Math 1000
Mathematical Literacy in Today’s World

Lecture 10
Winter 2010
When individuals have two data values associated

- Are the values related?
- Does one value depend on another?
- How strong is the dependence? (I.e., how strong is the relationship?)
- Can we give some indication of what the relationship is? (numerically)
Names for Values/Variables

- If changes in one variable is supposed to explain or cause changes in a related variable, the first one is called the **explanatory** variable.

- The second one is the **response** variable, because we expect it to respond to changes in the first.
The scatterplot may show “positive association” (slope)
Or negative association (slope)
Regression Lines

- Regression is a computation which computes the “best-fitting” equation from the data points.
- Your two-variable calculator should do this after you key in the data values.
- The calculator should give you slope and y-intercept.
But Remember:

- We say \( y = mx + b \)
- The book says \( y = a + bx \) (so \( b \) is slope and \( a \) is y-intercept) – also TI-83
- The TI-30X(II) calculator says \( y = ax + b \) (so \( a \) is slope and \( b \) is y-intercept)
Regression depends on which variable is explanatory

- Regression computes by minimizing errors in predictions of $y$-values from $x$-values.
- So switching which variable is explanatory switches to predicting $x$-values from $y$-values and minimizing errors in those predictions.
About the Correlation $r$

- Measures strength of relationship between variables
- Ignores difference between explanatory and response variables
- Has values between $-1$ and $+1$
- Has no units (not changed by scaling)
- Strongly affected by outliers
Correlation $r = +0.95$
Correlation $r = -0.82$
Correlation $r = -0.10$
Correlation $r = +0.38$
When Correlation is close to +1 or -1

- Data points are close to the regression line
- Predictions are more likely to be useful
When Correlation is close to 0

- Data points are often far from the regression line
- Predictions are less likely to be useful
Least-squares regression picks out smallest sum of squares

- Sum of squares can only be zero if points are all on an actual line
- The least-squares regression line always goes through the point (xbar, ybar) of averages of the x-values and of the y-values
Some formulas we remember

- Number of values = n
- X-values are x₁, x₂, …, xn
- For x-values, mean is
  \[ \bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} \]
- Standard deviation of x-values is
  \[ S_x = \sqrt{\frac{(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}} \]
Some formulas we remember

- Number of values = n
- y-values are y₁, y₂, …, yn
- For y-values, mean is \( \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} \)
- Standard deviation of y-values is \( S_y = \sqrt{\frac{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1}} \)
Some formulas (you don’t need to remember)

- Slope $m$ is sum of $(x_i-x\bar{)}(y_i-y\bar{)}$ divided by sum of $(x_i-x\bar{)}^2$
- Sum of $(x_i-x\bar{)}^2$ is $(n-1)(S_x)^2$
- Correlation coefficient $r = (m*S_x)/S_y$
- Intercept $b = y\bar{)} - (m*x\bar{)}$
Median Income by Family Size

- For Size 2, Median Income is $50,441
- For Size 3, Median Income is $60,085
- For Size 4, Median Income is $72,591
- For Size 5, Median Income is $70,028
- For Size 6, Median Income is $60,035
Which is explanatory?

- Possibly family size, since having a larger family might cause wage-earners to try to earn more money.
- Then the family would respond to increases in family size by (attempting to) increase family income.
Scatterplot

Number in Family

Median Income
Regression gives $y = 2913x + 50984$
Using $y = 2913x + 50984$

- At $x = 0$, $y = 50984$, so $(0, 50984)$
- At $x = 6$, $y = 2913(6) + 50984 = 17478 + 50984 = 68462$, so $(6, 68462)$
- Draw line through $(0, 50984)$ and $(6, 68462)$
Why are these numbers so big?

- In $y = 2913x + 50984$, intercept 50984 is the predicted income for a family with 0 members (so still $50,984, \text{ a big number}$)
- The slope 2913 is the amount that the family income increases for each extra family member (so $2,913$ for each)
Plotting (0, 50984) and (6, 68462)
Predicting for Size 7

- Using $y = 2913x + 50984$, we could predict that when $x = 7$, $y = 2913(7) + 50984 = 71375$
- So a family of 7 people is predicted to have a family income of $71,375$
But not that good a prediction
Strength of the linear relationship is not that good.
In fact Correlation $r = 0.52$
So only a moderate relationship
Which is explanatory?

- Possibly family income, since having a larger family income might cause families to support more people (or have more children)

- Then the family would respond to increases in family income by (attempting to) increase family size
Scatterplot the other way
Regression gives  \( y = (9.23 \times 10^{-05})x - 1.78 \)

- \(9.23 \times 10^{-05}\) means 9.23 times 10 to the \(-5\) power, so 0.0000923 is the slope.
- Intercept -1.78 means that the regression line starts 1.78 below the positive values.
Why is the slope 0.0000923 so small?

- Here the x-values are incomes, which are large numbers
- The y-values are family sizes, small numbers
- Say, \( x = 50000 \), so \( y = 0.0000923(50000) - 1.78 = 4.615 - 1.78 = 2.835 \)
- So predicted point (50000, 2.835)
One more point needed

- Say for \( x = 80000, \ y = 0.0000923(80000) - 1.78 = 7.384 - 1.78 = 5.604 \)
- So points (50000, 2.835) and (80000, 5.604)
Plotting (50000, 2.835), (80000, 5.604)
Draw Regression Line
Extended Line has intercept = $-1.78$
Note prediction for 7 Family members is off-scale
Using $y = 0.0000923x - 1.78$

- If you want 7 family members, in this case that means that $y = 7$
- So we can solve $7 = 0.0000923x - 1.78$
- Next $7 + 1.78 = 8.78 = 0.0000923x$
- Finally $x = 8.78/0.0000923 = $95,125
- Way off from $71,375 which was done the other way
Infant Mortality Rates throughout the World
Correlation $r = 0.993$
Regression line: $y = 0.87x - 1.33$
Why slope = 0.87?

- Suggests mortality for males is higher by a factor of 0.87 than that for females
- And that this is fairly consistent across different countries
- So males have approximately a 13% higher infant mortality rate
Applet of Textbook

- At [www.whfreeman.com/fapp8e](http://www.whfreeman.com/fapp8e)
- Allows you to make scatterplots and look at possible regression lines
- Can see how outliers weaken relationships
Reasons for more variation “at ends” (= far from (xbar, ybar))

- No data points there to confirm that the mildly positive association continues in this area of the plot
- Small error in computing m would result in larger error in computing predicted y
When does linear regression give a bad answer?

- When points do not really appear to have a linear relationship
- (But the method still works, just gives a bad answer)
Not a linear relationship
Not a linear relationship
Association does not imply causation

- There is a high correlation among nations of the world between \((x = \text{number of television sets per person})\) and \((y = \text{average life expectancy})\)
  - That is, nations with more televisions per person have a longer average life expectancy

- But the an increase in the number of televisions does not cause an increase in life expectancy; Instead both are more likely related to a third variable, that of money
Another example

- Suppose there is a strong negative association between number of sick days and quality of classroom performance
- Maybe students performing better are less likely to miss school when sick, or
- Maybe students performing better are less likely to get sick in the first place, or
- Maybe some students happened to get sick and performed worse because of missing class.
Interpretation is needed

- Allow for reasons which may explain an association or correlation without necessarily assuming a causal relationship
So when looking at data relationships

- Always draw a scatterplot, so as to see
  - If relationship **not** linear (in which case even well-correlated data may give bad predictions)
  - If there are outliers which will control the regression line more than they deserve
  - If data occur in bunches (which makes prediction outside the bunches less accurate)
When looking at data relationships

- Remember to consider which variable is explanatory, and whether changes in the other variable is responding to changes in the explanatory one
- Also remember there may be a hidden variable to which both are responding
The End for Today

Good luck on the Quiz!