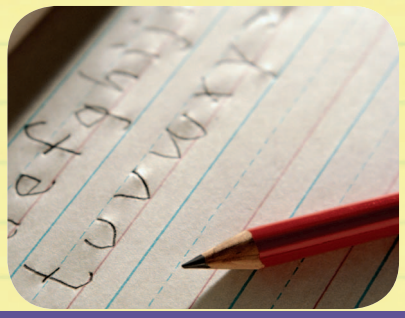
PEARSON NEW
INTERNATIONAL EDITION

Intro Stats
Richard D. De Veaux
Paul F. Velleman David E. Bock
Fourth Edition



Pearson New International Edition

Intro Stats
Richard D. De Veaux
Paul F. Velleman David E. Bock
Fourth Edition

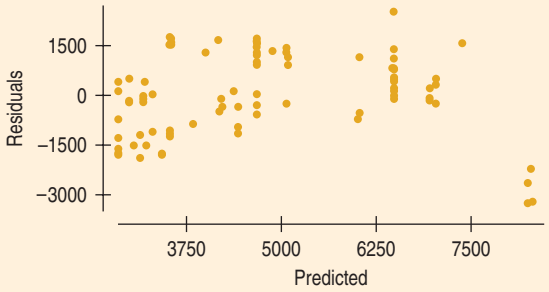
PEARSON®

The outlier is the Husqvarna TE 510 Centennial. Most of its components are handmade exclusively for this model, including extensive use of carbon fiber throughout. That may explain its $19,500 price tag! Clearly, the TE 510 is not like the other bikes. We'll set it aside for now and look at the data for the remaining dirt bikes.

QUESTION: What effect will removing this outlier have on the regression? Describe how the slope, R^2, and s_e will change.

ANSWER: The TE 510 was an influential point, tilting the regression line upward. With that point removed, the regression slope will get smaller. With that dirt bike omitted, the pattern becomes more consistent, so the value of R^2 should get larger and the standard deviation of the residuals, s_e, should get smaller.

With the outlier omitted, here's the new regression and a scatterplot of the residuals:



Dependent variable is MSRP
R-squared = 61.3% s = 1237

Variable	Coefficient
Intercept	2411.02
Displacement	9.05450

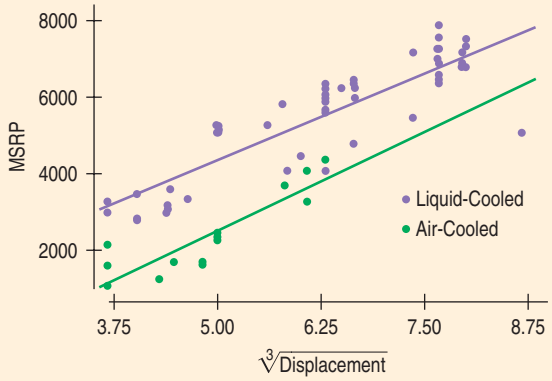
QUESTION: What do you see in the residuals plot?

ANSWER: The points at the far right don't fit well with the other dirt bikes. Overall, there appears to be a bend in the relationship, so a linear model may not be appropriate.

Let's try a re-expression. Here's a scatterplot showing *MSRP* against the cube root of *Displacement* to make the relationship closer to straight. (Since displacement is measured in cubic centimeters, its cube root has the simple units of centimeters.) In addition, we've colored the plot according to the cooling method used in the bike's engine: liquid or air. Each group is shown with its own regression line, as we did for the cereals on different shelves.

QUESTION: What does this plot say about dirt bikes?

ANSWER: There appears to be a positive, linear relationship between *MSRP* and the cube root of *Displacement*. In general, the larger the engine a bike has, the higher the suggested price. Liquid-cooled dirt bikes, however, typically cost more than air-cooled bikes with comparable displacement. A few liquid-cooled bikes appear to be much less expensive than we might expect, given their engine displacements (but without separating the groups we might have missed that because they look more like air-cooled bikes.)



Source: lib.stat.cmu.edu/datasets/dirtbike_aug.csv, "The Dirt on Bikes: An Illustration of CART Models for Brand Differentiation," Jiang Lu, Joseph B. Kadane, and Peter Boatwright.

WHAT CAN GO WRONG?

This entire chapter has held warnings about things that can go wrong in a regression analysis. So let's just recap. When you make a linear model:

- **Make sure the relationship is straight.** Check the Straight Enough Condition. Always examine the residuals for evidence that the Linearity Assumption has failed. It's often easier to see deviations from a straight line in the residuals plot than in the scatterplot of the original data. Pay special attention to the most extreme residuals because they may have something to add to the story told by the linear model.
- **Be on guard for different groups in your regression.** Check for evidence that the data consist of separate subsets. If you find subsets that behave differently, consider fitting a different linear model to each subset.
- **Beware of extrapolating.** Beware of extrapolation beyond the x-values that were used to fit the model. Although it's common to use linear models to extrapolate, the practice is dangerous.
- **Beware especially of extrapolating into the future!** Be especially cautious about extrapolating into the future with linear models. To predict the future, you must assume that future changes will continue at the same rate you've observed in the past. Predicting the future is particularly tempting and particularly dangerous.
- **Look for unusual points.** Unusual points always deserve attention and may well reveal more about your data than the rest of the points combined. Always look for them and try to understand why they stand apart. A scatterplot of the data is a good way to see high-leverage and influential points. A scatterplot of the residuals against the predicted values is a good tool for finding points with large residuals.
- **Beware of high-leverage points and especially of those that are influential.** Influential points can alter the regression model a great deal. The resulting model may say more about one or two points than about the overall relationship.
- **Consider comparing two regressions.** To see the impact of outliers on a regression, it's often wise to run two regressions, one with and one without the extraordinary points, and then to discuss the differences.
- **Treat unusual points honestly.** If you remove enough carefully selected points, you will eventually get a regression with a high R^2, but it won't give you much understanding. Some variables are not related in a way that's simple enough for a linear model to fit very well. When that happens, report the failure and stop.
- **Beware of lurking variables.** Think about lurking variables before interpreting a linear model. It's particularly tempting to explain a strong regression by thinking that the x-variable *causes* the y-variable. A linear model alone can never demonstrate such causation, in part because it cannot eliminate the chance that a lurking variable has caused the variation in both x and y.
- **Watch out when dealing with data that are summaries.** Be cautious in working with data values that are themselves summaries, such as means or medians. Such statistics are less variable than the data on which they are based, so they tend to inflate the impression of the strength of a relationship.

CONNECTIONS

We should always be alert to things that could go wrong if we were to use statistics without thinking carefully. Regression opens new vistas of potential problems. But each one relates to issues we've thought about before.

It is always important that our data be from a single homogeneous group and not made up of disparate groups. We looked for multiple modes in single variables. Now we

check scatterplots for evidence of subgroups in our data. As with modes, it's often best to split the data and analyze the groups separately.

Our concern with unusual points and their potential influence also harkens back to our earlier concern with outliers in histograms and boxplots—and for many of the same reasons. As we've seen here, regression offers such points new scope for mischief.

What Have We Learned?

Learning Objectives

Be skeptical of regression models. Always plot and examine the residuals for unexpected behavior. Be alert to a variety of possible violations of the standard regression assumptions and know what to do when you find them.

Be alert for subgroups in the data.

- Often these will turn out to be separate groups that should not be analyzed together in a single analysis.
- Often identifying subgroups can help us understand what is going on in the data.

Be especially cautious about extrapolating beyond the data.

Look out for unusual and extraordinary observations.

- Cases that are extreme in x have high leverage and can affect a regression model strongly.
- Cases that are extreme in y have large residuals, and are called outliers.
- Cases that have both high leverage and large residuals are influential. Setting them aside will change the regression model, so you should consider whether their influence on the model is appropriate or desirable.

Interpret regression models appropriately. Don't infer causation from a regression model.

Notice when you are working with summary values.

- Summaries vary less than the data they summarize, so they may give the impression of greater certainty than your model deserves.

Diagnose and treat nonlinearity.

- If a scatterplot of y vs. x isn't straight, a linear regression model isn't appropriate.
- Re-expressing one or both variables can often improve the straightness of the relationship.
- The powers, roots, and the logarithm provide an ordered collection of re-expressions so you can search up and down the "ladder of powers" to find an appropriate one.

Review of Terms

Extrapolation Although linear models provide an easy way to predict values of y for a given value of x, it is unsafe to predict for values of x far from the ones used to find the linear model equation. Such extrapolation may pretend to see into the future, but the predictions should not be trusted.

Outlier Any data point that stands away from the others can be called an outlier. In regression, cases can be extraordinary in two ways: by having a large residual—being an outlier—or by having high leverage.

Leverage Data points whose x-values are far from the mean of x are said to exert leverage on a linear model. High-leverage points pull the line close to them, and so they can have a large effect on the line, sometimes completely determining the slope and intercept. With high enough leverage, their residuals can be deceptively small.

Influential point A point that, if omitted from the data, results in a very different regression model.

Lurking variable A variable that is not explicitly part of a model but affects the way the variables in the model appear to be related. Because we can never be certain that observational data are not hiding a lurking variable that influences both x and y, it is never safe to conclude that a linear model demonstrates a causal relationship, no matter how strong the linear association.

On the Computer REGRESSION DIAGNOSIS

Most statistics technology offers simple ways to check whether your data satisfy the conditions for regression. We have already seen that these programs can make a simple scatterplot. They can also help check the conditions by plotting residuals.

DATA DESK

- Click on the **HyperView** menu on the **Regression** output table. A menu drops down to offer scatterplots of residuals against predicted values, Normal probability plots of residuals, or just the ability to save the residuals and predicted values.
- Click on the name of a predictor in the regression table to be offered a scatterplot of the residuals against that predictor.

COMMENTS

If you change any of the variables in the regression analysis, Data Desk will offer to update the plots of residuals.

EXCEL

The Data Analysis add-in for Excel includes a Regression command. The dialog box it shows offers to make plots of residuals.

COMMENTS

Do not use the Normal probability plot offered in the regression dialog. It is not what it claims to be and is wrong.

JMP

- From the **Analyze** menu, choose **Fit Y by X**. Select **Fit Line**.
- Under Linear Fit, select **Plot Residuals**. You can also choose to **Save Residuals**.
- Subsequently, from the **Distribution** menu, choose **Normal quantile plot** or **histogram** for the residuals.

MINITAB

- From the **Stat** menu, choose **Regression**.
- From the **Regression** submenu, select **Regression** again.
- In the Regression dialog, enter the response variable name in the "Response" box and the predictor variable name in the "Predictor" box.
- To specify saved results, in the Regression dialog, click **Storage**.
- Check "Residuals" and "Fits." Click **OK**.
- To specify displays, in the Regression dialog, click **Graphs**.
- Under "Residual Plots," select "Individual plots" and check "Residuals versus fits."
- Click **OK**. Now back in the Regression dialog, click **OK**. Minitab computes the regression and the requested saved values and graphs.

R

Save the regression model object by giving it a name, such as "myreg":

- **myreg = lm(Y~X) or myreg = lm(Y~X, data = DATA)** where DATA is the name of the data frame.
- **plot(residuals(myreg)~predict(myreg))** plots the residuals against the predicted values.
- **qqnorm(residuals(myreg))** gives a normal probability plot of the residuals.
- **plot(myreg)** gives similar plots (but not exactly the same).

SPSS

- From the **Analyze** menu, choose **Regression**.
- From the Regression submenu, choose **Linear**.
- After assigning variables to their roles in the regression, click the "**Plots ...**" button.

In the Plots dialog, you can specify a Normal probability plot of residuals and scatterplots of various versions of standardized residuals and predicted values.

COMMENTS

A plot of ***ZRESID** against ***PRED** will look most like the residual plots we've discussed. SPSS standardizes the residuals by dividing by their standard deviation. (There's no need to subtract their mean; it must be zero.) The standardization doesn't affect the scatterplot.

STATCRUNCH

To create a residuals plot:

- Click on **Stat**.
- Choose **Regression » Simple Linear** and choose X and Y.
- Click on **Next** and click on **Next** again.
- Indicate which type of residuals plot you want.
- Click on **Calculate.**

COMMENTS

Note that before you click on **Next** for the second time you may indicate that you want to save the values of the residuals. Residuals becomes a new column, and you may use that variable to create a histogram or residuals plot.

TI-83/84 PLUS

- To make a residuals plot, set up a **STATPLOT** as a scatterplot.
- Specify your explanatory data list as **Xlist**.
- For **Ylist**, import the name **RESID** from the **LIST NAMES** menu. **ZoomStat** will now create the residuals plot.

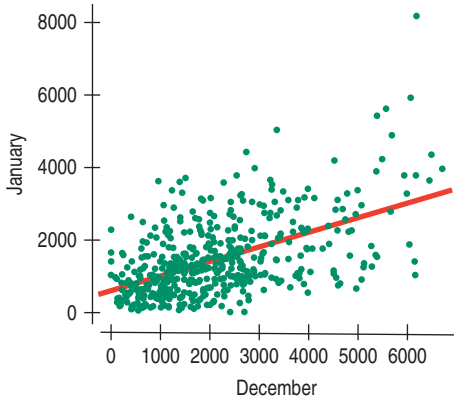
COMMENTS

Each time you execute a **LinReg** command, the calculator automatically computes the residuals and stores them in a data list named **RESID**. If you want to see them, go to **STAT EDIT**. Space through the names of the lists until you find a blank. Import **RESID** from the **LIST NAMES** menu. Now every time you have the calculator compute a regression analysis, it will show you the residuals.

Exercises

Section 1

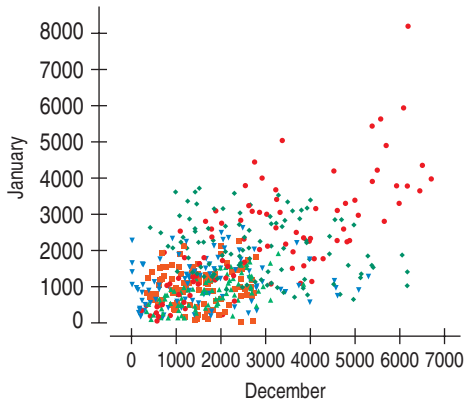
1. Credit card spending An analysis of spending by a sample of credit card bank cardholders shows that spending by cardholders in January (*Jan*) is related to their spending in December (*Dec*):



The assumptions and conditions of the linear regression seemed to be satisfied and an analyst was about to predict January spending using the model

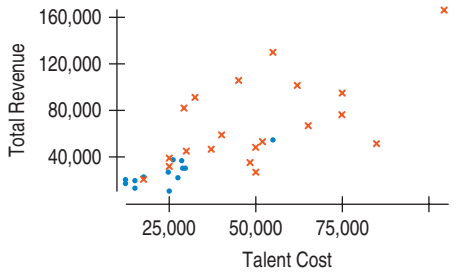
$$\widehat{Jan} = \$612.07 + 0.403\,Dec.$$

Another analyst worried that different types of cardholders might behave differently. She examined the spending patterns of the cardholders and placed them into five market *Segments*. When she plotted the data using different colors and symbols for the five different segments, she found the following:



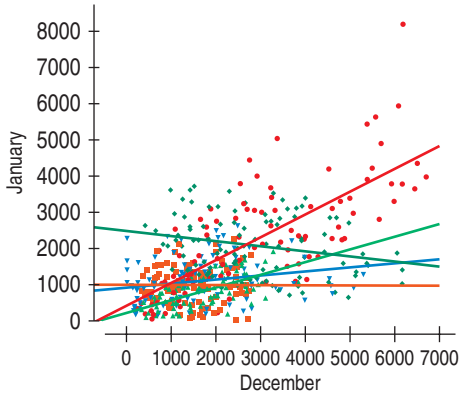
Look at this plot carefully and discuss why she might be worried about the predictions from the model $\widehat{Jan} = \$612.07 + 0.403\,Dec$.

2. Revenue and talent cost A concert production company examined its records. The manager made the following scatterplot. The company places concerts in two venues, a smaller, more intimate theater (plotted with blue circles) and a larger auditorium-style venue (red x's).



a) Describe the relationship between *Talent Cost* and *Total Revenue*. (Remember: direction, form, strength, outliers.)
b) How are the results for the two venues similar?
c) How are they different?

3. Market segments The analyst in Exercise 1 tried fitting the regression line to each market segment separately and found the following:



What does this say about her concern in Exercise 1? Was she justified in worrying that the overall model $\widehat{Jan} = \$612.07 + 0.403\,Dec$ might not accurately summarize the relationship? Explain briefly.