CHAPTER 29

Multiple Regression

WHO	250 Male subjects
WHAT	Body fat and waist size
UNITS	%Body fat and inches
WHEN	1990s
WHERE	United States
WHY	Scientific research

In Chapter 27 we tried to predict the percent body fat of male subjects from their waist size, and we did pretty well. The R^2 of 67.8% says that we accounted for almost 68% of the variability in *%body fat* by knowing only the *waist* size. We completed the analysis by performing hypothesis tests on the coefficients and looking at the residuals.

But that remaining 32% of the variance has been bugging us. Couldn't we do a better job of accounting for *%body fat* if we weren't limited to a single predictor? In the full data set there were 15 other measurements on the 250 men. We might be able to use other predictor variables to help us account for that leftover variation that wasn't accounted for by waist size.

What about *height*? Does *height* help to predict *%body fat*? Men with the same *waist* size can vary from short and corpulent to tall and emaciated. Knowing a man has a 50-inch waist tells us that he's likely to carry a lot of body fat. If we found out that he was 7 feet tall, that might change our impression of his body type. Knowing his *height* as well as his *waist* size might help us to make a more accurate prediction.

Just Do It

Does a regression with *two* predictors even make sense? It does—and that's fortunate because the world is too complex a place for simple linear regression alone to model it. A regression with two or more predictor variables is called a **multiple regression.** (When we need to note the difference, a regression on a single predictor is called a *simple* regression.) We'd never try to find a regression by hand, and even calculators aren't really up to the task. This is a job for a statistics program on a computer. If you know how to find the regression of *%body fat* on *waist* size with a statistics package, you can usually just add *height* to the list of predictors without having to think hard about how to do it.

A Note on Terminology

When we have two or more predictors and fit a linear model by least squares, we are formally said to fit a least squares linear multiple regression. Most folks just call it "multiple regression." You may also see the abbreviation OLS used with this kind of analysis. It stands for "Ordinary Least Squares."

For simple regression we found the **Least Squares** solution, the one whose coefficients made the sum of the squared residuals as small as possible. For multiple regression, we'll do the same thing but this time with more coefficients. Remarkably enough, we can still solve this problem. Even better, a statistics package can find the coefficients of the least squares model easily.

Here's a typical example of a multiple regression table:

```
Dependent variable is: Pct BF
R-squared = 71.3%    R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom

Variable     Coefficient    SE(Coeff)    t-ratio    P-value
Intercept    −3.10088       7.686        −0.403      0.6870
Waist         1.77309       0.0716       24.8       ≤0.0001
Height       −0.60154       0.1099       −5.47      ≤0.0001
```

You should recognize most of the numbers in this table. Most of them mean what you expect them to.

Metalware Prices. Multiple regression is a valuable tool for businesses. Here's the story of one company's analysis of its manufacturing process.

R^2 gives the fraction of the variability of *%body fat* accounted for by the *multiple* regression model. (With *waist* alone predicting *%body fat*, the R^2 was 67.8%.) The multiple regression model accounts for 71.3% of the variability in *%body fat*. We shouldn't be surprised that R^2 has gone up. It was the hope of accounting for some of that leftover variability that led us to try a second predictor.

The standard deviation of the residuals is still denoted s (or sometimes s_e to distinguish it from the standard deviation of y).

The degrees of freedom calculation follows our rule of thumb: the degrees of freedom is the number of observations (250) minus one for each coefficient estimated—for this model, 3.

Compute a Multiple Regression. We always find multiple regressions with a computer. Here's a chance to try it with the statistics package you've been using.

For each predictor we have a coefficient, its standard error, a t-ratio, and the corresponding P-value. As with simple regression, the t-ratio measures how many standard errors the coefficient is away from 0. So, using a Student's t-model, we can use its P-value to test the null hypothesis that the true value of the coefficient is 0.

Using the coefficients from this table, we can write the regression model:

$$\widehat{\%\textit{body fat}} = -3.10 + 1.77\,\textit{waist} - 0.60\,\textit{height}.$$

As before, we define the residuals as

$$\textit{residuals} = \%\textit{body fat} - \widehat{\%\textit{body fat}}.$$

We've fit this model with the same least squares principle: The sum of the squared residuals is as small as possible for any choice of coefficients.

So, What's New?

So what's different? With so much of the multiple regression looking just like simple regression, why devote an entire chapter (or two) to the subject?

There are several answers to this question. First—and most important—the *meaning* of the coefficients in the regression model has changed in a subtle but important way. Because that change is not obvious, multiple regression coefficients

A S **Reading the Multiple Regression Table.** You may be surprised to find that you already know how to interpret most of the values in the table. Here's a narrated review.

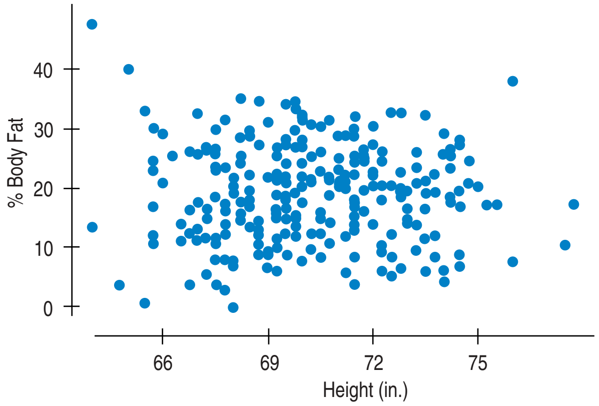
are often misinterpreted. We'll show some examples to help make the meaning clear.

Second, multiple regression is an extraordinarily versatile calculation, underlying many widely used Statistics methods. A sound understanding of the multiple regression model will help you to understand these other applications.

Third, multiple regression offers our first glimpse into statistical models that use more than two quantitative variables. The real world is complex. Simple models of the kind we've seen so far are a great start, but often they're just not detailed enough to be useful for understanding, predicting, and decision making. Models that use several variables can be a big step toward realistic and useful modeling of complex phenomena and relationships.

What Multiple Regression Coefficients Mean

We said that height might be important in predicting body fat in men. What's the relationship between *%body fat* and *height* in men? We know how to approach this question; we follow the three rules. Here's the scatterplot:

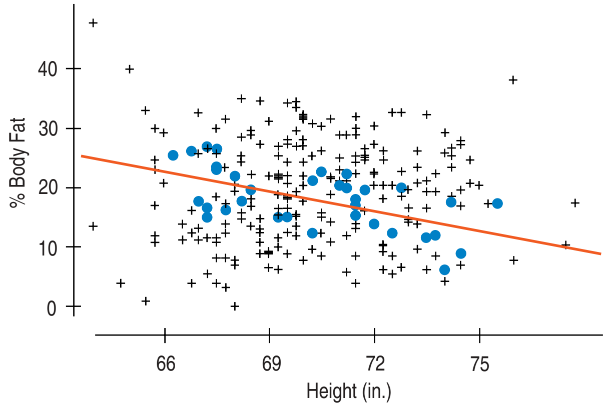


The scatterplot of *%body fat* against *height* seems to say that there is little relationship between these variables. **Figure 29.1**

It doesn't look like *height* tells us much about *%body fat*. You just can't tell much about a man's *%body fat* from his *height*. Or can you? Remember, in the multiple regression model, the coefficient of *height* was −0.60, had a *t*-ratio of −5.47, and had a very small P-value. So it *did* contribute to the *multiple* regression model. How could that be?

The answer is that the multiple regression coefficient of *height* takes account of the other predictor, *waist size*, in the regression model.

To understand the difference, let's think about all men whose waist size is about 37 inches—right in the middle of our sample. If we think only about *these* men, what do we expect the relationship between *height* and *%body fat* to be? Now a negative association makes sense because taller men probably have less body fat than shorter men *who have the same waist size*. Let's look at the plot:



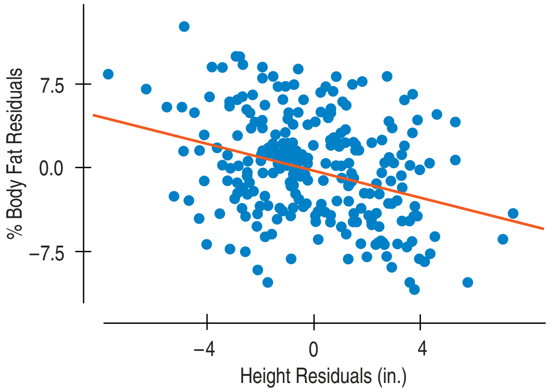
When we restrict our attention to men with waist sizes between 36 and 38 inches (points in blue), we can see a relationship between *%body fat* and *height*. **Figure 29.2**

Here we've highlighted the men with waist sizes between 36 and 38 inches. Overall, there's little relationship between *%body fat* and *height*, as we can see from the full set of points. But when we focus on *particular* waist sizes, there *is* a relationship between body fat and height. This relationship is *conditional* because we've restricted our set to only those men within a certain range of waist sizes. For men with that waist size, an extra inch of height is associated with a decrease of about 0.60% in body fat. If that relationship is consistent for each *waist* size, then the multiple regression coefficient will estimate it. The simple regression coefficient simply couldn't see it.

We've picked one particular *waist* size to highlight. How could we look at the relationship between *%body fat* and *height* conditioned on *all waist* sizes *at the same time*? Once again, residuals come to the rescue.

As their name reminds us, residuals are what's left over after we fit a model. That lets us remove the effects of some variables. The residuals are what's left.

We plot the residuals of *%body fat* after a regression on *waist size* against the residuals of *height* after regressing *it* on *waist* size. This display is called a *partial regression plot*. It shows us just what we asked for: the relationship of *%body fat* to *height* after removing the linear effects of *waist* size.



A partial regression plot for the coefficient of *height* in the regression model has a slope equal to the coefficient value in the multiple regression model. **Figure 29.3**

A **partial regression plot** for a particular predictor has a slope that is the same as the *multiple* regression coefficient for that predictor. Here, it's −0.60. It also has the same residuals as the full multiple regression, so you can spot any outliers or influential points and tell whether they've affected the estimation of this particular coefficient.

Many modern statistics packages offer partial regression plots as an option for any coefficient of a multiple regression. For the same reasons that we always look at a scatterplot before interpreting a simple regression coefficient, it's a good idea to make a partial regression plot for any multiple regression coefficient that you hope to understand or interpret.

The Multiple Regression Model

We can write a multiple regression model like this, numbering the predictors arbitrarily (we don't care which one is x_1), writing β's for the model coefficients (which we will estimate from the data), and including the errors in the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Of course, the multiple regression model is not limited to two predictor variables, and regression model equations are often written to indicate summing any number (a typical letter to use is k) of predictors. That doesn't really change anything, so we'll often stick with the two-predictor version just for simplicity. But don't forget that we can have many predictors.

The assumptions and conditions for the multiple regression model sound nearly the same as for simple regression, but with more variables in the model, we'll have to make a few changes.

Assumptions and Conditions

Linearity Assumption

We are fitting a linear model.[1] For that to be the right kind of model, we need an underlying linear relationship. But now we're thinking about several predictors. To see whether the assumption is reasonable, we'll check the Straight Enough Condition for *each* of the predictors.

Multiple Regression Assumptions. The assumptions and conditions we check for multiple regression are much like those we checked for simple regression. Here's an animated discussion of the assumptions and conditions for multiple regression.

Straight Enough Condition: Scatterplots of y against each of the predictors are reasonably straight. As we have seen with *height* in the body fat example, the scatterplots need not show a strong (or any!) slope; we just check that there isn't a bend or other nonlinearity. For the *%body fat* data, the scatterplot is beautifully linear in *waist* as we saw in Chapter 27. For *height*, we saw no relationship at all, but at least there was no bend.

As we did in simple regression, it's a good idea to check the residuals for linearity after we fit the model. It's good practice to plot the residuals against the

[1] By *linear* we mean that each x appears simply multiplied by its coefficient and added to the model. No x appears in an exponent or some other more complicated function. That means that as we move along any x-variable, our prediction for y will change at a constant rate (given by the coefficient) if nothing else changes.

predicted values and check for patterns, especially for bends or other nonlinearities. (We'll watch for other things in this plot as well.)

If we're willing to assume that the multiple regression model is reasonable, we can fit the regression model by least squares. But we must check the other assumptions and conditions before we can interpret the model or test any hypotheses.

Check the Residual Plot (Part 1)

The residuals should appear to have no pattern with respect to the predicted values.

Independence Assumption

As with simple regression, the errors in the true underlying regression model must be independent of each other. As usual, there's no way to be sure that the Independence Assumption is true. Fortunately, even though there can be many predictor variables, there is only one response variable and only one set of errors. The Independence Assumption concerns the errors, so we check the corresponding conditions on the residuals.

Check the Residual Plot (Part 2)

The residuals should appear to be randomly scattered and show no patterns or clumps when plotted against the predicted values.

Randomization Condition: The data should arise from a random sample or randomized experiment. Randomization assures us that the data are representative of some identifiable population. If you can't identify the population, you can't interpret the regression model or any hypothesis tests because they are about a regression model for that population. Regression methods are often applied to data that were not collected with randomization. Regression models fit to such data may still do a good job of modeling the data at hand, but without some reason to believe that the data are representative of a particular population, you should be reluctant to believe that the model generalizes to other situations.

We also check displays of the regression residuals for evidence of patterns, trends, or clumping, any of which would suggest a failure of independence. In the special case when one of the x-variables is related to time, be sure that the residuals do not have a pattern when plotted against that variable.

The *%body fat* data were collected on a sample of men. The men were not related in any way, so we can be pretty sure that their measurements are independent.

Equal Variance Assumption

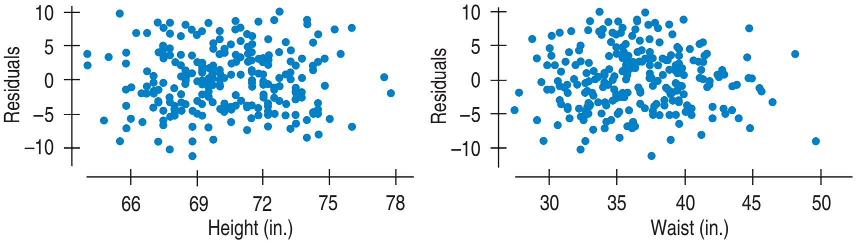
The variability of the errors should be about the same for all values of *each* predictor. To see if this is reasonable, we look at scatterplots.

Check the Residual Plot (Part 3)

The spread of the residuals should be uniform when plotted against any of the *x*'s or against the predicted values.

Does the Plot Thicken? Condition: Scatterplots of the regression residuals against each x or against the predicted values, $\hat{y}$, offer a visual check. The spread around the line should be nearly constant. Be alert for a "fan" shape or other tendency for the variability to grow or shrink in one part of the scatterplot.

Here are the residuals plotted against *waist* and *height*. Neither plot shows patterns that might indicate a problem.

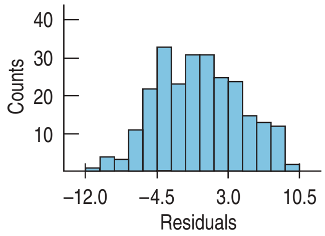


Residuals plotted against each predictor show no pattern. That's a good indication that the Straight Enough Condition and the "Does the Plot Thicken?" Condition are satisfied.
Figure 29.4

If residual plots show no pattern, if the data are plausibly independent, and if the plots don't thicken, we can feel good about interpreting the regression model. Before we test hypotheses, however, we must check one final assumption.

Normality Assumption

We assume that the errors around the idealized regression model at any specified values of the x-variables follow a Normal model. We need this assumption so that we can use a Student's t-model for inference. As with other times when we've used Student's t, we'll settle for the residuals satisfying the Nearly Normal Condition.



Check a histogram of the residuals. The distribution of the residuals should be unimodal and symmetric. Or check a Normal probability plot to see whether it is straight.

Figure 29.5

Nearly Normal Condition: Because we have only one set of residuals, this is the same set of conditions we had for simple regression. Look at a histogram or Normal probability plot of the residuals. The histogram of residuals in the *%body fat* regression certainly looks nearly Normal, and the Normal probability plot is fairly straight. And, as we have said before, the Normality Assumption becomes less important as the sample size grows.

Let's summarize all the checks of conditions that we've made and the order that we've made them:

1. Check the Straight Enough Condition with scatterplots of the y-variable against each x-variable.
2. If the scatterplots are straight enough (that is, if it looks like the regression model is plausible), fit a multiple regression model to the data. (Otherwise, either stop or consider re-expressing an x- or the y-variable.)
3. Find the residuals and predicted values.
4. Make a scatterplot of the residuals against the predicted values.[2] This plot should look patternless. Check in particular for any bend (which would suggest that the data weren't all that straight after all) and for any thickening. If there's a bend and especially if the plot thickens, consider re-expressing the y-variable and starting over.
5. Think about how the data were collected. Was suitable randomization used? Are the data representative of some identifiable population? If the data are measured over time, check for evidence of patterns that might suggest they're not independent by plotting the residuals against time to look for patterns.
6. If the conditions check out this far, feel free to interpret the regression model and use it for prediction. If you want to investigate a particular coefficient, make a partial regression plot for that coefficient.
7. If you wish to test hypotheses about the coefficients or about the overall regression, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.

A S **Partial Regression Plots vs. Scatterplots.** When should you use a partial regression plot? And why? This activity shows you.

[2] In Chapter 27 we noted that a scatterplot of residuals against the predicted values looked just like the plot of residuals against x. But for a multiple regression, there are several x's. Now the predicted values, $\hat{y}$, are a combination of the x's—in fact, they're the combination given by the regression equation we have computed. So they combine the effects of all the x's in a way that makes sense for our particular regression model. That makes them a good choice to plot against.

Multiple Regression Step-By-Step

Let's try finding and interpreting a multiple regression model for the body fat data.

Think

Plan Name the variables, report the W's, and specify the questions of interest.

I have body measurements on 250 adult males from the BYU Human Performance Research Center. I want to understand the relationship between % body fat, height, and waist size.

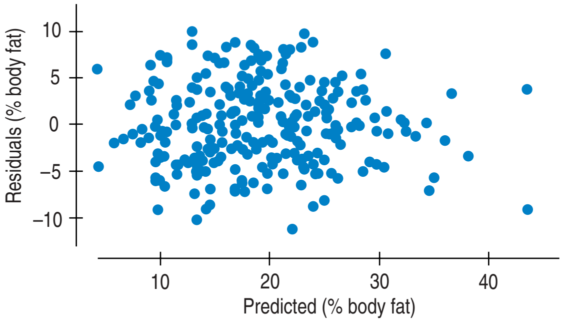
Model Check the appropriate conditions.

✔ **Straight Enough Condition:** There is no obvious bend in the scatterplots of %body fat against either x-variable. The scatterplot of residuals against predicted values below shows no patterns that would suggest nonlinearity.

✔ **Independence Assumption:** These data are not collected over time, and there's no reason to think that the %body fat of one man influences that of another. I don't know whether the men measured were sampled randomly, but the data are presented as being representative of the male population of the United States.

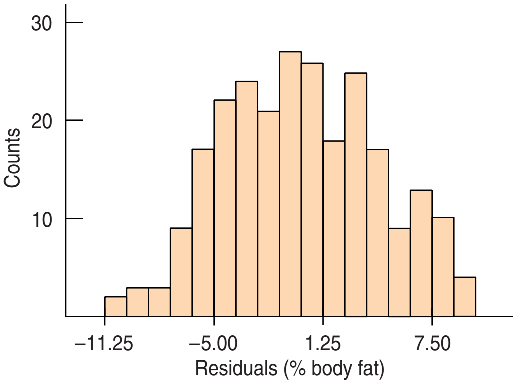
Now you can find the regression and examine the residuals.

✔ **Does the Plot Thicken? Condition:** The scatterplot of residuals against predicted values shows no obvious changes in the spread about the line.

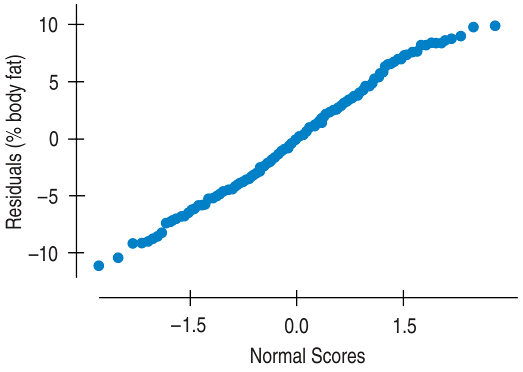


Actually, you need the Nearly Normal Condition only if we want to do inference.

✔ **Nearly Normal Condition:** A histogram of the residuals is unimodal and symmetric.



The Normal probability plot of the residuals is reasonably straight:



Choose your method.

Under these conditions a full multiple regression analysis is appropriate.

Show

Mechanics

Here is the computer output for the regression:

```
Dependent variable is: %BF
R-squared = 71.3%   R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom

                  Sum of          Mean
Source            Squares   DF    Square     F-ratio   P-value
Regression        12216.6     2   6108.28    307       <0.0001
Residual           4912.26  247   19.8877

Variable     Coefficient    SE(Coeff)   t-ratio   P-value
Intercept    −3.10088       7.686       −0.403    0.6870
Waist         1.77309       0.0716      24.8      <0.0001
Height       −0.60154       0.1099      −5.47     <0.0001
```

The estimated regression equation is

$$\widehat{\%\text{body fat}} = -3.10 + 1.77\,\textit{waist} - 0.60\,\textit{height}.$$

Tell

Conclusion Interpret the regression in the proper context.

The R^2 for the regression is 71.3%. *Waist* size and *height* together account for about 71% of the variation in *%body fat* among men. The regression equation indicates that each inch in *waist* size is associated with about a 1.77 increase in *%body fat* among men who are of a particular *height*. Each inch of *height* is associated with a decrease in *%body fat* of about 0.60 among men with a particular *waist* size.

The standard errors for the slopes of 0.07 (*waist*) and 0.11 (*height*) are both small compared with the slopes themselves, so it looks like the coefficient estimates are fairly precise. The residuals have a standard deviation of 4.46%, which gives an indication of how precisely we can predict *%body fat* with this model.

Multiple Regression Inference I: I Thought I Saw an ANOVA Table . . .

A S **Mean Squares and More.** Here's an animated tour of the rest of the regression table. The numbers work together to help us understand the analysis.

There are several hypothesis tests in the multiple regression output, but all of them talk about the same thing. Each is concerned with whether the underlying model parameters are actually zero.

The first of these hypotheses is one we skipped over for simple regression (for reasons that will be clear in a minute). Now that we've looked at ANOVA (in Chapter 28),[3] we can recognize the ANOVA table sitting in the middle of the regression output. Where'd that come from?

The answer is that now that we have more than one predictor, there's an overall test we should consider before we do more inference on the coefficients. We ask the global question "Is this multiple regression model any good at all?" That is, would we do as well using just $\bar{y}$ to model y? What would that mean in terms of the regression? Well, if all the coefficients (except the intercept) were zero, we'd have

$$\hat{y} = b_0 + 0x_1 + \cdots + 0x_k$$

and we'd just set $b_0 = \bar{y}$.

To address the overall question, we'll test

$$\mathrm{H}_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

(That null hypothesis looks very much like the null hypothesis we tested in the Analysis of Variance in Chapter 28.)

We can test this hypothesis with a statistic that is labeled with the letter F (in honor of Sir Ronald Fisher, the developer of Analysis of Variance). In our example, the F-value is 307 on 2 and 247 degrees of freedom. The alternative hypothesis is just that the slope coefficients aren't all equal to zero, and the test is one-sided—bigger F-values mean smaller P-values. If the null hypothesis were true, the F-statistic would be near 1. The F-statistic here is quite large, so we can easily reject the null hypothesis and conclude that the multiple regression model is better than just using the mean.[4]

Why didn't we do this for simple regression? Because the null hypothesis would have just been that the lone model slope coefficient was zero, and we were already testing that with the t-statistic for the slope. In fact, the *square* of that t-statistic is equal to the F-statistic for the simple regression, so it really was the identical test.

Multiple Regression Inference II: Testing the Coefficients

Once we check the F-test and reject the null hypothesis—and, if we are being careful, *only* if we reject that hypothesis—we can move on to checking the test statistics

[3] If you skipped over Chapter 28, you can just take our word for this and read on.

[4] There are F tables on the CD, and they work pretty much as you'd expect. Most regression tables include a P-value for the F-statistic, but there's almost never a need to perform this particular test in a multiple regression. Usually we just glance at the F-statistic to see that it's reasonably far from 1.0, the value it would have if the true coefficients were really all zero.

for the individual coefficients. Those tests look like what we did for the slope of a simple regression. For each coefficient we test

$$H_0\colon \beta_j = 0$$

against the (two-sided) alternative that it isn't zero. The regression table gives a standard error for each coefficient and the ratio of the estimated coefficient to its standard error. If the assumptions and conditions are met (and now we need the Nearly Normal condition), these ratios follow a Student's *t*-distribution.

$$t_{n-k-1} = \frac{b_j - 0}{SE(b_j)}$$

How many degrees of freedom? We have a rule of thumb and it works here. The degrees of freedom is the number of data values minus the number of predictors (in this case, counting the intercept term). For our regression on two predictors, that's $n - 3$. You shouldn't have to look up the *t*-values. Almost every regression report includes the corresponding P-values.

We can build a confidence interval in the usual way, as an estimate ± a margin of error. As always, the margin of error is just the product of the standard error and a critical value. Here the critical value comes from the *t*-distribution on $n - k - 1$ degrees of freedom. So a confidence interval for β_j is

$$b_j \pm t^*_{n-k-1}\, SE(b_j).$$

The tricky parts of these tests are that the standard errors of the coefficients now require harder calculations (so we leave it to the technology) and the meaning of a coefficient, as we have seen, depends on all the *other* predictors in the multiple regression model.

That last bit is important. If we fail to reject the null hypothesis for a multiple regression coefficient, it does **not** mean that the corresponding predictor variable has no linear relationship to y. It means that the corresponding predictor contributes nothing to modeling y *after allowing for all the other predictors.*

How's That, Again?

This last point bears repeating. The multiple regression model looks so simple and straightforward:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

It *looks* like each β_j tells us the effect of its associated predictor, x_j, on the response variable, y. But that is not so. This is, without a doubt, the most common error that people make with multiple regression:

A S **How Regression Coefficients Change with New Variables.** When the regression model grows by including a new prdictor, all the coefficients are likely to change. That can help us understand what those coefficients mean.

- It is possible for there to be no simple relationship between y and x_j, and yet β_j in a *multiple* regression can be significantly different from 0. We saw this happen for the coefficient of *height* in our example.
- It is also possible for there to be a strong two-variable relationship between y and x_j, and yet β_j in a multiple regression can be almost 0 with a large P-value so that we must retain the null hypothesis that the true coefficient is zero. If

we're trying to model the horsepower of a car, using both its weight and its engine size, it may turn out that the coefficient for *engine size* is nearly 0. That *doesn't* mean that engine size isn't important for understanding horsepower. It simply means that after allowing for the weight of the car, the engine size doesn't give much *additional* information.

A S Multiple Regression Coefficients. You may be thinking that multiple regression coefficients must be more consistent than this discussion suggests. Here's a hands-on analysis for you to investigate.

- It is even possible for there to be a significant linear relationship between y and x_j in one direction, and yet β_j can be of the *opposite* sign and strongly significant in a multiple regression. More expensive cars tend to be bigger, and since bigger cars have worse fuel efficiency, the price of a car has a slightly negative association with fuel efficiency. But in a multiple regression of fuel efficiency on *weight* and *price,* the coefficient of *price* may be positive. If so, it means that *among cars of the same weight,* more expensive cars have better fuel efficiency. The simple regression on *price,* though, has the opposite direction because, *overall,* more expensive cars are bigger. This switch in sign may seem a little strange at first, but it's not really a contradiction at all. It's due to the change in the *meaning* of the coefficient of *price* when it is in a multiple regression rather than a simple regression.

So we'll say it once more: The coefficient of x_j in a multiple regression depends as much on the *other* predictors as it does on x_j. Remember that when you interpret a multiple regression model.

Another Example: Modeling Infant Mortality

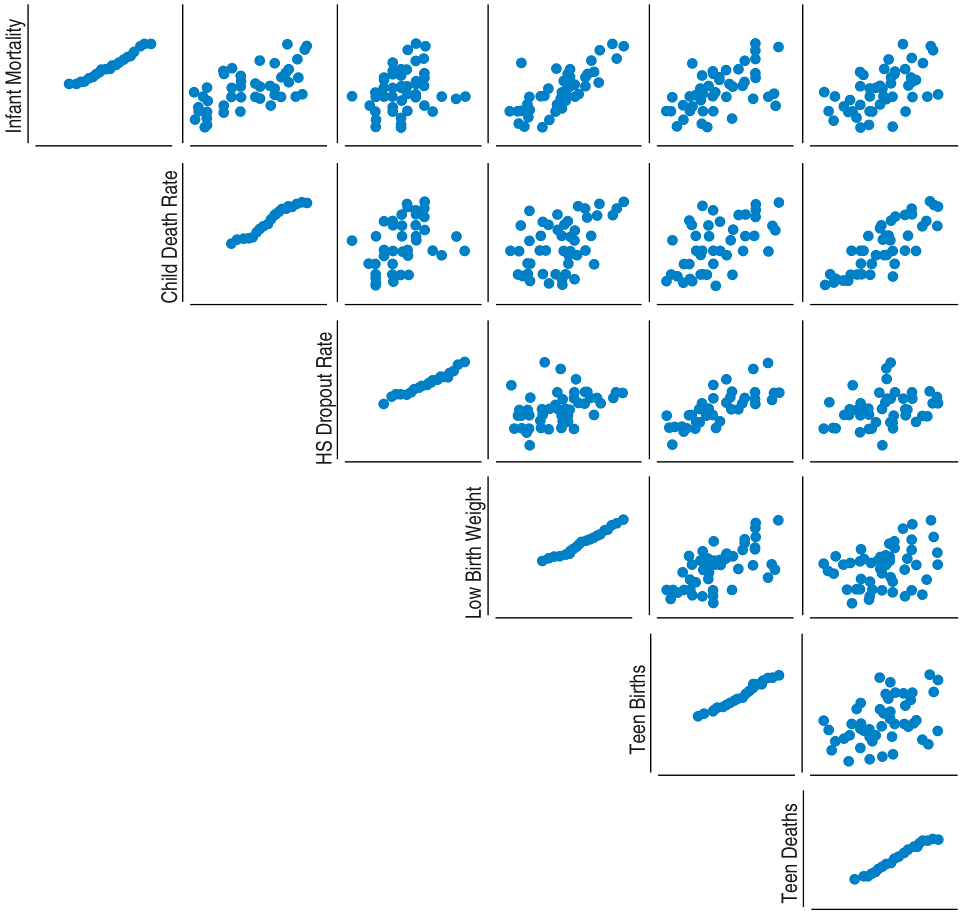
WHO	U.S. states
WHAT	Various measures relating to children and teens
WHEN	1999
WHY	Research and policy

Infant mortality is often used as a general measure of the quality of healthcare for children and mothers. It is reported as the rate of deaths of newborns per 1000 live births. Data recorded for each of the 50 states of the United States may allow us to build regression models to help understand or predict infant mortality. The variables available for our model are *child death rate* (deaths per 100,000 children aged 1–14), percent of teens who are *high school dropouts* (ages 16–19), percent of *low–birth weight babies* (*lbw*), *teen birth rate* (births per 100,000 females ages 15–17), and *teen deaths* by accident, homicide, and suicide (deaths per 100,000 teens ages 15–19).[5]

All of these variables were displayed and found to have no outliers and nearly Normal distributions.[6] One useful way to check many of our conditions is with a **scatterplot matrix.** This is an array of scatterplots set up so that the plots in each row have the same variable on their y-axis and those in each column have the same variable on their x-axis. This way every pair of variables is graphed. On the diagonal, rather than plotting a variable against itself, you'll usually find either a Normal probability plot or a histogram of the variable to help us assess the Nearly Normal Condition.

[5] The data are available from the Kids Count section of the Annie E. Casey Foundation, and are all for 1999.

[6] In the interest of complete honesty, we should point out that the original data include the District of Columbia, but it proved to be an outlier on several of the variables, so we've restricted attention to the 50 states here.



A scatterplot matrix shows a scatterplot of each pair of variables arrayed so that the vertical and horizontal axes are consistent across rows and down columns. The diagonal cells may hold Normal probability plots (as they do here), histograms, or just the names of the variables. These are a great way to check the Straight Enough Condition and to check for simple outliers. **Figure 29.6**

The individual scatterplots show at a glance that each of the relationships is straight enough for regression. There are no obvious bends, clumping, or outliers. And the plots don't thicken. So it looks like we can examine some multiple regression models with inference.

Inference for Multiple Regression Step-By-Step

Let's try to model *infant mortality* with all of the available predictors.

Think

Plan State what you want to know.

I wonder whether all or some of these predictors contribute to a useful model for *infant mortality*.

Hypotheses Specify your hypotheses.

(Hypotheses on the intercept are not particularly interesting for these data.)

First, there is an overall null hypothesis that asks whether the entire model is better than just modeling y with its mean:

H_O: The model itself contributes nothing useful, and all the slope coefficients,

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

H_A: At least one of the β_j is not 0.

If I reject this hypothesis, then I'll test a null hypothesis for each of the coefficients of the form:

Model State the null model.

H_0: The *j*-th variable contributes nothing useful, after allowing for the other predictors in the model: $\beta_j = 0$.

H_A: The *j*-th variable makes a useful contribution to the model: $\beta_j \neq 0$.

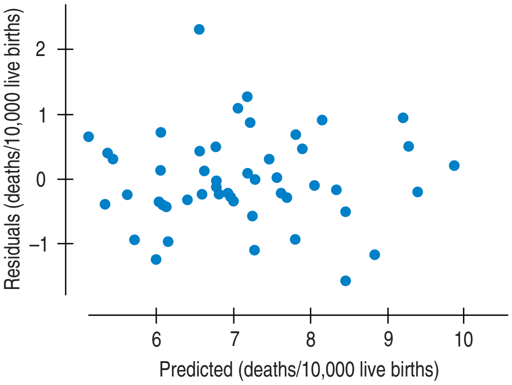
Check the appropriate assumptions and conditions.

✔ **Straight Enough Condition:** The scatterplot matrix shows no bends, clumping, or outliers.

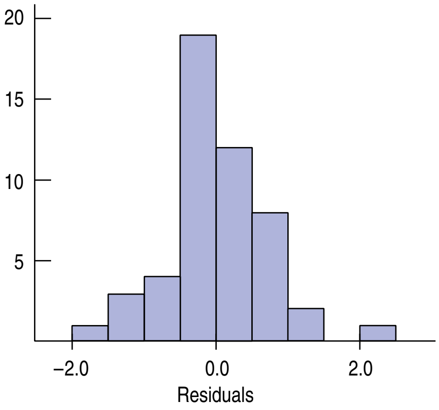
✔ **Independence Assumption:** These data are based on random samples and can be considered independent.

These conditions allow me to compute the regression model and find residuals.

✔ **Does the Plot Thicken? Condition:** The residual plot shows no obvious trends in the spread:



✔ **Nearly Normal Condition:** A histogram of the residuals is unimodal and symmetric.



The one possible outlier is South Dakota. I may repeat the analysis after removing South Dakota to see whether it changes substantially.

Choose your method.

Under these conditions I can continue with a ***multiple regression analysis.***

Show

Mechanics

Multiple regressions are always found from a computer program.

Computer output for this regression looks like this:

```
Dependent variable is: Infant mort
R-squared = 71.3 %   R-squared (adjusted) 68.0 %
s = 0.7520 with 50 − 6 = 44 degrees of freedom

                 Sum of             Mean
Source           Squares    DF     Square     F-ratio
Regression       61.7319     5    12.3464      21.8
Residual         24.8843    44     0.565553

Variable     Coefficient    SE(Coeff)   t-ratio    P-value
Intercept      1.63168       0.9124      1.79       0.0806
CDR            0.03123       0.0139      2.25       0.0292
HS drop       −0.09971       0.0610     −1.63       0.1096
Low BW         0.66103       0.1189      5.56      <0.0001
Teen
 births        0.01357       0.0238      0.57       0.5713
Teen
 deaths        0.00556       0.0113      0.49       0.6245
```

The P-values given in the regression output table are from the Student's t-distribution on $(n - 6) = 44$ degrees of freedom. They are appropriate for two-sided alternatives.

Consider the hypothesis tests. Under the assumptions we're willing to accept, and considering the conditions we've checked, the individual coefficients follow Student's t-distributions on 44 degrees of freedom.

The F-ratio of 21.8 on 5 and 44 degrees of freedom is certainly large enough to reject the default null hypothesis that the regression model is no better than using the mean infant mortality rate. So I'll go on to examine the individual coefficients.

Tell

Conclusion

Interpret your results in the proper context.

Most of these coefficients have relatively small t-ratios, so I can't be sure that their underlying values are not zero. Two of the coefficients, child death rate (cdr) and low birth weight (lbw), have P-values less than 5%. So I am confident that in this model both of these variables are unlikely to really have zero coefficients.

Overall the R^2 indicates that more than 71% of the variability in infant mortality can be accounted for with this regression model.

After allowing for the linear effects of the other variables in the model, an increase in the child death rate of 1 death per 100,000 is associated with an increase of 0.03 deaths per 1000 live births in the infant mortality rate. And an increase of 1% in the percentage of live births that are low birth weight is associated with an increase of 0.66 deaths per 1000 live births.

Comparing Multiple Regression Models

We have more variables available to us than we used when we modeled infant mortality. Moreover, several of those we tried don't seem to contribute to the model. How do we know that some other choice of predictors might not provide a better model? What exactly *would* make an alternative model better?

These are not easy questions. There is no simple measure of the success of a multiple regression model. Many people look at the R^2 value, and certainly we are not likely to be happy with a model that accounts for only a small fraction of the variability of y. But that's not enough. You can always drive the R^2 up by piling on more and more predictors, but models with many predictors are hard to understand. Keep in mind that the meaning of a regression coefficient depends on all the *other* predictors in the model, so it is best to keep the number of predictors as small as possible.

Regression models should make sense. Predictors that are easy to understand are usually better choices than obscure variables. Similarly, if there is a known mechanism by which a predictor has an effect on the response variable, that predictor is usually a good choice for the regression model.

How can we know whether we have the best possible model? The simple answer is that we can't. There's always the chance that some other predictors might bring an improvement (in higher R^2 or fewer predictors or simpler interpretation).

Adjusted R^2

You may have noticed that the full regression tables shown in this chapter include another statistic we haven't discussed. It is called adjusted R^2 and sometimes appears in computer output as R^2(adjusted). The **adjusted R^2** statistic is a rough attempt to adjust for the simple fact that when we add another predictor to a multiple regression, the R^2 can't go down and will most likely get larger. Only if we were to add a predictor whose coefficient turned out to be exactly zero would the R^2 remain the same. This fact makes it difficult to compare alternative regression models that have different numbers of predictors.

We can write a formula for R^2 using the sums of squares in the ANOVA table portion of the regression output table:

$$R^2 = \frac{SS_{Regression}}{SS_{Regression} + SS_{Residual}} = \frac{SS_{Regression}}{SS_{Total}}.$$

Adjusted R^2 simply substitutes the corresponding *mean squares* for the SS's:

$$R^2_{adj} = \frac{MS_{Regression}}{MS_{Total}}.$$

Because the mean squares are sums of squares divided by their degrees of freedom, they are adjusted for the number of predictors in the model. As a result, the adjusted R^2 value won't necessarily increase when a new predictor is added to the multiple regression model. That's fine. But adjusted R^2 no longer tells the fraction of variability accounted for by the model and it isn't even bounded by 0 and 100%, so it can be awkward to interpret.

Comparing alternative regression models is a challenge, especially when they have different numbers of predictors. The search for a summary statistic to help us choose among models is the subject of much contemporary research in Statistics. Adjusted R^2 is one common—but not necessarily the best—choice often found in computer regression output tables. Don't use it as the sole decision criterion when you compare different regression models.

What Can Go Wrong?

Interpreting Coefficients

- ***Don't claim to "hold everything else constant" for a single individual.*** It's often meaningless to say that a regression coefficient says what we expect to happen if all variables but one were held constant for an individual and the predictor in question changed. While it's mathematically correct, it often just doesn't make any sense. We can't gain a year of experience or have another child without getting a year older. Instead, we *can* think about all those who fit given criteria on some predictors and ask about the conditional relationship between y and one x for those individuals. The coefficient -0.60 of *height* for predicting *%body fat* says that among men of the same *waist* size, those who are one inch taller in *height* tend to be, on average, 0.60% lower in *%body fat*. The multiple regression coefficient measures that average conditional relationship.
- ***Don't interpret regression causally.*** Regressions are usually applied to observational data. Without deliberately assigned treatments, randomization, and control, we can't draw conclusions about causes and effects. We can never be certain that there are no variables lurking in the background, causing everything we've seen. Don't interpret b_1, the coefficient of x_1 in the multiple regression, by saying, "If we were to change an individual's x_1 by 1 unit (holding the other x's constant) it would change his y by b_1 units." We have no way of knowing what applying a change to an individual would do.
- ***Be cautious about interpreting a regression model as predictive.*** Yes, we do call the x's predictors, and you can certainly plug in values for each of the x's and find a corresponding *predicted value,* $\hat{y}$. But the term "prediction" suggests extrapolation into the future or beyond the data, and we know that we can get into trouble when we use models to estimate $\hat{y}$ values for x's not in the range of the data. Be careful not to extrapolate very far from the span of your data. In simple regression it was easy to tell when you extrapolated. With many predictor variables, it's often harder to know when you are outside the bounds of your original data.[7] We usually think of fitting models to the data more as modeling than as prediction, so that's often a more appropriate term.
- ***Don't think that the sign of a coefficient is special.*** Sometimes our primary interest in a predictor is whether it has a positive or negative association with y. As we have seen, though, the sign of the coefficient also depends on the other predictors in the model. Don't look at the sign in isolation and conclude that "the direction of the relationship is positive (or negative)." Just like the value of the coefficient, the sign is about the relationship after

[7] With several predictors we can wander beyond the data because of the *combination* of values even when individual values are not extraordinary. For example, both 28-inch waists and 76-inch heights can be found in men in the body fat study, but a single individual with both these measurements would not be at all typical. The model we fit is probably not appropriate for predicting the % body fat for such a tall and skinny individual.

allowing for the linear effects of the other predictors. The sign of a variable can change depending on which other predictors are in or out of the model. For example, in the regression model for infant mortality, the coefficient of *high school dropout rate* was negative and its P-value was fairly small, but the simple association between dropout rate and infant mortality is positive. (Check the plot matrix.)

- ***If a coefficient's t-statistic is not significant, don't interpret it at all.*** You can't be sure that the value of the corresponding parameter in the underlying regression model isn't really zero. ⊘

What Else Can Go Wrong?

- ***Don't fit a linear regression to data that aren't straight.*** This is the most fundamental regression assumption. If the relationship between the x's and y isn't approximately linear, there's no sense in fitting a linear model to it. What we mean by "linear" is a model of the form we have been writing for the regression. When we have two predictors, this is the equation of a plane, which is linear in the sense of being flat in all directions. With more predictors, the geometry is harder to visualize, but the simple structure of the model is consistent; the predicted values change consistently with equal size changes in any predictor.

  Usually we're satisfied when plots of y against each of the x's are straight enough. We'll also check a scatterplot of the residuals against the predicted values for signs of nonlinearity.
- ***Watch out for the plot thickening.*** The estimate of the error standard deviation shows up in all the inference formulas. If s_e changes with x, these estimates won't make sense. The most common check is a plot of the residuals against the predicted values. If plots of residuals against several of the predictors all show a thickening, and especially if they also show a bend, then consider re-expressing y. If the scatterplot against only one predictor shows thickening, consider re-expressing that predictor.
- ***Make sure the errors are nearly Normal.*** All of our inferences require that the true errors be modeled well by a Normal model. Check the histogram and Normal probability plot of the residuals to see whether this assumption looks reasonable.
- ***Watch out for high-influence points and outliers.*** We always have to be on the lookout for a few points that have undue influence on our model, and regression is certainly no exception. Partial regression plots are a good place to look for influential points and to understand how they affect each of the coefficients. ⊘

CONNECTIONS

We would never consider a regression analysis without first making scatterplots. The aspects of scatterplots that we always look for—their direction, shape, and scatter—relate directly to regression.

Regression inference is connected to just about every inference method we have seen for measured data. The assumption that the spread of data about the line is constant is essentially the same as the assumption of equal variances required for the pooled-t methods. Our use of all the residuals together to estimate their standard deviation is a form of pooling.

Of course, the ANOVA table in the regression output connects to our consideration of ANOVA in Chapter 28. This, too, is not coincidental. Multiple Regression, ANOVA, pooled *t*-tests, and inference for means are all part of a more general statistical model known as the General Linear Model (GLM).

What have we learned?

We first met regression in Chapter 8 and its inference in Chapter 27. Now we add more predictors to our equation.

We've learned that there are many similarities between simple and multiple regression:

- We fit the model by least squares.
- The assumptions and conditions are essentially the same. For multiple regression:
  1. The relationship of *y* with each *x* must be straight (check the scatterplots).
  2. The data values must be independent (think about how they were collected).
  3. The spread about the line must be the same across the *x*-axis for each predictor variable (make a scatterplot or check the plot of residuals against predicted values).
  4. The errors must follow a Normal model (check a histogram or Normal probability plot of the residuals).
- R^2 still gives us the fraction of the total variation in *y* accounted for by the model.
- We perform inference on the coefficients by looking at the *t*-values, created from the ratio of the coefficients to their standard errors.

But we've also learned that there are some profound differences in interpretation when adding more predictors:

- The coefficient of each *x* indicates the average change in *y* we'd expect to see for a unit change in that *x for particular values of all the other x-variables*.
- The coefficient of a predictor variable can change sign when another variable is entered or dropped from the model.
- Finding a suitable model from among the possibly hundreds of potential models is not straightforward.

TERMS

Multiple regression A linear regression with two or more predictors whose coefficients are found to minimize the sum of the squared residuals is a least squares linear multiple regression. But it is usually just called a multiple regression. When the distinction is needed, a least squares linear regression with a single predictor is called a simple regression. The multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

Least squares We still fit multiple regression models by choosing the coefficients that make the sum of the squared residuals as small as possible. This is called the method of least squares.

Partial regression plot The partial regression plot for a specified coefficient is a display that helps in understanding the meaning of that coefficient in a multiple regression. It has a slope equal to the coefficient value and shows the influences of each case on that value. A partial regression plot for a specified *x* displays the residuals when *y* is regressed on the *other* predictors against the residuals when the specified *x* is regressed on the other predictors.

Assumptions for inference in regression (and conditions to check for some of them)

- Linearity. Check that the scatterplots of *y* against each *x* are straight enough and that the scatterplot of residuals against predicted values has no obvious pattern. (If we find the relationships straight enough, we may fit the regression model to find residuals for further checking.)
- Independent errors. Think about the nature of the data. Check a residual plot. Any evident pattern in the residuals can call the assumption of independence into question.
- Constant variance. Check that the scatterplots show consistent spread across the ranges of the *x*-variables and that the residual plot has constant variance too. A common problem is increasing spread with increasing predicted values—*the plot thickens!*
- Normality of the residuals. Check a histogram or a Normal probability plot of the residuals.

ANOVA

The Analysis of Variance table that is ordinarily part of the multiple regression results offers an *F*-test to test the null hypothesis that the overall regression is no improvement over just modeling *y* with its mean:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

If this null hypothesis is not rejected, then you should not proceed to test the individual coefficients.

***t*-ratios for the coefficients**

The *t*-ratios for the coefficients can be used to test the null hypotheses that the true value of each coefficient is zero against the alternative that it is not.

Scatterplot matrix

A scatterplot matrix displays scatterplots for all pairs of a collection of variables, arranged so that all the plots in a row have the same variable displayed on their *y*-axis and all plots in a column have the same variable on their *x*-axis. Usually, the diagonal holds a display of a single variable such as a histogram or Normal probability plot, and identifies the variable in its row and column.

Adjusted R^2

An adjustment to the R^2 statistic that attempts to allow for the number of predictors in the model. It is sometimes used when comparing regression models with different numbers of predictors.

SKILLS

When you complete this lesson you should:

Think

- Understand that the "true" regression model is an idealized summary of the data.
- Know how to examine scatterplots of *y* vs. each *x* for violations of assumptions that would make inference for regression unwise or invalid.
- Know how to examine displays of the residuals from a multiple regression to check that the conditions have been satisfied. In particular, know how to judge linearity and constant variance from a scatterplot of residuals against predicted values. Know how to judge Normality from a histogram and Normal probability plot.
- Remember to be especially careful to check for failures of the independence assumption when working with data recorded over time. Examine scatterplots of the residuals against time and look for patterns.

Show

- Be able to use a statistics package to perform the calculations and make the displays for multiple regression, including a scatterplot matrix of the variables, a scatterplot of residuals vs. predicted values, and partial regression plots for each coefficient.
- Know how to use the ANOVA *F*-test to check that the overall regression model is better than just using the mean of *y*.

- Know how to test the standard hypotheses that each regression coefficient is really zero. Be able to state the null and alternative hypotheses. Know where to find the relevant numbers in standard computer regression output.

Tell

- Be able to summarize a regression in words. In particular, be able to state the meaning of the regression coefficients, taking full account of the effects of the other predictors in the model.
- Be able to interpret the *F*-statistic for the overall regression.
- Be able to interpret the P-value of the *t*-statistics for the coefficients to test the standard null hypotheses.

Regression Analysis on the Computer

All statistics packages make a table of results for a regression. If you can read a package's regression output table for simple regression, then you can read its table for a multiple regression. You'll want to look at the ANOVA table, and you'll see information for each of the coefficients, not just for a single slope.

Most packages offer to plot residuals against predicted values. Some will also plot residuals against the *x*'s. With some packages you must request plots of the residuals when you request the regression. Others let you find the regression first and then analyze the residuals afterward. Either way, your analysis is not complete if you don't check the residuals with a histogram or Normal probability plot and a scatterplot of the residuals against the *x*'s or the predicted values.

One good way to check assumptions before embarking on a multiple regression analysis is with a scatterplot matrix. This is sometimes abbreviated SPLOM in commands.

Multiple regressions are always found with a computer or programmable calculator. Before computers were available, a full multiple regression analysis could take months or even years of work.

DATA DESK

- Select Y- and X-variable icons.
- From the **Calc** menu, choose **Regression.**
- Data Desk displays the regression table.
- Select plots of residuals from the Regression table's HyperView menu.

Comments

You can change the regression by dragging the icon of another variable over either the Y- or an X-variable name in the table and dropping it there. You can add a predictor by dragging its icon into that part of the table. The regression will recompute automatically.

EXCEL

- From the **Tools** menu, select Data Analysis.
- Select Regression from the **Analysis Tools** list.
- Click the **OK** button.
- Enter the data range holding the Y-variable in the box labeled "Y-range."
- Enter the range of cells holding the X-variables in the box labeled "X-range."
- Select the **New Worksheet Ply** option.
- Select **Residuals** options. Click the **OK** button.

Comments

The Y and X ranges do not need to be in the same rows of the spreadsheet, although they must cover the same number of cells. But it is a good idea to arrange your data in parallel columns as in a data table. The X-variables must be in adjacent columns. No cells in the data range may hold non-numeric values.

Although the dialog offers a Normal probability plot of the residuals, the data analysis add-in does not make a correct probability plot, so don't use this option.

JMP

- From the **Analyze** menu select **Fit Model.**
- Specify the response, Y. Assign the predictors, X, in the **Construct Model Effects** dialog box.
- Click on **Run Model.**

Comments

JMP chooses a regression analysis when the response variable is "Continuous." The predictors can be any combination of quantitative or categorical. If you get a different analysis, check the variable types.

MINITAB

- Choose **Regression** from the **Stat** menu.
- Choose **Regression. . .** from the **Regression** submenu.
- In the Regression dialog, assign the Y-variable to the Response box and assign the X-variables to the Predictors box.
- Click the **Graphs** button.
- In the Regression-Graphs dialog, select **Standardized residuals,** and check **Normal plot of residuals** and **Residuals versus fits.**
- Click the **OK** button to return to the Regression dialog.
- To specify displays, click **Graphs,** and check the displays you want.
- Click the **OK** button to return to the Regression dialog.
- Click the **OK** button to compute the regression.

SPSS

- Choose **Regression** from the **Analyze** menu.
- Choose **Linear** from the **Regression** submenu.
- When the Linear Regression dialog appears, select the Y-variable and move it to the dependent target. Then move the X-variables to the independent target.
- Click the **Plots** button.
- In the Linear Regression Plots dialog, choose to plot the *SRESIDs against the *ZPRED values.
- Click the **Continue** button to return to the Linear Regression dialog.
- Click the **OK** button to compute the regression.

TI-83/84 Plus

Comments

You need a special program to compute a multiple regression on the TI-83.

TI-89

Under **STAT Tests** choose **B:MultREg Tests**

- Specify the number of predictor variables, and which lists contain the response variable and predictor variables.
- Press $\boxed{\div}$ to perform the calculations.

Comments

- The first portion of the output gives the *F*-statistic and its P-value as well as the values of R^2, $AdjR^2$, the standard deviation of the residuals (s), and the Durbin-Watson statistic, which measures correlation among the residuals.
- The rest of the main output gives the components of the *F*-test, as well as values of the coefficients, their standard errors, and associated *t*-statistics along with P-values. You can use the right arrow to scroll through these lists (if desired).
- The calculator creates several new lists that can be used for assessing the model and its conditions: Yhatlist, resid, sresid (standardized residuals), leverage, and cookd, as well as lists of the coefficients, standard errors, *t*'s, and P-values.

EXERCISES

1. Interpretations. A regression performed to predict selling price of houses found the equation

$$\widehat{price} = 169328 + 35.3\,area + 0.718\,lotsize - 6543\,age$$

where *price* is in dollars, *area* is in square feet, *lotsize* is in square feet, and *age* is in years. The R^2 is 92%. One of the interpretations below is correct. Which is it? Explain what's wrong with the others.

a) Each year a house *age*s it is worth $6543 less.

b) Every extra square foot of *area* is associated with an additional $35.30 in average price, for houses with a given *lotsize* and *age*.

c) Every dollar in price means *lotsize* increases 0.718 square feet.

d) This model fits 92% of the data points exactly.

2. More interpretations. A household appliance manufacturer wants to analyze the relationship between total sales and the company's three primary means of advertising (television, magazines, and radio). All values were in millions of dollars. They found the regression equation

$$\widehat{sales} = 250 + 6.75\,TV + 3.5\,radio + 2.3\,magazines.$$

One of the interpretations below is correct. Which is it? Explain what's wrong with the others.

a) If they did no advertising, their income would be $250 million.

b) Every million dollars spent on radio makes sales increase $3.5 million, all other things being equal.

c) Every million dollars spent on magazines increases TV spending $2.3 million.

d) Sales increase on average about $6.75 million for each million spent on TV, after allowing for the effects of the other kinds of advertising.

3. Predicting final exams. How well do exams given during the semester predict performance on the final? One class had three tests during the semester. Computer output of the regression gives

```
Dependent variable is Final
s =13.46  R-Sq = 77.7%   R-Sq(adj) = 74.1%

Predictor     Coeff     SE(Coeff)      t       P-value
Intercept    −6.72      14.00        −0.48      0.636
Test1         0.2560     0.2274       1.13      0.274
Test2         0.3912     0.2198       1.78      0.091
Test3         0.9015     0.2086       4.32     <0.0001

Analysis of Variance
Source       DF     SS        MS        F     P-value
Regression    3   11961.8   3987.3   22.02  <0.0001
Error        19    3440.8    181.1
Total        22   15402.6
```

a) Write the equation of the regression model.
b) How much of the variation in final exam scores is accounted for by the regression model?
c) Explain in context what the coefficient of *Test3* scores means.
d) A student argues that clearly the first exam doesn't help to predict final performance. She suggests that this exam not be given at all. Does Test 1 have no effect on the final exam score? Can you tell from this model? (*Hint:* Do you think test scores are related to each other?)

4. Scottish hill races. Hill running—races up and down hills—has a written history in Scotland dating back to the year 1040. Races are held throughout the year at different locations around Scotland. A recent compilation of information for 71 races (for which full information was available and omitting two unusual races) includes the *distance* (miles), the *climb* (ft), and the *record time* (seconds). A regression to predict the men's records as of 2000 looks like this:

Dependent variable is: Men's record
R-squared = 98.0% R-squared (adjusted) = 98.0%
s = 369.7 with 71 − 3 = 68 degrees of freedom

	Sum of		Mean	
Source	**Squares**	**df**	**Square**	**F-ratio**
Regression	458947098	2	229473549	1679
Residual	9293383	68	136667	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−521.995	78.39	−6.66	<0.0001
Distance	351.879	12.25	28.7	<0.0001
Climb	0.643396	0.0409	15.7	<0.0001

a) Write the regression equation. Give a brief report on what it says about men's record times in hill races.
b) Interpret the value of R^2 in this regression.
c) What does the coefficient of *climb* mean in this regression?

5. Home prices. Many variables have an impact on determining the price of a house. A few of these are *size* of the house (square feet), *lot size,* and number of *bathrooms.* Information for a random sample of homes for sale in the Statesboro, GA, area was obtained from the Internet. Regression output modeling the *asking price* with *square footage* and number of *bathrooms* gave the following result:

Dependent Variable is: Price
s = 67013 R-Sq = 71.1% R-Sq (adj) = 64.6%

Predictor	Coeff	SE(Coeff)	T	P-value
Intercept	−152037	85619	−1.78	0.110
Baths	9530	40826	0.23	0.821
Sq ft	139.87	46.67	3.00	0.015

Analysis of Variance

Source	DF	SS	MS	F	P-value
Regression	2	99303550067	49651775033	11.06	0.004
Residual	9	40416679100	4490742122		
Total	11	1.39720E+11			

a) Write the regression equation.
b) How much of the variation in home asking prices is accounted for by the model?
c) Explain in context what the coefficient of *square footage* means.
d) The owner of a construction firm, upon seeing this model, objects because the model says that the number of bathrooms has no effect on the price of the home. He says that when *he* adds another bathroom, it increases the value. Is it true that the number of bathrooms is unrelated to house price? (*Hint:* Do you think bigger houses have more bathrooms?)

6. More hill races. Here is the regression for the women's records for the same Scottish hill races we considered in Exercise 4:

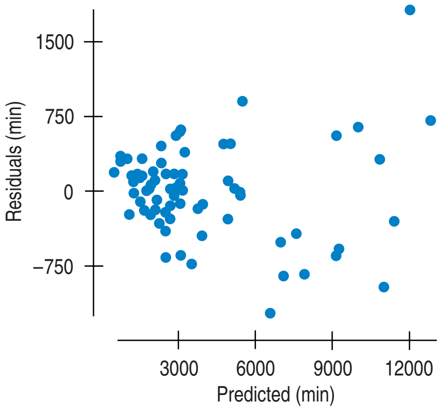
Dependent variable is: Women's record
R-squared = 97.7% R-squared (adjusted) = 97.6%
s = 479.5 with 71 − 3 = 68 degrees of freedom

	Sum of		Mean	
Source	**Squares**	**df**	**Square**	**F-ratio**
Regression	658112727	2	329056364	1431
Residual	15634430	68	229918	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−554.015	101.7	−5.45	<0.0001
Distance	418.632	15.89	26.4	<0.0001
Climb	0.780568	0.0531	14.7	<0.0001

a) Compare the regression model for the women's records with that found for the men's records in Exercise 4.

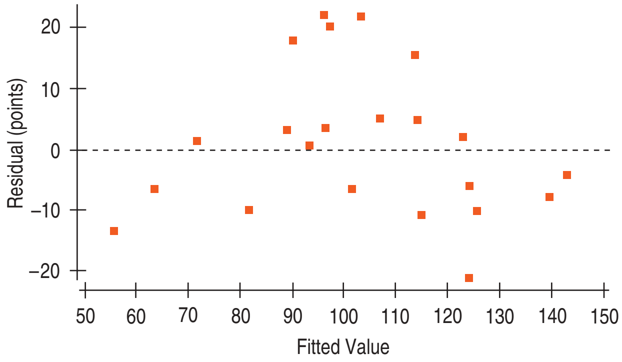
Here's a scatterplot of the residuals for this regression:



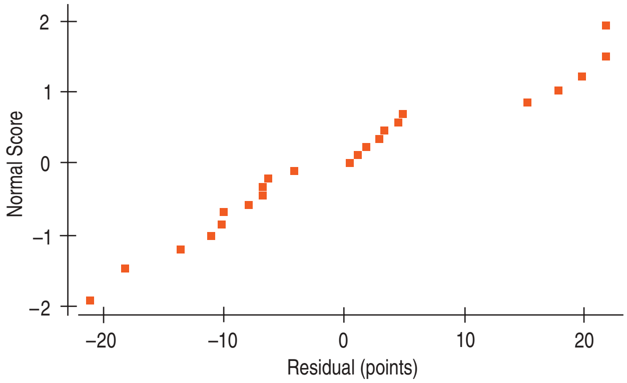
b) Discuss the residuals and what they say about the assumptions and conditions for this regression.

7. Predicting finals II. Here are some diagnostic plots for the final exam data from Exercise 3. These were generated by a computer package and may look different from the plots generated by the packages you use. (In particular, note that the axes of the Normal probability plot are swapped relative to the plots we've made in the text. We only care about the pattern of this plot, so it shouldn't affect your interpretation.) Examine these plots and discuss whether the assumptions and conditions for the multiple regression seem reasonable.

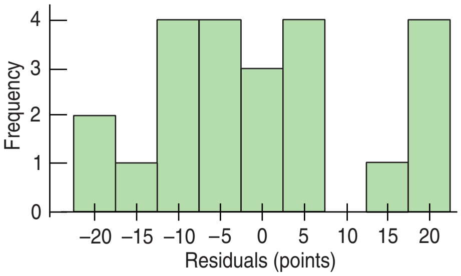
Residuals vs. the Fitted Values
(Response is Final)



Normal Probability Plot of the Residuals
(Response is Final)



Histogram of the Residuals
(Response is Final)



8. Secretary performance. The AFL-CIO has undertaken a study of 30 secretaries' yearly salaries (in thousands of dollars). The organization wants to predict salaries from several other variables.

The variables considered to be potential predictors of salary are:

X1 = months of service
X2 = years of education
X3 = score on standardized test
X4 = words per minute (wpm) typing speed
X5 = ability to take dictation in words per minute

A multiple regression model with all five variables was run on a computer package, resulting in the following output:

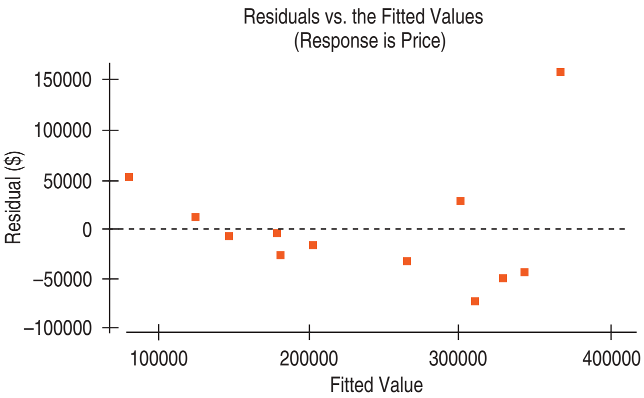
Variable	Coefficient	Std. Error	t-value
Constant	9.788	0.377	25.960
X1	0.110	0.019	5.178
X2	0.053	0.038	1.369
X3	0.071	0.064	1.119
X4	0.004	0.307	0.013
X5	0.065	0.038	1.734

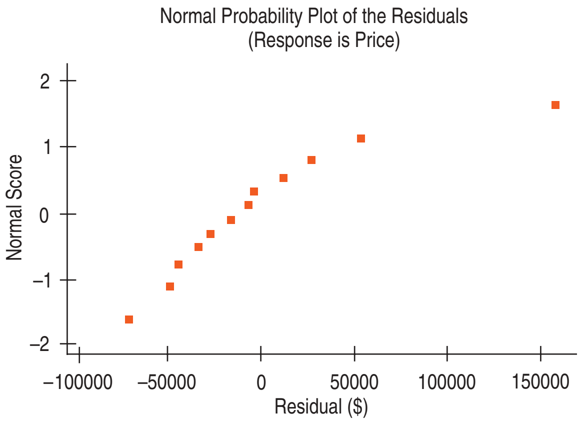
s = 0.430 $R^2 = 0.863$

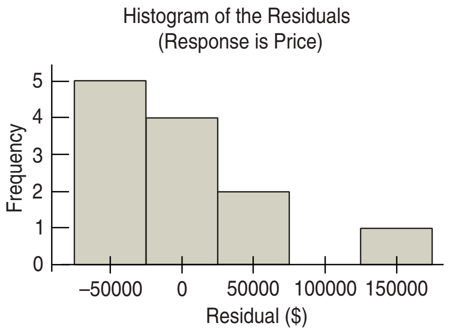
Assume that the residual plots show no violations of the conditions for using a linear regression model.

a) What is the regression equation?
b) From this model, what is the predicted *salary* (in thousands of dollars) of a secretary with 10 years (120 months) of experience, 9th grade education (9 years of education), a 50 on the standardized test, 60 wpm typing speed, and the ability to take 30 wpm dictation?
c) Test whether the coefficient for words per minute of typing speed (*X4*) is significantly different from zero at $\alpha = 0.05$.
d) How might this model be improved?
e) A correlation of *age* with *salary* finds $r = 0.682$, and the scatterplot shows a moderately strong positive linear association. However, if X6 = *age* is added to the multiple regression, the estimated coefficient of *age* turns out to be $b_6 = -0.154$. Explain some possible causes for this apparent change of direction in the relationship between age and salary.

9. Home prices II. Here are some diagnostic plots for the home prices data from Exercise 5. These were generated by a computer package and may look different from the plots generated by the packages you use. (In particular, note that the axes of the Normal probability plot are swapped relative to the plots we've made in the text. We only care about the pattern of this plot, so it shouldn't affect your interpretation.) Examine these plots and discuss whether the assumptions and conditions for the multiple regression seem reasonable.


Residuals vs. the Fitted Values
(Response is Price)


Normal Probability Plot of the Residuals
(Response is Price)


Histogram of the Residuals
(Response is Price)

10. GPA and SATs. A large section of Stat 101 was asked to fill out a survey on grade point average and SAT scores. A regression was run to find out how well Math and Verbal SAT scores could predict academic performance as measured by GPA. The regression was run on a computer package with the following output:

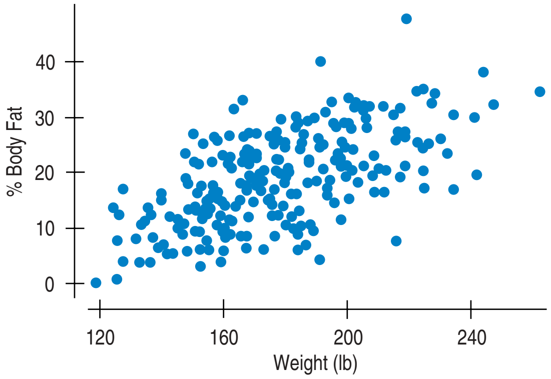
```
Response: GPA

              Coefficient   Std Error   t-ratio   Prob > |t|
Constant      0.574968      0.253874    2.26      0.0249
SAT Verbal    0.001394      0.000519    2.69      0.0080
SAT Math      0.001978      0.000526    3.76      0.0002
```

a) What is the regression equation?

b) From this model, what is the predicted GPA of a student with an SAT Verbal score of 500 and an SAT Math score of 550?

c) What else would you want to know about this regression before writing a report about the relationship between SAT scores and grade point averages? Why would these be important to know?

T **11. Body fat revisited.** The data set on body fat contains 15 body measurements on 250 men from 22 to 81 years old. Is average *%body fat* related to *weight?* Here's a scatterplot:



And here's the simple regression:

```
Dependent variable is: Pct BF
R-squared = 38.1%   R-squared (adjusted) = 37.9%
s = 6.538 with 250 − 2 = 248 degrees of freedom

Variable    Coefficient   SE(Coeff)   t-ratio   P-value
Intercept   −14.6931      2.760       −5.32     <0.0001
Weight        0.18937     0.0153      12.4      <0.0001
```

a) Is the coefficient of *%body fat* on *weight* statistically distinguishable from 0? (Perform a hypothesis test.)

b) What does the slope coefficient mean in this regression*?*

We saw before that the slopes of both *waist* size and *height* are statistically significant when entered into a multiple regression equation. What happens if we add *weight* to that regression? Recall that we've already checked the assumptions and conditions for regression on *waist* size and *height* in the chapter. Here is the output from a regression on all three variables:

```
Dependent variable is: Pct BF
R-squared = 72.5%   R-squared (adjusted) = 72.2%
s = 4.376 with 250 − 4 = 246 degrees of freedom

              Sum of                Mean
Source        Squares      df       Square      F-ratio
Regression    12418.7        3      4139.57     216
Residual       4710.11     246        19.1468
```

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−31.4830	11.54	−2.73	0.0068
Waist	2.31848	0.1820	12.7	<0.0001
Height	−0.224932	0.1583	−1.42	0.1567
Weight	−0.100572	0.0310	−3.25	0.0013

c) Interpret the slope for *weight*. How can the coefficient for *weight* in this model be negative when its coefficient was positive in the simple regression model?

d) What does the P-value for *height* mean in this regression? (Perform the hypothesis test.)

12. Breakfast cereals. We saw in Chapter 8 that the calorie content of a breakfast cereal is linearly associated with its sugar content. Is that the whole story? Here's the output of a regression model that regresses *calories* for each serving on its *protein(g)*, *fat(g)*, *fiber(g)*, *carbohydrate(g)*, and *sugars(g)* content.

Dependent variable is: calories
R-squared = 84.5% R-squared (adjusted) = 83.4%
s = 7.947 with 77 − 6 = 71 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	24367.5	5	4873.50	77.2
Residual	4484.45	71	63.1613	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	20.2454	5.984	3.38	0.0012
Protein	5.69540	1.072	5.32	<0.0001
Fat	8.35958	1.033	8.09	<0.0001
Fiber	−1.02018	0.4835	−2.11	0.0384
Carbo	2.93570	0.2601	11.3	<0.0001
Sugars	3.31849	0.2501	13.3	<0.0001

Assuming that the conditions for multiple regression are met,

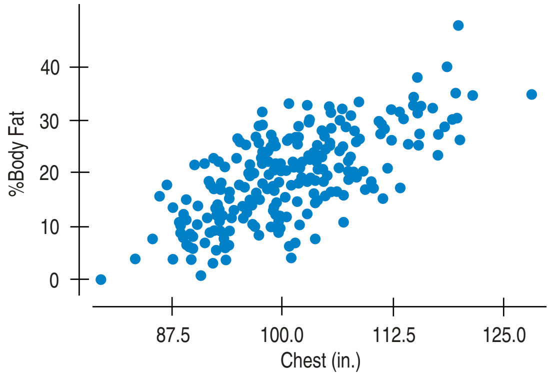
a) What is the regression equation?

b) Do you think this model would do a reasonably good job at predicting calories? Explain.

c) To check the conditions, what plots of the data might you want to examine?

d) What does the coefficient of *fat* mean in this model?

13. Body fat again. Chest size might be a good predictor of body fat. Here's a scatterplot of *%body fat* vs. *chest size.*



A regression of *%body fat* on *chest size* gives the following equation:

Dependent variable is: Pct BF
R-squared = 49.1% R-squared (adjusted) = 48.9%
s = 5.930 with 250 − 2 = 248 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−52.7122	4.654	−11.3	<0.0001
Chest	0.712720	0.0461	15.5	<0.0001

a) Is the slope of *%body fat* on *chest size* statistically distinguishable from 0? (Perform a hypothesis test.)

b) What does the answer in part a mean about the relationship between *%body fat* and *chest size?*

We saw before that the slopes of both *waist* size and *height* are statistically significant when entered into a multiple regression equation. What happens if we add *chest size* to that regression? Here is the output from a regression on all three variables:

Dependent variable is: Pct BF
R-squared = 72.2% R-squared (adjusted) = 71.9%
s = 4.399 with 250 − 4 = 246 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio	P
Regression	12368.9	3	4122.98	213	<0.0001
Residual	4759.87	246	19.3491		

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	2.07220	7.802	0.266	0.7908
Waist	2.19939	0.1675	13.1	<0.0001
Height	−0.561058	0.1094	−5.13	<0.0001
Chest	−0.233531	0.0832	−2.81	0.0054

c) Interpret the coefficient for *chest.*

d) Would you consider removing any of the variables from this regression model? Why or why not?

14. Grades. The table below shows the five scores from an introductory Statistics course. Find a model for predicting final exam score by trying all possible models with two predictor variables. Which model would you choose? Be sure to check the conditions for multiple regression.

Name	Final	Midterm 1	Midterm 2	Project	Home work
Timothy F.	117	82	30	10.5	61
Karen E.	183	96	68	11.3	72
Verena Z.	124	57	82	11.3	69
Jonathan A.	177	89	92	10.5	84
Elizabeth L.	169	88	86	10.6	84
Patrick M.	164	93	81	10.0	71
Julia E.	134	90	83	11.3	79
Thomas A.	98	83	21	11.2	51
Marshall K.	136	59	62	9.1	58
Justin E.	183	89	57	10.7	79

continued

Name	Final	Midterm 1	Midterm 2	Project	Home work
Alexandra E.	171	83	86	11.5	78
Christopher B.	173	95	75	8.0	77
Justin C.	164	81	66	10.7	66
Miguel A.	150	86	63	8.0	74
Brian J.	153	81	86	9.2	76
Gregory J.	149	81	87	9.2	75
Kristina G.	178	98	96	9.3	84
Timothy B.	75	50	27	10.0	20
Jason C.	159	91	83	10.6	71
Whitney E.	157	87	89	10.5	85
Alexis P.	158	90	91	11.3	68
Nicholas T.	171	95	82	10.5	68
Amandeep S.	173	91	37	10.6	54
Irena R.	165	93	81	9.3	82
Yvon T.	168	88	66	10.5	82
Sara M.	186	99	90	7.5	77
Annie P.	157	89	92	10.3	68
Benjamin S.	177	87	62	10.0	72
David W.	170	92	66	11.5	78
Josef H.	78	62	43	9.1	56
Rebecca S.	191	93	87	11.2	80
Joshua D.	169	95	93	9.1	87
Ian M.	170	93	65	9.5	66
Katharine A.	172	92	98	10.0	77
Emily R.	168	91	95	10.7	83
Brian M.	179	92	80	11.5	82
Shad M.	148	61	58	10.5	65
Michael R.	103	55	65	10.3	51
Israel M.	144	76	88	9.2	67
Iris J.	155	63	62	7.5	67
Mark G.	141	89	66	8.0	72
Peter H.	138	91	42	11.5	66
Catherine R.M.	180	90	85	11.2	78
Christina M.	120	75	62	9.1	72
Enrique J.	86	75	46	10.3	72
Sarah K.	151	91	65	9.3	77
Thomas J.	149	84	70	8.0	70
Sonya P.	163	94	92	10.5	81
Michael B.	153	93	78	10.3	72
Wesley M.	172	91	58	10.5	66
Mark R.	165	91	61	10.5	79
Adam J.	155	89	86	9.1	62
Jared A.	181	98	92	11.2	83
Michael T.	172	96	51	9.1	83
Kathryn D.	177	95	95	10.0	87
Nicole M.	189	98	89	7.5	77
Wayne E.	161	89	79	9.5	44
Elizabeth S.	146	93	89	10.7	73
John R.	147	74	64	9.1	72
Valentin A.	160	97	96	9.1	80
David T. O.	159	94	90	10.6	88
Marc I.	101	81	89	9.5	62
Samuel E.	154	94	85	10.5	76
Brooke S.	183	92	90	9.5	86

T **15. Fifty states.** Here is a data set on various measures of the 50 United States. The *murder* rate is per 100,000, *HS graduation* rate is in %, *income* is per capita income in dollars, *illiteracy* rate is per 1000, and *life expectancy* is in years. Find a regression model for *life expectancy* with three predictor variables by trying all four of the possible models.

a) Which model appears to do the best?

b) Would you leave all three predictors in this model?

c) Does this model mean that by changing the levels of the predictors in this equation, we could affect life expectancy in that state? Explain.

d) Be sure to check the conditions for multiple regression. What do you conclude?

State name	Murder	HS grad	Income	Illiteracy	Life exp
Alabama	15.1	41.3	3624	2.1	69.05
Alaska	11.3	66.7	6315	1.5	69.31
Arizona	7.8	58.1	4530	1.8	70.55
Arkansas	10.1	39.9	3378	1.9	70.66
California	10.3	62.6	5114	1.1	71.71
Colorado	6.8	63.9	4884	0.7	72.06
Connecticut	3.1	56.0	5348	1.1	72.48
Delaware	6.2	54.6	4809	0.9	70.06
Florida	10.7	52.6	4815	1.3	70.66
Georgia	13.9	40.6	4091	2.0	68.54
Hawaii	6.2	61.9	4963	1.9	73.60
Idaho	5.3	59.5	4119	0.6	71.87
Illinois	10.3	52.6	5107	0.9	70.14
Indiana	7.1	52.9	4458	0.7	70.88
Iowa	2.3	59.0	4628	0.5	72.56
Kansas	4.5	59.9	4669	0.6	72.58
Kentucky	10.6	38.5	3712	1.6	70.10
Louisiana	13.2	42.2	3545	2.8	68.76
Maine	2.7	54.7	3694	0.7	70.39
Maryland	8.5	52.3	5299	0.9	70.22
Massachusetts	3.3	58.5	4755	1.1	71.83
Michigan	11.1	52.8	4751	0.9	70.63
Minnesota	2.3	57.6	4675	0.6	72.96
Mississippi	12.5	41.0	3098	2.4	68.09
Missouri	9.3	48.8	4254	0.8	70.69
Montana	5.0	59.2	4347	0.6	70.56
Nebraska	2.9	59.3	4508	0.6	72.60
Nevada	11.5	65.2	5149	0.5	69.03
New Hampshire	3.3	57.6	4281	0.7	71.23
New Jersey	5.2	52.5	5237	1.1	70.93
New Mexico	9.7	55.2	3601	2.2	70.32
New York	10.9	52.7	4903	1.4	70.55
North Carolina	11.1	38.5	3875	1.8	69.21
North Dakota	1.4	50.3	5087	0.8	72.78
Ohio	7.4	53.2	4561	0.8	70.82
Oklahoma	6.4	51.6	3983	1.1	71.42
Oregon	4.2	60.0	4660	0.6	72.13

State name	Murder	HS grad	Income	Illiteracy	Life exp
Pennsylvania	6.1	50.2	4449	1.0	70.43
Rhode Island	2.4	46.4	4558	1.3	71.9
South Carolina	11.6	37.8	3635	2.3	67.96
South Dakota	1.7	53.3	4167	0.5	72.08
Tennessee	11.0	41.8	3821	1.7	70.11
Texas	12.2	47.4	4188	2.2	70.90
Utah	4.5	67.3	4022	0.6	72.90
Vermont	5.5	57.1	3907	0.6	71.64
Virginia	9.5	47.8	4701	1.4	70.08
Washington	4.3	63.5	4864	0.6	71.72
West Virginia	6.7	41.6	3617	1.4	69.48
Wisconsin	3.0	54.5	4468	0.7	72.48
Wyoming	6.9	62.9	4566	0.6	70.29

16. Breakfast cereals again. We saw in Chapter 8 that the calorie count of a breakfast cereal is linearly associated with its sugar content. Can we predict the calories of a serving from its vitamins and mineral content? Here's a multiple regression model of *calories* per serving on its *sodium (mg)*, *potassium (mg)*, and *sugars (g)*:

```
Dependent variable is: Calories
R-squared = 38.9%   R-squared (adjusted) = 36.4%
s = 15.74 with 75 – 4 = 71 degrees of freedom

             Sum of           Mean
Source       Squares    df    Square     F-ratio   P-value
Regression   11211.1     3    3737.05     15.1     <0.0001
Residual     17583.5    71    247.655

Variable    Coefficient    SE(Coeff)   t-ratio     P-value
Intercept   81.9436         5.456       15.0      <0.0001
Sodium       0.05922        0.0218       2.72      0.0082
Potassium   –0.01684        0.0260      –0.648     0.5193
Sugars       2.44750        0.4164       5.88     <0.0001
```

Assuming that the conditions for multiple regression are met,

a) What is the regression equation?

b) Do you think this model would do a reasonably good job at predicting calories? Explain.

c) Would you consider removing any of these predictor variables from the model? Why or why not?

d) To check the conditions, what plots of the data might you want to examine?

17. Burger King revisited. Recall the Burger King menu data from Chapter 8. BK's nutrition sheet lists many variables. Here's a multiple regression to predict calories for Burger King foods from *protein* content *(g)*, *total fat (g)*, *carbohydrate (g)*, and *sodium (mg)* per serving:

```
Dependent variable is: Calories
R-squared = 100.0%   R-squared (adjusted) = 100.0%
s = 3.140 with 31 – 5 = 26 degrees of freedom

               Sum of              Mean
Source         Squares     df     Square     F-ratio
Regression     1419311      4     354828     35994
Residual       256.307     26     9.85796

Variable     Coefficient    SE(Coeff)   t-ratio    P-value
Intercept     6.53412       2.425        2.69      0.0122
Protein       3.83855       0.0859      44.7      <0.0001
Total fat     9.14121       0.0779     117        <0.0001
Carbs         3.94033       0.0336     117        <0.0001
Na/S         –0.69155       0.2970      –2.33      0.0279
```

a) Do you think this model would do a good job of predicting calories for a new BK menu item? Why or why not?

b) The mean of *calories* is 455.5 with a standard deviation of 217.5. Discuss what the value of *s* in the regression means about how well the model fits the data.

c) Does the R^2 value of 100.0% mean that the residuals are all actually equal to zero?