Regression Analysis

Bivariate Data

Bivariate data is data for which there are two variables for each individual.
Below is an example of bivariate data. It shows the high school and college GPA's for nine college graduates.

Name	*Asia*	*Garrett*	*Ray*	*Evan*	*Alex*	*Natalie*	*Raul*	*MacKenzie*	*Sandra*
High school GPA	3.3	4.0	3.4	3.5	3.8	3.7	3.5	3.6	3.3
College GPA	3.4	3.9	3.5	3.6	3.9	4.0	3.4	3.6	3.7

- ✓ Is there a relationship between a student's college GPA and their high school GPA?
- ✓ Can we predict a student's GPA in college by using their high school GPA?

Our Mission (should you choose to accept it) in this lesson...

To learn methods for describing the relationship between two numerical variables and for assessing the strength of the relationship between them.

This will allow us to be able to answer questions like the ones above.

SCATTERPLOT...displays any relationship between bivariate data in a graphical manner.

Each point on a scatterplot marks a pair of observations taken from one individual in the sample. Typically, the **independent** or **explanatory variable** (*X*) is plotted on the horizontal axis and the **dependent** or **response variable** (*Y*) is plotted on the vertical axis.

Try This...

The table below gives the weights (in hundreds of pounds) and highway fuel usage rates (in miles per gallon) for a sample of new domestic cars.

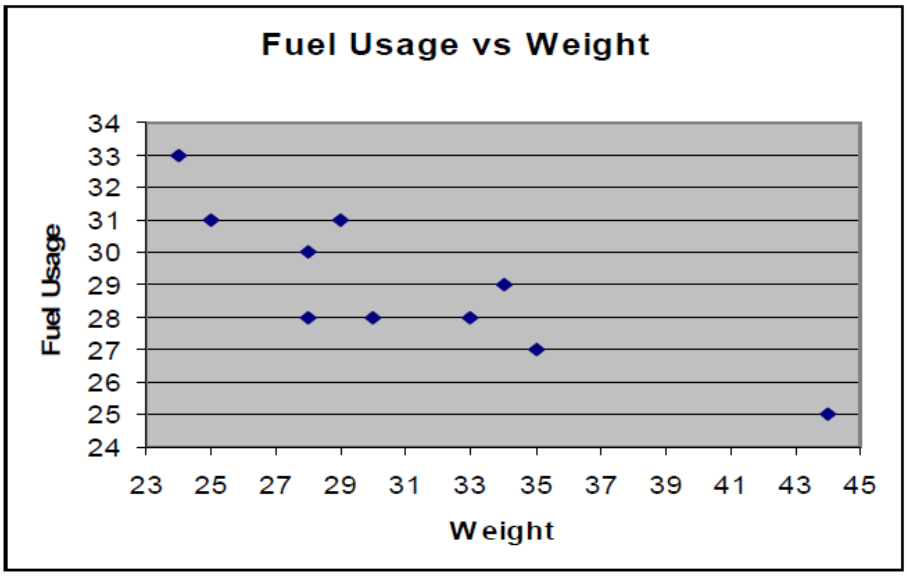
Weight	29	35	28	44	25	34	30	33	28	24
Fuel Usage	31	27	30	25	31	29	28	28	28	33

Which variable do you think is the response, *Y*? _____

Which variable do you think is the explanatory, *X*? _____

(Hint: Y depends on X)

Here's the Scatterplot for the Fuel Usage data:



Based on the data, can you expect to use more gas if you buy a heavier car?

_____

What might the fuel usage be for a car that weighs 4000 pounds? _____

Interpreting scatterplots

After plotting two variables on a scatterplot, we describe the relationship by examining the *Form*, *Direction* and *Strength of association*. We also look for an overall pattern and any deviations from that pattern, called *Outliers*.

Form (pattern formed by the dots): linear, non-linear, or no form.

Direction (of the dots): positive, negative, or no direction.

- ✓ Positive direction → goes from bottom left to top right indicating that the values of *y increase (decrease) as the values of x increase (decrease)* (positive association between the two variables.)

- ✓ Negative direction → goes from top left to bottom right indicating that the values of *y decrease (increase) as the values of x increase (decrease)* (negative association between the two variables.)

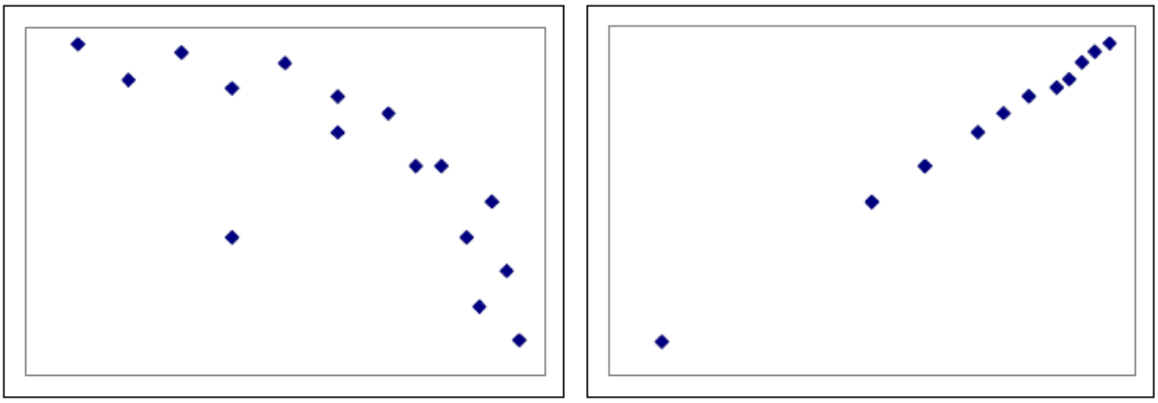
Strength (How the points are scattered): strong or weak association between the variables.

- Strong association — less scatter, clearer pattern
- Weak association — more scatter, less clear pattern

Outliers: Points that fall outside of the pattern of the scatterplot.

Try This...

Identity the Form, Direction, Strength, and Outliers of the scatterplots below:



Plot 1 **Plot 2**

Form: _____ _____

Direction: _____ _____

Strength: _____ _____

Outliers: _____ _____

Nonlinear associations are covered in a more advanced statistics course.

If a scatterplot (or statistical calculation) indicates that there is an association between the two variables, this does not prove that there is a cause-and-effect relationship between them. All it implies is that there is an association between the two variables and nothing more.
There could be a lurking variable affecting the predictor or the response or both resulting in the association between them.

Correlation

A measure of the relationship between two statistical variables measured (bivariate data) from the same population.

Positive Linear Correlation

High (low) values for one variable correspond to high (low) values for the second variable.

Example: Blood Alcohol Level vs. Reaction Time.

Negative Linear Correlation

High (low) values for one variable correspond to low (high) values for the second variable.

Example: Age of a Ford F-150 Truck vs. Retail Value of a Ford F-150 Truck.

No Linear Correlation

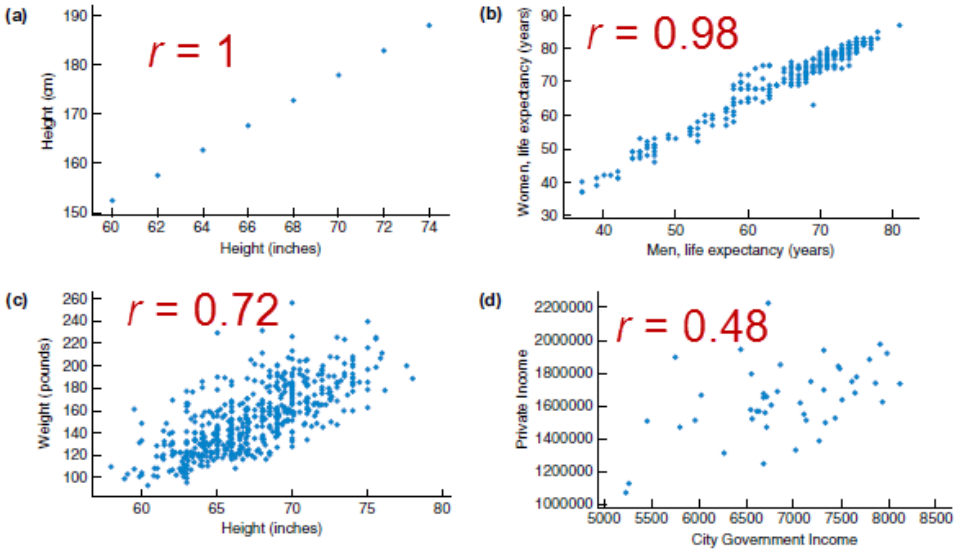
No relationship between the variables or a non-linear relationship.

Example: Height vs. Number of pets owned**.**

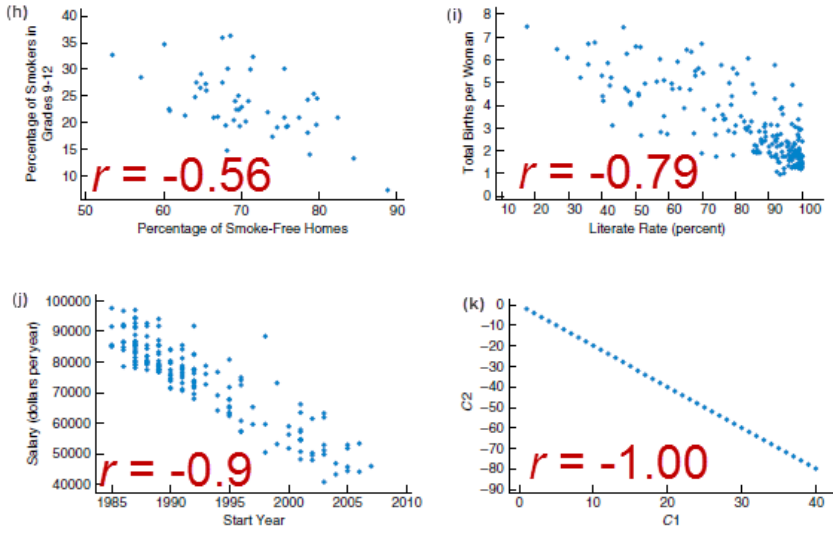
Correlation Coefficient *(r)*... is a measure of the *direction* and *strength* of the linear association between two quantitative variables.

- ✓ r is always a number in the interval $-1 \leq r \leq 1$.
- ✓ A negative value of r indicates that there is a negative linear correlation.
- ✓ A positive value of r indicates that there is a positive linear correlation.
- ✓ A value of r close to 0 indicates almost no linear association (but not necessarily no association at all).
- ✓ r does not have any units.

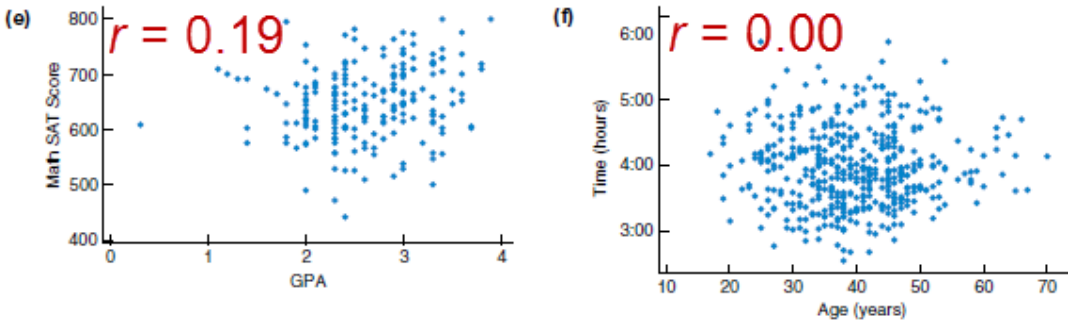
Positive Correlation



Negative Correlation



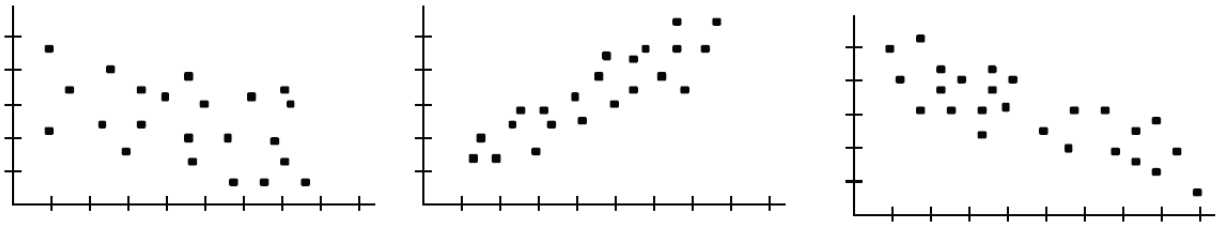
Weak or No correlation



Try This...

Under each scatterplot, write the corresponding correlation coefficient *r*.

Choose from *r* = 0.88, *r* = -0.65, *r* = -0.33.



_____ _____ _____

Some more things we should know about *r* :

1. The linear correlation coefficient *r* measures the strength of only a linear association between two quantitative variables.

2. The value of *r will not change if we shift the data* (add or subtract a constant) or rescale the data (multiply or divide by a constant).

3. The value of *r will not change if* all the values of a variable are converted to different unit (e.g. grams to pounds).

4. The value of *r will not change if* we interchange our choice of *X* and *Y* **—** the strength of the relationship remains the same no matter how we label the variables.

5. *r* is not resistant to outliers.
Outliers can influence the correlation coefficient in different and significant ways. An outlier can inflate an otherwise small correlation or reduce a large one. It can give a negative association a positive correlation and vice versa.

Calculating the Correlation Coefficient from Raw Data

Here's the formula for *r* ... $\boldsymbol{r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}}$

s_x and s_y are the standard deviations of the *x* and *y* values respectively.
n = sample size.

But we can use technology to assist us in calculating the correlation coefficient, *r*.

Try This...

Calculate the correlation coefficient for the data of weight vs fuel usage of domestic cars.

Weight	29	35	28	44	25	34	30	33	28	24
Fuel Usage	31	27	30	25	31	29	28	28	28	33

$\bar{x} = 31$ $s_x = 5.83$ $\bar{y} = 29$ $s_y = 2.28$ $n = 10$

TI-84 Tip: ***Calculating the correlation coefficient r***

- ✓ **Enter the *X* data into List 1 (*L1*) and the *Y* data into List 2 (*L2*).**
- ✓ **STAT > CALC > #8:LinReg(a +b*x*)**

Note: The first time you do this, you might have to set the following to get "*r*" to appear on screen:
From a clear screen: Press **2nd 0** (Catalog), scroll down to **Diagnostic On**, press **Enter**, **Enter**.
When it says done, the above commands should print out "*r*".

Try This (using your calculator)...

"*Text Book Talk*"

Exercise 5.9

"*Text Book Talk*"

Exercise 5.6