4 Describing Relationships

Investigating relationships between variables is central to what we do in statistics. When we understand the relationship between two variables, we can use the value of one variable to help us make predictions about the other variable.

4.1 Scatterplots and Correlation

4.1.1 What You Will Learn

- Identify explanatory and response variables in situations where one variable helps to explain or influences the other.
- Make a scatterplot to display the relationship between two quantitative variables.
- Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify outliers in a scatterplot.
- Interpret the correlation.
- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Use technology to calculate correlation.
- Explain why association does not imply causation.

4.1.2 Explanatory and Response Variables

Explanatory Variable

Response Variable

Identify the explanatory and response variable from the examples.

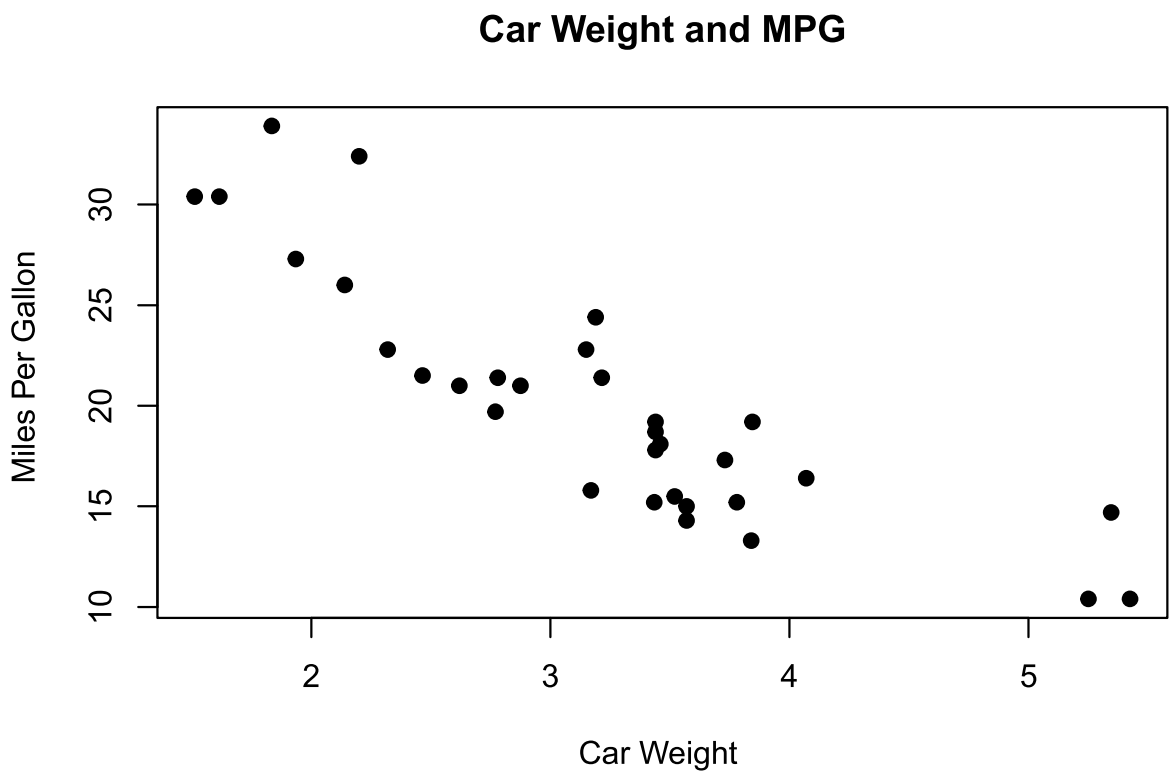
1. We think that car weight helps explain accident deaths
2. Smoking influences life expectancy.
3. Grades may be higher if students attend school more.
4. How does drinking beer affect the level of alcohol in people's blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.
5. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

Do all correlations have explanatory and response variables?

4.1.3 Displaying Relationships: Scatterplots

Although there are many ways to display the distribution of a single quantitative variable, a scatterplot is the best way to display the relationship between two quantitative variables.

Scatterplot Definition:



How to make a Scatterplot
1.

2.

3.

Mr Vanderkin thinks there may be a correlation between students attendance and their grade. below is a table of 20 students' number of absences and grades. Create a scatterplot of the data to look for a correlation.

Grades	Absences
89	9
71	14
92	6
83	9
79	9
84	9
92	6
85	9
81	11
85	9
69	13
76	7
86	10
85	8
104	4
78	11
89	6
83	9
66	14
95	4

Do you think there is a correlation between the number of absences and students' grades?

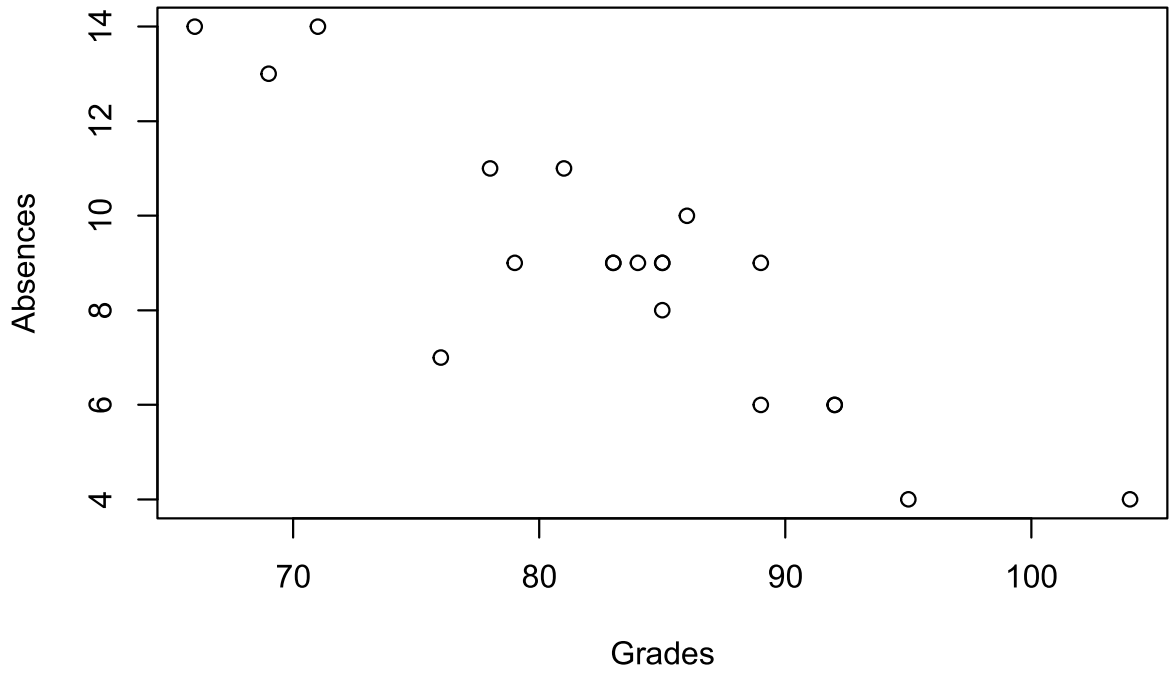
As absences increase grades _____.

As grades increase absences _____.

Does this data point to which is the explanatory and which is the response variable?

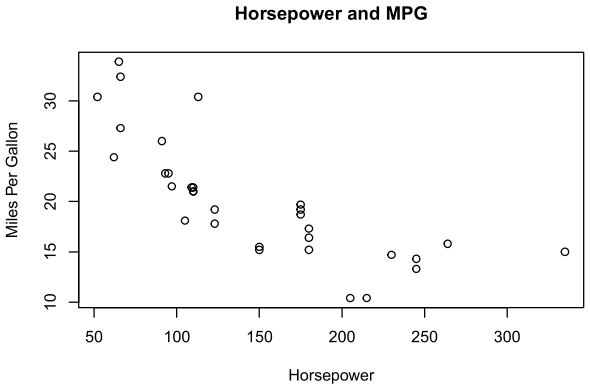
4.1.4 Describing Scatterplots

To describe a scatterplot, follow the basic strategy of data analysis from Chapter 1: look for patterns and important departures from those patterns.



This scatterplot shows a _____ association between absences and grade averages.

mtcars is a popular data set from the 1974 Motor Trend magazine the compares 10 variables from 32 cars. Describe the association between mpg and horsepower



Positive Association	Negative Association	No Association

How to describe a Scatterplot
1. Direction

2. Form

3. Strength

4. Unusual Features

Describe the 2 Scatterplots below:

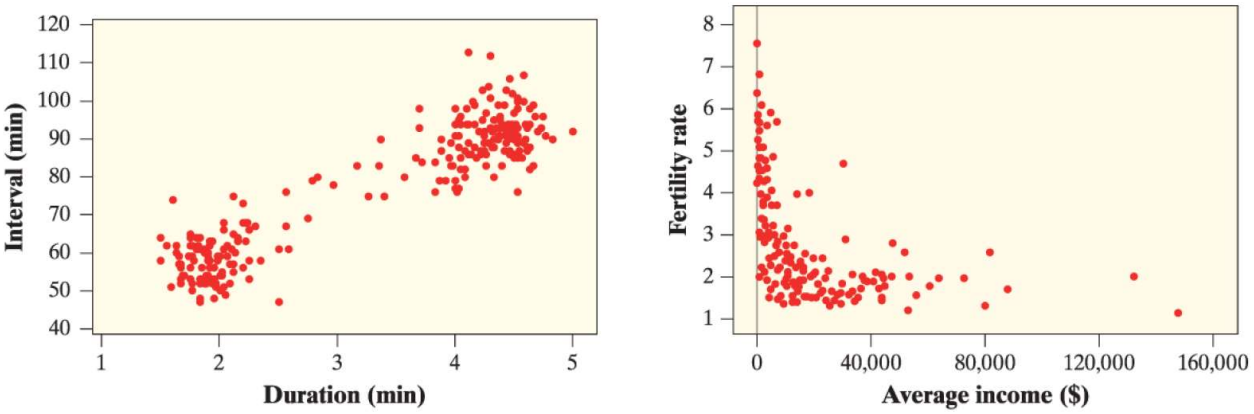


Figure 1: 2 Scatterplots from AP Stats Book

Check Your Understanding: Is there a relationship between the amount of sugar (in grams) and the number of calories in movie-theater candy? Here are the data from a sample of 12 types of candy.

Name	Sugar (g)	Calories
Butterfinger Minis	45	450
Junior Mints	107	570
M&M'S®	62	480
Milk Duds	44	370
Peanut M&M'S®	79	790
Raisinets	60	420
Reese's Pieces	61	580
Skittles	87	450
Sour Patch Kids	92	490
SweeTarts	136	680
Twizzlers	59	460
Whoppers	48	350

1. Identify the explanatory and response variables. Explain your reasoning.

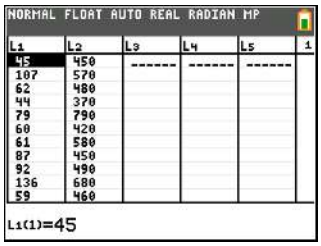
2. Make a scatterplot to display the relationship between amount of sugar and the number of calories in movie theater candy.

3. Describe the relationship shown in the scatterplot.

Using a ti-84 to create scatterplots

Step 1) [stat] 1: edit...

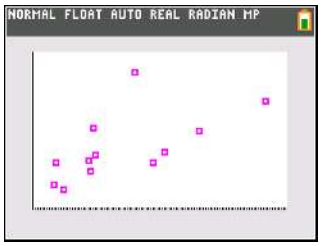
enter data, make sure to maintain the same order



Step 2) [2nd] [y=] 1: Plot 1...

Select the scatterplot and use L1 and L2 for your lists

Step 3) [zoom] [9] to set the window to your data.



4.1.5 Measuring Linear Association: Correlation

Linear relationships are particularly important because a straight line is a simple pattern that is quite common. A linear relationship is considered strong if the points lie close to a straight line and is considered weak if the points are widely scattered about the line. When the association between two quantitative variables is linear, we can use the **correlation r** to help describe the strength and direction of the association.

Definition of Correlation r:

Important properties of correlation r:

The following scatterplots have the correlations -0.99, -0.7, -.03, 0, 0.5, and 0.9. Match the correlations to the pictures.

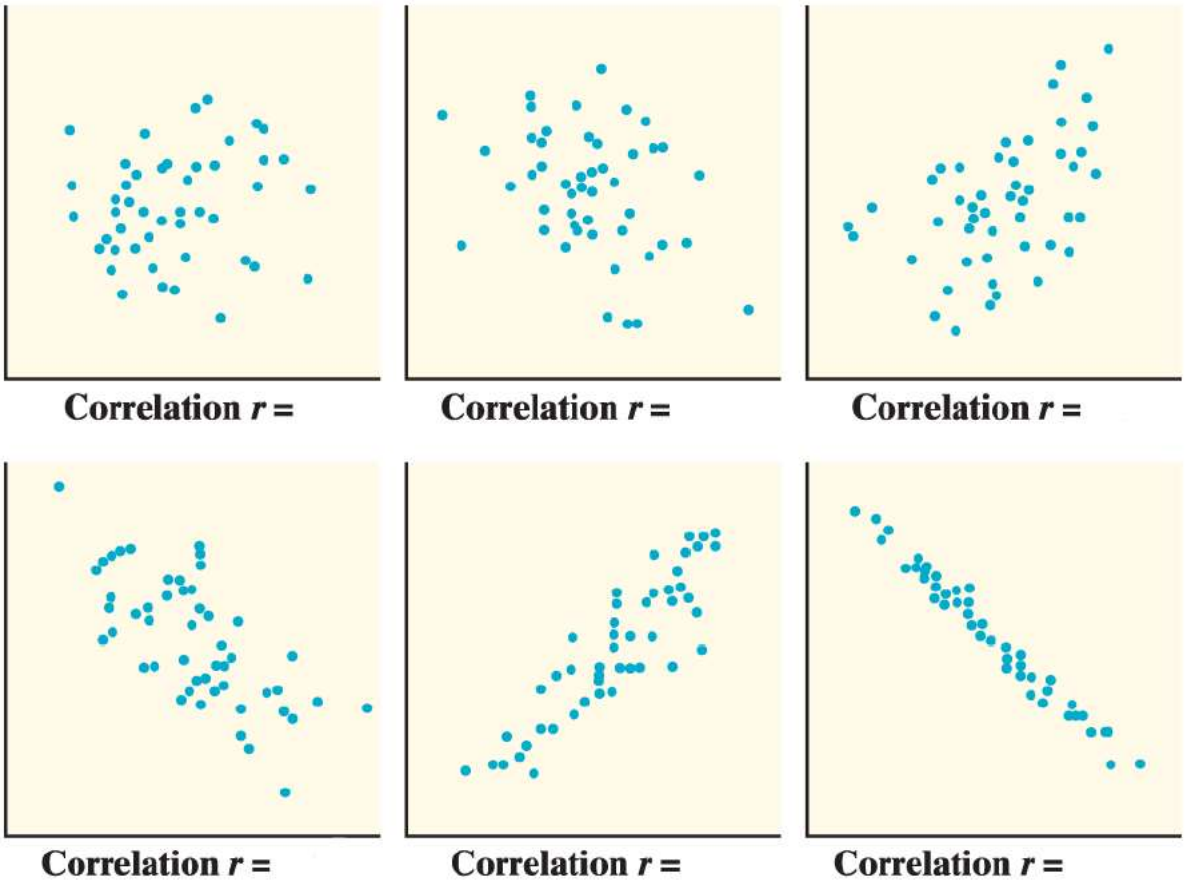


Figure 2: Correlation Guessing

Limitations of correlations

Correlations and Causation: The following scatterplot compares the total revenue generated by skiing facilities and the number of people who died by being tangled in their bedsheets. The correlation is r = 0.97

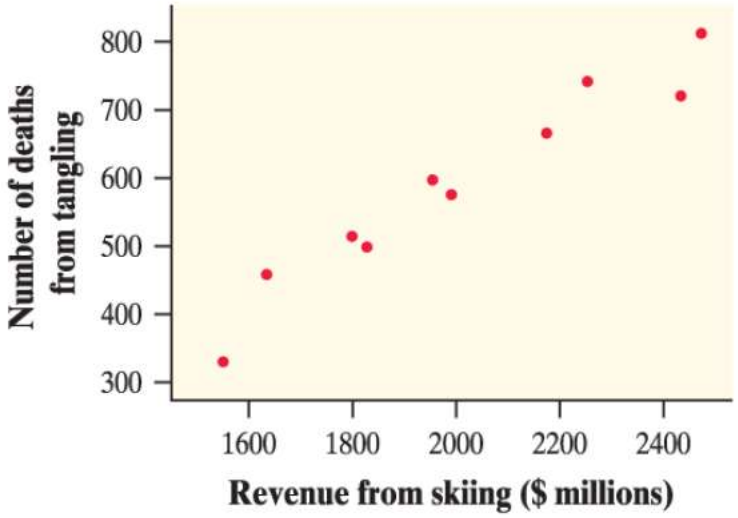
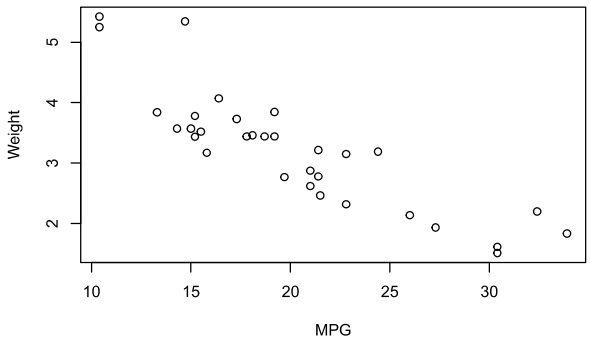


Figure 3: Correlation Guessing

Describe the relationship and correlation of the scatterplot.

Does this mean that skiing causes death by tangling?
Correlation does not measure form.

The scatterplot below compares the MPG and the weight of cars.



The correlation r is -0.87

Despite the strong correlation coefficient, does this scatterplot have a linear shape?

Here is a more extreme example:

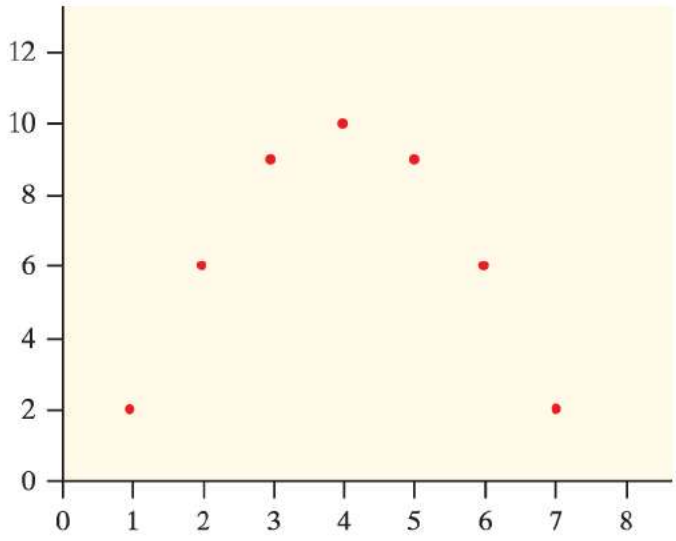


Figure 4: Quadratic Correlation

This graph has a correlation of r = 0.

Correlations should only be used to describe _____

Section 4.1 Summary

- A scatterplot displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.
- If we think that a variable x may help predict, explain, or even cause changes in another variable y, we call x an explanatory variable and y a response variable. Always plot the explanatory variable on the x axis of a scatterplot. Plot the response variable on the y axis.
- When describing a scatterplot, look for an overall pattern (direction, form, strength) and departures from the pattern (unusual features) and always answer in context.
  - Direction: A relationship has a positive association when values of one variable tend to increase as the values of the other variable increase, a negative association when values of one variable tend to decrease as the values of the other variable increase, or no association when knowing the value of one variable doesn't help predict the value of the other variable.
  - Form: The form of a relationship can be linear or nonlinear (curved).
  - Strength: The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
  - Unusual features: Look for individual points that fall outside the pattern and distinct clusters of points.
- For linear relationships, the correlation r measures the strength and direction of the association between two quantitative variables x and y.

4.2 Least Squares Regression

4.2.1 What You Will Learn

- Make predictions using regression lines, keeping in mind the dangers of extrapolation.
- Calculate and interpret a residual.
- Interpret the slope and y intercept of a regression line.
- Determine the equation of a least-squares regression line using technology or computer output.
- Construct and interpret residual plots to assess whether a regression model is appropriate.
- Interpret the standard deviation of the residuals and r2 and use these values to assess how well a least-squares regression line models the relationship between two variables.
- Describe how the least-squares regression line, standard deviation of the residuals, and r2 are influenced by unusual points.
- Find the slope and y intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.

4.2.2 Regression line

No more guessing what line will fit the functions, we can use our calculators to create lines of best fit will fit our data. These lines are called regression lines.

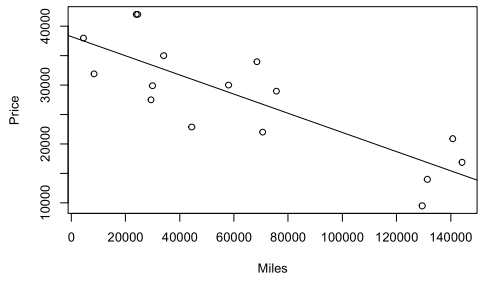
Regression line:

Below is the data for 16 used Ford Supercrew 4x4s create a scatterplot of the data.

Create a scatterplot of the data

Miles	Price
70583	21994
129484	9500
29932	29875
23953	41995
24495	41995
75678	28986
8359	31891
4447	37991
34077	34995
58023	29988
44447	22896
68474	33961
144162	16883
140776	20897
29397	27495
131385	13997

Calculate a linear regression line and a corellation r.



4.2.3 Prediction

regression from previous page:

What does y represent?

What does x represent?

How can we use this regression?

According to the model, how much would you predict a truck with 100,000 miles to sell for?

According to the model, how many miles would you predict a $30,000 truck to have?

According to the model, how much would you predict a truck with 300,000 miles to sell for?

According to the model, how many miles would you expect a $50,000 truck to have?

what are some problems and advantages of regression models?

Why were the 300,000 miles and $50,000 predictions such weird numbers?

4.2.4 Residuals

Regression lines are always predictions! Your regression lines will almost never line up with all the data points. Every prediction will have an error, the difference between the guess and the actual data.

This difference is called a:

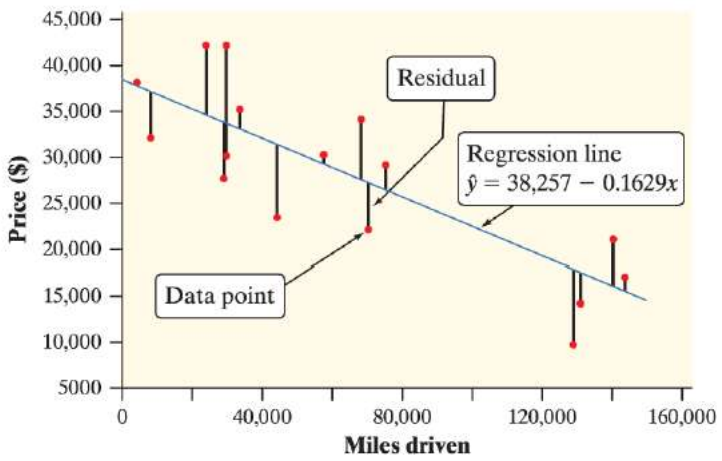


Figure 5: Residual Visual

To find the residual of a certain value, we find the actual y value and the predicted y value for the residual formula

residual = y - $\hat{y}$

Find the residual of the truck that had 70,583 miles driven.

What does the residual mean for this truck?

What reasons could there be for this difference?

Practice

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of y = weight (in grams) and x = time since birth (in weeks) shows a fairly strong, positive linear relationship. The regression equation $\hat{y} = 100 + 40x$ models the data fairly well.

1. Predict the rat's weight at 16 weeks old.
2. Calculate and interpret the residual if the rat weighed 700 grams at 16 weeks old.
3. Should you use this line to predict the rat's weight at 2 years old? Use the equation to make the prediction and discuss your confidence in the result. (There are 454 grams in a pound.)

4.2.5 Interpreting a regression line

Lets go back to the truck regression line.

$\hat{y}$ = 38,257 - 0.1629x

What does 38,257 represent on a graph?

What does 38,257 represent for the truck market?

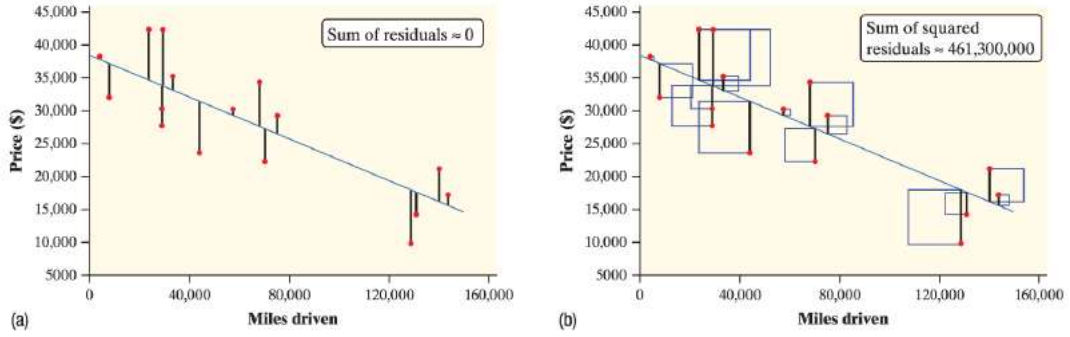
What does -0.1629x represent on a graph?

What does -0.1629x represent for the truck market?

4.2.6 Least-Squares Regression line

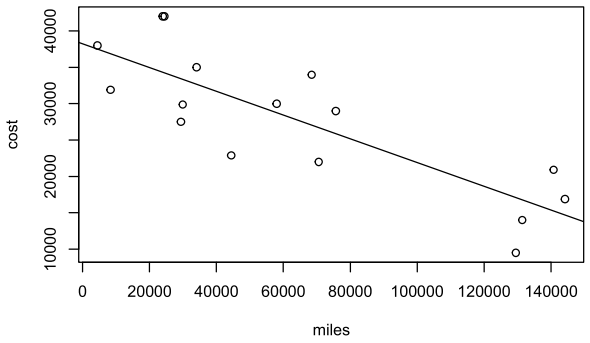
A good regression line makes the residuals as small as possible, why is this important?

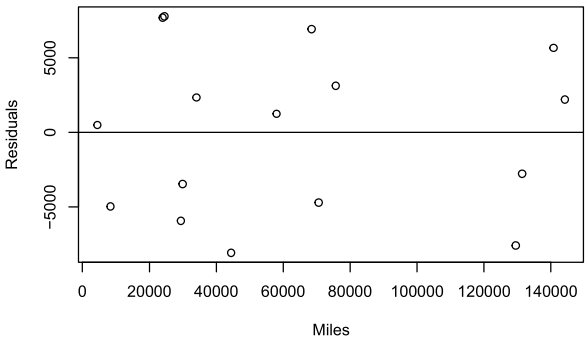
We measure this "as small as possible" goal using the average r^2, hence the name **Least-Squares** Regression line



4.2.7 Is a linear model right for you? introducting Residual Plots!

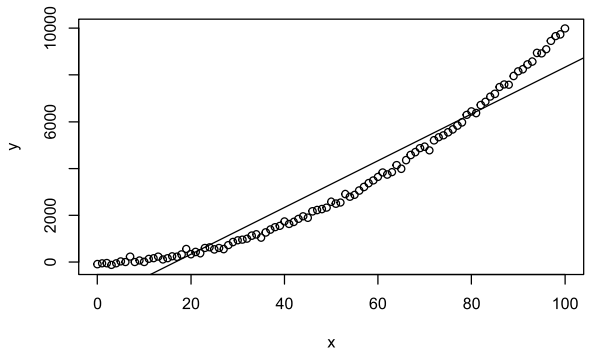
Here is the plotted graph from our truck data and a Residual plot of our truck data. What do you think a residual plot is?

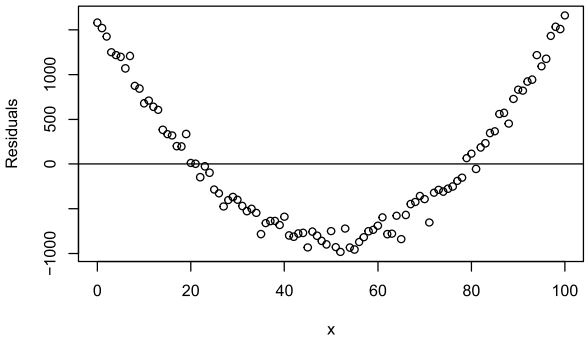




We we're already pretty sure that the linear regression was right for our truck data, because it looks like linear.

lets look at some data that looks non-linear





What is the difference between the 2 residual plots?

What features of a residual plot indicate that linear regression is the correct choice or regression?

To make a regression plot in your TI-84, create a 3rd list of residuals using [2nd] [stat] (list) and selecting the RESID list.

4.2.8 The coefficient of determination: r^2

ever wondered which plot is best for you and your data? Are all the regression options on your calculator overwhelming? Allow me to introduce r^2, your go to measure of regression effectiveness.

r^2 is called the coefficient of determination. The closer r^2 is to 1, the better that regression is.

Here is some data on baby ages (in months) and their weights (in pounds). We want to create a model so we can predict how heavy a baby will be based on how many months old it is.

1	4.3
2	5.1
3	5.7
4	6.3
5	6.8
6	7.1
7	7.2
8	7.2
9	7.2
10	7.2
11	7.5
12	7.8

lets go through the steps of data analysis.

should a linear regression be used for this data? how can we find out?

If we shouldnt use linear, which model should we use and why?

What model has the best r^2?

What is the scope of our model? can it predict a child who is 6 months old? 2 years? 5 pounds? 2 pounds? What is the difference between extrapolation and predictions here?

Any glaring issues or warnings that we could give to people who might use the model we have created? how confident are we in the model?

4.2.9 Calculating the least squares regression line without data

the least squares regression line, $\hat{y} = a + bx$, can be calculated without seeing the data with the standard deviations, means, and r.

$b = r\frac{s_y}{s_x}$, $a = \overline{y} - b\overline{x}$

Example: Does foot size relate to height? The mean and standard deviation of the foot lengths are $\overline{x}$ = 24.76 cm and s_x = 2.71 cm. The mean and standard deviation of the heights are $\overline{y}$ = 171.43 cm and s_y = 10.69 cm. The correlation between foot length and height is r = 0.697.

Calculate the least squares regression line.

4.2.10 Putting it all together

Does the age at which a child begins to talk predict a later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later.

Age	Score
15	95
26	71
10	83
9	91
15	102
20	87
18	93
11	100
8	104
20	94
7	113
9	96
10	83
11	84
11	102
10	100
12	105
42	57
17	121
11	86
10	100

Should we use a linear model to predict a child's Gesell score from his or her age at first word?

If so, how accurate will our predictions be?

4.3 Unit 4 Review

1. Late bloomers? Japanese cherry trees tend to blossom early when spring weather is warm and later when spring weather is cool. Here are some data on the average March temperature (in °C) and the day in April when the first cherry blossom appeared over a 24-year period:

(a) Make a well-labeled scatterplot that's suitable for predicting when the cherry trees will bloom from the temperature. Which variable did you choose as the explanatory variable? Explain.

(b) Use technology to calculate the correlation and the equation of the least-squares regression line. Interpret the correlation, slope, and y intercept of the line in this setting.

Temp (C)	Days
4.0	14
5.4	8
3.2	11
2.6	19
4.2	14
4.7	14
4.9	14
4.0	21
4.9	9
3.8	14
4.0	13
5.1	11
4.3	13
1.5	28
3.7	17
3.8	19
4.5	10
4.1	17
6.1	3
6.2	3
5.1	11
5.0	6
4.6	9
4.0	11

(c) Suppose that the average March temperature this year was 8.2°C. Would you be willing to use the equation in part (b) to predict the date of first bloom? Explain.

(d) Calculate and interpret the residual for the year when the average March temperature was 4.5°C. Show your work.

(e) Construct a residual plot. What does the residual plot tell you?

2. Penguins diving A study of king penguins looked for a relationship between how deep the penguins dive to seek food and how long they stay under water.31 For all but the shallowest dives, there is a linear relationship that is different for different penguins. The study gives a scatterplot for one penguin titled "The Relation of Dive Duration (y) to Depth (x)." Duration y is measured in minutes and depth x is in meters. The report then says, "The regression equation for this bird is: $\hat{y} = 2.69 + 0.0138x$.

(a) What is the slope of the regression line? Interpret this value.

(b) Does the y intercept of the regression line make any sense? If so, interpret it. If not, explain why not.

(c) According to the regression line, how long does a typical dive to a depth of 200 meters last?

(d) Suppose that the researchers reversed the variables, using x = dive duration and y = depth. What effect will this have on the correlation? On the equation of the least-squares regression line?

3. Husbands and wives The mean height of married American women in their early twenties is 64.5 inches and the standard deviation is 2.5 inches. The mean height of married men the same age is 68.5 inches, with standard deviation 2.7 inches. The correlation between the heights of husbands and wives is about r = 0.5.

(a) Find the equation of the least-squares regression line for predicting a husband's height from his wife's height for married couples in their early 20s. Show your work.

(b) Suppose that the height of a randomly selected wife was 1 standard deviation below average. Predict the height of her husband