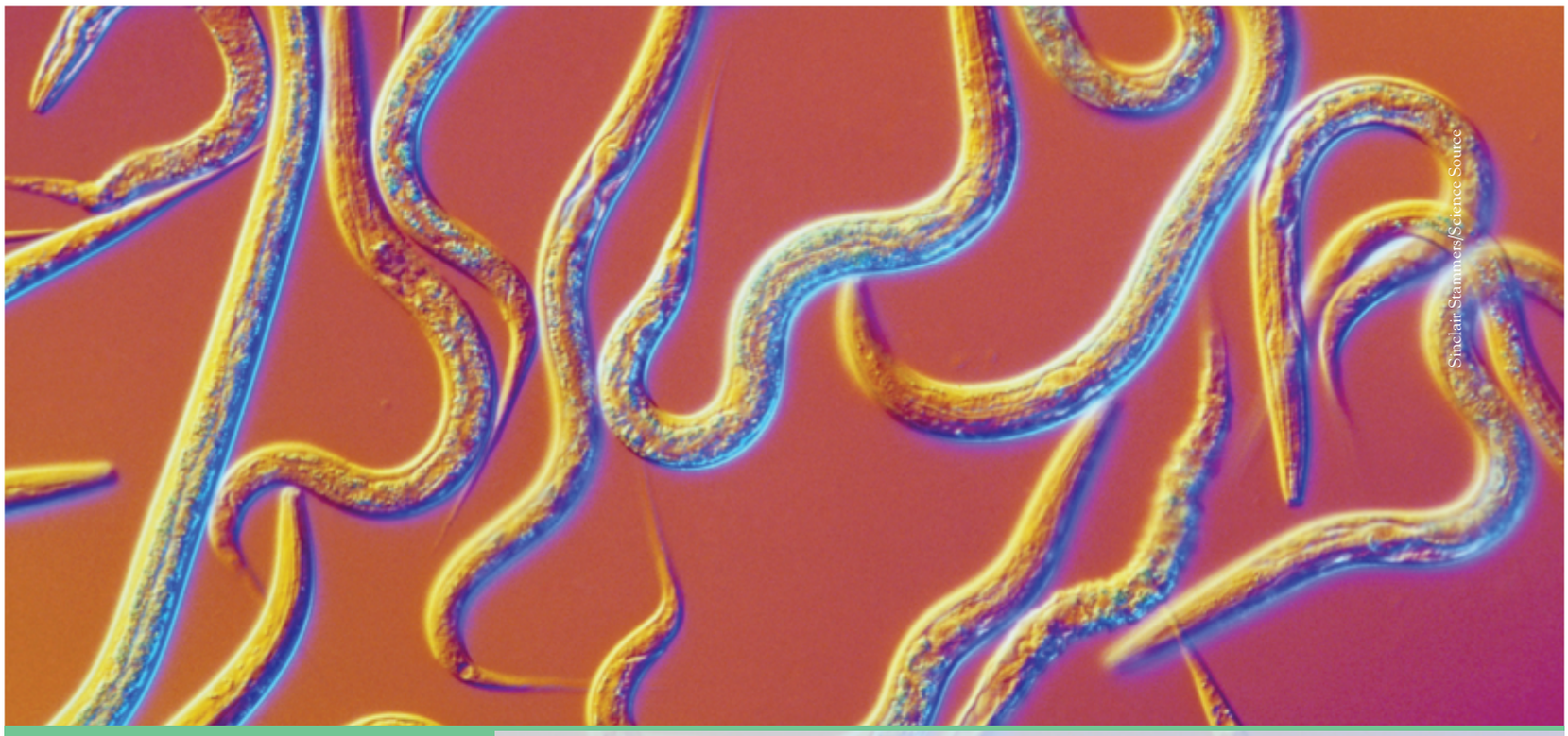


CHAPTER **28**

Multiple and Logistic Regression

When a scatterplot shows a linear relationship between a quantitative explanatory variable x and a quantitative response variable y, we fit a regression line to the data to describe the relationship. We can also use the line to predict the value of y for a given value of x. For example, Chapter 4 uses regression lines to describe relationships between

- the gain in body fat y of overfed individuals and the amount of nonexercise activity x in their daily routine;
- the mating-call frequency y of frogs and outside temperature x; and
- the number y of new adults that join a colony of birds and the percent x of adult birds that return from the previous year.

In all these cases, other explanatory variables might improve our understanding of the response y and help us to better predict y.

- The gain in body fat y of an overfed individual may depend on the amount of daily nonexercise activity x_1 and on the individual's age x_2 and gender x_3.
- Mating-call frequency y may depend on outside temperature x_1 and also on the species x_2 of the frogs we study.

IN THIS CHAPTER WE COVER...

- Parallel regression lines
- Estimating parameters
- Using technology
- Conditions for inference
- Inference for multiple regression
- Interaction
- A case study for multiple regression
- Logistic regression
- Inference for logistic regression

- The number y of new adults in a bird colony may depend on the percent x_1 of returning adults and also on the species x_2 of the birds we study.

simple linear regression
multiple regression

We will now call regression with just one explanatory variable **simple linear regression** to remind us that this is a special case. This chapter introduces the more general case of **multiple regression,** which allows several explanatory variables to combine in explaining a response variable. We start with the simplest case by comparing two parallel regression lines and will gradually build up to increasingly complex models.

Parallel regression lines

In Chapter 3 we learned how to add a categorical variable to a scatterplot by using different colors or plot symbols to indicate the different values of the categorical variable. In Example 3.6 we examined the relationship between the longevity y and the thorax length of male fruit flies under two experimental conditions. The conditions make-up a categorical variable (call it x_2) that takes just two values. The scatterplot shows two *parallel* straight-line patterns linking the response y (longevity) to a quantitative explanatory variable x_1 (thorax length), with one pattern for each value of x_2 (experimental condition).

EXAMPLE 28.1 Cost of reproduction in male fruit flies

Yoav Levy/Phototake

STATE: Longevity in male fruit flies is positively associated with adult size. But reproduction has a high physiological cost that could impact longevity. A study looks at the association between longevity and adult size in male fruit flies kept under one of two conditions. One group is kept with sexually active females over the male's life span. The other group is cared for in the same way but kept with females that are not sexually active. Table 28.1 gives the longevity in days and thorax length in millimeters (mm) for the male fruit flies given an opportunity to reproduce (IndReprod = 1) and for those deprived of the opportunity (IndReprod = 0).[1] Is there evidence that reproduction impacts male fruit fly longevity?

PLAN: Make a scatterplot to display the relationship between longevity y and thorax length x_1. Use different colors for the two reproductive conditions. (So reproduction is a categorical variable x_2 that takes two values.) If both reproductive conditions show linear patterns, fit two separate least-squares regression lines to describe them.

SOLVE: Figure 28.1 shows a scatterplot with two separate regression lines, one for each experimental group. The longevities for fruit flies that never reproduced are higher than the longevities of reproducing fruit flies of similar size. Both regression lines have a positive slope, indicating that larger adult males live longer, but the males prevented from reproducing clearly have greater longevity than reproducing males of similar size. Software gives the following regression lines:

$$\text{For nonreproducing males: } \hat{y} = -52.381 + 143.504x_1$$

$$\text{For reproducing males: } \hat{y} = -59.816 + 123.370x_1$$

We notice from the estimated regression lines and Figure 28.1 that the intercepts for the two regression lines are clearly different, but the two regression lines are roughly

TABLE 28.1 Thorax length (mm) and longevity (days) of reproducing and nonreproducing male fruit flies

Longevity	ThxLength	IndReprod	Longevity	ThxLength	IndReprod
35	0.64	0	16	0.64	1
37	0.68	0	19	0.64	1
49	0.68	0	19	0.68	1
46	0.72	0	33	0.72	1
64	0.72	0	34	0.72	1
39	0.76	0	34	0.74	1
46	0.76	0	30	0.76	1
56	0.76	0	42	0.76	1
64	0.76	0	42	0.76	1
65	0.76	0	34	0.78	1
56	0.80	0	26	0.80	1
65	0.80	0	30	0.80	1
70	0.80	0	40	0.82	1
64	0.84	0	54	0.82	1
65	0.84	0	35	0.84	1
70	0.84	0	35	0.84	1
76	0.84	0	46	0.84	1
81	0.84	0	46	0.84	1
85	0.84	0	42	0.88	1
70	0.88	0	46	0.88	1
70	0.88	0	54	0.88	1
76	0.92	0	54	0.88	1
76	0.92	0	56	0.88	1
76	0.94	0	61	0.88	1
			44	0.92	1

parallel. Therefore, the difference in the two intercepts indicates how much longer the nonreproducing males live after we take into account thorax length. We will soon learn how to formally estimate parameters and make inferences for parallel regression lines.

CONCLUDE: Our preliminary analysis clearly shows that the nonreproducing male fruit flies live longer than reproducing males of similar size. We know that both groups of male fruit flies lived in similar conditions, except the nonreproducing males lived their lives in the company of females that were not fertile and could not reproduce, whereas the reproducing males lived with fertile females. So the physiological cost of reproducing was the sole difference between the two groups. We will learn later in this chapter how to formally test if the difference observed is statistically significant. ■

We now think that the relationship between longevity and thorax length has about the same rate for both experimental conditions, but that the physiological cost of reproduction results in shorter life spans. This difference in life span appears to be fairly constant for all thorax lengths. We would like to have a single regression model that captures this insight.

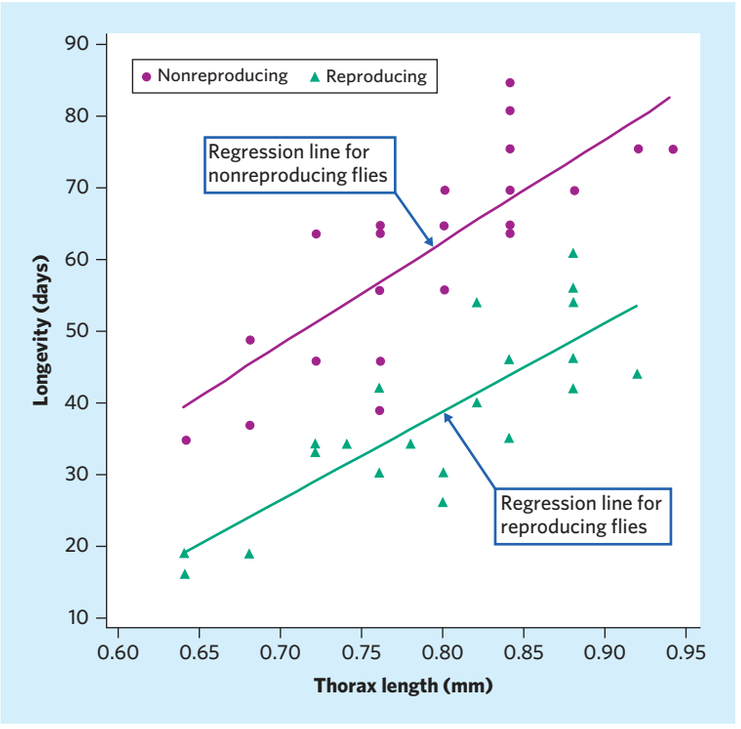


FIGURE 28.1 A scatterplot of longevity of male fruit flies of different adult sizes, with two separate regression lines, for Example 28.1.

To do this, introduce a second explanatory variable x_2 for "reproductive status." This is a categorical variable that takes the values "reproductive" and "nonreproductive." To incorporate this categorical variable x_2 into a regression model, we simply use values 0 and 1 to distinguish the two conditions. Now we have an *indicator variable:*

$$x_2 = 0 \text{ for nonreproductive males}$$
$$x_2 = 1 \text{ for reproductive males}$$

INDICATOR VARIABLE

An **indicator variable** places individuals into one of two categories, usually coded by the two values 0 and 1.

An indicator variable is like an indicator light on the dash of a car. If the fuel is above a certain level, the indicator light for fuel is off, but if the fuel drops below a certain level, the light switches on to indicate that there is a low amount of fuel in the tank. An indicator variable for fuel, say x_2, could be coded as $x_2 = 0$ if the fuel is above a certain level and $x_2 = 1$ if the fuel is below that level. Indicator variables are commonly used to indicate categorical characteristics such as gender (0 = male, 1 = female) or condition of patient (0 = good, 1 = poor). When a categorical variable can take more than two values, it is necessary to use more than one indicator variable to encode it. For instance, if we describe a patient's condition as

stationary, improving, or deteriorating, we would need two indicator variables, say x_2 and x_3, each taking only values 0 or 1; we could then use stationary = (1, 0), improving = (0, 1), and deteriorating = (0, 0) in the regression model.

The conditions for inference in simple linear regression (Chapter 23, page 563) describe the relationship between the explanatory variable x and the mean response μ_y in the population by a *population regression line* $\mu_y = \alpha + \beta x$. In this chapter we switch to the notation $\mu_y = \beta_0 + \beta_1 x$ because it is easier to extend to the multiple regression model. Now we add a second explanatory variable, so that our *regression model* for the population becomes

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The conditions for inference are the same as in the simple linear regression setting: For any fixed values of the explanatory variables, y varies about its mean according to a Normal distribution with unknown standard deviation σ, which is the same for all values of x_1 and x_2. We will look in detail at conditions for inference in multiple regression later on.

EXAMPLE 28.2 Interpreting a multiple regression model

Multiple regression models are no longer simple straight lines, so we must think a bit harder in order to interpret what they say. Consider our model

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

in which y is the longevity of male fruit flies in the lab, x_1 is their thorax length, and x_2 is an indicator variable for reproductive status. For nonreproducing males, $x_2 = 0$, so the model becomes

$$\mu_y = \beta_0 + \beta_1 x_1$$

For reproducing males, $x_2 = 1$, so the model is

$$\begin{aligned} \mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned}$$

Look carefully: The slope that describes how the mean longevity changes as thorax length x_1 varies is β_1 in both groups. The intercepts differ: β_0 for nonreproducing males and $(\beta_0 + \beta_2)$ for reproducing males. So β_2 is of particular interest, because it is the fixed change between nonreproducing and reproducing male fruit flies.

Figure 28.2 is a graph of this model with all three β's identified. We have succeeded in giving a single model for two parallel straight lines. ■

APPLY YOUR KNOWLEDGE

28.1 Bird colonies. Suppose that the number y of new birds that join a colony this year has the same straight-line relationship with the percent x_1 of returning birds in colonies for two different bird species. An indicator variable shows which species we observe; $x_2 = 0$ for one, and $x_2 = 1$ for the other. Write a population regression model that describes this setting. Explain in words what each β in your model means.

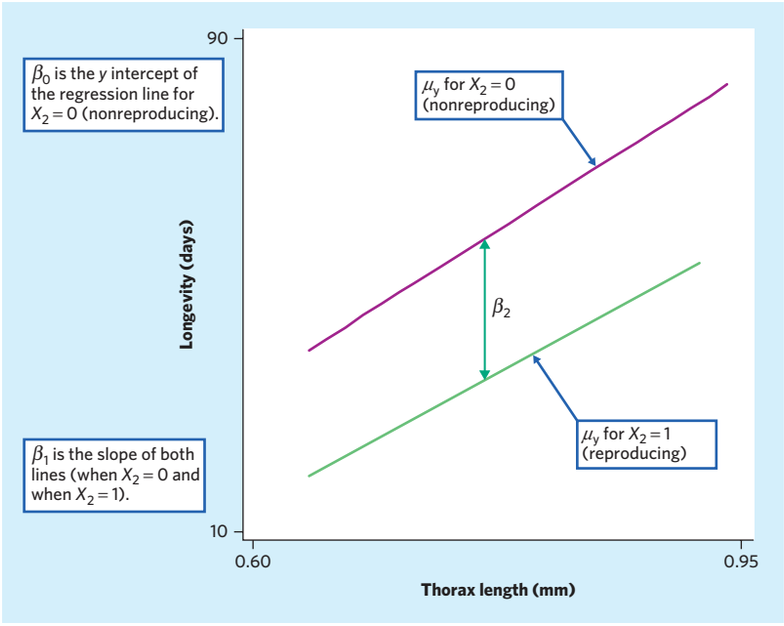


FIGURE 28.2 Multiple regression model with two parallel straight lines, for Example 28.2.

Estimating parameters

How shall we estimate the β's in the model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$? Because we hope to predict y, we want to make the errors in the y direction small. We can't call this the vertical distance from the points to *a line* as we did for a simple linear regression model, because we now have two lines. But we still concentrate on the prediction of y and therefore on the deviations between the observed responses y and the responses predicted by the regression model. These deviations are residual errors, similar to the residuals we first saw in simple linear regression:

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

The method of least squares estimates the β's in the model by choosing the values that minimize the sum of the squared residuals. Call b's the estimates of the β's. Then the b's minimize

$$\begin{aligned}\sum (\text{residual})^2 &= \sum (\text{observed } y - \text{predicted } y)^2 \\ &= \sum (y - \hat{y})^2 \\ &= \sum (y - b_0 - b_1 x_1 - b_2 x_2)^2\end{aligned}$$

The least-squares regression model $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ estimates the population regression model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Remember from Chapter 23 that the population model describes the mean response μ_y for given values of the explanatory variable(s).

The remaining parameter to be estimated is the standard deviation σ, which describes the variability of the response y about the mean μ_y given by the population regression model. Since the residuals estimate how much y varies about the

mean of the regression model, the standard deviation s of the residuals is used to estimate σ. The value of s is also referred to as the *regression standard error*.

REGRESSION STANDARD ERROR

The **regression standard error** for the multiple regression **model with two parallel lines** is

$$s = \sqrt{\frac{1}{n-3}\sum \text{residual}^2} = \sqrt{\frac{1}{n-3}\sum (y - \hat{y})^2}$$

where n is the number of observations in the data set. Use s to estimate the standard deviation σ of the responses about the mean given by the population regression model.

Notice that instead of dividing by $(n - 2)$, as we did for the simple linear regression model in Chapter 23, we are now dividing by $(n - 3)$. Since we are estimating three β parameters in our population regression model, the degrees of freedom must reflect this change. In general, the **degrees of freedom** for the regression standard error will be the number of data points minus the number of β parameters in the population regression model.

degrees of freedom

Why do we prefer one regression model with parallel lines to the two separate regression models? Simplicity is one reason—why use separate models with four β's if a single model with three β's describes the data well? Looking at the regression standard error provides another reason: The n in the formula for s includes all the observations in both groups. As usual, more observations produce a more precise estimate of σ. (Of course, using one model for both groups assumes that σ describes the scatter about the line in both groups.)

EXAMPLE 28.3 Cost of reproduction in male fruit flies

We developed a single model with parallel straight lines in Example 28.2 for the longevities in the two experimental groups shown in Figure 28.1. The mean longevity is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is the individual thorax length (the value on the x axis in Figure 28.1) and x_2 is an indicator variable to identify the experimental group (different symbols in Figure 28.1). The estimated regression model obtained from statistical software is

$$\hat{y} = -44.285 + 133.395x_1 - 23.551x_2$$

By substituting the two values of the indicator variable into our estimated regression equation, we can obtain a least-squares line for each group. The predicted longevities are

$$\hat{y} = -67.836 + 133.395x_1 \text{ for reproducing males } (x_2 = 1)$$

and

$$\hat{y} = -44.285 + 133.395x_1 \text{ for nonreproducing males } (x_2 = 0)$$

Comparing these estimated regression equations with the two separate regression lines obtained in Example 28.1, we see that the intercept parameters are very close to one

another (−67.836 is close to −59.816, and −44.285 is close to −52.381) for both groups. The big change, as intended, is that the slope 133.395 is now the same for both lines. In other words, the estimated change in mean longevity for a 1-unit change in thorax length is now the same for both models, 133.395. A closer look reveals that 133.395 is roughly the average of the two slope estimates (143.504 and 123.370) obtained in Example 28.1.

Finally, the regression standard error $s = 7.846$ indicates the size of the error between the observations and the estimated model. ■

APPLY YOUR KNOWLEDGE

28.2 LDL cholesterol. A study examined the change in blood LDL cholesterol level (in millimoles per liter, mmol/l) as a function of baseline LDL cholesterol level when subjects took either a placebo or the drug Pravastatin for two years.[2] One model of the results shows two parallel lines and gives a multiple linear regression model $\mu_y = -1.875 + 0.4375x_1 + 1.4x_2$, where x_1 represents baseline cholesterol level and x_2 is an indicator variable for treatment that takes value 0 for the placebo group and 1 for the Pravastatin group. Identify the parameter estimates in the model. Find the straight-line equations for the placebo group and for the Pravastatin group, and then make a graph that shows these two lines.

Using technology

Table 28.1 provides a compact way to display data in a textbook, but this is typically not the best way to enter your data into a statistical software package for analysis. The primary disadvantage to entering the data into a worksheet as it appears in Table 28.1 is that the response variable of interest, longevity, is entered in two separate columns, one for each experimental group.

As problems become more complex, we often collect information on more than two variables. This means that data management becomes a much larger part of our job. When fitting and analyzing models, it is usually best to have one column for each variable and to use the rows to identify the individuals.

EXAMPLE 28.4 Organizing data

To fit the multiple regression model with equal slopes in Example 28.3, three columns were created. The 49 longevities y for reproducing and nonreproducing males were stacked into a column labeled *Longevity*, values of the explanatory variable x_1 were entered into a column labeled *ThxLength*, and values of the indicator variable x_2 were entered into a column labeled *IndReprod*. Here are the first five rows of the worksheet:

Row	Longevity	ThxLength	IndReprod
1	35	0.64	0
2	37	0.68	0
3	49	0.68	0
4	46	0.72	0
5	64	0.72	0

■

Statistical software performs all the necessary computations after we identify the response variable *Longevity* and the two explanatory variables *ThxLength* and *IndReprod*. Figure 28.3 shows the regression output from Excel, Minitab, R, and SPSS. The format of the output differs slightly, but each package provides parameter estimates, standard errors, t statistics, and P-values, an analysis of variance (ANOVA) table, the regression standard error, and R^2. We will digest this output one piece at a time: first describing the model, then looking at the conditions needed for inference, and finally interpreting the results of inference.

Microsoft Excel

Book1

	A	B	C	D	E	F	G
1	*Regression statistics*						
2	Multiple R	0.900006132					
3	R Square	0.810011038					
4	Adjusted R Square	0.801750649					
5	Standard Error	7.845843899					
6	Observations	49					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	2	12072.56982	6036.284912	98.0596647	2.57485E-17	
11	Residual	46	2831.634258	61.55726649			
12	Total	48	14904.20408				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	-44.28543	11.71411366	-3.78051907	0.000449471	-67.8646953	-20.7061648
16	ThxLength	133.3949179	14.49005467	9.205963738	5.2827E-12	104.2279785	162.5618572
17	IndReprod	-23.5505042	2.242169166	-10.5034467	8.33921E-14	-28.0637524	-19.0372561

Minitab

Session

```
Regression Analysis: Longevity versus ThxLength, IndReprod

The regression equation is
Longevity = - 44.3 + 133 ThxLength - 23.6 Ind_Reprod

Predictor       Coef  SE Coef       T      P
Constant     -44.29    11.71   -3.78  0.000
ThxLength    133.39    14.49    9.21  0.000
IndReprod   -23.551    2.242  -10.50  0.000

S = 7.84584   R-Sq = 81.0%   R-Sq(adj) = 80.2%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        2  12072.6  6036.3  98.06  0.000
Residual Error   46   2831.6    61.6
Total            48  14904.2
```

FIGURE 28.3 Output from Excel, Minitab, SPSS, and R for the model with parallel regression lines in Example 28.3. (*Continued on next page.*)

SPSS

Session

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.900[a]	.810	.802	7.846

a. Predictors: (Constant), IndReprod, ThxLength

b. Dependent Variable: Longevity

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12072.570	2	6036.285	98.060	.000[a]
	Residual	2831.634	46	61.557		
	Total	14904.204	48			

a. Predictors: (Constant), IndReprod, ThxLength

b. Dependent Variable: Longevity

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-44.285	11.714		-3.781	.000	-67.865	-20.706
	ThxLength	133.395	14.490	.592	9.206	.000	104.228	162.562
	IndReprod	-23.551	2.242	-.675	-10.503	.000	-28.064	-19.037

a. Dependent Variable: Longevity

R

```
Im(formula = Longevity ~ ThxLength + IndReprod)

Residuals:
     Min       1Q    Median      3Q     Max
 -18.095   -5.106    -1.537   5.792  17.234

Coefficients:
            Estimate  Std. Error t value   Pr(>|t|)
(Intercept)  -44.285      11.714  -3.781  0.000449 ***
ThxLength    133.395      14.490   9.206  5.28e-12 ***
IndReprod    -23.551       2.242 -10.503  8.34e-14 ***
---
Signif. codes: 0 '***' 0.001 '***'0.01 '***'0.05  '.' 0.1 ' ' 1

Residual standard error: 7.846 on 46 degrees of freedom
Multiple R-squared: 0.81, Adjusted R-squared: 0.8018
F-statistic: 98.06 on 2 and 46 DF, p-value: <2.2e-16
```

FIGURE 28.3 (*Continued*)

EXAMPLE 28.5 Parameter estimates on statistical output

On the Excel output in Figure 28.3, the parameter estimates $b_0 = -44.28543$, $b_1 = 133.3949179$, and $b_2 = -23.5505042$ are clearly labeled in a column. Thus, our multiple regression model for predicting longevity (after rounding) is $\hat{y} = -44.285 + 133.395x_1 - 23.551x_2$. SPSS and R have a similar output format. Minitab provides the estimated regression equation first and then gives more detailed estimates in a column labeled

"Coef" for coefficients. After rounding, all four sets of estimates match the estimates provided in Example 28.3.

Although the labels differ again, the regression standard error is provided by all four packages:

Excel:	Standard Error = 7.845843899
Minitab:	$S = 7.84584$
R:	Residual standard error = 7.846
SPSS:	Std. Error of the Estimate = 7.846 ■

For simple linear regression models, the square of the correlation coefficient r^2 between y and x provides a statistic that can be used along with residual plots and other techniques to assess the fit of the model. In particular, r^2 measures the proportion of variation in the response variable that is explained by using the explanatory variable. For our multiple regression model with parallel regression lines, we do not have one correlation coefficient. However, by squaring the correlation coefficient between the observed responses y and the predicted responses $\hat{y}$, we obtain the *squared multiple correlation coefficient* R^2.

Alternative computation formulas based on values in the ANOVA table of the output help us interpret this new statistic. You may want to refer to Chapter 24 to refresh your memory about the details of the ANOVA procedure. The ANOVA table breaks the total variability in the responses into two pieces. One piece summarizes the variation explained by the model, and the other piece summarizes the residual variation traditionally labeled "error." In short, we have

$$\text{total variation} = \text{variation explained by model} + \text{residual variation}$$

The value of R^2 is the ratio of *model* to *total* variation, so R^2 tells us how much variation in the response variable y we explained by using the set of explanatory variables in the multiple regression model.

SQUARED MULTIPLE CORRELATION COEFFICIENT

The **squared multiple correlation coefficient** R^2 is the square of the correlation coefficient between the observed responses y and the predicted responses $\hat{y}$ and can be computed as

$$R^2 = \frac{\text{variability explained by model}}{\text{total variability in } y} = \frac{\sum(\hat{y} - \overline{y})^2}{\sum(y - \overline{y})^2} = \frac{\text{SSModel}}{\text{SSTotal}}$$

The denominator measures the deviation of the observed responses about their mean. Just as in simple linear regression, the predicted responses $\hat{y}$ have the same mean $\overline{y}$ as the observed responses. So the numerator is the variability we would see if the model fitted perfectly and there were no spread of the y's about the model. We can think of this as the variability explained by the model.

EXAMPLE 28.6 Using R^2

The value of $R^2 = 0.810$ for our multiple regression model with parallel lines in Example 28.3 can be found in the output for all four packages in Figure 28.3 (labeled as Multiple R-squared, R Square, or R-Sq). Thus, the proportion of variation in the response variable *Longevity* that is explained by the explanatory variable *ThxLength* and the indicator variable *IndReprod* using the regression model with parallel lines is 0.810, or 81.0%. Converting R^2 to the percent of variation explained by the explanatory variables in a multiple regression model is a common practice. The value of R^2 indicates that 81% of the variation in the longevity of male fruit flies in the lab is explained by using multiple regression with an explanatory variable for their thorax length and an indicator variable for the experimental condition determining their reproductive status. ■

The squared multiple correlation coefficient takes only values between 0 and 1 (0% and 100%) and is a very useful statistic to help us assess the fit of a multiple regression model.

APPLY YOUR KNOWLEDGE

28.3 Heights and weights for boys and girls. Suppose you are designing a study to investigate the relationship between height and weight for boys and girls.

(a) Specify a model with two regression lines that could be used to predict height separately for boys and for girls. Be sure to identify all variables and describe all parameters in your model.

(b) How many columns in a worksheet would be required to fit this model with statistical software? Describe each column.

28.4 Nestling mass and nest humidity. Researchers investigated the relationship between nestling mass, measured in grams, and nest humidity index, measured as the

Minitab

Session

```
Regression Analysis: Mass versus Humidity, Exposed

The regression equation is
Mass = 18.1 - 5.41 Humidity + 0.848 Exposed

Predictor      Coef  SE Coef       T       P
Constant    18.0848   0.6592   27.43   0.000
Humidity     -5.411    1.377   -3.93   0.000
Exposed      0.8484   0.3587    2.37   0.024

S = 1.01583   R-Sq = 47.7%   R-Sq(adj) = 44.6%

Analysis of Variance

Source           DF       SS      MS       T       P
Regression        2   32.008  16.004   15.51   0.000
Residual Error   34   35.085   1.032
Total            36   67.092
```

ratio of total mass of water in the nest divided by nest dry mass, for two different groups of great titmice parents.[3] One group was exposed to fleas during egg laying, and the other was not. Exposed parents were coded as 1, and unexposed parents were coded as 0. Use the output on the previous page, obtained by fitting a multiple regression model with parallel lines for the two groups of parents, to answer the following questions.

(a) Identify the regression model for predicting nestling mass from nest humidity index for the two groups of great titmice parents.

(b) Based on your model, do you think that nestling mass was higher in nests of birds exposed to fleas during egg laying? Explain.

(c) What is the value of the regression standard error? Interpret this value.

(d) What is the value of the squared multiple correlation coefficient? Interpret this value.

Conditions for inference

We have seen in a simple but useful case how adding another explanatory variable can fit patterns more complex than the single straight line of simple linear regression. So far we have included two explanatory variables: a quantitative variable x_1 and an indicator variable x_2. Multiple linear regression, however, can allow any number of explanatory variables, each of which can be either quantitative or indicator. You can also work with data from just one random sample, or you can pull together separate random samples represented by an indicator variable (as in the case of the fruit fly data in Example 28.1). Here is a statement of the general model that includes the conditions needed for inference.

THE MULTIPLE LINEAR REGRESSION MODEL

We have n observations on k explanatory variables $x_1, x_2, \ldots, x_k$ and a response variable y. Our goal is to study or predict the behavior of y for a given set of the explanatory variables.

- For any set of fixed values of the explanatory variables, the response y varies according to a **Normal distribution.** Repeated responses y are **independent** of each other.
- The mean response μ_y has a **linear relationship** given by the **population regression model**

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The β's are unknown parameters.

- The **standard deviation** of y (call it σ) is the same for all values of the explanatory variables. The value of σ is unknown.

This model has $k + 2$ parameters that we must estimate from data: the $k + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and the standard deviation σ.

These conditions are similar to those we have described in Chapter 23 for simple linear regression (page 563). However, the challenge in checking inference conditions for multiple regression is that there is no one simple plot that can give the entire picture. We will present a few simple basics here because regression diagnostics is a subject that could be expanded to several chapters.

First, plot the response variable against each of the explanatory variables. These plots help you explore and understand potential relationships. Multiple regression models allow curvature and other interesting features that are not simple to visually check, especially when we get beyond two explanatory variables. For example, if you see a quadratic pattern, you should consider adding a quadratic term (x^2) for that explanatory variable.

Then, plot the residuals against the predicted values and all of the explanatory variables in the model. These plots will allow you to check the condition that **the standard deviation of the response about the multiple regression model is the same everywhere.** They should show an unstructured horizontal band of points centered at 0. The mean of the residuals is always 0, just as in simple linear regression. Funnel or cone shapes indicate that this condition is not met and that one or more variables should be transformed to stabilize the standard deviation of the residuals before doing inference.

Look for outliers and influential observations in all residual plots. To check how much influence a particular observation has, you can fit your model with and without this observation. If the estimates and statistics do not change much, you can safely proceed. However, if there are substantial changes, you must begin a more careful investigation. Do not simply throw out observations to improve the fit and increase R^2. (See Chapter 2, page 53, for an in-depth discussion of outlier treatment.)

Ideally, we would like all the explanatory variables to be independent and the observations on the response variable to be independent. As you will see in this chapter, practical problems include explanatory variables that are not independent. Association between two or more explanatory variables can create serious problems in the model, so use correlations and scatterplots to check relationships.

The response should vary about the multiple regression model according to a Normal distribution. This condition is checked by making a histogram, stemplot, or Normal quantile plot of the residuals. Once again, we will rely on the robustness of the regression methods when there is a slight departure from Normality, except for prediction intervals. As we did with simple linear regression, we view prediction intervals from multiple regression models as rough approximations.

EXAMPLE 28.7 Checking the conditions

Four plots displaying the residuals for the multiple regression model with parallel lines in Example 28.3 are shown in Figure 28.4. The conditions for inference are linearity, Normality, constant variance, and independence. We will check these conditions one at a time.

Linear trend: The scatterplot in Figure 28.1 shows parallel linear patterns for the two experimental groups, so the model is reasonable.

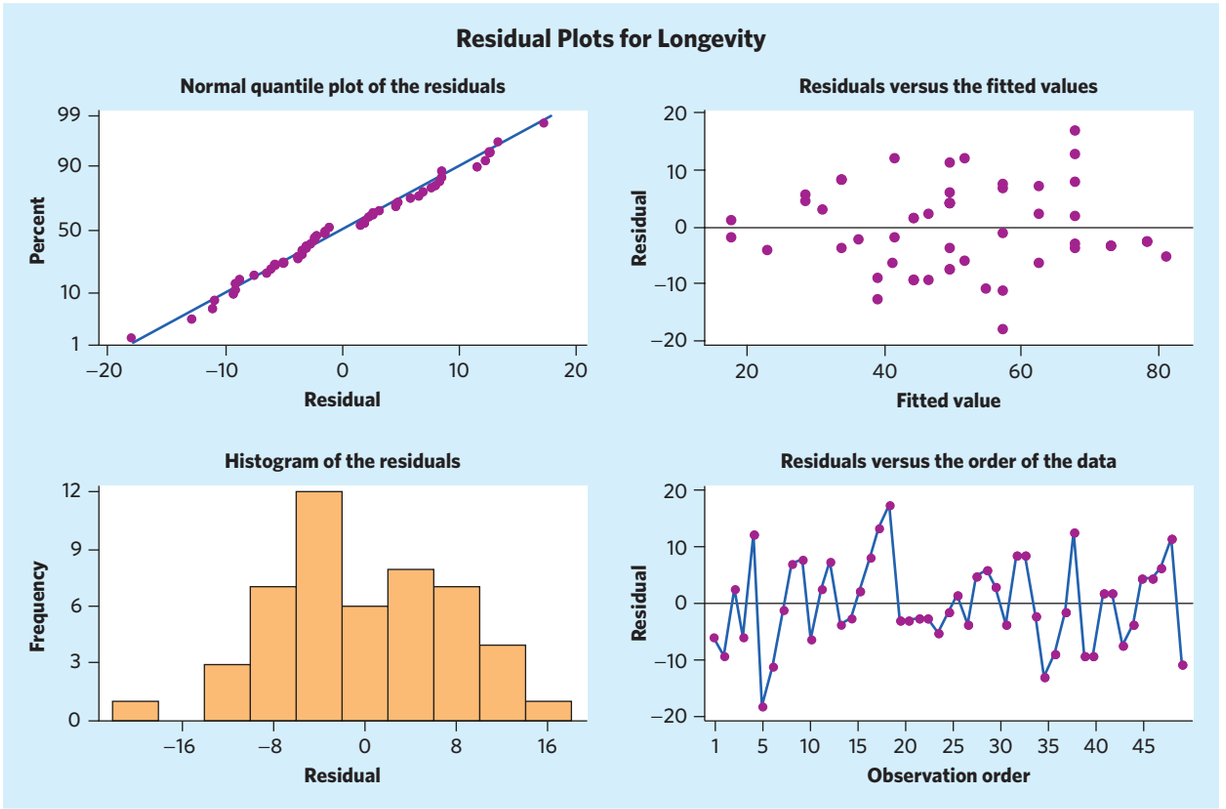


FIGURE 28.4 Residual plots to check the conditions for inference for the model with parallel regression lines in Example 28.3.

Normality: The normal quantile plot (upper left) and histogram (lower left) of the model's residuals in Figure 28.4 indicate that the residuals are symmetric about zero and approximately Normal.

Constant variance: The residual plot (upper right, residuals versus $\hat{y}$ values) in Figure 28.4 does not have any obvious pattern and shows a fairly constant amount of variability. Note, though, that since there are fewer points at the edges of the plot, it is harder to assess variability for the fruit flies with the shortest and the longest life spans. Overall, this residual plot supports the model's condition that a single σ describes the scatter about the reproducing and nonreproducing lines.

Independence: The male fruit flies were randomly assigned to the two experimental conditions. The individuals should therefore be independent. In addition, the plot of ordered residuals (lower right, ordered as they appear in the data set) provides a quick check to see if there is a pattern in the residuals based on the order in which they were entered into the worksheet. The lack of any obvious pattern in this last graph reinforces our belief that the individuals used for this model are independent.

We conclude that a linear model provides a reasonable summary of the male fruit fly longevities and that the conditions for inference appear to be satisfied. We can now proceed to inference for this model. ■

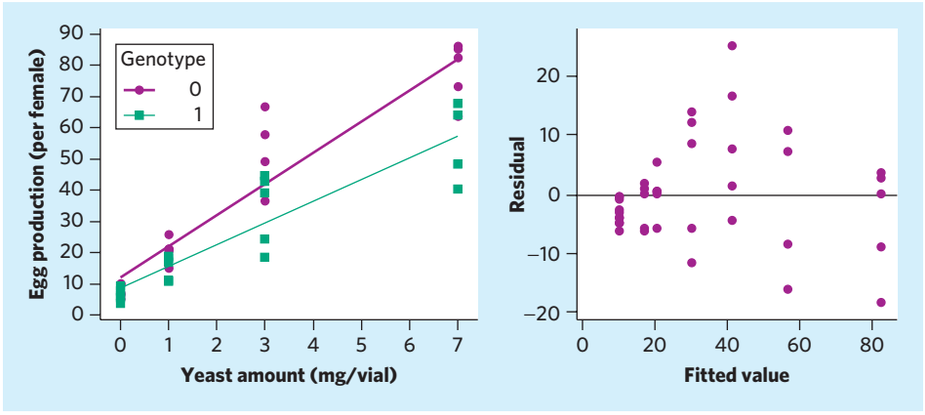
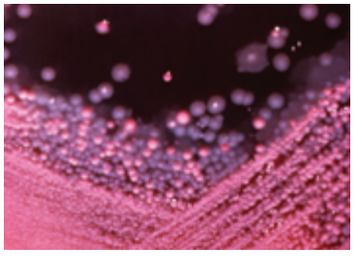


FIGURE 28.5 Scatterplot and residual plot to check the conditions for inference in Exercise 28.5.

APPLY YOUR KNOWLEDGE

28.5 Nutrition and reproduction. A research team looked at the relationship between protein intake (in the form of yeast supplements) and reproductive output (mean number of eggs per female) in wild-type versus mutant fruit flies with higher longevity.[4] Figure 28.5 shows the scatterplot and residual plot of this relationship. Do the plots suggest any potential problems with the conditions for inference? Explain your answer.



CDC/Dr. Gilda Jones

28.6 Evolution in bacteria. Biologists designed an experiment to test the theory of evolution in *E. coli* bacteria. These bacteria can be grown for 2000 generations ("evolved") in less than a year. Bacteria were evolved in either a neutral or an acidic pH (x_1) and then compared with the ancestral line to compute a relative fitness score (y) in environments of various pH (x_2).[5] A multiple regression model with three variables was found significant with $R^2 = 74.1\%$ and equation $\hat{y} = 2.12 - 0.957x_1 - 0.166x_2 + 0.147x_1x_2$. The residual plot and a histogram of the residuals are shown in Figure 28.6. Do you see any reason for concern in making inferences from this model, which has a high R^2-value? Comment on both plots.

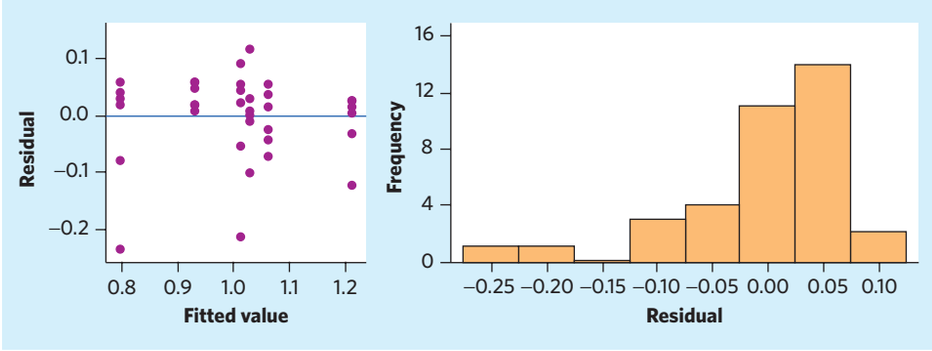


FIGURE 28.6 Residual plot and histogram of the residuals to check the conditions for inference in Exercise 28.6.

Inference for multiple regression

To this point, we have concentrated on understanding the model, estimating parameters, and verifying the conditions for inference that are part of a regression model. Inference in multiple regression begins with tests that help us to decide if a model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ adequately fits the data and to choose between several possible models.

The first inference for a multiple regression model examines the overall model. The ANOVA table summarizes the breakdown of the variability in the response variable when k explanatory variables are used. There is one row for each of the three sources of variation: model, error, and total. Each source of variation has a degrees of freedom associated with it. The F statistic for the overall model is the ratio of the model variation (MSM) over the error variation (MSE). Refer back to Chapter 24 on ANOVA if you need to refresh your memory.

F STATISTIC FOR REGRESSION MODEL

The **ANOVA *F* statistic** for testing the null hypotheses that all the regression coefficients (β's) except β_0 are equal to zero has the form

$$F = \frac{\text{variation due to model}}{\text{variation due to error}} = \frac{\text{MSM}}{\text{MSE}}$$

This F statistic is used to compute the P-value and determine whether all the k regression coefficients (one for each of the k explanatory variables, but not the intercept) are significantly different from zero. The P-value is obtained by finding the area to the right of F under the $F(k, n-k-1)$ distribution. Software typically provides the ANOVA table, which contains all the information you need.

EXAMPLE 28.8 Overall *F* test for parallel lines

The mean longevity is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is labeled as *ThxLength* and x_2 is labeled as *IndReprod* on the output in Figure 28.3. The null and alternative hypotheses for the overall F test are

$$H_0: \beta_1 = \beta_2 = 0 \text{ (that is, } \mu_y = \beta_0)$$

$$H_a: \text{at least one of } \beta_1 \text{ and } \beta_2 \text{ is not } 0$$

null model

The null hypothesis H_0 specifies a model, called the **null model,** where the response variable y is a constant (its mean) plus random variation. In other words, the null model says that x_1 and x_2 together do not help predict y.

The value of the F statistic reported in all four ANOVA tables in Figure 28.3 is $F = 98.06$ (up to rounding error). The P-value is obtained from an F distribution with 2 degrees of freedom for the numerator and 46 degrees of freedom for the denominator. SPSS and Minitab report a P-value of 0.000, which does not mean that its value is zero but that it is less than 0.0005. R shows that $P < 2.2e\text{-}16$, and Excel reports that the P-value is actually 2.57485E-17, which is extremely small. All four software packages indicate that the P-value is extremely significant. Since the P-value is less than any reasonable significance level, say, $\alpha = 0.01$, we reject the null hypothesis and conclude that at least one of the x's helps explain the variation in the longevities y. ■

Rejecting the null hypothesis with the F statistic tells us that at least one of our β parameters (excluding β_0) is not equal to zero, but it doesn't tell us which parameters are not equal to zero. We turn to individual tests for each parameter to answer that question.

INDIVIDUAL *t* TESTS FOR COEFFICIENTS

To test the null hypothesis that one of the β's in a specific regression model is zero, compute the t statistic

$$t = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{b}{\text{SE}_b}$$

When the conditions for inference are met, the t distribution with $(n - k - 1)$ degrees of freedom can be used to compute confidence intervals and conduct hypothesis tests for each of the β's.

EXAMPLE 28.9 Individual *t* tests

The output in Figure 28.3 provides parameter estimates and standard errors for the coefficients β_0, β_1, and β_2. The individual t statistic for x_1 (*ThxLength*) tests the hypotheses

H_0: $\beta_1 = 0$ in this model (that is, $\mu_y = \beta_0 + \beta_2 x_2$)

H_a: $\beta_1 \neq 0$ in this model

We must state the model explicitly in the null hypothesis because the bare statement H_0: $\beta_1 = 0$ can be misleading. The hypothesis of interest is that *in this model* the coefficient of x_1 is 0. If the same x_1 is used in a different model with different explanatory variables, the hypothesis H_0: $\beta_1 = 0$ has a different meaning even though we would write it the same way. More on this later.

Using the Minitab output, we see that the test statistic is

$$t = \frac{133.39}{14.49} = 9.21 \text{ for } \beta_1$$

The P-value is obtained by finding the area under a t distribution with $49 - 2 - 1 = 46$ degrees of freedom below –9.21 or above 9.21. Since this value is so small, Minitab simply reports the P-value as being 0.000 (< 0.0005).

The test statistic for the other variable coefficient is

$$t = \frac{-23.551}{2.242} = -10.50 \text{ for } \beta_2$$

The P-value is again obtained using the t distribution with 46 degrees of freedom. It is so small that it is reported by Minitab as being 0.000 (< 0.0005).

We have very strong evidence that thorax length x_1 (*ThxLength*) helps explain the longevity y even after we allow the reproductive status x_2 (*IndReprod*) to explain longevity. Similarly, reproductive status x_2 adds to our ability to explain longevity even after we take thorax length x_1 into account. ■

Statistical software reports two-sided P-values for the individual t tests. If you had an alternative hypothesis that would require a one-sided P-value, you can

simply divide the P-value on the output by 2 if the effect is in the direction of the alternative.

Example 28.9 illustrates one of the easiest situations you will encounter. The overall F test tells us that at least one of the coefficients (except for β_0) is not equal to zero; that is, both explanatory variables together help explain the response. Then the individual t tests indicate that both coefficients are significantly different from zero. This means that each explanatory variable significantly improves the explanation when added to a model that uses only the other explanatory variable.

Individual t tests are helpful in identifying the explanatory variables that are useful predictors, but *extreme caution* is necessary when interpreting the results of these tests. Remember that an individual t assesses the contribution of its variable in the presence of the other variables in this specific model. That is, individual t's depend on the model in use, not just on the direct association between an explanatory variable and the response. The case study described later in this chapter illustrates some of these challenges.

CONFIDENCE INTERVALS FOR COEFFICIENTS

A **level C confidence interval for $\boldsymbol{\beta}$** is $b \pm t^*\text{SE}_b$.

The critical value t^* is obtained from the $t(n-k-1)$ density curve, where k is the number of explanatory variables used in the multiple regression model.

EXAMPLE 28.10 Parameter confidence intervals when all predictors are helpful

The individual t statistics and the corresponding P-values in Figure 28.3 indicate that both explanatory variables are useful predictors. Only Excel and SPSS provide a confidence interval for the regression coefficients. With 95% confidence, both programs determine that β_1 is between 104.228 and 162.562 days per millimeter (mm) of thorax length and that β_2 is within the interval -28.064 to -19.037. Interpretation is relatively simple here because the model has only two parallel lines. We are 95% confident that, in this particular model, the longevity of male fruit flies in each group increases by about 104 to 163 days per mm of thorax length (or, more practically, by 10 to 16 days per tenth of mm) and that reproducing males live about 19 to 28 days less than nonreproducing males of similar size.

If the software you use does not provide confidence intervals for the regression coefficients, you can compute them yourself from the parameter estimates provided. The regression model for the cost of reproduction in male fruit flies has $n-k-1 = 49-2-1 = 46$ degrees of freedom. Using Table C in the back of the textbook, we find that the t critical value for degrees of freedom 40 (since 46 is not in the table) and confidence level 95% is $t^* = 2.021$. Thus, an approximate 95% confidence interval for β_1 is

$$b_1 \pm t^*\text{SE}_{b_1} = 133.395 \pm (2.021)(14.490) = 133.395 \pm 29.284, \text{ or } 104.111 \text{ to } 162.679$$

and an approximate 95% confidence interval for β_2 is

$$b_2 \pm t^*\text{SE}_{b_2} = -23.551 \pm (2.021)(2.242) = -23.551 \pm 4.531, \text{ or } -28.082 \text{ to } -19.020$$

These approximate intervals are very close to those provided by software. ■

Interpreting coefficients, beyond significance, can get very challenging in more complex models. We will see an example later in the chapter.

Beyond model building, prediction can also be an important objective of scientific studies. Construction of *confidence intervals for a mean response* and *prediction intervals for a future observation* with multiple regression models is similar to the methods we used for simple linear regression in Chapter 23. The main difference is that we must now specify a list of values for all the explanatory variables in the model. As we learned in simple linear regression, the additional uncertainty in predicting future observations will result in prediction intervals that are wider than confidence intervals.

CONFIDENCE AND PREDICTION INTERVALS FOR MULTIPLE REGRESSION RESPONSE

A level C **confidence interval for the mean response** μ_y is $\hat{y} \pm t^* \text{SE}_{\hat{\mu}}$.

A level C **prediction interval for a single response** y is $\hat{y} \pm t^* \text{SE}_{\hat{y}}$.

The critical value t^* is obtained from the $t(n-k-1)$ density curve, where k is the number of explanatory variables used in the multiple regression model.

EXAMPLE 28.11 Predicting means and future values

Figure 28.7 provides the predicted values ("Fit"), 95% confidence limits for the mean longevity ("95% CI"), and 95% prediction limits ("95% PI") for two observations. As you can see, we had to specify a numerical value for both of the model's explanatory variables in order to get these intervals. Here we have obtained confidence and prediction intervals for male fruit flies with thorax lengths of 0.8 mm (whether allowed to reproduce or not). As expected, the prediction limits for future fruit fly longevities are wider than the corresponding confidence limits for the mean fruit fly longevities. ■

Minitab

Session

```
Predicted Values for New Observations

New
Obs    Fit     SE Fit        95% CI             95% PI
  1   62.43      1.60   (59.21, 65.65)   (46.31, 78.55)
  2   38.88      1.57   (35.72, 42.04)   (22.77, 54.99)

Values of Predictors for New Observations

New
Obs   ThxLengt   IndReprod
  1      0.800    0.000000
  2      0.800        1.00
```

FIGURE 28.7 Predicted values, confidence limits for mean longevity, and prediction limits for future longevity for two fruit fly observations, for Example 28.11.

APPLY YOUR KNOWLEDGE

28.7 Bone density in elderly men. Bone fractures are a substantial concern among the elderly. A study examined the impact of a number of variables on bone mineral density (BMD) assessed from the femur in a random sample of 43 elderly men hospitalized for hip fracture. The researchers found that BMD could be predicted from serum albumin level and body weight according to the following multiple regression model:[6]

	b	SE	*t*	*P*
Constant	−8.14	1.00	−8.12	<0.001
Albumin	0.098	0.026	3.71	<0.01
Weight	0.042	0.01	3.69	<0.01

$F = 15.19$, $(P < 0.001)$, $R^2 = 0.451$

Use the information provided to answer the following questions.

(a) What is the multiple linear regression equation for this model? What is the number k of explanatory variables involved?

(b) What tells you that the overall model is significant?

(c) What percent of variation in BMD is explained by the model?

(d) What is a 95% confidence interval for the slope parameter for albumin level?

(e) What is a 95% confidence interval for the slope parameter for body weight?

28.8 Bone density in elderly women. Men and women age differently. The study described in the previous exercise also examined a random sample of 243 elderly women hospitalized for hip fracture. The researchers found that BMD in elderly women with hip fracture could be modeled from body weight, fracture type (cervical or trochanteric), age, and total lymphocyte count (TLC) according to the following multiple regression model:

	b	SE	*t*	*P*
Constant	−5.03	0.644	−7.82	<0.001
Weight	0.055	0.005	10.01	<0.001
Fracture	−0.51	0.113	−4.56	<0.001
Age	0.019	0.006	−3.00	<0.01
TLC	3.1E-04	0.000	2.39	<0.05

$F = 42.2$, $(P < 0.001)$, $R^2 = 0.422$

Use the information provided to answer the following questions.

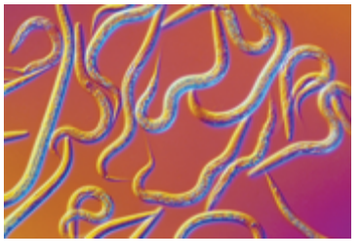
(a) What is the multiple linear regression equation for this model? What is the number k of explanatory variables involved?

(b) Which, if any, of the explanatory variables in the model is an indicator variable?

(c) What tells you that the overall model is significant?

(d) What percent of variation in BMD in elderly women is explained by the model?

(e) What is a 95% confidence interval for the slope parameter for body weight?



Sinclair Stammers/Science Source

28.9 Neural mechanism of nociception. The roundworm *Caenorhabditis elegans* is a widely studied animal model, in part because of its small number of neurons and easily manipulated genome. Nociception is the neural perception of an actually or potentially harmful stimulus. In *C. elegans*, it evokes a self-preserving withdrawal behavior. However, repeated stimulation can result in reduced withdrawal response, or habituation. Researchers compared the withdrawal response to disturbing light stimuli in wild-type *C. elegans* and a mutant *C. elegans* line that exhibits a slower response of PVD sensory neurons. Failure to react indicates habituation. Here are the percents of animals tested that exhibited a withdrawal reaction to a noxious stimulus consisting of varying numbers of consecutive light pulses (IndMutant = 1 for the mutant line, and 0 for the wild type):[7]

% Reacting	Pulses	IndMutant	% Reacting	Pulses	IndMutant	% Reacting	Pulses	IndMutant
93	1	1	54	15	1	91	9	0
81	2	1	56	16	1	82	10	0
73	3	1	50	17	1	86	11	0
64	4	1	42	18	1	82	12	0
73	5	1	42	19	1	91	13	0
68	6	1	29	20	1	85	14	0
64	7	1	96	1	0	83	15	0
65	8	1	100	2	0	70	16	0
50	9	1	92	3	0	74	17	0
68	10	1	96	4	0	64	18	0
57	11	1	92	5	0	68	19	0
50	12	1	92	6	0	78	20	0
62	13	1	84	7	0			
44	14	1	96	8	0			

(a) Make a scatterplot of the percent reacting as a function of the number of light pulses with two separate lines representing the two subpopulations. Describe the shape of the relationship. What does it suggest about the pattern of withdrawal responses in the two subpopulations of *C. elegans* for increasing numbers of light pulses? How does your answer fit in the context of habituation? Explain why these results suggest an involvement of PVD sensory neurons in nociception and habituation.

(b) Use software to obtain the statistical analysis of the linear regression of the percent reacting (y) as a function of the number of light pulses (x_1) and the indicator variable IndMutant (x_2). What is the regression equation for this model?

(c) Is the overall model significant? What percent of variation in percent reacting is explained by the model?

(d) Are the regression coefficients for the two explanatory variables significant?

28.10 Neural mechanism of nociception: prediction. Minitab gives the following information for mutants exposed to 10 light pulses ($x_1 = 10$, $x_2 = 1$):

Predicted Values for New Observations

NewObs	Fit	SE Fit	95% CI	95% PI
1	60.18	1.47	(52.20, 63.17)	(46.52, 73.84)

With 95% confidence, what is the average percent of mutants reacting to 10 light pulses in similar experimental conditions? What range of percents reacting would we predict for 95% of experiments in which mutant *C. elegans* are exposed to 10 light pulses?

Interaction

We have first examined simple examples with two parallel linear patterns for two values of an indicator variable. We now turn to situations with two linear patterns that are not parallel. To write a regression model for this setting, we need to consider the **interaction** between two explanatory variables. We first described the concept of interaction between two factors in Chapter 26 for two-way ANOVA. In multiple linear regression, interaction between variables x_1 and x_2 appears as a product term x_1x_2, making a new variable x_3 in the model. The product term means that *the relationship between the mean response and one explanatory variable x_1 changes when we change the value of the other explanatory variable x_2*. Here is an example.

interaction

EXAMPLE 28.12 Lung capacity in boys and girls

STATE: Children's respiratory health has become a major concern due to the effects of airborne pollution and secondhand smoke. To understand the impact of these negative factors, however, it is important first to understand the respiratory system of healthy children. A common way to assess respiratory function is by measuring the forced expiratory volume (FEV), which represents the amount of air that can be actively exhaled in one second. A study recorded the FEV of a random sample of 589 boys and girls aged 3 to 19 years to determine how FEV is related to age and gender in children.[8]

PLAN: Fit and evaluate a model with two regression lines for predicting FEV. ■

Since boys and girls have different growth patterns, it is likely that the pattern of FEV over time will differ somewhat for boys and girls. The scatterplot in Figure 28.8 confirms this suspicion. The pattern is linear for both boys and girls, but the two lines are not parallel. Let's see how adding an interaction term can help us model the data. Consider the following model:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

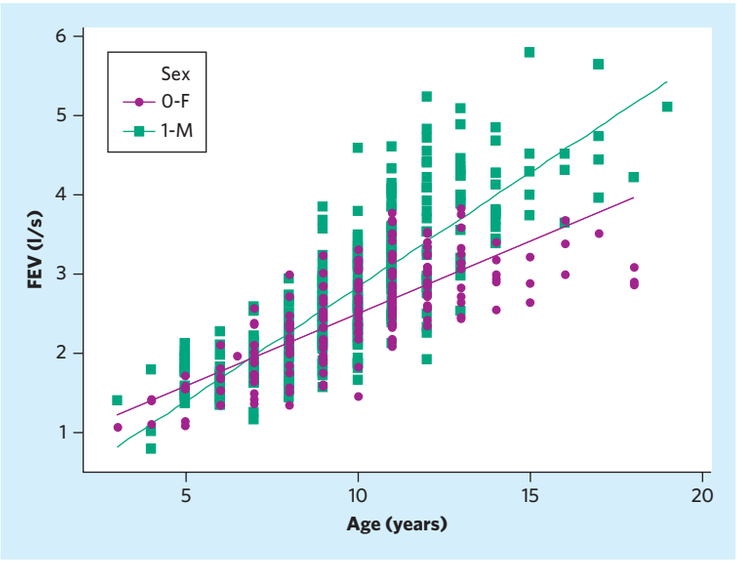


FIGURE 28.8 Scatterplot of FEV versus age for male and female children, for Example 28.12.

Here y is the FEV value, x_1 is the child's age, x_2 is an indicator variable that is 1 if the child is a boy and 0 otherwise, and x_1x_2 is the interaction term. For girls, $x_2 = 0$, so the model is

$$\mu_y = \beta_0 + \beta_1 x_1$$

For boys, $x_2 = 1$, so the model is

$$\begin{aligned}\mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1\end{aligned}$$

A careful look allows us to interpret all four parameters: β_0 and β_1 are the intercept and slope, respectively, for girls; β_2 and β_3 indicate the fixed change in the intercept and slope, respectively, for boys. Be careful not to interpret β_2 as the intercept and β_3 as the slope for boys. The indicator variable allows us to change the intercept as we did before, and the new interaction term allows us to change the slope. Thus, if β_3 is zero, the two lines are parallel. If β_3 is nonzero, then the lines are not parallel. The test for H_0: $\beta_3 = 0$ is sometimes called the "test of parallelism."

INTERACTION IN A MODEL WITH TWO REGRESSION LINES

We have n observations on an explanatory variable x_1, an indicator variable x_2 coded as 0 for some individuals and as 1 for other individuals, and a response variable y. The mean response μ_y is a linear function of the four parameters β_0, β_1, β_2, and β_3:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

EXAMPLE 28.13 Lung capacity in boys and girls

SOLVE: Figure 28.8 shows the two regression lines, one for girls and one for boys. The fitted model, as shown by the two regression lines in this case, appears to provide a good visual summary for the two groups.

Figure 28.9 provides the regression output from Minitab. By substituting 0 and 1 for the indicator variable x_2, we can easily obtain the two estimated regression lines. The estimated regression lines are

$$\hat{y} = 0.674 + 0.182x_1 \text{ for girls}$$

and

$$\hat{y} = (0.674 - 0.731) + (0.182 + 0.106)x_1 = -0.057 + 0.288x_1 \text{ for boys}$$

The overall F statistic 389.61 and the corresponding P-value in the ANOVA table clearly indicate that the model with the two variables and their interaction is helpful in predicting FEV. Looking at the individual t tests for the coefficients, we notice that all are highly significant and the model explains 66.6% of variations in FEV (R^2 = 66.6%). Therefore, all three variables are useful in explaining FEV, and the two regression lines (for boys and for girls) are not parallel. That is, the lung capacity of boys and girls does not develop similarly as children grow up. Lung capacity increases faster in boys.

The residual plots (not shown) indicate no major deviation from Normality or linearity, but possibly some minor concern with the assumption of constant variance.

CONCLUDE: The model with two regression lines, one for girls and one for boys, and interaction between age and gender explains approximately 67% of the variation in FEV values. This model provides a better fit than the simple linear regression model predicting FEV from age alone (R^2 = 61.1%, not shown). ■

Minitab

Session

```
Regression Analysis: fev versus age, sex, Age*Sex

The regression equation is
fev = 0.674 + 0.182 age - 0.731 sex + 0.106 age*sex

Predictor      Coef  SE Coef      T      P
Constant     0.6739   0.1069   6.31  0.000
age         0.18209  0.01097  16.60  0.000
sex         -0.7314   0.1475  -4.96  0.000
age*sex     0.10613  0.01490   7.12  0.000

S = 0.492471   R-Sq = 66.6%   R-Sq(adj) = 66.5%

Analysis of Variance

Source           DF       SS      MS       F      P
Regression        3  283.473  94.491  389.61  0.000
Residual Error  585  141.879   0.243
Total           588  425.351
```

FIGURE 28.9 Output from Minitab for the model with two regression lines in Example 28.13.

Even though, for simplicity, we have discussed models without interaction first, it is better in practice to consider models with interaction terms before going to the more restrictive model with parallel regression lines. If you begin your model fitting with the more restrictive model with parallel regression lines, you are basically assuming that there is no interaction. We won't discuss model selection formally, but deciding which model to use is an important skill.

The concept of interaction is fairly simple to visualize when there are only two variables and one of them is an indicator variable because we can simply see whether the two lines are parallel or not. But *multiple regression* models can include any number of explanatory variables, and interaction need not necessarily involve a simple indicator variable. In fact, the multiple linear regression model can even include squares or higher powers of quantitative explanatory variables, as long as the model is *linear* and each variable is multiplied by a constant β. Here are some conceptual examples that illustrate the flexibility of multiple regression models.

EXAMPLE 28.14 Two interacting explanatory variables

Suppose that we have n observations on two continuous explanatory variables x_1 and x_2 and a response variable y. Our goal is to predict the behavior of y for given values of x_1 and x_2 such that the mean response is given by

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Because there are two explanatory variables x_1 and x_2, we can graph the relationship of y with x_1 and x_2 in three dimensions. Figure 28.10 shows y vertically above a plane in which x_1 and x_2 take their values. The result is a surface in space. Figure 28.10(a) shows the easiest extension of our simple linear regression model from Chapter 23. Instead of fitting a line to the data, we are now fitting a plane. This figure shows the plane $\mu_y = x_1 + x_2$. The plane is a population model, and when we collect data on our explanatory variables, we will see vertical deviations from the points to the plane. The goal of our least-squares regression model is to minimize the vertical distances from the points to the plane.

Figure 28.10(b) adds a slight twist created by the interaction term in the model. The mean response in Figure 28.10(b) is $\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. The coefficients in front of the explanatory variables indicate part of the effect of a 1-unit change on the mean response for each 1-unit change in one of the explanatory variables. But the interpretation of the effect of a 1-unit change on the mean response for one variable also depends on the other variable. For example, if $x_2 = 1$, the mean response increases by 12 ($\mu_y = 2x_1 + 2(1) + 10x_1(1) = 2 + 12x_1$) for a 1-unit increase in x_1. However, when $x_2 = 2$, the mean response increases by 22 ($\mu_y = 2x_1 + 2(2) + 10x_1(2) = 4 + 22x_1$) for a 1-unit increase in x_1. *To interpret the parameters in multiple regression models, we think about the impact of one variable on the mean response while all the other variables are held fixed.*

Another way to think about possible changes in the mean response for different possible multiple regression models for two explanatory variables is to take a piece of paper and hold it as shown in Figure 28.10(a). Now begin moving the corners of the paper to get different surfaces. You will discover that a wide variety of surfaces are possible with only two explanatory variables.

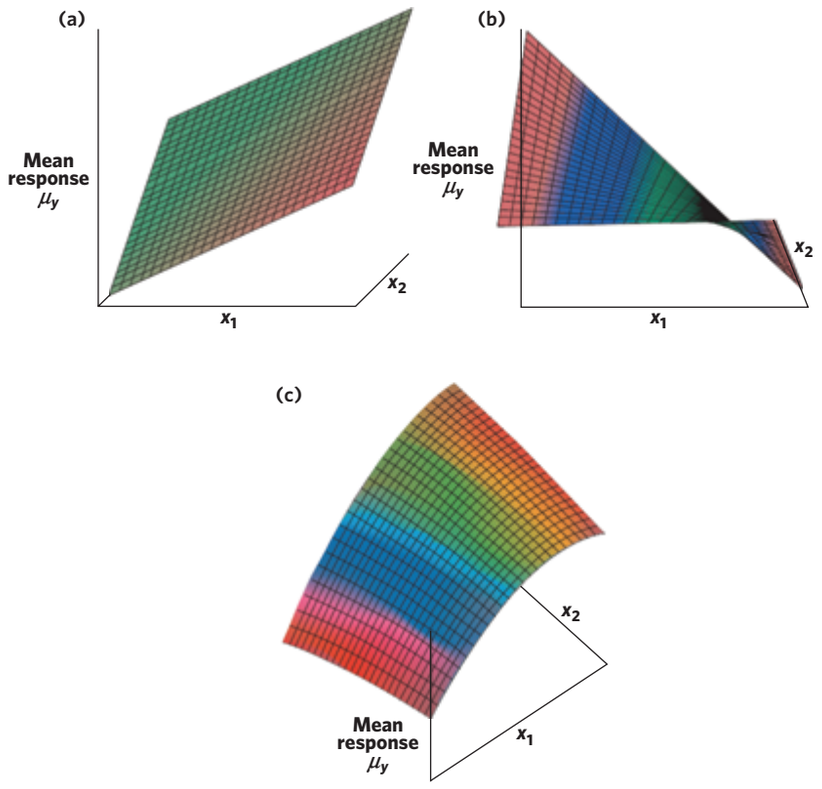


FIGURE 28.10 Some possible surfaces for multiple regression models, for Example 28.14. (a) Plane $\mu_y = x_1 + x_2$. (b) Surface $\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. (c) Surface $\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$.

Another possible response surface is shown in Figure 28.10(c). A quick inspection of this figure reveals some curvature in the mean response. Multiple linear regression models are linear in the parameters, so we can fit quadratic models by squaring the explanatory variables, or higher-order polynomial models by considering higher-order terms for the explanatory variables. However, these polynomial models require special care, beyond the scope of this textbook. The mean response in Figure 28.10(c) is $\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$. Notice that this mean response has two linear terms, two quadratic terms, and one interaction term. Models of this form are known as second-order polynomial regression models. ■

Software estimates the parameters just as before, finding the β's by the least-squares method and estimating σ by the regression standard error based on the residuals. Nothing is new there except more complicated calculations that software does for us and, of course, much more challenging interpretations.

One of the greatest practical challenges lies with regression coefficients because *the relationship between the response y and any one explanatory variable can change greatly, depending on what other explanatory variables are present in the model.* In fact, when the explanatory variables are correlated, multiple regression models can produce some very odd and counterintuitive results, so it is important to check carefully for correlation among a potential set of explanatory variables. We shall see a concrete example of this in the following case study.

APPLY YOUR KNOWLEDGE

28.11 Physiology of men and women. High blood pressure and arterial stiffness are risk factors for cardiovascular disease, the leading cause of death worldwide. The sympathetic nervous system is known to affect vasoconstriction. Researchers investigated the relationship between sympathetic nerve activity (in number of bursts per 100 heartbeats) and an indirect measure of arterial stiffness (augmented aortic pressure, in mm Hg) in 44 healthy young adults. The results are displayed in Table 28.2.[9]

(a) Plot arterial stiffness versus sympathetic nerve activity using different symbols for men and women. Does a multiple regression model with parallel slopes for men and women make sense? Explain.

(b) Create a new variable representing the interaction $x_1 x_2$ between the two explanatory variables. Then fit a model that includes both explanatory variables and their interaction. What do you conclude?

(c) Are the conditions for inference met for your model in part (b)? Construct appropriate residual plots and comment.

28.12 Neural mechanism of nociception: interaction. You can add an interaction term to test whether the lines in a model with two lines are likely to be parallel.

TABLE 28.2 Sympathetic nerve activity (bursts/100 heartbeats) and arterial stiffness (mm Hg) of men and women

IndMen	Sympathetic	Stiffness	IndMen	Sympathetic	Stiffness
1	12.8	−2.0	0	17.5	10.0
1	14.2	−1.1	0	20.1	10.0
1	17.7	−1.0	0	18.8	8.3
1	19.9	0.0	0	18.1	7.0
1	31.1	0.5	0	16.9	4.9
1	34.6	2.0	0	14.3	4.5
1	35.9	2.0	0	9.9	4.0
1	37.9	2.0	0	9.6	3.4
1	45.0	8.0	0	17.3	0.0
1	46.1	2.7	0	17.9	2.0
1	66.7	7.5	0	22.5	2.5
1	63.8	0.0	0	23.4	3.5
1	47.2	−4.0	0	23.4	4.3
1	36.5	0.0	0	27.7	3.0
1	34.7	−4.0	0	27.1	2.0
1	33.3	−4.5	0	34.2	−0.5
1	28.2	−5.5	0	34.9	1.5
1	29.1	−2.9	0	36.9	4.0
1	29.8	−2.5	0	37.1	4.4
1	24.1	−3.0	0	37.5	8.0
1	18.4	−5.5	0	49.2	−3.0
1	5.3	9.0	0	52.7	−4.0

Exercise 28.9 described the changes in percent reacting to light pulses in two subpopulations *C. elegans*. The scatterplot created for part (a) of that exercise showed two regression lines with negative slopes.

(a) Create a new variable equal to the product of Pulses and IndMutant. This is your interaction term, Pulses*Mutant. Then use software to create a multiple linear regression model of percent reacting as a function of Pulses, IndMutant, and Pulses*Mutant. Compare it with the Minitab output provided below and make sure that your results are similar, up to rounding errors.

(b) Interpret the software output for the model with interaction. Is the model significant overall? What percent of the variations in percent reacting is explained by the model?

(c) Obtain the equation of the estimated regression line for each subpopulation. Are the slopes significantly different for the two subpopulations? Explain your answers.

Session

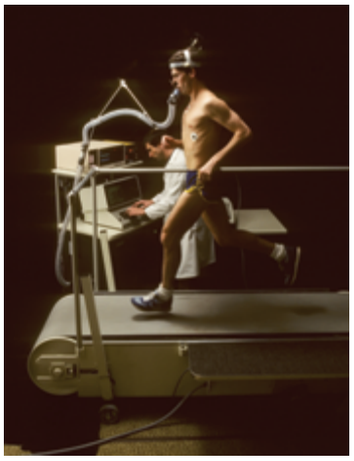
```
Regression Analysis: Reacting versus Pulses, IndMutant, Pulses*Mutant

The regression equation is
PercentReacting = 101 - 1.47 Pulses - 17.6 IndMutant - 0.789 Pulses*Mutant

Predictor          Coef    SE Coef       T       P
Constant         100.542      2.891   34.78   0.000
Pulses           -1.4707     0.2413   -6.09   0.000
IndMutant        -17.568      4.089   -4.30   0.000
Pulses*Mutant    -0.7887     0.3413   -2.31   0.027

S = 6.22379  R-Sq = 89.2%  R-Sq(adj) = 88.3%

Analysis of Variance

Source            DF        SS       MS       F       P
Regression         3   11515.3   3838.4   99.09   0.000
Residual Error    36    1394.5     38.7
Total             39   12909.8
```

A case study for multiple regression

We will now illustrate step-by-step the process of arriving at a suitable multiple regression model from raw data. To build a multiple regression model, first examine the data for outliers and other deviations that might unduly influence your conclusions. Next, use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response and to find out if some of the explanatory variables are also correlated. Fit several models using combinations of these variables, paying attention to the individual t statistics to see if any variables contribute little in this particular model.

Always think about the real-world setting of your data and use common sense as part of the process. When building a regression model, it is far more efficient

to start with a selection of variables that make sense biologically, than to search systematically through a large number of variables in a blind fashion.



Tom Tracy Photography/Alamy

EXAMPLE 28.15 Energy cost of running: exploratory analysis

If you run on a computerized treadmill at the gym and check the calories you have just burned, chances are that the machine's answer is based on the following 1963 study. Researchers asked athletes to run on a treadmill at various speeds and inclines and assessed the athletes' energy expenditure (computed indirectly via oxygen consumption and individual body measurements). Table 28.3 provides data from 25 measurements taken under various conditions of speed (in kilometers per hour) and treadmill incline (in percent).[10] We want to find a model that can explain the variations in energy expenditure.

The response variable y is energy expenditure. A careful examination of these values reveals no outliers or skew. We plot the energy values against speed and against incline in two separate scatterplots. They are shown in Figure 28.11(a) and (b), respectively. As expected, the treadmill's incline influences energy expenditure. This relationship is positive and linear. Surprisingly though, there is no direct relationship between energy expenditure and running speed in the data set.

A correlation analysis gives the following correlation coefficients and associated P-values:

Energy and Incline	$r = 0.726$, $P < 0.001$
Energy and Speed	$r = -0.022$, $P = 0.916$
Incline and Speed	$r = -0.686$, $P < 0.001$

The correlation analysis confirms what we saw in the scatterplots, but it also reveals that incline and speed are strongly correlated ($r = -0.686$). For the significant relationship between energy and incline, software gives the following simple linear regression model:

$$\hat{y} = 10.2 + 0.352\ Incline \quad ■$$

TABLE 28.3 Energy expenditure while running at various speeds (km/h) and treadmill inclines (%)

Speed	Energy	Incline	Speed	Energy	Incline
8.7	3.0	−10	13.8	13.5	0
11.3	4.4	−10	18.9	18.1	0
13.8	6.2	−10	10.0	13.5	5
16.3	7.8	−10	8.8	12.2	5
18.8	8.4	−10	6.3	9.1	5
21.4	10.0	−10	5.0	10.6	10
6.2	3.7	−5	6.9	13.1	10
8.8	5.9	−5	7.5	14.3	10
13.8	9.1	−5	8.8	16.7	10
16.3	12.0	−5	4.4	14.0	15
6.2	6.1	0	5.0	15.0	15
8.8	8.4	0	6.0	16.8	15
11.3	11.1	0			

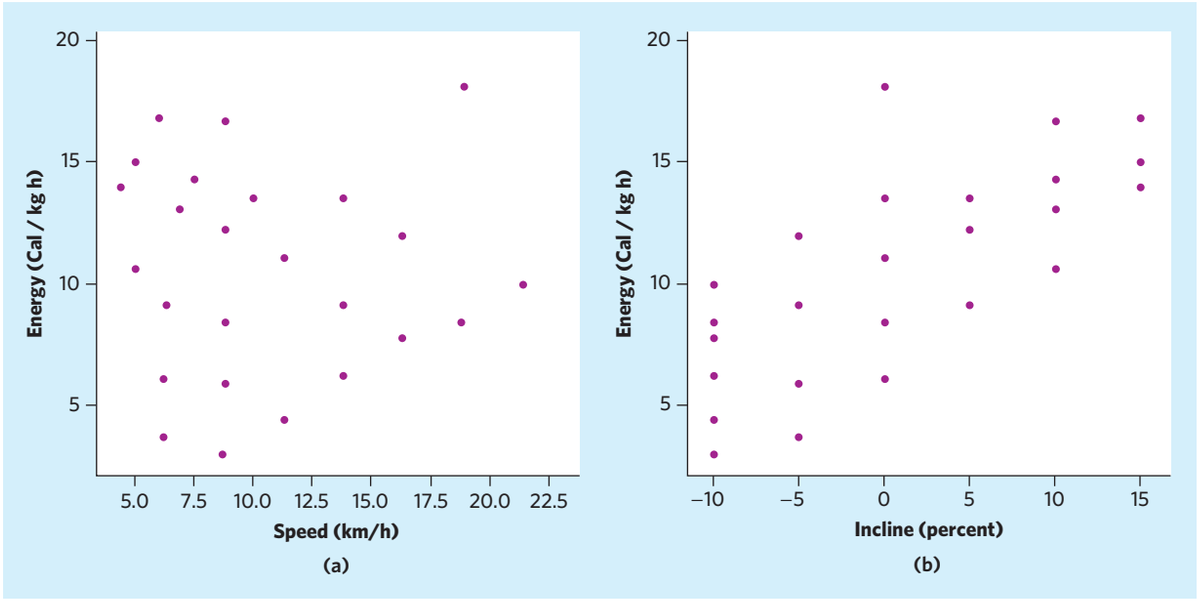


FIGURE 28.11 Scatterplots of energy expenditure against (a) speed and against (b) incline for Example 28.15.

The exploratory analysis shows that, in this particular context when *both* speed and incline are varied, speed alone cannot explain the observed variations in energy expenditure. Incline alone, on the other hand, explains 53% ($R^2 = 0.527$) of the variations in energy expenditure. Could we obtain an even better model?

The finding that speed and incline are correlated is also very important. It tells us that the effect of one may be offset or modified by the other. Therefore, although the relationship between energy and speed is not significant, we shouldn't discount speed in building a model of running energy expenditure. Also, if common sense or previous studies suggest that a variable may be influential, you should make sure to study it carefully. Common sense certainly points to some impact of running speed on energy expenditure.

EXAMPLE 28.16 Energy cost of running: displaying more complex relationships

One of the challenges of multiple regression is visualizing complex relationships. At this point, we would like to see how energy expenditure relates to *both* speed and incline. Figure 28.12(a) shows a 3D scatterplot of the three variables with a surface fit. The dots on this graph follow a clear pattern that depends on both speed and incline, supporting our suspicion that speed should impact energy expenditure in some way. If your software program creates 3D plots, try rotating the graph to better see the pattern.

In our case, incline is discrete, taking only 6 possible values. Therefore, we can create a scatterplot of energy against speed and color-code the dots for each incline value. Figure 28.12(b) shows such a plot. The relationship between energy, speed, and

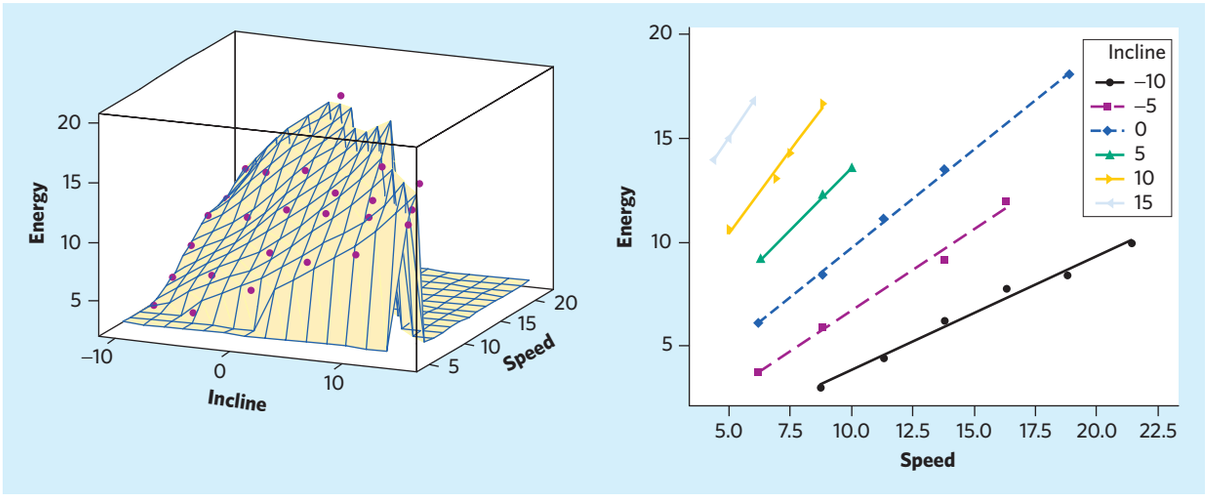


FIGURE 28.12 Scatterplots of energy expenditure against speed and incline using (a) a 3D-surface plot or a (b) 2D plot with regression lines of different colors for various inclines.

incline is now much clearer. For any fixed incline value, energy expenditure increases linearly with speed. For similar speeds, energy expenditure is higher for higher inclines. We also notice that the lines corresponding to different inclines appear to be somewhat steeper for higher inclines than for lower ones. ■

The more advanced graphing led us to expose the relationship between speed and energy in this experiment, even though we found that the correlation between the two is not significant. Therefore, we definitely want to include both speed and incline to build a model of energy expenditure. Because the lines in Figure 28.12(b) are not all parallel, we will also check whether an interaction term between speed and incline helps predict energy.

EXAMPLE 28.17 Energy cost of running: comparing models

We create an interaction variable x_1x_2, *Speed*Incline*. Figure 28.13 shows the multiple regression output using *Speed*, *Incline*, and their interaction *Speed*Incline*. The overall model is significant (P-value of ANOVA test < 0.0005), and residual plots (not shown) give no reason to discard the model. All regression coefficients in the model are significant too ($P < 0.0005$) and, together, they explain 98.2% of the variations in energy expenditure. The fitted model using all three explanatory variables is

$$\hat{y} = 1.25 + 0.471\ Incline + 0.888\ Speed + 0.0232\ Speed*Incline$$

The model using speed, incline, and their interaction is outstanding and does a much better job of explaining energy expenditure than the model using incline alone, discussed in Example 28.15. The next question is whether a simpler model could be

Minitab

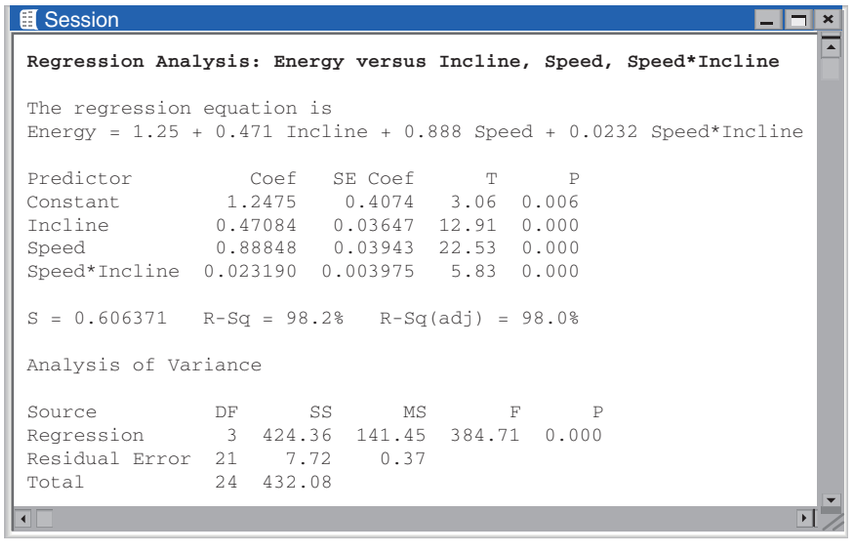


FIGURE 28.13 Output from Minitab for the model with two variables and their interaction term in Example 28.17.

almost as successful. (We know that removing an explanatory variable can only reduce R^2.)

Energy versus Speed, Incline	$R^2 = 95.3\%$
Energy versus Speed, Speed*Incline	$R^2 = 84.0\%$
Energy versus Incline, Speed*Incline	$R^2 = 55.0\%$

Two of the models have a drastically reduced squared multiple correlation coefficient. However, the model with speed and incline but no interaction term comes quite close, with $R^2 = 95.3\%$. The fitted model using only speed and incline is

$$\hat{y} = 1.77 + 0.782\ Speed + 0.651\ Incline$$

with significant regression coefficients for both speed and incline ($P < 0.0005$).

While a model with only speed and incline is quite satisfactory, the full model with incline, speed, and their interaction performs substantially better. The complete model shows that, when both speed and incline are taken into account, the interaction of these two variables plays a significant role in explaining variations in energy expenditure. This also reflects our observation that the color-coded lines in Figure 28.12(b) are not all parallel. ■

automated algorithms

Some statistical programs provide **automated algorithms** to choose regression models. All possible regression algorithms are very useful. However, automated programs do not add interaction terms, create new variables, or notice curved patterns in the residuals. *Automated algorithms that add or remove variables one at a time often miss good models*. Try building models by considering and evaluating various possible subsets of models.

It is helpful to start with clear objectives, if you can. Choose variables that fit these objectives, even in choosing the response variable. Notice that in

Example 28.17, energy expenditure is measured in Calories per kilogram of body weight and per hour. The researchers could have used a simpler measure of Calories burned. However, the duration of the runs and the physiological characteristics of the runners would have added a lot of variability and made the model selection a lot more challenging. The model we have created can be used quite successfully to predict the number of Calories burned per unit of time for a person with a given weight.

APPLY YOUR KNOWLEDGE

28.13 Fish sizes. Table 28.4 contains data on the size of perch caught in a lake in Finland.[11] Use statistical software to help you analyze these data.

TABLE 28.4 Measurements on 56 perch

Observation number	Weight (g)	Length (cm)	Width (cm)	Observation number	Weight (g)	Length (cm)	Width (cm)
104	5.9	8.8	1.4	132	197.0	27.0	4.2
105	32.0	14.7	2.0	133	218.0	28.0	4.1
106	40.0	16.0	2.4	134	300.0	28.7	5.1
107	51.5	17.2	2.6	135	260.0	28.9	4.3
108	70.0	18.5	2.9	136	265.0	28.9	4.3
109	100.0	19.2	3.3	137	250.0	28.9	4.6
110	78.0	19.4	3.1	138	250.0	29.4	4.2
111	80.0	20.2	3.1	139	300.0	30.1	4.6
112	85.0	20.8	3.0	140	320.0	31.6	4.8
113	85.0	21.0	2.8	141	514.0	34.0	6.0
114	110.0	22.5	3.6	142	556.0	36.5	6.4
115	115.0	22.5	3.3	143	840.0	37.3	7.8
116	125.0	22.5	3.7	144	685.0	39.0	6.9
117	130.0	22.8	3.5	145	700.0	38.3	6.7
118	120.0	23.5	3.4	146	700.0	39.4	6.3
119	120.0	23.5	3.5	147	690.0	39.3	6.4
120	130.0	23.5	3.5	148	900.0	41.4	7.5
121	135.0	23.5	3.5	149	650.0	41.4	6.0
122	110.0	23.5	4.0	150	820.0	41.3	7.4
123	130.0	24.0	3.6	151	850.0	42.3	7.1
124	150.0	24.0	3.6	152	900.0	42.5	7.2
125	145.0	24.2	3.6	153	1015.0	42.4	7.5
126	150.0	24.5	3.6	154	820.0	42.5	6.6
127	170.0	25.0	3.7	155	1100.0	44.6	6.9
128	225.0	25.5	3.7	156	1000.0	45.2	7.3
129	145.0	25.5	3.8	157	1100.0	45.5	7.4
130	188.0	26.2	4.2	158	1000.0	46.0	8.1
131	180.0	26.5	3.7	159	1000.0	46.6	7.6


Design Pics Inc./Alamy

(a) Use the multiple regression model with two explanatory variables, length and width, to predict the weight of a perch. Provide the estimated multiple regression equation.

(b) How much of the variation in the weight of perch is explained by the model in part (a)?

(c) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.

(d) Do the individual t tests indicate that both β_1 and β_2 are significantly different from zero? Explain.

(e) Create a new variable, called interaction, that is the product of length and width. Use the multiple regression model with three explanatory variables—length, width, and interaction—to predict the weight of a perch. Provide the estimated multiple regression equation.

(f) How much of the variation in the weight of perch is explained by the model in part (e)?

(g) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.

(h) Describe how the individual t statistics changed when the interaction term was added.

28.14 Tasting cheddar cheese. What determines the taste of cheddar cheese? Experimenters assessed the concentration of lactic acid, acetic acid, and hydrogen sulfide (H_2S) in 30 randomly chosen pieces of cheddar cheese that were also rated for taste by a panel of food tasters (the higher the score, the tastier). Data from the study are shown in Table 28.5. "Acetic" and "H_2S" are actually log-transformed concentrations.[12]

TABLE 28.5 Chemical composition and taste score of aged cheddar cheese

Taste	Acetic	H_2S	Lactic	Taste	Acetic	H_2S	Lactic
12.3	4.543	3.135	0.86	40.9	6.365	9.588	1.74
20.9	5.159	5.043	1.53	15.9	4.787	3.912	1.16
39.0	5.366	5.438	1.57	6.4	5.412	4.700	1.49
47.9	5.759	7.496	1.81	18.0	5.247	6.174	1.63
5.6	4.663	3.807	0.99	38.9	5.438	9.064	1.99
25.9	5.697	7.601	1.09	14.0	4.564	4.949	1.15
37.3	5.892	8.726	1.29	15.2	5.298	5.220	1.33
21.9	6.078	7.966	1.78	32.0	5.455	9.242	1.44
18.1	4.898	3.850	1.29	56.7	5.855	10.199	2.01
21.0	5.242	4.174	1.58	16.8	5.366	3.664	1.31
34.9	5.740	6.142	1.68	11.6	6.043	3.219	1.46
57.2	6.446	7.908	1.90	26.5	6.458	6.962	1.72
0.7	4.477	2.996	1.06	0.7	5.328	3.912	1.25
25.9	5.236	4.942	1.30	13.4	5.802	6.685	1.08
54.9	6.151	6.752	1.52	5.5	6.176	4.787	1.25

(a) For each of the three explanatory variables in turn, make a scatterplot with taste on the y axis and find the correlation coefficient. Which relationships are linear? Which have the strongest correlation with taste?

(b) Use software to obtain the regression equation and run inference for a regression model that includes all three explanatory variables. Interpret the software output, including the meaning of the value taken by R^2.

(c) The model in (b) has one nonsignificant regression coefficient. Which explanatory variable does it describe? Now create a new regression model that excludes this explanatory variable. Interpret the software output and compare it with your findings in (b).

(d) Both regression coefficients in (c) are significant. Which explanatory variable of the two has the less significant (larger) P-value? Create a new regression model that excludes this explanatory variable and keeps only the more significant one. How does this last model compare with the model in (c)?

(e) Which model best explains cheddar cheese taste? Check the conditions for inference for this model and conclude. The approach you have used here to select a multiple linear regression model is called **backward selection** because you start with all reasonable explanatory variables and gradually prune the model back until you get a satisfactory one.

backward selection

Logistic regression

So far, we have examined how a quantitative response variable can be expressed as a linear function of any number of either quantitative or categorical explanatory variables. A simple linear regression model, for instance, represents the mean response μ_y for a given value of the explanatory variable x in the form of $\mu_y = \beta_0 + \beta_1 x$. Can we apply the same principles to a categorical response variable?

We have seen in Chapters 12 and 19 that when a variable y is categorical, we can describe it in a binary fashion such that any outcome is either a success or a failure (arbitrarily defined). For example, we can ask whether a patient's health improves after treatment (success) or not (failure). For any given value of an explanatory variable x, the mean response is the population proportion of successes $p_y = p$ for that particular value of x. This is also the probability of randomly obtaining a success when x takes this particular value.

If there is no relationship between x and y, the mean response is simply p for any value of x. But what if there is an association between x and y? Can we find a simple linear model that expresses p as a function of x? Unfortunately, this is rather unlikely. Indeed, as a probability or a proportion, p can take only values between 0 and 1. On the other hand, a linear function of a quantitative variable x is likely to create values over a whole range of intervals, especially for extreme values of x. So the relationship between p and a quantitative variable x is unlikely to be truly linear.

A number of examples in this and other chapters on regression discussed relationships that are not linear. It is sometimes possible to reveal a linear relationship by using a simple mathematical transformation, such as computing the logarithm

of one or both quantitative variables. We now apply the same idea to describing relationships in which the response variable is categorical.

In Chapter 9 we discussed the concept of odds in probability. Odds are the ratio of two probabilities where the numerator is the probability p of an event and the denominator is the complementary probability of that event not occurring. That is,

$$\text{odds} = \frac{p}{1-p}$$

Provided that neither probability is equal to zero, odds can take any strictly positive value. In turn, the natural logarithm of odds can take any real value, positive or negative. We can therefore hope to find a linear relationship between the natural logarithm of odds (called log odds, logit, or L) and some relevant continuous random variable x.

LOGISTIC REGRESSION

The **logistic regression model** is

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where p is the proportion or probability of a given outcome in a population, x is an explanatory variable, and L is the natural logarithm of the odds of that outcome in the population.

The model can be extended to include k explanatory variables, such that

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \cdots + \beta_k x_k$$

As with the relationship between two quantitative variables, not all (L, x) pairs of variables studied will show a linear trend, but those that do can be modeled and studied with the tools of linear regression. Note also that the logistic model does not represent values of p that are either 0 or 1. Instead, it provides models in which p can come arbitrarily close to 0 or 1.

Figure 28.14(a) illustrates how a linear relationship between L and x translates when p is expressed as a function of x. Figure 28.14(b) shows that the logistic regression model corresponds to a relationship between p and x that has an "S," or "sigmoidal," shape. In practice, you may observe only a subset of that sigmoidal curve if your explanatory variable x takes values over a limited interval only. We can derive the equation for that sigmoidal curve by using the antilog transformation:

$$\frac{p}{1-p} = e^{\ln(p/(1-p))} = e^{\beta_0+\beta_1 x}$$

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

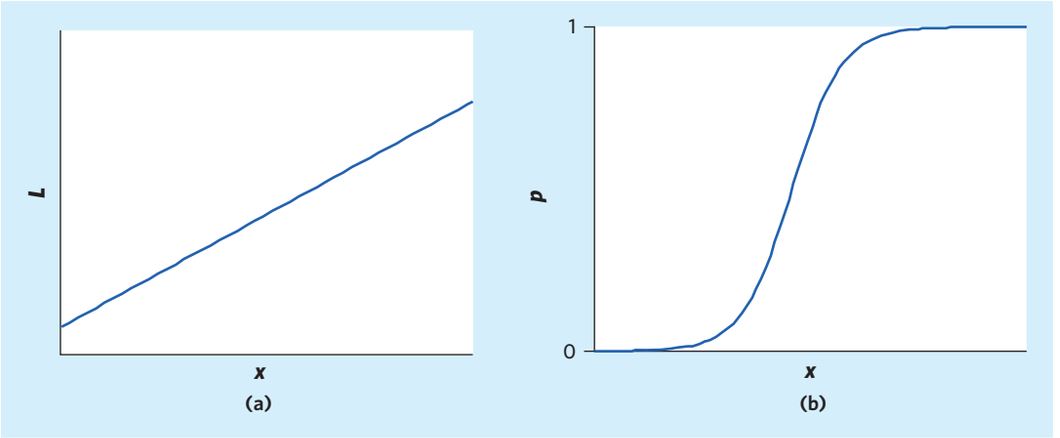


FIGURE 28.14 (a) Linear relationship between the logit L and an explanatory variable x and (b) its equivalent when L is untransformed to obtain p, the probability of a given outcome. The relationship between p and x has a sigmoidal, or "S," shape.

Therefore, there are three versions of the same model:

$$(1) \quad L = \ln(\text{odds}) = \beta_0 + \beta_1 x$$

$$(2) \quad \text{odds} = \frac{p}{1-p} = e^{\beta_0+\beta_1 x}$$

$$(3) \quad p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

We are often more interested in the odds (2) or the probability p of success (3), but we first estimate the logit model (1) because of its linear property. The mathematical process needed to obtain parameter estimates for this linear model is somewhat computationally intensive. Explaining these computations is beyond the scope of an introductory statistics textbook, and we will rely solely on technology to obtain these estimates.

EXAMPLE 28.18 Diagnosing meningitis

Meningitis is a generic term that describes an inflammation of the meninges, the outer membranes protecting the brain. The source of the inflammation is usually either a viral or a bacterial infection. Meningitis is potentially deadly and must be treated promptly.

Researchers examined the test results of 352 patients with acute meningitis that were later unambiguously identified as either a viral or a bacterial infection.[13] They defined the variable *abm* (acute bacterial meningitis) as a binary response variable (y) that equals 0 when the infection is viral and equals 1 when the infection is bacterial. To help with diagnosis, the immune system's response to infection can be assessed by examining the white cell count per cubic millimeter of cerebrospinal fluid filling the meninges. Does this white cell count, *wcsf (x)*, help predict whether a case of acute meningitis is viral or bacterial?

Figure 28.15 shows dotplots of *wcsf* for viral ($abm = 0$) and for bacterial ($abm = 1$) meningitis cases. The two plots are stacked and share a common x axis, making the comparison particularly easy. The two distributions overlap, but all individuals who had a high white cell count ($wcsf > 500$) had bacterial meningitis. On the other hand, all

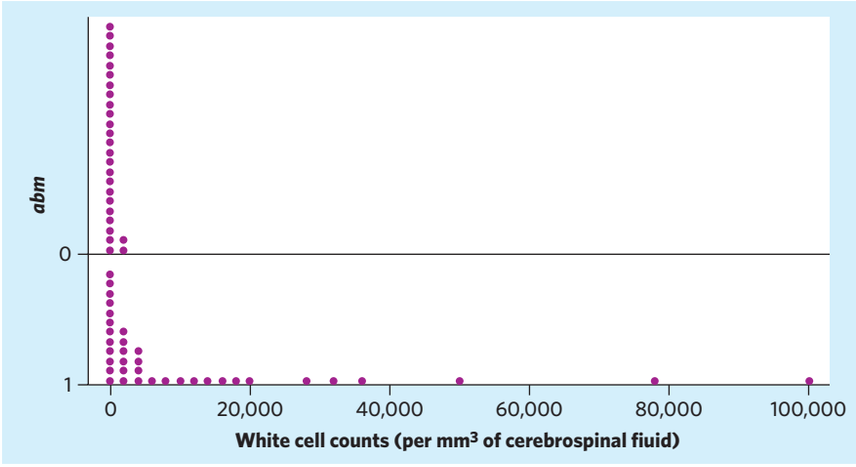


FIGURE 28.15 Dotplots of white cell counts (per mm^3 of cerebrospinal fluid) in patients with viral ($abm = 0$) or bacterial ($abm = 1$) meningitis, for Example 28.18. Note that each dot represents up to seven observations.

Minitab

Session

```
Binary Logistic Regression: abm versus wcsf

Logistic Regression Table
                                                    Odds      95% CI
Predictor       Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant    -1.19627   0.170705   -7.01  0.000
wcsf       0.0018239  0.0002824    6.46  0.000   1.00   1.00   1.00
```

FIGURE 28.16 Minitab output for the simple logistic regression model in Example 28.18.

patients with viral meningitis had relatively low white cell counts in the cerebrospinal fluid. Therefore, there is clearly some relationship between *abm* and *wcsf*.

Figure 28.16 shows part of Minitab's output for the logistic regression of *abm* as a function of *wcsf*. The estimates for β_0 and β_1 are -1.1963 and 0.0018, respectively, so that the fitted logistic regression model is

$$L = \ln\left(\frac{p}{1-p}\right) = -1.1963 + 0.0018x$$

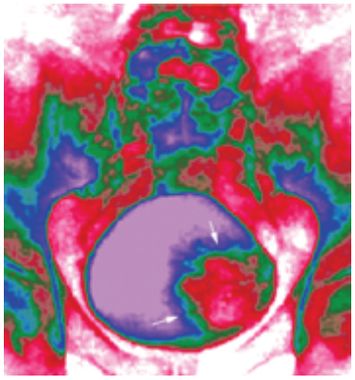
If we convert this equation to express the estimated probability p that a given case of acute meningitis is due to a bacterial infection as a function of the white cell count per cubic millimeter x, we find that

$$p = \frac{e^{-1.1963+0.0018x}}{1+e^{-1.1963+0.0018x}}$$

The logistic regression model allows us to estimate the probability p that a given case is due to bacterial infection for different test result values of x. Here are some results:

x(*wcsf*)	0	100	500	1000	10,000
p(bacterial)	0.2321	0.2657	0.4265	0.6465	0.99999995

Low white cell counts in the cerebrospinal fluid indicate that a viral infection is more likely, whereas high counts indicate that a bacterial infection is more likely, and very high counts indicate that a bacterial infection is almost certain. ■



James Cavallini/Science Source

APPLY YOUR KNOWLEDGE

28.15 Recurring bladder tumors. Why do some cancers come back after a period of remission while others don't? Can we predict which cancers are more likely to recur? Researchers examined 86 patients with bladder cancer, looking for a relationship between cancer recurrence y (0 no, 1 yes) and the number x of superficial bladder tumors identified and removed right after the initial cancer diagnosis.[14] Use the Minitab output for the logistic regression analysis to find the linear equation linking the estimated logit L to the response variable x.

Session

```
Binary Logistic Regression: recurrence versus number

Logistic Regression Table

                                                  Odds    95%     CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.596302  0.376309  -1.58  0.113
number      0.400574  0.169022   2.37  0.018   1.49   1.07   2.08
```

28.16 Recurring bladder tumors, continued. Return to the previous exercise. We are interested in the probability p that a patient with x number of initial tumors will experience a cancer recurrence. Convert the logistic equation so that you express p as a function of x. Then use this equation to estimate the probability p of cancer recurrence when $x = 0$, when $x = 2$, and when $x = 6$.

Inference for logistic regression

Inference for logistic regression is very similar to inference for regression with a continuous response variable, except that the actual computations are more complex. In particular, the sample statistic follows approximately a Normal distribution only when the sample size is large enough.[15] (This is again an issue related to the central limit theorem we first discussed in Chapter 13.) We will rely on software for the computations and focus on interpreting the results.

INFERENCE FOR THE SLOPE OF THE LOGISTIC REGRESSION MODEL

To **test the hypothesis** H_0: $\beta_1 = 0$, compute the test statistic

$$z = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{b_1}{\text{SE}_{b1}}$$

Approximate P-values for this test come from the standard Normal distribution.

An approximate **level C confidence interval for the slope** β_1 has the form

$$b_1 \pm z^* \text{SE}_{b1}$$

where z^* is the critical value for the standard Normal curve with area C between $-z^*$ and z^*.

EXAMPLE 28.19 Diagnosing meningitis

The output in Figure 28.16 provides parameter estimates and standard errors for the slope β_1: $b_1 = 0.0018$ and $SE_{b1} = 0.00028$. The test of H_0: $\beta_1 = 0$ indicates that the linear relationship between L and x is significant ("P 0.000" implies that $P < 0.0005$). Minitab does not provide a confidence interval for β_1, but we can compute one very easily. A 95% confidence interval for the slope β_1 is $b_1 \pm z^* SE_{b1} = 0.0018 \pm (1.96)(0.00028)$, or 0.0013 to 0.0023. That is, for every increase of 1 white cell per mm^3, the logit L increases by 0.0013 to 0.0023, with 95% confidence. ■

odds ratio

The log of odds is not an intuitive concept, and we often prefer to translate it back into odds. In particular, the ratio of two odds, or **odds ratio (OR)**, is a commonly used comparative measure of health risk in medical science. We first discussed this concept in Chapter 20.

Suppose that the explanatory variable x increases by 1. In simple linear regression, the corresponding change in y is the slope. In logistic regression, we measure the impact of this change in the odds of success. This is the odds ratio of $(x + 1)$ to x:

$$OR = \frac{\text{odds}_{x+1}}{\text{odds}_x}$$

Because $e^a/e^b = e^{a-b}$ and $e^{\ln(a)} = a$, it follows that

$$OR = \frac{\text{odds}_{x+1}}{\text{odds}_x} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

An odds ratio equal to 1 indicates no effect (same risk for x and $x + 1$) and is equivalent to $\beta_1 = 0$.

CONFIDENCE INTERVALS FOR THE ODDS RATIO

A **level C confidence interval for the odds ratio** e^{β_1} can be obtained by transforming the confidence interval for the slope, resulting in the interval

$$(e^{b_1 - z^* SE_{b1}}, e^{b_1 + z^* SE_{b1}})$$

where z^* is the critical value for the standard Normal curve with area C between $-z^*$ and z^*.

This particular odds ratio may or may not have a practical meaning. In the meningitis example, for instance, an increase of 1 white cell per mm^3 is not medically relevant and brings us no further understanding. However, when the explanatory variable in a logistic regression model is categorical, the odds ratio is commonly used to compare the odds between two conditions.

EXAMPLE 28.20 Preventing blood clots in immobilized patients

Patients immobilized for a substantial amount of time can develop deep vein thrombosis (DVT), a blood clot in a leg or pelvis vein. DVT can have serious adverse health effects and can be difficult to diagnose. On its website, drug manufacturer Pfizer reports the outcome of a study looking at the effectiveness of the drug Fragmin (dalteparin) compared with that of a placebo in preventing DVT in immobilized patients.

In a double-blind study, severely immobilized patients were randomly assigned to receive daily subcutaneous injections of either Fragmin or a placebo for two weeks. The number of patients experiencing a complication from DVT (including death) over the next 90 days is shown in the following table:

	Treatment Outcome		
	Complication	No complication	Sample size
Fragmin	42	1476	1518
Placebo	73	1400	1473

The proportions of subjects experiencing DVT complications in the two samples are

$$\hat{p}_{\text{Fragmin}} = 42/1518 = 0.0277$$
$$\hat{p}_{\text{placebo}} = 73/1473 = 0.0496$$

The odds of a subject's experiencing DVT complications in the two samples are

$$\text{odds}_{\text{Fragmin}} = 42/1476 \approx 0.02846 \text{ and } \ln(\text{odds}_{\text{Fragmin}}) \approx -3.5594$$
$$\text{odds}_{\text{placebo}} = 73/1400 \approx 0.05214 \text{ and } \ln(\text{odds}_{\text{placebo}}) \approx -2.9538$$

If we set the indicator variable *treatment*, or x, equal to 0 for the placebo group and 1 for the Fragmin group, we can derive from the above information the equation for the logistic regression model:

$$L = \ln(\text{odds}_{\text{DVT}}) = \ln(\text{odds}_{\text{placebo}}) + [\ln(\text{odds}_{\text{Fragmin}}) - \ln(\text{odds}_{\text{placebo}})]x$$
$$L = \ln(\text{odds}_{\text{DVT}}) = -2.9538 + (-3.5594 + 2.9538)x = -2.9538 - 0.6056x$$

Minitab gives the same result, up to rounding error, as can be seen from the output shown here:

Session

```
Binary Logistic Regression: DVT, n versus treatment

Logistic Regression Table

                                                    Odds    95%     CI
Predictor       Coef   SE Coef       Z      P  Ratio  Lower  Upper
Constant    -2.95377  0.120054  -24.60  0.000
treatment  -0.605653  0.197231   -3.07  0.002   0.55   0.37   0.80
```

The estimated odds ratio of DVT complication is

$$\frac{\text{odds}_{x=1}}{\text{odds}_{x=0}} = e^{b_1} = e^{-0.6056} \approx 0.546$$

This estimate is also the observed ratio of the two experimental odds, $\text{odds}_{\text{Fragmin}}/\text{odds}_{\text{placebo}} \approx 0.02846/0.05214 \approx 0.546$. A 95% confidence interval for the odds ratio can be calculated to be

$$\begin{aligned}(e^{b_1 - z^*\text{SE}_{b1}}, e^{b_1 + z^*\text{SE}_{b1}}) &= (e^{-0.6056-(1.96)(0.1972)}, e^{-0.6056+(1.96)(0.1972)}) \\ &= (e^{-0.9921}, e^{-0.2291}) \approx (0.371, 0.803)\end{aligned}$$

Our calculations agree with the Minitab output, up to rounding error. Notice that the confidence interval for *OR* is not symmetric around the estimate. This is because the exponential function is nonlinear.

The *OR* confidence interval means that, with 95% confidence, the odds of DVT complication in the Fragmin group are somewhere between 0.37 and 0.80 times the corresponding odds for the placebo group. Because we are 95% confident that *OR* is less than 1, we can conclude that the Fragmin treatment helps reduce the odds of DVT complication in immobilized patients. ■

APPLY YOUR KNOWLEDGE

28.17 Recurring bladder tumors, continued. Go back to Exercise 28.15 and use Minitab's output in your answers. Is the slope of the logistic regression model significant? Give the 95% confidence interval for the slope of the logistic regression. Can you conclude with 95% confidence that the slope is positive? What does this mean in terms of the impact of the number of initial tumors on the probability of cancer recurrence?

28.18 Antibiotics in the ICU. Patients who are admitted to the intensive care unit (ICU) and require mechanical ventilation can develop fatal bacterial infections. A clinical trial randomly assigned to two groups 934 patients who were admitted to an ICU and required mechanical ventilation. The control group received a standard medical treatment, whereas the antibiotics group received antibiotics prophylactically in addition to standard treatment. The study found that 107 of the 468 patients in the control group died, compared with 69 of the 466 patients in the antibiotics group.[16]

(a) What are the odds of dying in each of the two groups? Use this information to compute the equation of the logistic regression model if the indicator variable x equals 0 for the control group and 1 for the antibiotics group.

(b) Below is Minitab's logistic regression output for these data. Show how you would compute the odds ratio for death and its 95% confidence interval. Check that your calculations agree with Minitab's.

(c) Interpret your results in the context of the study.

Session

Binary Logistic Regression: death versus antibiotic

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-1.21605	0.110072	-11.05	0.000			
antibiotic	-0.533781	0.170668	-3.13	0.002	0.59	0.42	0.82

CHAPTER 28 SUMMARY

- An indicator variable x_2 can be used to fit a regression model with **two parallel lines.** The mean response is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where x_1 is a continuous explanatory variable.

- A multiple regression model with **two regression lines** includes a continuous variable x_1, an indicator variable x_2, and an interaction term $x_1 x_2$. The mean response is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- The mean response μ_y for a general **multiple regression model** based on k explanatory variables $x_1, x_2, \ldots, x_k$ is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- The **estimated regression model** is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

where the b's are obtained by the method of least squares. Use software to obtain these estimates.

- The **regression standard error** s has $n - k - 1$ degrees of freedom and is used to estimate σ.
- The **analysis of variance (ANOVA) table** breaks the total variability in the responses into two pieces. One piece summarizes the variability due to the model, and the other piece summarizes the variability due to error.

total variation = variation explained by model + residual ("error") variation

- The **squared multiple correlation coefficient** R^2 represents the proportion of variability in the response variable y that is explained by the explanatory variables $x_1, x_2, \ldots, x_k$ in a multiple regression model.
- To test the hypothesis that all the regression coefficients (β's) except β_0 are equal to zero, use the **ANOVA *F* statistic.** The null hypothesis says that the x's do not help predict y. The alternative is that at least one of the explanatory variables is helpful in predicting y.
- **Individual *t* procedures** in regression inference have $n - k - 1$ degrees of freedom. These individual t procedures are dependent on the other explanatory variables specified in a multiple regression model. Individual t tests assess the contribution of one explanatory variable in the presence of the other variables in a model. The null hypothesis is written as H_0: $\beta = 0$ but interpreted as "the coefficient of x is 0 *in this particular model.*"
- **Confidence intervals** for the mean response μ_y have the form $\hat{y} \pm t^* \mathrm{SE}_{\hat{\mu}}$.
- **Prediction intervals** for an individual future response y have the form $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$.

- When the **response variable is categorical,** it can be coded as $y = 1$ for a given outcome (often arbitrarily defined as "success") and $y = 0$ otherwise ("failure"). The population mean response is the probability p of the given outcome (success) in the population. We use **logistic regression** methods to describe the relationship between p and any set of explanatory variables.
- The simple **logistic regression model** with one explanatory variable x is

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where L is the natural logarithm of the odds of a given outcome in the population. The estimated model is $L_{\text{estimate}} = b_0 + b_1 x$, which can be rewritten as

$$p_{\text{estimate}} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

- A Normal z procedure is used to **test the hypothesis** that the slope β_1 is equal to zero and to compute an approximate **level C confidence interval for the slope** of the form $b_1 \pm z^* \text{SE}_{b1}$.
- Logistic regression models often report an approximate **level C confidence interval for the odds ratio** $\mathbf{OR = e^{\beta_1}}$**,** which is obtained by transforming the confidence interval for the slope, giving the interval

$$(e^{b_1 - z^* \text{SE}_{b1}},\ e^{b_1 + z^* \text{SE}_{b1}})$$

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. Preliminaries

1. Examine the data for outliers and other deviations that might influence your conclusions.
2. Use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response.
3. Make scatterplots to examine the relationships between explanatory variables and a response variable.
4. Use software to compute a correlation matrix to explore the relationships between pairs of variables.

B. Recognition

1. Recognize when a multiple regression model with parallel regression lines is appropriate.
2. Recognize when an interaction term needs to be added to fit a multiple regression model with two separate regression lines.

3. Recognize when a multiple regression model with several explanatory variables is appropriate.
4. Recognize the difference between the overall F test and the individual t tests.
5. Recognize when the explanatory variable is categorical, requiring the use of a logistic regression model.
6. Recognize that the parameter estimates, t statistics, and P-values for each explanatory variable depend on the specific model.
7. Inspect the data to recognize situations in which inference isn't safe: influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression model.

C. Doing Inference Using Computer Output

1. Use software to find the estimated multiple regression model or logistic model.
2. Explain the meaning of the regression parameters (β's) in any specific multiple regression model.
3. Understand the software output for regression. Find the regression standard error, the squared multiple correlation coefficient R^2, and the overall F test and P-value. Identify the parameter estimates, standard errors, individual t tests, and P-values.
4. Use that information to carry out tests and calculate confidence intervals for the β's.
5. Use R^2 and residual plots to assess the fit of a model.
6. Choose a model by comparing R^2-values, regression standard errors, and individual t statistics.
7. Explain the distinction between a confidence interval for the mean response and a prediction interval for an individual response.
8. Understand what the odds ratio represents in a logistic model, and use the confidence interval to determine whether a change in condition results in a significant increase or decrease in odds.

THIS CHAPTER IN CONTEXT

Chapters 3 and 4 described the relationship between *two quantitative variables*. We used scatterplots to visualize patterns and identify linear relationships in particular. We obtained the equation of the least-square regression line to model linear relationships. In Chapter 23 we used statistical inference procedures to ask whether the linear pattern in the sample data would hold for the entire population and to estimate parameters such as the regression slope and the mean response.

In this chapter we expand the concept of simple linear regression to include multiple explanatory variables used to describe and predict one response variable. We use inference procedures to determine which of these explanatory variables,

alone or in sets, has a statistically significant effect on the response variable. Some of these statistical procedures are variants of the t procedures from Chapter 17 and ANOVA from Chapter 24. In fact, in this chapter we find out how a categorical variable can be coded with zeros and ones to allow it into a linear regression model, instead of being used as a factor as in ANOVA.

A categorical response variable can also be coded with zeros and ones for the purpose of regression, as we describe here in the context of logistic regression. Logistic regression relies on the concepts of odds and odds ratios, which were first examined in Chapters 9 and 20, respectively.

CHECK YOUR SKILLS

Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, Calories burned per hour (or total Calories), and duration of the workout. The data in Table 28.6 show the treadmill display's claimed Calories per hour by speed (miles per hour) for a 175-pound male using a Cybex treadmill at inclines of 0%, 2%, and 4%.

The relationship between speed and Calories is different for walking and running, so we need an indicator for slow/fast. The variables created from Table 28.6 are

$Calories = Calories\ burned\ per\ hour$
$MPH = speed\ of\ the\ treadmill$
$Incline = the\ incline\ percent (0, 2, or\ 4)$
$Ind\_slow = 1\ for\ MPH \leq 3\ and$
$Ind\_slow = 0\ for\ MPH > 3.0$

Here is part of the Minitab output from fitting a multiple regression model to predict Calories from MPH, Ind_slow, and Incline for the Cybex treadmill:

Minitab

Session

```
Predictor      Coef  SE Coef       T      P
Constant    -80.41    18.99   -4.24  0.000
MPH         145.841    2.570   56.74  0.000
Ind_slow    -50.01    16.04   -3.12  0.003
Incline      36.264    2.829   12.82  0.000

S = 33.9422    R-Sq = 99.3%    R-Sq(adj) = 99.3%

Analysis of Variance

Source           Df       SS       MS        F      P
Regression        3  8554241  2851414  2475.03  0.000
Residual Error   50    57604     1152
Total            53  8611845

Predicted Values for New Observations
New
Obs     Fit   SE Fit         95% CI              95% PI
  1  940.09     5.28    (929.49, 950.69)    (871.09, 1009.08)

Values of Predictors for New Observations
New
Obs    MPH    Ind_slow   Incline
  1   6.50   0.000000      2.00
```

TABLE 28.6 Cybex treadmill display's claimed Calories per hour by speed and incline for a 175-lb man

	Incline (%)		
MPH	0	2	4
1.5	174	207	240
2.0	205	249	294
2.5	236	291	347
3.0	267	333	400
3.5	372	436	503
4.0	482	542	607
4.5	592	649	709
5.0	701	756	812
5.5	763	824	885
6.0	825	892	959
6.5	887	960	1032
7.0	949	1027	1105
7.5	1011	1094	1178
8.0	1073	1163	1252
8.5	1135	1230	1325
9.0	1197	1298	1398
9.5	1259	1365	1470
10.0	1321	1433	1544

Exercises 28.19 to 28.26 are based on this output.

28.19 The equation for predicting Calories from these explanatory variables is

(a) $Calories = -80.41 + 145.84MPH - 50.01Ind\_slow + 36.26Incline$.
(b) $Calories = -4.24 + 56.74MPH - 3.12Ind\_slow + 12.82Incline$.
(c) $Calories = 18.99 + 2.57MPH + 16.04Ind\_slow + 2.83Incline$.

28.20 The regression standard error for these data is

(a) 0.993. (b) 33.94. (c) 1152.

28.21 To predict Calories when walking with no incline, use the line

(a) $-80.41 + 145.84MPH$.
(b) $(-80.41 - 50.01) + 145.84MPH$.
(c) $[-80.41 + (2 \times 36.26)] + 145.84MPH$.

28.22 To predict Calories when running with no incline, use the line

(a) $-80.41 + 145.84MPH$.
(b) $(-80.41 - 50.01) + 145.84MPH$.
(c) $[-80.41 + (2 \times 36.26)] + 145.84MPH$.

28.23 To predict Calories when running on a 2% incline, use the line

(a) $-80.41 + 145.84MPH$.
(b) $(-80.41 - 50.01) + 145.84MPH$.
(c) $[-80.41 + (2 \times 36.26)] + 145.84MPH$.

28.24 Is there significant evidence that more Calories are burned for higher speeds? To answer this question, you test the hypotheses

(a) $H_0: \beta_0 = 0$ versus $H_a: \beta_0 > 0$.
(b) $H_0: \beta_1 = 0$ versus $H_a: \beta_1 > 0$.
(c) $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$.

28.25 Confidence intervals and tests for these data use the t distribution with degrees of freedom

(a) 3. (b) 50. (c) 53.

28.26 Orlando, a 175-pound man, plans to run 6.5 mph for 1 hour on a 2% incline. He can be 95% confident that he will burn between

(a) 871 and 1009 Cal.
(b) 929 and 950 Cal.
(c) 906 and 974 Cal.

Physical and hormonal factors contribute to bone weakness in aging women. A study examined the number of bone fractures among a random sample of 1469 elderly men and women. We set the response variable fracture to be equal to 1 for one or more fractures and 0 for no fracture in the previous three years. In this context, p is the probability of having a bone fracture in the past three years for an elderly person.[17] *Below is the Minitab output for the logistic regression of fracture as a function of gender (0 for men and 1 for women). Exercises 28.27 and 28.28 are based on this output.*

28.27 The estimated logistic regression equation is
(a) $L = -4.758 + 1.089gender$.
(b) $L = 1.089 + 0.471gender$.
(c) $p = -4.758 + 1.089gender$.

28.28 Which of the following statements is true?
(a) The odds of a fracture are not significantly different for men and for women.
(b) The odds of a fracture are significantly higher for women than for men, about 3 to 7.5 times with 95% confidence.
(c) The odds of a fracture are significantly higher for women than for men, about 1.2 to 7.5 times with 95% confidence.

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-4.75789	0.409997	-11.60	0.000			
gender	1.08895	0.471244	2.31	0.021	2.97	1.18	7.48

CHAPTER 28 EXERCISES

28.29 Children's feet. A mother was told by a shoe salesman that girls' shoes are narrower than boys' shoes because girls have narrower feet. Table 28.7 shows foot measurements (in cm) collected on 39 fourth-graders.[18] Run a simple t test to determine whether there is a significant difference in foot width between boys and girls in the fourth grade. Does your analysis support the salesman's assertion? Also use software to create a regression model of *width* using only *sex_B*. Because boys and girls have similar variability in this data set, you should find that both analyses give very *similar* P-values.

28.30 Children's feet, continued. Go back to the previous exercise. Feet are always longer than they are wide, and we can expect that foot width is somehow associated with foot length.
(a) Make a scatterplot of width (y) against length and describe the relationship. Use software to determine whether this relationship is statistically significant.
(b) Now use software to obtain a multiple linear regression model of *width* that includes both *length* and *sex_B* as explanatory variables. Create a scatterplot using different symbols for boys and girls. Discuss the results shown in the ANOVA table and individual t tests in the context of your scatterplot. Check the conditions for inference. Does this analysis support the salesman's assertion?
(c) Contrast the conclusions you reached here and in Exercise 28.29. The conclusions are somewhat different. Explain why. This illustrates why regression coefficients must be interpreted in the context of the specific model used.

28.31 Caterpillars. Scientists have long been interested in the question of how body mass determines characteristics such as metabolic rate. Tobacco hornworm caterpillars (*Manduca sexta*) were chosen to study this relationship because they maintain their overall shape throughout the five stages of larval development. For theoretical reasons (explored in Exercises 28.44 and 28.45), researchers were interested more specifically in the relationship between the log of metabolic rate $y = \log(MR)$ and the log of body mass $x_1 = \log(BM)$. The full data set for caterpillars at the fourth and fifth stages of development is available in the file *ex28-31.dat* available on the companion website.[19] What is the relationship between the log of metabolic rate and the log

TABLE 28.7 Foot width and length of fourth-graders (in cm)

Width	Length	Sex_B	Width	Length	Sex_B
8.4	22.4	1	9.2	22.0	1
8.8	23.4	1	8.6	22.4	1
9.7	22.5	1	8.3	22.0	0
9.8	23.2	1	9.0	22.5	0
8.9	23.1	1	8.1	22.2	0
9.7	23.7	1	9.4	25.1	1
9.6	24.1	1	9.5	24.1	0
8.8	21.0	0	9.5	23.5	1
9.3	21.6	0	8.9	22.2	1
8.8	20.9	1	9.3	21.9	1
9.8	25.5	1	9.3	22.0	0
8.9	22.8	1	8.6	20.5	0
9.1	24.1	1	8.6	22.5	0
9.8	25.0	1	9.0	21.6	1
9.3	24.0	0	8.6	22.7	1
7.9	21.7	0	8.5	20.9	0
8.7	22.0	0	9.0	24.0	0
8.8	22.7	0	7.9	19.6	0
9.0	24.7	0	8.8	22.6	0
9.5	23.5	0			

of body mass for tobacco hornworm caterpillars? Is the relationship the same for the two different stages? Use the four-step process as in Examples 28.1, 28.12, and 28.13. Make sure to check the conditions for inference and to include a statistical analysis.

28.32 Correlated explanatory variables. Suppose that $x_1 = 2x_2 - 4$, so that x_1 and x_2 are positively correlated. Let $y = 3x_2 + 4$, so that y and x_2 are positively correlated.

(a) Use the relationship between x_1 and x_2 to find the linear relationship between y and x_1. Are y and x_1 positively correlated?

(b) Add the equations $x_1 = 2x_2 - 4$ and $y = 3x_2 + 4$ and solve for y to obtain an equation relating y to both x_1 and x_2. Are the coefficients of both x's positive? Combining explanatory variables that are correlated can produce surprising results.

28.33 Body fat for men. You are interested in predicting the amount of body fat y of a man using the explanatory variables waist size x_1 and height x_2.

(a) Do you think body fat y and waist size x_1 are positively correlated? Explain.

(b) For a fixed waist size, height x_2 is negatively correlated with body fat y. Explain why.

(c) The slope of the simple linear regression line for predicting body fat from height for a sample of men is almost 0, say, 0.13. Knowing a man's height does not tell you much about his body fat. Do you think this parameter estimate would become negative if a multiple regression model with height x_2 and waist size x_1 was used to predict body fat? Explain.

28.34 Metabolic rate and body mass. Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below 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 kilocalories (Cal) 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.

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) Use the model with two regression lines to predict metabolic rate from lean body mass for the different genders. Summarize the results.

(c) The parameter associated with the interaction term is often used to decide if a model with parallel regression lines can be used. Test the hypothesis that this parameter is equal to zero, and comment on whether you would be willing to use the more restrictive model with parallel regression lines for these data.

28.35 World record running times. Table 28.8 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

(a) Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.

(b) Fit the model with two regression lines, one for women and one for men, and identify the estimated regression lines.

TABLE 28.8 World record running times for men and women

Men				Women	
Record year	Time (seconds)	Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1962	1698.2	1967	2286.4
1921	1840.2	1963	1695.6	1970	2130.5
1924	1835.4	1965	1659.3	1975	2100.4
1924	1823.2	1972	1658.4	1975	2041.4
1924	1806.2	1973	1650.8	1977	1995.1
1937	1805.6	1977	1650.5	1979	1972.5
1938	1802.0	1978	1642.4	1981	1950.8
1939	1792.6	1984	1633.8	1981	1937.2
1944	1775.4	1989	1628.2	1982	1895.2
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1993	1618.4	1983	1887.6
1949	1761.2	1994	1612.2	1984	1873.8
1950	1742.6	1995	1603.5	1985	1859.4
1953	1741.6	1996	1598.1	1986	1813.7
1954	1734.2	1997	1591.3	1993	1771.8
1956	1722.8	1997	1587.8		
1956	1710.4	1998	1582.7		
1960	1698.8	2004	1580.3		

(c) Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

28.36 World record running times, continued. The previous exercise shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

(a) Provide the ANOVA table for the regression model with two regression lines, one for men and one for women.

(b) Are all the individual coefficients significantly different from zero? Set up the appropriate hypotheses, identify the test statistics and P-values, and make conclusions in the context of the problem.

28.37 Mycorrhizal colonies and plant nutrition. Mycorrhizal fungi are present in the roots of many plants. This is a symbiotic relationship in which the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared the effects of adding various amounts of nitrogen fertilizer (in kilograms per hectare) to two genotypes of tomato plants, a wild-type variety with mycorrhizal colonies and a mutant variety without the colonies. The percent of phosphorus in the plant at harvest time was then assessed, and the results are shown in Table 28.9.[20]

(a) Produce a scatterplot of the amount of phosphorous in the plant against the amount of fertilizer provided (nitrogen), and use different symbols for the two plant genotypes. Does the graph suggest that a multiple linear regression might be appropriate? Explain.

(b) Use software to obtain the estimated multiple linear regression equation when the two explanatory variables nitrogen and genotype are included. Create a residual plot for this model. Are the conditions for multiple linear regression satisfied? Explain.

(c) Create a new variable called "Interaction" by multiplying the explanatory variables nitrogen and genotype. Add this new variable to your regression model. Provide the estimated multiple linear regression equation. Create a residual plot for this new model and discuss whether the conditions for multiple linear regression are met.

(d) Does the ANOVA table for the model with the interaction term indicate that at least one of the explanatory variables is helpful in predicting the amount of phosphorus in the plant? Do the individual t tests indicate that all coefficients are significantly different from zero? Explain.

TABLE 28.9 Percent of phosphorus in tomato plants

Phosphorus	Nitrogen	Genotype	Phosphorus	Nitrogen	Genotype
0.29	0	0	0.64	0	1
0.25	0	0	0.54	0	1
0.27	0	0	0.53	0	1
0.24	0	0	0.52	0	1
0.24	0	0	0.41	0	1
0.2	0	0	0.43	0	1
0.21	28	0	0.41	28	1
0.24	28	0	0.37	28	1
0.21	28	0	0.5	28	1
0.22	28	0	0.43	28	1
0.19	28	0	0.39	28	1
0.17	28	0	0.44	28	1
0.18	160	0	0.34	160	1
0.20	160	0	0.31	160	1
0.19	160	0	0.36	160	1
0.19	160	0	0.37	160	1
0.16	160	0	0.26	160	1
0.17	160	0	0.27	160	1

28.38 Mycorrhizal colonies and plant nutrition, continued. In Chapter 26, we ran a two-way ANOVA on this same data set. Go back to Example 26.12 and read it carefully. Explain why we can run either a two-way ANOVA or a multiple linear regression analysis on this data set. Compare your conclusions from Exercise 28.37 with those reached in Example 26.12. What are the advantages and disadvantages of either method?

28.39 Reaction time. A learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject clicks in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 28.10 gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.[21]

(a) Specify the population multiple regression model for predicting time from distance separately for each hand. Make sure you include the interaction term that is necessary to allow for the possibility of having different slopes. Explain in words what each β in your model means.

(b) Use statistical software to find the estimated multiple regression equation for predicting time from distance separately for each hand. What percent of variation in the distances is explained by this multiple regression model?

(c) Explain how to use the estimated multiple regression equation in part (b) to obtain the least-squares line for each hand. Draw these lines on a scatterplot of time versus distance.

28.40 Reaction time, continued. Go back to the previous exercise. Would you be willing to use your model to predict reaction times for individuals in the general population? Explain your answer. A statistical model represents only the population from which the sample was taken.

28.41 Burning calories with exercise. Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, Calories burned per hour (or total Calories), and duration of the workout. Let's take another look at the data in Table 28.6 that were used for the Check Your Skills

TABLE 28.10 Reaction times (ms) in a computer game

Time	Distance	Hand	Time	Distance	Hand
115	190.70	Right	240	190.70	Left
96	138.52	Right	190	138.52	Left
110	165.08	Right	170	165.08	Left
100	126.19	Right	125	126.19	Left
111	163.19	Right	315	163.19	Left
101	305.66	Right	240	305.66	Left
111	176.15	Right	141	176.15	Left
106	162.78	Right	210	162.78	Left
96	147.87	Right	200	147.87	Left
96	271.46	Right	401	271.46	Left
95	40.25	Right	320	40.25	Left
96	24.76	Right	113	24.76	Left
96	104.80	Right	176	104.80	Left
106	136.80	Right	211	136.80	Left
100	308.60	Right	238	308.60	Left
113	279.80	Right	316	279.80	Left
123	125.51	Right	176	125.51	Left
111	329.80	Right	173	329.80	Left
95	51.66	Right	210	51.66	Left
108	201.95	Right	170	201.95	Left

exercises. Scatterplots show different linear relationships for each incline, one for slow speeds and another for faster speeds, so the following indicator variables were created:

Ind_slow = 1 for $MPH \leq 3$ and *Ind_slow* = 0 for $MPH > 3.0$
NonIncline = 1 for 0% incline and *NonIncline* = 0 for other inclines
2%Incline = 1 for a 2% incline and *2%Incline* = 0 for other inclines

Figure 28.17 shows part of the Minitab output from fitting a multiple regression model to predict *Calories* from *MPH*, *Ind_slow*, *NoIncline*, and *2%Incline* for a 175-pound man using the Cybex machine.

(a) Use the Minitab output to estimate each parameter in this multiple regression model for predicting Calories burned with the Cybex machine. Don't forget to estimate σ.
(b) How many separate lines are fitted with this model? Do the lines all have the same slope? Identify each fitted line.
(c) Do you think that this model provides a good fit for these data? Explain.
(d) Is there significant evidence that more calories are burned for higher speeds? State the hypotheses, identify the test statistic and *P*-value, and provide a conclusion in the context of this question.

28.42 Children's perception of reading difficulty. Table 28.11 contains measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school.[22] The variables are

Variable	Description
OBS	Observation number for each individual
SEX	Gender of the individual
LSS	Median grade level of student's selection of "best for me to read" (8 repetitions, each with four choices at grades 3, 5, 7, and 9 level)
IQ	IQ score
READ	Score on reading subtest of the Metropolitan Achievement Test
EST	Student's own estimate of his or her reading ability, scale 1 to 5 (1 = low)

(a) Is the relationship between measured (*READ*) and self-estimated (*EST*) reading ability the same for both boys and girls? Create an indicator variable for gender and fit an appropriate multiple regression model to answer the question.
(b) Fit a multiple regression model for predicting *IQ* from the explanatory variables *LSS*, *READ*, and *EST*. Are you happy with the fit of this model? Explain.
(c) Use residual plots to check the appropriate conditions for your model.
(d) Only two of the three explanatory variables in your model in part (b) have parameters that are significantly

Minitab

Session

```
Regression Analysis: Calories versus MPH, Ind_slow, NoIncline, 2% Incline

Predictor      Coef   SE Coef       T       P
Constant      64.75     19.46    3.33   0.002
MPH         145.841     2.596   56.17   0.000
Ind_slow     -50.01     16.20   -3.09   0.003
NoIncline   -145.06     11.43  -12.69   0.000
2%Incline    -72.83     11.43   -6.37   0.000

S = 34.2865    R-Sq = 99.3%    R-Sq(adj) = 99.3%

Analysis of Variance

Source           Df       SS       MS        F       P
Regression        4  8554242  2138561  1819.18   0.000
Residual Error   49    57603     1176
Total            53  8611845
```

FIGURE 28.17 Minitab output for Exercise 28.42.

TABLE 28.11 Measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school

OBS	SEX	LSS	IQ	READ	EST	OBS	SEX	LSS	IQ	READ	EST
1	F	5.00	145	98	4	31	M	7.00	106	55	4
2	F	8.00	139	98	5	32	M	6.00	124	70	4
3	M	6.00	126	90	5	33	M	8.00	115	82	5
4	F	5.33	122	98	5	34	M	8.40	133	94	5
5	F	5.60	125	55	4	35	F	5.00	116	75	4
6	M	9.00	130	95	3	36	F	6.66	102	80	3
7	M	5.00	96	50	4	37	F	5.00	127	85	4
8	M	4.66	110	50	4	38	M	6.50	117	88	5
9	F	4.66	118	75	4	39	F	5.00	109	70	3
10	F	8.20	118	75	5	40	M	5.50	137	80	4
11	M	4.66	101	65	4	41	M	6.66	117	55	4
12	M	7.50	142	68	5	42	M	6.00	90	65	2
13	F	5.00	134	80	4	43	F	4.00	103	30	1
14	M	7.00	124	10	4	44	F	5.50	114	74	5
15	M	6.00	112	67	4	45	M	5.00	139	80	5
16	M	6.00	109	83	3	46	M	6.66	101	70	2
17	F	5.33	134	90	4	47	F	8.33	122	60	4
18	M	6.00	113	90	5	48	F	6.50	105	45	2
19	M	6.00	81	55	3	49	F	4.00	97	45	1
20	F	6.00	113	83	4	50	M	5.50	89	55	4
21	M	6.00	123	65	4	51	M	5.00	102	30	2
22	F	4.66	94	25	3	52	F	4.00	108	10	4
23	M	4.50	100	45	3	53	M	4.66	110	40	1
24	F	6.00	136	97	4	54	M	5.33	128	65	1
25	M	5.33	109	75	4	55	M	5.20	114	15	2
26	F	3.60	131	70	4	56	M	4.00	112	62	2
27	M	4.00	117	23	3	57	F	3.60	114	98	4
28	M	6.40	110	45	3	58	M	6.00	102	52	2
29	F	6.00	127	70	2	59	F	4.60	82	23	1
30	F	6.00	124	85	5	60	M	5.33	101	35	2

different from zero according to the individual t tests. Drop the explanatory variable that is not significant, and add the interaction term for the two remaining explanatory variables. Are you surprised by the results from fitting this new model? Explain what happened to the individual t tests for the two explanatory variables.

28.43 Fish sizes. In Exercise 28.13 you built a multiple regression model with three explanatory variables (length, width, and their interaction) to predict the weight of a perch. Minitab provides the following information for a perch with length 21 cm and width 2.8 cm (which creates an interaction term of $21 \times 2.8 = 58.8$):

```
      Predicted Values for New Observations
NewObs Fit   SE Fit 95% CI          95% PI
1      84.02 10.41  (63.13, 104.91) (-7.18, 175.21)
```

Interpret both intervals in the context of the data set.

28.44 More on caterpillars. Exercise 28.31 described the relationship between basal metabolic rate (MR) and body mass

(*BM*) in tobacco hornworm caterpillars, or, more specifically, the relationship between response variable $y = \log(MR)$ and explanatory variable $x_1 = \log(BM)$. Both experimental and theoretical research have suggested the general relationship $MR = \alpha(BM)^{\beta}$.[23] However, there is still considerable debate on whether the scaling exponent is $\beta = 2/3$ or $\beta = 3/4$.

(a) Use software to estimate α and β in the general relationship $MR - \alpha(BM)^{\beta}$, which is the same as $\mu_{\log(MR)} = \log(\alpha) + \beta \log(BM)$. The predicted model is $\hat{y} = a + b \log(BM)$, so that b_0 estimates $\log(\alpha)$ and b_1 estimates β in the original model. What percent of variation in $\log(MR)$ is explained by using linear regression with the explanatory variable $\log(MR)$?

(b) Find a 95% confidence interval for the slope parameter β. Are the values $\beta = 2/3$ and $\beta = 3/4$ contained in your confidence interval?

28.45 More on caterpillars, continued. Go back to the previous exercise. This time, use the multiple regression model with two explanatory variables: the continuous variable $x_1 = \log(BM)$ and an indicator variable reflecting instar stage ($x_2 = 1$ for instar 5, $x_2 = 0$ for instar 4). This should reflect what you did in Exercise 28.31.

(a) Find a 95% confidence interval for the slope parameter β for caterpillars during stage 4.

(b) If you were asked to report a confidence interval for the slope parameter β for caterpillars during stage 5, would you report the same interval that you calculated in part (a)? Explain why or why not.

(c) Are the values $\beta = 2/3$ and $\beta = 3/4$ contained in your confidence interval from part (a)?

(d) How does your confidence interval in part (a) compare with the confidence interval you computed in part (b) of Exercise 28.44?

28.46 Recurring bladder tumors, continued. Go back to Exercises 28.15 and 28.16. You already computed the estimated value of p when x is 0, 2, or 6. Complete your analysis for the other values of x (1, 3, 4, 5, 7, 8). Use your results to create a plot that shows how p changes as a function of x. Interpret this plot.

28.47 FDA warning on insomnia prescriptions. In a January 2013 safety announcement, the Food and Drug Administration warned drug manufacturers that the recommended dose of the widely prescribed insomnia drug zolpidem should be halved for women. The warning came after pharmacokinetic trials found that 38 of 250 women and 8 of 250 men who had taken the recommended dose of zolpidem still had enough drug in their blood 8 hours later to cause impairment.

(a) What are the odds of impairment 8 hours after dosing for women and for men? Use this information to compute the equation of the logistic regression model if the indicator variable x equals 0 for men and 1 for women.

(b) Minitab provides the standard error of the slope, $SE_{b1} = 0.4002$. Use this information to compute a 95% confidence interval for the slope of the logistic regression model. Then compute a 95% confidence interval for the odds ratio for impairment 8 hours after dosing. Interpret your results in the context of the FDA warning.

28.48 Treating AIDS. The drug AZT was the first drug that seemed effective in delaying the onset of AIDS in people infected with HIV. A study randomly assigned 435 HIV-positive volunteers who had not yet developed AIDS to take 500 mg of AZT each day ($AZT = 1$) and another 435 to take a placebo ($AZT = 0$). At the end of the study, 38 of the placebo subjects and 17 of the AZT subjects had developed AIDS.

(a) What are the odds of developing AIDS in each of the two groups? Use this information to compute the equation of the logistic regression model using the indicator variable *AZT*.

(b) Minitab provides the standard error of the slope, $SE_{b1} = 0.3001$. Use this information to compute a 95% confidence interval for the slope of the logistic regression model. Then compute a 95% confidence interval for the odds ratio for developing AIDS. Do the data support the claim that taking AZT lowers the odds of developing AIDS in HIV-positive patients? Explain.

A study screened a random sample of adult subjects for adult-onset, type 2 diabetes with the objective of identifying some of the risk factors associated with the disease. The file ex28-49.dat on the companion website contains data about diabetes status, weight (in pounds), waist circumference (in inches), and cholesterol ratio (ratio of total cholesterol to HDL cholesterol in blood) for all 386 subjects.[24] *Fifty-nine of the 386 were identified as having type 2 diabetes (diabetes = 1). Exercises 28.49 to 28.51 are based on this study.*

28.49 Type 2 diabetes. Adult-onset, or type 2, diabetes has been associated with obesity. The Minitab output in Figure 28.18 shows the results of a logistic regression analysis with response variable *diabetes* and explanatory variable *weight*. Give the equation of the logistic regression model. Is the coefficient for the slope significantly different from zero? Does a higher weight appear to increase the odds of having type 2 diabetes?

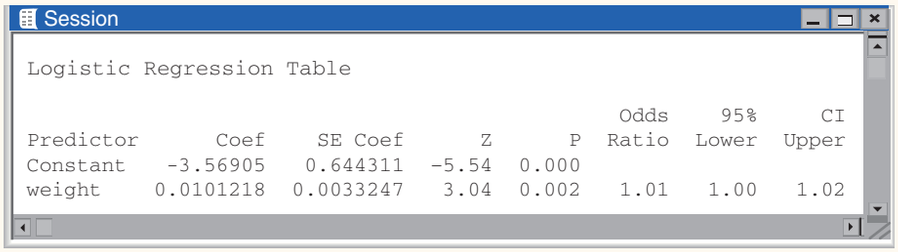


FIGURE 28.18 Minitab output for Exercises 28.49 to 28.51.

28.50 Type 2 diabetes, continued. Use software to obtain the logistic regression analysis with response variable *diabetes* and explanatory variable *weight*. Make sure that you obtain the same results as those shown in the previous exercise, up to rounding error.

(a) Research suggests that where body fat is stored is an important factor in predicting diabetes. Use software to run a model with both *weight* and *waist* as explanatory variables. What is the equation of this new logistic regression model? Do both variables contribute significantly to the model? Explain how adding the variable *waist* could change the relative contribution of the variable *weight*.

(b) Use software to fit a logistic regression model to predict *diabetes* from *waist*. Interpret the output in the context of the study. Which of the three models described in this exercise would you favor and why?

28.51 Type 2 diabetes, continued. Among the list of potential risk factors for type 2 diabetes is a person's cholesterol ratio (*cholratio*), the ratio of total cholesterol to good, or HDL, cholesterol. Use software to fit a logistic regression model to predict *diabetes* from both *waist* and *cholratio* as explanatory variables. What is the equation of this new logistic regression model? Are both coefficients significant? Explain in your own words what the two corresponding odds ratios mean in the context of this study.

28.52 Sleep apnea. Sleep apnea is a potentially dangerous condition in which patients cease to breathe for at least 10 seconds at a time while asleep. Researchers collected data on 133 individuals in order to find risk factors for sleep apnea.

(a) Neck size (measured in centimeters) may predict the risk of sleep apnea because a larger neck size might constrict the airway. Below is the software output for the logistic regression model explaining sleep apnea ($y = 1$ for sleep apnea, 0 otherwise) as a function of neck size. Is neck size significant in predicting sleep apnea? Explain.

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	CI Upper
Constant	-8.15026	1.90008	-4.29	0.000			
Neck Size	0.227170	0.0511498	4.44	0.000	1.26	1.14	1.39

(b) Being overweight has been linked to sleep apnea. Below is the software output for the logistic regression model explaining sleep apnea as a function of an individual's body mass index (BMI, measured in kg/m^2). Is BMI significant in predicting sleep apnea? Explain.

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	CI Upper
Constant	-8.37612	1.66623	-5.03	0.000			
BMI	0.322748	0.0627380	5.14	0.000	1.38	1.22	1.56

(c) Based on the results obtained in (a) and (b), we also ran a logistic regression model using both neck size and BMI as explanatory variables. The software output is given below. Are both variables still significant when the other is taken into consideration? What conclusions do you reach?

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	CI Upper
Constant	-9.31099	2.28744	-4.07	0.000			
BMI	0.298216	0.0729842	4.09	0.000	1.35	1.17	1.55
Neck Size	0.0428161	0.0691903	0.62	0.536	1.04	0.91	1.20

28.53 Student achievement and self-concept. In order to LARGE DATA SET determine if student achievement is related to self-concept, as measured by the Piers-Harris Children's Self-Concept Scale, data were collected on 78 seventh-grade students from a rural midwestern school. The large data file *Large.SelfConcept* available on the companion website contains data on the following variables for all 78 students:[25]

Variable	Description
OBS	Observation number ($n = 78$, some gaps in numbers)
GPA	GPA from school records
IQ	Test score from school records
AGE	Age in years, self-reported
GENDER	1 = F, 2 = M, self-reported
RAW	Raw score on Piers-Harris Childrens' Self-Concept Scale
C1	Cluster 1 within self-concept: behavior
C2	Cluster 2: school status
C3	Cluster 3: physical
C4	Cluster 4: anxiety
C5	Cluster 5: popularity
C6	Cluster 6: happiness

We will investigate the relationship between *GPA* and only three of the explanatory variables:

- *IQ*, the student's score on a standard IQ test
- C2, the student's self-assessment of his or her school status
- C5, the student's self-assessment of his or her popularity

Use statistical software to analyze the relationship between students' *GPA* and their *IQ*, self-assessed school status (C2), and self-assessed popularity (C5).

(a) One observation is an extreme outlier when all three explanatory variables are used. Which observation is this? Give the observation number and explain how you found it using regression output. Find this observation in the data list. What is unusual about it?

(b) Software packages often identify unusual or influential observations. Have any observations been identified as unusual or influential? If so, identify these points on a scatterplot of *GPA* versus *IQ*.

(c) C2 (school status) is the aspect of self-concept most highly correlated to *GPA*. If we carried out the simple linear regression of *GPA* on C2, what percent of the variation in students' GPAs would be explained by the straight-line relationship between *GPA* and C2?

(d) You know that *IQ* is associated with *GPA*, and you are not studying that relationship. Because C2 and *IQ* are positively correlated ($r = 0.547$), a significant relationship between C2 and *GPA* might occur just because C2 can "stand in" for *IQ*. Does C2 still contribute significantly to explaining *GPA* after we have allowed for the relationship between *GPA* and *IQ*? (Give a test statistic, its P-value, and your conclusion.)

(e) A new student in this class has *IQ* 115 and C2 score 14. What do you predict this student's *GPA* to be? (Just give a point prediction, not an interval.)

28.54 Growth of pine trees. LARGE DATA SET The Department of Biology at Kenyon College conducted an experiment to study the growth of pine trees at a site located just south of Gambier on a hill overlooking the Kokosing River. In April 1990, student and faculty volunteers planted 1000 white pine (*Pinus strobus*) seedlings at the Brown Family Environmental Center. These seedlings were planted in two grids, distinguished by 10- and 15-foot spacings between the seedlings. A subset of the data collected by students at Kenyon College is available in the *Large.PineTrees* file on the companion website.[26] A description of the variables is provided below.

Variable	Description
Row	Row number in pine plantation
Col	Column number in pine plantation
Hgt90	Tree height at the time of planting (cm)
Hgt96	Tree height in September 1996 (cm)
Diam96	Tree trunk diameter in September 1996 (cm)
Grow96	Leader growth during 1996 (cm)
Hgt97	Tree height in September 1997 (cm)
Diam97	Tree trunk diameter in September 1997 (cm)
Spread97	Widest lateral spread in September 1997 (cm)
Needles97	Needle length in September 1997 (mm)
Deer95	Type of deer damage in September 1995: 1 = none, 2 = browsed
Deer97	Type of deer damage in September 1997: 1 = none, 2 = browsed
Cover95	Amount of thorny cover in September 1995: 0 = none, 1 = <1/3, 2 = between 1/3 and 2/3, 3 = >2/3
Fert	Indicator for fertilizer: 0 = no, 1 = yes
Spacing	Distance (in feet) between trees (10 or 15)

(a) Use tree height at the time of planting (*Hgt90*) and the indicator variable for fertilizer (*Fert*) to fit a multiple regression model for predicting *Hgt97*. Specify the estimated regression model and the regression standard error. Are you happy with the fit of this model? Comment on the value of R^2 and the plot of the residuals against the predicted values.

(b) Construct a correlation matrix with *Hgt90*, *Hgt96*, *Diam96*, *Grow96*, *Hgt97*, *Diam97*, *Spread97*, and *Needles97*. Which variable is most strongly correlated with the response variable of interest (*Hgt97*)? Does this make sense to you?

(c) Add tree height in September 1996 (*Hgt96*) to the model in part (a). Does this model do a better job of predicting tree height in 1997? Explain.

(d) What happened to the individual t statistic for *Hgt90* when *Hgt96* was added to the model? Explain why this change occurred.

(e) Fit a multiple regression model for predicting *Hgt97* based on the explanatory variables *Diam97*, *Hgt96*, and *Fert*. Summarize the results of the individual t tests. Does this model provide a better fit than the previous models? Explain by comparing the values of R^2 and s for each model.

(f) Does the parameter estimate for the variable indicating whether a tree was fertilized or not have the sign you expected? Explain. (Experiments can produce surprising results!)

(g) Do you think that the model in part (e) should be used for predicting growth in other pine seedlings? Think carefully about the conditions for inference.

NOTES AND DATA SOURCES

1. Data were estimated from a scatterplot in L. Partridge and M. Farquhar, "Sexual activity reduces lifespan of male fruitflies," *Nature*, 294 (1981), pp. 580–582.
2. Data were estimated from a scatterplot in T. J. Cleophas, "The sense and nonsense of regression modeling for increasing precision of clinical trials," *Clinical Pharmacology and Therapeutics*, 74 (2003), pp. 295–297.
3. Data were estimated from a scatterplot in P. Heeb, M. Kolliker, and H. Richner, "Bird-ectoparasite interactions, nest humidity, and ectoparasite community structure," *Ecology*, 81 (2000), pp. 958–968.
4. F. H. Simmons, "Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation," MS thesis, University of California at Irvine, 1996.
5. Data courtesy of Brad Hughes, Department of Ecology and Evolutionary Biology, University of California, Irvine.
6. M. Di Monacoa et al., "Biochemical markers of nutrition and bone mineral density in the elderly," *Gerontology*, 49 (2003), pp. 50–54.
7. S. J. Husson et al., "Optogenetic analysis of a nociceptor neuron and network reveals ion channels acting downstream of primary sensors," *Current Biology*, 22 (2012), pp. 743–752, doi:10.1016/j.cub.2012.02.066.
8. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by M. J. Kahn of Wheaton College.
9. D. P. Casey et al., "Relationship between muscle sympathetic nerve activity and aortic wave reflection characteristics in young men and women," *Hypertension*, 57 (2011), pp. 421–427.
10. R. Margaria et al., "Energy cost of running," *Journal of Applied Physiology*, 18 (1963), pp. 367–370.
11. The data in Table 28.4 are part of a larger data set in the *Journal of Statistics Education* online archive. The original source is P. Brofeldt, "Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi," in T. H. Jaervi, *Finlands fiskeriet*, vol. 4, *Meddelanden utgivna av fiskerifoereningen i Finland*, Helsinki, 1917. The data were contributed to the archive (with information in English) by J. Puranen of the University of Helsinki.
12. These data come from the Data and Story Library online at `lib.stat.cmu.edu/DASL`. They were contributed by D. S. Moore and G. P. McCabe.
13. A. Spanos, F. E. Harrell, and D. T. Durack, "Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations," *Journal of the American Medical Association*, 262 (1989), pp. 2700–2707.
14. These data come from the Data and Story Library online at `lib.stat.cmu.edu/DASL` and were first published in L. J. Wei, D. Y. Lin,

and L. Weissfeld, "Tumor recurrence data for patients with bladder cancer," *Journal of the American Statistical Association,* 84 (1989), pp. 1065–1073.

15. The test statistic for logistic regression is z, not t (unlike for simple and multiple linear regression). Interestingly, when z follows the standard Normal distribution, the value z^2 actually follows a chi-square distribution with df = 1. For this reason, inference for logistic regression sometimes uses the chi-square statistic (such that $X^2 = z^2$) and the chi-square distribution with df = 1. Both methods give the same P-value.

16. E. de Jonge et al., "Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial," *Lancet,* 362 (2003), pp. 1011–1016.

17. A. M. Tromp et al., "Predictors for falls and fractures in the longitudinal aging study, Amsterdam," *Journal of Bone and Mineral Research,* 13 (1998), pp. 1932–1939.

18. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by M. C. Meyer from the University of Georgia.

19. We thank Haruhiko Itagaki and his students Andrew Vreede and Marissa Stearns for providing data on tobacco hornworm caterpillars (*Manduca sexta*).

20. Data courtesy of David LeBauer, University of California at Irvine.

21. P. Velleman, *ActivStats 2.0,* Addison-Wesley Interactive, 1997.

22. J. T. Fleming, "The measurement of children's perception of difficulty in reading materials," *Research in the Teaching of English*, 1 (1967), pp. 136–156.

23. For more details, see H. Hoppeler and E. Weibel, "Scaling functions to body size: theories and facts," *Journal of Experimental Biology*, 208 (2005), pp. 1573–1574.

24. J. B. Schorling et al., "A trial of church-based smoking cessation interventions for rural African Americans," *Preventive Medicine,* 26 (1997), pp. 92–101. Data posted on the website of the Department of Biostatistics at Vanderbilt University.

25. D. Gordon, "The relationships among academic self-concept, academic achievement, and persistence with academic self-attribution, study habits, and perceived school environment," PhD thesis, Purdue University, 1997.

26. We thank Ray and Pat Heithaus for providing data on the pine seedlings at the Brown Family Environmental Center.