

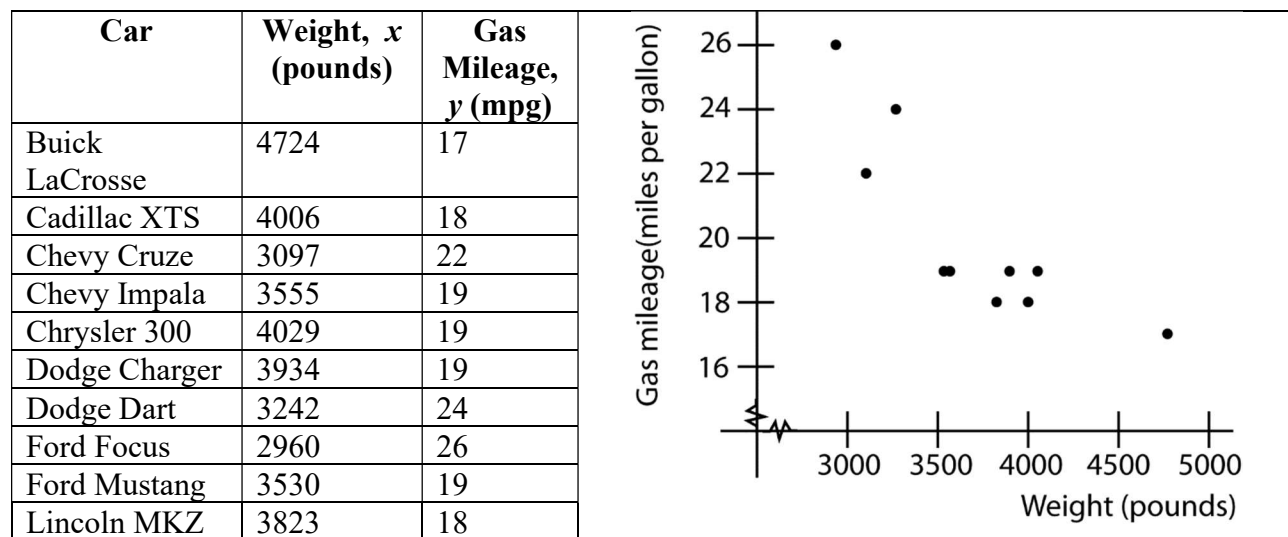
Statistics
Class Notes
Least-Squares Regression (Section 4.2)

We will now find the one line that best fits the data on a scatter plot.

We have seen how two variables can be correlated by a linear pattern. We will now work to find the equation that fits the pattern the best. The method we use finds us the **least-squares regression equation**. This can also be called the **line of best fit**.

We may want to review equations of lines and slope before our work here. There are materials on www.stlmath.com that can help. You can find them under Archived Worksheets > Algebra > Elementary Algebra Class Notes (Second Half of Class). Relevant material can be found in the notes Graphing Linear Equations, Slope of a Line and Rate of Change, and Equations of Lines.

expl 1: Consider the data in the table below. It gives the weights of various cars along with their gas mileages. How do you think those two variables are related? (source: each manufacturer's website through textbook)



Look at the scatter plot and determine if a linear pattern exists between car weight and gas mileage. Is it a positive or negative association?

Which is the explanatory and which is the response variable?

expl 2: Considering the linear pattern we see in the scatter plot, we will find a line that fits this data. We will do this by selecting two points and finding the line that goes through them. Follow the steps below. This is *not* the least-squares method.

a.) Consider *only* the points (3242, 24) and (4006, 18). Find the slope between these two points. Round to five decimal places.

$$\begin{aligned} m &= \frac{y_2 - y_1}{x_2 - x_1} \\ &= \frac{24 - 18}{3242 - 4006} \\ &= \frac{6}{-764} \\ &\approx -.00785 \end{aligned}$$

The slope between the two points

(x_1, y_1) and (x_2, y_2) is $m = \frac{y_2 - y_1}{x_2 - x_1}$.

Parts *a* and *b* are done for us.

b.) Now, find the equation of the line through these two points. Write your final answer in slope-intercept form which is $y = mx + b$. Try to use unrounded values until the very end.

$$\begin{aligned} y &= mx + b \\ y &= -.00785x + b \\ 18 &= -.00785(4006) + b \\ 18 &= -31.4607 + b \\ 49.4607 &= b \end{aligned}$$

We enter the slope and one of our points.

You can use either the $y - y_1 = m(x - x_1)$ or the $y = mx + b$ form.

$$y = -.00785x + 49.4607$$

c.) Use the equation to predict the gas mileage of my Buick LaCrosse which weighs 4,724 pounds.

Round predictions to one more decimal place than the response variable (in the data) has.

d.) Find the “observed” gas mileage for this Buick LaCrosse in the original data table. (It’s the first entry.) What is the difference between the observed gas mileage from the table and your prediction? This is called the **residual**. Would you say your prediction was a good one?

Least-Squares Regression Method:

We found a line that fits the data and it did yield good predictions of gas mileage given weight. But could we do better? We will now look at the least-squares regression method. This method finds the line that minimizes the sum of the squared residuals for all of the data. We will not go into the details of how it does this but we define the line below.

Definition: The equation of the **least-squares regression line** is given by

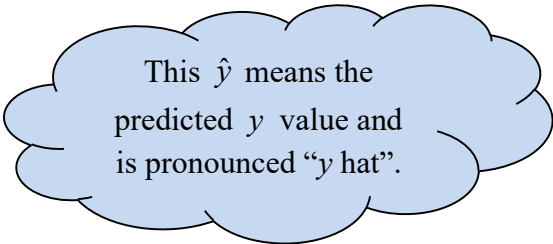
$$\hat{y} = b_1x + b_0$$

Where

$$b_1 = r \cdot \frac{s_y}{s_x}$$
 is the slope of the least-squares regression line

Where

$$b_0 = \bar{y} - b_1\bar{x}$$
 is the **y-intercept** of the least-squares regression line



This \hat{y} means the predicted y value and is pronounced “y hat”.

We will use technology to calculate these lines. These formulas are here for reference only. Know that this is what the calculator is doing.

Instructions for StatCrunch:

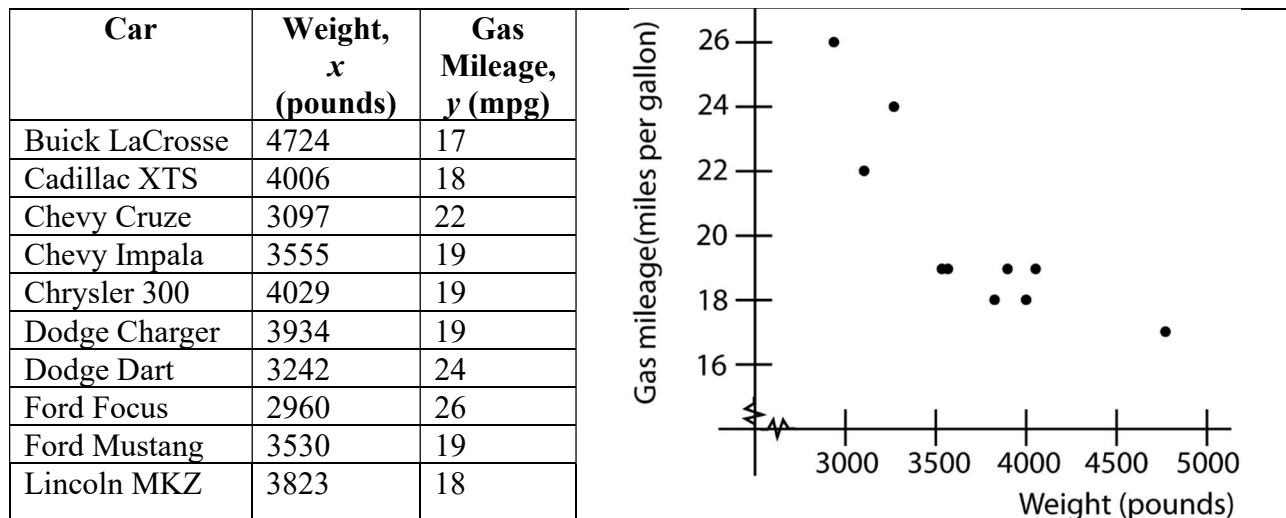
Enter the values of the explanatory (independent) variable in the column **var1**. The values of the response (dependent) variable go in column **var2**. If you are working in MML, use the overlapping rectangles next to the data to import the data automatically into StatCrunch (**Open in StatCrunch**). If the data has meaning (and is not just made up numbers), it is a good idea to rename the columns if they do not automatically.

From the top menu, select **Stat > Regression > Simple Linear**. You will need to tell it which columns contain the **X variable** and **Y variable**.

You can even use this screen to have it to predict a y -value given an x -value. Do this under the **Prediction of Y** option. Leave the **Level** at 0.95 and enter the given x -value.

Also, tell it "Fitted line plot" under **Graphs** to see the line and scatter plot. Click **Compute** at the bottom of the dialog box. You may have to scroll down the output window.

expl 3: We will again examine the data concerning car weight and gas mileage.



a.) Find the least-squares regression line using StatCrunch. Record the equation of the line and the value of the correlation coefficient r .

b.) Use the Critical Values for Correlation Coefficient table to evaluate if the line is a good fit.

Critical Values for Correlation Coefficient	
Sample size, n	Critical Value
3	0.997
4	0.950
5	0.878
10	0.632
15	0.514
30	0.361

What is our
sample size?
Is the line a
good fit?

c.) Once again, use your equation to predict the gas mileage of my Buick LaCrosse which weighs 4,724 pounds. (How does your prediction do against the observed value of 17 miles per gallon? In other words, find the **residual** of this prediction.)

Round predictions to one
more decimal
place than the response
variable (in the data) has.

Worksheet: Linear regression on your calculator:

We will explore a couple of examples with step-by-step instructions on how to find the regression equations using the calculator. Instructions include entering the data, drawing a scatter plot, finding the least-squares regression equation, and various other details. Video available on www.stlmath.com under your class' Assorted Handouts and Tutorials.

Interpreting Slope and y -intercept of the Regression Line:

expl 4: Let's return to the example of car weight and gas mileage.

a.) We know that slope tells us the ratio of $\frac{\text{rise}}{\text{run}}$ along our line. This number can be interpreted as the change in y divided by the change in x (from one point on the line to another point). What was the slope of your regression equation in the previous example? Include units. To do this, consider the units of the x and y variables.

b.) When we interpret the slope in algebra, sometimes called **average rate of change**, we are more definitive. However in statistics, these values are *not* considered to be 100% true. A different sample will yield different results. So we will state our conclusion like this:

If the weight of a car increases by 1 pound, the gas mileage would decrease by _____ on average.

OR

If the weight of a car increases by 1 pound, the *expected* gas mileage would decrease by _____.

c.) What is the y -intercept of our line? What real-world meaning does this number have?

Notice we have the word *decrease* here. Why?

Include units in your sentences.

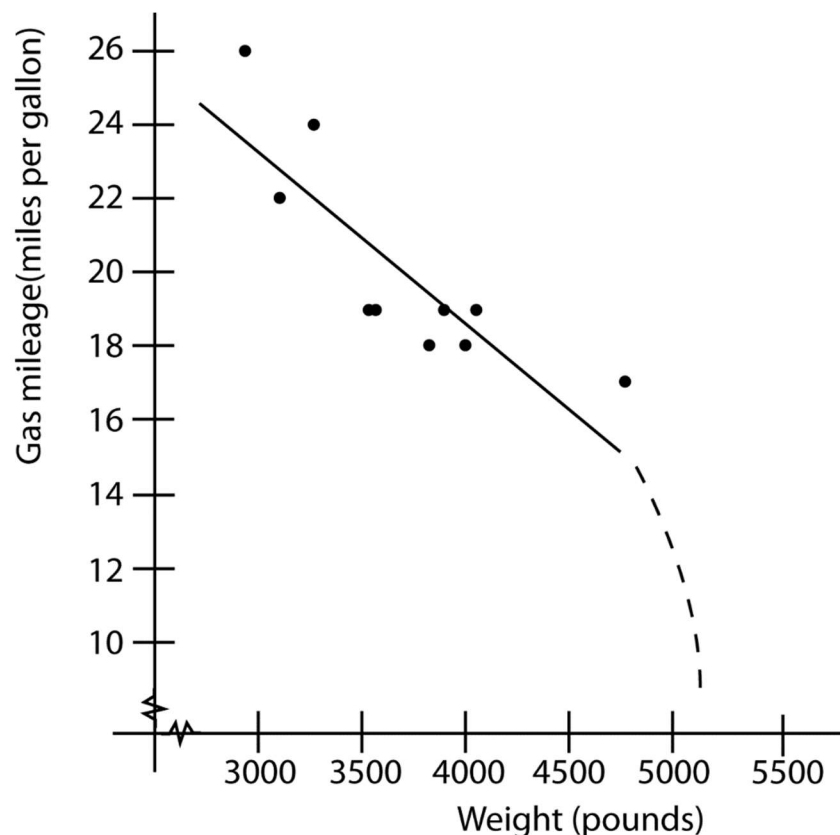
Recall the y -intercept is where $x = 0$. Is this a reasonable value for the explanatory variable? Do we have any data values near $x = 0$?

Interpolation Versus Extrapolation:

Interpolation is when we use our regression line to find a value that is within our original data set. This is a good use of regression.

Extrapolation is when we use our line to find a value that is *not* within our original data set. It is considered to be bad practice. Interpreting the y -intercept when there really are no x values near 0 is an example of extrapolation and should *not* be done.

We cannot be sure that the data even continues in a linear pattern. Consider the possible graph of car weight versus gas mileage below. Here we see the regression line drawn in (solid line) over the scatter plot. Since we have no evidence for cars over the weight of 4,724 pounds, we cannot say that the relationship does *not* take a nose dive as supposed here (dashed line).



What do we do if there is no linear relationship?

If there is *not* a linear pattern (as shown by the Critical Values for Correlation Coefficient table), we use the mean of the response variable to predict any values needed. That is, $\hat{y} = \bar{y}$. It's just as good as we can do. So average the response variable data (y -values) and use that as the prediction for any and all x -values.