available for inclusion in a regression model, we can decide which ones to include by fitting *all* the possible models and then choosing the one that gives the largest R^2_{adj} (or smallest AIC) value. This is called an ***all-subsets variable selection procedure***.

Example 13.22: All-Subsets Variable Selection

Recall that in the study of the energy content of waste (Example 13.2), the response variable is

Y = Energy content (kcal/kg)

and the variables that are available for use as predictors in a regression model are

X_1 = Percent plastics by weight
X_2 = Percent paper by weight
X_3 = Percent garbage by weight
X_4 = Percent moisture by weight

There are 15 possible models, listed below along with their R^2_{adj} and AIC values (obtained using software).

Variables in the Model	R^2_{adj}	AIC
X_1	0.32	292.5
X_2	-0.03	305.2
X_3	-0.03	305.1
X_4	0.81	255.1
X_1, X_2	0.32	293.7
X_1, X_3	0.30	294.4
X_1, X_4	0.94	218.4
X_2, X_3	-0.07	307.0
X_2, X_4	0.80	256.9
X_3, X_4	0.80	257.0
X_1, X_2, X_3	0.30	295.5
X_1, X_2, X_4	0.95	215.0
X_1, X_3, X_4	0.94	220.4
X_2, X_3, X_4	0.79	258.9
X_1, X_2, X_3, X_4	0.96	211.5

The model with all four predictors has the highest R^2_{adj} value (0.96) and the lowest AIC value (211.5), so this model is deemed best according to both criteria.

If there are a total of p variables available for inclusion in a model, there will be a total of $2^p - 1$ possible models that will need to be fitted (not including the one with no predictors) for the all-subsets procedure. To see why, notice that there are two options for each variable: either include it in the model or leave it out. Thus, since there are p variables, there are $2 \cdot 2 \cdots 2 = 2^p$ models (including the one without any predictors). Subtracting one discounts the model with no predictors.

13.19.4 Best-Subsets Variable Selection

If the number of variables available for inclusion in a regression model is large, the number of models that would need to be evaluated for the all-subsets procedure can be overwhelming. In ***best-subsets variable selection procedures***, the software only reports the "best" few models for each possible number of predictors. The user is free to choose how many "best" models the software reports for each model size, and is usually given the choice as to whether R^2_{adj} or AIC is used deem a model among the "best."

Example 13.23: Best-Subsets Variable Selection

Continuing from the previous example, if we specify to statistical software that we want the two best models according to the R^2_{adj} criterion for each possible number of predictors that the model could contain, the software reports the following.

	Variables in the Model	R^2_{adj}
The two best one-variable models:	X_4	0.81
	X_1	0.32
The two best two-variable models:	X_1, X_4	0.94
	X_2, X_4	0.80
The two best three-variable models:	X_1, X_2, X_4	0.95
	X_1, X_3, X_4	0.94
The only four-variable model:	X_1, X_2, X_3, X_4	0.96

We see that the best two-variable model ($R^2_{\text{adj}} = 0.94$) and the best three-variable model ($R^2_{\text{adj}} = 0.95$)

are almost as good as the four-variable model ($R^2_{\text{adj}} = 0.96$). Note that X_1 (plastics) and X_4 (moisture) show up in all three of these models, and are therefore apparently are the most important predictors of energy content.

13.19.5 Stepwise Variable Selection

In ***stepwise variable selection procedures***, a *sequence* of models is fitted the data, each one obtained from the previous one by adding or removing a variable so as to improve the suitability of the model for the data. The procedures terminate when neither adding nor dropping a variable from the current model will lead to any improvement. Although they're not guaranteed to produce the best possible model (unlike the all-subsets and best-subsets procedures), they'll generally produce a good one.

We'll look at three stepwise procedures, the *forward selection procedure*, the *backward elimination procedure*, and the *bidirectional search procedure*. In each case, we'll use the R^2_{adj} to decide which variable, if any, to add or drop from the model. In practice, the AIC could be used instead.

Forward Selection Stepwise Procedure

In the ***forward selection stepwise procedure***, we start with a model that doesn't contain any predictor variables (it just has an intercept), and add variables one at a time as long as doing so increases the model's R^2_{adj} (or AIC) value. At each step, the variable added to the model is the one that gives the biggest increase in R^2_{adj} (or decrease in AIC). The process terminates when none of the remaining variables not yet in the model can improve the R^2_{adj} (or AIC) value.

Example 13.24: Forward Selection Procedure

To illustrate the forward selection procedure, we'll use the data from the study of dioxin emissions from waste incineration given in Example 13.17.

Recall that the response variable is

Y = Dioxin at the outlet

and the variables available for use as predictor in a model are

X_1 = Furnace bed temperature
X_2 = Furnace top temperature
X_3 = Oxygen
X_4 = Secondary/primary air ratio
X_5 = Total air supply

Recall also (from Example 13.18) that if we fit the model with all five predictors, not all of them are statistically significant. This suggests that they don't all need to be included in the model.

We'll use the forward selection procedure to decide which variables to include in the model. Statistical software carries out the procedure and organizes the results in the following manner.

	Current Model	Possible Actions	R^2_{adj}
Start:	Dioxin = <none>		**0.000**
		+ Total Air	0.364
		+ SP Ratio	0.357
		+ O2	0.269
		+ Top Temp	0.035
		+ Bed Temp	-0.060
Step 1:	Dioxin = Total Air		**0.364**
		+ O2	0.423
		+ Top Temp	0.335
		+ SP Ratio	0.334
		+ Bed Temp	0.323
Step 2:	Dioxin = Total Air + O2		**0.423**
		+ Top Temp	0.468
		+ SP Ratio	0.408
		+ Bed Temp	0.388
Step 3:	Dioxin = Total Air + O2		
	+ Top Temp		**0.468**
		+ Bed Temp	0.484
		+ SP Ratio	0.437
Step 4:	Dioxin = Total Air + O2		
	+ Top Temp		
	+ Bed Temp		**0.484**
		+ SP Ratio	0.461
Final Model:	Dioxin = Total Air + O2		
	+ Top Temp		
	+ Bed Temp		**0.484**

Starting from the model containing no predictors (just an intercept), the variables that are available for adding to the model are listed along with the R^2_{adj} values for the models that would result. The model at each step was obtained by adding to the previous model the variable that raises the R^2_{adj} the most. The final model is the one for which no other variable additions can increase the R^2_{adj}.

In this case, the final model includes all the variables except secondary/primary ratio. It should come as no surprise that either secondary/primary ratio or total air would be left out of the model because they're so highly correlated (Example 13.17)

Backward Elimination Stepwise Procedure

The ***backward elimination stepwise procedure*** is similar to the forward selection procedure except that it starts with the *full* model (containing *all* the variables that are available) and *drops* variables one at a time as long as doing so increases the model's R^2_{adj} (or AIC) value. At each step, the variable dropped from the model is the one whose elimination leads to the largest increase in R^2_{adj} (or decrease in AIC). The

process terminates when none of the variables still remaining in the model can be dropped to improve the R^2_{adj} (or AIC) value.

Example 13.25: Backward Elimination Procedure

Using the dioxin emissions data from the last example, the backward elimination procedure, carried out by statistical software, leads to the following.

	Current Model	Possible Actions	R^2_{adj}
Start:	Dioxin = Bed Temp + Top Temp + O2 + SP Ratio + Total Air		**0.461**
		- SP Ratio	0.484
		- Bed Temp	0.437
		- Total Air	0.424
		- Top Temp	0.379
		- O2	0.240
Step 1:	Dioxin = Bed Temp + Top Temp + O2 + Total Air		**0.484**
		- Bed Temp	0.468
		- Top Temp	0.388
		- O2	0.293
		- Total Air	0.187
Final Model:	Dioxin = Bed Temp + Top Temp + O2 + Total Air		**0.484**

Starting from the full model with all five predictors, the variables that could be dropped from the model are listed along with the R^2_{adj} values for the models that would result. The model at each step was obtained from the previous model by dropping the variable that raises the R^2_{adj} the most. The final model is the one for which no other variable removals can increase the R^2_{adj}.

As seen above, secondary/primary ratio is dropped in the first step because it's the only variable among the five whose removal increases R^2_{adj} (from 0.461 to 0.484). No further steps are taken because after refitting the model with just the four remaining predictors, dropping any one of them would increase the R^2_{adj}, not lower it.

In the last example, the final model selected by backward elimination was the same as the one obtained by forward selection in Example 13.24. In practice, though, there's no guarantee that the two procedures will lead to the same model.

Bidirectional Search Stepwise Procedure

In forward selection, once a variable is added to the model, it can't be removed in a later step. Likewise in backward elimination, once a variable is removed, it can't be added back later. In the ***bidirectional search stepwise procedure***, variables added to the model in one step are candidates for removal in later steps, and variables removed in one step are candidates for being added back in later.

The procedure starts with either the model containing no predictors (just an intercept) or the full model (containing all available predictors). At each step either a variable not already in the model is added or one in the model is removed, whichever action results in the largest increase in R^2_{adj} (or decrease in AIC).

The process terminates when neither adding nor dropping a variable from the model can give a bigger R^2_{adj} (or smaller AIC).

The next example illustrates the procedure starting from the model with no predictors.

Example 13.26: Bidirectional Search Procedure

Continuing with the study of dioxin emissions from the last example, the bidirectional search procedure, carried out by statistical software, leads to the following.

	Current Model	Possible Actions	R^2_{adj}
Start:	Dioxin = <none>		**0.000**
		+ TotalAir	0.364
		+ SPAir	0.357
		+ O2	0.269
		+ TopTemp	0.035
		+ BedTemp	-0.060
Step 1:	Dioxin = Total Air		**0.364**
		+ O2	0.423
		+ TopTemp	0.335
		+ SPAir	0.334
		+ BedTemp	0.323
		- TotalAir	0.000
Step 2:	Dioxin = Total Air + O2		**0.423**
		+ TopTemp	0.468
		+ SPAir	0.408
		+ BedTemp	0.388
		- O2	0.364
		- TotalAir	0.269
Step 3:	Dioxin = Total Air + O2 + TopTemp		**0.468**
		+ BedTemp	0.484
		+ SPAir	0.437
		- TopTemp	0.423
		- O2	0.335
		- TotalAir	0.222

Step 4:	Dioxin = Total Air + O2		
	+ TopTemp		
	+ BedTemp		**0.484**
		- BedTemp	0.468
		+ SPAir	0.461
		- TopTemp	0.388
		- O2	0.293
		- TotalAir	0.187
Final Model:	Dioxin = Total Air + O2		
	+ TopTemp		
	+ BedTemp		**0.484**

Starting from the model with no predictors (just an intercept), the variables that could be added are listed along with the resulting R^2_{adj} values. At each step, the model was obtained from the previous one by adding or dropping the variable that raised the R^2_{adj} the most, and the variables available to be added or dropped from the current model are listed along with the resulting R^2_{adj} values. The final model is the one for which no other variable additions or removals can increase the R^2_{adj}.

In this case, at each step another predictor was added to the model, which is to say no variables were ever removed.

The final model in the previous example turned out to be the same as the ones obtained by forward selection and backward elimination (Examples 13.24 and 13.25). In practice, though, there's no guarantee that the three procedures will lead to the same model.

Comment: Because there's no guarantee that a stepwise variable selection procedure will produce the "best" model, a better approach to variable selection is to use the stepwise procedure to suggest *how many* predictors to include in the model, but not necessarily *which* ones. Then, once the number of predictors has been determined, we choose from among *all* models that have *that many predictors* the one whose R^2_{adj} is highest (or AIC is lowest), for example by using the best-subsets procedure.

13.20 Problems

13.1 If we add another predictor to a regression model, will the R^2 necessarily increase? What about R^2_{adj}? If the R^2_{adj}, what would that indicate about the predictor that was added to the model?

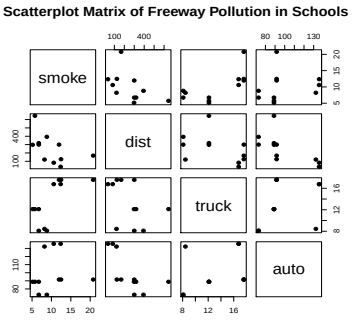
13.2 In a study of the effects of freeway air pollution on the respiratory health of children in nearby schools, black smoke and nitrogen dioxide (NO_2) were measured in the air inside 11 schools in the Province of South Holland, Netherlands [13].

The table below shows, for each school, the indoor black smoke and NO_2 concentrations ($\mu g/m^3$), the distance of the school from a freeway (m), and the truck and automobile traffic volume (thousands of vehicles per day) of the freeway.

Freeway Pollution in Schools

School	Black Smoke	NO_2	Distance from Freeway	Truck Traffic Volume	Automobile Traffic Volume
1	8.81	9.2	393	8.10	72.88
2	6.73	13.0	300	8.10	72.88
3	8.21	23.6	121	8.44	132.29
4	6.73	14.8	318	12.12	88.87
5	5.15	9.2	298	12.12	88.87
6	5.74	14.7	645	12.12	88.87
7	12.37	32.8	35	16.77	135.69
8	10.59	27.7	83	16.77	135.69
9	20.78	30.0	168	17.58	91.61
10	12.47	22.0	125	17.58	91.61
11	11.97	21.5	300	17.58	91.61

In this problem, we'll analyze the black smoke data. A scatterplot matrix of the black smoke, distance from freeway, and truck and automobile volumes is below.



a) Carry out a multiple regression analysis, with black smoke as the response and distance from freeway, truck volume, and automobile volume as predictors. Give the equation of the fitted regression model.

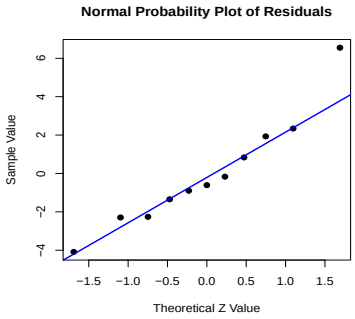
b) Based on the fitted regression model, by how much does the black smoke concentration decrease, on average, for each one-meter increase in distance from the freeway (while holding the other predictors constant)?

c) Based on the fitted regression model, by how much does the black smoke concentration increase, on average, for each one-thousand-vehicle increase in truck traffic volume (while holding the other predictors constant)?

d) Based on the fitted regression model, by how much does the black smoke concentration decrease, on average, for each one-thousand-vehicle increase in automobile traffic volume (while holding the other predictors constant)?

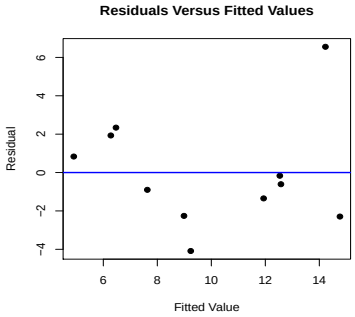
e) Based on the regression analysis of part *a*, which (if any) of the three predictors exhibit a statistically significant relationship to black smoke? Use a level of significance $\alpha = 0.05$.

f) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t tests, that the error term ϵ is normally distributed appear to be met?

g) A plot of the residuals versus the fitted values is below.



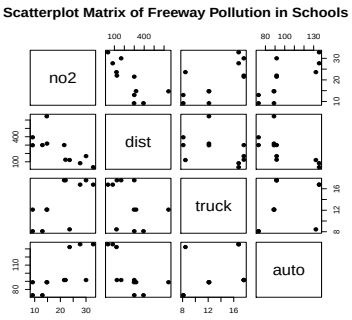
Based on the plot, does the assumption, required by the t tests, that the standard deviation σ of the error distribution is the same for different values of the predictors appear to be met?

h) Assuming that σ is the same for the different values of the predictors, what's the estimated value of σ?

i) Compute the coefficient of multiple determination R^2. What proportion of the variation in black smoke can be explained by the three predictors distance from freeway, truck volume, and automobile volume?

13.3 Refer to the study of the effects of freeway air pollution on the respiratory health of children in nearby schools described in Problem 13.2.

In this problem, we'll analyze the NO_2 data. A scatterplot matrix of the NO_2, distance from freeway, and truck and automobile volumes is below.



a) Carry out a multiple regression analysis, with NO_2 as the response and distance from freeway, truck volume, and automobile volume as predictors. Give the equation of the fitted regression model.

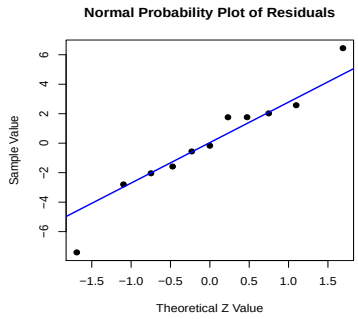
b) Based on the fitted regression model, by how much does the NO_2 concentration decrease, on average, for each one-meter increase in distance from the freeway (while holding the other predictors constant)?

c) Based on the fitted regression model, by how much does the NO_2 concentration increase, on average, for each one-thousand-vehicle increase in truck traffic volume (while holding the other predictors constant)?

d) Based on the fitted regression model, by how much does the NO_2 concentration increase, on average, for each one-thousand-vehicle increase in automobile traffic volume (while holding the other predictors constant)?

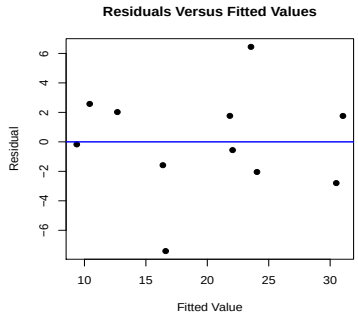
e) Based on the regression analysis of part *a*, which (if any) of the three predictors exhibit a statistically significant relationship to NO_2? Use a level of significance $\alpha = 0.05$.

f) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t tests, that the error term ϵ is normally distributed appear to be met?

g) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the t tests, that the standard deviation σ of the error distribution is the same for different values of the predictors appear to be met?

h) Assuming that σ is the same for the different values of the predictors, what's the estimated value of σ?

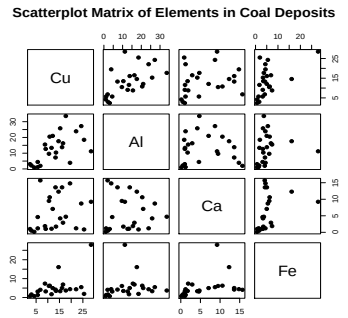
i) Compute the coefficient of multiple determination R^2. What proportion of the variation in NO_2 can be explained by the three predictors distance from freeway, truck volume, and automobile volume?

13.4 Trace elements and three major elements were measured in coal samples from 24 worldwide deposits to identify relationships between these elements [15]. The table below gives the data on the trace element copper (Cu, μg/g) and the major elements aluminum, calcium, and iron (Al, Ca, and Fe, all mg/g).

Elements in Coal Deposits

Coal Specimen	Country	Cu	Al	Ca	Fe
1	S. Africa	12.1	21.00	7.08	5.00
2	S. Africa	10.8	20.50	10.60	6.28
3	USA	5.7	4.44	4.62	3.17
4	USA	13.5	10.40	1.73	3.48
5	Japan	14.6	17.60	12.30	16.10
6	India	5.7	1.17	1.20	1.34
7	China	24.3	27.20	8.71	5.49
8	Colombia	4.2	0.96	0.15	0.72
9	India	9.1	12.70	1.03	3.23
10	China	17.7	33.60	4.72	3.71
11	Australia	16.5	16.30	2.91	6.88
12	China	6.8	1.98	15.70	4.10
13	Australia	22.1	23.90	1.21	4.38
14	China	15.3	25.80	4.21	3.34
15	India	3.0	2.18	0.51	1.62
16	India	2.4	3.08	1.01	0.69
17	Australia	25.5	18.50	0.87	1.93
18	Australia	11.6	13.40	1.09	1.74
19	India	8.8	15.50	1.88	7.37
20	China	10.5	9.62	9.49	6.11
21	Australia	16.1	13.80	13.60	5.04
22	USA	28.5	11.20	9.24	27.90
23	USA	13.3	7.26	13.60	4.21
24	USA	19.5	3.90	14.80	4.43

We'll investigate the relationship between the trace element Cu and the three major elements Al, Ca, and Fe. A scatterplot matrix of these variables is below.



a) Perform a multiple regression analysis with Cu as the response and Al, Ca, and Fe as predictors. Give the equation of the fitted regression model.

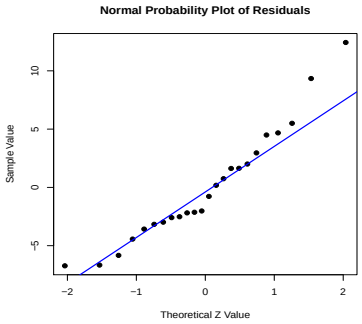
b) Based on the fitted regression model, by how many μg/g and does the Cu concentration increase, on average, for each one-mg/g increase in Al (while holding the other predictors constant)?

c) Based on the fitted regression model, by how many μg/g and does the Cu concentration increase, on average, for each one-mg/g increase in Ca (while holding the other predictors constant)?

d) Based on the fitted regression model, by how many μg/g and does the Cu concentration increase, on average, for each one-mg/g increase in Fe (while holding the other predictors constant)?

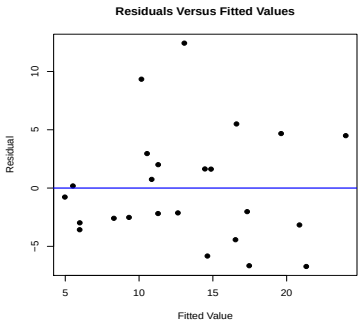
e) Based on the regression analysis of part *a*, which (if any) of the three predictors exhibit a statistically significant relationship to Cu? Use a level of significance $\alpha = 0.05$.

f) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t tests, that the error term ϵ is normally distributed appear to be met?

g) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the t tests, that the standard deviation σ of the error distribution is the same for different values of the predictors appear to be met?

h) Assuming that σ is the same for the different values of the predictors, what's the estimated value of σ?

i) Compute the coefficient of multiple determination R^2. What proportion of the variation in Cu can be explained by the three predictors distance from freeway, truck volume, and automobile volume?

13.5 A *groundwater budget analysis* is an attempt to quantify the amount of water entering and leaving an aquifer via recharge from rainfall and surface water sources, pumping for human usage, and loss through evaporation and discharge into rivers.

Near coastal areas, fresh groundwater is lost not only by discharge into rivers but also by submarine groundwater discharge, which refers to underground seepage into the oceans and flow from submarine springs. In such areas, understanding the relationship between submarine groundwater discharge and the other variables that affect an aquifer's groundwater supply is important for a groundwater budget analysis.

A study was carried out near the southwestern coast of Taiwan at Fangsan, where fresh groundwater is in short supply, to investigate the relationship between submarine groundwater discharge and three other variables that impact groundwater supply [9]. The data below are values of the following variables (all in $10^4\,\mathrm{m}^3$) for each of the years 1997 - 2003:

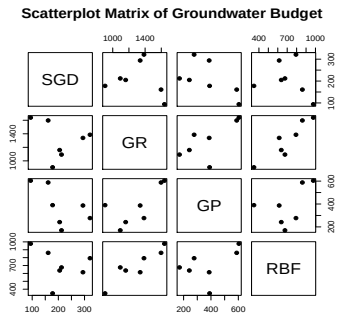
Submarine groundwater discharge (SGD).

Groundwater recharge (GR), for example from rainfall and surface water.
Groundwater pumped for human usage (GP).
River base flow (RBF), or groundwater discharged into rivers.

Groundwater Budget

Year	SGD	GR	GP	RBF
1997	321	1386	277	793
1998	205	1160	242	637
1999	161	1597	588	861
2000	294	1339	387	615
2001	94	1639	604	977
2002	178	905	390	347
2003	212	1093	170	676

A scatterplot matrix of these variables is below.



a) Carry out a multiple regression analysis, with submarine groundwater discharge (SGD) as the response and groundwater recharge (GR), groundwater pumped (GP), and river base flow (RBF) as predictors. Give the equation of the fitted regression model.

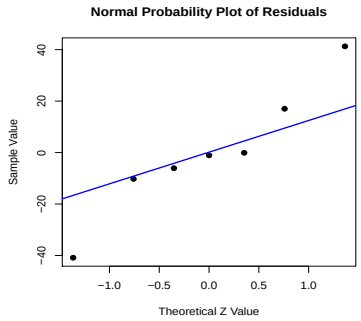
b) Based on the fitted regression model, by how much does the submarine groundwater discharge increase, on average, for each $10^4\,\mathrm{m}^3$ increase in groundwater recharge (while holding the other predictors constant)?

c) Based on the fitted regression model, by how much does the submarine groundwater discharge decrease, on average, for each $10^4\,\mathrm{m}^3$ increase in groundwater pumped for human usage (while holding the other predictors constant)?

d) Based on the fitted regression model, by how much does the submarine groundwater discharge decrease, on average, for each $10^4\,\mathrm{m}^3$ increase in river base flow (while holding the other predictors constant)?

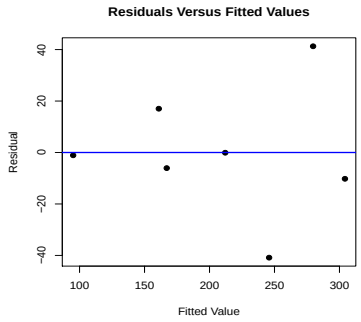
e) Based on the regression analysis of part *a*, which (if any) of the three predictors exhibit a statistically significant relationship to submarine groundwater discharge? Use a level of significance $\alpha = 0.05$.

f) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t tests, that the error term ϵ is normally distributed appear to be met?

g) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the t tests, that the standard deviation σ of the error distribution is the same for different values of the predictors appear to be met?

h) Assuming that σ is the same for the different values of the predictors, what's the estimated value of σ?

i) Compute the coefficient of multiple determination R^2. What proportion of the variation in submarine groundwater discharge can be explained by the three predictors groundwater recharge, groundwater pumped, and river base flow?

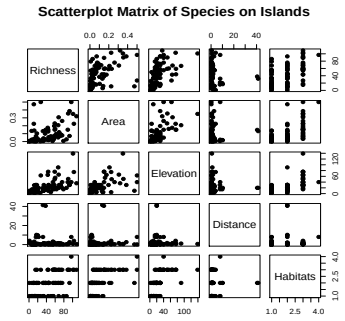
13.6 In the study of the relationship between plant species diversity and various features of small islands cited in Problem 11.7 in Chapter 11, for each of 86 islands in the Aegean archipelago, Greece, the *species richness* (number of unique species) was measured along with the the island's area (km^2), elevation (m), distance from the nearest large island (km), and number of habitat types [8]. All of the islands investigated were 0.5 km^2 or smaller. The table below shows data.

Plant Species Richness on Small Islands

Island	Species Richness	Area	Elevation	Distance to Nearest Island	Number of Habitat Types
Ag. Kyriaki	59	0.150	76	1.3	3
Ag. Nikolaos	69	0.119	30	0.4	3
Antidragonera	89	0.150	40	0.6	3
Archontonisi	11	0.028	15	0.2	1
Arefousa	43	0.175	65	0.7	3
Aspronisi (east)	11	0.007	15	2.1	2
Aspronisi (east 1)	34	0.037	30	1.4	2
Aspronisi (north)	45	0.056	30	1.8	2
Aspronisi (northwest)	46	0.048	25	1.1	2
Aspronisi (west)	7	0.010	15	1.9	2
Diabates (east)	50	0.067	5	0.04	1
Diabates (west)	16	0.067	5	0.06	1
East Gourna	33	0.008	10	0.5	1
Faradonisi (northwest)	33	0.040	10	1.3	2
Faradonisi (south)	22	0.025	5	0.7	1
Faradonisi (southwest)	19	0.020	10	0.9	1
Faradonisi megalo	60	0.160	55	0.8	3
Fragkonisi	103	0.225	75	4.0	3
Glaronisi (north)	57	0.030	15	0.7	2
Glaronisi (south)	73	0.090	28	1.3	3
Ilias	6	0.023	10	0.03	1
Imia (east)	17	0.017	20	10.2	2
Imia (west)	20	0.020	15	9.8	2
Kalapodi megalo	55	0.039	25	3.2	3
Kalapodi mikro	12	0.005	5	3.1	1
Kalovolos	68	0.307	66	1.6	3
Kapelo	1	0.009	10	4.0	1
Kapparonisi	57	0.068	18	1.8	3
Katsaganaki	16	0.002	10	0.1	1
Katsagani	72	0.090	30	0.2	2
Kombi	68	0.090	20	0.4	1
Kommeno nisi	34	0.028	10	1.2	1
Koukonisi	11	0.472	10	0.2	2
Kouloura 1	76	0.078	20	0.7	2
Kouloura 2	45	0.020	30	0.2	2
Koumaro	37	0.100	20	0.1	2
Kounelonisi	59	0.230	50	1.7	3
Lidia	15	0.035	27	0.8	1
Lyra	55	0.050	40	0.2	3
Makronisi	15	0.034	30	0.4	3
Makronisi 1	76	0.261	40	0.4	3
Makronisi 2	58	0.197	30	2.1	3
Marathi	90	0.355	51	0.6	3
Mavra (east)	38	0.148	20	40.6	2
Mavra (west)	32	0.132	20	41.4	2
Megali Dragonera	109	0.320	36	0.6	3
Megalo Stroggylo	12	0.030	29	1.2	1
Megalo Trachili	11	0.225	5	0.2	2
Mikro Trachili	6	0.135	5	0.1	1
Minaronisi	45	0.021	20	0.4	2
Neronisi	27	0.500	63	0.2	3
Nisida Manoli	55	0.029	30	1.0	2
Paplomata	26	0.004	3	0.1	1
Patelidi	8	0.025	5	0.7	1
Piato	56	0.060	20	1.8	1
Piganousa	101	0.350	139	0.7	3
Pitta	22	0.024	20	8.0	2
Plakousa	17	0.050	10	0.6	2
Plochoros	60	0.067	20	0.7	2
Pontikos	18	0.103	30	0.8	3
Prassonisi	14	0.013	10	1.0	1
Prassonisi 3	15	0.040	13	0.2	1
Prassonisi 1	13	0.011	2	1.3	1
Prassonisi 2	32	0.012	15	8.0	1
Prassou	97	0.500	40	8.0	4
Psathi	67	0.052	20	1.5	2
Psathonisi	44	0.127	10	0.7	2
Psonos	93	0.071	30	1.3	2
Saraki	16	0.007	30	7.8	1
Spartonisi	39	0.025	15	1.3	2
Stroggyli	62	0.096	20	1.1	2
Stroggyli 1	67	0.207	91	1.1	2
Stroggyli 2	44	0.150	76	0.5	3
Thimonies	8	0.010	10	0.1	1
Tiganaki	55	0.042	20	0.1	1
Tigani	12	0.140	5	0.03	1
Trypiti megali	72	0.072	30	1.0	2
Trypiti mikri	44	0.020	15	0.8	1
Vatopoula	54	0.007	15	2.4	1
Vatos	93	0.386	30	2.6	3
Velona	63	0.070	15	0.3	3
West Gourna	7	0.006	8	0.5	1
Zouka (Megali)	86	0.028	20	0.4	1
Zouka (Mikri)	79	0.008	15	0.2	1
(unnamed 1)	12	0.005	10	1.2	1
(unnamd 2)	1	0.0005	2	1.7	2

We'll investigate the relationship between the species richness of an island and its area, elevation, distance to nearest large island, and number of habitat types. A scatterplot matrix of these variables is below.

a) Carry out a multiple regression analysis, with species richness as the response and area, elevation,



distance to the nearest large island, and number of habitat types as predictors. Give the equation of the fitted regression model.

b) Based on the fitted regression model, by how much does the species richness increase, on average, for each one-km^2 increase in area (while holding the other predictors constant)?

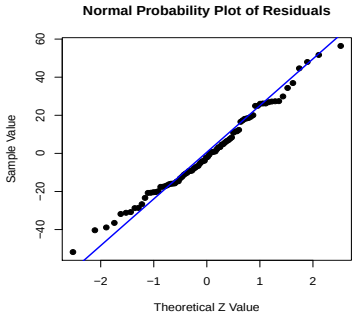
c) Based on the fitted regression model, by how much does the species richness increase, on average, for each one-m increase in elevation (while holding the other predictors constant)?

d) Based on the fitted regression model, by how much does the species richness decrease, on average, for each one-km increase in the distance to the nearest large island (while holding the other predictors constant)?

e) Based on the fitted regression model, by how much does the species richness increase, on average, for each one-habitat increase in the number of habitats (while holding the other predictors constant)?

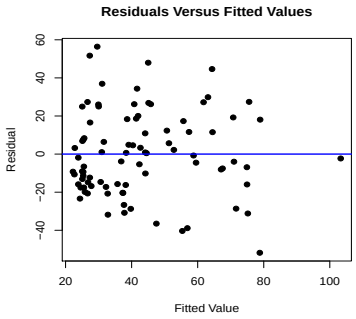
f) Based on the regression analysis of part *a*, which (if any) of the four predictors exhibit a statistically significant relationship to species richness? Use a level of significance $\alpha = 0.05$.

g) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t tests, that the error term ϵ is normally distributed appear to be met?

h) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the t tests, that the standard deviation σ of the error distribution is the same for different values of the predictors appear to be met?

i) Assuming that σ is the same for the different values of the predictors, what's the estimated value of σ?

j) Compute the coefficient of multiple determination R^2. What proportion of the variation in species richness can be explained by the four predictors area, elevation, distance to the nearest large island, and number of habitat types?

13.7 The fish stock off the coast of Greenland has declined dramatically since the early 1920s due to fishing activities. The data below are from a study of the progressive commercial extinction of Atlantic cod [10]. The response variable is the number of recruits at age three, that is, fish reaching their third year (in millions), a measure of the reproductive success of the fish. The two predictor variables are the spawning stock biomass (SSB), or total weight of all fish mature enough to reproduce (in thousands of tons), and the sea surface water temperature (degrees C) in mid June. The data are for the years 1955 - 1989.

Atlantic Cod off Greenland

Year	Recruits	SSB	Temperature
1955	134.5	1817.5	1.2
1956	463.7	1519.5	0.9
1957	531.7	1331.3	2.3
1958	226.9	1469.3	2.2
1959	93.6	1042.4	1.6
1960	409.6	1228.8	2.7
1961	703.4	1083.5	3.2
1962	286.7	1035.9	2.2
1963	330.0	1020.4	1.6
1964	105.6	887.2	2.3
1965	37.5	716.2	2.1
1966	39.1	715.5	1.6
1967	22.8	828.7	1.5
1968	88.0	775.9	2.1
1969	4.2	572.0	0.3
1970	9.2	467.0	0.3
1971	6.2	378.3	0.8
1972	24.6	248.1	0.6
1973	154.6	109.5	1.7
1974	16.6	88.9	1.4
1975	20.1	54.8	1.9
1976	26.8	30.1	1.4
1977	71.1	20.6	2.2
1978	14.3	37.8	0.9
1979	56.5	78.8	2.3
1980	7.7	94.1	1.9
1981	13.8	71.1	1.6
1982	2.0	57.2	0.8
1983	10.9	46.6	0.4
1984	265.7	35.6	1.0
1985	85.1	29.9	2.1
1986	1.4	32.9	2.2
1987	1.6	36.2	2.1
1988	0.6	56.4	2.0
1989	0.3	83.6	0.9

a) Carry out a multiple regression analysis, with recruits as the response and SSB and temperature as predictors. Give the equation of the fitted regression model.

b) Based on the fitted regression model, by how much does the number of recruits increase, on average, with each one-thousand-ton increase in SSB?

c) Based on the fitted regression model, by how much does the number of recruits increase, on average, with each one-degree increase in temperature?

d) Based on the regression analysis of part *a*, which (if any) of the two predictors are important for determining the number of recruits? Use a level of significance $\alpha = 0.05$.

e) Compute the coefficient of multiple determination R^2. What proportion of the variation in recruits can be explained by the two predictors SSB and temperature?

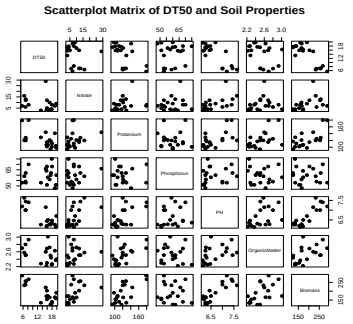
13.8 Biodegradation of pesticides in the environment results from the activities of soil microorganisms. Degradation rates can vary spatially according to the abundance of the microorganisms and the chemical properties of the soil such as its organic matter content, pH, and nutrient supply.

A study was carried out to investigate the influence of these soil characteristics on the degradation rate of the herbicide isoproturon [14]. Soil specimens collected from 20 sites in a research field were analyzed for nitrate, potassium, phosphorus (all in mg/kg), pH, organic matter (%), and microbial biomass (mg C/kg). Each soil specimen was then treated with isoproturon and incubated for 65 days. The response variable is DT_{50}, the time (in days) required for the isoproturon concentration to decrease by 50%. The data are below.

Degradation Rates and Soil Properties

Site	DT_{50}	Nitrate	Potassium	Phosphorus	pH	Organic Matter	Biomass
A2	16.3	4	107	58	6.50	3.01	173
A3	19.6	8	104	51	6.14	2.37	159
A4	18.0	8	111	68	6.47	2.27	140
A5	15.9	12	115	68	6.44	2.69	203
B1	7.1	13	128	62	7.44	2.51	257
B2	17.6	4	117	62	6.31	2.44	124
B3	17.4	3	94	60	6.26	2.50	125
C1	7.1	12	120	53	7.00	2.51	254
C2	19.8	9	98	60	6.70	2.40	191
C3	17.9	4	101	75	6.36	2.25	134
C6	19.2	7	141	48	6.26	2.63	166
D1	7.0	6	125	54	7.00	2.80	237
D2	17.0	10	121	56	6.28	2.72	165
D4	15.9	5	102	51	6.20	2.30	154
D5	13.4	3	129	53	6.26	2.28	135
E1	6.5	16	120	59	7.63	2.58	233
E3	15.7	3	107	64	6.17	2.22	172
E4	8.3	8	180	70	7.40	2.93	267
E5	15.4	29	144	66	7.15	2.60	221
E6	5.6	7	179	53	7.19	2.70	288

A scatterplot matrix of the data is below.



a) Carry out a multiple regression analysis with DT_{50} as the response and nitrate, potassium, phosphorus, pH, organic matter, and biomass as predictors. Give the equation of the fitted regression model.

b) Based on the regression analysis of part *a*, which (if any) of the six predictors exhibit a statistically significant relationship to DT_{50}? Use a level of significance $\alpha = 0.05$.

c) Carry out a backward elimination procedure to reduce the number of predictors in the model. Which predictors end up in the final model?

d) Now use a backward stepwise procedure to reduce the number of predictors in the model from the original six. Which predictors end up in the final model? Compare this final model to the one in part *c*.

13.9 In a study of heavy metals deposited by precipitation and wind-blown particles in the Alps mountain range, mosses were collected at sites along five transects following altitudinal gradients and the metal concentrations measured in the moss shoots [16]. For each site, the altitude and total precipitation during

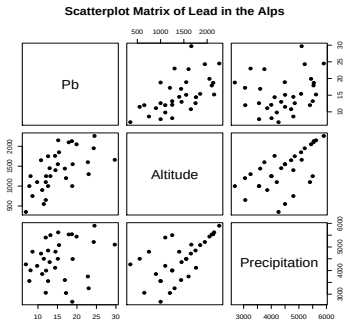
the moss growth period (1989 - 1991) were recorded.

Of interest were the relationships between the metals and the two predictors, altitude and precipitation. The table below shows the concentrations of lead, vanadium, iron, copper, nickel, and arsenic (Pb, V, Fe, Cu, Ni, As, all μg/g), the altitudes (m), and the precipitation levels (mm).

Metals in the Alps

Site	Pb	V	Fe	Cu	Ni	As	Altitude	Precipitation
1/1	11.1	1.3	215	5.3	1.1	0.30	900	3855
1/2	9.8	1.4	235	5.2	0.3	0.23	1100	4200
1/3	14.5	5.2	931	4.7	2.4	0.71	1400	4800
1/4	29.7	6.5	1239	6.6	2.0	0.81	1660	5100
2/1	12.0	1.4	844	5.6	2.2	0.94	650	3060
2/2	7.8	1.6	952	6.1	2.7	0.95	1000	3560
2/3	8.1	0.6	230	5.6	0.7	0.34	1250	4010
2/4	18.9	0.7	394	5.8	1.4	0.42	1550	4490
2/5	12.6	0.9	273	5.6	1.4	0.29	1750	4850
2/6	24.3	1.4	364	5.9	1.3	0.34	1950	5210
2/7	18.7	0.9	322	5.4	1.5	0.24	2130	5540
2/8	24.5	1.6	521	5.9	3.2	0.54	2260	5900
3/1	12.6	0.8	672	5.4	2.3	1.33	1000	3555
3/2	12.1	1.0	503	5.6	3.8	0.59	1250	4000
3/3	15.1	0.5	288	5.4	2.1	0.39	1550	4500
3/4	13.0	1.4	391	6.3	1.9	0.63	1450	4360
3/5	10.8	0.6	369	5.2	3.3	0.66	1650	4720
3/6	15.4	1.0	336	5.8	2.1	0.29	1850	5080
3/7	17.9	1.8	717	5.9	2.6	0.64	2100	5530
3/8	19.9	1.7	613	6.6	2.1	0.75	2050	5440
3/9	15.2	1.2	609	6.5	3.0	0.80	2150	5620
4/1	6.9	1.4	368	4.4	2.2	0.21	350	4263
4/2	11.5	5.9	2087	7.7	6.9	0.84	550	4500
4/3	8.6	5.4	1248	6.2	7.0	0.82	750	4800
4/4	12.0	2.4	1144	6.9	4.0	0.69	1100	5400
4/5	13.2	2.1	392	6.3	1.8	0.48	1250	5500
5/1	18.8	2.5	470	6.3	1.9	0.40	1000	2681
5/2	17.2	1.6	477	4.9	1.4	0.36	1200	3050
5/3	23.0	1.7	527	7.4	2.2	0.52	1300	3250
5/4	16.9	1.8	466	5.9	1.9	0.39	1440	3600
5/5	22.8	1.7	355	7.3	1.9	0.77	1600	3750
5/6	14.4	3.1	1970	4.8	4.6	1.19	1760	4120

In this problem, we'll investigate the relationship between the response variable Pb and the two predictors, altitude and precipitation. A scatterplot matrix of the data is below.



a) Carry out a multiple regression analysis with response variable Pb and predictors altitude and precipitation.

b) Based on the regression analysis of part *b*, which (if any) of the predictors are important for determining the Pb value?

c) By how much does the Pb concentration tend to increase, on average, with each one-meter increase in altitude?

d) What proportion of the total variation in Pb is explained by the regression model that includes altitude and precipitation as predictors?

13.10 A common method for disposing of household and industrial waste is incineration. But waste incineration emits pollutants, including dioxin. Several components of the incineration process can be adjusted to try to reduce dioxin emissions. In a study of the relationship between several incinerator combustion conditions and dioxin emissions, an incineration process was run 18 times under different operating conditions and dioxin measurements made for each test run [3].

The components of the incineration process that were adjusted included:

X_1 = furnace bed temperature
X_2 = furnace top temperature
X_3 = oxygen (O_2)
X_4 = secondary/primary air ratio
X_5 = total air supply
X_6 = nitrogen oxide (NO_x)

The resulting data are shown in the table below.

Test Run	Dioxin at Outlet	Total Air Supply	Secondary/Primary Ratio	Furnace Bed Temperature	Furnace Top Temperature	O_2	NO_x
1	13.70	29800	1.48	618	845	11.3	101
2	29.53	31500	1.63	609	783	12.1	116
3	14.98	31800	1.65	597	746	13.2	130
4	18.15	32100	1.79	589	858	12.5	118
5	8.89	31800	1.77	583	876	12.5	123
6	6.61	31500	1.76	604	874	12.7	147
7	20.12	29600	1.47	612	858	12.5	92
8	7.08	33580	2.04	603	864	11.9	107
9	11.49	36250	2.31	601	866	11.6	93
10	19.34	30520	1.78	603	890	12.1	NA
11	19.4	30520	1.78	603	890	12.1	NA
12	5.20	29110	1.64	601	915	12.8	128
13	11.00	32740	1.98	605	840	12.1	96
14	14.00	31270	1.82	603	861	11.8	100
15	25.78	26790	1.23	600	930	10.8	134
16	26.38	26780	1.23	600	930	10.8	132
17	30.41	27700	1.50	593	961	11.0	113
18	31.27	26500	1.39	596	948	11.5	93

a) Make a scatterplot matrix that includes the seven numerical variables in the table above.

b) Based on the scatterplot matrix of part *a*, which two predictors appear to be the most highly correlated (either positively or negatively)?

c) Perform a multiple regression analysis with Dioxin at Outlet as the response and Secondary/Primary Air Ratio, Furnace Bed Temperature, Furnace Top Temperature, O_2, and NO_x as the predictors. Give the equation of the fitted regression model.

d) Based on the regression analysis of part *c*, which of the seven predictors appear to be important in determining the Dioxin at Outlet? Use a level of significance $\alpha = 0.05$.

e) Calculate the R^2 for the model of part *c*. What proportion of the total variation in Dioxin at Outlet can be explained by the model?

f) The scatterplot matrix of part *a* shows that total air supply and secondary/primary air ratio are highly correlated, which suggests that it may not be necessary to include *both* of these variables in the model. Carry out the multiple regression analysis again, this time omitting secondary/primary air ratio from the model. Give the equation of the fitted regression model.

g) Based on the regression analysis of part *f*, which of the six predictors appear to be important in determining the dioxin at outlet? Use a level of significance $\alpha = 0.05$.

h) Why do the p-values for the six predictors left in the model change so dramatically when Secondary/Primary Air Ratio is removed from the model?

i) Calculate the R^2 for the model of part *f*. What proportion of the total variation in Dioxin at Outlet can be explained by the model?

j) Compare the R^2 of part *i* to the R^2 of part *e*. Does it appear that omitting secondary/primary air ratio from the model results in a much poorer fit to the data? Explain your answer.

13.11 Data were collected over the years 1980 - 1989 to document the recovery of a sparrowhawk population in Rockingham Forest, east-central England, in relation to declining levels of DDE and HEOD pesticide residues and declining levels of polychlorinated biphenyls (PCBs) in the region [7]. Each year during the study period, sparrowhawk nests were examined and unhatched eggs were collected for physical and chemical analysis. The following variables were recorded for each year:

DDE = Geometric mean of levels in sparrowhawk eggs (μg/g).
HEOD = Geometric mean of levels in sparrowhawk eggs (μg/g).
PCB = Geometric mean of levels in sparrowhawk eggs (μg/g)
Young Per Clutch = Mean number of surviving offspring per clutch, a measure of breeding success.
Clutches Hatched = Percentage of clutches for which one or more eggs hatched.
Shell Index = Mean of an index of egg shell thickness defined as $\frac{\text{Shell Weight (mg)}}{\text{Shell Length (mm)} \times \text{Shell Width (mm)}}$.

Geometric means were used to represent the pesticide residues and PCBs because the distributions of these variables were right skewed. The table below shows the data.

Year	Young Per Clutch	Clutches Hatched	Shell Index	DDE	HEOD	PCB
1980	1.8	58.8	1.18	10.66	0.94	5.69
1981	1.8	50.0	1.15	14.22	0.79	1.57
1982	1.8	50.0	1.20	9.55	0.98	2.50
1983	2.2	68.2	1.28	5.79	0.49	3.20
1984	2.1	71.1	1.28	5.94	0.78	4.61
1985	2.2	67.3	1.26	7.65	0.56	1.32
1986	2.8	75.9	1.33	5.52	0.64	2.84
1987	3.2	77.4	1.35	3.83	0.36	3.02
1988	2.9	76.4	1.33	4.17	0.82	4.56
1989	2.9	84.6	1.33	4.40	0.33	4.94

In this problem, we'll examine the relationship between the response variable young per clutch and the three predictors DDE, HEOD, and PCB.

a) Make a scatterplot matrix of the variables young per clutch, DDE, HEOD, and PCB.

b) Calculate the correlation matrix of the variables young per clutch, DDE, HEOD, and PCB.

c) From the scatterplot and correlation matrices of parts *a* and *b*, describe the relationship between young per clutch and DDE, between young per clutch and HEOD, and between young per clutch and PCB.

d) Based on the correlation matrix of part *b*, which of the predictors are most correlated with each other?

e) Perform a multiple regression analysis with young per clutch as the response variable and DDE, HEOD, and PCB as predictors. Give the equation of the fitted regression model.

f) Make a normal probability plot of the residuals from the regression analysis of part *c*. Based on the plot, is there any strong indication of non-normality in the errors?

g) Based on the regression analysis of part *c*, which (if any) of the three predictors exhibit a statistically significant relationship to young per clutch? Use a level of significance $\alpha = 0.05$.

h) Perform the multiple regression analysis again, but this time omitting HEOD from the model. Which (if any) of the two remaining predictors shows a significant relationship to young per clutch? Use a level of significance $\alpha = 0.05$.

i) Explain why DDE is statistically significant in the regression analysis of part *j*, but it wasn't significant in part *d*.

13.12 In a study of the composition of municipal solid waste worldwide, data from cities throughout the world or nation averages were compiled from several sources [1]. The goal of the study was to investigate the relationship between food waste and waste from materials used in packaging food. One potential question was whether lower levels of food waste are associated with higher levels of packaging waste, which would suggest that more packaging reduces the amount of food discarded. The table below shows the data, in proportions of total waste, along with the years the data were collected or first reported.

Country	City	Year	Paper and Board	Metal	Glass	Food Waste	Plastics
Austria	Vienna	1975	0.383	0.081	0.092	0.186	0.061
Austria	Vienna	1982	0.403	0.049	0.081	0.244	0.090
Belgium	Average	1976	0.300	0.053	0.080	0.400	0.050
Bulgaria	Sofia	1977	0.100	0.017	0.016	0.540	0.017
Columbia	Medellin	1979	0.220	0.010	0.020	0.560	0.050
Czechoslovakia	Prague	1975	0.134	0.062	0.066	0.418	0.042
Denmark	Average	1978	0.329	0.041	0.061	0.440	0.068
Denmark	Average	1970	0.450	0.040	0.080	0.130	NA
England	Average	1969	0.380	0.097	0.105	0.195	0.014
England	Average	1935-6	0.143	0.040	0.034	0.137	NA
England	Average	1963	0.230	0.082	0.086	0.141	NA
England	Average	1967	0.295	0.080	0.081	0.155	0.012
England	Average	1968	0.369	0.089	0.091	0.176	0.011
England	Doncaster	1985	0.210	0.070	0.060	0.150	0.050
England	Doncaster	1982	0.240	0.080	0.080	0.280	0.050
England	Doncaster	1985	0.280	0.090	0.080	0.200	0.070
England	London	1980	0.421	0.110	0.117	0.170	0.040
England	Stevenage	1979	0.330	0.070	0.090	0.160	0.030
Finland	Average	1978	0.550	0.050	0.060	0.200	0.060
France	Laval	1985	0.340	0.050	0.120	0.300	0.060
France	Paris	1979	0.340	0.040	0.090	0.150	0.040
Gabon	Average	1977	0.060	0.050	0.090	0.770	0.030
Germany (FRG)	Aachen	1974	0.308	0.069	0.135	0.164	0.045
Germany (FRG)	Aachen	1979	0.310	0.030	0.130	0.160	0.040
Germany (FRG)	Berlin	1978	0.218	0.049	0.191	0.314	0.060
Germany (FRG)	Dusseldorf	1974	0.278	0.044	0.164	0.342	0.062
Germany (FRG)	Hamburg	1975	0.231	0.045	0.227	0.300	0.046
Germany (FRG)	Munich	1974	0.406	0.061	0.069	0.075	0.075
Germany (FRG)	Stuttgart	1974	0.147	0.053	0.099	0.524	0.062
Germany (FRG)	Tubingen	1974	0.137	0.047	0.138	0.443	0.076
India	Calcutta	1976	0.030	0.010	0.080	0.360	0.010
India	Lucknow	1980	0.020	0.030	0.060	0.800	0.040
Indonesia	Bandung	1979	0.100	0.020	0.010	0.720	0.060
Indonesia	Bandung	1978	0.096	0.022	0.004	0.716	0.055
Indonesia	Bogor	1985	0.060	NA	NA	0.800	0.040
Indonesia	Jakarta	1978	0.020	0.040	0.010	0.820	0.030
Indonesia	Jakarta	1978	0.080	0.014	0.005	0.795	0.037
Indonesia	Surabaya	1983	0.020	0.005	0.010	0.940	0.020
Iran	Teheran	1978	0.172	0.018	0.021	0.698	0.038
Italy	Average	1979	0.310	0.070	0.030	0.360	0.070
Italy	Milan	1984	0.300	0.030	0.080	0.390	0.100
Italy	Rome	1980	0.250	0.025	0.013	0.500	0.060
Italy	Rome	1979	0.180	0.030	0.040	0.500	0.040
Japan	Gifu	1985	0.210	0.057	0.039	0.500	0.062
Japan	Mito	1985	0.301	0.015	0.011	0.418	0.056
Japan	Sakai (new area)	1985	0.230	0.022	0.053	0.541	0.081
Japan	Sakai (old area)	1985	0.295	0.039	0.049	0.404	0.071
Japan	Tokyo	1972	0.382	0.041	0.071	0.227	0.073
Japan	Tokyo	1978	0.436	0.012	0.010	0.340	0.056
Japan	Utsunomiya	1985	0.249	0.016	0.015	0.502	0.073
Kenya	Mombasa	1974	0.122	0.027	0.013	0.426	0.010
Netherlands	Amsterdam	1979	0.260	0.030	0.140	0.460	0.060
Netherlands	Average	1974	0.341	0.036	0.055	0.376	0.057
Netherlands	Average	1978	0.222	0.032	0.119	0.500	0.062
Netherlands	Average	1971	0.223	NA	0.081	0.536	0.068
Nigeria	Kano	1980	0.170	0.050	0.020	0.430	0.040
Nigeria	Lagos	NA	0.140	0.040	0.030	0.600	NA
Norway	Oslo	1985	0.382	0.020	0.075	0.304	0.065
Pakistan	Lahore	1980	0.040	0.040	0.030	0.490	0.020
Philippine Is.	Manilla	1978	0.170	0.020	0.050	0.430	0.040
Spain	Average	1978	0.180	0.040	0.030	0.500	0.040
Spain	Madrid	1979	0.190	0.060	0.030	0.500	0.080
Sri Lanka	Colombo	1981	0.080	0.010	0.060	0.800	0.010
Sudan	Khartoum	1984	0.040	0.030	NA	0.300	0.026
Sweden	Average	1977	0.500	0.070	0.080	0.150	0.080
Sweden	Stockholm	1985	0.390	0.050	0.140	0.150	0.080
U.S.A.	Average	1975	0.289	0.093	0.104	0.178	0.034
U.S.A.	Average	1973	0.427	0.092	0.103	0.146	0.017
U.S.A.	Berkeley, CA	1967	0.446	0.087	0.113	0.125	0.019
U.S.A.	Estimated	1975	0.290	0.091	0.104	0.178	0.034
U.S.A.	Estimated	1971	0.295	0.091	0.096	0.176	0.034
U.S.A.	Estimated	1975	0.272	0.153	0.103	0.154	0.032
U.S.A.	Estimated	1971	0.293	0.155	0.090	0.164	0.026
U.S.A	Johnson City, TN	1968	0.349	0.093	0.090	0.346	0.034
U.S.A.	New Orleans, LA	1972	0.394	0.122	0.146	0.189	0.038
U.S.A.	N. Little Rock, AK	1978	0.541	0.117	0.082	0.068	0.087
U.S.A.	Several	1970	0.442	0.087	0.085	0.116	0.012
U.S.A.	Wilmington, DE	1973	0.337	0.066	0.147	0.165	0.033

a) Make a scatterplot matrix with the variables food waste, paper and board, metal, glass, and plastics.

b) Compute the correlation matrix for the five variables of part *a*.

c) Perform a multiple regression analysis with food waste as the response and paper and board, metal, glass, and plastics as predictors. Give the equation of the fitted regression model.

d) Make a histogram of the residuals. Does the assumption of normality in the error terms appear to be met?

e) Which (if any) of the four predictors show a statistically significant relationship to clutches hatched? Use a level of significance $\alpha = 0.05$.

f) The cited paper includes the following quote:

> The fraction of food waste decreases as the fractions of waste from paper and board, metals and glass increase.

Do your regression analysis agree with this conclusion? Explain your answer.

13.13 The study on municipal waste composition cited in Problem 13.12 also reported data specific to the U.S.

Year	Paper and Board	Glass	Steel	Aluminum	Plastics	Food Waste
1960	0.144	0.077	0.060	0.002	0.002	0.147
1965	0.162	0.087	0.051	0.003	0.011	0.131
1970	0.158	0.106	0.048	0.005	0.019	0.115
1975	0.144	0.108	0.042	0.006	0.024	0.118
1980	0.145	0.105	0.027	0.007	0.034	0.093
1981	0.151	0.104	0.025	0.006	0.034	0.089
1982	0.144	0.102	0.023	0.006	0.033	0.088
1983	0.149	0.095	0.021	0.007	0.035	0.085
1984	0.156	0.089	0.021	0.007	0.037	0.081
1990	0.153	0.080	0.019	0.008	0.043	0.076
1995	0.157	0.074	0.016	0.009	0.048	0.073
2000	0.158	0.068	0.014	0.009	0.052	0.068

a) Make a scatterplot matrix of year and the six municipal waste variables.

b) Perform a multiple regression analysis with food waste as the response and year, paper and board, glass, steel, aluminum, and plastics as predictors. Give the equation of the fitted regression model.

c) Note that none of the seven predictor show a statistically significant relationship to food waste, but the overall model F test is highly significant. Explain how this situation arose.

13.14 Refer to the study of the relationship between water usage and wealth and size for metropolitan areas described in Example 13.1. The table below shows data on water consumption for commercial, industrial, and residential uses for each of 26 Chinese cities. Also shown for each city are the population (city and its surrounding area), a wealth score, and a measure of the local water resources.

City	Population (2000)	Area (km^2)	Population Density (pop/km^2)	Wealth Score	Water Usage (million liters/day)	Local Water Resources (kiloliters/day)
Shanghai	10,185,900	6341	1606.36	1.361	7895	11.61
Beijing	9,201,200	16,808	547.43	1.216	3064	3.94
Tianjin	5,937,500	11,920	498.11	0.150	1964	2.72
Chongqing	5,584,000	82,403	67.76	-0.619	1609	6.38
Wuhai	5,241,300	8467	619.03	-0.496	2823	11.92
Shenyang	4,791,000	12,980	369.11	-0.712	1720	5.84
Guangchou	3,956,500	7434	532.22	2.094	3369	22.33
Xi-An	3,734,900	9983	374.13	-0.708	1024	3.05
Chengdu	3,219,200	12,390	259.82	-0.164	1266	12.47
Haerbin	2,965,200	53,067	55.88	-0.895	976	2.37
Nanjing	2,732,600	6516	419.37	0.503	3688	10.09
Ji-nan	2,548,200	8154	312.51	0.177	840	4.26
Lanzhou	1,724,400	13,086	131.77	-0.487	1050	1.14
Kunming	1,697,100	15,561	109.06	0.075	605	6.81
Hangzhou	1,692,900	16,596	102.01	0.762	865	18.15
Fuzhou	1,416,800	11,968	118.38	-0.023	841	11.02
Wu-lu-mo-qi	1,361,200	12,000	113.43	-0.060	387	3.89
Datong	1,234,400	14,127	87.38	-0.731	363	1.76
Shenzhen	1,094,600	2020	541.88	3.090	1028	5.03
Huhehaote	986,800	17,224	57.29	-0.929	306	1.55
Liuzhou	875,400	5284	165.67	-0.434	1156	10.28
Mudanjiang	789,200	33,569	23.51	-1.163	400	2.41
Qinhuangdao	664,200	7523	88.29	-0.134	491	6.54
Guilin	603,500	4195	143.86	-0.595	403	16.13
Lianyungang	593,900	7444	79.78	-0.569	223	5.75
Yinchuan	573,400	3499	163.88	-0.709	302	0.65

a) Convert the populations to units of millions of people and take the logs of the water usage values, then fit a multiple linear regression model with response log of water usage and two predictors, wealth (z-score) and city size (population in millions). Write out the equation of the fitted regression model.

b) Controlling for the size of a city, by how much does water usage increase, on average, for each one unit increase in wealth?

c) Make a histogram and normal probability plot of the residuals. Do the plots provide any indication that the normality assumption of the error term in the multiple regression model is not met?

d) What proportion of the variation in water usage can be explained by the model with wealth and city size?

13.15 Particulate matter smaller than 2.5 μm in diameter ($PM_{2.5}$) in the air can trigger asthma attacks in asthmatic children. It is desirable, therefore, that children not be exposed to high $PM_{2.5}$ levels immediately upon being dismissed from schools at the end of each day. To assist schools in developing end-of-day dismissal procedures that limit children's exposure to $PM_{2.5}$, a study investigated the impact of car and truck traffic and meteorological variables on $PM_{2.5}$ concentrations near schools [11].

For each of 13 days in the fall of 2006, $PM_{2.5}$ concentrations were measured at one minute intervals between 1:45 and 3:30 pm near the intersection of Madison Avenue and E. 104th Street in the East Harlem neighborhood, New York City. The site was in close proximity to two schools, the Reece School and Public School 171, and the sampling period 1:45 - 3:30 encompasses the end-of-day dismissals for the schools.

Also recorded were the number of gasoline vehicles (cars and pickup trucks) and diesel vehicles (buses and large trucks) passing by (through a green light) or idling (at a red light or while parking). In addition, during the sampling period 1:45 - 3:30 for each day, the temperature, wind speed, relative humidity, barometric pressure and background $PM_{2.5}$ concentrations were averaged over several sites in New York City. The data are shown below.

Date (mm/dd)	$PM_{2.5}$ Avg (ng/m^3)	$PM_{2.5}$ Background (ng/m^3)	Temp (°C)	Wind Speed (m/s)	Relative Humidity (%)	Barometric Pressure (mm Hg)	Idling Diesel	Idling Gasoline	Passing Diesel	Passing Gasoline
10/31	49245	19367	20	7.4	56	756	60	117	158	152
11/01	44873	14094	21	5.0	61	758	73	52	99	162
11/02	20769	8237	12	5.8	53	759	85	201	244	193
11/03	15864	5007	9	6.3	45	766	138	287	334	169
11/06	52143	18377	14	5.2	43	769	109	156	219	130
11/07	11689	9022	14	5.9	85	764	10	60	84	146
11/09	15146	6880	20	7.3	68	749	55	148	187	133
11/10	6519	7501	16	5.1	54	759	73	173	207	157
11/13	30108	10099	16	10.6	94	756	45	88	125	134
11/14	19075	9558	16	5.8	86	754	76	154	204	154
11/15	69899	18047	16	4.5	95	759	81	79	120	134
11/16	22328	10032	19	14.4	93	752	83	146	192	244
11/17	8177	7822	16	9.7	61	754	93	121	165	143

We want to choose a model that adequately explains variation in $PM_{2.5}$ concentrations in terms of just a few of the the traffic, meteorological, and background $PM_{2.5}$ explanatory variables. Carry out a stepwise model selection procedure to select an appropriate model, and state the model suggested by the procedure.

13.16 Dioxins are toxic chemical compounds that are released into the environment during incineration of municipal waste, during forest fires, and via some industrial processes. They are poorly soluble, persistent, and can bioaccumulate. Dioxins can pollute water bodies through stormwater runoff and wet and dry atmospheric deposition of emissions.

To quantify dioxin levels in runoff in Houston, Texas, dioxins were measured in water from 10 small flood control drainage channels in the Houston area during stormwater runoff events [12]. The table below shows concentrations of one of the dioxins along with several variables characterizing the sampling sites, storm events, and water flowing through the channels. The variables included in the data set are:

Sample collection information:
- Site = Sampling site identification code
- Date = Date of the storm event and sample collection from the site

Dioxin:
- HpCDFs = Concentration(pg/L) of the heptachlorodibenzofuran dioxins in the sampled water

Sampling site and storm event characteristics:
- Downwind = Is the sampling site down wind of a known atmospheric dioxin source (Yes/No)?
- Dry Days = Antecedent dry period (days) preceding the storm event
- Rainfall = Total rainfall (mm) during the storm event
- Channel Flow = Total water flow (m^3) through the channel during the storm event

Characteristics of the watershed (area from which water drains into the channel):
- Area = The total area (ha) of the channel's watershed
- Developed = Percent of the channel's watershed composed of developed land
- Grass/Ag = Percent of the channel's watershed composed of grassland or agriculture
- Woodland = Percent of the channel's watershed composed of woodland
- Open Water = Percent of the channel's watershed composed of open water
- Wetlands = Percent of the channel's watershed composed of wetlands
- Bare = Percent of the channel's watershed composed of bare land

Characteristics of the water sampled from the channel:
- TSS = Total suspended sediment (mg/L, solids retained after filtering the water

through a 0.45 μm filter)
TOC = Total organic carbon (mg/L, total amount of carbon bound in organic compounds in the water)
DOC = Dissolved organic carbon (mg/L, amount of carbon bound in organic compounds and dissolved in the water)
Temp = Temperature of the water (°C)
Conductivity = Conductivity (μS/cm) of the water (ability to pass electrical current, an indicator of the presence of inorganic dissolved solids)
Salinity = Salinity (%, the amount of dissolved salts in the water)
DO = Dissolved oxygen (mg/L, the concentration of oxygen incorporated in the water)
PH = Acidity of the water (pH)

Site	Date	HpCDFs	Down-wind	Dry Days	Rain-fall	Channel Flow	Area	Devel-oped	Grass/ Ag
SS-7	12/9/02	18.376	No	2	21.3	17032	1227	68	15
SS-8	12/9/02	32.645	Yes	2	20.3	16275	1059	48	13
SS-9	12/12/02	48.815	Yes	3	59.7	96139	1069	14	12
SS-10	12/3/02	40.709	Yes	6	18	12869	310	70	17
SS-11	1/26/03	60.224	Yes	7	31	56397	645	63	22
SS-12	3/3/03	23.955	Yes	4	45.2	11733	833	15	67
SS-14	2/6/03	22.159	No	7	12.2	14761	1000	72	19
SS-15	3/3/03	46.143	No	4	17.3	186222	1152	89	8
SS-16	12/4/02	2.688	No	6	27.2	9084	876	63	26
SS-17	11/3/02	0.785	No	2	20.3	9462	1005	22	47

(cont'd)

Wood-land	Open Water	Wet-lands	Bare	TSS	TOC	DOC	Temp	Conduc-tivity	Salinity	DO	PH
14	1	1	0	37.6	10.5	8.3	11.31	242.3	0.12	9	7.45
36	1	1	0	421.7	11.1	8.8	12.11	414.7	0.21	9.4	7.36
38	7	27	1	97	16.9	13.6	12.62	132.5	0.06	9.1	6.98
12	0	1	0	28.8	11.5	8	20.02	381.3	0.18	4.5	7.76
13	0	1	0	66.1	11.8	9.6	8.45	237.3	0.11	11.3	7.81
7	0	11	0	48.3	7.3	5.6	12.33	368.8	0.18	8.7	7.35
6	1	1	1	66.3	11.1	9.1	12.14	180.7	0.09	12.3	7.51
1	1	0	1	55.3	147.7	126.4	13.62	427.3	0.21	4.4	9.25
8	0	2	1	34.2	14.4	10	18.47	85.0	0.04	4.8	8.05
26	0	3	1	24.0	15.3	14.6	17.03	192.0	0.09	7.3	7.65

We want to fit a model, with dioxin (HpCDFs) as the response variable and no more than four predictors from among the variables characterizing the sampling site, storm event and water sampled from the channel.

Use a forward stepwise regression procedure to determine which four predictors should be included in the model.

13.17 Studies have shown that the number of eggs laid by females of the bird species blue tit (*Parus caeruleus*) is that number which maximizes the number of offspring that survive to enter the population (the recruitment rate). Laying more eggs or fewer would lead to smaller numbers of offspring surviving each year.

It's hypothesized that although the number of eggs laid is optimal for the survival of offspring, rearing that many young may take its toll on parents and lead to lower survival and reproductive rates among the parents. In other words, the parents may be paying a price for not rearing a smaller brood.

To test this hypothesis, an experiment was carried in which adult females were made rear one of five brood sizes (3, 6, 9, 12, and 15). Allocation of females to brood sizes was done randomly. The survival status after one year (yes/no) was then determined for each adult female. The data are shown below.

Female Adult	Brood Size	Survived
1	3	No
2	3	No
3	3	No
4	3	No
5	3	Yes
6	3	No
7	3	No
8	3	Yes
9	3	Yes
10	3	No
11	6	No
12	6	No
13	6	No
14	6	No
15	6	Yes
16	6	No
17	6	No
18	6	No
19	6	No
20	6	Yes
21	9	No
22	9	No
23	9	No
24	9	Yes
25	9	No
26	9	No
27	9	No
28	9	No
29	9	No
30	9	No
31	12	No
32	12	No
33	12	No
34	12	No
35	12	No
36	12	No
37	12	No
38	12	Yes
39	12	No
40	12	No
41	15	No
42	15	No
43	15	No
44	15	No
45	15	No
46	15	No
47	15	No
48	15	No
49	15	No
50	15	No
51	15	No

Carry out a logistic regression to determine if there's a relationship between survival status and brood size. Use a level of significance $\alpha = 0.05$.

Bibliography

[1] H. Alter. The origins of municipal solid waste: The relations between residues from packaging materials and food. *Waste Management and Research*, 7:103 – 114, 1989.

[2] H. Gan, M. Zhuo, D. Li, and Y. Zhou. Quality characterization and impact assessment of highway runoff in urban and rural area of Guangzhou, China. *Environmental Monitoring and Assessment*. DOI 10.1007/s10661-007-9856-2.

[3] R. Ishikawa, A. Buekens, H. Huang, and K. Watanabe. Influence of combustion conditions on dioxin in an industrial scale fluidized-bed incinerator: Experimental study and statistical modelling. *Chemosphere*, 35(3):465 – 477, 1997.

[4] G. Darrel Jenerette, Wanli Wu, Susan Goldsmith, Wendy A. Marussich, and W. John Roach. Contrasting water footprints of cities in China and the United States. *Ecological Economics*, 57:346 – 358, 2006.

[5] Michael H. Kutner, Christopher J. Nachtsheim, and John Neter. *Applied Linear Regression Models*. McGraw-Hill/Irwin, fourth edition, 2004.

[6] Juni I. Liu, Rajendra D. Paode, and Thomas M. Holsen. Modeling the energy content of municipal solid waste using multiple regression analysis. *Journal of the Air and Waste Management Association*, 46(7):650 – 656, 1996.

[7] I. Newton and I. Wyllie. Recovery of a sparrowhawk population in relation to declining pesticide contamination. *Journal of Applied Ecology*, 29(2):476 – 484, 1992.

[8] M. Panitsa, D. Tzanoudakis, K. A. Triantis, and S. Sfenthourakis. Patterns of species richness on very small islands: the plants of the Aegean archipelago. *Journal of Biogeography*, 33:1223 – 1234, 2006.

[9] Tsung-Ren Peng, Chen-Tung Arthur Chen, Chung-Ho Wang, Jing Zhang, and Yi-Jie Lin. Assessment of terrestrial factors controlling the submarine groundwater discharge in water shortage and highly deformed island of Taiwan, western Pacific Ocean. *Journal of Oceanography*, 64:323 – 337, 2008.

[10] H. Ratz, M. Stein, and J. Lloret. Variation in growth and recruitment of Atlantic cod (*Gadus morhua*) off Greenland during the second half of the twentieth century. *Journal of Northwest Atlantic Fishery Science*, 25:161 – 170, 2003.

[11] J. Richmond-Bryant, C. Saganich, L. Bukiewicz, and R. Kalin. Associations of $pm_{2.5}$ and black carbon concentrations with traffic, idling, background pollution, and meteorology during school dismissals. *Science of the Total Environment*, 407:3357 – 3364, 2009.

[12] Monica P. Suarez, Hanadi S. Rifai, Jennifer Schimek, Michael Bloom, Paul Jensen, and Larry Koenig. Dioxin in storm-water runoff in Houston, Texas. *Journal of Environmental Engineering*, 132(12):1633 – 1643, Dec 2006.

[13] P. Van Vliet, M. Knape, J. De Hartog, N. Janssen, H. Harssema, and B. Brunekreef. Motor vehicle exhaust and chronic respiratory symptoms in children living near freeways. *Environmental Research*, 74:122 – 132, 1997.

[14] A. Walker, M. Jurado-Exposito, G.D. Bending, and V.J.R. Smith. Spatial variability in the degradation rate of isoproturon in soil. *Environmental Pollution*, 111:407 – 415, 2001.

[15] Jie Wang, Osamu Yamada, Tetsuya Nakazato, Zhan-Guo Zhang, Yoshizo Suzuki, and Kinya Sakanishi. Statistical analysis of the concentrations of trace elements in a wide diversity of coals and its implications for understanding elemental modes of occurrence. *Fuel*, 87:2211 – 2222, 2008.

[16] H. G. Zechmeister. Correlation between altitude and heavy metal deposition in the Alps. *Environmental Pollution*, 89:73 – 80, 1995.