chapter

29 Multiple Regression Wisdom

Where have we been?

We've looked ahead in each of the preceding chapters, but this is a good time to take stock. Wisdom in building and interpreting multiple regressions uses all that we've discussed throughout this book—even histograms and scatterplots. But most important is to keep in mind that we use models to help us understand the world with data. This chapter is about building that understanding even when we use powerful, complex methods. And that's been our purpose all along.

Roller coasters are an old thrill that continues to grow in popularity. Engineers and designers compete to make them bigger and faster. For a two-minute ride on the best roller coasters, fans will wait hours. Can we learn what makes a roller coaster fast? Or how long the ride will last? Here are data on some of the fastest roller coasters in the world:

Name	Park	Country	Type	Duration (sec)	Speed (mph)	Height (ft)	Drop (ft)	Length (ft)	Inversion?
New Mexico Rattler	Cliff's Amusement Park	USA	Wooden	75	47	80	75	2750	No
Fujiyama	Fuji-Q Highlands	Japan	Steel	216	80.8	259.2	229.7	6708.67	No
Goliath	Six Flags Magic Mountain	USA	Steel	180	85	235	255	4500	No
Great American Scream Machine	Six Flags Great Adventure	USA	Steel	140	68	173	155	3800	Yes
Hangman	Wild Adventures	USA	Steel	125	55	115	95	2170	Yes
Hayabusa	Tokyo SummerLand	Japan	Steel	108	60.3	137.8	124.67	2559.1	No
Hercules	Dorney Park	USA	Wooden	135	65	95	151	4000	No
Hurricane	Myrtle Beach Pavilion	USA	Wooden	120	55	101.5	100	3800	No

Table 29.1
A small selection of coasters from the larger data set available on the DVD.

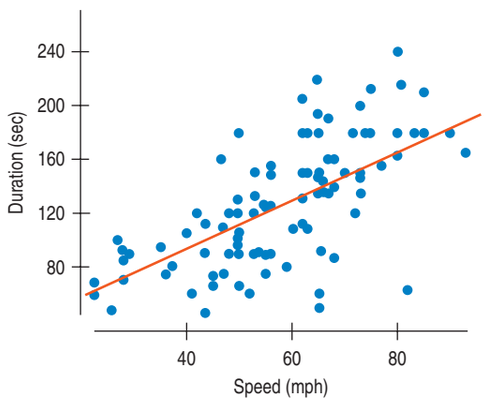
Who	Roller coasters
What	See Table 29.1. (For multiple regression we have to know "What" and the units for each variable.)
Where	Worldwide
When	All were in operation in 2014.
Source	The Roller Coaster DataBase, www.rcdb.com

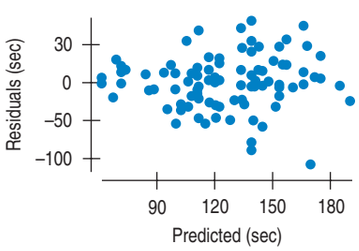
Here are the variables and their units:

- *Type* indicates what kind of track the roller coaster has. The possible values are "wooden" and "steel." (The frame usually is of the same construction as the track, but doesn't have to be.)
- *Duration* is the duration of the ride in seconds.
- *Speed* is top speed in miles per hour.
- *Height* is maximum height above ground level in feet.
- *Drop* is greatest drop in feet.
- *Length* is total length of the track in feet.
- *Inversions* reports whether riders are turned upside down during the ride. It has the values "yes" or "no."

It's always a good idea to explore the data before starting to build a model. Let's first consider the ride's *Duration.* We have that information for only 136 of the 195 coasters in our data set, but there's no reason to believe that the data are missing in any patterned way so we'll look at those 136 coasters. The average *Duration* for these coasters is 124.5 seconds, but one ride is as short as 28 seconds and another as long as 240 seconds. We might wonder whether the duration of the ride should depend on the maximum speed of the ride. Here's the scatterplot of *Duration* against *Speed* and the regression:

Figure 29.1
Duration of the ride appears to be linearly related to the maximum *Speed* of the ride.





Response variable is: Duration
R-squared = 34.5% R-squared (adjusted) = 34.0%
s = 36.36 with 134 − 2 = 132 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	91951.7	1	91951.7	69.6
Residual	174505 1	32	1322.01	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	20.4744	12.93	1.58	0.1156
Speed	1.82262	0.2185	8.34	<0.0001

The regression conditions seem to be met, and the regression makes sense. We'd expect longer tracks to give longer rides. From a base of 20.47 seconds, the duration of the ride increases by about 1.82 seconds per mile per hour of speed—faster coasters actually have rides that last *longer*.

29.1 Indicators

Of course, there's more to these data. One interesting variable might not be one you'd naturally think of. Many modern coasters have "inversions." That's a nice way of saying that they turn riders upside down, with loops, corkscrews, or other devices. These inversions add excitement, but they must be carefully engineered, and that enforces some speed limits on that portion of the ride.

We'd like to add the information of whether the roller coaster has an inversion to our model. Until now, all our predictor variables have been quantitative. Whether or not a roller coaster has any inversions is a categorical variable ("yes" or "no"). Can we introduce the categorical variable *Inversions* as a predictor in our regression model? What would it mean if we did?

Let's start with a plot. Figure 29.2 shows the same scatterplot of duration against speed, but now with the roller coasters that have inversions shown as red x's and a separate regression line drawn for each type of roller coaster.

It's easy to see that, for a given length, the roller coasters with inversions take a bit longer, and that for each type of roller coaster, the slopes of the relationship between duration and length are not quite equal but are similar.

We could split the data into two groups—coasters without inversions and those with inversions—and compute the regression for each group. That would look like this:

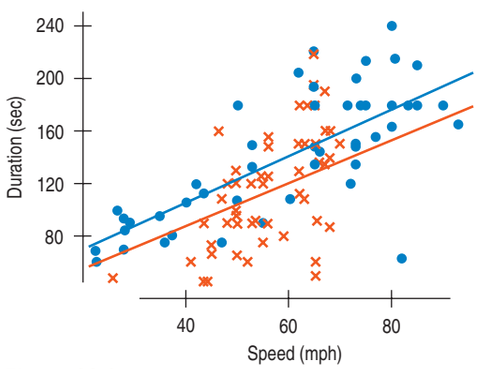


Figure 29.2
The two lines fit to coasters with inversions and without are roughly parallel.

Response variable is: Duration
Cases selected according to no Inversions
R-squared = 57.0% R-squared (adjusted) = 56.2%
s = 31.40 with 55 − 2 = 53 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	33.9448	13.31	2.55	0.0137
Speed	1.79522	0.2142	8.38	<0.0001

Response variable is: Duration
Cases selected according to Inversions
R-squared = 12.8% R-squared (adjusted) = 11.7%
s = 37.61 with 79 − 2 = 77 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	22.7725	27.71	0.822	0.4137
Speed	1.63518	0.4865	3.36	0.0012

As the scatterplot showed, the slopes are very similar, but the intercepts are different.

When we have a situation like this with roughly parallel regressions for each group,[1] there's an easy way to add the group information to a single regression model. We make up a special variable that *indicates* what type of roller coaster we have, giving it the value 1 for roller coasters that have inversions and the value 0 for those that don't. (We could

[1]The fact that the individual regression lines are nearly parallel is really a part of the Straight Enough Condition. You should check that the lines are nearly parallel before using this method. Or read on to see what to do if they are not parallel enough.

have reversed the coding; it's an arbitrary choice.[2]) Such variables are called **indicator variables** or *indicators* because they indicate which of two categories each case is in.[3]

When we add our new indicator, *Inversions*, to the regression model, the model looks like this:

Response variable is: Duration
R-squared = 39.5% R-squared (adjusted) = 38.5%
s = 35.09 with 134 − 3 = 131 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	35.9969	13.34	2.70	0.0079
Speed	1.76038	0.2118	8.31	<0.0001
Inversions	−20.2726	6.187	−3.28	0.0013

This looks like a better model than the simple regression for all the data. The R^2 is larger, the *t*-ratios of both coefficients are large, and the residuals look reasonable. But what does the coefficient for *Inversions* mean?

Let's see how an indicator variable works when we calculate predicted values for two of the roller coasters given at the start of the chapter:

Name	Park	Country	Type	Duration	Speed	Height	Drop	Length	Inversion?
Hangman	Wild Adventures	USA	Steel	125	55	115	95	2170	Yes
Hayabusa	Tokyo SummerLand	Japan	Steel	108	60.3	137.8	124.67	2559.1	No

The model says that for all coasters, the predicted *Duration* is

$$36 + 1.76 \times Speed - 20.2726 \times Inversions$$

For *Hayabusa*, the speed is 55 mph and the value of *Inversions* is 0, so the model predicts a duration of [4]

$$35.9969 + 1.76 \times 55 - 20.2726 \times 0 = 132.79 \text{ seconds}$$

That's not far from the actual duration of 108 seconds.

For the *Hangman*, the speed is 60.3 mph. It has an inversion, so the value of *Inversions* is 1, and the model predicts a duration of

$$35.9969 + 1.76 \times 60.3 - 20.2726 \times 1 = 121.85 \text{ seconds}$$

That compares well with the actual duration of 125 seconds.

Notice how the indicator works in the model. When there is an inversion (as in *Hangman*), the value 1 for the indicator causes the amount of the indicator's coefficient, −20.2726, to be added to the prediction. When there is no inversion (as in *Hayabusa*), the indicator is zero, so nothing is added. Looking back at the scatterplot, we can see that this is exactly what we need. The difference between the two lines is a vertical shift of about 20 seconds.

This may seem a bit confusing at first. We usually think of the coefficients in a multiple regression as slopes. For indicator variables, however, they act differently. They're vertical shifts that keep the slopes for the other variables apart.

[2]Some implementations of indicator variables use −1 and 1 for the levels of the categories.

[3]They are also commonly called *dummies* or *dummy variables*. But this sounds like an insult, so the more politically correct term is *indicator variable.*

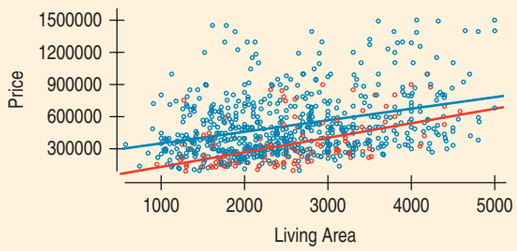
[4]We round coefficient values when we write the model but calculate with the full precision, rounding at the end of the calculation.

For Example USING INDICATOR VARIABLES

As a class project, students in a large Statistics class collected publicly available information on recent home sales in their hometowns. There are 894 properties. These are not a random sample, but they may be representative of home sales during a short period of time, nationwide. In Chapter 28 we looked at these data and constructed a multiple regression model. Let's look further. Among the variables available is an indication of whether the home was in an urban, suburban, or rural setting.

QUESTION: How can we incorporate information such as this in a multiple regression model?

ANSWER: We might suspect that homes in rural communities might differ in price from similar homes in urban or suburban settings. We can define an indicator (dummy) variable to be 1 for homes in rural communities and 0 otherwise. A scatterplot shows that rural homes have, on average, lower prices for a given living area:



The multiple regression model is

Dependent variable is: Price
R-squared = 18.4% R-squared (adjusted) = 18.2%
s = 260996 with 894 − 3 = 891 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	230945	25706	8.98	<0.0001
Living area	112.534	9.353	12.0	<0.0001
Rural	−172359	23749	−7.26	<0.0001

The coefficient of *Rural* indicates that, for a given living area, rural homes sell for on average about $172,000 less than comparable homes in urban or suburban settings.

Adjusting for Different Slopes

What if the lines aren't parallel? An indicator variable that is 0 or 1 can only shift the line up and down. It can't change the slope, so it works only when we have lines with the same slope and different intercepts.

Let's return to the Burger King data we looked at in Chapter 7 and look at how *Calories* are related to *Carbohydrates* (*Carbs* for short). Figure 29.3 shows the scatterplot.

It's not surprising to see that more *Carbs* goes with more *Calories*, but the plot seems to thicken as we move from left to right. Could there be something else going on?[5]

Burger King foods can be divided into two groups: those with meat (including chicken and fish) and those without. When we color the plot (red for meat, blue for non-meat) and look at the regressions for each group, we see a different picture.

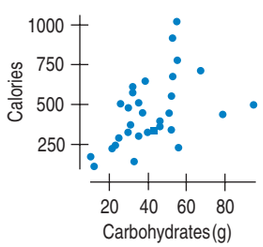


Figure 29.3
Calories of Burger King foods plotted against *Carbohydrates* seems to fan out.

[5]Would we even ask if there weren't?

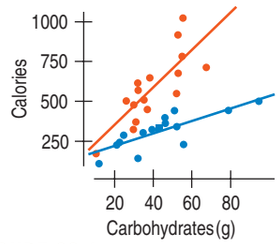


FIGURE 29.4
Plotting the meat-based and non-meat items separately, we see two distinct linear patterns.

Clearly, meat-based dishes have more calories for each gram of carbohydrate than do other Burger King foods. But the regression model can't account for the kind of difference we see here by just including an indicator variable. It isn't just the height of the lines that is different; they have entirely different slopes. How can we deal with that in our regression model?

The trick is to adjust the slopes with another constructed variable. This one is the *product* of an indicator for one group and the predictor variable. The coefficient of this constructed **interaction term** in a multiple regression gives an adjustment to the slope, b_1, to be made for the individuals in the indicated group.[6] Here we have the indicator variable *Meat*, which is 1 for meat-containing foods and 0 for the others. We then construct an interaction variable, *Carbs*Meat,* which is just the product of those two variables. That's right; just multiply them. The resulting variable has the value of *Carbs* for foods containing meat (those coded 1 in the *Meat* indicator) and the value 0 for the others. By including the interaction variable in the model, we can adjust the slope of the line fit to the meat-containing foods. Here's the resulting analysis:

Dependent variable is: Calories
R-squared = 78.1% R-squared (adjusted) = 75.7%
s = 106.0 with 32 − 4 = 28 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	1119979	3	373326	33.2
Residual	314843	28	11244.4	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	137.395	58.72	2.34	0.0267
Carbs(g)	3.93317	1.113	3.53	0.0014
Meat	−26.1567	98.48	−0.266	0.7925
Carbs*Meat	7.87530	2.179	3.61	0.0012

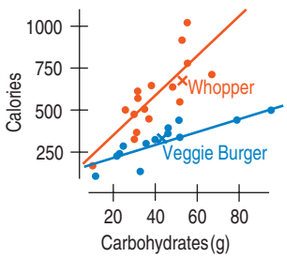


Figure 29.5
The Whopper and Veggie Burger belong to different groups.

What does the coefficient for the indicator *Meat* mean? It provides a different intercept to separate the meat and non-meat items at the origin (where *Carbs* = 0). For these data, there is a different slope, but the two lines nearly meet at the origin, so there seems to be no need for an additional adjustment. The estimated difference of 26.16 calories is small. That's why the coefficient for *Meat* has a small *t*-statistic.

By contrast, the coefficient of the interaction term, *Carbs*Meat*, says that the slope relating calories to carbohydrates is steeper by 7.875 calories per carbohydrate gram for meat-containing foods than for meat-free foods. Its small P-value suggests that this difference is real.

$$137.40 + 3.93\,Carbs - 26.16\,Meat + 7.88\,Carbs*Meat$$

Let's see how these adjustments work. A BK Whopper has 53g of *Carbohydrates* and is a meat dish. The model predicts its *Calories* as

$$137.395 + 3.93317 \times 53 - 26.1567 \times 1 + 7.8753 \times 53 \times 1 = 737.1,$$

not far from the measured calorie count of 680. By contrast, the Veggie Burger, with 43g of *Carbohydrates,* is predicted to have

$$137.395 + 3.93317 \times 43 - 26.1567 \times 0 + 7.87530 \times 0 \times 43 = 306.5 \text{ calories},$$

not far from the 330 measured officially. The last two terms in the equation for the Veggie Burger are just zero because the indicator for *Meat* is 0 for the Veggie Burger.

[6]Chapter 27 discussed interaction effects in two-way ANOVA. Interaction terms such as these are exactly the same idea.

29.2 Diagnosing Regression Models: Looking at the Cases

We often use regression analyses to try to understand the world. By working with the data and creating models, we can learn a great deal about the relationships among variables. As we saw with simple regression, sometimes we can learn as much from the cases that *don't* fit the model as from the bulk of cases that do. Extraordinary cases often tell us more about the world simply by the ways in which they fail to conform and the reasons we can discover for those deviations.

If a case doesn't conform to the others, we should identify it and, if possible, understand why it is different. As in simple regression, a case can be extraordinary by standing away from the model in the y direction or by having unusual values in an x-variable. In multiple regression it can also be extraordinary by having an unusual *combination* of values in the x-variables. Deviations in the y direction show up in the residuals. Deviations in the x's show up as *leverage.*

Leverage

Recent events have focused attention on airport screening of passengers. But screening has a longer history. The *Sourcebook of Criminal Justice Statistics Online* lists the numbers of various violations found by airport screeners for each of several types of violations in each year from 1977 to 2000. Here's a regression of the number of long guns (rifles and the like) found vs. the number of times false information was discovered.

Response variable is: Long guns
R-squared = 3.8% R-squared (adjusted) = −0.6%
s = 56.34 with 24 − 2 = 22 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio
Intercept	87.0071	19.68	4.42
false information	0.246762	0.2657	0.929

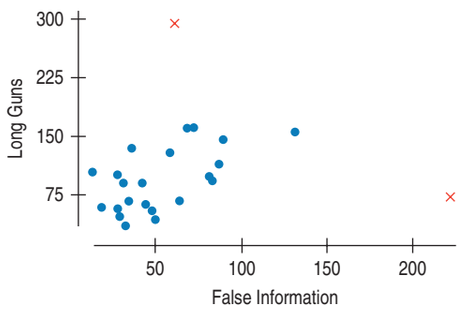


Figure 29.6
A high-leverage point can hide a strong relationship, so that you can't see it in the regression. Make a plot.

That summary doesn't look like it's a particularly successful regression. The R^2 is only 3.8%, and the P-value for *False Info* is large. But a look at the scatterplot tells us more.

The unusual cases are from 1988 and 2000. In 2000, there were nearly 300 long gun confiscations. But because this point is at a typical value for false information, it doesn't have a strong effect on the slope of the regression line. But in 1988, the number of false information reports jumped over 200. The resulting case has high leverage because it is so far from the x-values of the other points. It's easy to see the influence of that one high-leverage case if we look at the regression lines with and without that case (Figure 29.7).

The **leverage** of a case measures its ability to move the regression model all by itself by just moving in the y direction. In Chapter 8, when we had only one predictor variable, we could *see* high leverage points in a scatterplot because they stood far from the mean of x. But now, with several predictors, we can't count on seeing them in our plots.

Fortunately, we can put a number on the leverage. If we keep everything else the same, change the y-value of a case by 1.0, and find a new regression, the leverage of that case is the amount by which the *predicted* value at that case would change. Leverage can never be greater than 1.0—we wouldn't expect the line to move *farther* than we move the case, only to try to keep up. Nor can it be less than 0.0—we'd hardly expect the line to move in the *opposite* direction. A point with zero leverage has no effect at all on the regression model, although it does participate in the calculations of R^2, s, and the F- and t-statistics.

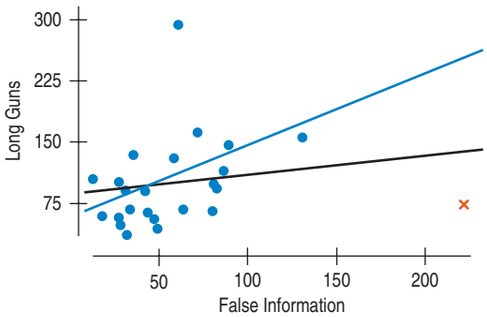


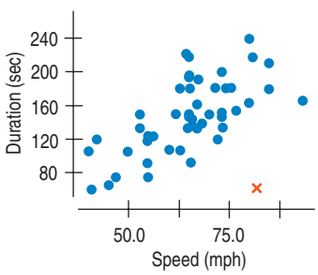
Figure 29.7

A single high-leverage point can change the regression slope quite a bit. The line omitting the point for 1988 is quite different from the line that includes the outlier.

For the airport inspections, the leverage of 1988 is 0.63. That's quite high. If there had been even one fewer long gun discovered that year (decreasing the *observed* y-value by 1), the *predicted* y-value for 1988 would have decreased by 0.63, dragging the regression line down still farther. For comparison, the point for 2000 that has an extraordinary value for Long guns only has leverage 0.42. We would consider it an outlier because it is far from the other values in the y (here, Long gun) direction. But because it is near the mean in x (False information) it doesn't have enough leverage to change the slope of the regression line.

The leverage of a case is a measure of how far that case is from the center of the x's. As always in Statistics, we expect to measure that distance with a ruler based on a standard deviation—here, the standard deviation of the x's. And that's really all the leverage is: an indication of how far each point is away from the center of all the x-values, measured in standard deviations. Fortunately, there's a less tedious way to calculate leverage than moving each case in turn, but it's beyond the scope of this book and you'd never want to do it by hand anyway. So just let the computer do the computing and think about what the result *means*. Most statistics programs calculate leverage values, and you should examine them.

A case can have large leverage in two different ways:



- It might be extraordinary in one or more individual variables. For example, the fastest or slowest roller coaster may stand out.
- It may be extraordinary in a *combination* of variables. For example, one roller coaster stands out in the scatterplot of *Duration* against *Speed*. It isn't extraordinarily fast and others have shorter duration, but the combination of high speed and short duration is unusual. Looking at leverage values can be a very effective way to discover cases that are extraordinary on a combination of x-variables.

There are no tests for whether the leverage of a case is too large. The average leverage value among all cases in a regression is $1/n$, but that doesn't give us much of a guide. One common approach is to just make a histogram of the leverages. Any case whose leverage stands out in a histogram of leverages probably deserves special attention. You may decide to leave the case in the regression or to see how the regression model changes when you delete the case, but you should be aware of its potential to influence the regression.

For Example DIAGNOSING A REGRESSION

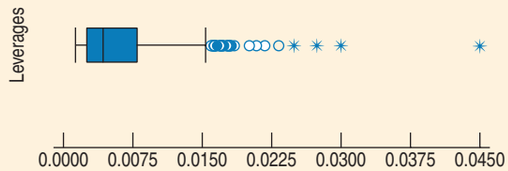
Here's another regression model for the real estate data we looked at in the previous For Example.

Dependent variable is: Price
R-squared = 23.1% R-squared (adjusted) = 22.8%
s = 253709 with 893 − 5 = 888 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	322470	40192	8.02	<0.0001
Living area	92.6272	13.09	7.08	<0.0001
Bedrooms	−69720.6	12764	−5.46	<0.0001
Bathrooms	82577.6	13410	6.16	<0.0001
Rural	−161575	23313	−6.93	<0.0001

QUESTION: What do diagnostic statistics tell us about these data and this model?

ANSWER: A boxplot of the leverage values shows one extraordinarily large leverage:



Investigation of that case reveals it to be a home that sold for $489,900. It has 8 bedrooms and only 2.5 baths. This is a particularly unusual combination, especially for a home with that value. If we were pursuing this analysis further, we'd want to check the records for this house to be sure that the number of bedrooms and bathrooms were recorded accurately.

Residuals and Standardized Residuals

Residuals are not all alike. Consider a point with leverage 1.0. That's the highest a leverage can be, and it means that the line follows the point perfectly. So, a point like that must have a zero residual. And since we know the residual exactly, that residual has zero standard deviation. This tendency is true in general: The larger the leverage, the *smaller* the standard deviation of its residual.[7]

When we want to compare values that have differing standard deviations, it's a good idea to standardize them.[8] We can do that with the regression residuals, dividing each one by an estimate of its own standard deviation. When we do that, the resulting values follow a Student's *t*-distribution. In fact, these standardized residuals are called **Studentized residuals.** It's a good idea to examine the Studentized residuals (rather than the simple residuals) to assess the Nearly Normal Condition and the **Does the Plot Thicken? Condition.** Any Studentized residual that stands out from the others deserves our attention.[9]

It may occur to you that we've always plotted the *unstandardized* residuals when we made regression models. And we've treated them as if they all had the same standard deviation when we checked the **Nearly Normal Condition**. It turns out that this was a simplification. It didn't matter much for simple regression, but for multiple regression models, it's a better idea to use the Studentized residuals when checking the Nearly Normal Condition. (Of course, Student's *t* isn't exactly Normal either—that's why we say "nearly" Normal.)

IT ALL FITS TOGETHER DEPARTMENT

Make an indicator variable for a single case—that is, construct a variable that is 0 everywhere except that it is 1 just for the case in question. When you include that indicator in the regression model, its *t*-ratio will be what that case's externally Studentized residual was in the original model without the indicator. That tells us that an externally Studentized residual can be used to perform a *t*-test of the null hypothesis that a case is *not* an outlier. If we reject that null hypothesis, we can call the point an outlier.[10]

[7]Technically, $SD(e_i) = \sigma\sqrt{1 - h_i}$ where h_i is the leverage of the i-th case, e_i is its residual, and σ is the standard deviation of the regression model errors.

[8]Be cautious when you encounter the term "standardized residual." It is used in different books and by different statistics packages to mean quite different things. Be sure to check the meaning.

[9]There's more than one way to Studentize residuals according to how you estimate σ. You may find statistics packages referring to *externally Studentized residuals* and *internally Studentized residuals*. It is the *externally* Studentized version that follows a *t*-distribution, so those are the ones we recommend.

[10]Finally we have a test to decide whether a case is an outlier. Up until now, all we've had was our judgment based on how the plots looked. But you must still use your common sense and understanding of the data to decide *why* the case is extraordinary and whether it should be corrected or removed from the analysis. That important decision is *still* a judgment call.

Influential Cases

A case that has *both* high leverage and large Studentized residuals is likely to have changed the regression model substantially all by itself. Such a case is said to be **influential.** An influential case cries out for special attention because removing it is likely to give a very different regression model.

The surest way to tell whether a case is influential is to omit it[11] and see how much the regression model changes. You should call a case "influential" if omitting it changes the regression model by enough to matter for *your* purposes. To identify possibly influential cases, check the leverage and Studentized residuals. Two statistics that combine leverage and Studentized residuals into a single measure of influence, Cook's distance and DFFITs, are offered by many statistics programs. If either measure is unusually large for a case, that case should be checked as a possibly influential point.

When a regression analysis has cases that have both high leverage and large Studentized residuals, it would be irresponsible to report only the regression on all the data. You should also compute and discuss the regression found with such cases removed, and discuss the extraordinary cases individually if they offer additional insight. If your interest is to understand the world, the extraordinary cases may well tell you more than the rest of the model. If your only interest is in the model (for example, because you hope to use it for prediction), then you'll want to be certain that the model wasn't determined by only a few influential cases, but instead was built on the broader base of the bulk of your data.

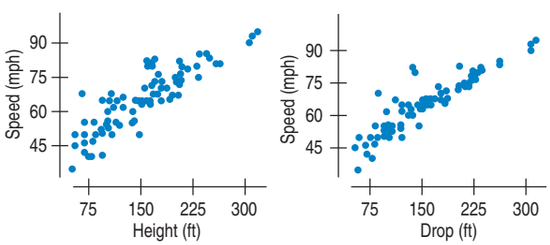
Step-By-Step Example DIAGNOSING A MULTIPLE REGRESSION

Let's consider what makes a roller coaster fast and then diagnose the model to understand more. Roller coasters get their *Speed* from gravity (the "coaster" part), so we'd naturally look to such variables as the *Height* and largest *Drop* as predictors. Let's make and diagnose that multiple regression.

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

I have data on 75 roller coasters that give their top *Speed* (mph), maximum *Height*, and largest *Drop* (both in feet).

Plot



Plan Think about the assumptions and check the conditions.

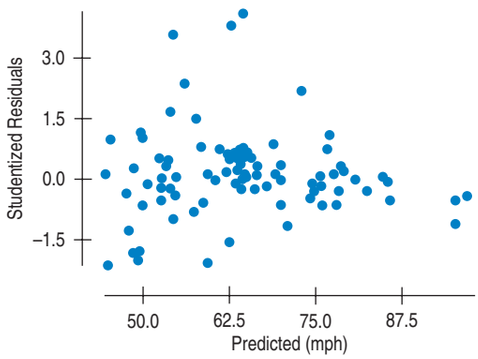
✓ **Straight Enough Condition:** The plots look reasonably straight.

✓ **Independence Assumption:** There are only a few manufacturers of roller coasters worldwide, and coasters made by the same

[11] Or, equivalently, include an indicator variable that selects only for that case.

company may be similar in some respects, but each roller coaster in our data is individualized for its site, so the coasters are likely to be independent.

Because these conditions are met I computed the regression model and found the Studentized residuals.

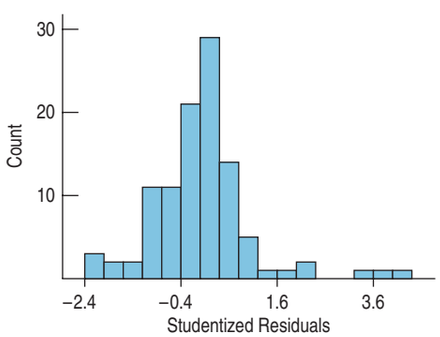


✓ **Straight Enough Condition (2):** The values for one roller coaster don't seem to affect the values for the others in any systematic fashion. This makes the independence assumption more plausible.

✓ **Does the Plot Thicken? Condition:** The scatterplot of Studentized residuals against predicted values shows no obvious changes in the spread about the line. There do seem to be some large residuals that might be outliers.

✓ **Nearly Normal Condition:** A histogram of the Studentized residuals is unimodal and reasonably symmetric, but has three high outliers.

Actually, we need the Nearly Normal Condition only if we want to do inference, but it's hard not to look at the P-values, so we usually check it out. In a multiple regression, it's best to check the Studentized residuals, although the difference is rarely large enough to change our assessment of the normality.



✗ **Outlier Condition:** The histogram suggests that there may be a few large positive residuals. I'll want to look at those.

Choose your method.

Under these conditions, the multiple regression model is appropriate as long as we are cautious about the possible outliers.

SHOW ➡ Mechanics

Here is the computer output for the regression:

Response variable is: Speed
R-squared = 85.7% R-squared (adjusted) = 85.4%
s = 4.789 with 105 − 3 = 102 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	14034.3	2	7017.15	306
Residual	2338.88	102	22.9302	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	34.7035	1.288	27.0	<0.0001
Height	0.050380	0.0185	2.72	0.0077
Drop	0.150264	0.0183	8.20	<0.0001

The estimated regression equation is

$\widehat{Speed} = 34.7 + 0.05\,Height + 0.15\,Drop.$

TELL ➡ Interpretation

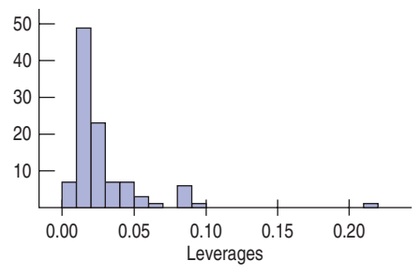
The R^2 for the regression is 85.7%. *Height* and *Drop* account for 86% of the variation in *Speed* in roller coasters like these. Both *Height* and *Drop* contribute significantly to the *Speed* of a roller coaster.

Diagnosis

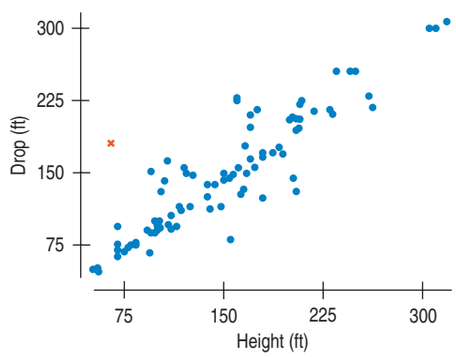
Leverage Most computer regression programs will calculate leverages. There is a leverage value for each case.

It may not be necessary to remove high leverage points from the model, but it's certainly wise to know where they are and, if possible, why they are unusual.

A histogram of the leverages shows one roller coaster with a rather high leverage of more than 0.21.



This high-leverage point is the *Oblivion* coaster in Alton, England. Neither the *Height* nor the *Drop* is extraordinary. To see what's going on, I made a scatterplot of *Drop* against *Height* with *Oblivion* shown as a red x.



Residuals At this point, we might consider recomputing the regression model after removing these three coasters. That's what we do in the next section.

Although *Oblivion's* maximum height is a modest 65 feet, it has a surprisingly long drop of 180 feet. At first, that seemed like an error, but looking it up, I discovered that the unusual feature of the *Oblivion* coaster is that it plunges riders down a deep hole below the ground.

The histogram of the Studentized residuals (above) also nominates some cases for special attention. That bar on the right of the histogram holds three roller coasters with large positive residuals: the *Xcelerator, Hypersonic XCL,* and *Volcano, the Blast Coaster.* New technologies such as hydraulics or compressed air are used to launch all three roller coasters. These three coasters are different in that their speed doesn't depend only on gravity.

Diagnosis Wrap-Up

The *Oblivion* roller coaster plunges into a hole in the ground.

What have we learned from diagnosing the regression? We've discovered four roller coasters that may be influencing the model. And for each of them, we've been able to understand why and how they differed from the others. The oddness of *Oblivion* in plunging into a hole in the ground may cause us to prefer *Drop* as a predictor of speed rather than *Height*.

The three influential cases turned out to be different from the other roller coasters because they are "blast coasters" that don't rely only on gravity for their acceleration. Although we can't count on always discovering why influential cases are special, diagnosing influential cases raises the question of what about them might be different. Understanding influential cases can help us understand our data better.

When there are influential cases, we always want to consider the regression model without them:

Response variable is: Speed
R-squared = 91.5% R-squared (adjusted) = 91.3%
s = 3.683 with 102 − 3 = 99 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	34.3993	0.9926	34.7	<0.0001
Drop	0.198348	0.0153	13.0	<0.0001
Height	0.00183	0.0155	0.118	0.9062

Without the three blast coasters, *Height* no longer appears to be important in the model, so we might try omitting it:

Response variable is: Speed
R-squared = 91.5% R-squared (adjusted) = 91.4%
s = 3.664 with 102 − 2 = 100 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	34.4285	0.9567	36.0	<0.0001
Drop	0.200004	0.0061	32.7	<0.0001

That looks like a good model. It seems that our diagnosis has led us back to a simple regression.

INDICATORS FOR INFLUENCE

One good way to examine the effect of an extraordinary case on a regression is to construct a special indicator variable that is zero for all cases *except* the one we want to isolate. Including such an indicator in the regression model has the same effect as removing the case from the data, but it has two special advantages. First, it makes it clear to anyone looking at the regression model that we have treated that case specially. Second, the *t*-statistic for the indicator variable's coefficient can be used as a test of whether the case is influential. If the P-value is small, then that case really didn't fit well with the rest of the data. Typically, we name such an indicator with the identifier of the case we want to remove. Here's the last roller coaster model in which we have removed the influence of the three blast coasters by constructing indicators for them instead of by removing them from the data. Notice that the coefficients for the other predictors are just the same as the ones we found by omitting the cases.

Response variable is: Speed
R-squared = 91.8% R-squared (adjusted) = 91.5%
s = 3.664 with 105 − 5 = 100 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	34.4285	0.9567	36.0	<0.0001
Drop	0.200004	0.0061	32.7	<0.0001
Xcelerator	21.5711	3.684	5.86	<0.0001
HyperSonic	18.9711	3.683	5.15	<0.0001
Volcano	19.5713	3.704	5.28	<0.0001

The P-values for the three indicator variables confirm that each of these roller coasters doesn't fit with the others.

29.3 Building Multiple Regression Models

"It is the mark of an educated mind to be able to entertain a thought without accepting it."

—Aristotle

When many possible predictors are available, we will naturally want to select only a few of them for a regression model. But which ones? The first and most important thing to realize is that often there is no such thing as the "best" regression model. (After all, all models are wrong.) Several alternative models may be useful or insightful. The "best" for one purpose may not be best for another. And the one with the highest R^2 may well not be best for many purposes. There is nothing wrong with continuing to work with several models without choosing among them.

Multiple regressions are subtle. The coefficients often don't mean what they at first appear to mean. The choice of which predictors to use determines almost everything about the regression.

Predictors interact with each other, which complicates interpretation and understanding. So it is usually best to build a *parsimonious* model, using as few predictors as you can. On the other hand, we don't want to leave out predictors that are theoretically or practically important. Making this trade-off is the heart of the challenge of selecting a good model. The best regression models, in addition to satisfying the assumptions and conditions of multiple regression, have:

- Relatively few predictors to keep the model simple.
- A relatively high R^2 indicating that much of the variability in y is accounted for by the regression model.
- A relatively small value of s, the standard deviation of the residuals, indicating that the magnitude of the errors is small.

- Relatively small P-values for their *F*- and *t*-statistics, showing that the overall model is better than a simple summary with the mean, and that the individual coefficients are reliably different from zero.
- No cases with extraordinarily high leverage that might dominate and alter the model.
- No cases with extraordinarily large residuals, and Studentized residuals that appear to be Nearly Normal. Outliers can alter the model and certainly weaken the power of any test statistics. And the Nearly Normal Condition is required for inference.
- Predictors that are reliably measured and relatively unrelated to each other.

The term "relatively" in this list is meant to suggest that you should favor models with these attributes over others that satisfy them less, but, of course, there are many trade-offs and no absolute rules.

Cases with high leverage or large residuals can be dealt with by introducing indicator variables.

In addition to favoring predictors that can be measured reliably, you may want to favor those that are less expensive to measure, especially if your model is intended for prediction with values not yet measured.

Seeking Multiple Regression Models Automatically

How can we find the best multiple regression model? The list of desirable features we just looked at should make it clear that there is no simple definition of the "best" model. A computer can try all combinations of the predictors to find the regression model with the highest R^2, or optimize some other criterion,[12] but models found that way are not best for all purposes, and may not even be particularly good for many purposes.

Another alternative is to have the computer build a regression "stepwise." In a **stepwise regression,** at each step, a predictor is either added to or removed from the model. The predictor chosen to add is the one whose addition increases the R^2 the most (or similarly improves some other measure). The predictor chosen to remove is the one whose removal reduces the R^2 least (or similarly loses the least on some other measure). The hope is that by following this path, the computer can settle on a good model. The model will gain or lose a predictor only if that change in the model makes a big enough improvement in the performance measure. The changes stop when no more changes pass the criterion.

STEPPING IN THE WRONG DIRECTION

Here's an example of how stepwise regression can go astray. We might want to find a regression to model *Horsepower* in a sample of cars from the car's engine size (*Displacement*) and its *Weight*. The simple correlations are as follows:

	HP	Disp	Wt
Horsepower	1.000		
Displacement	0.872	1.000	
Weight	0.917	0.951	1.000

Because *Weight* has a slightly higher correlation with *Horsepower*, stepwise regression will choose it first. Then, because *Weight* and engine size (*Displacement*) are so highly correlated, once *Weight* is in the model, *Displacement* won't be added to the model. But *Weight* is, at best, a lurking variable leading to both the need for more horsepower and a larger engine. Don't try to tell an engineer that the best way to increase horsepower is to add weight to the car and that the engine size isn't important! From an engineering standpoint, *Displacement* is a far more appropriate predictor of *Horsepower*, but stepwise regression for these data doesn't find that model.

[12]This is literally true. Even for many variables and a moderately large number of cases, it is computationally possible to find the "best subset" of predictors that maximizes R^2. Many statistics programs offer this capability.

Stepwise methods can be valuable when there are hundreds or thousands of potential predictors, as can happen in data mining applications. They can build models that are useful for prediction or as starting points in the search for better models. Because they do each step *automatically*, however, stepwise methods are inevitably affected by influential points and nonlinear relationships. A better strategy might be to mimic the stepwise procedure yourself, but more carefully. You could consider adding or removing a variable yourself with a careful look at the assumptions and conditions each time a variable is considered. That kind of guided stepwise method is still not guaranteed to find a good model, but it may be a sensible way to search among the potential candidates.

Building Regression Models Sequentially

You can build a regression model by adding variables to a growing regression. Each time you add a predictor, you hope to account for a little more of the variation in the response. What's left over is the residuals. At each step, consider the predictors still available to you. Those that are most highly correlated with the current residuals are the ones that are most likely to improve the model.

If you see a variable with a high correlation at this stage and it is *not* among those that you thought were important, stop and think about it. Is it correlated with another predictor or with several other predictors? Don't let a variable that doesn't make sense enter the model just because it has a high correlation, but at the same time, don't exclude a predictor just because you didn't initially think it was important. (That would be a good way to make sure that you never learn anything new.) Finding the balance between these two choices underlies the art of successful model building.

Alternatively, you can start with all available predictors in the model and remove those with small *t*-ratios. At each step make a plot of the residuals to check for outliers, and check the leverages (say, with a histogram of the leverage values) to be sure there are no high-leverage points. Influential cases can strongly affect which variables appear to be good or poor predictors in the model. It's also a good idea to check that a predictor doesn't appear to be unimportant in the model only because it's correlated with other predictors in the model. It may (as is true of *Displacement* in the example of predicting *Horsepower*) actually be a more useful or meaningful predictor than some of those in the model.

In either method, adding or removing a predictor will usually change *all* of the coefficients, sometimes by quite a bit.

Step-By-Step Example BUILDING MULTIPLE REGRESSION MODELS

Let's return to the Kids Count infant mortality data. In Chapter 28, we fit a large multiple regression model in which several of the t-ratios for coefficients were too small to be discernibly different from zero. Maybe we can build a more parsimonious model. Which model should we build?

The most important thing to do is to think about the data. Regression models can and should make sense. Many factors can influence your choice of a model, including the cost of measuring particular predictors, the reliability or possible biases in some predictors, and even the political costs or advantages to selecting predictors.

Think ➡ Variables Name the available variables, report the W's, and specify the question of interest or the purpose of finding the regression model.

I have data on the 50 states. The available variables are (all for 1999):

Infant Mortality (deaths per 1000 live births)

Low Birth Weight (Low BW%—%babies with low birth weight)

Child Deaths (deaths per 100,000 children ages 1–14)

%Poverty (percent of children in *poverty* in the previous year)

HS Drop% (percent of teens who are high school *dropouts*, ages 16–19)

Teen Births (births per 100,000 females ages 15–17)

Teen Deaths (by accident, homicide, and suicide; deaths per 100,000 teens ages 15–19)

We've examined a scatterplot matrix and the regression with all potential predictors in Chapter 28.

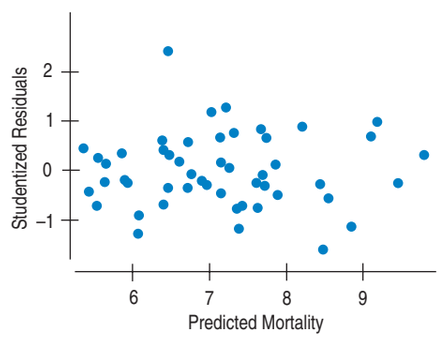
I hope to gain a better understanding of factors that affect infant mortality.

Plan Think about the assumptions and check the conditions.

✓ **Straight Enough Condition:** The scatterplot matrix shows no bends, clumping, or outliers in any of the scatterplots.

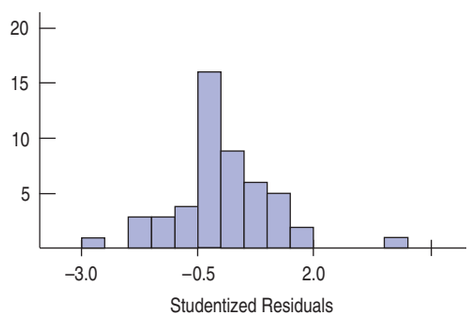
✓ **Independence Assumption:** These data are based on random samples.

With this assumption and condition satisfied, I can compute the regression model and find residuals.



Remember that in a multiple regression, rather than plotting residuals against each of the predictors, we usually plot Studentized residuals against the predicted values.

✓ **Does the Plot Thicken? Condition:** This scatterplot of Studentized residuals vs. predicted values for the full model (all predictors) shows no obvious trends in the spread.



✘ ***Nearly Normal Condition, Outlier Condition:*** A histogram of the Studentized residuals from the full model is unimodal and symmetric, but it seems to have an outlier. The unusual state is South Dakota. I'll test whether it really is an outlier by making an indicator variable for South Dakota and including it in the predictors.

Show ➡ Mechanics

Multiple regressions are always found from a computer program.

I'll start with the full regression and work backward:

Dependent variable is: Infant mort
R-squared = 78.7% R-squared (adjusted) = 75.2%
s = 0.6627 with 50 − 8 = 42 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	1.31183	0.8639	1.52	0.1364
Low BW%	0.73272	0.1067	6.87	<0.0001
Child Deaths	0.02857	0.0123	2.31	0.0256
%Poverty	−5.3026e-3	0.0332	−0.160	0.8737
HS Drop%	−0.10754	0.0540	−1.99	0.0531
Teen Births	0.02402	0.0234	1.03	0.3111
Teen Deaths	−1.5516e-4	0.0101	−0.015	0.9878
S. Dakota	2.74813	0.7175	3.83	0.0004

For model building, look at the P-values only as general indicators of how much a predictor contributes to the model.

The coefficient for the S. Dakota indicator variable has a very small P-value, so that case is an outlier in this regression model. *Teen Births, Teen Deaths, and %Poverty* have large P-values and look like they are less successful predictors in this model.

You shouldn't remove more than one predictor at a time from the model because each predictor can influence how the others contribute to the model. If removing a predictor from the model doesn't change the remaining coefficients very much (or reduce the R^2 by very much), that predictor wasn't contributing very much to the model.

I'll remove *Teen Deaths* first:

Dependent variable is: Infant mort
R-squared = 78.7% R-squared (adjusted) = 75.7%
s = 0.6549 with 50 − 7 = 43 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	1.30595	0.7652	1.71	0.0951
Low BW%	0.73283	0.1052	6.97	<0.0001
Child Deaths	0.02844	0.0085	3.34	0.0018
%Poverty	−5.3548e-3	0.0326	−0.164	0.8703
HS Drop%	−0.10749	0.0533	−2.02	0.0501
Teen Births	0.02402	0.0231	1.04	0.3053
S. Dakota	2.74651	0.7014	3.92	0.0003

Removing *Teen Births* and *%Poverty*, in turn, gives this model:

Dependent variable is: Infant mort
R-squared = 78.1% R-squared (adjusted) = 76.2%
s = 0.6489 with 50 − 5 = 45 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	1.03782	0.6512	1.59	0.1180
Low BW%	0.78334	0.0934	8.38	<0.0001
Child Deaths	0.03104	0.0075	4.12	0.0002
HS Drop%	−0.06732	0.0381	−1.77	0.0837
S. Dakota	2.66150	0.6899	3.86	0.0004

Adjusted R^2 can increase when you remove a predictor if that predictor wasn't contributing very much to the regression model.

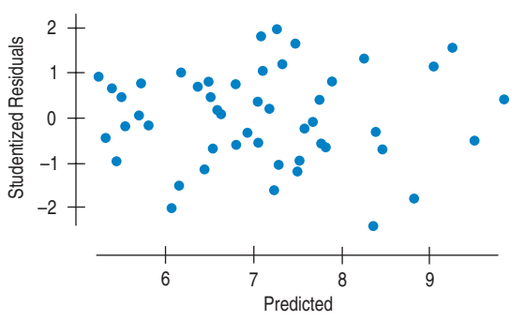
Compared with the full model, the R^2 has come down only very slightly, and the adjusted R^2 has actually increased. The P-value for HS Drop% is bigger than the standard .05 level, but more to the point, Child Deaths and Low Birth Weight are both variables that look at birth and early childhood. HS Drop% seems not to belong with them. When I take that variable out, the model looks like this:

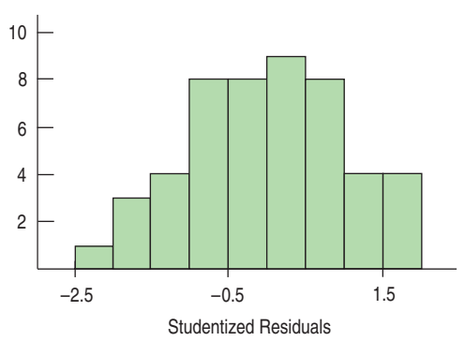
Dependent variable is: Infant mort
R-squared = 76.6% R-squared (adjusted) = 75.1%
s = 0.6638 with 50 − 4 = 46 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	0.760145	0.6465	1.18	0.2457
Child Deaths	0.026988	0.0073	3.67	0.0006
Low BW%	0.750461	0.0937	8.01	<0.0001
S.Dakota	2.74057	0.7042	3.89	0.0003

This looks like a good model. It has a reasonably high R^2 and small P-values for each of the coefficients.

Before deciding that any regression model is a "keeper," remember to check the residuals.





TELL Summarize the features of this model.

Here's an example of an outlier that might help us learn something about the data or the world. Whatever makes South Dakota's infant mortality rate so much higher than the model predicts, it might be something we could address with new policies or interventions.

The scatterplot of Studentized residuals against predicted values shows no structure, and the histogram of Studentized residuals is Nearly Normal. So this looks like a good model for Infant Mortality. The coefficient for S. Dakota is still very significant, so I'd prefer to keep South Dakota separate and look into why its Infant Mortality rate is so much higher (2.74 deaths per 1000 live births) than we would otherwise expect from its Child Death Rate and Low Birth Weight percent.

SHOW Let's try the other way and build a regression model "forward" by selecting variables to add to the model.

One way to select variables to add to a growing regression model is to find the correlation of the residuals of the current state of the model with the potential new predictors. Predictors with higher correlations can be expected to account for more of the remaining residual variation if we include them in the regression model.

Notice that adding a predictor that does not contribute to the model can reduce the adjusted R^2.

The data include variables that concern young adults: *Teen Births*, *Teen Deaths*, and the *HS Drop%*.

Both *Teen Births* and *Teen Deaths* are promising predictors, but births to teens seem more directly relevant. Here's the regression model:

Dependent variable is: Infant mort
R-squared = 29.3% R-squared (adjusted) = 27.9%
s = 1.129 with 50 − 2 = 48 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	4.96399	0.5098	9.74	<0.0001
Teen Births	0.081217	0.0182	4.47	<0.0001

The correlations of the residuals with other predictors look like this:

	Resids
HS Drop%	−0.188
Teen Deaths	0.333
%Poverty	0.105

Teen Deaths looks like a good choice to add to the model:

Dependent variable is: Infant mort
R-squared = 39.1% R-squared (adjusted) = 36.5%
s = 1.059 with 50 − 3 = 47 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	3.98643	0.5960	6.69	<0.0001
Teen Births	0.057880	0.0191	3.04	0.0039
Teen Deaths	0.028228	0.0103	2.75	0.0085

Finally, I'll try adding *HS Drop%* to the model:

Dependent variable is: Infant mort
R-squared = 44.0% R-squared (adjusted) = 40.4%
s = 1.027 with 50 − 4 = 46 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	4.51922	0.6358	7.11	<0.0001
Teen Births	0.097855	0.0272	3.60	0.0008
Teen Deaths	0.026844	0.0100	2.69	0.0099
HS Drop%	−0.164347	0.0819	−2.01	0.0506

Here is one more step, adding *%Poverty* to the model:

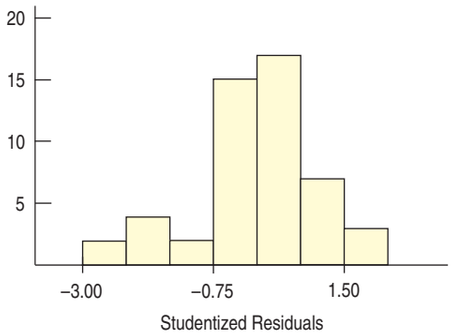
Dependent variable is: Infant mort
R-squared = 44.0% R-squared (adjusted) = 39.1%
s = 1.038 with 50 − 5 = 45 degrees of freedom

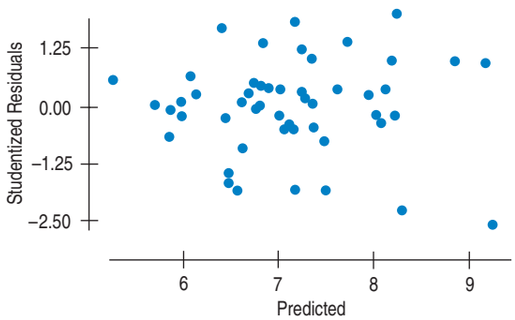
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	4.49810	0.7314	6.15	<0.0001
Teen Births	0.09690	0.0317	3.06	0.0038
Teen Deaths	0.02664	0.0106	2.50	0.0160
HS Drop%	−0.16397	0.0830	−1.98	0.0544
%Poverty	3.1053e-3	0.0513	0.061	0.9520

The P-value for *%Poverty* is quite high, so I prefer the previous model.

The regression that models *Infant Mortality* on *Teen Births, Teen Deaths,* and *HS Drop%* may be worth keeping as well. But, of course, we're not finished until we check the residuals:

Here are the residuals:





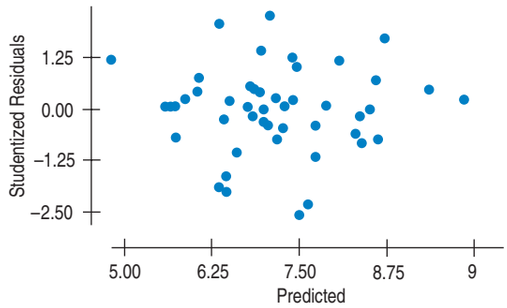
This histogram hints of a low mode holding some large negative residuals, and the scatterplot shows two in particular that trail off at the bottom right corner of the plot. They are Texas and New Mexico. These states are neighbors and may share some regional attributes. To be careful, I'll try removing them from the model. I'll construct two indicator variables that are 1 for the named state and 0 for all others:

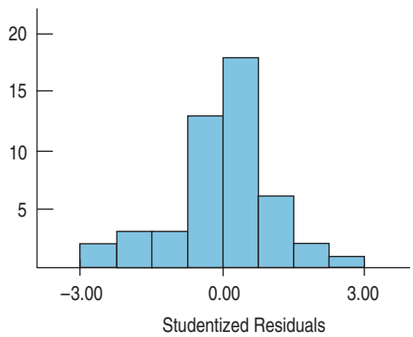
Dependent variable is: Infant mort
R-squared = 58.9% R-squared (adjusted) = 54.2%
s = 0.8997 with 50 − 6 = 44 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	4.15748	0.5673	7.33	<0.0001
Teen Births	0.13823	0.0259	5.33	<0.0001
Teen Deaths	0.02669	0.0090	2.97	0.0048
HS Drop%	−0.22808	0.0735	−3.10	0.0033
New Mexico	−3.01412	0.9755	−3.09	0.0035
Texas	−2.74363	0.9748	−2.81	0.0073

Removing the two outlying states has improved the model noticeably. The indicators for both states have small P-values, so I conclude that they were in fact outliers for this model. The R^2 has improved to 58.9%, and the P-values of all the other coefficients have been reduced.

A final check on the residuals from this model shows that they satisfy the regression conditions:





This model is an alternative to the first one I found. It has a smaller R^2 (58.9%) and larger s value, but it might be useful for understanding the relationships between these variables and infant mortality.

TELL Compare and contrast the models.

For a more complete understanding of infant mortality, we should look into South Dakota's early childhood variables and the teen-related variables in New Mexico and Texas. We might well learn as much about infant mortality by understanding why these states stand out—and how they differ from each other—as we would from the regression models themselves.

I have found two reasonable regression models for infant mortality. The first finds that *Infant Mortality* can be modeled by *Child Deaths* and *%Low Birth Weight*, removing the influence of South Dakota:

$$\widehat{\textit{Infant Mortality}} = 0.76 + 0.027\ \textit{Child Deaths} + 0.75\ \textit{LowBW\%}.$$

It may be worthwhile to look into why South Dakota is so different from the other states. The other model focused on teen behavior, modeling *Infant Mortality* by *Teen Births*, *Teen Deaths*, and *HS Drop%*, removing the influence of Texas and New Mexico:

$$\widehat{\textit{Infant Mortality}} = 4.16 + 0.138\ \textit{Teen Births} + 0.027\ \textit{Teen Deaths} - 0.228\ \textit{HS Drop\%}$$

The coefficient of *HS Drop%* is the opposite sign of the simple relationship between *Infant Deaths* and *HS Drop%*.

Each model has nominated different states as outliers. For a more complete understanding of infant mortality, it might be worthwhile to look into why these states are outliers in these models.

Which model is better? That depends on what you want to know. Remember—all models are wrong. But both may offer useful information and insights about infant mortality and its relationship with other variables and about the states that stood out and why they differ from the others.

Just Checking

1. Give two ways that we use histograms to support the construction, inference, and understanding of multiple regression models.
2. Give two ways that we use scatterplots to support the construction, inference, and understanding of multiple regression models.
3. What role does the Normal model play in the construction, inference, and understanding of multiple regression models?

Regression Roles

We build regression models for a number of reasons. One reason is to model how variables are related to each other in the hope of understanding the relationships. Another is to build a model that might be used to predict values for a response variable when given values for the predictor variables.

When we hope to understand, we are often particularly interested in simple, straightforward models in which predictors are as unrelated to each other as possible. We are especially happy when the *t*-statistics are large, indicating that the predictors each contribute to the model. We are likely to want to look at partial regression plots to understand the coefficients and to check that no outliers or influential points are affecting them.

When prediction is our goal, we are more likely to care about the overall R^2. Good prediction occurs when much of the variability in y is accounted for by the model. We might be willing to keep variables in our model that have relatively small *t*-statistics simply for the stability that having several predictors can provide. We care less whether the predictors are related to each other because we don't intend to interpret the coefficients anyway.

In both roles, we may include some predictors to "get them out of the way." Regression offers a way to approximately control for factors when we have observational data because each coefficient measures effects *after removing the effects* of the other predictors. Of course, it would be better to control for factors in a randomized experiment, but often that's just not possible.

*Indicators for Three or More Levels

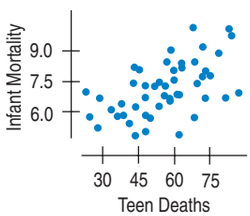
It's easy to construct indicators for a variable with two levels; we just assign 0 to one level and 1 to the other. But variables like *Month* or *Class* often have several levels. You can construct indicators for a categorical variable with several levels by constructing a separate indicator for each of these levels. There's just one trick: You have to choose one of the categories as a "baseline" and *leave out* its indicator. Then the coefficients of the other indicators can be interpreted as the amount by which their categories differ from the baseline, after allowing for the linear effects of the other variables in the model.[13]

[13]There are alternative coding schemes that compare all the levels with the mean. Make sure you know how the indicators are coded.

Make sure your collection of indicators doesn't exhaust all the categories. One category must be left out to serve as a baseline or the regression model can't be found. For the two-category variable *Inversions*, we used "no inversion" as the baseline and coasters with an inversion got a 1. We needed only one variable for two levels. If we wished to represent *Month* with indicators, we would need 11 of them. We might, for example, define *January* as the baseline, and make indicators for *February, March, … , November*, and *December.* Each of these indicators would be 0 for all cases except for the ones that had that value for the variable *Month.* Why not just a single variable with "1" for *January,* "2" for *February,* and so on? That might work. But it would impose the pretty strict assumption that the responses to the months are ordered and equally spaced—that is, that the change in our response variable from January to February is the same in both direction and amount as the change from July to August. That's a pretty severe restriction and may not be true for many kinds of data. Using 11 indicators releases the model from that restriction, but, of course, at the expense of having 10 fewer degrees of freedom for all of our *t*-tests.

Collinearity

Let's look at the infant mortality data one more time. One good predictor of *Infant Mortality* is *Teen Deaths.*



Dependent variable is: Infant mort
R-squared = 27.2% R-squared (adjusted) = 25.7%
s = 1.146 with 50 − 2 = 48 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	4.73979	0.5866	8.08	<0.0001
Teen Deaths	0.042129	0.0100	4.23	0.0001

Teen Deaths has a positive coefficient (as we might expect) and a very small P-value. Suppose we now add *Child Deaths Rate (CDR)* to the regression model:

Dependent variable is: Infant mort
R-squared = 42.6% R-squared (adjusted) = 40.1%
s = 1.029 with 50 − 3 = 47 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	5.79561	0.6049	9.58	<0.0001
Teen Deaths	−1.86877e-3	0.0153	−0.122	0.9032
Child Deaths	0.059398	0.0168	3.55	0.0009

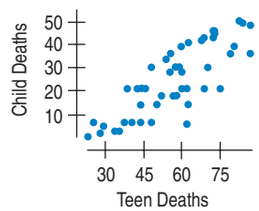


Figure 29.8
Child Deaths and *Teen Deaths* are linearly related.

Suddenly *Teen Deaths* has a small negative coefficient and a very large P-value. What happened? The problem is that *Teen Deaths* and *Child Deaths* are closely associated. The coefficient of *Teen Deaths* now reports how *Infant Mortality* is related to *Teen Deaths after allowing for the linear effects of Child Deaths* on both variables.

When we have several predictors, we must think about how the predictors are related to each other. When predictors are unrelated to each other, each provides new information to help account for more of the variation in *y.* Just as we need a predictor to have a large enough variability to provide a stable base for simple regression, when we have several predictors, we need for them to vary in different directions for the multiple regression to have a stable base. If you wanted to build a deck on the back of your house, you wouldn't build it with supports placed just along one diagonal. Instead, you'd want the supports spread out in different directions as much as possible to make the deck stable. We're in a similar situation with multiple regression. When predictors are highly correlated, they line up together, which makes the regression they support balance precariously. Even small variations can rock it one way or the other. A more stable model can be built when predictors have low correlation and the points are spread out.

MULTI-COLLINEARITY?

You may find this problem referred to as "multicollinearity." But there is no such thing as "unicollinearity"—we need at least two predictors for there to be a linear association between them—so there is no need for the extra two syllables.

WHY NOT JUST LOOK AT THE CORRELATIONS?

It's sometimes suggested that we examine the table of correlations of all the predictors to search for collinearity. But this will find only associations among *pairs* of predictors. Collinearity can—and does—occur among *several* predictors working together. You won't find that more subtle collinearity with a correlation table.

When two or more predictors are linearly related, they are said to be **collinear.** The general problem of predictors with close (but perhaps not perfect) linear relationships is called the problem of **collinearity.**

Fortunately, there's an easy way to assess collinearity. To measure how much one predictor is linearly related to the others, just find the regression of that predictor on the others[14] and look at the R^2. That R^2 gives the fraction of the variability of the predictor in question that is accounted for by the other predictors. So $1 - R^2$ is the amount of the predictor's variance that is left after we allow for the effects of the other predictors. That's what the predictor has left to bring to the regression model. And we know that a predictor with little variance can't do a good job of predicting.[15]

Collinearity can hurt our analysis in yet another way. We've seen that the variance of a predictor plays a role in the standard error of its associated coefficient. Small variance leads to a larger SE. In fact, it's exactly this leftover variance that shows up in the formula for the SE of the coefficient. That's what happened in the infant mortality example.

As a final blow, when a predictor is collinear with the other predictors, it's often difficult to figure out what its coefficient means in the multiple regression. We've blithely talked about "removing the effects of the other predictors," but now when we do that, there may not be much left. What is left is not likely to be about the original predictor, but more about the fractional part of that predictor not associated with the others. In a regression of horsepower on weight and engine size, once we've removed the effect of weight on horsepower, engine size doesn't tell us anything *more* about horsepower. That's certainly not the same as saying that engine size doesn't tell us anything about horsepower. It's just that most cars with big engines also weigh a lot.

When a predictor is collinear with the other predictors in the model, two things can happen:

1. Its coefficient can be surprising, taking on an unanticipated sign or being unexpectedly large or small.
2. The standard error of its coefficient can be large, leading to a smaller t-statistic and correspondingly large P-value.

One telltale sign of collinearity is the paradoxical situation in which the overall F-test for the multiple regression model is significant, showing that at least one of the coefficients is discernably different from zero, and yet most or all of the individual coefficients have small t-values, each in effect, denying that *it* is the significant one.

What should you do about a collinear regression model? The simplest cure is to remove some of the predictors. That both simplifies the model and generally improves the t-statistics. And, if several predictors give pretty much the same information, removing some of them won't hurt the model. Which should you remove? Keep the predictors that are most reliably measured, least expensive to find, or even those that are politically important.

CHOOSING A SENSIBLE MODEL

The mathematics department at a large university built a regression model to help them predict success in graduate study. They were shocked when the coefficient for Mathematics GRE score was not significant. But the Math GRE was collinear with some of the other predictors, such as math course GPA and Verbal GRE, which made its slope not significant. They decided to omit some of the other predictors and retain Math GRE as a predictor because that model seemed more appropriate—even though it predicted no better (and no worse) than others without Math GRE.

[14]The residuals from this regression are plotted as the x-axis of the partial regression plot for this variable. So if they have a very small variance, you can see it by looking at the x-axis labels of the partial regression plot, and get a sense of how precarious a line fit to the partial regression plot—and its corresponding multiple regression coefficient—may be.

[15]The statistic $1 - R^2$ for the R^2 found from the regression of one predictor on the other predictors in the model is also called the *Variance Inflation Factor,* or *VIF,* in some computer programs and books.

WHAT CAN GO WRONG?

In the Oscar-winning movie *The Bridge on the River Kwai* and in the book on which it is based,[16] the character Colonel Green famously says, "As I've told you before, in a job like yours, even when it's finished, there's always one more thing to do." It is wise to keep Colonel Green's advice in mind when building, analyzing, and understanding multiple regression models.

- ***Beware of collinearities.*** When the predictors are linearly related to each other, they add little to the regression model after allowing for the contributions of the other predictors. Check the R^2s when each predictor is regressed on the others. If these are high, consider omitting some of the predictors.
- ***Don't check for collinearity only by looking at pairwise correlations.*** Collinearity is a relationship among any number of the predictors. Pairwise correlations can't always show that. (Of course, a high pairwise correlation between two predictors does indicate collinearity of a special kind.)
- ***Don't be fooled when high-influence points and collinearity show up together.*** A single high-influence point can be the difference between your predictors being collinear and seeming not to be collinear. (Picture that deck supported only along its diagonal and with a single additional post in another corner. Supported in this way, the deck is stable, but the height of that single post completely determines the tilt of the deck, so it's very influential.) Removing a high-influence point may surprise you with unexpected collinearity. Alternatively, a single value that is extreme on several predictors can make them appear to be collinear when in fact they would not be if you removed that point. Removing that point may make apparent collinearities disappear (and would probably result in a more useful regression model).
- ***Beware missing data.*** Values may be missing or unavailable for any case in any variable. In simple regression, when the cases are missing for reasons that are unrelated to the variable we're trying to predict, that's not a problem. We just analyze the cases for which we have data. But when several variables participate in a multiple regression, any case with data missing on any of the variables will be omitted from the analysis. You can unexpectedly find yourself with a much smaller set of data than you started with. Be especially careful, when comparing regression models with different predictors, that the cases participating in the models are the same.
- ***Remember linearity.*** The **Linearity Assumption** (and the **Straight Enough Condition**) require linear relationships among the variables in a regression model. As you build and compare regression models, be sure to plot the data to check that it is straight. Violations of this assumption make everything else about a regression model invalid.
- ***Check for parallel regression lines.*** When you introduce an indicator variable for a category, check the underlying assumption that the other coefficients in the model are essentially the same for both groups. If not, consider adding an interaction term.

CONNECTIONS

Now that we understand indicator variables, we can see that multiple regression and ANOVA are really the same analysis. If the only predictor in a regression is an indicator variable that is 1 for one group and 0 for the other, the t-test for its coefficient is just the pooled t-test for the difference in the means of those groups. In fact, most of the Student's t–based methods in this book can be seen as part of a more general statistical model known as the General Linear Model (GLM). That accounts for why they seem to be so connected, using the same general ideas and approaches.[17] We've generalized the concept of leverage that we first saw in Chapter 8. Everything we said about how to think about these ideas back in Chapters 8 and 25 still applies to the multiple regression model.

Don't forget that the Straight Enough Condition is essential to all of regression. At any stage in developing a model, if the scatterplot that you check is not straight, consider re-expressing the variables to make the relationship straighter. The topics of Chapter 9 will help you with that.

[16]The author of the book, Pierre Boulle, also wrote the book and script for *Planet of the Apes.* The director, David Lean, also directed *Lawrence of Arabia.*

[17]It has been wistfully observed that if only we could start the course by teaching multiple regression, everything else would just be simplifications of the general method. Now that you're here, you might try reading the book backward, contradicting the White King's advice to Alice, which we quoted in Chapter 1.

What Have We Learned?

In Chapter 28, we learned that multiple regression is a natural way to extend what we knew about linear regression models to include several predictors. Now we've learned that multiple regression is both more powerful and more complex than it may appear at first. As with other chapters in this book whose titles spoke of greater "wisdom," this chapter has drawn us deeper into the uses and cautions of multiple regression.

Learning Objectives

- Know how to incorporate categorical data by using indicator variables, modeling relationships that have parallel slopes but at different levels for different groups.
- Know how to use interaction terms, to allow for different slopes. We can create identifier variables that isolate individual cases to remove their influence from the model while exhibiting how they differ from the other points and testing whether that difference is statistically significant.
- Beware unusual cases. A single case can have high leverage, allowing it to influence the entire regression. Such cases should be treated specially, possibly by fitting the model both with and without them or by including indicator variables to isolate their influence.
- Be cautious in complex models because one has to be careful in interpreting the coefficients. Associations among the predictors can change the coefficients to values that can be quite different from the coefficient in the simple regression of a predictor and the response, even changing the sign.
- Understand that building multiple regression models is an art that speaks to the central goal of statistical analysis: understanding the world with data. We've learned that there is no right model. We've seen that the same response variable can be modeled with several alternative models, each showing us different aspects of the data and of the relationships among the variables and nominating different cases as special and deserving of our attention.
- We've also seen that everything we've discussed throughout this book fits together to help us understand the world. The graphical methods are the same ones we learned in the early chapters, and the inference methods are those we originally developed for means. In short, there's been a consistent tale of how we understand data to which we've added more and more detail and richness, but which has been consistent throughout.

What *Else* Have We Learned?

We, the authors, hope that you've also learned to see the world differently, to understand what has been measured and about whom, to be skeptical of untested claims and curious about patterns and relationships. We hope that you find the world a more interesting, more nuanced place that can be understood and appreciated with the tools of Statistics and Science.

Finally, we hope you'll consider further study in Statistics. Whatever your field, whatever your job, whatever your interests, you can use Statistics to understand the world better.

Review of Terms

Indicator variable A variable constructed to indicate for each case whether it is in a designated group or not. A common way to assign values to indicator variables is to let them take on the values 0 and 1, where 1 indicates group membership (p. 862).

Interaction term A constructed variable found as the product of a predictor and an indicator variable. An interaction term adjusts the *slope* of the cases identified by the indicator against the predictor (p. 864).

Leverage	The leverage of a case measures how far its *x*-values are from the center of the *x*'s and, consequently, how much influence it can exert on the regression model. Points with high leverage can determine a regression model and should, therefore, be examined carefully (p. 865).
Studentized residual	When a residual is divided by an independent estimate of its standard deviation, the result is a Studentized residual. The type of Studentized residual that has a *t*-distribution is an *externally* Studentized residual (p. 867).
Influential case	A case is *influential* on a multiple regression model if, when it is omitted, the model changes by enough to matter for your purposes. (There is no specific amount of change defined to declare a case influential.) Cases with high leverage and large Studentized residual are likely to be influential (p. 868).
Stepwise regression	An automated method of building regression models in which predictors are added to or removed from the model one at a time in an attempt to optimize a measure of the success of the regression. Stepwise methods rarely find the best model and are easily influenced by influential cases, but they can be valuable in winnowing down a large collection of candidate predictors (p. 873).
Collinearity	When one (or more) of the predictors can be fit closely by a multiple regression on the other predictors, we have collinearity. When collinear predictors are in a regression model, they may have unexpected coefficients and often have inflated standard errors (and correspondingly small *t*-statistics) (p. 883).

On the Computer REGRESSION ANALYSIS

Statistics packages differ in how much information they provide to diagnose a multiple regression. Most packages provide leverage values. Many provide far more, including statistics that we have not discussed. But for all, the principle is the same. We hope to discover any cases that don't behave like the others in the context of the regression model and then to understand why they are special.

Many of the ideas in this chapter rely on the concept of examining a regression model and then finding a new one based on your growing understanding of the model and the data. Regression diagnosis is meant to provide steps along that road. A thorough regression analysis may involve finding and diagnosing several models.

DATA DESK

Request diagnostic statistics and graphs from the HyperView menus in the regression output table. Most will update and can be set to update automatically when the model or data change.

COMMENTS

You can add a predictor to the regression by dragging its icon into the table, or replace variables by dragging the icon over their name in the table. Click on a predictor's name to drop down a menu that lets you remove it from the model.

EXCEL

Excel does not offer diagnostic statistics with its regression function.

COMMENTS

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. The "standardized residuals" are just the residuals divided by their standard deviation (with the wrong df), so they too should be ignored.

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.**
- Click on the red triangle in the title of the Model output to find a variety of plots and diagnostics available.

COMMENTS

JMP chooses a regression analysis when the response variable is "Continuous."

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 **Storage** button.
- In the Regression Storage dialog, you can select a variety of diagnostic statistics. They will be stored in columns of your worksheet.
- 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.

COMMENTS

You will probably want to make displays of the stored diagnostic statistics. Use the usual Minitab methods for creating displays.

R

Suppose the response variable *y* and predictor variables *x1,…,xk* are in a data frame called mydata. After fitting a multiple regression of *y* on *x1* and *x2* via:

- mylm = lm(y~x1+x2,data=mydata)
- summary(mylm) # gives the details of the fit, including the ANOVA table
- plot(mylm) #gives a variety of plots
- lm.influence(mylm) #gives a variety of regression diagnostic values

To get partial regression plots (called Added Variable plots in **R**), you need the library car:

- library(car)

Then to get the partial regression plots:

- avPlots(mylm) #one plot for each predictor variable – interactions not permitted

STATCRUNCH

StatCrunch offers some of the diagnostic statistics discussed in this chapter in the regression dialog. It does not currently make partial regression plots.

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 **Save** button.
- In the Linear Regression Save dialog, choose diagnostic statistics. These will be saved in your worksheet along with your data.
- 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.

Exercises

Section 29.1

1. Indicators For each of these potential predictor variables say whether they should be represented in a regression model by indicator variables. If so, then suggest what specific indicators should be used (that is, what values they would have).

a) In a regression to predict income, the sex of respondents in a survey.

b) In a regression to predict the square footage available for rent, the number of stories in a commercial building.

c) In a regression to predict the amount an individual's medical insurance would pay for an operation, whether the individual was over 65 (and eligible for Medicare).

2. More indicators For each of these potential predictor variables say whether they should be represented in a regression model by indicator variables. If so, then suggest what specific indicators should be used (that is, what values they would have).

a) In a regression to predict income, the age of respondents in a survey.

b) In a regression to predict the square footage available for rent, whether a commercial building has an elevator or not.

c) In a regression to predict annual medical expenses, whether a person was a child (in pediatric care), an adult, or a senior (over 65 years old).

Section 29.2

3. Residual, leverage, influence For each of the following cases, would your primary concern about them be that they had a large residual, large leverage, or likely large influence on the regression model? Explain your thinking.

a) In a regression to predict the construction cost of roller coasters from the length of track, the height of the highest point, and the type of construction (metal or wood), the *Kingda Ka* coaster, which opened in 2005 and, at (456 ft), is currently the tallest.

b) In a regression to predict income of graduates of a college five years after graduation, a graduate who created a high-tech start-up company based on his senior thesis, and sold it for several million dollars.

4. Residual, leverage, influence, 2 For each of the following cases, would your primary concern about them be that they had a large residual, large leverage, or likely large influence on the regression model?

a) In a regression to predict Freshman grade point averages as part of the admissions process, a student whose Math SAT was 750, whose Verbal SAT was 585, and who had a 4.0 GPA at the end of her Freshman year.

b) In a regression to predict life expectancy in countries of the world from generally-available demographic, economic, and health statistics, a country that, due to a high prevelance of HIV, has an unusually low life expectancy.

Section 29.3

5. Significant coefficient? In a regression to predict compensation of employees in a large firm, the predictors in the regression were *Years with the firm*, *Age*, and *Years of Experience*. The coefficient of *Age* is negative and statistically significantly different from zero. Does this mean that the company pays workers less as they get older? Explain.

6. Better model? Joe wants to impress his boss. He builds a regression model to predict sales that has 20 predictors and an R^2 of 80%. Sally builds a competing model with only 5 predictors, but an R^2 of only 78%. Which model is likely to be most useful for understanding the drivers of sales? How could the boss tell? Explain.

Chapter Exercises

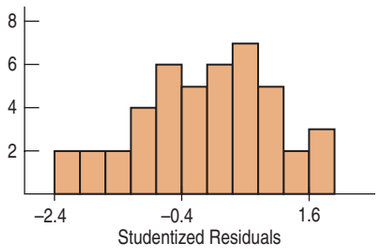
7. Climate change 2013 again Recent concern with the rise in global temperatures has focused attention on the level of carbon dioxide (CO_2) in the atmosphere. The National Oceanic and Atmospheric Administration (NOAA) records the CO_2 levels in the atmosphere atop the Mauna Loa volcano in Hawaii, far from any industrial contamination, and calculates the annual overall temperature of the atmosphere and the oceans using an established method. (See data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts+dSST.txt and ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_annmean_mlo.txt. We examined these data in Chapter 7 and again in Chapter 25. There we saw a strong relationship between global mean temperature and the level of CO_2 in the atmosphere.) Here is a regression predicting Mean Annual Temperature from annual CO_2 levels (parts per million). We'll examine the data from 1970 to 2013.

Dependent variable is: Global Temperature Anomaly
Response variable is: Global Avg Temp
R-squared = 73.6% R-squared (adjusted) = 72.3%
s = 0.1331 with 44 − 3 = 41 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	2.02629	2	1.01315	57.2
Residual	0.726571	41	0.017721	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	18.8061	35.03	0.537	0.5942
Year	−4.60135e-3	0.0197	−0.233	0.8169
CO_2	0.013127	0.0121	1.09	0.2830

A histogram of the externally Studentized residuals looks like this:



a) Comment on the distribution of the Studentized residuals.
b) It is widely understood that global temperatures have been rising consistently during this period. But the coefficient of *Year* is negative and its *t*-ratio is small. Does this contradict the common wisdom?

8. Pizza Consumers' Union rated frozen pizzas. Their report includes the number of *Calories, Fat* content, and *Type* (cheese or pepperoni, represented here as an indicator variable that is 1 for cheese and 0 for pepperoni). Here's a regression model to predict the *"Score"* awarded each pizza from these variables:

Dependent variable is: Score
R-squared = 28.7%
R-squared (adjusted) = 20.2%
s = 19.79 with 29 − 4 = 25 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	3947.34	3	1315.78	3.36
Residual	9791.35	25	391.654	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−148.817	77.99	−1.91	0.0679
Calories	0.743023	0.3066	2.42	0.0229
Fat	−3.89135	2.138	−1.82	0.0807
Type	15.6344	8.103	1.93	0.0651

a) What is the interpretation of the coefficient of cheese in this regression?
b) What displays would you like to see to check assumptions and conditions for this model?

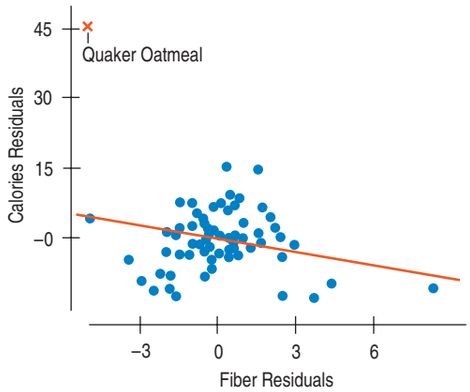
9. Healthy breakfast, sick data A regression model for data on breakfast cereals originally looked like this:

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

Let's take a closer look at the coefficient for *Fiber*. Here's the partial regression plot for *Fiber* in that regression model:



a) The line on the plot is the least squares line fit to this plot. What is its slope? (You may need to look back at the facts about partial regression plots in Chapter 28.)
b) One point is labeled as corresponding to Quaker Oatmeal. What effect does this point have on the slope of the line? (Does it make it larger, smaller, or have no effect at all?)

Here is the same regression with Quaker Oatmeal removed from the data:

Dependent variable is: Calories
77 total cases of which 1 is missing
R-squared = 93.9% R-squared (adjusted) = 93.5%
s = 5.002 with 76 − 6 = 70 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Regression	27052.4	5	5410.49	216	
Residual	1751.51	70	25.0216		

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−1.25891	4.292	−0.293	0.7701
Protein	3.88601	0.6963	5.58	<0.0001
Fat	8.69834	0.6512	13.4	<0.0001
Fiber	0.250140	0.3277	0.763	0.4478
Carbo	4.14458	0.2005	20.7	<0.0001
Sugars	3.96806	0.1692	23.4	<0.0001

c) Compare this regression with the previous one. In particular, which model is likely to make the best predictions of calories? Which seems to fit the data better?
d) How would you interpret the coefficient of *Fiber* in this model? Does *Fiber* contribute significantly to modeling calories?

(In fact, the data for Quaker Oatmeal was determined to be in error and was corrected for the subsequent analyses seen elsewhere in this book.)

10. Fifty states In Exercise 25 of Chapter 28 we looked at data from the 50 states. Here's an analysis of the same data from a few years earlier. 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. We are trying to find a regression model for *Life Expectancy*.

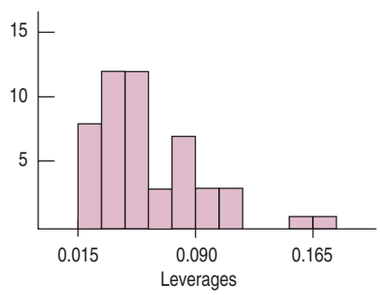
Here's the result of a regression on all the available predictors:

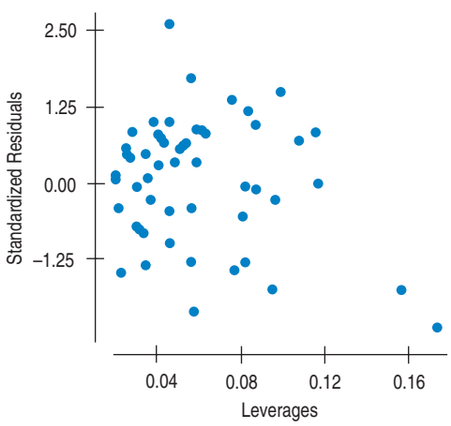
Dependent variable is: Lifeexp
R-squared = 67.0% R-squared (adjusted) = 64.0%
s = 0.8049 with 50 − 5 = 45 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	59.1430	4	14.7858	22.8
Residual	29.1560	45	0.6479	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	69.4833	1.325	52.4	<0.0001
Murder	−0.261940	0.0445	−5.89	<0.0001
HS grad	0.046144	0.0218	2.11	0.0403
Income	1.24948e–4	0.0002	0.516	0.6084
Illiteracy	0.276077	0.3105	0.889	0.3787

Here's a histogram of the leverages and a scatterplot of the externally Studentized residuals against the leverages:





a) The two states with high leverages and large (negative) Studentized residuals are Nevada and Alaska. Do you think they are likely to be influential in the regression? From just the information you have here, why or why not?

Here's the regression with indicator variables for Alaska and Nevada added to the model to remove those states from affecting the model:

Dependent variable is: Lifeexp
R-squared = 74.1% R-squared (adjusted) = 70.4%
s = 0.7299 with 50 − 7 = 43 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	66.9280	1.442	46.4	<0.0001
Murder	−0.207019	0.0446	−4.64	<0.0001
HS grad	0.065474	0.0206	3.18	0.0027
Income	3.91600e-4	0.0002	1.63	0.1105
Illiteracy	0.302803	0.2984	1.01	0.3159
Alaska	−2.57295	0.9039	−2.85	0.0067
Nevada	−1.95392	0.8355	−2.34	0.0241

b) What evidence do you have that Nevada and Alaska are outliers with respect to this model? Do you think they should continue to be treated specially? Why?

c) Would you consider removing any of the predictors from this model? Why or why not?

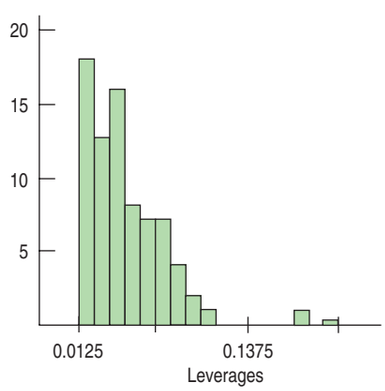
11. Cereals, part 2 In Exercise 26 of Chapter 28, we considered a multiple regression model for predicting calories in breakfast cereals. The regression looked like this:

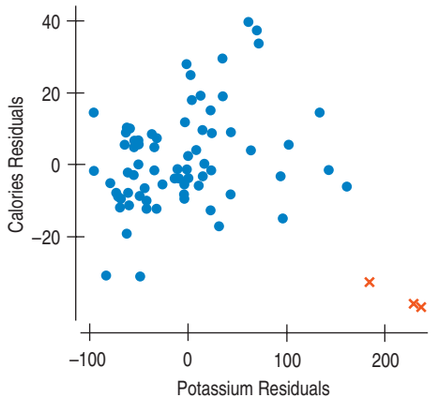
Dependent variable is: Calories
R-squared = 38.4% R-squared (adjusted) = 35.9%
s = 15.60 with 77 − 4 = 73 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Regression	11091.8	3	3697.28	15.2	<0.0001
Residual	17760.1	73	243.289		

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	83.0469	5.198	16.0	<0.0001
Sodium	0.057211	0.0215	2.67	0.0094
Potassium	−0.019328	0.0251	−0.769	0.4441
Sugars	2.38757	0.4066	5.87	<0.0001

Here's a histogram of the leverages and a partial regression plot for *Potassium* in which the three high-leverage points are plotted with red x's. (They are *All-Bran, 100% Bran,* and *All-Bran with Extra Fiber.)*



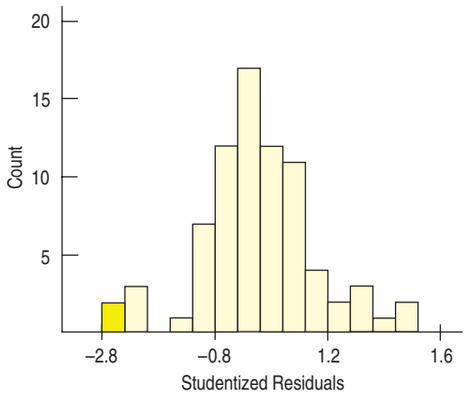


With this additional information, answer the following:

a) How would you interpret the coefficient of *Potassium* in the multiple regression?

b) *Without doing any calculating,* how would you expect the coefficient and *t*-statistic for *Potassium* to change if we were to omit the three high-leverage points?

Here's a histogram of the externally Studentized residuals. The selected bar, holding the two most negative residuals, holds the two bran cereals that had the largest leverages.



With this additional information, answer the following:

c) What term would you apply to these two cases? Why?

d) Do you think they should be omitted from this analysis? Why or why not? (*Note*: There is no correct choice. What matters is your reasons.)

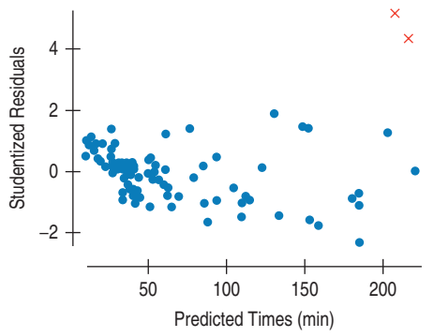
12. Scottish hill races 2008 In Chapter 28, Exercises 14 and 16, we considered data on hill races in Scotland. These are overland races that climb and descend hills—sometimes several hills in the course of one race. Here is a regression analysis to predict the *Women's Record* times from the *Distance* and total vertical *Climb* of the races:

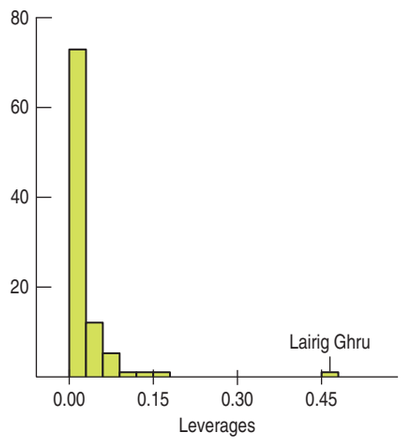
Dependent variable is: Women's record
R-squared = 96.7% R-squared (adjusted) = 96.7%
s = 10.06 with 90 − 3 = 87 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	261029	2	130515	1288
Residual	8813.02	87	101.299	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−11.6545	1.891	−6.16	<0.0001
Distance	4.43427	0.2200	20.2	<0.0001
Climb	0.045195	0.0033	13.7	<0.0001

Here is the scatterplot of externally Studentized residuals against predicted values, as well as a histogram of leverages for this regression:





a) Comment on what these diagnostic displays indicate.

b) The two races with the largest Studentized residuals are the Arochar Alps race and the Glenshee 9. Both are relatively new races, having been run only one or two times with relatively few participants. What effects can you be reasonably sure they have had on the regression? What displays would you want to see to investigate other effects? Explain.

c) If you have access to a suitable statistics package, make the diagnostic plots you would like to see and discuss what you find.

13. Traffic delays 2011 The Texas Transportation Institute studies traffic delays. Data the institute published for the year 2011 include information on the Cost of Congestion per auto commuter ($) (hours per year spent delayed by

traffic), *Congested% (Percent of vehicle miles traveled that were congested)*, and the *Size* of the city (small, medium, large, very large). The regression model based on these variables looks like this. The variables *Small, Large*, and *Very Large* are indicators constructed to be 1 for cities of the named size and 0 otherwise.

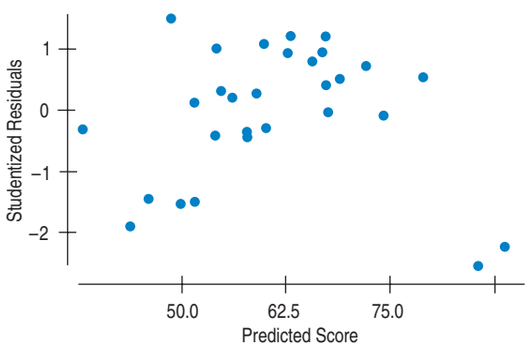
Response variable is: Congestion per auto commuter ($)
R-squared = 63.3% R-squared (adjusted) = 61.8%
s = 152.4 with 101 − 5 = 96 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	3852487	4	963122	41.5
Residual	2229912	96	23228.2	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	501.490	43.27	11.6	<0.0001
Small	−104.239	43.27	−2.41	0.0179
Large	106.026	41.46	2.56	0.0121
Very Large	348.147	59.67	5.83	<0.0001
Congested% ...	3.13481	0.9456	3.32	0.0013

a) Explain how the coefficients of *Small, Large,* and *Very Large* account for the size of the city in the model. Why is there no coefficient for *Medium*?
b) What is the interpretation of the coefficient of *Large* in this regression model?

14. Gourmet pizza Here's a plot of the Studentized residuals against the predicted values for the regression model found in Exercise 8:



The two extraordinary cases in the plot of residuals are Reggio's and Michelina's, two gourmet pizzas.
a) Interpret these residuals. What do they say about these two brands of frozen pizza? Be specific—that is, talk about the *Scores* they received and might have been expected to receive.

We can create indicator variables to isolate these cases. Adding them to the model results in the following model:

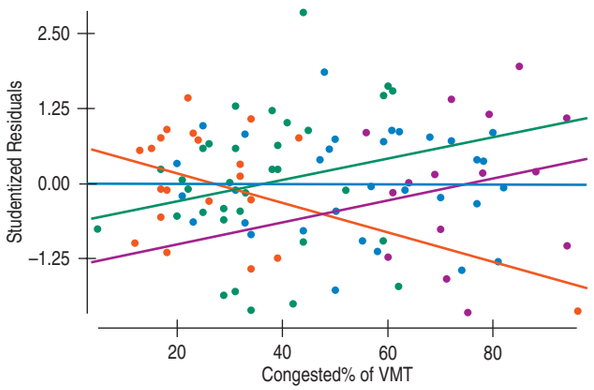
Dependent variable is: Score
R-squared = 65.2% R-squared (adjusted) = 57.7%
s = 14.41 with 29 − 6 = 23 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	8964.13	5	1792.83	8.64
Residual	4774.56	23	207.590	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−363.109	72.15	−5.03	<0.0001
Calories	1.56772	0.2824	5.55	<0.0001
Fat	−8.82748	1.887	−4.68	0.0001
Cheese	25.1540	6.214	4.05	0.0005
Reggio's	−67.6401	17.86	−3.79	0.0010
Michelina's	−67.0036	16.62	−4.03	0.0005

b) What does the coefficient of *Michelina's* mean in this regression model? Do you think that Michelina's pizza is an outlier for this model for these data? Explain.

15. More traffic Here's a plot of Studentized residuals against *Congested%* for the model of Exercise 13. The plot is colored according to *City Size,* and regression lines are fit for each size.



a) The model of Exercise 13 includes indicators for *City Size.* Considering this display, have these indicator variables accomplished what is needed for the regression model? Explain.

We constructed additional indicators as the product of *Small* with *Arterial mph* and the product of *Very Large* with *Arterial mph.* Here's the resulting model:

Response variable is: Congestion per auto commuter ($)
R-squared = 64.9% R-squared (adjusted) = 63.1%
s = 149.9 with 101 − 6 = 95 degrees of freedom

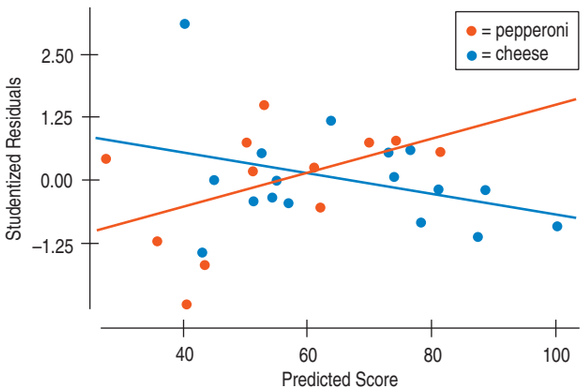
Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	3947593	5	789519	35.1
Residual	2134805	95	22471.6	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	461.515	46.79	9.86	<0.0001
Small	27.8150	77.02	0.361	0.7188
Large	86.1516	41.90	2.06	0.0425
Very Large	305.853	62.19	4.92	<0.0001
Congested% ...	4.24056	1.074	3.95	0.0002
Sml*C%V	−4.41683	2.147	−2.06	0.0424

b) What does the predictor *Sml*C%V* (Small by Congestion%) do in this model? Interpret the coefficient.

c) Does this appear to be a good regression model? Would you consider removing any predictors? Why or why not?

16. Another slice of pizza A plot of Studentized residuals against predicted values for the regression model found in Exercise 14 now looks like this. It has been colored according to *Type* of pizza and separate regression lines fitted for each type:



a) Comment on this diagnostic plot. What does it say about how the regression model deals with cheese and pepperoni pizzas?

Based on this plot, we constructed yet another variable consisting of the indicator *cheese* multiplied by *Calories*:

Dependent variable is: Score
R-squared = 73.7% R-squared (adjusted) = 66.5%
s = 12.82 with 29 − 7 = 22 degrees of freedom

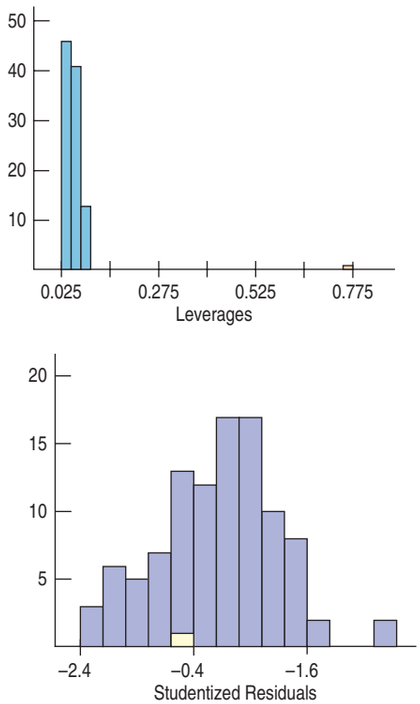
Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	10121.4	6	1686.90	10.3
Residual	3617.32	22	164.424	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−464.498	74.73	−6.22	<0.0001
Calories	1.92005	0.2842	6.76	<0.0001
Fat	−10.3847	1.779	−5.84	<0.0001
Cheese	183.634	59.99	3.06	0.0057
Cheese*cals	−0.461496	0.1740	−2.65	0.0145
Reggio's	−64.4237	15.94	−4.04	0.0005
Michelina's	−51.4966	15.90	−3.24	0.0038

b) Interpret the coefficient of *Cheese*cals* in this regression model.

c) Would you prefer this regression model to the model of Exercise 14? Explain.

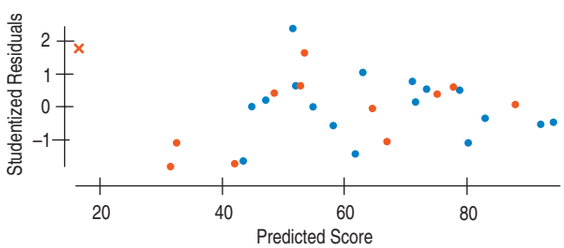
17. Influential traffic? Here are histograms of the leverage and Studentized residuals for the regression model of Exercise 15.



The city with the highest leverage is Laredo, TX. It's highlighted in both displays.

Do you think Laredo is an influential case? Explain your reasoning.

18. The final slice Here's the residual plot corresponding to the regression model of Exercise 16:

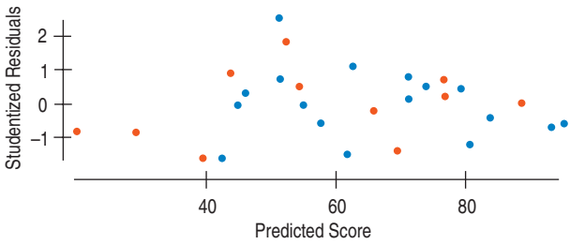


The extreme case this time is *Weight Watchers Pepperoni* (makes sense, doesn't it?). We can make one more indicator for *Weight Watchers*. Here's the model:

Dependent variable is: Score
R-squared = 77.1% R-squared (adjusted) = 69.4%
s = 12.25 with 29 − 8 = 21 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	10586.8	7	1512.41	10.1
Residual	3151.85	21	150.088	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−525.063	79.25	−6.63	<0.0001
Calories	2.10223	0.2906	7.23	<0.0001
Fat	−10.8658	1.721	−6.31	<0.0001
Cheese	231.335	63.40	3.65	0.0015
Cheese*cals	−0.586007	0.1806	−3.24	0.0039
Reggio's	−66.4706	15.27	−4.35	0.0003
Michelina's	−52.2137	15.20	−3.44	0.0025
Weight W…	28.3265	16.09	1.76	0.0928



a) Compare this model with the others we've seen for these data. In what ways does this model seem better or worse than the others?

b) Do you think the indicator for *Weight Watchers* should be in the model? (Consider the effect that including it has had on the other coefficients also.)

c) What do the Consumers' Union tasters seem to think makes for a really good pizza?

Just Checking ANSWERS

1. Histograms are used to examine the shapes of distributions of individual variables. We check especially for multiple modes, outliers, and skewness. They are also used to check the shape of the distribution of the residuals for the Nearly Normal Condition.

2. Scatterplots are used to check the Straight Enough Condition in plots of *y vs.* any of the *x's.* They are used to check plots of the residuals or Studentized residuals against the predicted values, against any of the predictors, or against *Time* to check for patterns. Scatterplots are also the display used in partial regression plots, where we check for influential points and unexpected subgroups.

3. The Normal model is needed only when we use inference; it isn't needed for computing a regression model. We check the Nearly Normal Condition on the residuals.