This measure tells us how well the line predicts y values.

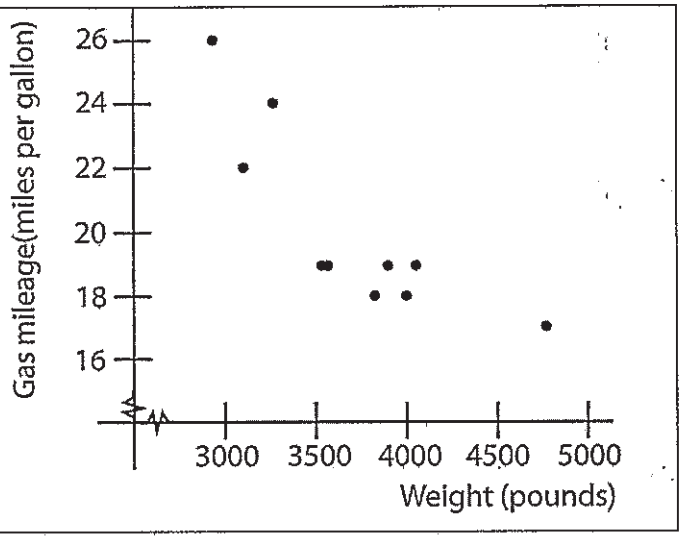
Statistics
Class Notes

Diagnostics on the Least-Squares Regression Line: The Coefficient of Determination and Residual Analysis (Section 4.3)

Here, we analyze a measure of the least-squares regression line.

expl 1: Consider the car weight and gas mileage data we have worked with previously.
(source: each manufacturer's website through textbook)

Car	Weight, x (pounds)	Gas Mileage, y (mpg)
Buick LaCrosse	4724	17
Cadillac XTS	4006	18
Chevy Cruze	3097	22
Chevy Impala	3555	19
Chrysler 300	4029	19
Dodge Charger	3934	19
Dodge Dart	3242	24
Ford Focus	2960	26
Ford Mustang	3530	19
Lincoln MKZ	3823	18



You'll recall, in the last section of notes, we found the least-squares regression line to be $\hat{y} = -.00468x + 37.357$ where x represents the weight of the car (in pounds) and $\hat{y}$ represents the gas mileage (in miles per gallon).

Which is the explanatory and which is the response variable?

explan: weight

response: gas mileage

We found r, the coefficient of correlation, to be -.84158. This told us the line fit the data well with a negative slope. The calculator also gives us a value for r^2. What is that?

We know that as the weight of the car increases, the gas mileage decreases. This line shows this relationship. But how well does it do?

This value of r^2 measures how well the regression line describes the relationship between the explanatory and response variables.

Remember that r^2 is just our old value of r, squared.

Definition: Coefficient of Determination: The **coefficient of determination**, R^2, measures the *proportion of total variation* in the response variable that is explained by the least-squares regression line.

Properties of the Coefficient of Determination, r^2 or R^2:

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \le R^2 \le 1$.

If $R^2 = 0$, then the line has *no* explanatory value.

The closer R^2 is to 1, the better the line describes how changes in the explanatory variable affect the value of the response variable.

If $R^2 = 1$, then the line explains 100% of the variation in the response variable.

There is a technical difference between r^2 and R^2 but for the least-squares linear regression we will use them interchangeably.

expl 1 continued: The value for r^2 for the car weight/gas mileage example is given, by the calculator, as .708. You could also get that by taking the value for r (-.8416) and squaring it. Turn this into a percent and complete the sentence below. Round to the nearest tenth of a percent.

70.8 % of the variation in gas mileage can be explained by the least-squares regression line.

Turn the value into a percentage and you can say…

$r^2 = 0.708$

70.8%

This means the remainder of the variation (29.2%) in gas mileage is explained by other factors.

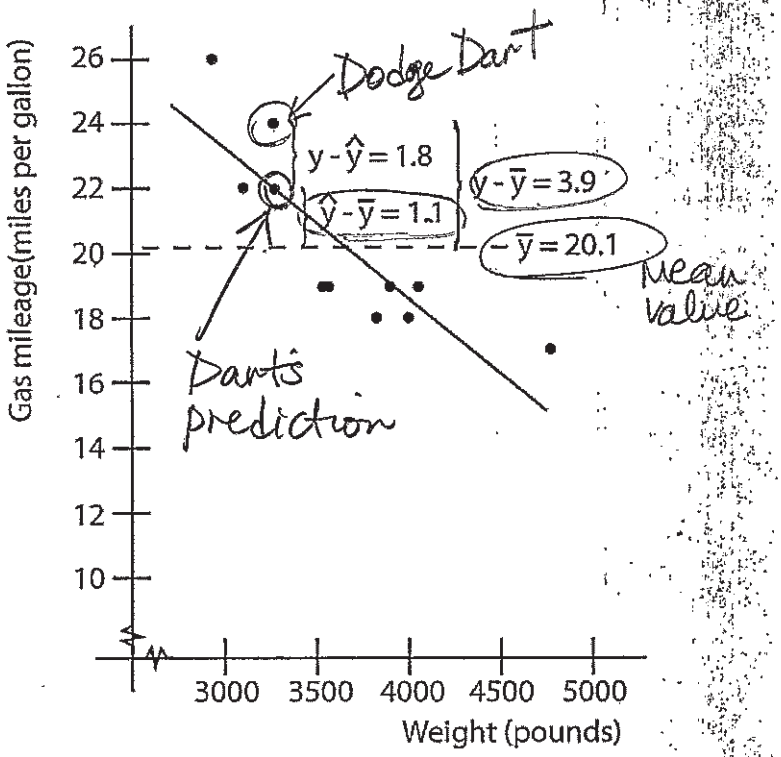
$100 - 70.8 = 29.2$

- age of car
- going uphill / downhill
- tire pressure
- car shape
- in town vs highway

Deviations:

We return to this car weight/gas mileage example. The mean value of the response variable is $\bar{y} = 20.1$ miles per gallon. This is a simple average of the values in the table.

Consider the Dodge Dart (which has an observed gas mileage of 24 miles per gallon and a weight of 3,242 pounds). Find this point on the graph now and circle it. Let's compare our regression prediction against reality and against the mean of the sample.



The difference between the observed value and the mean value is $y - \bar{y} = 24 - 20.1 = 3.9$ miles per gallon. This is called the **total deviation**. Do you see it labeled on the graph?

On the other hand, the least-squares line gives us

$$\hat{y} = -.00468x + 37.357$$
$$= -.00468(3242) + 37.357$$
$$\approx 22.2$$

The difference between this predicted value (shown as a red dot) and the mean, or $\hat{y} - \bar{y} = 22.2 - 20.1 = 1.1$ miles per gallon, is called the **explained deviation**.

Finally, the difference between the observed value (of 24 miles per gallon) and the predicted value (of 22.2 miles per gallon), or $y - \hat{y} = 24 - 22.2 = 1.8$, is called the **unexplained deviation**. We have seen this also called the **residual**. ★ residual $= y - \hat{y}$.

Now we see why the first is called the total deviation. **Notice how the total deviation equals the explained deviation plus the unexplained deviation.**

It is beyond what we want to discuss here, but it can be shown that the closer the observed y-values are to the regression line (the predicted y-values), the larger R^2 will be. In other words, the value of R^2 will be closer to 1 if the points line up closer to a perfect line. You can use this to estimate the value of R^2 given a graph of points and their regression equation.

We will *not* calculate R^2 by hand. However, just know that you would use these deviations to do so. We will rely on technology to calculate R^2. ★

4.3

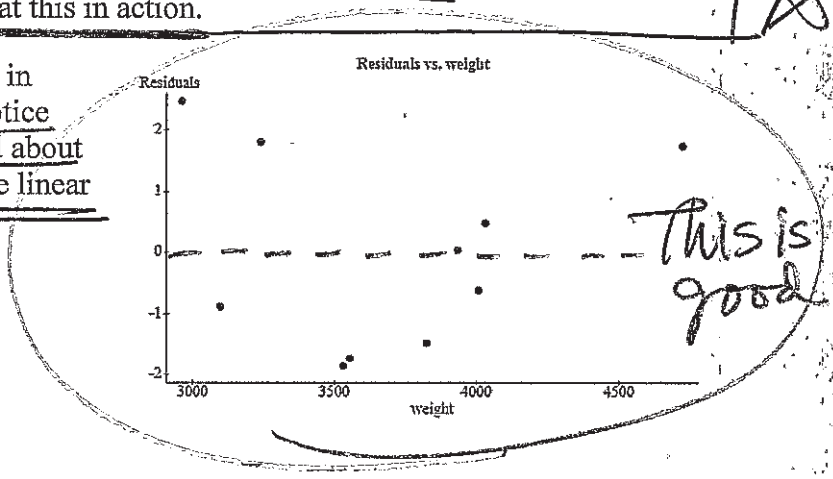
Residual Analysis of a Linear Regression Model:

Even though the correlation coefficient from earlier can be used to tell us if there is a linear relationship between x and y, it is *not* the whole story. This value may show a linear relation even though it is *not* linear. We need to draw a **residual plot** to be more confident.

Definition: Residual plot: A **residual plot** is a scatter diagram with the residuals $(y-\hat{y})$ on the vertical axis and the explanatory variable (x) on the horizontal axis.

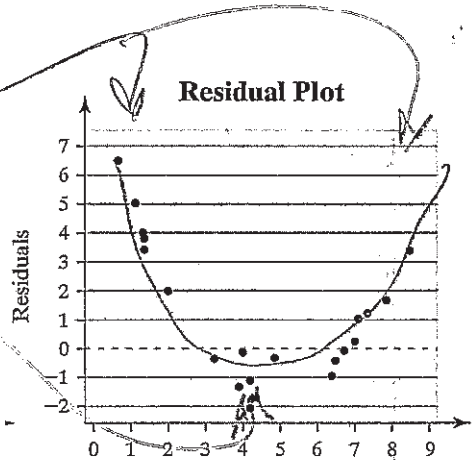
If this plot shows a discernible pattern (like a curve), then the response and predictor variable may *not* be linearly related. Let's look at this in action.

Here we see the residual plot as given in StatCrunch for our data on page 1. Notice how the points are pretty much spread about with *no pattern*. This indicates that the linear relationship really does exist.



Instructions for StatCrunch:
Within the **Stat > Regression > Simple Linear** dialog box, we select **Residuals vs. X-values** under **Graphs**. You may need to scroll down the list to get that option. You will page through the output to see the graph.

As a counterexample, look at the residual plot here for another data set. Do you notice a U-shaped pattern?



For both small and large values of x, the residuals are positive, indicating the regression line underestimates the true values. For intermediate values of x, the residuals are negative, indicating the regression line overestimates the true values.

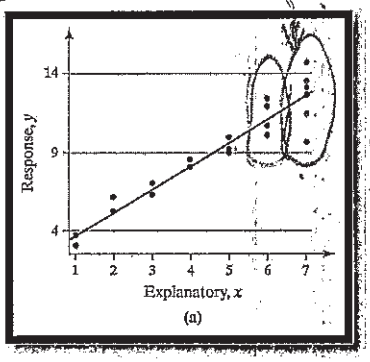
This indicates some other (non-linear) model would be more appropriate.

That is *not* all that can go wrong. Let's explore another phenomena with this residual plot. It has to do with how the residuals vary as you travel left to right on the residual plot.

Constant Error Variance:

It so happens that a strict requirement of the linear regression model is that the residuals do *not* increase or decrease as x increases. We call this **constant error variance**. Take a look at the following residual plot given along with the data's scatter plot.

This plot shows the data and regression line. All looks good, right?

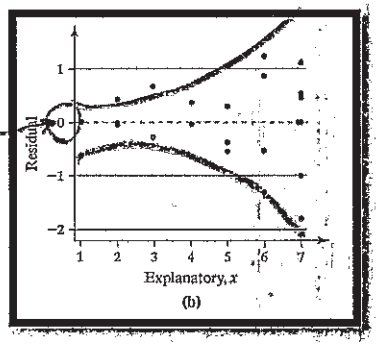


data's scatter plot

But take a closer look at the x-values of 6 and 7. Notice how they are spread out from the regression line, more so than for smaller x-values.

For large values of x, the regression line will *not* predict y-values as well.

Now take a look at the residual plot. Can you see how the residuals get further from the center value of 0 (red dashed line) as we travel to larger x-values?



residual plot

That indicates a *non*constant error variance and hence the linear model is *not* appropriate.

Oh, the fun we are having!

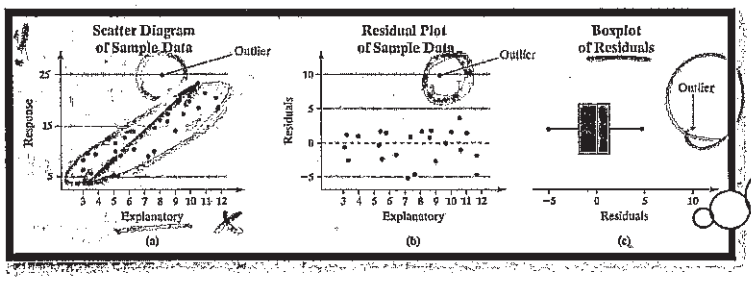
A Search for Outliers:

Recall that outliers are values (points) that do *not* fit the general pattern, or extreme observations. They will reveal themselves sometimes on the scatter plot of the data. But to be sure the point is a true outlier, we will use a boxplot of the residuals (where an asterisk will denote an outlier).

Instructions for StatCrunch:

Within the **Stat > Regression > Simple Linear** dialog box, select **Residuals** under the **Save** option. That will create a column called **Residuals** in the spreadsheet. Then follow **Graph > Boxplot** (and select the Residuals column, of course) to make the boxplot.

Here is an example.



That outlier sticks out on plots *a* and *b*. We verify it is truly an outlier by drawing the residual boxplot.

Influential Observations:

An **influential observation** is an observation that significantly affects the least-squares regression line's slope and/or y-intercept, or the value of the correlation coefficient.

To determine if a point is influential, remove the point and recalculate the regression line. If the correlation coefficient, slope, or y-intercept changes significantly, the removed point is influential.

We will *not* be doing this.