Correlation and Regression

Correlation and Regression

The test you choose depends on level of measurement:

Independent	Dependent	Statistical Test
Dichotomous	Interval-ratio Dichotomous	Independent Samples t-test
Nominal Dichotomous	Nominal Dichotomous	Cross Tabs
Nominal Dichotomous	Interval-ratio Dichotomous	ANOVA
Interval-ratio **Dichotomous**	**Interval-ratio**	**Correlation and OLS Regression**

Correlation and Regression

- Correlation is a statistic that assesses the strength and direction of linear association of two interval-ratio variables . . . It is created through a technique called “regression”
- Bivariate regression is a technique that fits a straight line as close as possible between all the coordinates of two interval-ratio variables plotted on a two-dimensional graph--to summarize the relationship between the variables

Correlation and Regression

- For example:

  A sociologist may be interested in the relationship between education and self-esteem or Income and Number of Children in a family.

Independent Variables		Dependent Variables
Education	$\longrightarrow$	Self-Esteem
Family Income	$\longrightarrow$	Number of Children

Correlation and Regression

- For example:
  - May expect: As education increases, self-esteem increases (positive relationship).
  - May expect: As family income increases, the number of children in families declines (negative relationship).

<u>Independent Variables</u>		<u>Dependent Variables</u>
Education	$\xrightarrow{+}$	Self-Esteem
Family Income	$\xrightarrow{-}$	Number of Children

Correlation and Regression

- For example:
  - Null Hypothesis: There is no relationship between education and self-esteem.
  - Null Hypothesis: There is no relationship between family income and the number of children in families.
  - H_o: $b = 0$
  - $H_{a:}$ $b \neq 0$

"b" is a symbol for a statistic that describes the relationship

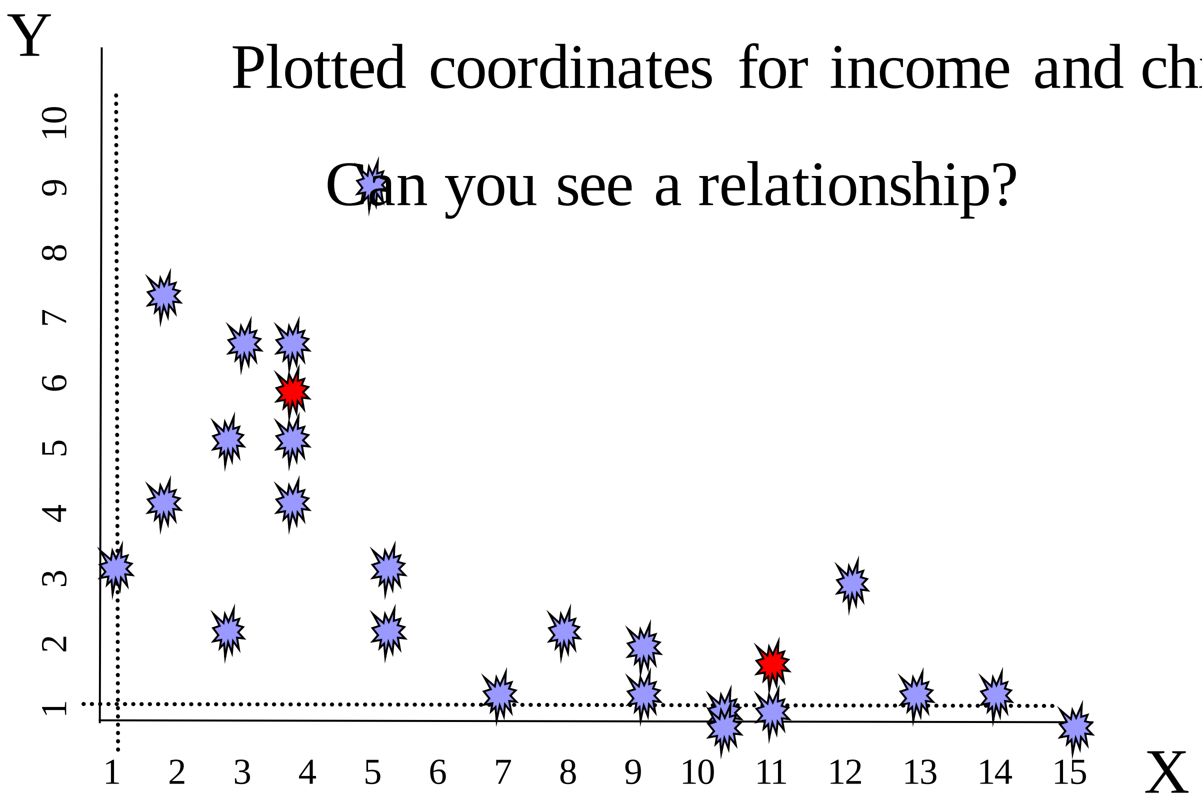
<u>Independent Variables</u>		<u>Dependent Variables</u>
Education	→	Self-Esteem
Family Income	→	Number of Children

Correlation and Regression

- Let's look at the relationship between income and number of children.
- Regression will start with plotting the coordinates in your data (although you will hardly ever "plot" your data in reality).
- Some data:

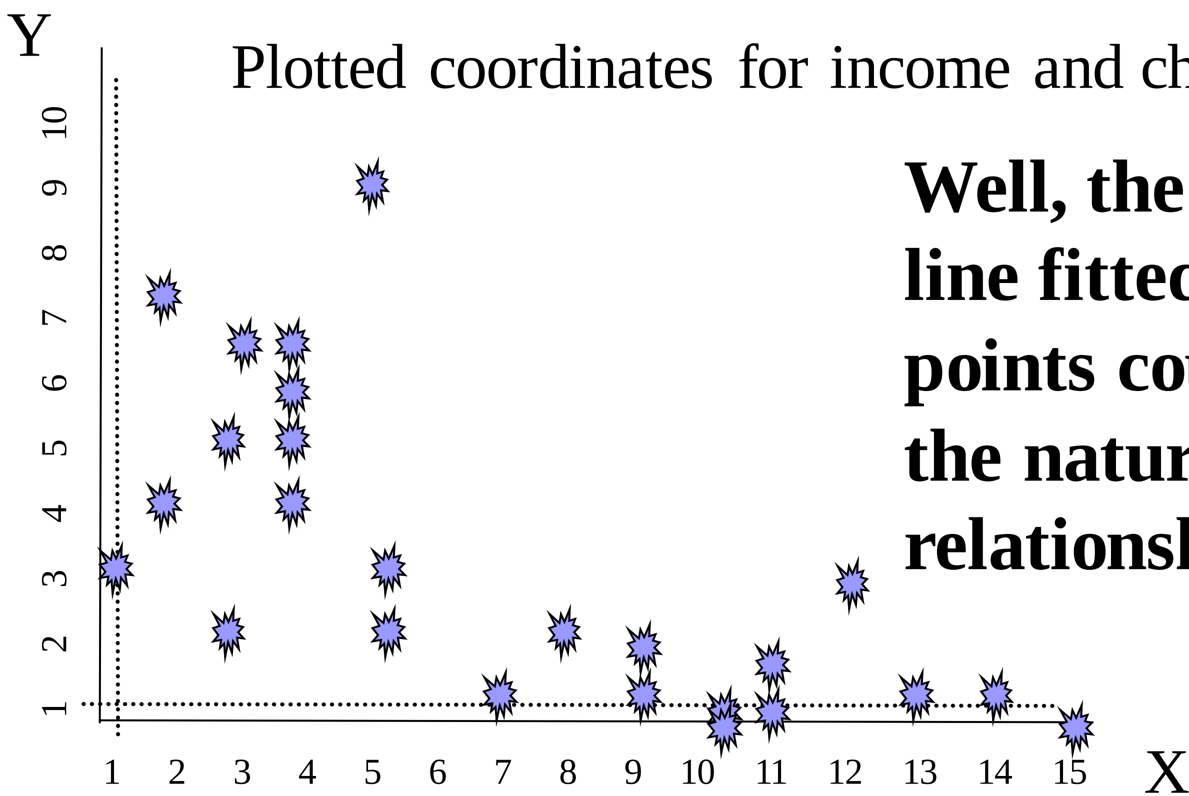
Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression



Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

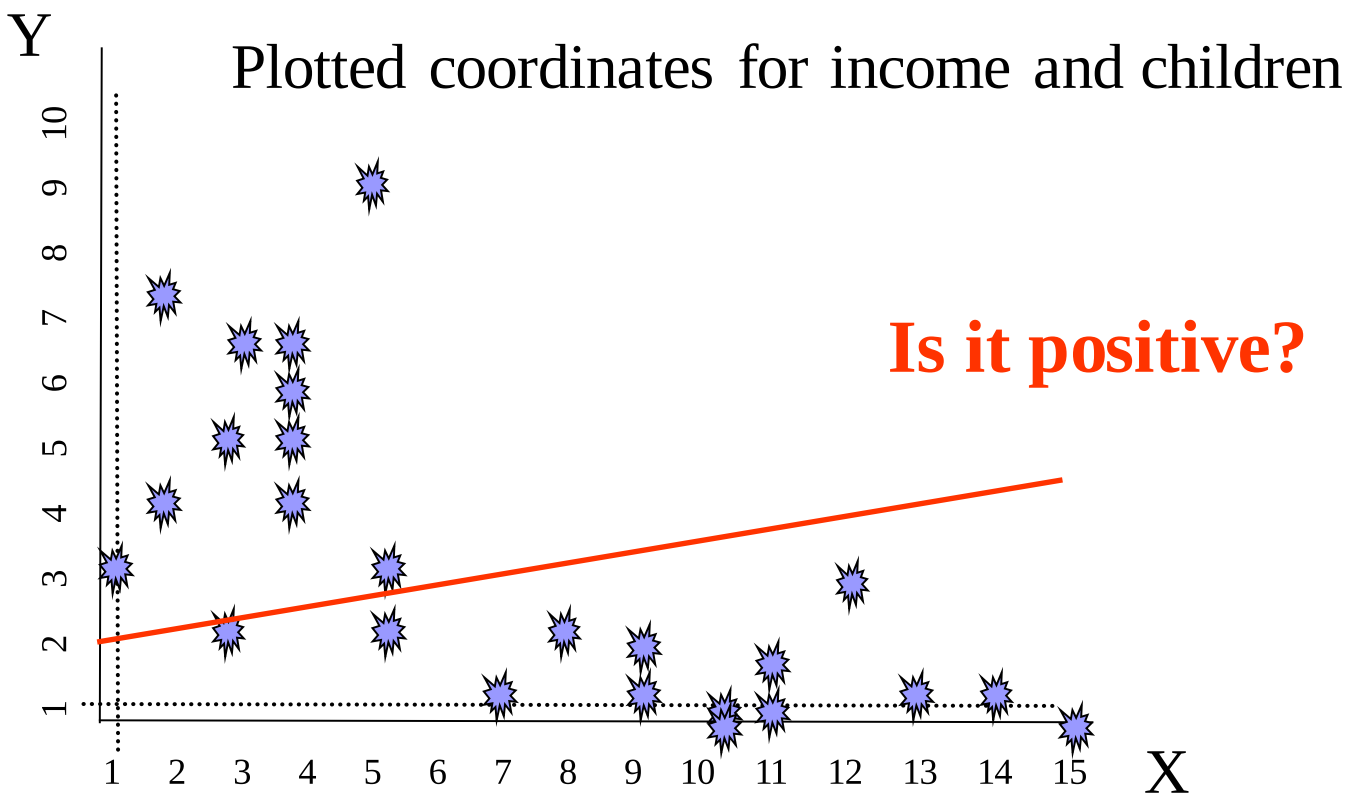
Correlation and Regression



Well, the slope of a line fitted to the points could tell us the nature of the relationship!

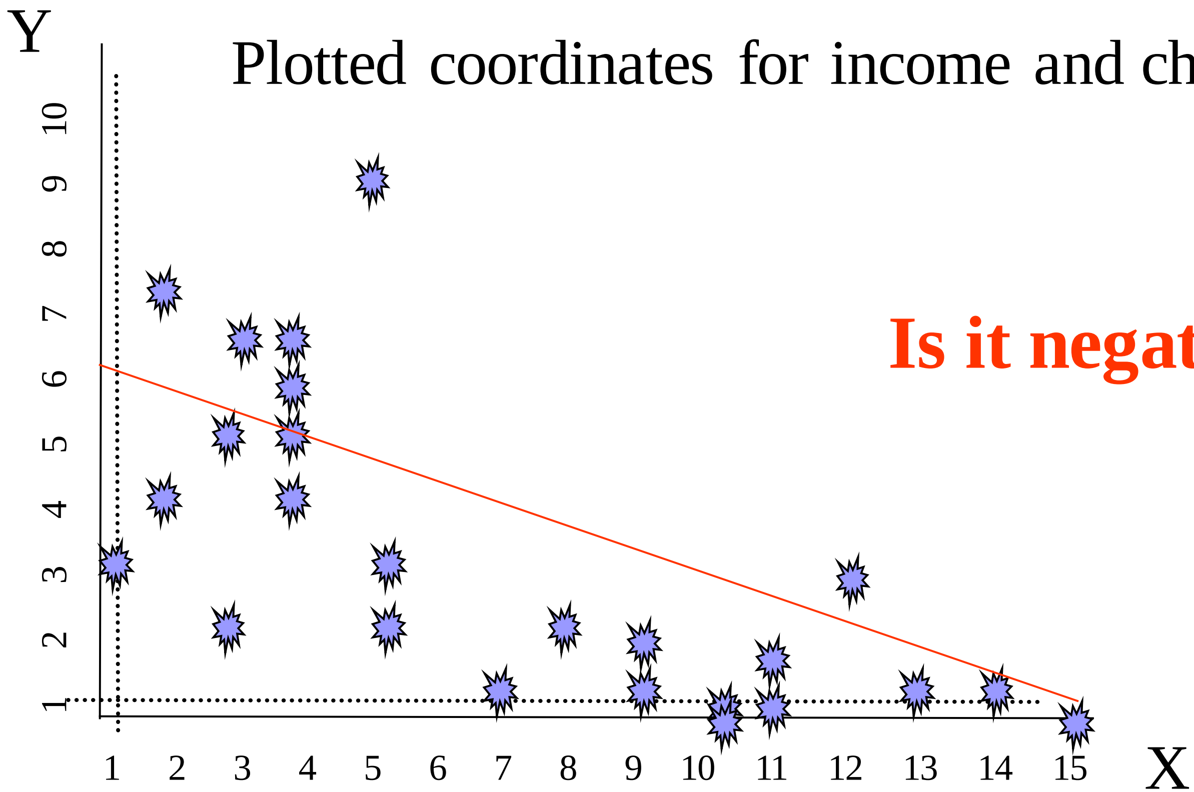
Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression



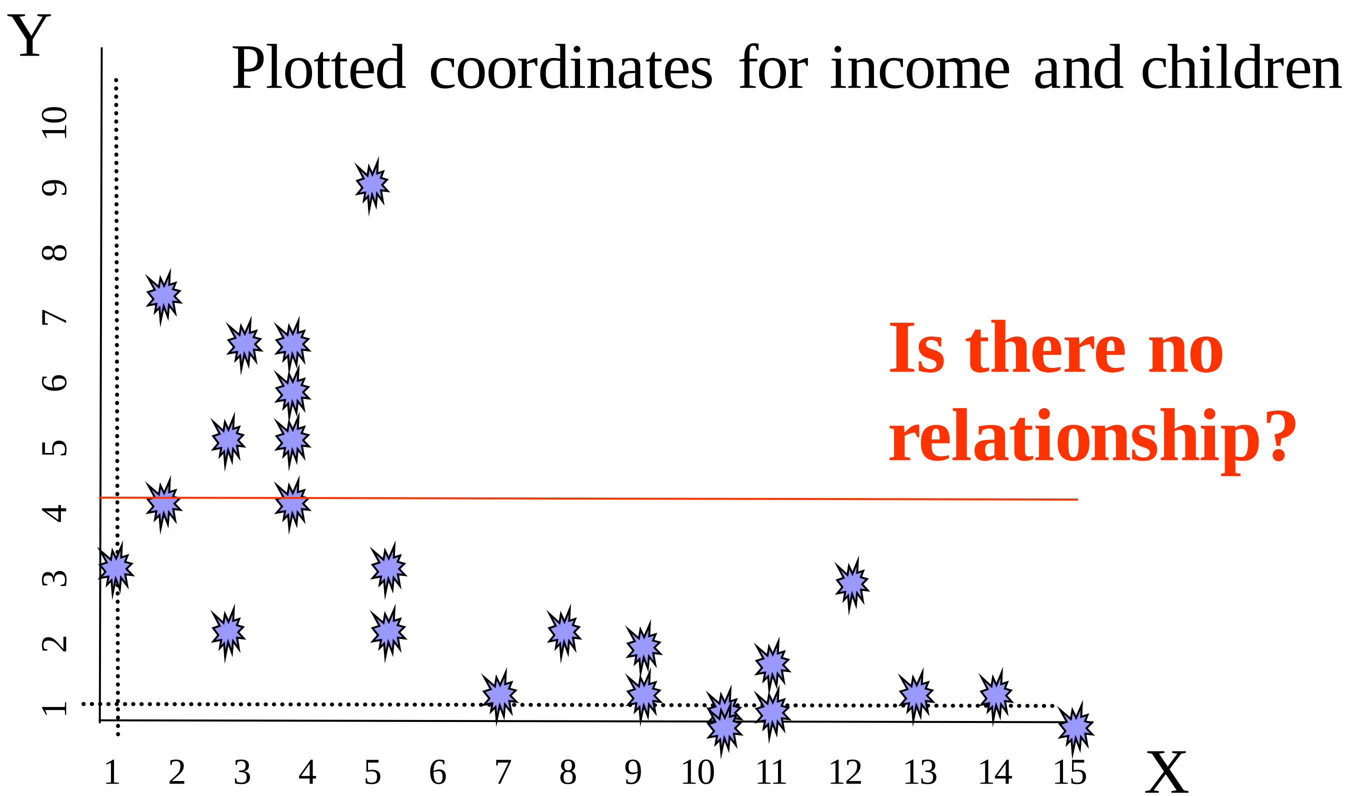
Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression



Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression



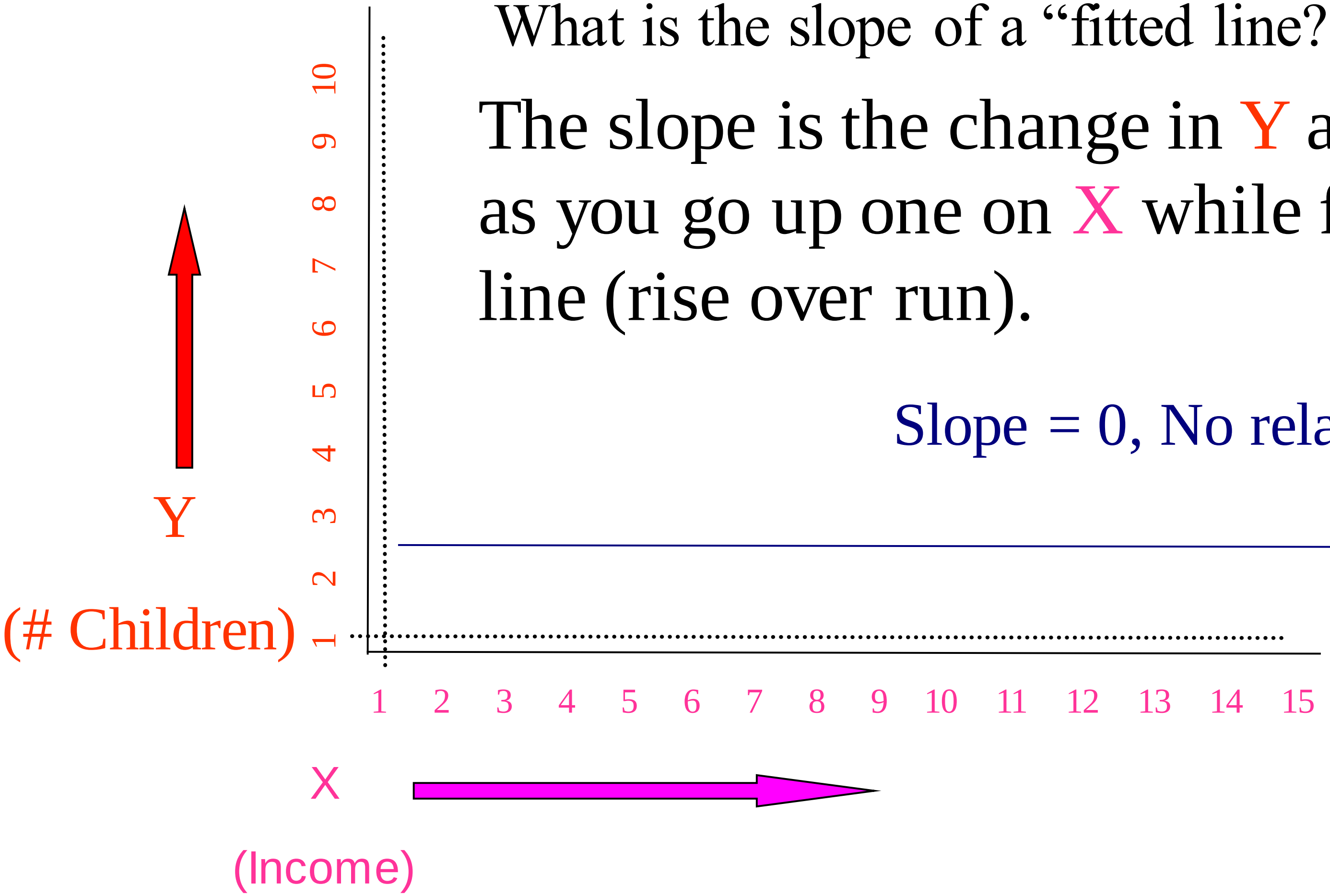
Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

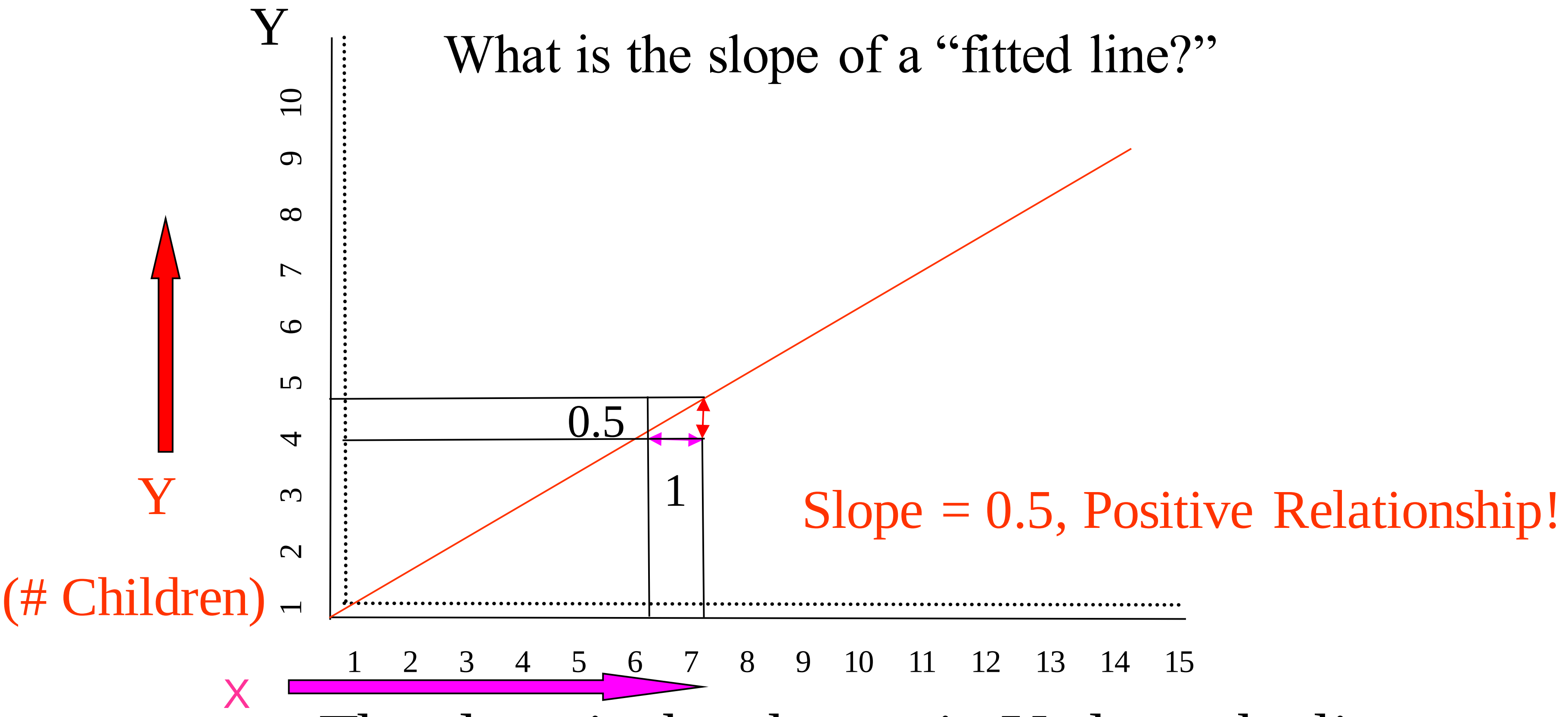
What is the slope of a “fitted line?”

The slope is the change in Y along the line as you go up one on X while following the line (rise over run).

Slope = 0, No relationship!

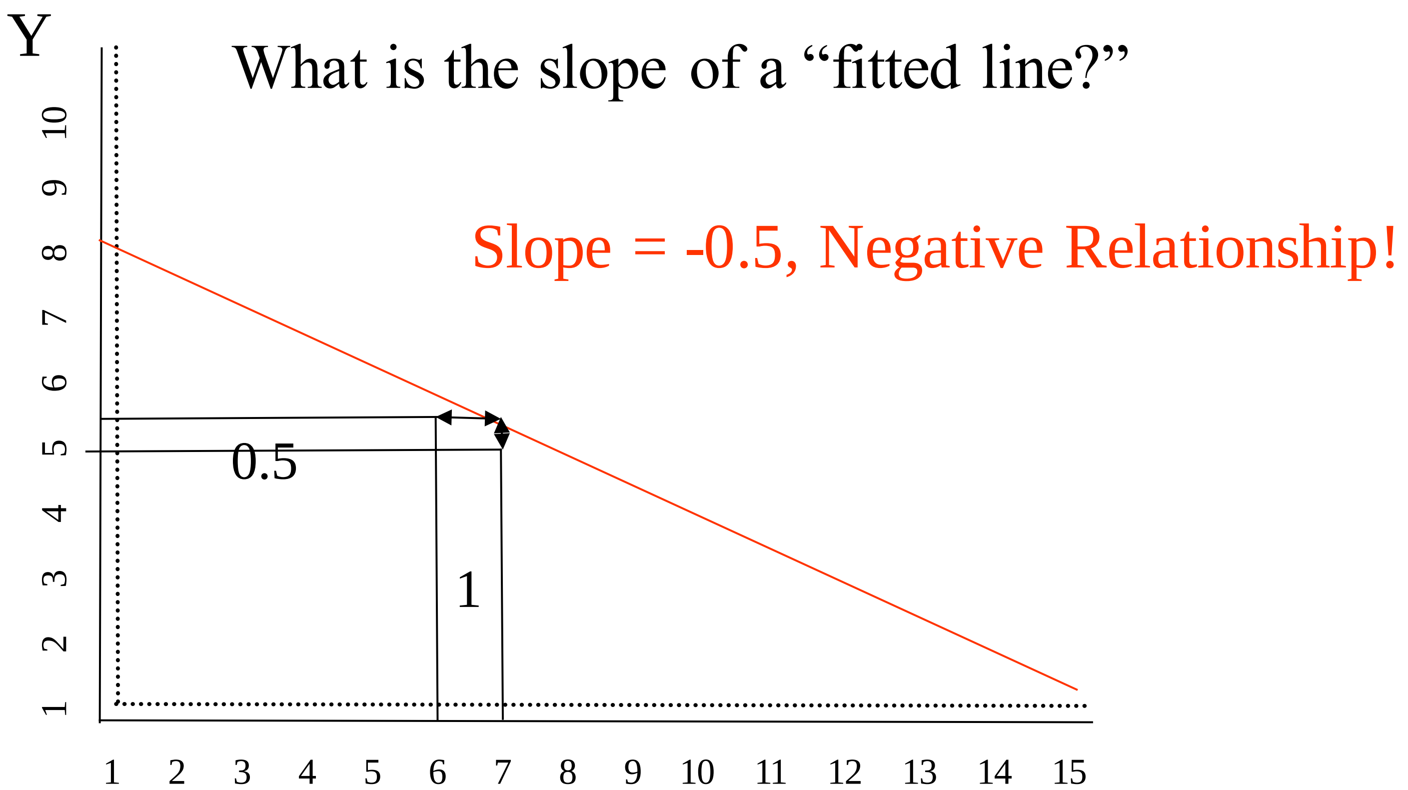


Correlation and Regression



The slope is the change in Y along the line as you go up one on X while following the line (rise over run).

Correlation and Regression



The slope is the change in Y along the line as you go up one on X while following the line (rise over run).

Correlation and Regression

- The mathematical equation for a line:

  $Y = mx + b$

Where: Y = the line's position on the vertical axis at any point

X = the line's position on the horizontal axis at any point

m = the slope of the line

b = the intercept with the Y axis, where X equals zero

Correlation and Regression

- The statistics equation for a line:

  $\hat{Y} = a + bx$

Where: $\hat{Y}$ = the line's position on the vertical axis at any point (estimated value of dependent variable)

X = the line's position on the horizontal axis at any point (value of the independent variable for which you want an estimate of Y)

b = the slope of the line (called the coefficient)

a = the intercept with the Y axis, where X equals zero

Correlation and Regression

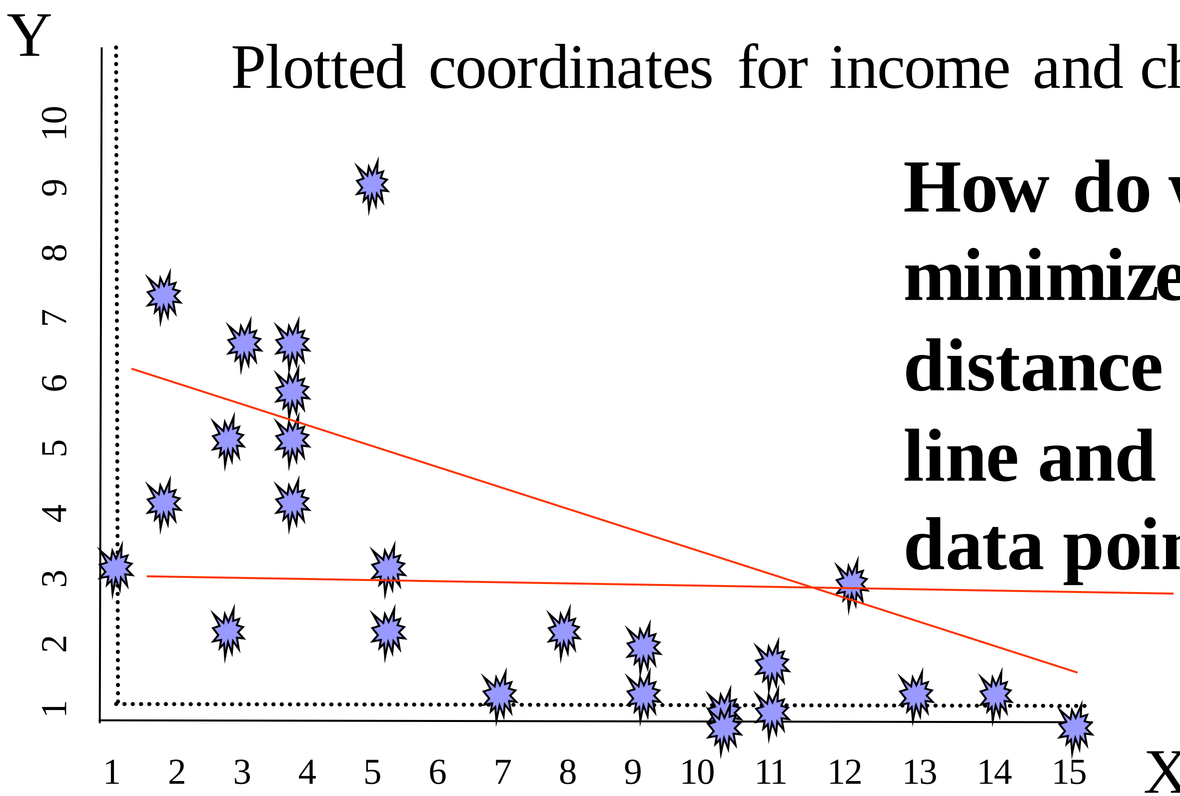
- The next question:

  How do we draw the line???

- Our goal for the line:

  Fit the line as close as possible to all the data points for all values of X.

Correlation and Regression



How do we minimize the distance between a line and all the data points?

Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

How do we minimize the distance between a line and all the data points?

You already know of a statistic that minimizes the distance between itself and all data values for a variable--the mean!

The mean minimizes the sum of squared deviations--it is where deviations sum to zero and where the squared deviations are at their lowest value. $\Sigma(Y - \text{Y-bar})^2$

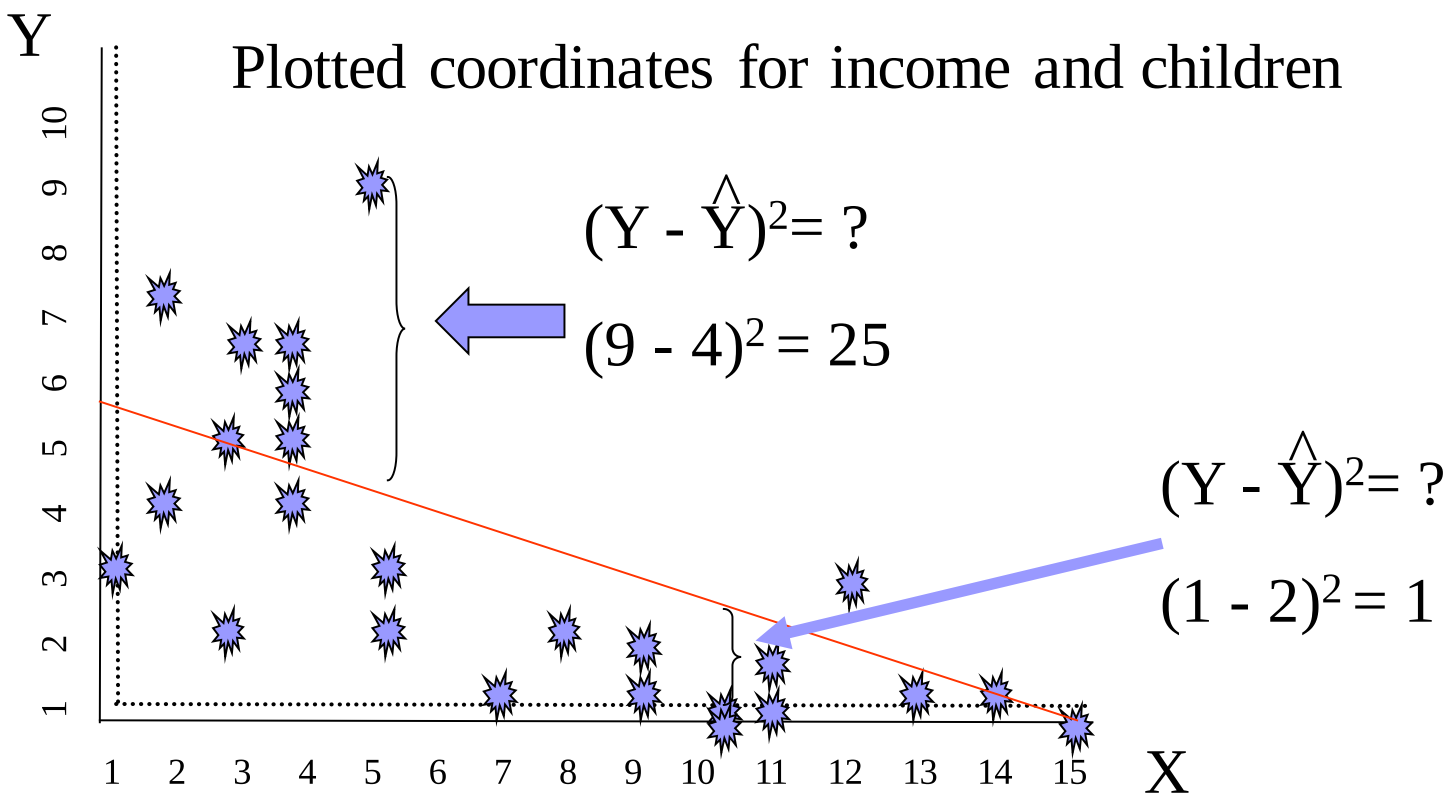
Correlation and Regression

Let's "fit the line" to the place where squared deviations from the line (vertically) are at their lowest value (across all X's).

Minimize this: $\Sigma(Y - \hat{Y})^2$ $\qquad$ $\hat{Y}$ **= line**

Minimizing the sum of squared errors gives you the unique, best fitting line for all the data points. It is the line that is closest to all points.

Correlation and Regression



Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

- $\Sigma(Y - \hat{Y})^2$ aka "sum of squared errors"
- There is a simple, elegant formula for "discovering" the line that minimizes the sum of squared errors—You don't have to memorize!

$$b = \frac{\Sigma((X - \bar{X})(Y - \bar{Y}))}{\Sigma(X - \bar{X})^2} \qquad a = \bar{Y} - b\bar{X} \qquad \hat{Y} = a + bX$$

- This is the method of least squares, it gives our least squares estimate and indicates why we call this technique "ordinary least squares" or OLS regression

Correlation and Regression

In fact, this is the output that would give you for the data values:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.679[a]	.460	.437	1.9048

a. Predictors: (Constant), INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.194	1	71.194	19.623	.000[a]
	Residual	83.446	23	3.628		
	Total	154.640	24			

a. Predictors: (Constant), INCOME

$$\hat{Y} = a + bX$$

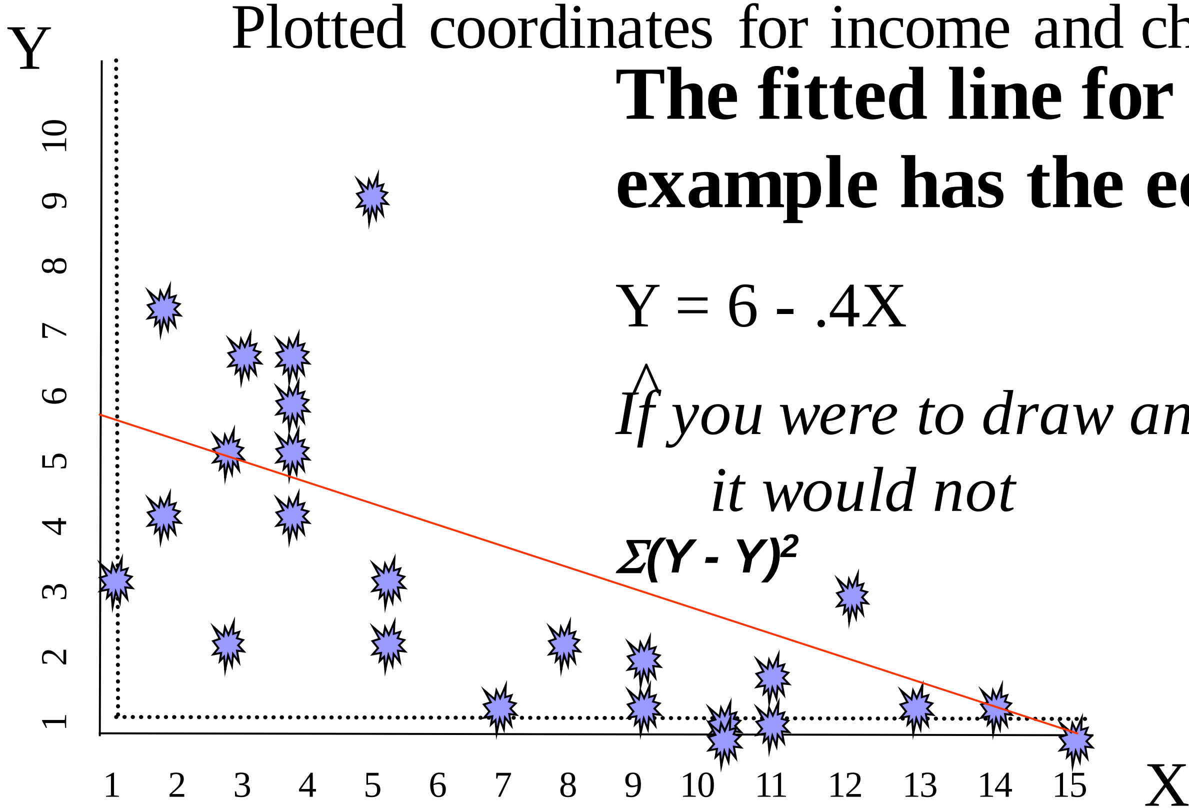
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.003	.754		7.960	.000
	INCOME	-.414	.094	-.679	-4.430	.000

a. Dependent Variable: CHILD

Correlation and Regression

Plotted coordinates for income and children



The fitted line for our example has the equation:

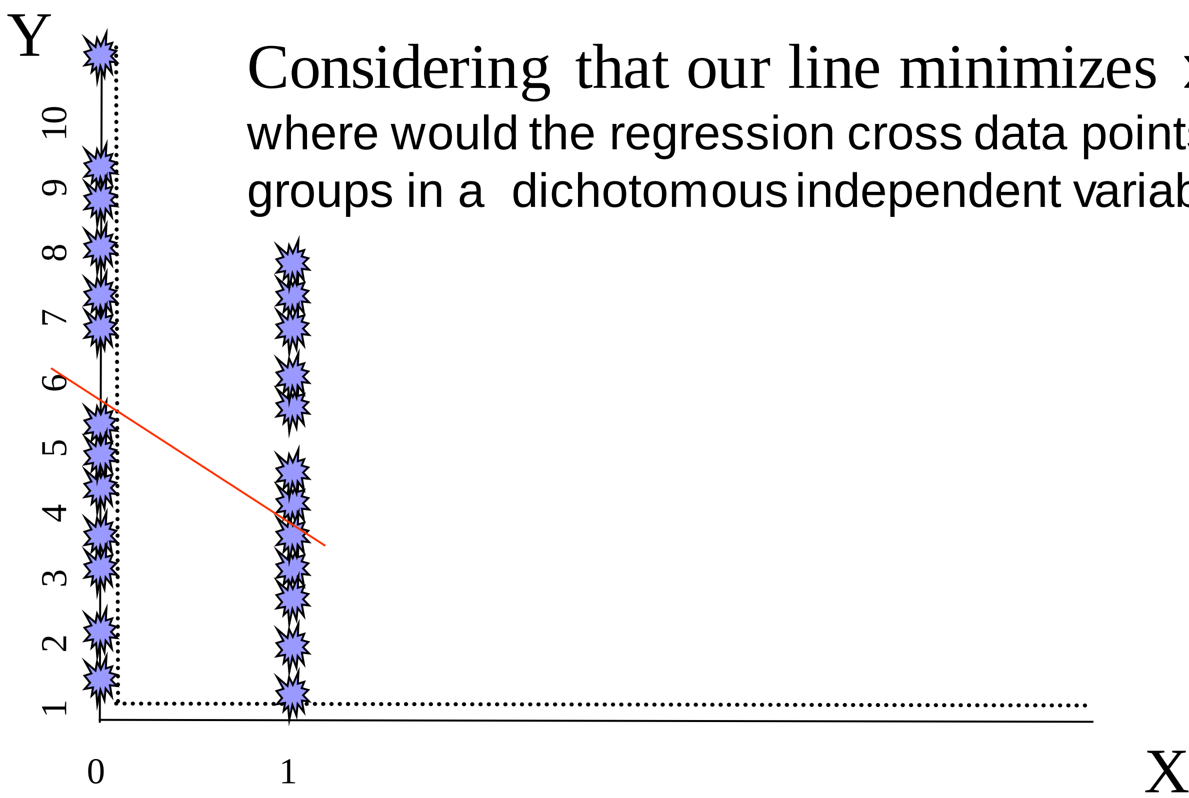
$\hat{Y} = 6 - .4X$

If you were to draw any other line, it would not minimize $\Sigma(Y - \hat{Y})^2$

Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

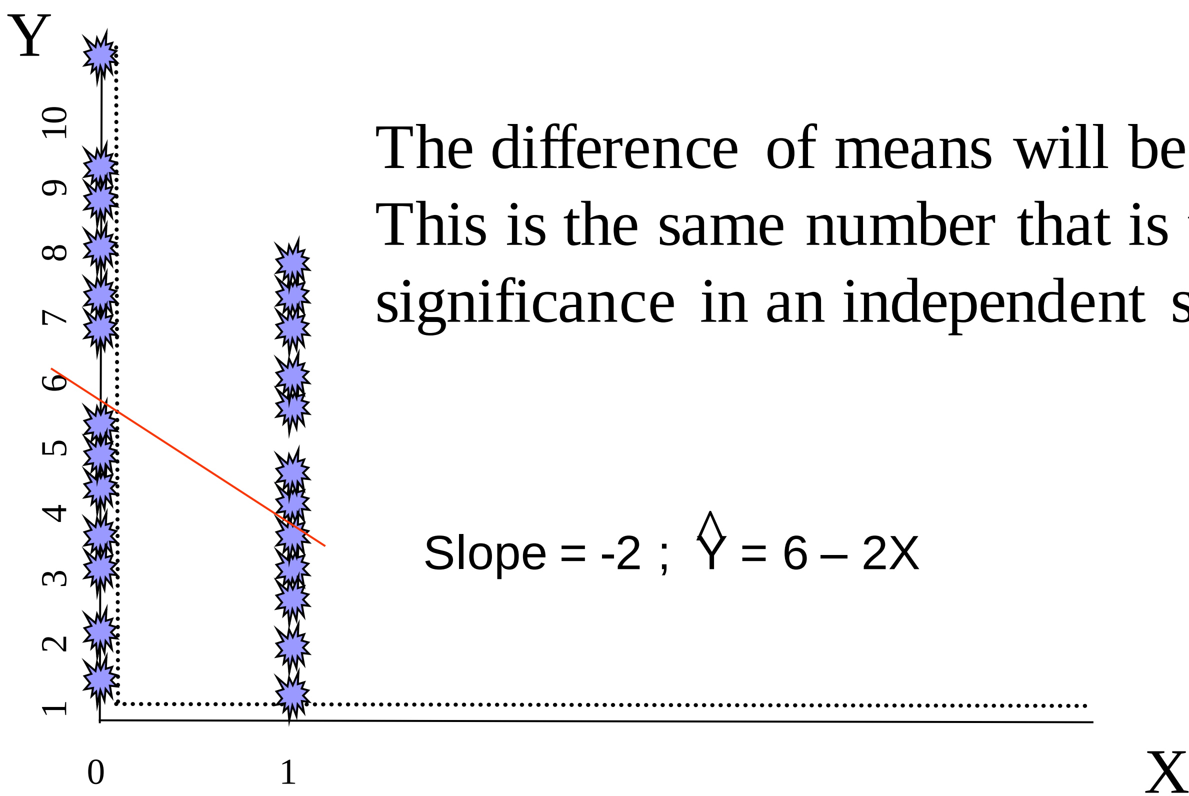
Considering that our line minimizes $\Sigma (Y - \hat{Y})^2$, where would the regression cross data points for two groups in a dichotomous independent variable?



0=Men: Mean = 6

1=Women: Mean = 4

Correlation and Regression



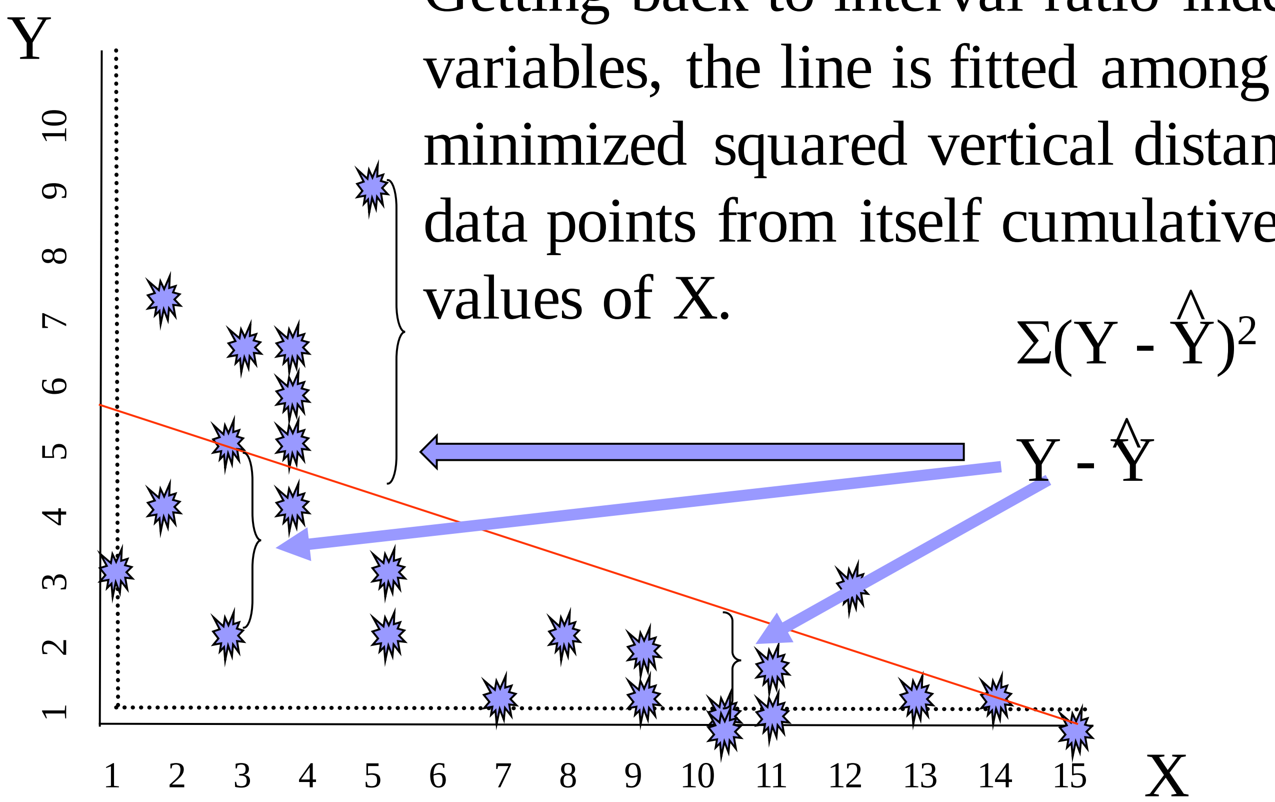
The difference of means will be the slope. This is the same number that is tested for significance in an independent samples t-test.

0=Men: Mean = 6

1=Women: Mean = 4

Correlation and Regression

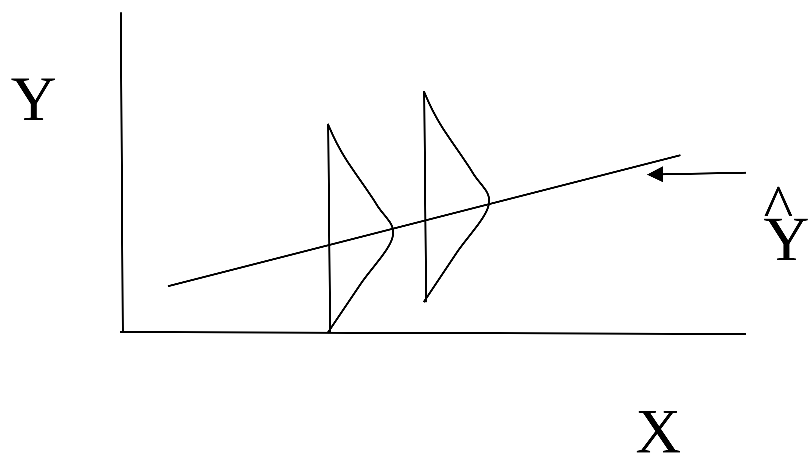
Getting back to interval-ratio independent variables, the line is fitted among the minimized squared vertical distance of all data points from itself cumulatively for all values of X.



Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

- $\hat{Y} = a + bX$ This equation gives the conditional mean of Y at any given value of X.

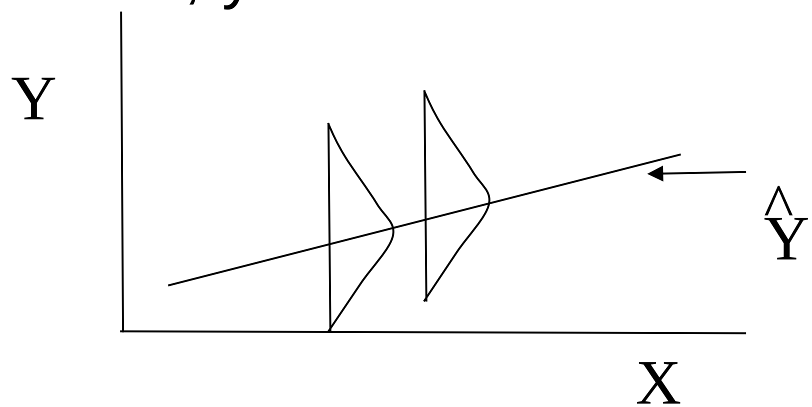


$\hat{Y}$ = Average for Y at each level of X

- So… In reality, our line gives us the expected mean of Y given each value of X
- The line's equation tells you how the mean on your dependent variable changes as your independent variable goes up.

Correlation and Regression

- As you know, every mean has a distribution around it--so there is a standard deviation. This is true for conditional means as well. So, you also have a conditional standard deviation.

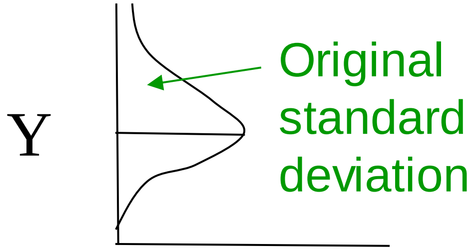


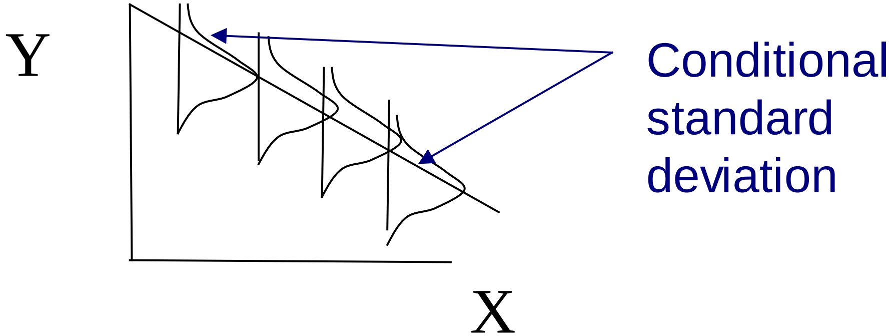
- "Conditional Standard Deviation" or "Root Mean Square Error" equals "approximate average deviation from the line."

- $$\hat{\sigma} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{\Sigma\,(Y - \hat{Y})^2}{n - 2}}$$

Correlation and Regression

- The Assumption of **Homoskedasticity**:
  - The variation around the line is the same no matter the X.
  - The conditional standard deviation is for any given value of X.





- If there is a relationship between X and Y, the conditional standard deviation is going to be less than the standard deviation of Y--if this is so, you have improved prediction of the mean value of Y by taking into account each level of X.
- If there were no relationship, the conditional standard deviation would be the same as the original, and the regression line would be flat at the mean of Y.

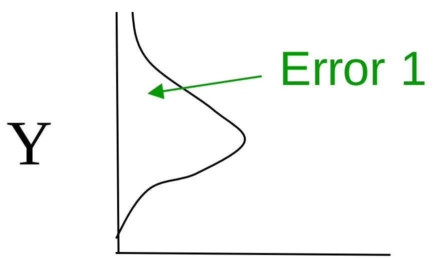
Correlation and Regression

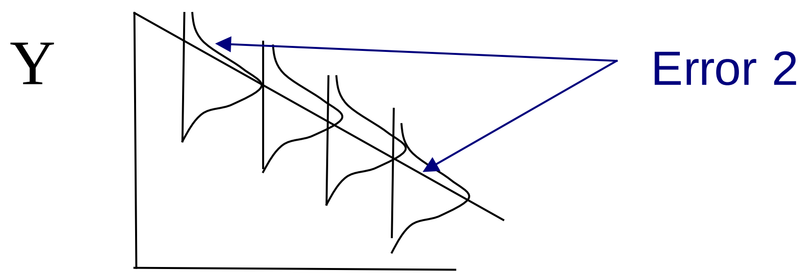
- So guess what?

- We have a way to determine how much our understanding of Y is improved when taking X into account—it is based on the fact that conditional standard deviations should be smaller than Y's original standard deviation.

Correlation and Regression

- Proportional Reduction in Error
  - Let's call the variation around the mean in Y "Error 1."
  - Let's call the variation around the line when X is considered "Error 2."





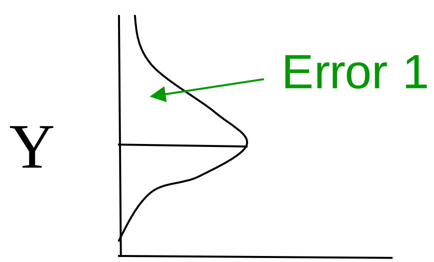
  - But rather than going all the way to standard deviation to determine error, let's just stop at the basic measure, Sum of Squared Deviations.
  - Error 1 (E1) = $\Sigma (Y - \bar{Y})^2$ also called "Sum of Squares"
  - Error 2 (E2) = $\Sigma (Y - \hat{Y})^2$ also called "Sum of Squared Errors"

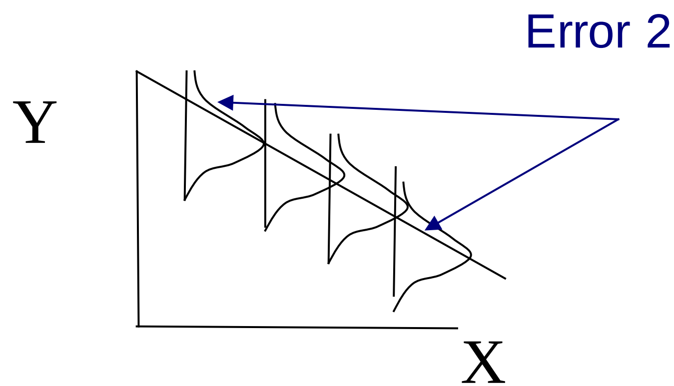
Correlation and Regression

- Proportional Reduction in Error
  - To determine how much taking X into consideration reduces the variation in Y (at each level of X) we can use a simple formula:

$$\frac{E1 - E2}{E1}$$

Which tells us the proportion or percentage of original error that is Explained by X.





  - Error 1 (E1) = $\Sigma (Y - \bar{Y})^2$
  - Error 2 (E2) = $\Sigma (Y - \hat{Y})^2$

Correlation and Regression

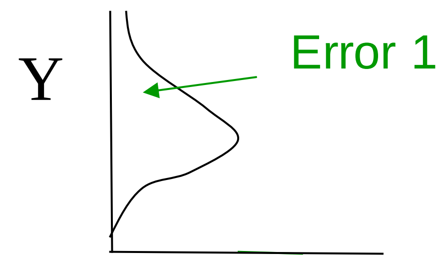
$$r^2 = \frac{E1 - E2}{E1}$$

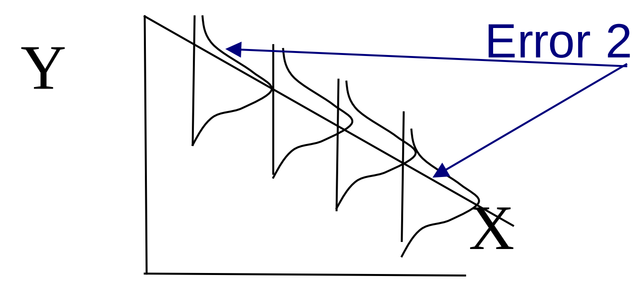
$$= \frac{TSS - SSE}{TSS}$$

$$= \frac{\Sigma (Y - \bar{Y})^2 - \Sigma (Y - \hat{Y})^2}{\Sigma (Y - \bar{Y})^2}$$

r^2 is called the "coefficient of determination"…

It is also the square of the Pearson correlation





Correlation and Regression

- R^2
  - Is the improvement obtained by using X (and drawing a line through the conditional means) in getting as near as possible to everybody's value for Y over just using the mean for Y alone.
  - Falls between 0 and 1
    - 1 means an exact fit (and there is no variation of scores around the regression line)
    - 0 means no relationship (and as much scatter around the line as in the original Y variable and a flat regression line (slope = 0) through the mean of Y)
  - Would be the same for X regressed on Y as for Y regressed on X
  - Can be interpreted as the percentage of variability in Y that is explained by X.
- *Some people get hung up on maximizing R^2, but this is too bad because any effect is still a finding—a small R^2 only indicates that you haven't told the whole (or much of the) story of the relationship between your variables.*

Correlation and Regression

Back to the output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.679[a]	.460	.437	1.9048

a. Predictors: (Constant), INCOME

r^2

$$\frac{\Sigma (Y - \bar{Y})2 \;-\; \Sigma (Y - \hat{Y})2}{\Sigma (Y - \bar{Y})2}$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.194	1	71.194	19.623	.000[a]
	Residual	83.446	23	3.628		
	Total	154.640	24			

a. Predictors: (Constant), INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.003	.754		7.960	.000
	INCOME	-.414	.094	-.679	-4.430	.000

a. Dependent Variable: CHILD

71.194 ÷ 154.64 = .460

Correlation and Regression

Q: So why did I see an ANOVA Table?

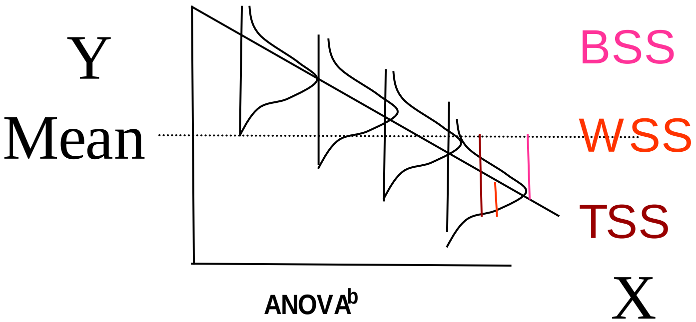
A: Levels of X can be thought of like groups in ANOVA

…and the squared distance from the line to the mean (Regression SS) is equivalent to BSS—group mean to big mean (but df = 1)

…and the squared distance from the line to the data values on Y (Residual SS) is equivalent to WSS—data value to the group's mean

… and the ratio of these forms an F distribution in repeated sampling

If F is significant, X is explaining some of the variation in Y.



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.194	1	71.194	19.623	.000[a]
	Residual	83.446	23	3.628		
	Total	154.640	24			

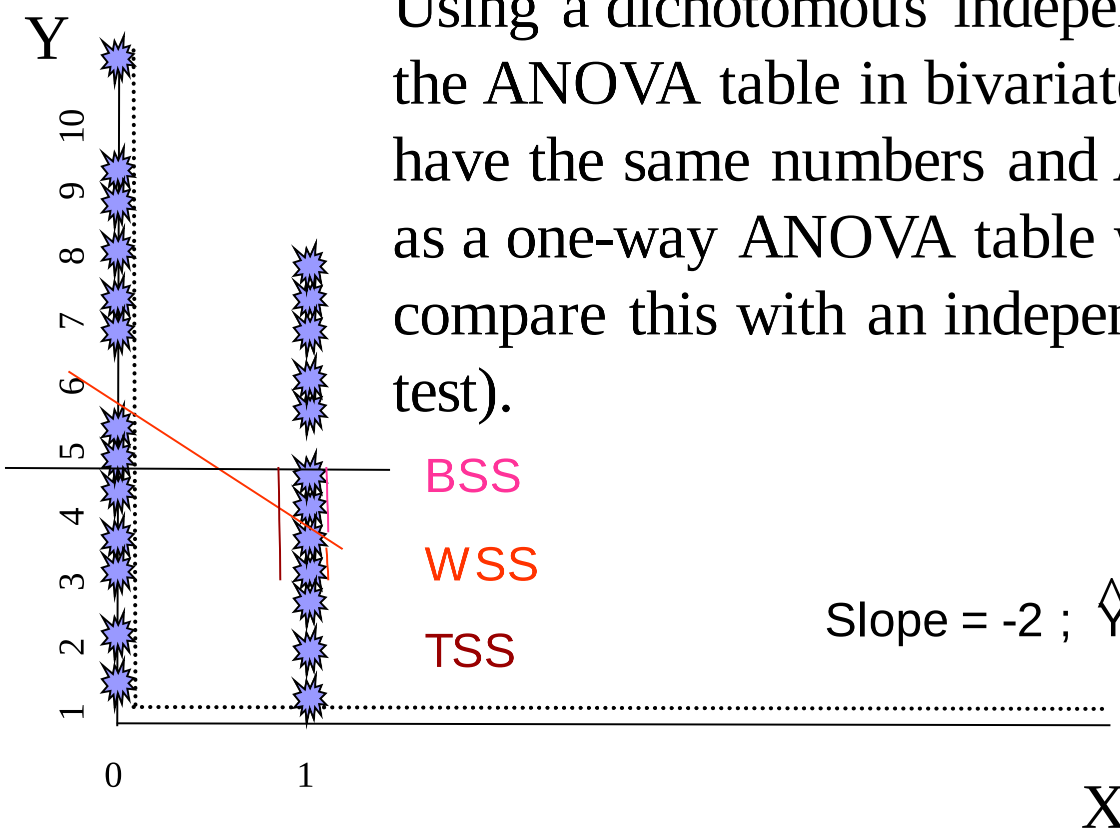
a. Predictors: (Constant), INCOME

b. Dependent Variable: CHILD

Correlation and Regression

Using a dichotomous independent variable, the ANOVA table in bivariate regression will have the same numbers and ANOVA results as a one-way ANOVA table would (and compare this with an independent samples t-test).

Mean = 5



Slope = -2 ; $\hat{Y} = 6 - 2X$

0=Men: Mean = 6

1=Women: Mean = 4

Correlation and Regression

Recall that statistics are divided between descriptive and inferential statistics.

Descriptive:

- The equation for your line is a descriptive statistic. It tells you the real, best-fitted line that minimizes squared errors.

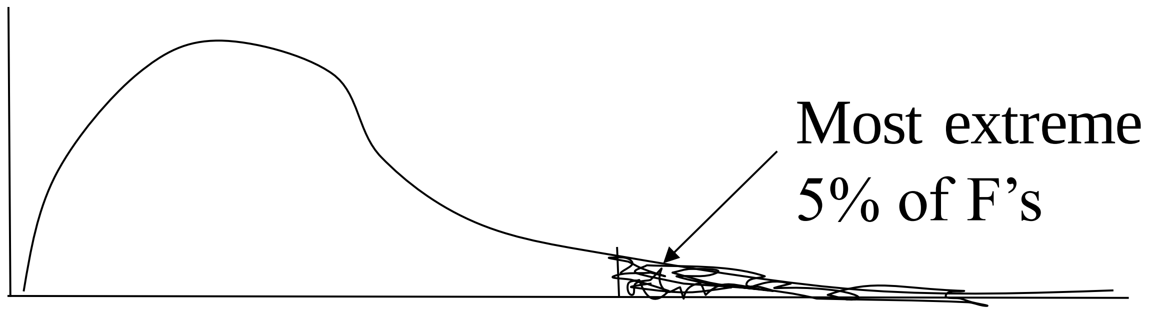
Inferential:

- But what about the population? What can we say about the relationship between your variables in the population???
- The inferential statistics are estimates based on the best-fitted line.

Correlation and Regression

- The significance of F, you already understand.

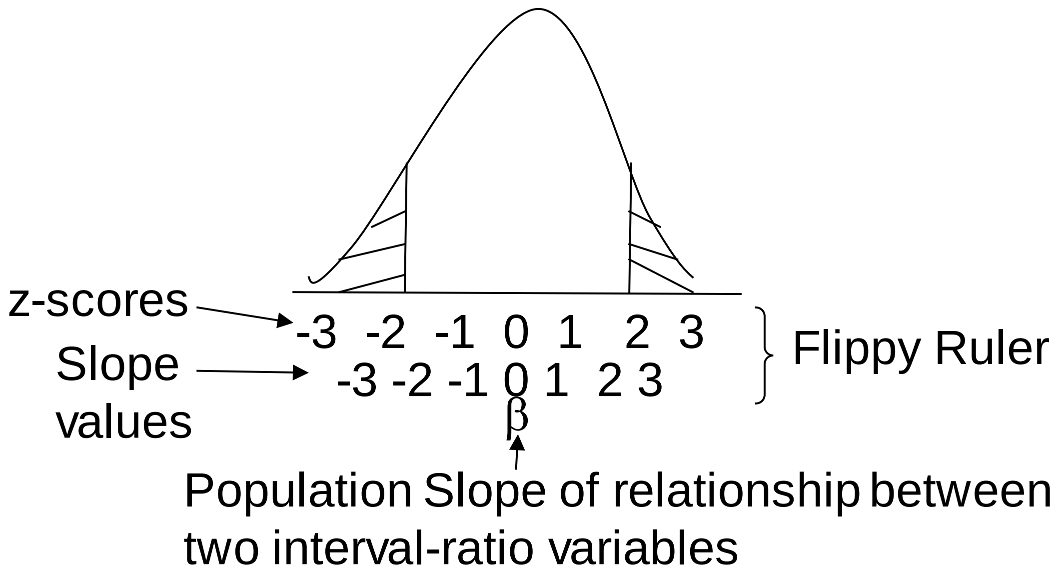
F = Regression SS / Residual SS



- The ratio of Regression (line to the mean of Y) to Residual (line to data point) Sums of Squares forms an F ratio in repeated sampling.
- Null: $r^2 = 0$ in the population. If F exceeds critical F, then your variables have a relationship in the population (X explains some of the variation in Y).

Correlation and Regression

- What about the Slope (called "Coefficient")?
  - The slope has a sampling distribution that is normally distributed.
  - So we can do a significance test.



Correlation and Regression

Conducting a Test of Significance for the slope of the Regression Line

By slapping the sampling distribution for the slope over a guess of the population's slope, H_o, we can find out whether our sample could have been drawn from a population where the slope is equal to our guess.

1. Two-tailed significance test for α-level = .05
2. Critical t = +/- 1.96
3. To find if there is a significant slope in the population,
   H_o: $\beta = 0$
   H_a: $\beta \neq 0$
4. Collect Data
5. Calculate t (z): $t = \dfrac{b - \beta_o}{\text{s.e.}}$ s.e. = $\dfrac{\sqrt{\dfrac{\Sigma(Y - \hat{Y})^2}{n - 2}}}{\sqrt{\Sigma(X - \bar{X})^2}}$
6. Make decision about the null hypothesis
7. Find P-value

Correlation and Regression

Back to the output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.679[a]	.460	.437	1.9048

a. Predictors: (Constant), INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.194	1	71.194	19.623	.000[a]
	Residual	83.446	23	3.628		
	Total	154.640	24			

a. Predictors: (Constant), INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.003	.754		7.960	.000
	INCOME	-.414	.094	-.679	-4.430	.000

a. Dependent Variable: CHILD

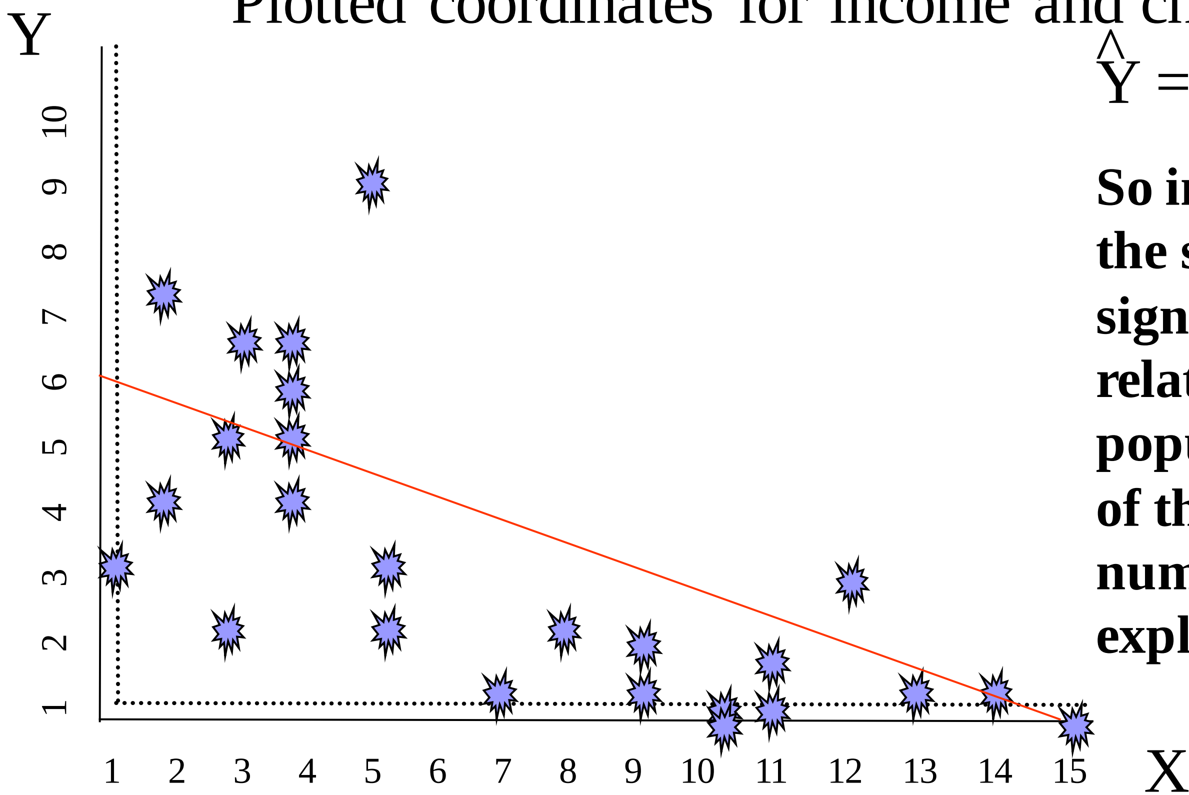
Of course, you get the standard error and

t on your output,

…and the p-value too!

Correlation and Regression

Plotted coordinates for income and children



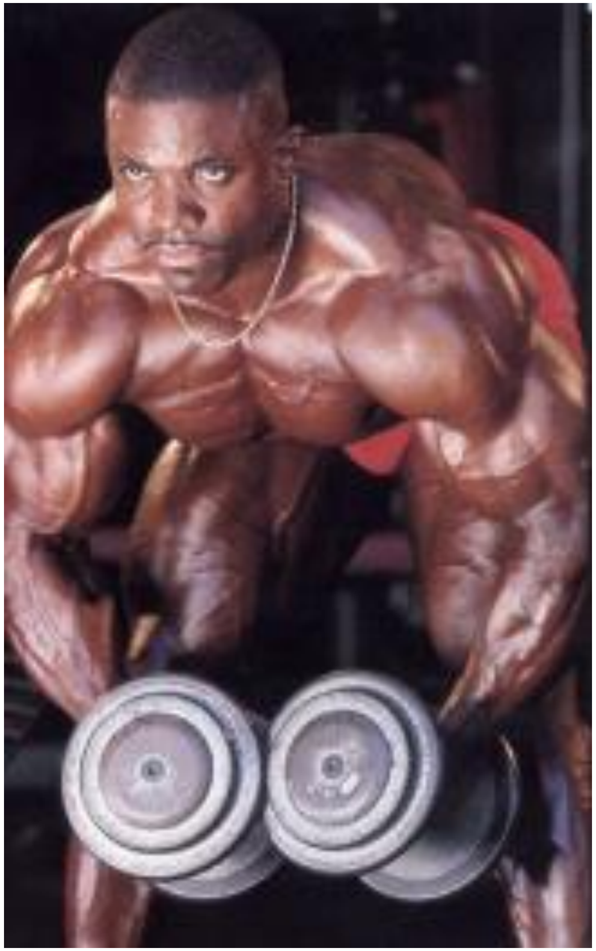
$$\hat{Y} = 6 - .4X$$

So in our example, the slope is significant, there is a relationship in the population, and 46% of the variation in number of children is explained by income.

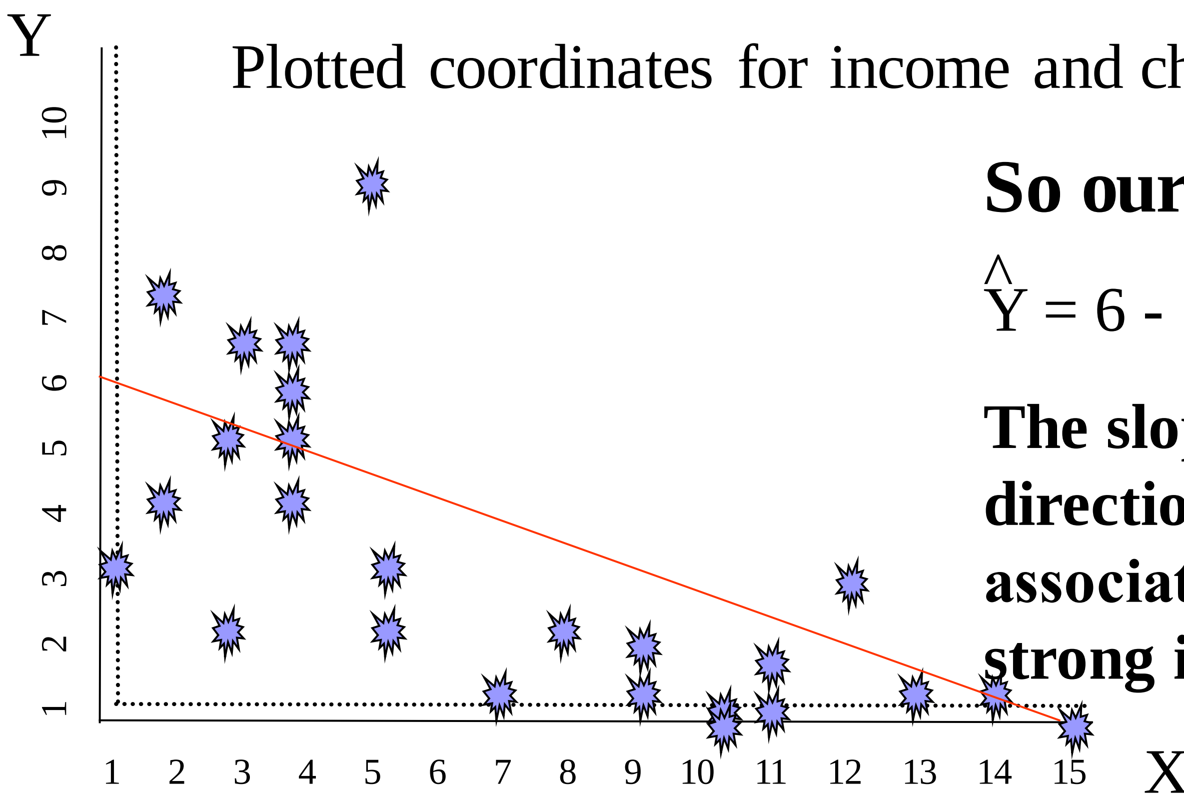
Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

- We've talked about the summary of the relationship, but not about strength of association.
- How strong is the association between our variables?
- For this we need correlation.



Correlation and Regression



So our equation is:

$$\hat{Y} = 6 - .4X$$

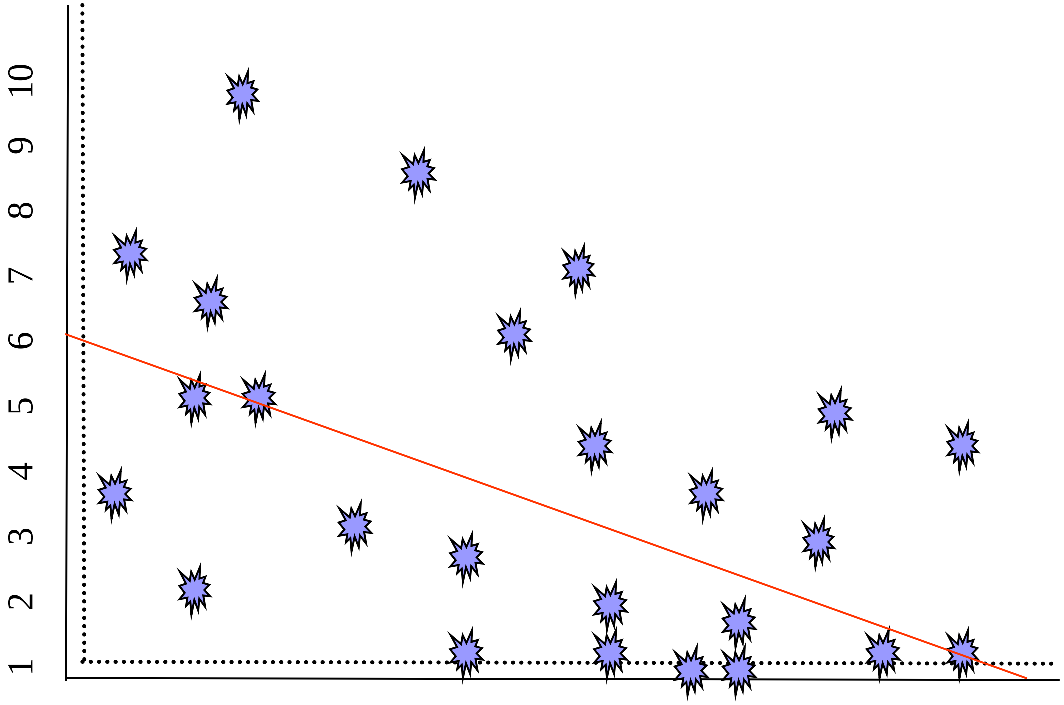
The slope tells us direction of association… How strong is that?

Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Children (Y):	2	5	1	9	6	3	1	0	3	7	7	2	4	2	1	0	1	2	4	3	0	1	2	5	7
Income 1=$10K (X):	3	4	9	5	4	12	14	10	1	4	3	11	4	9	13	10	7	5	2	5	15	11	8	3	2

Correlation and Regression

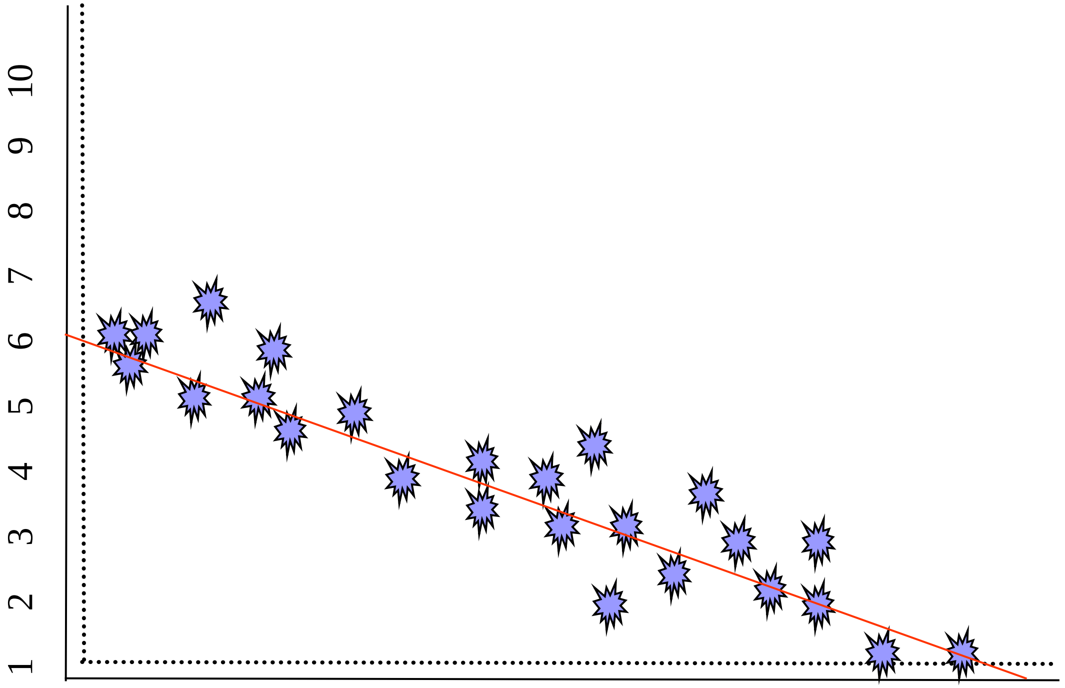
- To find the strength of the relationship between two variables, we need correlation.
- The correlation is the standardized slope… it refers to the standard deviation change in Y when you go up a standard deviation in X.

Correlation and Regression



Example of *Low* Negative Correlation

Correlation and Regression



Example of *High* Negative Correlation

Correlation and Regression

- The correlation is the standardized slope… it refers to the standard deviation change in Y when you go up a standard deviation in X.
- Recall that s.d. of x, $Sx = \sqrt{\dfrac{\Sigma(X - \bar{X})^2}{n - 1}}$
- and the s.d. of y, $Sy = \sqrt{\dfrac{\Sigma(Y - \bar{Y})^2}{n - 1}}$
- Pearson correlation, $r = \left(\dfrac{Sx}{Sy}\right) b$

Correlation and Regression

- The Pearson Correlation, r:
  - tells the direction and strength of the relationship between continuous variables
  - ranges from -1 to +1
  - is + when the relationship is positive and - when the relationship is negative
  - the higher the absolute value of r, the stronger the association
  - a standard deviation change in x corresponds with r standard deviation change in Y

Correlation and Regression

- The Pearson Correlation, r:
  - The pearson correlation is a statistic that is an inferential statistic too.
  - $t_{n-2} = \dfrac{r - (null = 0)}{\sqrt{(1-r^2)\,(n-2)}}$
  - When it is significant, there is a linear relationship between the two variables in the population—it is *not* non-existent!

Correlation and Regression

Our data's correlation is .679. How strong is that?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.679[a]	.460	.437	1.9048

a. Predictors: (Constant), INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.194	1	71.194	19.623	.000[a]
	Residual	83.446	23	3.628		
	Total	154.640	24			

a. Predictors: (Constant), INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.003	.754		7.960	.000
	INCOME	-.414	.094	-.679	-4.430	.000

a. Dependent Variable: CHILD

Correlation, r, is significant.

Correlation and Regression

If you were to use the "correlate, bivariate" command, you'd get this ouput…

Correlations

Correlation, r, is significant.

Correlations

		CHILDREN	INCOME
CHILDREN	Pearson Correlation	1	-.679**
	Sig. (2-tailed)	.	.000
	N	25	25
INCOME	Pearson Correlation	-.679**	1
	Sig. (2-tailed)	.000	.
	N	25	25

**. Correlation is significant at the 0.01 level (2-tailed).