GLOBAL EDITION

Stats

Data and Models

FIFTH EDITION

De Veaux • Velleman • Bock

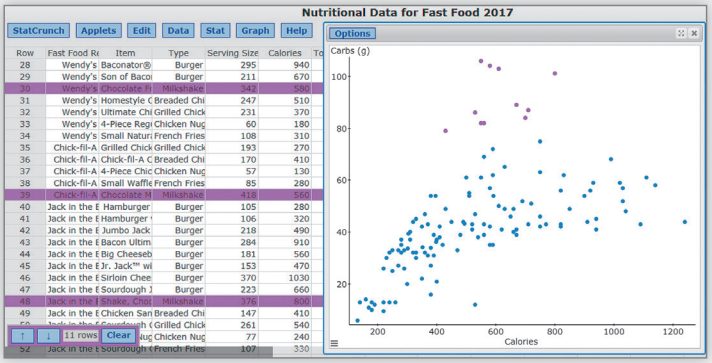
Get the Most Out of MyLab Statistics

MyLab™ Statistics is the teaching and learning platform that empowers instructors to reach *every* student. By combining trusted author content with digital tools and a flexible platform, MyLab Statistics personalizes the learning experience and improves results for each student.

Collect, crunch, and communicate with StatCrunch

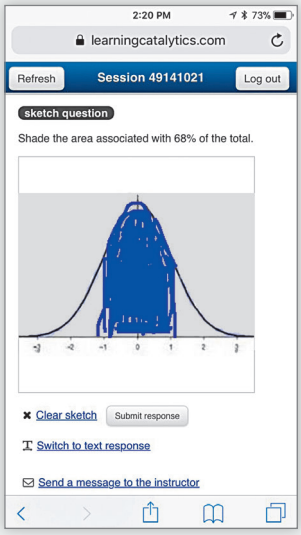
With StatCrunch®, Pearson's powerful web-based statistical software, instructors and students can access tens of thousands of data sets including those from the textbook, perform complex analyses, and generate compelling reports. StatCrunch is integrated directly into MyLab Statistics.

Beyond StatCrunch, MyLab Statistics makes learning and using a variety of statistical software packages seamless and intuitive by allowing users to download and copy data sets directly into other programs. Students can access instructional tools including tutorial videos, and study cards.



Give every student a voice with Learning Catalytics

Learning Catalytics™ is an interactive classroom tool that allows every student to participate. Instructors ask a variety of questions that help students recall ideas, apply concepts, and develop critical-thinking skills. Students answer using their smartphones, tablets, or laptops to show that they do—or don't—understand. Instructors monitor responses to adjust their teaching approach, and even set up peer-to-peer learning. More importantly, they use real-time analytics to address student misconceptions the moment they occur and ensure they hear from every student when it matters most.



Visit pearsonmylabandmastering.com and click Training & Support to make sure you're getting the most out of MyLab Statistics.

T ***68. Planets** Here is a table of the 9 sun-orbiting objects formerly known as planets.

Planet	Position Number	Distance from Sun (million miles)	Length of Year (Earth years)
Mercury	1	36.254	0.24
Venus	2	66.931	0.62
Earth	3	92.960	1
Mars	4	141.299	1.88
Jupiter	5	483.392	11.86
Saturn	6	886.838	29.46
Uranus	7	1782.97	84.01
Neptune	8	2794.37	164.8
Pluto	9	3671.92	248

a) Plot the *Length* of the year against the *Distance* from the sun. Describe the shape of your plot.
b) Re-express one or both variables to straighten the plot. Use the re-expressed data to create a model describing the length of a planet's year based on its distance from the sun.
c) Comment on how well your model fits the data.

T ***69. Is Pluto a planet?** Let's look again at the pattern in the locations of the planets in our solar system seen in the table in Exercise 68.

a) Re-express the distances to create a model for the *Distance* from the sun based on the planet's *Position*.
b) Based on this model, would you agree with the International Astronomical Union that Pluto is not a planet? Explain.

T ***70. Planets and asteroids** The asteroid belt between Mars and Jupiter may be the remnants of a failed planet. If so, then Jupiter is really in position 6, Saturn is in 7, and so on. Repeat Exercise 69, using this revised method of numbering the positions. Which method seems to work better?

T ***71. Planets and Eris** In July 2005, astronomers Mike Brown, Chad Trujillo, and David Rabinowitz announced the discovery of a sun-orbiting object, since named Eris,[8] that is 5% larger than Pluto. Eris orbits the sun once every 560 earth years at an average distance of about 6300 million miles from the sun. Based on its *Position*, how does Eris's *Distance* from the sun (re-expressed to logs) compare with the prediction made by your model of Exercise 70?

T ***72. Planets, models, and laws** The model you found in Exercise 68 is a relationship noted in the 17th century by Kepler as his Third Law of Planetary Motion. It was subsequently explained as a consequence of Newton's Law of Gravitation. The models for Exercises 69, 70, and 71 relate to what is sometimes called the Titius-Bode "law," a pattern noticed in the 18th century but lacking any scientific explanation.

Compare how well the re-expressed data are described by their respective linear models. What aspect of the model of Exercise 68 suggests that we have found a physical law? In the future, we may learn enough about planetary systems around other stars to tell whether the Titius-Bode pattern applies there. If you discovered that another planetary system followed the same pattern, how would it change your opinion about whether this is a real natural "law"? What would you think if some of the extrasolar planetary systems being discovered do not follow this pattern?

***73. Logs (not logarithms)** The value of a log is based on the number of board feet of lumber the log may contain. (A board foot is the equivalent of a piece of wood 1 inch thick, 12 inches wide, and 1 foot long. For example, a $2'' \times 4''$ piece that is 12 feet long contains 8 board feet.) To estimate the amount of lumber in a log, buyers measure the diameter inside the bark at the smaller end. Then they look in a table based on the Doyle Log Scale. The table below shows the estimates for logs 16 feet long.

Diameter of Log	8	12	16	20	24	28
Board Feet	169	441	841	1369	2025	2809

a) What model does this scale use?
b) How much lumber would you estimate that a log 25 inches in diameter contains?
c) What does this model suggest about logs 30 inches in diameter?

T ***74. Weightlifting 2016** Listed below are the world record men's weightlifting performances as of 2016.

Weight Class (kg)	Record Holder	Country	Total Weight
56	Long Quingquan	China	307
62	Kim Un-Guk	North Korea	327
69	Galain Boevski	Bulgaria	357
77	Lu Xiaojun	China	379
85	Kianoush Rostami	Iran	396
94	Ilya Ilyin	Kazakhstan	418
105	Andrei Aramnau	Belarus	436
105+	Lasha Tlakhadze	Georgia	473

a) Create a linear model for the *Weight Lifted* in each *Weight Class*, leaving out the 105+ unlimited class.
b) Check the residuals plot. Is your linear model appropriate?
c) Create a better model by re-expressing *Weight Lifted* and explain how you found it.
d) Explain why you think your new model is better.
e) The record holder of the Unlimited weight class weighs 157 kg. Predict how much he can lift with the models from parts a and c.

***75. Life expectancy history** The table gives the *Life Expectancy* for a certain demographic group in a certain region for every decade during the past 120 years (1 = 1900 to 1910, 2 = 1911 to 1920, etc.). Create a model to predict future increases in life expectancy. Hint: Try "Plan B."

Decade	1	2	3	4	5	6	7	8	9	10	11	12
Life exp	46.8	48.4	54.5	60.0	62.3	66.3	67.8	68.5	70.7	72.6	75.2	76.0

[8]Eris is the Greek goddess of warfare and strife who caused a quarrel among the other goddesses that led to the Trojan War. In the astronomical world, Eris stirred up trouble when the question of its proper designation led to the raucous meeting of the IAU in Prague where IAU members voted to demote Pluto and Eris to dwarf-planet status (www.gps.caltech.edu/~mbrown/planetlila/#paper).

T ***76. Lifting more weight** In Exercise 74 you examined the record weightlifting performances for the Olympics. You found a re-expression of *Weight Lifted*.

a) Find a model for *Weight Lifted* by re-expressing *Weight Class* instead of *Weight Lifted*.
b) Compare this model to the one you found in Exercise 74.
c) Predict the *Weight Lifted* by the 157 kg record holder in Exercise 74, part e.
d) Which prediction do you think is better? Explain.
e) The record holder is Lasha Talakhadze, who lifted 473 kg at the 2016 Rio de Janeiro Olympics. Which model predicted it better?

T ***77. Slower is cheaper?** Researchers studying how a car's *Fuel Efficiency* varies with its *Speed* drove a compact car 200 miles at various speeds on a test track. Their data are shown in the table.

Speed (mph)	35	40	45	50	55	60	65	70	75
Miles per gal	15.9	19.4	22.8	24.9	25.9	26.2	25.2	23.9	21.7

Create a linear model for this relationship, and report any concerns you may have about the model.

T ***78. Oranges** The table below shows that as the number of oranges on a tree increases, the fruit tends to get smaller. Create a model for this relationship, and express any concerns you may have.

Number of Oranges/Tree	Average Weight/Fruit (lb)
50	0.61
100	0.59
150	0.57
200	0.56
250	0.53
300	0.52
350	0.51
400	0.49
450	0.48
500	0.46
600	0.45
700	0.43
800	0.41
900	0.40

T ***79. Years to live, 2016** Insurance companies and other organizations use actuarial tables to estimate the remaining lifespans of their customers. The data file gives life expectancy and estimated additional years of life for black males in a particular country.

Age	10	20	30	40	50	60	70	80	90	100
Years Left	61.2	52.4	40.8	33.7	27.9	19.5	13.3	8.6	4.8	2.5

a) Fit a regression model to predict *Life expectancy* from *Age*. Does it look like a good fit? Now plot the residuals.
b) Find a re-expression to create a better model. Predict the life expectancy of a 21-year-old black man.
c) Are you satisfied that your model has accounted for the relationship between *Life expectancy* and *Age*? Explain.

***80. Tree growth** A 1996 study examined the growth of grapefruit trees in Texas, determining the average trunk *Diameter* (in inches) for trees of varying *Ages:*

Age (yr)	2	4	6	8	10	12	14	16	18	20
Diameter (in.)	2.1	3.9	5.2	6.2	6.9	7.6	8.3	9.1	10.0	11.4

a) Fit a linear model to these data. What concerns do you have about the model?
b) If data had been given for individual trees instead of averages, would you expect the fit to be stronger, less strong, or about the same? Explain.

JUST CHECKING

Answers

1. Not high leverage, not influential, large residual
2. High leverage, not influential, small residual
3. High leverage, influential, not large residual
4. Counts are often best transformed by using the square root.
5. None. The relationship is already straight.
6. Even though, technically, the population values are counts, you should probably try a stronger transformation like log(population) because populations grow in proportion to their size.

9 Multiple Regression

WHERE ARE WE GOING?

Linear regression models are often useful, but the world is usually not so simple that a two-variable model does the trick. For a more realistic understanding, we need models with several variables.

Three percent of a man's body is essential fat. For a woman, the percentage is closer to 12.5%. As the name implies, essential fat is necessary for a normal, healthy body. Fat is stored in small amounts throughout the body. Too much body fat, however, can be dangerous to health. For men between 18 and 39 years old, a healthy percentage of body fat varies from 8% to 19%. For women of the same age, it's 21% to 32%.

Measuring body fat can be tedious and expensive. The "standard reference" measurement is by dual-energy X-ray absorptiometry (DEXA), which involves two low-dose X-ray generators and takes from 10 to 20 minutes.

How close can we get to a usable prediction of body fat from easily measurable variables such as *Height, Weight*, and *Waist* size? Here's a scatterplot of *%Body Fat* plotted against *Waist* size for a sample of 250 men of various ages. (Data in **Bodyfat**)

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

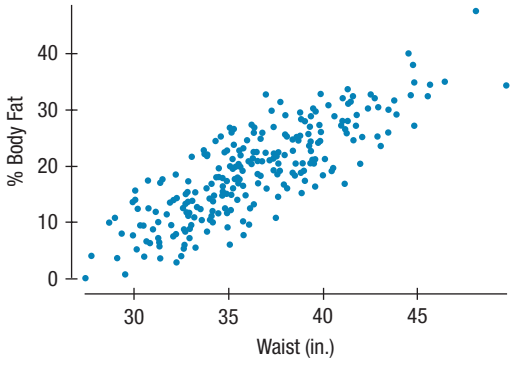


Figure 9.1
The relationship between *%Body Fat* and *Waist* size for 250 men

The plot is clearly straight, so we can find a least squares regression line. The equation of the least squares line for these data is $\widehat{\%Body\ Fat} = -42.7 + 1.7\ Waist$. The slope says that, on average, men who have an additional inch around the waist are expected to have about 1.7% more body fat.

This regression does pretty well. The standard deviation of the residuals is just 4.713 *%Body Fat*. The R^2 of 67.8% says that the regression model accounts for almost 68% of the variability in *%Body Fat* just by knowing *Waist* size.

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

9.1 What Is Multiple Regression?

A Note on Terminology

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

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

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

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

Table 9.1
A multiple regression table provides many values of interest. But for now, we'll just look at the R^2, the standard deviation of the residuals, and the coefficients. You can ignore the grayed-out parts of this table until we come back to the topic later in the text.

Dependent variable is: %Body Fat
R-squared = 71.3,
s = 4.460 with 250 − 3 = 247 degrees of freedom

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

We have already seen many of the important numbers in this table. Most of them continue to mean what they did in a simple regression.

R^2 gives the fraction of the variability of *%Body Fat* accounted for by the multiple regression model. The multiple regression model accounts for 71.3% of the variability in *%Body Fat*. That's an improvement over 67.8% with *Waist* alone predicting *%Body Fat*. We shouldn't be surprised that R^2 has gone up. It was the hope of accounting for some of the leftover variability that led us to try a second predictor.

The standard deviation of the residuals is still denoted s (or sometimes s_e to distinguish it from the standard deviation of y). It has gone down from 4.713 to 4.460. As before, we can think of the 68–95–99.7 Rule. The regression model will, of course, make errors—indeed, it is unlikely to fit any of the 250 men perfectly. But about 68% of those errors are likely to be smaller than 4.46 *%Body Fat*, and 95% are likely to be less than twice that.

Other values are usually reported as part of a regression. We've shown them here because you are likely to see them in any computer-generated regression table. But we've made them gray because we aren't going to talk about them yet. You can safely ignore them for now. Don't worry, we'll discuss them all in later chapters.

For each predictor, the table gives a coefficient next to the name of the variable it goes with. Using the coefficients from this table, we can write the regression model:

$$\widehat{\%Body\ Fat} = -3.10 + 1.77\ Waist - 0.60\ Height.$$

As before, we define the residuals as

$$Residuals = \%Body\ Fat - \widehat{\%Body\ Fat}.$$

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

So what's different and why is understanding multiple regression important? There are several answers to these questions. First—and most important—the meaning of the coefficients in the regression model has changed in a subtle but important way. Because that change is not obvious, multiple regression coefficients are often misinterpreted.

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

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

EXAMPLE 9.1 Modeling Home Prices

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. (Data in **Real estate**)

Variables available include the price paid, the size of the living area (sq ft), the number of bedrooms, the number of bathrooms, the year of construction, the lot size (acres), and a coding of the location as urban, suburban, or rural made by the student who collected the data.

Here's a regression to model the sale price from the living area (sq ft) and the number of bedrooms.

Dependent variable is: Price
R-squared = 14.6%
s = 266899 with 894 − 3 = 891 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	308100	41148	7.49	<0.0001
Living area	135.089	11.48	11.8	<0.0001
Bedrooms	−43346.8	12844	−3.37	0.0008

QUESTION: How should we interpret the regression output?

ANSWER: The model is

$$\widehat{Price} = 308{,}100 + 135\ Living\ Area - 43{,}347\ Bedrooms.$$

The *R*-squared says that this model accounts for 14.6% of the variation in *Price*. But the value of *s* leads us to doubt that this model would provide very good predictions because the standard deviation of the residuals is more than $266,000. Nevertheless, we may be able to learn about home prices.

9.2 Interpreting Multiple Regression Coefficients

The multiple regression model suggests that height might be important in predicting body fat in men, but it isn't immediately obvious why that should be true. What's the relationship between *%Body Fat* and *Height* in men? We know how to approach this question; we follow the three rules from Chapter 2 and make a picture. Here's the scatterplot:

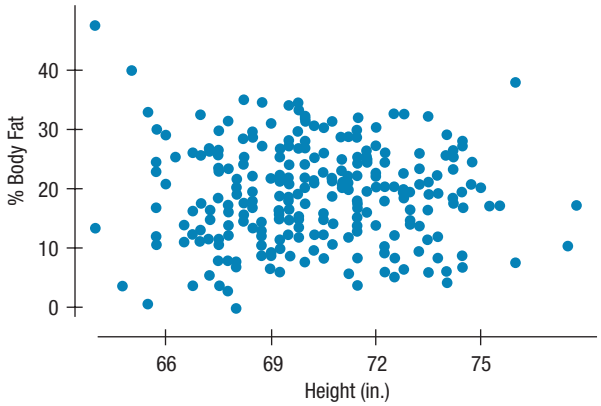


Figure 9.2
The scatterplot of *%Body Fat* against *Height* seems to say that there is little relationship between these variables.

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

That doesn't look very promising. It doesn't look like *Height* tells us much about *%Body Fat* at all. You just can't tell much about a man's *%Body Fat* from his *Height*. But in the multiple regression it certainly seemed that *Height did* contribute to the multiple regression model. How could that be?

The answer is that the multiple regression coefficient of *Height* takes account of the other predictor, *Waist* size, in the regression model. In a multiple regression, each coefficient must be interpreted as the relationship between *y* and that *x after allowing for the linear effects of the other x's on both variables*. So the coefficient for *Height* is about the relationship between *%Body Fat* and *Height* after we allow for *Waist* size. But what does that mean?

Think about all men whose waist size is about 37 inches—right in the middle of our data. If we think only about these men, what do we expect the relationship between *Height* and *%Body Fat* to be? Now a negative association makes sense because taller men probably have less body fat than shorter men who have the same waist size. Let's look at the plot. Figure 9.3 shows the men with waist sizes between 36 and 38 inches as blue dots.

What about the coefficient of *Waist*? Well, it no longer means what it did in the simple regression. Multiple regression treats all the predictors alike, so the coefficient of *Waist* now tells about how the body fat% of men of the same height tends to vary with their waist size. In the simple regression, the coefficient of *Waist* was 1.7. It hasn't changed much in the multiple regression. (It is now 1.77.) But its meaning *has* changed.

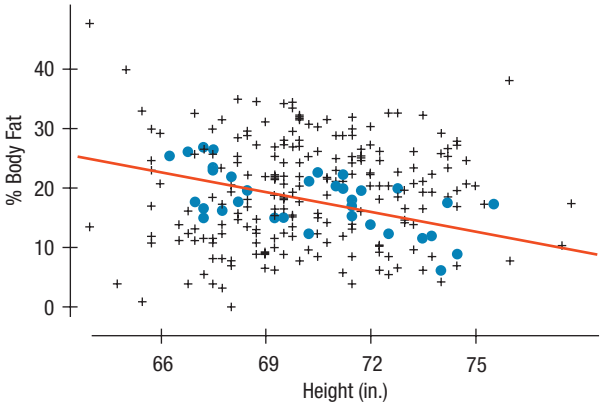


Figure 9.3
When we restrict our attention to men with waist sizes between 36 and 38 inches (points in blue), we can see a relationship between *%Body Fat* and *Height*.