Relationships between two quantitative variables

Last unit (chapters 3-6) focused on analyzing a single variable and considering some variation of count vs. category:

Consider our height and step data. We could make a histogram to display it...

In a way, histograms still have x-axis as a category, because they divide the quantitative variable into bins (that are like categories).

Relationships between two quantitative variables

This unit focuses on analyzing two variables which are both quantitative. We will not be dividing into 'bins' and looking at counts. Instead, we will look at how the value of one variable is related to the value of the other variable.

Our main tool to visualize relationships between two variables is a Scatterplot:

Scatterplots give us a visual indication of whether there is an association between the variables, the direction of the association, and a rough idea of the strength. The 'shape' of the cluster of points is called the form.

Associations may be linear or non-linear:

Positive Association (y increases as x increases)

Negative Association (y decreases as x increases)

No Association

Associations may be 'strong' or 'weak':

'stronger'

'weaker'
Roles of the variables

Which variable should be x, and which y? It depends upon how we think about the variables and if we believe variation in one can predict variation in the other.

**x-variable (a.k.a. explanatory, predictor, independent variable)**

The variable we can vary and expect that the other variable will depend upon the value of this variable. The variable associated with ‘cause’ if there is a causal relationship. Often this variable is ‘time’ or ‘year’.

**y-variable (a.k.a. response, dependent variable)**

The variable we expect will be affected by the value of the x-variable. Often, this is the variable we are studying in an analysis (e.g. if studying auto accidents per year, auto accidents would be the y-variable). The variable associated with ‘effect’ if there is a causal relationship.

**Scatterplots using a calculator**

Enter data for x variable in L1, y variable in L2. Let’s use our height vs steps data:

![Scatterplot](image)

**Correlation and Association**

**Association** is a general term meaning there appears to be some relationship between the variables.

**Correlation** is a precise term describing the strength and direction of a linear relationship.

The strength of association (correlation) can be found by computing the **correlation coefficient** of a dataset:

\[
\text{correlation coefficient: } \ r = \frac{\sum z_x z_y}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

Correlation coefficient is a number between -1 (perfect negative correlation) and +1 (perfect positive correlation). Correlation coefficient of 0 means there is no association between the variables.

**How does correlation coefficient work?**

Points in quadrants 1 and 3 tend to indicate positive direction and \( z_x z_y \) will be positive, so will tend to increase the sum.

Points in quadrants 2 and 4 tend to indicate negative direction and \( z_x z_y \) will be negative, so will tend to decrease the sum.

**Correlation Coefficient on Ti84**

Calculator can compute correlation coefficient, but you need to run linear regression (explained more fully later):

1) Mode, Stat Diagnostics: set to ON (only done once)

2) Data entered into L1 and L2.

3) Stat, -> Calc, 8: LinReg(\(a+b\)X)

4) or LinReg(\(a+b\)X) L1,L2,Y1

5) Screen should display r value.
Correlation Coefficient properties

- Sign of $r$ gives direction of association.
- $r$ values from -1 to 1. 0 = no association, closer to 1 or -1 = stronger association.
- $r$ has no units (although some people, incorrectly, report it as a percentage).
- $r$ is unaffected by changes in center or scale of either variable.
- There is no formal agreement about what constitutes 'strong' or 'weak' association. It depends upon context. But here is a general rule of thumb:
  
  $|r| \text{ close to } 1$: 'strong association'
  
  $|r| \approx 0.6$: 'somewhat weak association'
  
  $|r| \text{ close to } 0$: 'no association'

Correlation Conditions

- Both variables must be quantitative.
- Correlation coefficient is very sensitive to outliers. Consider computing with and without outliers included.
- Correlation coefficient assumes a linear relationship. Results are not meaningful for non-linear associations.

(Check for linearity and outliers with a scatterplot).

Linear Regression

Associations which are approximately linear on a scatterplot can be modelled with a line called the Least Squares Regression Line (LSRL). (Also known less accurately as 'linear model', 'line of best fit').

A residual is the error between what the LSRL line predicts the $y$ value will be (for a given $x$) and the actual $y$ value.

$$ e = y - \hat{y} $$

Residuals can be positive or negative. To minimize the overall error, we square each residual and sum all the squared residuals. We then adjust the line so that the total sum of squares is minimized.

Let's look at a specific, small data set:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

(use L3, L4)

The LSRL is the equation of the line which has the smallest sum of squares of the residuals (errors)

https://www.geogebra.org/m/PQfrXv99#material/crBa6TAWhtt
Linear Regression on Ti84

If you have the full data set, you can find the LSRL with a calculator:

1) Mode, Stat Diagnostics: set to ON (only done once)

2) Data entered into L1 and L2.

3) Stat, -> Calc, 8: LinReg(ax+b) or LinReg(ax+b) L1,L2,Y1

(careful, there is an ax+b also)

4) \( \hat{y} = (53.8471) - (0.5728)x \)

5) steps = \( (53.8471) - (0.5728) \text{height} \)

Linear Regression

Equation of the LSRL:

\[ \hat{y} = a + bx \quad \text{(calculator)} \]

\[ \hat{y} = b_0 + b_1x \quad \text{(textbook)} \]

\[ \hat{y} = (53.8471) - (0.5728)x \]

b: slope
For each increase of 1 explanatory variable unit, the LSRL predicts an increase (or decrease) of b units of the response variable.

"For every 1 added inch in height, the number of steps decreases by 0.5728 steps, on average."

a: y-intercept
Most of the time, a doesn't have any particular meaning.

"A person who is zero inches tall is predicted to take 53.8471 steps, on average."

Linear Regression: Using the LSRL

The LSRL is a model that can be used to predict the y value for a given x value.

Interpolation
Using an LSRL to predict when x is inside the range of the data set.

Extrapolation
Using an LSRL to predict when x is outside the range of the data set.

Caution: extrapolation makes the assumption that data in a range we have no evidence about will continue following the same pattern.

What does the LSRL predict will be the number of steps a person 59" tall will take?

\[ \hat{y} = (53.8471) + (-0.5728)x \]

What does the LSRL predict will be the number of steps a person 82" tall will take?
\( r^2 \): Coefficient of determination

The value \( r^2 \) is called the **coefficient of determination** and it is a measure of how 'good' the LSRL is at explaining the variation in \( y \) as \( x \) varies.

You can think of \( r^2 \times 100 \) as the percentage of variation in the **response variable** that is explained by the LSRL which relates **explanatory variable** to **response variable**.

The reason why this is true is explained very well in the math box on p.172-173 of your textbook, but we can also explain intuitively here with our LSRL applet.

One way to explain this is to consider if there were no association between \( x \) and \( y \):

This would mean that the \( y \) values are varying randomly, but not in any way connected with the \( x \) values. It would also mean if there is no association, the LSRL slope=0.

The sum of squares of the areas then represents the total 'error' or variation in \( y \) (16.65).

If we then adjust the LSRL for minimum error...

...we're recognizing that the variation in \( y \) is, at least partly, connected with variation in \( x \). The sum of squares goes down, and now represents just the residual error which is the variation in \( y \) that is not accounted for by the LSRL. (10.49)

That means \( \frac{10.49}{16.65} = 0.63 \) or 63% of the variation in \( y \) is still present after we account for as much as we can with the LSRL. **Which means about 37% of the variation in \( y \) is explained by the LSRL relating \( y \) to \( x \).**

(The calculator shows this is 39.5%).

(Read the math box on p.172-173 of your textbook if you want to know why this is always exactly equal to \( r^2 \))

\( r^2 = 1.0 \) would be a perfect model (every data point is exactly on the LSRL).

\( r^2 = 0.0 \) would be a terrible model.

Real models are between these extremes. Let's check the coefficient of determination for our steps vs. height model.

This means that approximately 76% of the variation in number of steps, from person to person, is associated with a person's height. The remaining 24% of variation in number of steps must be due to other factors.
Getting Equation of LSRL from software output

Equation of the LSRL: \( \hat{y} = a + bx \)  (calculator)

Sometimes, it isn't written as an equation at all:

<table>
<thead>
<tr>
<th>Dependent variable: steps</th>
<th>Coefficient</th>
<th>Intercept</th>
<th>53.847</th>
<th>this row is about the intercept term (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Coefficient</td>
<td>height</td>
<td>-0.5728</td>
<td>this row is about the 'height' variable term (b)</td>
</tr>
</tbody>
</table>

This software output means: \( \text{steps} = 53.846 - 0.5728 \times \text{height} \)

Relationship between slope b, r, and Scatter plot shape:

Slope b is related to the slope of the LSRL:

| r value doesn't change with slope, but it does match the sign of the slope |

Correlation r is related to how tightly grouped the points are around the LSRL:

Correlation coefficient only measures the strength of linear association and it becomes invalid when slope approaches 0 or infinity

\[
 r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum_{i=1}^{n} z_x z_y
\]

Since the correlation coefficient is a sum of the products of z scores, any point that has a z score near 0 in either x or y doesn't contribute to r.

z, is nearly 0 for all these points, so r is nearly 0 even though there is a strong linear association.

r is nearly zero because even though there is a strong association, it isn't linear.

Standardizing a Scatterplot

The general shape of a scatterplot remains the same even if the units are changed, but may look stretched depending upon the scale size. But a consistent scatterplot shape can be produced by using the z-score standardized values for each variable, so that each axis plots number of standard deviations a value is away from that variable's mean:

\[
 (z_x, z_y) = \left( \frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y} \right)
\]

origin (0,0) corresponds to (\( \bar{x}, \bar{y} \)).
Linear Regression: How LSRL, b and r are related
(If you don't have the full data set, you can still calculate various useful information.)

A very useful equation: \( b = r \frac{s_y}{s_x} \)

The math box on p.172-173 derives this - please read this for more detail.
But the following examples provides some intuitive sense about this equation and ways we can use it.

1) Given summary statistics (but no data) find the slope and LSRL

<table>
<thead>
<tr>
<th>height</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.8333 in</td>
<td>16.1389</td>
</tr>
<tr>
<td>4.9598 in</td>
<td>16.1389</td>
</tr>
</tbody>
</table>

\( r = -0.8734 \)

a) Use formula to calculate \( b \).
b) Find y-intercept \( a \) - the centroid \((\overline{x}, \overline{y})\) will always lie on the LSRL.
c) Write out the LSRL.

2) Given data value std dev in \( x \), find corresponding std dev in \( y \)

If a student's height is 1.2 standard deviations above the mean height of the class, how many standard deviations above the mean number of steps would you predict this student's number of steps to be?

(When considering standard deviations, these are z-scores, so \( s_y = s_x = 1 \))

3) Find LSRL if \( x \) and \( y \) are swapped

\( b = r \frac{s_y}{s_x} \)

Does \( r \) change? \( \text{no} \) (\( r \) measures strength of association, swap does not affect \( r \))

Does \( b \) change? \( \text{yes} \)  
Before swap: \( b = r \frac{s_y}{s_x} \)  
After swap: \( b = r \frac{s_x}{s_y} \)
The effect of outliers on linear regression  
https://www.geogebra.org/m/NZpWpCW8

Any outlier deserves attention due to the unduly large effect it may have on results. Outliers may (or may not) affect correlation and/or slope.

Outlier effect on slope - the concept of 'leverage'
The point $(\bar{x}, y)$ is always on the LSRL and you can think of it like a 'fulcrum' of a lever. A data point whose x value is the same as the fulcrum will have no effect on the LSRL, but a line whose x value is far away from the fulcrum has a large effect and we say this is a high leverage point.

But for the point to cause a change in slope of the LSRL, it needs to have a high residual. A point already near the LSRL won't 'push' the LSRL and cause much change. If a point has high leverage and high residual (so that removing it or adding it causes a large change in the LSRL slope) we say that point is influential.

Outlier effect on correlation
• Correlation is the measure of the strength of the association and is high (close to +1 or -1) if the points are grouped tightly around the LSRL.
• Correlation is calculated as sum of products of standardized distances x and y from the mean, so a point has a large effect on correlation if it is far from the 'fulcrum' in both x and y.

Residual Plots
A residual plot for a given linear regression shows the residual (r) vs. x.
Examples: Sketch residual plots by hand for each data set

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>51</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>6</td>
<td>31</td>
</tr>
</tbody>
</table>

Scatter plots
Residual plots

If the residuals are randomly scattered around '0' then you know that a linear model is appropriate. (Residuals make it easier to see non-linearity compared to scatterplots.)
To do residual plot on Ti84, after running a LinReg as normal:
- 2nd, Y= (Stat Plots)
- Re-enter the stat plot you are using, or enable Stat Plot 1
- Change the Ylist for RESID (2nd LIST)
- Zoom, 9:ZoomStat

(Or you can use Stat Plot 1 for the scatterplot and Stat Plot 2 for the residuals)

A residual plot for a given linear regression shows the residual (r) vs. x.

- Slope of an LSRL through residuals is always zero.
- Mean of the residuals is always zero.
- Standard deviation of the residuals is a measure of how points are spread around the LSRL.

Association (correlation) does not imply Causation

So, if we want people to live longer, we should just give them more TVs, right?

A strong association between two variables does not mean that one variable is causing a change in the other variable.

What might be going on here to explain this?

wealth \[\rightarrow\] life expectancy

number of TVs \[\leftarrow\] life expectancy

Both life expectancy and number of TVs might be changing according to the value of a 3rd ‘lurking variable’ such as wealth.

(be wary of lurking variables)