GLOBAL EDITION

INTRO STATS

SIXTH EDITION

De Veaux | Velleman | Bock

SIXTH EDITION
GLOBAL EDITION

Intro Stats

Richard D. De Veaux
Williams College

Paul F. Velleman
Cornell University

David E. Bock
Ithaca High School (Retired)

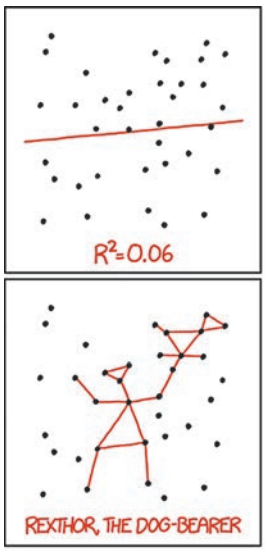
with contributions from

Brianna Heggeseth
Macalaster College

and

Susan Wang
Google Inc.

Pearson



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

EXAMPLE 7.4 Interpreting R^2

RECAP: Our regression model that predicts maximum wind speed in hurricanes based on the storm's central pressure has $R^2 = 80.6\%$.

QUESTION: What does that say about our regression model?

ANSWER: An R^2 of 80.6% indicates that 80.6% of the variation in maximum wind speed can be accounted for by the hurricane's central pressure. Other factors, such as temperature and whether the storm is over water or land, may account for some of the remaining variation.

JUST CHECKING

Back to our regression of house *Price* ($) on house *Size* (square feet):

$$\widehat{Price} = 8400 + 88.67\ Size.$$

The R^2 value is reported as 57.8%, and the standard deviation of the residuals is $53,790.

7. What does the R^2 value mean about the relationship of price and size?
8. Is the correlation of price and size positive or negative? How do you know?
9. You find that your house is worth $50,000 more than the regression model predicts. You are undoubtedly pleased, but is this actually a surprisingly large residual?

SOME EXTREME TALES

One major company developed a method to differentiate between proteins. To do so, they had to distinguish between regressions with R^2 of 99.99% and 99.98%. For this application, 99.98% was not high enough.

On the other hand, the head of a hedge fund reports that although his regressions give R^2 below 2%, they are highly successful because those used by his competition are even lower.

How Big Should R^2 Be?

R^2 is always between 0% and 100%. But what's a "good" R^2 value? The answer depends on the kind of data you are analyzing and on what you want to do with it. Just as with correlation, there is no value for R^2 that automatically determines that the regression is "good." Data from scientific experiments often have R^2 in the 80% to 90% range and even higher. Data from observational studies and surveys, though, often show relatively weak associations because it's so difficult to measure responses reliably. An R^2 of 50% to 30% or even lower might be taken as evidence of a useful regression. The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.

As we've seen, an R^2 of 100% is a perfect fit, with no scatter around the line. The s_e would be zero. All of the variance is accounted for by the model and none is left in the residuals at all. This sounds great, but it's too good to be true for real data.[10]

Along with the slope and intercept for a regression, you should always report R^2 and s_e so that readers can judge for themselves how successful the regression is at fitting the data. Statistics is about variation, and R^2 measures the success of the regression model in terms of the fraction of the variation of y accounted for by the linear model. The residual standard deviation, s_e, tells us how far the points are likely to be from the fitted line. R^2 is the first part of a regression that many people look at because, along with the scatterplot, it tells whether the regression model is even worth thinking about.

[10] If you see an R^2 of 100%, it's a good idea to figure out what happened. You may have discovered a new law of Physics, but it's much more likely that you accidentally regressed two variables that measure the same thing.

Predicting in the Other Direction—A Tale of Two Regressions

Regression slopes may not behave exactly the way you'd expect at first. Our regression model for the Burger King sandwiches was $Fat = 8.4 + 0.91\ Protein$. That equation allowed us to estimate that a sandwich with 31 grams of protein would have 36.6 grams of fat. Suppose, though, that we knew the fat content and wanted to predict the amount of protein. It might seem natural to think that by solving the equation for *Protein* we'd get a model for predicting *Protein* from *Fat*. But that doesn't work.

Our original model is $\hat{y} = b_0 + b_1 x$, but the new one needs to evaluate an $\hat{x}$ based on a value of y. We don't have y in our original model, only $\hat{y}$, and that makes all the difference. Our model doesn't fit the BK data values perfectly, and the least squares criterion focuses on the *vertical* (y) errors the model makes in using x to model y—not on *horizontal* errors related to x.

A quick look at the equations reveals why. Simply solving for x would give a new line whose slope is the reciprocal of ours. To model y in terms of x, our slope is $b_1 = r\frac{s_y}{s_x}$. To model x in terms of y, we'd need to use the slope $b_1 = r\frac{s_x}{s_y}$. That's *not* the reciprocal of ours.

Protein	Fat
$\bar{x} = 18.0$ g	$\bar{y} = 24.8$ g
$s_x = 13.5$ g	$s_y = 16.2$ g
$r = 0.76$	

Sure, if the correlation, r, were 1.0 or -1.0 the slopes *would* be reciprocals, but that would happen only if we had a perfect fit. Real data don't follow perfect straight lines, so in the real world y and $\hat{y}$ aren't the same, r is a fraction, and the slopes of the two models are not simple reciprocals of one another. Also, if the standard deviations were equal—for example, if we standardize both variables—the two slopes would be *the same*. Far from being reciprocals, both would be equal to the correlation—but we already knew that the correlation of x with y is the same as the correlation of y with x.

Otherwise, slopes of the two lines will not be reciprocals, so we can't derive one equation from the other. If we want to predict x from y, we need to create that model from the data. For example, to predict *Protein* from *Fat* we can't just invert the model we found before. Instead we find, the slope, $b_1 = 0.76\frac{13.5}{16.2} = 0.63$ grams of protein per gram of fat. The regression model is then $\widehat{Protein} = 2.29 + 0.63\ Fat$.

Moral of the story: Decide which variable you want to use (x) to predict values for the other (y). Then find the model that does that. If, later, you want to make predictions in the other direction, start over and create the other model from scratch.

7.7 Regression Assumptions and Conditions

The linear regression model may be the most widely used model in all of statistics. It has everything we could want in a model: two easily estimated parameters, a meaningful measure of how well the model fits the data, and the ability to predict new values. It even provides an easy way to see violations of conditions in plots of the residuals.

Like all models, though, linear models are only appropriate if some assumptions are true. We can't confirm assumptions, but we often can check related conditions.

First, be sure that both variables are quantitative. It makes no sense to perform a regression on categorical variables. After all, what could the slope possibly mean? Always check the **Quantitative Variables Condition**.[11]

The linear model only makes sense if the relationship is linear. It is easy to check the associated **Straight Enough Condition**. Just look at the scatterplot of y vs. x. You don't need a *perfectly* straight plot, but it must be straight enough for the linear model to make sense. If you try to model a curved relationship with a straight line, you'll usually get a regression model that misses all the interesting things in your data. If the scatterplot is not straight enough, stop here. You can't use a linear model for *any* two variables, even if they are related. They must have a *linear* association, or the model won't mean a thing.

[11] Actually, there are ways to introduce categorical predictors into a linear model. But we'll need a bigger linear model. We'll see that in Chapter 9.

Make a Picture (or Two)

You can't check the conditions just by checking boxes. You need to examine both the original scatterplot of *y* against *x* before you fit the model, and the plot of residuals afterward. These plots can save you from making embarrassing errors and losing points on the exam.

For the standard deviation of the residuals to summarize the scatter of all the residuals, the residuals must share the same spread. That's an assumption. But if the scatterplot of *y* vs. *x* looks equally spread out everywhere and (often more vividly) if the *residual plot* of residuals vs. predicted values also has a consistent spread, then the assumption is reasonable. The most common violation of that equal variance assumption is for the residuals to spread out more for *larger* values of *x*, so a good mnemonic for this check is the **Does the Plot Thicken? Condition**.

Outlying points can dramatically change a regression model. They can even change the sign of the slope, which would give a very different impression of the relationship between the variables if you only look at the regression model. So check the **Outlier Condition**. Check both the scatterplot of *y* against *x*, and the residual plot to be sure there are no outliers. The residual plot often shows violations more clearly and may reveal other unexpected patterns or interesting quirks in the data. Of course, any outliers are likely to be interesting and informative, so be sure to look into why they are unusual.

To summarize:

Before starting, be sure to check the

- **Quantitative Variable Condition** If either *y* or *x* is categorical, you can't make a scatterplot and you can't perform a regression. Stop.

From the scatterplot of *y* against *x*, check the

- **Straight Enough Condition** Is the relationship between *y* and *x* straight enough to proceed with a linear regression model?
- **Outlier Condition** Are there any outliers that might dramatically influence the fit of the least squares line?
- **Does the Plot Thicken? Condition** Does the spread of the data around the generally straight relationship seem to be consistent for all values of *x*?

After fitting the regression model, make a plot of residuals against the predicted values and look for

- Any bends that would violate the **Straight Enough Condition**,
- Any outliers that weren't clear before, and
- Any change in the spread of the residuals from one part of the plot to another.

STEP-BY-STEP EXAMPLE

Regression

If you plan to hit the fast-food joints for lunch, you should have a good breakfast. Nutritionists, concerned about "empty calories" in breakfast cereals, recorded facts about 77 cereals, including their *Calories* per serving and *Sugar* content (in grams).[12] (Data in **Cereals**)

QUESTION: How are calories and sugar content related in breakfast cereals?

THINK

PLAN State the problem.

I am interested in the relationship between sugar content and calories in cereals.

VARIABLES Name the variables and report the W's.

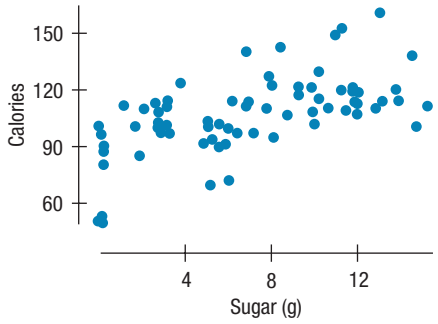
I have two quantitative variables, *Calories* and *Sugar* content per serving, measured on 77 breakfast cereals. The units of measurement are calories and grams of sugar, respectively.

[12]OK, we lied about the nutritionists. Some students went to a supermarket, read the nutrition labels of the cereals in the aisle, and recorded the information.

Check the conditions for a regression by making a picture. Never fit a regression without looking at the scatterplot first. Calories are reported to the nearest 10 calories and sugar content is reported to the nearest gram, so many cereals have the same values for both variables. In order to show them better, we have added a little random noise to both variables in the scatterplot, a process called *jittering*. Of course, we use the actual values for calculation.

✓ **Quantitative Variables**: Both variables are quantitative. The units of measurement for *Calories* and *Sugar* are calories and grams, respectively.

I'll check the remaining conditions from the scatterplot:



✓ **Outlier Condition**: There are no obvious outliers or groups.

✓ **Straight Enough**: The scatterplot looks straight to me.

✓ **Does the Plot Thicken?** The spread around the line looks about the same throughout, but I'll check it again in the residuals.

I will fit a regression model to these data.

SHOW **MECHANICS** If there are no clear violations of the conditions, fit a straight line model of the form $\hat{y} = b_0 + b_1 x$ to the data. Summary statistics give the building blocks of the calculation (though one would usually use software to calculate these quantities).

Calories

$\bar{y} = 106.88$ *calories*

$s_y = 19.48$ *calories*

Sugar

$\bar{x} = 6.94$ *grams*

$s_x = 4.42$ *grams*

Correlation

$r = 0.564$

Find the slope.

$$b_1 = r\frac{s_y}{s_x} = 0.564\,\frac{19.48}{4.42}$$

$= 2.49$ *calories per gram of sugar*

Find the intercept.

$$b_0 = \bar{y} - b_1\bar{x} = 106.88 - 2.49(6.94)$$

$= 89.6$ *calories*

Write the equation, using meaningful variable names.

So the least squares line is

$\hat{y} = 89.6 + 2.49\,x$,

or $\widehat{Calories} = 89.6 + 2.49$ Sugar.

State the value of R^2.

Squaring the correlation gives

$R^2 = 0.564^2 = 0.318$ or 31.8%.

TELL

CONCLUSION Describe what the model says in words and numbers. Be sure to use the names of the variables and their units.

The key to interpreting a regression model is to start with the phrase "b_1 *y*-units per *x*-unit," substituting the estimated value of the slope for b_1 and the names of the respective units.

The intercept is then a starting or base value. It may (as in this example) be meaningful, or (when $x = 0$ is not realistic) it may just be a starting value.

R^2 gives the fraction of the variability of *y* accounted for by the linear regression model.

Find the standard deviation of the residuals, s_e, and compare it to the original, s_y.

The scatterplot shows a positive, linear relationship and no outliers. The least squares regression line fit through these data has the equation

$$\widehat{Calories} = 89.6 + 2.49\ Sugar.$$

The slope says that we expect cereals to have, on average, about 2.49 calories per gram of sugar above a base of 89.6 calories.

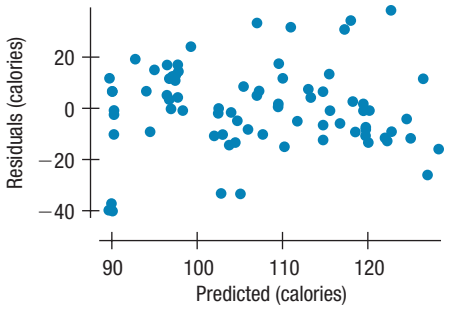
The intercept predicts that a serving of a sugar-free cereal would average about 89.6 calories.

The R^2 says that 31.8% of the variability in *Calories* is accounted for by variation in *Sugar* content.

$s_e = 16.2$ calories. That's smaller than the original SD of 19.5, but still fairly large. A prediction error of plus or minus 32.4 calories may be too large for the regression model to be useful.

THINK AGAIN

CHECK AGAIN Even though we looked at the scatterplot *before* fitting a regression model, a plot of the residuals is essential to any regression analysis because it is the best check for additional patterns and interesting quirks in the data. (As we did earlier, the points have been jittered to see the pattern more clearly.)



The residuals show a horizontal direction, a shapeless form, and roughly equal scatter for all predicted values. The linear model appears to be appropriate.

REGRESSION: ADJECTIVE, NOUN, OR VERB?

You may see the term *regression* used in different ways. There are many ways to fit a line to data, but the term "regression line" or "regression" without any other qualifiers always means least squares. People also use *regression* as a verb when they speak of *regressing* a *y*-variable on an *x*-variable to mean fitting a linear model.

Reality Check: Is the Regression Reasonable?

Statistics don't come out of nowhere. They are based on data. The results of a statistical analysis should reinforce your common sense, not fly in its face. If the results are surprising, then either you've learned something new about the world or your analysis is wrong.

Whenever you perform a regression, think about the coefficients and ask whether they make sense. Is a slope of 2.5 calories per gram of sugar reasonable? That's hard to say right off. We know from the summary statistics that a typical cereal has about 100 calories and 7 grams of sugar per serving. A gram of sugar contributes some calories (actually, 4, but you don't need to know that), so calories should go up with increasing sugar. The direction of the slope seems right.

To see if the *size* of the slope is reasonable, a useful trick is to consider its order of magnitude. Start by asking if deflating the slope by a factor of 10 seems reasonable. Is 0.25 calories per gram of sugar enough? The 7 grams of sugar found in the average cereal would contribute less than 2 calories. That seems too small.

Then try inflating the slope by a factor of 10. Is 25 calories per gram reasonable? The average cereal would have 175 calories from sugar alone. The average cereal has only 100 calories per serving, though, so that slope seems too big.

We have tried inflating the slope by a factor of 10 and deflating it by 10 and found both to be unreasonable. So, like Goldilocks, we're left with the value in the middle that's just right. And an increase of 2.5 calories per gram of sugar is certainly *plausible.*

The small effort of asking yourself whether the regression equation is plausible is repaid whenever you catch errors or avoid saying something silly or absurd about the data. It's too easy to take something that comes out of a computer at face value and assume that it makes sense.

Always be skeptical and ask yourself if the answer is reasonable.

EXAMPLE 7.5 Causation and Regression

RECAP: Our regression model predicting hurricane wind speeds from the central pressure was reasonably successful. The negative slope indicates that, in general, storms with lower central pressures have stronger winds.

QUESTION: Can we conclude that lower central barometric pressure *causes* the higher wind speeds in hurricanes?

ANSWER: No. While it may be true that lower pressure causes higher winds, a regression model for observed data such as these cannot demonstrate causation. Perhaps higher wind speeds reduce the barometric pressure, or perhaps both pressure and wind speed are driven by some other variable we have not observed.

(As it happens, in hurricanes it is reasonable to say that the low central pressure at the eye is responsible for the high winds because it draws moist, warm air into the center of the storm, where it swirls around, generating the winds. But as is often the case, things aren't quite that simple. The winds themselves contribute to lowering the pressure at the center of the storm as it becomes organized into a hurricane. The lesson is that to understand causation in hurricanes, we must do more than just model the relationship of two variables; we must study the mechanism itself.)

WHAT CAN GO WRONG?

There are many ways in which data that appear at first to be good candidates for regression analysis may be unsuitable. And there are ways that people use regression that can lead them astray. Here's an overview of the most common problems. We'll discuss them at length in the next chapter.

- **Don't fit a straight line to a nonlinear relationship.** Linear regression is suited only to relationships that are, well, *linear.* Fortunately, we can often improve the linearity easily by using re-expression. We'll come back to that topic in Chapter 8.
- **Don't ignore outliers.** Outliers can have a serious impact on the fitted model. You should identify them and think about why they are extraordinary. If they turn out not to be obvious errors, read the next chapter for advice.
- **Don't invert the regression.** The BK regression model was $\widehat{Fat} = 8.4 + 0.91\ Protein$. Knowing protein content, we can predict the amount of fat. But that doesn't let us switch the regression around. We can't use this model to predict protein values from fat values. To model y from x, the least squares slope is $b_1 = r\frac{s_y}{s_x}$. To model x in terms of y, we'd find $b_1 = r\frac{s_x}{s_y}$. That's not the