


to be popular/likable. As part of a recent study on adolescents, an experimenter looked at the relationship between the expression of a particular serotonin receptor gene, a person's "popularity," and the person's rule-breaking (RB) behaviors.²⁴ RB was measured by both a questionnaire and video observation. The composite score is an equal combination of these two assessments. Here is a table of the correlations:

Rule-breaking measure	Popularity	Gene expression
Sample 1 ($n = 123$)		
RB.composite	0.28	0.26
RB.questionnaire	0.22	0.23
RB.video	0.24	0.20
Sample 1 Caucasians only ($n = 96$)		
RB.composite	0.22	0.23
RB.questionnaire	0.16	0.24
RB.video	0.19	0.16



For each correlation, test the null hypothesis that the corresponding true correlation is zero. Reproduce the table and mark the correlations that have $P < 0.001$ with ***, those that have $P < 0.01$ with **, and those that have $P < 0.05$ with *. Write a summary of the results of your significance tests.

10.60 Resting metabolic rate and exercise. Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The following table gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.  METRATE

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

(a) Make a scatterplot of the data, using different symbols or colors for men and women. Summarize what you see in the plot.



(b) Run the regression to predict metabolic rate from lean body mass for the women in the sample and summarize the results. Do the same for the men.

 **10.61 Resting metabolic rate and exercise, continued.** Refer to the previous exercise. It is tempting to conclude that there is a strong linear relationship for the women but no relationship for the men. Let's look at this issue a little more carefully.  METRATE

(a) Find the confidence interval for the slope in the regression equation that you ran for the females. Do the same for the males. What do these suggest about the possibility that these two slopes are the same? (The formal method for making this comparison is a bit complicated and is beyond the scope of this chapter.)

(b) Examine the formula for the standard error of the regression slope given on page 593. The term in the denominator is $\sqrt{\sum(x_i - \bar{x})^2}$. Find this quantity for the females; do the same for the males. How do these calculations help to explain the results of the significance tests?

(c) Suppose that you were able to collect additional data for males. How would you use lean body mass in deciding which subjects to choose?

 **10.62 Inference over different ranges of X .** Think about what would happen if you analyzed a subset of a set of data by analyzing only data for a restricted range of values of the explanatory variable. What results would you expect to change? Examine your ideas by analyzing the fuel efficiency data described in Example 10.11 (page 581). First, run a regression of MPG versus MPH using all cases. This least-squares regression line is shown in Figure 10.9. Next run a regression of MPG versus MPH for only those cases with speed less than or equal to 30 mph. Note that this corresponds to 3.4 in the log scale. Finally, do the same analysis with a restriction on the response variable. Run the analysis with only those cases with fuel efficiency less than or equal to 20 mpg. Write a summary comparing the effects of these two restrictions with each other and with the complete data set results.  MPHMPG



Multiple Regression

Introduction

In Chapter 10 we presented methods for inference in the setting of a linear relationship between a response variable y and a *single* explanatory variable x . In this chapter, we use *more than one* explanatory variable to explain or predict a single response variable.

Many of the ideas that we encountered in our study of simple linear regression carry over to the multiple linear regression setting. For example, the descriptive tools we learned in Chapter 2—scatterplots, least-squares regression, and correlation—are still essential preliminaries to inference and also provide a foundation for confidence intervals and significance tests.

The introduction of several explanatory variables leads to many additional considerations. In this short chapter we cannot explore all these issues. Rather, we will outline some basic facts about inference in the multiple regression setting and then illustrate the analysis with a case study whose purpose was to predict success in college based on several high school achievement scores.

CHAPTER

11

- 11.1 Inference for Multiple Regression
- 11.2 A Case Study

11.1 Inference for Multiple Regression

When you complete this section, you will be able to

- Describe the multiple linear regression model in terms of a population regression line and the deviations of the response variable y from this line.
- Interpret regression output from statistical software to obtain the least-squares regression equation and model standard deviation, multiple correlation coefficient, ANOVA F test, and individual regression coefficient t tests.
- Explain the difference between the ANOVA F test and the t tests for individual coefficients.
- Interpret a level C confidence interval or significance test for a regression coefficient.
- Use diagnostic plots to check the assumptions of the multiple linear regression model.

Population multiple regression equation

The simple linear regression model assumes that the mean of the response variable y depends on the explanatory variable x according to a linear equation

$$\mu_y = \beta_0 + \beta_1 x$$

For any fixed value of x , the response y varies Normally around this mean and has a standard deviation σ that is the same for all values of x .

In the multiple regression setting, the response variable y depends on p explanatory variables, which we will denote by x_1, x_2, \dots, x_p . The mean response depends on these explanatory variables according to a linear function

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

population regression equation

Similar to simple linear regression, this expression is the **population regression equation**, and the observed values y vary about their means given by this equation.

Just as we did in simple linear regression, we can also think of this model in terms of subpopulations of responses. Here, each subpopulation corresponds to a particular set of values for *all* the explanatory variables x_1, x_2, \dots, x_p . In each subpopulation, y varies Normally with a mean given by the population regression equation. The regression model assumes that the standard deviation σ of the responses is the same in all subpopulations.

EXAMPLE

11.1 Predicting early success in college. Our case study is based on data collected on science majors at a large university.¹ The purpose of the study was to attempt to predict success in the early university years. One measure of success was the cumulative grade point average (GPA) after three semesters. Among the explanatory variables recorded at the time the students enrolled in the university were average high school grades in mathematics (HSM), science (HSS), and English (HSE).

We will use high school grades to predict the response variable GPA. There are $p = 3$ explanatory variables: $x_1 = \text{HSM}$, $x_2 = \text{HSS}$, and $x_3 = \text{HSE}$. The high school grades are coded on a scale from 1 to 10, with 10 corresponding to A, 9 to A-, 8 to B+, and so on. These grades define the subpopulations. For example, the straight-C students are the subpopulation defined by $\text{HSM} = 4$, $\text{HSS} = 4$, and $\text{HSE} = 4$.

One possible multiple regression model for the subpopulation mean GPAs is

$$\mu_{\text{GPA}} = \beta_0 + \beta_1 \text{HSM} + \beta_2 \text{HSS} + \beta_3 \text{HSE}$$

For the straight-C subpopulation of students, the model gives the subpopulation mean as

$$\mu_{\text{GPA}} = \beta_0 + \beta_1 4 + \beta_2 4 + \beta_3 4$$

Data for multiple regression

The data for a simple linear regression problem consist of observations (x_i, y_i) of the two variables. Because there are several explanatory variables in multiple regression, the notation needed to describe the data is more elaborate. Each observation or case consists of a value for the response variable and for each of the explanatory variables. Call x_{ij} the value of the j th explanatory variable for the i th case. The data are then

Case 1: $(x_{11}, x_{12}, \dots, x_{1p}, y_1)$

Case 2: $(x_{21}, x_{22}, \dots, x_{2p}, y_2)$

•

•

•

Case n : $(x_{n1}, x_{n2}, \dots, x_{np}, y_n)$

Here, n is the number of cases and p is the number of explanatory variables. Data are often entered into computer regression programs in this format. Each row is a case and each column corresponds to a different variable.

The data for Example 11.1, with several additional explanatory variables, appear in this format in the GPA data file. Figure 11.1 shows the first 5 rows entered into an Excel spreadsheet. Grade point average (GPA) is the response variable, followed by $p = 7$ explanatory variables. There are a total of $n = 150$ students in this data set.

FIGURE 11.1 Format of data set for Example 11.1 in an Excel spreadsheet.

[illegible]

USE YOUR KNOWLEDGE

11.1 Describing a multiple regression. Traditionally, demographic and high school academic variables have been used to predict college academic success. One study investigated the influence of emotional health on GPA.² Data from 242 students who had completed their first two semesters of college were obtained. The researchers were interested in describing how students' second-semester grade point averages are explained by gender, a standardized test score, perfectionism, self-esteem, fatigue, optimism, and depressive symptomatology.

- What is the response variable?
- What is n , the number of cases?
- What is p , the number of explanatory variables?
- What are the explanatory variables?

Multiple linear regression model

We combine the population regression equation and assumptions about variation to construct the multiple linear regression model. The subpopulation means describe the FIT part of our statistical model. The RESIDUAL part represents the variation of observations about the means.

We will use the same notation for the residual that we used in the simple linear regression model. The symbol ϵ represents the deviation of an individual observation from its subpopulation mean.

We assume that these deviations are Normally distributed with mean 0 and an unknown model standard deviation σ that does not depend on the values of the x variables. *These are assumptions that we can check by examining the residuals in the same way that we did for simple linear regression.*



MULTIPLE LINEAR REGRESSION MODEL

The **statistical model for multiple linear regression** is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $i = 1, 2, \dots, n$.

The **mean response** μ_y is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The **deviations** ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ . In other words, they are an SRS from the $N(0, \sigma)$ distribution.

The **parameters of the model** are $\beta_0, \beta_1, \beta_2, \dots, \beta_p$, and σ .

The assumption that the subpopulation means are related to the regression coefficients β by the equation

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

← LOOK BACK
DATA = FIT + RESIDUAL,
p. 567

implies that we can estimate all subpopulation means from estimates of the β 's. To the extent that this equation is accurate, we have a useful tool for describing how the mean of y varies with the collection of x 's.

We do, however, need to be cautious when interpreting each of the regression coefficients in a multiple regression. First, the β_0 coefficient represents the mean of y when *all* the x variables equal zero. Even more so than in simple linear regression, this subpopulation is rarely of interest. Second, the description provided by the regression coefficient of each x variable is similar to that provided by the slope in simple linear regression but only in a specific situation, namely, *when all other x variables are held constant*. We need this extra condition because with multiple x variables, it is quite possible that a unit change in one x variable may be associated with changes in other x variables. If that occurs, then the change in the mean of y is not described by just a single regression coefficient.

USE YOUR KNOWLEDGE

11.2 Understanding the fitted regression line. The fitted regression equation for a multiple regression is

$$\hat{y} = -1.8 + 6.1x_1 - 1.1x_2$$

- If $x_1 = 3$ and $x_2 = 1$, what is the predicted value of y ?
- For the answer to part (a) to be valid, is it necessary that the values $x_1 = 3$ and $x_2 = 1$ correspond to a case in the data set? Explain why or why not.
- If you hold x_2 at a fixed value, what is the effect of an increase of two units in x_1 on the predicted value of y ?

Estimation of the multiple regression parameters

← LOOK BACK
least squares, p. 113

Similar to simple linear regression, we use the method of least squares to obtain estimators of the regression coefficients β . The details, however, are more complicated. Let

$$b_0, b_1, b_2, \dots, b_p$$

denote the estimators of the parameters

$$\beta_0, \beta_1, \beta_2, \dots, \beta_p$$

For the i th observation, the predicted response is

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

← LOOK BACK
residual, p. 569

The i th residual, the difference between the observed and the predicted response, is therefore

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip} \end{aligned}$$

The method of least squares chooses the values of the b 's that make the sum of the squared residuals as small as possible. In other words, the parameter estimates $b_0, b_1, b_2, \dots, b_p$ minimize the quantity

$$\sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2$$



The formula for the least-squares estimates is complicated. We will be content to understand the principle on which it is based and to let software do the computations.

The parameter σ^2 measures the variability of the responses about the population regression equation. As in the case of simple linear regression, we estimate σ^2 by an average of the squared residuals. The estimator is

$$s^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}$$

LOOK BACK
degrees of freedom, p. 44

The quantity $n - p - 1$ is the degrees of freedom associated with s^2 . The degrees of freedom equal the sample size, n , minus $p + 1$, the number of β 's we must estimate to fit the model. In the simple linear regression case there is just one explanatory variable, so $p = 1$ and the degrees of freedom are $n - 2$. To the model standard deviation σ we use

$$s = \sqrt{s^2}$$

Confidence intervals and significance tests for regression coefficients

We can obtain confidence intervals and perform significance tests for each of the regression coefficients β_j as we did in simple linear regression. The standard errors of the b 's have more complicated formulas, but all are multiples of s . We again rely on statistical software to do the calculations.

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR β_j

A **level C confidence interval** for β_j is

$$b_j \pm t^* SE_{b_j}$$

where SE_{b_j} is the standard error of b_j and t^* is the value for the $t(n - p - 1)$ density curve with area C between $-t^*$ and t^* .

To test the hypothesis $H_0: \beta_j = 0$, compute the **t statistic**

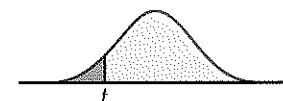
$$t = \frac{b_j}{SE_{b_j}}$$

In terms of a random variable T having the $t(n - p - 1)$ distribution, the P -value for a test of H_0 against

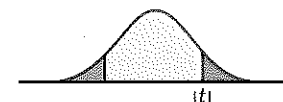
$$H_a: \beta_j > 0 \text{ is } P(T \geq t)$$



$$H_a: \beta_j < 0 \text{ is } P(T \leq t)$$



$$H_a: \beta_j \neq 0 \text{ is } 2P(T \geq |t|)$$



LOOK BACK
confidence intervals for mean
response, p. 577
prediction intervals, p. 579

Because regression is often used for prediction, we may wish to use multiple regression models to construct confidence intervals for a mean response and prediction intervals for a future observation. The basic ideas are the same as in the simple linear regression case.

In most software systems, the same commands that give confidence and prediction intervals for simple linear regression work for multiple regression. The only difference is that we specify a list of explanatory variables rather than a single variable. Modern software allows us to perform these rather complex calculations without an intimate knowledge of all the computational details. This frees us to concentrate on the meaning and appropriate use of the results.

ANOVA table for multiple regression

In simple linear regression the F test from the ANOVA table is equivalent to the two-sided t test of the hypothesis that the slope of the regression line is 0. For multiple regression there is a corresponding ANOVA F test, but it tests the hypothesis that *all* the regression coefficients (with the exception of the intercept) are 0. Here is the general form of the ANOVA table for multiple regression:

Source	Degrees of freedom	Sum of squares	Mean square	F
Model	p	$\sum (\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	$n - p - 1$	$\sum (y_i - \hat{y}_i)^2$	SSE/DFE	
Total	$n - 1$	$\sum (y_i - \bar{y})^2$	SST/DFT	

The ANOVA table is similar to that for simple linear regression. The degrees of freedom for the model increase from 1 to p to reflect the fact that we now have p explanatory variables rather than just one. As a consequence, the degrees of freedom for error decrease by the same amount. *It is always a good idea to calculate the degrees of freedom by hand and then check that your software agrees with your calculations. In this way you can verify that your software is using the number of cases and number of explanatory variables that you intended.*

The sums of squares represent sources of variation. Once again, both the sums of squares and their degrees of freedom add:

$$SST = SSM + SSE$$

$$DFT = DFM + DFE$$

The estimate of the variance σ^2 for our model is again given by the MSE in the ANOVA table. That is, $s^2 = \text{MSE}$.

The ratio MSM/MSE is an F statistic for testing the null hypothesis

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

against the alternative hypothesis

$$H_a: \text{at least one of the } \beta_j \text{ is not } 0$$

The null hypothesis says that none of the explanatory variables are predictors of the response variable when used in the form expressed by the multiple regression equation. The alternative states that *at least one* of them is a predictor of the response variable.

LOOK BACK
 F statistic, p. 588

As in simple linear regression, large values of F give evidence against H_0 . When H_0 is true, F has the $F(p, n - p - 1)$ distribution. The degrees of freedom for the F distribution are those associated with the model and error in the ANOVA table.



A common error in the use of multiple regression is to assume that all the regression coefficients are statistically different from zero whenever the F statistic has a small P -value. Be sure that you understand the difference between the F test and the t tests for individual coefficients.

ANALYSIS OF VARIANCE F TEST

In the multiple regression model, the hypothesis

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

is tested against the alternative hypothesis

$$H_a: \text{at least one of the } \beta_i \text{ is not } 0$$

by the analysis of variance **F statistic**

$$F = \frac{\text{MSM}}{\text{MSE}}$$

The P -value is the probability that a random variable having the $F(p, n - p - 1)$ distribution is greater than or equal to the calculated value of the F statistic.

Squared multiple correlation R^2

For simple linear regression we noted that the square of the sample correlation could be written as the ratio of SSM to SST and could be interpreted as the proportion of variation in y explained by x . A similar statistic is routinely calculated for multiple regression.

THE SQUARED MULTIPLE CORRELATION

The statistic

$$R^2 = \frac{\text{SSM}}{\text{SST}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

is the proportion of the variation of the response variable y that is explained by the explanatory variables x_1, x_2, \dots, x_p in a multiple linear regression.

Often, R^2 is multiplied by 100 and expressed as a percent. The square root of R^2 , called the **multiple correlation coefficient**, is the correlation between the observations y_i and the predicted values \hat{y}_i .

USE YOUR KNOWLEDGE

11.3 Significance tests for regression coefficients. As part of a study on undergraduate success among actuarial students a multiple regression was run using 82 students.³ The following table contains the estimated coefficients and standard errors:

Variable	Estimate	SE
Intercept	-0.764	0.651
SAT Math	0.00156	0.00074
SAT Verbal	0.00164	0.00076
High school rank	1.470	0.430
College placement exam	0.889	0.402

(a) All the estimated coefficients for the explanatory variables are positive. Is this what you would expect? Explain.

(b) What are the degrees of freedom for the model and error?

(c) Test the significance of each coefficient and state your conclusions.

11.4 ANOVA table for multiple regression. Use the following information and the general form of the ANOVA table for multiple regression on page 617 to perform the ANOVA F test and compute R^2 .

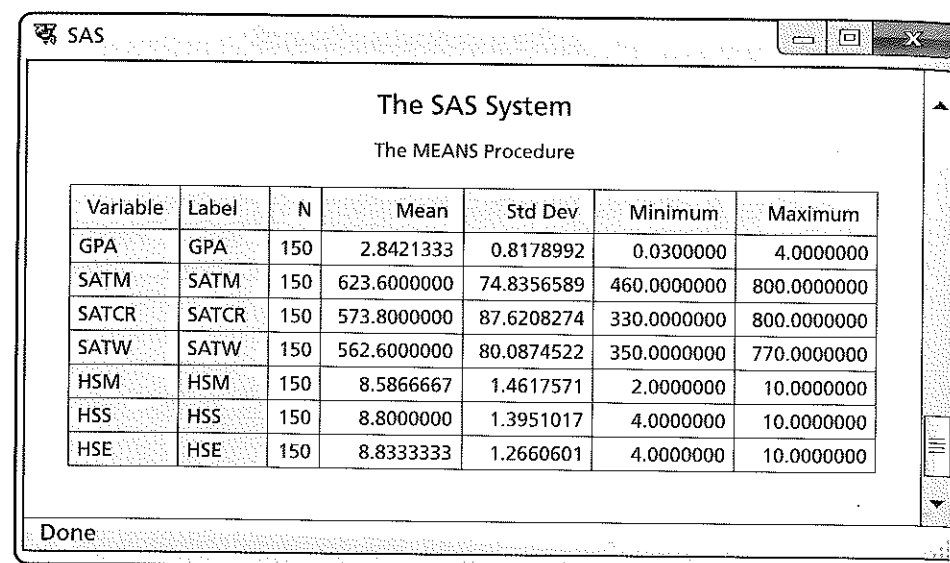
Source	Degrees of freedom	Sum of squares
Model		75
Error	53	
Total	57	594

11.2 A Case Study

Preliminary analysis

In this section we illustrate multiple regression by analyzing the data from the study described in Example 11.1. The response variable is the cumulative GPA, on a 4-point scale, after three semesters. The explanatory variables previously mentioned are average high school grades, represented by HSM, HSS, and HSE. We also examine the SAT Mathematics (SATM), SAT Critical Reading (SATCR), and SAT Writing (SATW) scores as explanatory variables. We have data for $n = 150$ students in the study. We use SAS, Excel, and Minitab to illustrate the outputs that are given by most software.

The first step in the analysis is to carefully examine each of the variables. Means, standard deviations, and minimum and maximum values appear in Figure 11.2. The minimum value for high school mathematics (HSM) appears to be rather extreme; it is $(8.59 - 2.00)/1.46 = 4.51$ standard deviations below

FIGURE 11.2 Descriptive statistics for the College of Science student case study.


The SAS System

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
GPA	GPA	150	2.8421333	0.8178992	0.0300000	4.0000000
SATM	SATM	150	623.6000000	74.8356589	460.0000000	800.0000000
SATCR	SATCR	150	573.8000000	87.6208274	330.0000000	800.0000000
SATW	SATW	150	562.6000000	80.0874522	350.0000000	770.0000000
HSM	HSM	150	8.5866667	1.4617571	2.0000000	10.0000000
HSS	HSS	150	8.8000000	1.3951017	4.0000000	10.0000000
HSE	HSE	150	8.8333333	1.2660601	4.0000000	10.0000000

Done

the mean. Similarly, the minimum value for GPA is 3.43 standard deviations below the mean. We do not discard either of these cases at this time but will take care in our subsequent analyses to see if they have an excessive influence on our results.

The mean for the SATM score is higher than the means for the Critical Reading (SATCR) and Writing (SATW) scores, as we might expect for a group of science majors. The three SAT standard deviations are all about the same.

Although mathematics scores were higher on the SAT, the means and standard deviations of the three high school grade variables are very similar. Since the level and difficulty of high school courses vary within and across schools, this may not be that surprising. The mean GPA is 2.842 on a 4-point scale, with standard deviation 0.818.

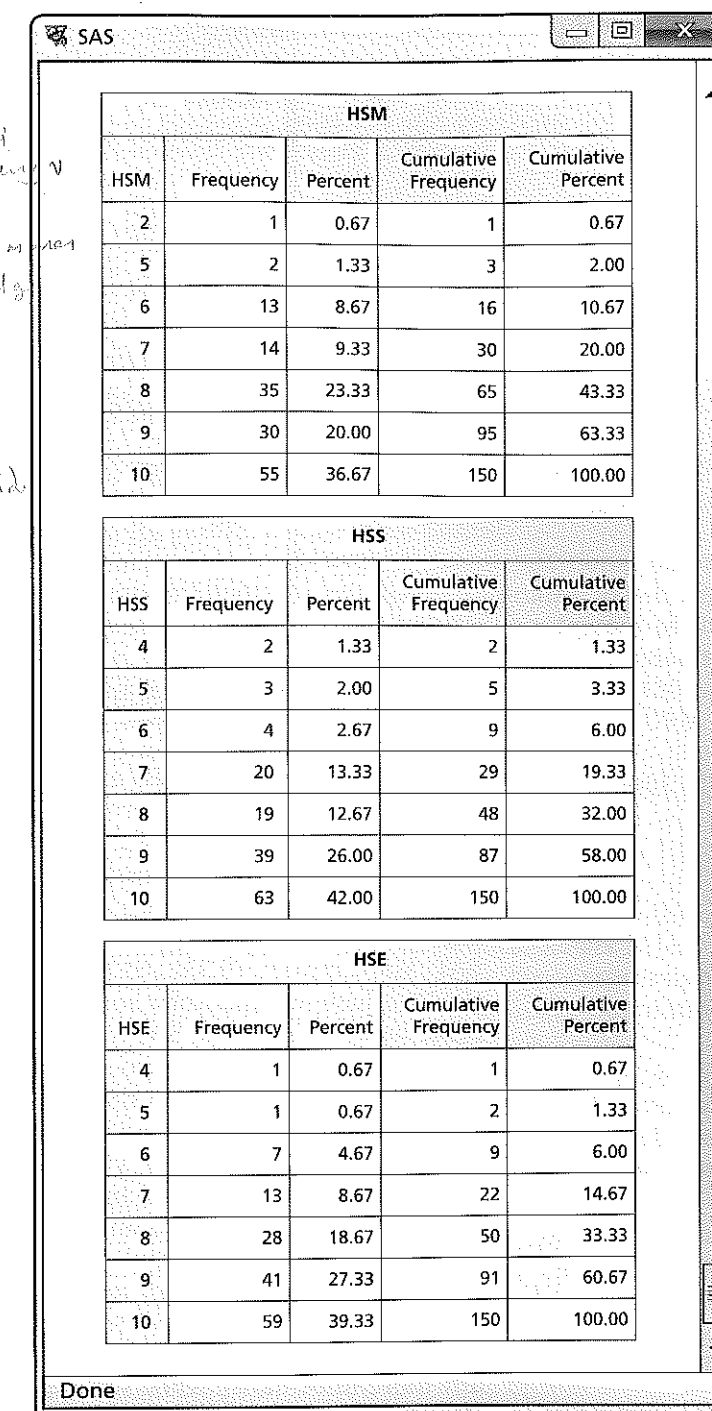
Because the variables GPA, SATM, SATCR, and SATW have many possible values, we could use stemplots or histograms to examine the shapes of their distributions. Normal quantile plots indicate whether or not the distributions look Normal. *It is important to note that the multiple regression model does not require any of these distributions to be Normal. Only the deviations of the responses y from their means are assumed to be Normal.*

The purpose of examining these plots is to understand something about each variable alone before attempting to use it in a complicated model. *Extreme values of any variable should be noted and checked for accuracy.* If found to be correct, the cases with these values should be carefully examined to see if they are truly exceptional and perhaps do not belong in the same analysis with the other cases. When our data on science majors are examined in this way, no obvious problems are evident.

The high school grade variables HSM, HSS, and HSE have relatively few values and are best summarized by giving the relative frequencies for each possible value. The output in Figure 11.3 provides these summaries. The distributions are all skewed, with a large proportion of high grades (10 = A and 9 = A-.) Again we emphasize that these distributions need not be Normal.

FIGURE 11.3 The distributions of the high school grade variables.

Lin = plot R mot Predikate vardi
 " R mot hves av frekvens v
 variabelen
 - uavhengighet: R mot observasjonsummer
 hvis dataene er hendet i rekkefølge
 - konstant varians: R mot PV
 absolute vardi av R mot PV
 - normalitet - normalplot av R
 - Histogram over Resid.



HSM

HSM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	1	0.67	1	0.67
5	2	1.33	3	2.00
6	13	8.67	16	10.67
7	14	9.33	30	20.00
8	35	23.33	65	43.33
9	30	20.00	95	63.33
10	55	36.67	150	100.00

HSS

HSS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	2	1.33	2	1.33
5	3	2.00	5	3.33
6	4	2.67	9	6.00
7	20	13.33	29	19.33
8	19	12.67	48	32.00
9	39	26.00	87	58.00
10	63	42.00	150	100.00

HSE

HSE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	1	0.67	1	0.67
5	1	0.67	2	1.33
6	7	4.67	9	6.00
7	13	8.67	22	14.67
8	28	18.67	50	33.33
9	41	27.33	91	60.67
10	59	39.33	150	100.00

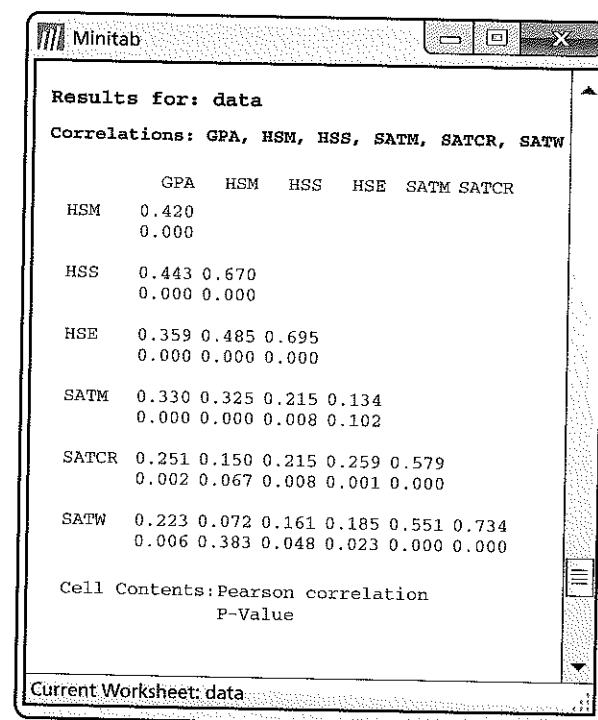
Done

Relationships between pairs of variables

The second step in our analysis is to examine the relationships between all pairs of variables. Scatterplots and correlations are our tools for studying two-variable relationships. The correlations appear in Figure 11.4. The output includes the P -value for the test of the null hypothesis that the population correlation is 0 versus the two-sided alternative for each pair. Thus, we see that

← LOOK BACK
 correlation, p. 103

FIGURE 11.4 Correlations among the case study variables.



the correlation between GPA and HSM is 0.42, with a P -value of 0.000 (that is, $P < 0.0005$), whereas the correlation between GPA and SATW is 0.22, with a P -value of 0.006. Because of the large sample size, even somewhat weak associations are found to be statistically significant.

As we might expect, math and science grades have the highest correlation with GPA ($r = 0.42$ and $r = 0.44$), followed by English grades (0.36) and then SAT Mathematics (0.33). SAT Critical Reading (SATCR) and SAT Writing (SATW) have comparable, somewhat weak, correlations with GPA. On the other hand, SATCR and SATW have a high correlation with each other (0.73). The high school grades also correlate well with each other (0.49 to 0.70). SATM correlates well with the other SAT scores (0.58 and 0.55), somewhat with HSM (0.32), less with HSS (0.22), and poorly with HSE (0.13). SATCR and SATW do not correlate well with any of the high school grades (0.07 to 0.26).

It is important to keep in mind that by examining pairs of variables we are seeking a better understanding of the data. *The fact that the correlation of a particular explanatory variable with the response variable does not achieve statistical significance does not necessarily imply that it will not be a useful (and statistically significant) predictor in a multiple regression.*

Numerical summaries such as correlations are useful, but plots are generally more informative when seeking to understand data. Plots tell us whether the numerical summary gives a fair representation of the data.

For a multiple regression, each pair of variables should be plotted. For the seven variables in our case study, this means that we should examine 21 plots. In general, there are $p + 1$ variables in a multiple regression analysis with p explanatory variables, so that $p(p + 1)/2$ plots are required. *Multiple regression is a complicated procedure. If we do not do the necessary preliminary work, we are in serious danger of producing useless or misleading results.* We leave the task of making these plots as an exercise.

USE YOUR KNOWLEDGE

11.5 Pairwise relationships among variables in the GPA data set. Using a statistical package, generate the pairwise correlations and scatterplots discussed previously. Comment on any unusual patterns or observations.

Regression on high school grades

To explore the relationship between the explanatory variables and our response variable GPA, we run several multiple regressions. The explanatory variables fall into two classes. High school grades are represented by HSM, HSS, and HSE, and standardized tests are represented by the three SAT scores. We begin our analysis by using the high school grades to predict GPA. Figure 11.5 gives the multiple regression output.

The output contains an ANOVA table, some additional descriptive statistics, and information about the parameter estimates. When examining any ANOVA table, it is a good idea to first verify the degrees of freedom. This ensures that we have not made some serious error in specifying the model for the software or in entering the data. Because there are $n = 150$ cases, we have $DFT = n - 1 = 149$. The three explanatory variables give $DFM = p = 3$ and $DFE = n - p - 1 = 150 - 3 - 1 = 146$.

The ANOVA F statistic is 14.35, with a P -value of < 0.0001 . Under the null hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

the F statistic has an $F(3, 146)$ distribution. According to this distribution, the chance of obtaining an F statistic of 14.35 or larger is less than 0.0001. We therefore conclude that at least one of the three regression coefficients for the high school grades is different from 0 in the population regression equation.

In the descriptive statistics that follow the ANOVA table we find that Root MSE is 0.726. This value is the square root of the MSE given in the ANOVA table and is s , the estimate of the parameter σ of our model. The value of R^2 is 0.23. That is, 23% of the observed variation in the GPA scores is explained by linear regression on high school grades.

Although the P -value of the F test is very small, the model does not explain very much of the variation in GPA. Remember, a small P -value does not necessarily tell us that we have a strong predictive relationship, particularly when the sample size is large.

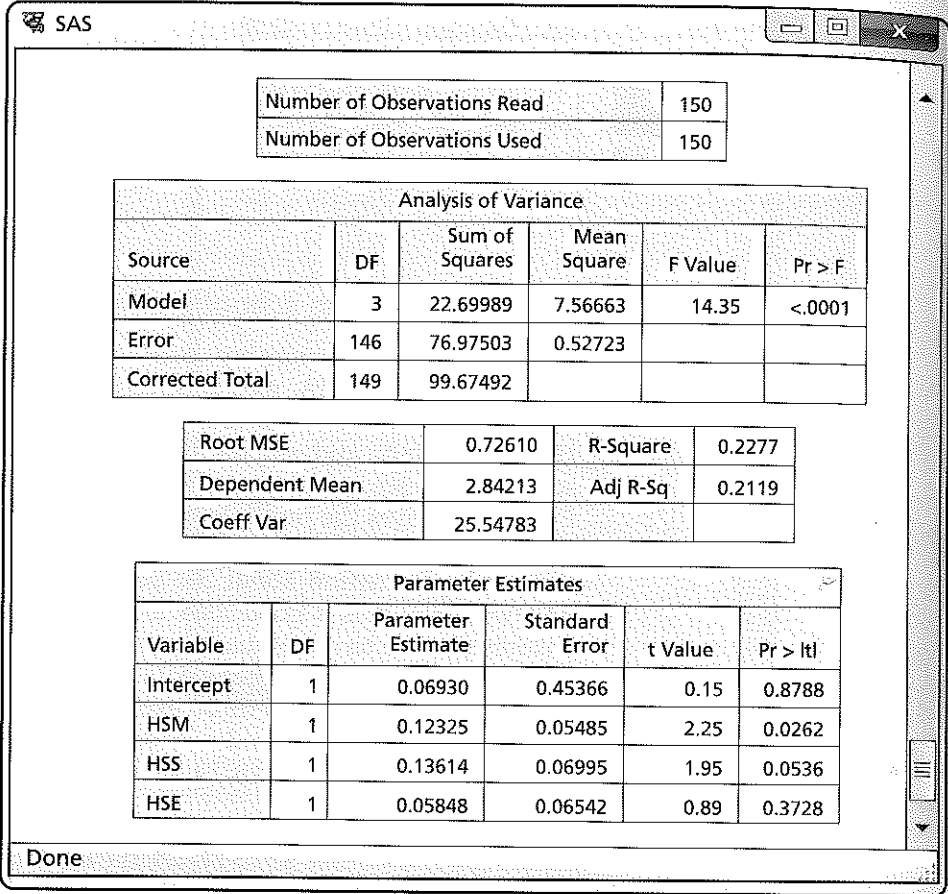
From the Parameter Estimates section of the computer output we obtain the fitted regression equation

$$\widehat{\text{GPA}} = 0.069 + 0.123\text{HSM} + 0.136\text{HSS} + 0.058\text{HSE}$$

Let's find the predicted GPA for a student with an A- average in HSM, B+ in HSS, and B in HSE. The explanatory variables are $\text{HSM} = 9$, $\text{HSS} = 8$, and $\text{HSE} = 7$. The predicted GPA is

$$\begin{aligned}\widehat{\text{GPA}} &= 0.069 + 0.123(9) + 0.136(8) + 0.058(7) \\ &= 2.67\end{aligned}$$

FIGURE 11.5 Multiple regression output for regression using high school grades to predict GPA.



Recall that the t statistics for testing the regression coefficients are obtained by dividing the estimates by their standard errors. Thus, for the coefficient of HSM we obtain the t -value given in the output by calculating

$$t = \frac{b}{SE_b} = \frac{0.12325}{0.05485} = 2.25$$

The P -values appear in the last column. Note that these P -values are for the two-sided alternatives. HSM has a P -value of 0.0262, and we conclude that the regression coefficient for this explanatory variable is significantly different from 0. The P -values for the other explanatory variables (0.0536 for HSS and 0.3728 for HSE) do not achieve statistical significance.

Interpretation of results

The significance tests for the individual regression coefficients seem to contradict the impression obtained by examining the correlations in Figure 11.4. In that display we see that the correlation between GPA and HSS is 0.44 and the correlation between GPA and HSE is 0.36. The P -values for both of these correlations are < 0.0005 . In other words, if we used HSS alone in a regression to predict GPA, or if we used HSE alone, we would obtain statistically significant regression coefficients.

This phenomenon is not unusual in multiple regression analysis. Part of the explanation lies in the correlations between HSM and the other two

explanatory variables. These are rather high (at least compared with most other correlations in Figure 11.4). The correlation between HSM and HSS is 0.67, and that between HSM and HSE is 0.49. Thus, when we have a regression model that contains all three high school grades as explanatory variables, there is considerable overlap of the predictive information contained in these variables.



The significance tests for individual regression coefficients assess the significance of each predictor variable assuming that all other predictors are included in the regression equation. Given that we use a model with HSM and HSS as predictors, the coefficient of HSE is not statistically significant. Similarly, given that we have HSM and HSE in the model, HSS does not have a significant regression coefficient. HSM, however, adds significantly to our ability to predict GPA even after HSS and HSE are already in the model.

Unfortunately, we cannot conclude from this analysis that the pair of explanatory variables HSS and HSE contribute nothing significant to our model for predicting GPA once HSM is in the model. Questions like these require fitting additional models.

The impact of relations among the several explanatory variables on fitting models for the response is the most important new phenomenon encountered in moving from simple linear regression to multiple regression. In this chapter, we can only illustrate some of the many complicated problems that can arise.

Residuals

As in simple linear regression, we should always examine the residuals as an aid to determining whether the multiple regression model is appropriate for the data. Because there are several explanatory variables, we must examine several residual plots. It is usual to plot the residuals versus the predicted values \hat{y} and also versus each of the explanatory variables. Look for outliers, influential observations, evidence of a curved (rather than linear) relation, and anything else unusual. Again, we leave the task of making these plots as an exercise. The plots all appear to show more or less random noise above and below the center value of 0.

If the deviations ϵ in the model are Normally distributed, the residuals should be Normally distributed. Figure 11.6 presents a Normal quantile plot and histogram of the residuals. Both suggest some skewness (shorter right tail) in the distribution. However, given our large sample size, we do not think this skewness is strong enough to invalidate this analysis.

USE YOUR KNOWLEDGE

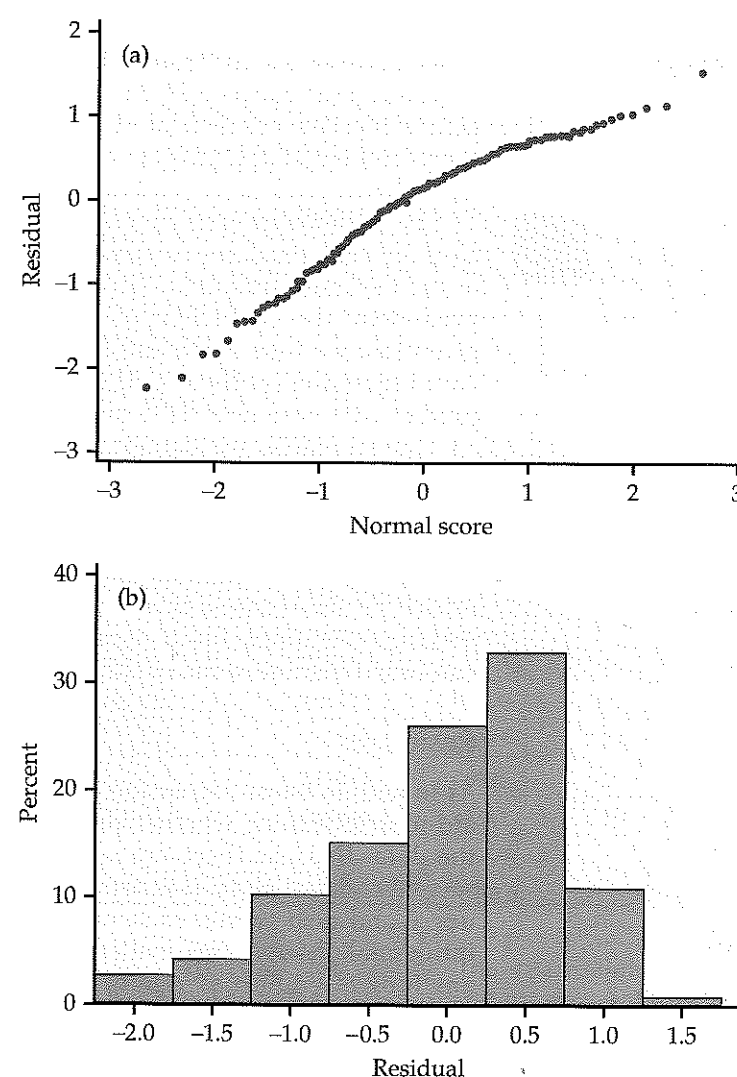


11.6 Residual plots for the GPA analysis. Using a statistical package, fit the linear model with HSM and HSE as predictors and obtain the residuals and predicted values. Plot the residuals versus the predicted values, HSM, and HSE. Are the residuals more or less randomly dispersed around zero? Comment on any unusual patterns.

Refining the model

Because the variable HSE has the largest P -value of the three explanatory variables (see Figure 11.5) and therefore appears to contribute the least to our explanation of GPA, we rerun the regression using only HSM and HSS as explanatory

FIGURE 11.6 (a) Normal quantile plot and (b) histogram of the residuals from the high school grades model. There are no important deviations from Normality.



variables. The SAS output appears in Figure 11.7. The F statistic indicates that we can reject the null hypothesis that the regression coefficients for the two explanatory variables are both 0. The P -value is still <0.0001 . The value of R^2 has dropped very slightly compared with our previous run, from 0.2277 to 0.2235. Thus, dropping HSE from the model resulted in the loss of very little explanatory power.

The measure s of variation about the fitted equation (Root MSE in the printout) is nearly identical for the two regressions, another indication that we lose very little when we drop HSE. The t statistics for the individual regression coefficients indicate that HSM is still significant ($P = 0.0240$), while the statistic for HSS is larger than before (2.99 versus 1.95) and is now statistically significant ($P = 0.0032$).

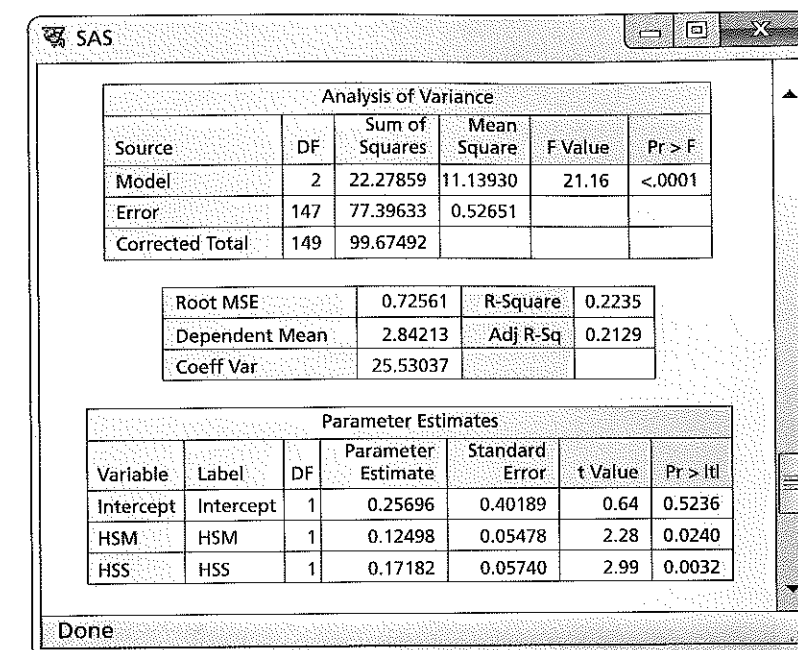
Comparison of the fitted equations for the two multiple regression analyses tells us something more about the intricacies of this procedure. For the first run we have

$$\widehat{\text{GPA}} = 0.069 + 0.123\text{HSM} + 0.136\text{HSS} + 0.058\text{HSE}$$

whereas the second gives us

$$\widehat{\text{GPA}} = 0.257 + 0.125\text{HSM} + 0.172\text{HSS}$$

FIGURE 11.7 Multiple regression output for regression using HSM and HSS to predict GPA.



Regression on SAT scores

We now turn to the problem of predicting GPA using the three SAT scores. Figure 11.8 gives the output. The fitted model is

$$\widehat{\text{GPA}} = 0.45797 + 0.00301\text{SATM} + 0.00080\text{SATCR} + 0.00008\text{SATW}$$

The degrees of freedom are as expected: 3, 146, and 149. The F statistic is 6.28, with a P -value of 0.0005. We conclude that the regression coefficients for SATM, SATCR, and SATW are not all 0. Recall that we obtained the P -value <0.0001 when we used high school grades to predict GPA. Both multiple regression equations are highly significant, but this obscures the fact that the two models have quite different explanatory power. For the SAT regression, $R^2 = 0.1143$, whereas for the high school grades model even with only HSM and HSS (Figure 11.7), we have $R^2 = 0.2235$, a value almost twice as large. *Stating that we have a statistically significant result is quite different from saying that an effect is large or important.*

Further examination of the output in Figure 11.8 reveals that the coefficient of SATM is significant ($t = 2.81$, $P = 0.0056$), and that SATCR ($t = 0.71$, $P = 0.4767$) and SATW ($t = 0.07$, $P = 0.9479$) are not. For a complete analysis we should carefully examine the residuals. Also, we might want to run the analysis without SATW and the analysis with SATM as the only explanatory variable.

FIGURE 11.8 Multiple regression output for regression using SAT scores to predict GPA.

SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11.38939	3.79646	6.28	0.0005
Error	146	88.28553	0.60470		
Corrected Total	149	99.67492			

Root MSE	0.77762	R-Square	0.1143
Dependent Mean	2.84213	Adj R-Sq	0.0961
Coeff Var	27.36049		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.45797	0.56657	0.81	0.4202
SATM	SATM	1	0.00301	0.00107	2.81	0.0056
SATCR	SATCR	1	0.00080324	0.00113	0.71	0.4767
SATW	SATW	1	0.00007882	0.00120	0.07	0.9479

Done

Regression using all variables

We have seen that fitting a model using either the high school grades or the SAT scores results in a highly significant regression equation. The mathematics component of each of these groups of explanatory variables appears to be a key predictor. Comparing the values of R^2 for the two models indicates that high school grades are better predictors than SAT scores. Can we get a better prediction equation using all the explanatory variables together in one multiple regression?

To address this question we run the regression with all six explanatory variables. The output from SAS, Minitab, and Excel appears in Figure 11.9. Although the format and organization of outputs differ among software packages, the basic results that we need are easy to find.

The degrees of freedom are as expected: 6, 143, and 149. The F statistic is 8.95, with a P -value < 0.0001, so at least one of our explanatory variables has a nonzero regression coefficient. This result is not surprising, given that we have already seen that HSM and SATM are strong predictors of GPA. The value of R^2 is 0.2730, which is about 0.05 higher than the value of 0.2235 that we found for the high school grades regression.

Examination of the t statistics and the associated P -values for the individual regression coefficients reveals a surprising result. None of the variables are significant! At first, this result may appear to contradict the ANOVA results. How can the model explain over 27% of the variation and have t tests that suggest none of the variables make a significant contribution?

Once again it is important to understand that these t tests assess the contribution of each variable when it is added to a model that already has the other five explanatory variables. This result does not necessarily mean that the regression coefficients for the six explanatory variables are all 0. It simply means that the contribution of each variable overlaps considerably with the contribution of the other five variables already in the model.

FIGURE 11.9 Multiple regression output for regression using all variables to predict GPA.

SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	27.21030	4.53505	8.95	<.0001
Error	143	72.46462	0.50675		
Corrected Total	149	99.67492			

Root MSE	0.71186	R-Square	0.2730
Dependent Mean	2.84213	Adj R-Sq	0.2425
Coeff Var	25.04670		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-1.18678	0.61641	-1.93	0.0562
SATM	SATM	1	0.00199	0.00106	1.88	0.0619
SATCR	SATCR	1	0.00015701	0.00105	0.15	0.8813
SATW	SATW	1	0.00047398	0.00112	0.42	0.6719
HSM	HSM	1	0.09148	0.05718	1.60	0.1119
HSS	HSS	1	0.13010	0.06877	1.89	0.0605
HSE	HSE	1	0.05679	0.06568	0.86	0.3887

Test SAT Results for Dependent Variable GPA				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	1.50347	2.97	0.0341
Denominator	143	0.50675		

Test HS Results for Dependent Variable GPA				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	5.27364	10.41	<.0001
Denominator	143	0.50675		

Done

Minitab

The regression equation is

$$\text{GPA} = -1.19 + 0.00199 \text{ SATM} + 0.00016 \text{ SATCR} + 0.00047 \text{ SATW} \\ + 0.0915 \text{ HSM} + 0.130 \text{ HSS} + 0.0568 \text{ HSE}$$

Predictor	Coef	SE Coef	T	P
Constant	-1.1868	0.6164	-1.93	0.056
SATM	0.001989	0.001057	1.88	0.062
SATCR	0.000157	0.001049	0.15	0.881
SATW	0.000474	0.001117	0.42	0.672
HSM	0.09148	0.05718	1.60	0.112
HSS	0.13010	0.06877	1.89	0.061
HSE	0.05679	0.06568	0.86	0.389

$$S = 0.711861 \quad R\text{-Sq} = 27.3\% \quad R\text{-Sq(adj)} = 24.2\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	27.2103	4.5350	8.95	0.000
Residual Error	143	72.4646	0.5067		
Total	149	99.6749			

Welcome to Minitab, press F1 for help.

The regression equation is					
GPA = -1.19 + 0.00199 SATM + 0.00016 SATCR + 0.00047 SATW + 0.0915 HSM + 0.130 HSS + 0.0568 HSE					

Predictor	Coef	SE Coef	T	P
Constant	-1.1868	0.6164	-1.93	0.056
SATM	0.001989	0.001057	1.88	0.062
SATCR	0.000157	0.001049	0.15	0.881
SATW	0.000474	0.001117	0.42	0.672
HSM	0.09148	0.05718	1.60	0.112
HSS	0.13010	0.06877	1.89	0.061
HSE	0.05679	0.06568	0.86	0.389

S = 0.711861 R-Sq = 27.3% R-Sq(adj) = 24.2%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	6	27.2103	4.5350	8.95	0.000
Residual Error	143	72.4646	0.5067		
Total	149	99.6749			

FIGURE 11.9 Continued

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.522484872					
5	R Square	0.272990441					
6	Adjusted R Square	0.242486543					
7	Standard Error	0.711860645					
8	Observations	150					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	6	27.21029964	4.53505	8.949363	2.69075E-08	
13	Residual	143	72.46461769	0.506746			
14	Total	149	99.67491733				
15							
16		Coefficients	Standard Error	t Stat	P-Value	Lower 95%	Upper 95%
17	Intercept	-1.18678259	0.616408136	-1.92532	0.056174	-2.40523174	0.031666565
18	SATM	0.001988735	0.001056719	1.881996	0.061869	-0.00010007	0.004077537
19	SATCR	0.000157014	0.001049462	0.149614	0.88128	-0.00191745	0.002231478
20	SATW	0.000473983	0.001116795	0.424414	0.671902	-0.00173358	0.002681542
21	HSM	0.091476942	0.057181116	1.599775	0.111855	-0.02155252	0.204506408
22	HSS	0.130096576	0.068767787	1.891824	0.060536	-0.00583617	0.266029324
23	HSE	0.056790708	0.065681417	0.864639	0.388685	-0.07304124	0.186622652

When a model has a large number of insignificant variables, it is common to refine the model. We prefer smaller models to larger models because they are easier to work with and understand. However, given the many complications that can arise in multiple regression, there is no universal “best” approach to refine a model. There is also no guarantee that there is just one acceptable refined model.

Many statistical software packages now provide the capability of summarizing all possible models from a set of P variables. We suggest using this capability to reduce the number of candidate models (for example, there are a total of 63 models when $p = 6$) and then carefully studying the remaining models before making a decision as to a best model or set of best models. If in doubt, consult an expert.

Test for a collection of regression coefficients

Many statistical software packages also provide the capability for testing whether a collection of regression coefficients in a multiple regression model

are all 0. We use this approach to address two interesting questions about our data set. We did not discuss such tests in the outline that opened this section, but the basic idea is quite simple and discussed in Exercise 11.26 (page 637).

In the context of the multiple regression model with all six predictors, we ask first whether or not the coefficients for the three SAT scores are all 0. In other words, do the SAT scores add any significant predictive information to that already contained in the high school grades? To be fair, we also ask the complementary question: Do the high school grades add any significant predictive information to that already contained in the SAT scores?

The answers are given in the last two parts of the SAS output in Figure 11.9. For the first test we see that $F = 2.97$. Under the null hypothesis that the three SAT coefficients are 0, this statistic has an $F(3, 143)$ distribution and the P -value is 0.0341. We conclude that the SAT scores (as a group) are significant predictors of GPA in a regression that already contains the high school scores as predictor variables. This means that we cannot just focus on refined models that involve the high school grades. Both high school grades and SAT scores appear to contribute to our explanation of GPA.

The test statistic for the three high school grade variables is $F = 10.41$. Under the null hypothesis that these three regression coefficients are 0, the statistic has an $F(3, 143)$ distribution and the P -value is < 0.0001 . Again this means that high school grades contain useful information for predicting GPA that is not contained in the SAT scores.

BEYOND THE BASICS

Multiple logistic regression

Many studies have yes/no or success/failure response variables. A surgery patient lives or dies; a consumer does or does not purchase a product after viewing an advertisement. Because the response variable in a multiple regression is assumed to have a Normal distribution, this methodology is not suitable for predicting such responses. However, there are models that apply the ideas of regression to response variables with only two possible outcomes.

One type of model that can be used is called **logistic regression**. We think in terms of a binomial model for the two possible values of the response variable and use one or more explanatory variables to explain the probability of success. Details are more complicated than those for multiple regression and are given in Chapter 14. However, the fundamental ideas are very much the same. Here is an example.

EXAMPLE

11.2 Tipping behavior in Canada. The Consumer Report on Eating Share Trends (CREST) contains data spanning all provinces of Canada and details away-from-home food purchases by roughly 4000 households per quarter. Some researchers accessed these data but restricted their attention to restaurants at which tips would normally be given.⁴ From a total of 73,822 observations, “high” and “low” tipping variables were created based on whether the observed tip rate was above 20% or below 10%, respectively. They then used logistic regression to identify explanatory variables associated with either “high” or “low” tips.

logistic regression



The model consisted of over 25 explanatory variables, grouped as “control” variables and “stereotype-related” variables. The stereotype-related explanatory variables were x_1 , a variable having the value 1 if the age of the diner was greater than 65 years, and 0 otherwise; x_2 , coded as 1 if the meal was on Sunday, and 0 otherwise; x_3 , coded as 1 to indicate English was a second language; x_4 , a variable coded 1 if the diner was a French-speaking Canadian; x_5 , a variable coded 1 if alcoholic drinks were served with the meal; and x_6 , a variable coded 1 if the meal involved a lone male.

Similar to the F test in multiple regression, there is a chi-square test for multiple logistic regression that tests the null hypothesis that *all* coefficients of the explanatory variables are zero. These results were not presented in the article because the focus was more on comparing the high- and low-tip models. In place of the t tests for individual coefficients in multiple regression, chi-square tests, each with 1 degree of freedom, are used to test whether individual coefficients are zero. The article does report these tests. A majority of the variables considered in the models have P -values less than 0.01.

Interpretation of the coefficients is a little more difficult in multiple logistic regression because of the form of the model. For example, the high-tip model (using only the stereotype-related variables) is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_6 x_6$$

The expression $p/(1-p)$ is the **odds** that the tip was above 20%. Logistic regression models the “log odds” as a linear combination of the explanatory variables. Positive coefficients are associated with a higher probability that the tip is high. These coefficients are often transformed back (e^{β_j}) to the odds scale, giving us an **odds ratio**. An odds ratio greater than 1 is associated with a higher probability that the tip is high. Here is the table of odds ratios reported in the article for the high-tip model:

Explanatory variable	Odds ratio
Senior adult	0.7420*
Sunday	0.9970*
English as second language	0.7360*
French-speaking Canadian	0.7840*
Alcoholic drinks	1.1250*
Lone male	1.0220

The starred values were significant at the 0.01 level. We see that the probability of a high tip is reduced (odds ratio less than 1) when the diner is over 65 years old, speaks English as a second language, and is a French-speaking Canadian. The probability of a high tip is increased (odds ratio greater than 1) if alcohol is served with the meal.

CHAPTER 11 Summary

Data for multiple linear regression consist of the values of a response variable y and P explanatory variables x_1, x_2, \dots, x_p for n cases. We write the data and enter them into software in the form

Individual	Variables				
	x_1	x_2	...	x_p	y
1	x_{11}	x_{12}	...	x_{1p}	y_1
2	x_{21}	x_{22}	...	x_{2p}	y_2
...
n	x_{n1}	x_{n2}	...	x_{np}	y_n

The statistical model for **multiple linear regression** with response variable y and P explanatory variables x_1, x_2, \dots, x_p is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where $i = 1, 2, \dots, n$. The ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ . The **parameters** of the model are $\beta_0, \beta_1, \beta_2, \dots, \beta_p$, and σ .

The **multiple regression equation** predicts the response variable by a linear relationship with all the explanatory variables:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

The β 's are estimated by $b_0, b_1, b_2, \dots, b_p$, which are obtained by the **method of least squares**. The parameter σ is estimated by

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\sum e_i^2}{n - p - 1}}$$

where the e_i are the **residuals**,

$$e_i = y_i - \hat{y}_i$$

Always examine the **distribution of the residuals** and plot them against the explanatory variables prior to inference.

A **level C confidence interval** for β_j is

$$b_j \pm t^* \text{SE}_{b_j}$$

where t^* is the value for the $t(n - p - 1)$ density curve with area C between $-t^*$ and t^* .

The test of the hypothesis $H_0: \beta_j = 0$ is based on the **t statistic**

$$t = \frac{b_j}{\text{SE}_{b_j}}$$

and the $t(n - p - 1)$ distribution.

The estimate b_j of β_j and the test and confidence interval for β_j are all based on a specific multiple linear regression model. The results of all these procedures change if other explanatory variables are added to or deleted from the model.

The **ANOVA table** for a multiple linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The **ANOVA F statistic** is the ratio MSM/MSE and is used to test the null hypothesis

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

If H_0 is true, this statistic has an $F(p, n - p - 1)$ distribution.

The **squared multiple correlation** is given by the expression

$$R^2 = \frac{SSM}{SST}$$

and is interpreted as the proportion of the variability in the response variable y that is explained by the explanatory variables x_1, x_2, \dots, x_p in the multiple linear regression.

CHAPTER 11 Exercises

For Exercise 11.1, see page 614; for Exercise 11.2, see page 615; for Exercises 11.3 and 11.4, see page 619; for Exercise 11.5, see page 623; and for Exercise 11.6, see page 625.

11.7 95% confidence intervals for regression coefficients. In each of the following settings, give a 95% confidence interval for the coefficient of x_1 .

- (a) $n = 26$, $\hat{y} = 1.6 + 6.4x_1 + 5.7x_2$, $SE_{b_1} = 3.1$
- (b) $n = 53$, $\hat{y} = 1.6 + 6.4x_1 + 5.7x_2$, $SE_{b_1} = 2.9$
- (c) $n = 26$, $\hat{y} = 1.6 + 4.8x_1 + 3.2x_2 + 5.2x_3$, $SE_{b_1} = 2.2$
- (d) $n = 124$, $\hat{y} = 1.6 + 4.8x_1 + 3.2x_2 + 5.2x_3$, $SE_{b_1} = 2.1$

11.8 Significance tests for regression coefficients. For each of the settings in the previous exercise, test the null hypothesis that the coefficient of x_1 is zero versus the two-sided alternative.

11.9 What's wrong? In each of the following situations, explain what is wrong and why.

- (a) In a multiple regression with a sample size of 39 and 3 explanatory variables, the test statistic for the null hypothesis $H_0: b_2 = 0$ is a t statistic that follows the $t(35)$ distribution when the null hypothesis is true.
- (b) The multiple correlation coefficient gives the proportion of the variation in the response variable that is explained by the explanatory variables.
- (c) A small P -value for the ANOVA F test implies that all explanatory variables are significantly different from zero.

11.10 What's wrong? In each of the following situations, explain what is wrong and why.

- (a) One of the assumptions for multiple regression is that the distribution of each explanatory variable is Normal.
- (b) The smaller the P -value for the ANOVA F test, the greater the explanatory power of the model.
- (c) All explanatory variables that are significantly correlated with the response variable will have a statistically significant regression coefficient in the multiple regression model.
- (d) The multiple correlation coefficient gives the average correlation between the response variable and each explanatory variable in the model.


11.11 Constructing the ANOVA table. Seven explanatory variables are used to predict a response variable using a multiple regression. There are 142 observations.

- (a) Write the statistical model that is the foundation for this analysis. Also include a description of all assumptions.
- (b) Outline the analysis of variance table giving the sources of variation and numerical values for the degrees of freedom.

11.12 More on constructing the ANOVA table. A multiple regression analysis of 78 cases was performed with 5 explanatory variables. Suppose that $SSM = 16.5$ and $SSE = 100.8$.


- (a) Find the value of the F statistic for testing the null hypothesis that the coefficients of all the explanatory variables are zero.
- (b) What are the degrees of freedom for this statistic?
- (c) Find bounds on the P -value using Table E. Show your work.

(d) What proportion of the variation in the response variable is explained by the explanatory variables?



11.13 Refining the GPA model using all variables. Figure 11.9 (page 629) summarizes the regression model using all variables. Let's now compare several reduced models. For each of the following models, report the fitted model, MSE, percent explained variation, and the P -values for each of the individual coefficients. Based on these results, which model do you think is "best"? Explain your answer. 

- (a) SATM and HSS
- (b) SATM, HSM, and HSS
- (c) SATM, HSM, HSS, and HSE
- (d) HSM and HSS

11.14 Predicting college debt: combining measures. Refer to Exercises 10.10 (page 601) and 10.14 (page 602) for a description of the problem. Let's now consider fitting a model using all the explanatory variables.

 **BESTVALUE**

- (a) Write out the statistical model for this analysis, making sure to specify all assumptions.
- (b) Run the multiple regression model and specify the fitted regression equation.
- (c) Obtain the residuals from part (b) and check assumptions. Comment on any unusual residuals or patterns in the residuals.
- (d) What percent of the variability in average debt is explained by this model?

 **11.15 Predicting college debt: a simpler model.** Refer to the previous exercise. In the multiple regression analysis using all seven variables, only one variable, StudPerFac, is significant at the 0.05 level. Remove the variable with the highest P -value one at a time until you end up with a multiple regression model that has only significant predictors. Summarize your final model in a short paragraph. 

11.16 Comparison of prediction intervals. Refer to the previous two exercises. The Ohio State University has Admit = 68, Yr4Grad = 49, StudPerFac = 19, InAfterAid = 12,680, OutAfterAid = 27,575, AvgAid = 7789, and PercBorrow = 52. Use your software to construct

- (a) a 95% prediction interval based on the model with all the predictors.
- (b) a 95% prediction interval based on the model using your simpler model.
- (c) Compare the two intervals. Do the models give similar predictions and intervals?

11.17 Predicting energy-drink consumption. Energy-drink advertising consistently emphasizes a physically active lifestyle and often features extreme sports and risk taking. Are these typical characteristics of an energy-drink consumer? A researcher decided to examine the links between energy-drink consumption, sport-related (jock) identity, and risk taking.⁵ She invited over 1500 undergraduate students enrolled in large introductory-level courses at a public university to participate. Each participant had to complete a 45-minute anonymous questionnaire. From this questionnaire jock identity and risk-taking scores were obtained, where the higher the score, the stronger the trait. She ended up with 795 respondents. The following table summarizes the results of a multiple regression analysis using the frequency of energy-drink consumption in the past 30 days as the response variable:

Explanatory variable	b
Age	-0.02
Sex (1 = female, 0 = male)	-0.11**
Race (1 = nonwhite, 0 = white)	-0.02
Ethnicity (1 = Hispanic, 0 = non-Hispanic)	0.10**
Parental education	0.02
College GPA	-0.01
Jock identity	0.05
Risk taking	0.19***

A superscript of ** means that the individual coefficient t test had a P -value less than 0.01, and a superscript of *** means that the test had a P -value less than 0.001. All other P -values were greater than 0.05.

- (a) The overall F statistic is reported to be 8.11. What are the degrees of freedom associated with this statistic?
- (b) R is reported to be 0.28. What percent of the variation in energy-drink consumption is explained by the model? Is this a highly predictive model? Explain.
- (c) Interpret each of the regression coefficients that are significant.
- (d) The researcher states, "Controlling for gender, age, race, ethnicity, parental educational achievement, and college GPA, each of the predictors (risk taking and jock identity) was positively associated with energy-drink consumption frequency." Explain what is meant by "controlling for" these variables and how this helps strengthen her assertion that jock identity and risk taking are positively associated with energy-drink consumption.


11.18 Consider the gender of the students. Refer to Exercise 11.13. The seventh predictor variable provided in the GPA data set is a gender indicator variable. This

3911 theaters during the first weekend, grossing \$38.7 million dollars, and had an IMDb rating of 6.8. Use your software to construct

- a 95% prediction interval based on the model with all four predictors.
- a 95% prediction interval based on the model using only opening-weekend revenue and IMDb rating.
- Compare the two intervals. Do the models give similar predictions?

11.28 Effect of potential outliers. Consider the simpler model of Exercise 11.26 for this analysis.

- Two movies have much larger U.S. revenues than predicted. Which are they and how much more revenue did they earn than predicted?
- Remove these two movies and redo the multiple regression. Make a table giving the regression coefficients and their standard errors, t statistics, and P -values.
- Compare these results with those from Exercise 11.26. How does the removal of these outlying movies impact the estimated model?
- Obtain the residuals from this reduced data set and graphically examine their distribution. Do the residuals appear approximately Normal? Explain your answer.

The following three exercises use the RANKINGS data file. Since 2004, The Times Higher Education Supplement has provided an annual ranking of the world universities. A total score for each university is calculated based on the scores for the following explanatory variables: Peer Review (40%); Faculty-to-Student Ratio (20%); Citations-to-Faculty Ratio (20%); Recruiter Review (10%); Percent International Faculty (5%); and Percent International Students (5%). The percents represent the contributions of each score to the total. For our purposes, we will assume that these weights are unknown and will focus on the development of a model for the total score based on the first three explanatory variables. The report includes a table for the top 200 universities.⁸ The RANKINGS data file contains a random sample of 75 of these universities. This is not a random sample of all universities but for our purposes here we will consider it to be.  RANKINGS

11.29 Annual ranking of world universities. Let's consider developing a model to predict total score based on the peer review score (PEER), faculty-to-student ratio (FtoS), and citations-to-faculty ratio (CtoF).


- Using numerical and graphical summaries, describe the distribution of each explanatory variable.
- Using numerical and graphical summaries, describe the relationship between each pair of explanatory variables.

11.30 Looking at the simple linear regressions. Now let's look at the relationship between each explanatory variable and the total score.

- Generate scatterplots for each explanatory variable and the total score. Do these relationships all look linear?
- Compute the correlation between each explanatory variable and the total score. Are certain explanatory variables more strongly associated with the total score?


11.31 Multiple linear regression model. Now consider a regression model using all three explanatory variables.

- Write out the statistical model for this analysis, making sure to specify all assumptions.
- Run the multiple regression model and specify the fitted regression equation.
- Generate a 95% confidence interval for each coefficient. Should any of these intervals contain 0? Explain.
- What percent of the variation in total score is explained by this model? What is the estimate for σ ?

11.32 Predicting GPA of seventh-graders. Refer to the educational data for 78 seventh-grade students given in Table 1.3 (page 29). We view GPA as the response variable. IQ, gender, and self-concept are the explanatory variables.  SEVENGR

- Find the correlation between GPA and each of the explanatory variables. What percent of the total variation in student GPAs can be explained by the straight-line relationship with each of the explanatory variables?
- The importance of IQ in explaining GPA is not surprising. The purpose of the study is to assess the influence of self-concept on GPA. So we will include IQ in the regression model and ask, "How much does self-concept contribute to explaining GPA after the effect of IQ on GPA is taken into account?" Give a model that can be used to answer this question.
- Run the model and report the fitted regression equation. What percent of the variation in GPA is explained by the explanatory variables in your model?
- Translate the question of interest into appropriate null and alternative hypotheses about the model parameters. Give the value of the test statistic and its P -value. Write a short summary of your analysis with an emphasis on your conclusion.

The following three exercises use the HAPPY data file. The World Database of Happiness is an online registry of scientific research on the subjective appreciation of life. It is available at worlddatabaseofhappiness.eur.nl, and the project is directed by Dr. Ruut Veenhoven, Erasmus University, Rotterdam. One inventory presents the "average happiness" score for various nations. This average is based

on individual responses from numerous general population surveys to a general life satisfaction (well-being) question. Scores range from 0 (dissatisfied) to 10 (satisfied). The NationMaster website, www.nationmaster.com, contains a collection of statistics associated with various nations. For our analysis, we will consider the GINI index, which measures the degree of inequality in the distribution of income (higher score = greater inequality); the degree of corruption in government (higher score = less corruption); average life expectancy; and the degree of democracy (higher score = more civil and political liberties).  HAPPY

11.33 Predicting a nation's "average happiness" score. Consider the five statistics for each nation: LSI, the average life-satisfaction score; GINI, the GINI index; CORRUPT, the degree of government corruption; LIFE, the average life expectancy; and DEMOCRACY, a measure of civil and political liberties.

- Using numerical and graphical summaries, describe the distribution of each variable.
- Using numerical and graphical summaries, describe the relationship between each pair of variables.


11.34 Building a multiple linear regression model. Let's now build a model to predict the life-satisfaction score, LSI.

- Consider a simple linear regression using GINI as the explanatory variable. Run the regression and summarize the results. Be sure to check assumptions.
- Now consider a model using GINI and LIFE. Run the multiple regression and summarize the results. Again be sure to check assumptions.
- Now consider a model using GINI, LIFE, and DEMOCRACY. Run the multiple regression and summarize the results. Again be sure to check assumptions.
- Now consider a model using all four explanatory variables. Again summarize the results and check assumptions.

11.35 Selecting from among several models. Refer to the results from the previous exercise.

- Make a table giving the estimated regression coefficients, standard errors, t statistics, and P -values.
- Describe how the coefficients and P -values change for the four models.
- Based on the table of coefficients, suggest another model. Run that model, summarize the results, and compare it with the other ones. Which model would you choose to explain LSI? Explain.

The following six exercises use the BIOMARK data file. Healthy bones are continually being renewed by two processes. Through bone formation, new bone is built; through

bone resorption, old bone is removed. If one or both of these processes are disturbed, by disease, aging, or space travel, for example, bone loss can be the result. The variables VO+ and VO- measure bone formation and bone resorption, respectively. Osteocalcin (OC) is a biochemical marker for bone formation: higher levels of bone formation are associated with higher levels of OC. A blood sample is used to measure OC, and it is much less expensive to obtain than direct measures of bone formation. The units are milligrams of OC per milliliter of blood (mg/ml). Similarly, tartrate-resistant acid phosphatase (TRAP) is a biochemical marker for bone resorption that is also measured in blood. It is measured in units per liter (U/l). These variables were measured in a study of 31 healthy women aged 11 to 32 years.⁹ Variables with the first letter "L" are the logarithms of the measured variables.  BIOMARK

11.36 Bone formation and resorption. Consider the following four variables: VO+, a measure of bone formation; VO-, a measure of bone resorption; OC, a biomarker of bone formation; and TRAP, a biomarker of bone resorption.

- Using numerical and graphical summaries, describe the distribution of each of these variables.
- Using numerical and graphical summaries, describe the relationship between each pair of variables.

11.37 Predicting bone formation. Let's use regression methods to predict VO+, the measure of bone formation.

- Since OC is a biomarker of bone formation, we start with a simple linear regression using OC as the explanatory variable. Run the regression and summarize the results. Be sure to include an analysis of the residuals.
- Because the processes of bone formation and bone resorption are highly related, it is possible that there is some information in the bone resorption variables that can tell us something about bone formation. Use a model with both OC and TRAP, the biomarker of bone resorption, to predict VO+. Summarize the results. In the context of this model, it appears that TRAP is a better predictor of bone formation, VO+, than the biomarker of bone formation, OC. Is this view consistent with the pattern of relationships that you described in the previous exercise? One possible explanation is that, although all these variables are highly related, TRAP is measured with more precision than OC.


11.38 More on predicting bone formation. Now consider a regression model for predicting VO+ using OC, TRAP, and VO-.


- Write out the statistical model for this analysis including all assumptions.
- Run the multiple regression to predict VO+ using OC, TRAP, and VO-. Summarize the results.


(c) Make a table giving the estimated regression coefficients, standard errors, and t statistics with P -values for this analysis and for the two that you ran in the previous exercise. Describe how the coefficients and the P -values differ for the three analyses.


(d) Give the percent of variation in VO+ explained by each of the three models and the estimate of σ . Give a short summary.

(e) The results you found in part (b) suggest another model. Run that model, summarize the results, and compare them with the results in part (b).

 **11.39 Predicting bone formation using transformed variables.** Because the distributions of VO+, VO-, OC, and TRAP tend to be skewed, it is common to work with logarithms rather than the measured values. Using the questions in the previous three exercises as a guide, analyze the log data.

 **11.40 Predicting bone resorption.** Refer to Exercises 11.36 to 11.38. Answer these questions with the roles of VO+ and VO- reversed; that is, run models to predict VO-, with VO+ as an explanatory variable.

 **11.41 Predicting bone resorption using transformed variables.** Refer to the previous exercise. Rerun using logs.

The following 11 exercises use the PCB data file. Polychlorinated biphenyls (PCBs) are a collection of synthetic compounds, called congeners, that are particularly toxic to fetuses and young children. Although PCBs are no longer produced in the United States, they are still found in the environment. Since human exposure to these PCBs is primarily through the consumption of fish, the Environmental Protection Agency (EPA) monitors the PCB levels in fish. Unfortunately, there are 209 different congeners, and measuring all of them in a fish specimen is an expensive and time-consuming process. You've been asked to see if the total amount of PCBs in a specimen can be estimated with only a few, easily quantifiable congeners.¹⁰ If this can be done, costs can be greatly reduced.  PCB

11.42 Relationships among PCB congeners. Consider the following variables: PCB (the total amount of PCB) and four congeners: PCB52, PCB118, PCB138, and PCB180.

(a) Using numerical and graphical summaries, describe the distribution of each of these variables.

(b) Using numerical and graphical summaries, describe the relationship between each pair of variables.

11.43 Predicting the total amount of PCB. Use the four congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict PCB.

(a) Write the statistical model for this analysis. Include all assumptions.

(b) Run the regression and summarize the results.

(c) Examine the residuals. Do they appear to be approximately Normal? When you plot them versus each of the explanatory variables, are any patterns evident?

11.44 Adjusting the analysis for potential outliers.

The examination of the residuals in part (c) of the previous exercise suggests that there may be two outliers, one with a high residual and one with a low residual.

(a) Because of safety issues, we are more concerned about underestimating PCB in a specimen than about overestimating. Give the specimen number for each of the two suspected outliers. Which one corresponds to an overestimate of PCB?

(b) Rerun the analysis with the two suspected outliers deleted, summarize these results, and compare them with those you obtained in the previous exercise.

11.45 More on predicting the total amount of PCB.

Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis that you did in Exercise 11.43, with the change that PCB180 is not included as an explanatory variable.

(a) Summarize the results.

(b) In this analysis, the regression coefficient for PCB118 is not statistically significant. Give the estimate of the coefficient and the associated P -value.

(c) Find the estimate of the coefficient for PCB118 and the associated P -value for the model analyzed in Exercise 11.43.

(d) Using the results in parts (b) and (c), write a short paragraph explaining how the inclusion of other variables in a multiple regression can have an effect on the estimate of a particular coefficient and the results of the associated significance test.

11.46 Multiple regression model for total TEQ.

Dioxins and furans are other classes of chemicals that can cause undesirable health effects similar to those caused by PCB. The three types of chemicals are combined using toxic equivalent scores (TEQs), which attempt to measure the health effects on a common scale. The PCB data file contains TEQs for PCB, dioxins, and furans. The variables are called TEQPCB, TEQDIOXIN, and TEQFURAN. The data file also includes the total TEQ, defined to be the sum of these three variables.

(a) Consider using a multiple regression to predict TEQ using the three components TEQPCB, TEQDIOXIN, and


TEQFURAN as explanatory variables. Write the multiple regression model in the form


$$\text{TEQ} = \beta_0 + \beta_1 \text{TEQPCB} + \beta_2 \text{TEQDIOXIN} + \beta_3 \text{TEQFURAN} + \epsilon$$

Give numerical values for the parameters β_0 , β_1 , β_2 , and β_3 .

(b) The multiple regression model assumes that the ϵ 's are Normal with mean zero and standard deviation σ . What is the numerical value of σ ?

(c) Use software to run this regression and summarize the results.


 **11.47 Multiple regression model for total TEQ, continued.** The information summarized in TEQ is used to assess and manage risks from these chemicals. For example, the World Health Organization (WHO) has established the tolerable daily intake (TDI) as 1 to 4 TEQs per kilogram of body weight per day. Therefore, it would be very useful to have a procedure for estimating TEQ using just a few variables that can be measured cheaply. Use the four PCB congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict TEQ. Give a description of the model and assumptions, summarize the results, examine the residuals, and write a summary of what you have found.

 **11.48 Predicting total amount of PCB using transformed variables.** Because distributions of variables such as PCB, the PCB congeners, and TEQ tend to be skewed, researchers frequently analyze the logarithms of the measured variables. Create a data set that has the logs of each of the variables in the PCB data file. Note that zero is a possible value for PCB126; most software packages will eliminate these cases when you request a log transformation.

(a) If you do not do anything about the 16 zero values of PCB126, what does your software do with these cases? Is there an error message of some kind?


(b) If you attempt to run a regression to predict the log of PCB using the log of PCB126 and the log of PCB52, are the cases with the zero values of PCB126 eliminated? Do you think that this is a good way to handle this situation?


(c) The smallest nonzero value of PCB126 is 0.0052. One common practice when taking logarithms of measured values is to replace the zeros by one-half of the smallest observed value. Create a logarithm data set using this procedure; that is, replace the 16 zero values of PCB126 by 0.0026 before taking logarithms. Use numerical and graphical summaries to describe the distributions of the log variables.

 **11.49 Predicting total amount of PCB using transformed variables, continued.** Refer to the previous exercise.


(a) Use numerical and graphical summaries to describe the relationships between each pair of log variables.

(b) Compare these summaries with the summaries that you produced in Exercise 11.42 for the measured variables.

 **11.50 Even more on predicting total amount of PCB using transformed variables.** Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of PCB. Use only log PCB variables for this analysis. Write a report summarizing your results.

 **11.51 Predicting total TEQ using transformed variables.** Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of TEQ. Use only log PCB variables for this analysis. Write a report summarizing your results and comparing them with the results that you obtained in the previous exercise.

11.52 Interpretation of coefficients in log PCB regressions. Use the results of your analysis of the log PCB data in Exercise 11.50 to write an explanation of how regression coefficients, standard errors of regression coefficients, and tests of significance for explanatory variables can change depending on what other explanatory variables are included in the multiple regression analysis.

The following nine exercises use the CHEESE data file. As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The variable "Case" is used to number the observations from 1 to 30. "Taste" is the response variable of interest. The taste scores were obtained by combining the scores from several tasters. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide, and lactic acid. For acetic acid and hydrogen sulfide (natural) log transformations were taken. Thus, the explanatory variables are the transformed concentrations of acetic acid ("Acetic") and hydrogen sulfide ("H2S") and the untransformed concentration of lactic acid ("Lactic").¹¹  CHEESE

11.53 Describing the explanatory variables. For each of the four variables in the CHEESE data file, find the mean, median, standard deviation, and interquartile range. Display each distribution by means of a stemplot and use a Normal quantile plot to assess Normality of the data. Summarize your findings. Note that when doing regressions with these data, we do not assume that these distributions are Normal. Only the residuals from our model need to be (approximately) Normal. The careful study of each variable to be analyzed is nonetheless an important first step in any statistical analysis.

11.54 Pairwise scatterplots of the explanatory variables. Make a scatterplot for each pair of variables in the CHEESE data file (you will have six plots). Describe the relationships. Calculate the correlation for each pair of variables and report the P -value for the test of zero population correlation in each case.

11.55 Simple linear regression model of Taste. Perform a simple linear regression analysis using Taste as the response variable and Acetic as the explanatory variable. Be sure to examine the residuals carefully. Summarize your results. Include a plot of the data with the least-squares regression line. Plot the residuals versus each of the other two chemicals. Are any patterns evident? (The concentrations of the other chemicals are lurking variables for the simple linear regression.)

11.56 Another simple linear regression model of Taste. Repeat the analysis of Exercise 11.55 using Taste as the response variable and H2S as the explanatory variable.

11.57 The final simple linear regression model of Taste. Repeat the analysis of Exercise 11.55 using Taste as the response variable and Lactic as the explanatory variable.

11.58 Comparing the simple linear regression models. Compare the results of the regressions performed in the three previous exercises. Construct a table with values of the F statistic, its P -value, R^2 , and the estimate s of the standard deviation for each model. Report the three regression equations. Why are the intercepts in these three equations different?

11.59 Multiple regression model of Taste. Carry out a multiple regression using Acetic and H2S to predict Taste. Summarize the results of your analysis. Compare the statistical significance of Acetic in this model with its significance in the model with Acetic alone as a predictor (Exercise 11.55). Which model do you prefer? Give a simple explanation for the fact that Acetic alone appears to be a good predictor of Taste, but with H2S in the model, it is not.

11.60 Another multiple regression model of Taste. Carry out a multiple regression using H2S and Lactic to predict Taste. When we compare the results of this analysis with the simple linear regressions using each of these explanatory variables alone, it is evident that a better result is obtained by using both predictors in a model. Support this statement with explicit information obtained from your analysis.

11.61 The final multiple regression model of Taste. Use the three explanatory variables Acetic, H2S, and Lactic in a multiple regression to predict Taste. Write a short summary of your results, including an examination of the residuals. Based on all the regression analyses you have carried out on these data, which model do you prefer and why?

11.62 Finding a multiple regression model on the Internet. Search the Internet to find an example of the use of multiple regression. Give the setting of the example, describe the data, give the model, and summarize the results. Explain why the use of multiple regression in this setting was appropriate or inappropriate.



One-Way Analysis of Variance

Introduction

CHAPTER

12

Many of the most effective statistical studies are comparative. For example, we may wish to compare customer satisfaction of men and women who use an online fantasy football site or compare the responses to various treatments in a clinical trial. With a quantitative response, we display these comparisons with back-to-back stemplots or side-by-side boxplots, and we measure them with five-number summaries or with means and standard deviations.

When only two groups are compared, Chapter 7 provides the tools we need to answer the question “Is the difference between groups statistically significant?” Two-sample t procedures compare the means of two Normal populations, and we saw that these procedures, unlike comparisons of spread, are sufficiently robust to be widely useful.

In this chapter, we will compare any number of means by techniques that generalize the two-sample t test and share its robustness and usefulness. These methods will allow us to address comparisons such as

- How does a user’s number of Facebook friends affect his or her social attractiveness?
- On average, which of 5 brands of automobile tires wears longest?
- Among three therapies for lung cancer, is there a difference in average progression-free survival?

- 12.1 Inference for One-Way Analysis of Variance
- 12.2 Comparing the Means