Lecture 18 - Correlation and Regression

Sta102 / BME102

Colin Rundel

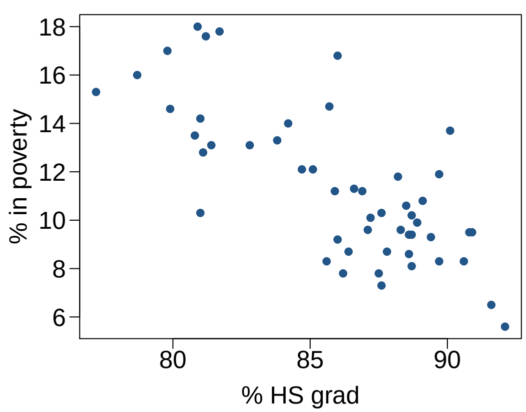
April 1, 2015

Modeling numerical variables

- So far we have worked with single numerical and categorical variables, and explored relationships between numerical and categorical, and two categorical variables.
- Today we will learn to quantify the relationship between two numerical variables.
- Next week we will learn to model numerical variables using many predictor (independent) variables (including both numerical and categorical) at once.

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response?

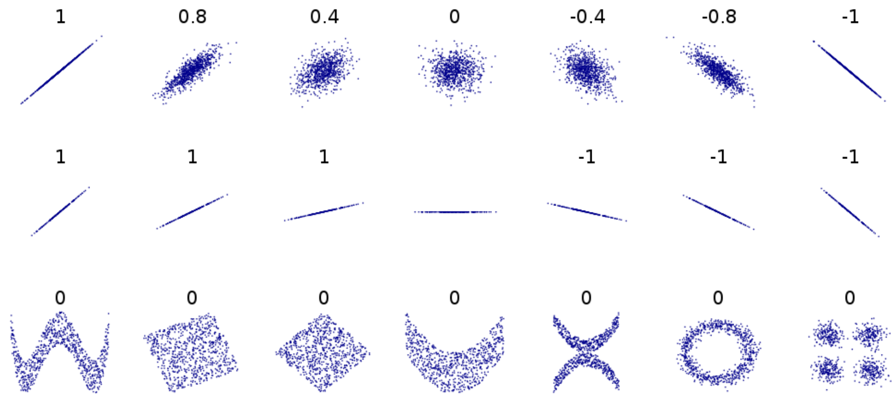
Predictor?

Relationship?

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.
- We use ρ to indicate the population correlation coefficient, and R or r to indicate the sample correlation coefficient.

Correlation Examples



From http://en.wikipedia.org/wiki/Correlation

Covariance

We have previously discussed the variance as a measure of uncertainty of a random variable:

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$$

In order to define correlation we first need to define covariance, which is a generalization of variance to two random variables

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

Covariance is not a measure of uncertainly but rather a measure of the degree to which X and Y tend to be large (or small) at the same time or the degree to which one tends to be large while the other is small.

Covariance, cont.

The magnitude of the covariance is not very informative since it is affected by the magnitude of both X and Y. However, the sign of the covariance tells us something useful about the relationship between X and Y.

Consider the following conditions:

- $x_i > \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i < \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i > \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be negative.
- $x_i < \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be negative.

Properties of Covariance

- $Cov(X,X) = Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,Y) = 0$ if X and Y are independent
- $Cov(X,c) = 0$
- $Cov(aX,bY) = ab\ Cov(X,Y)$
- $Cov(X+a,Y+b) = Cov(X,Y)$
- $Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)$

Correlation

Since $Cov(X, Y)$ depends on the magnitude of X and Y we would prefer to have a measure of association that is not affected by changes in the scales of the variables.

The most common measure of *linear* association is correlation which is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \, \sigma_Y}$$

$$-1 < \rho(X, Y) < 1$$

Where the magnitude of the correlation measures the strength of the *linear* association and the sign determines if it is a positive or negative relationship.

Correlation and Independence

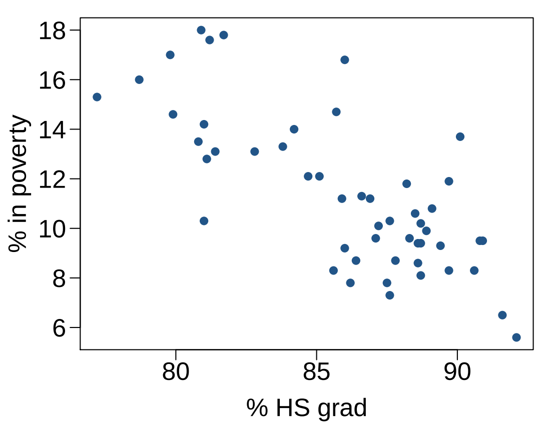
Given random variables X and Y

$$X \text{ and } Y \text{ are independent} \implies Cov(X, Y) = \rho(X, Y) = 0$$

$$Cov(X, Y) = \rho(X, Y) = 0 \;\not\Longrightarrow\; X \text{ and } Y \text{ are independent}$$

Guessing the correlation

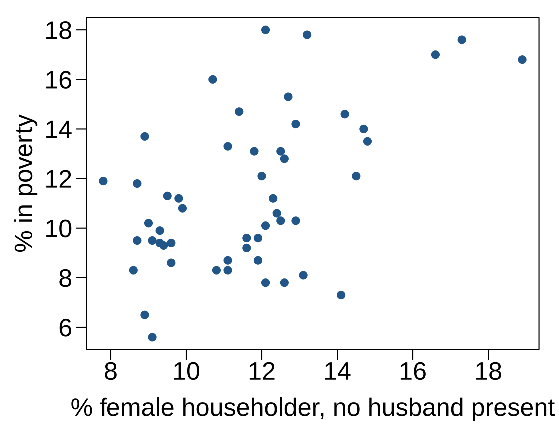
Which of the following is the best guess for the correlation between % in poverty and % HS grad?



(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5

Guessing the correlation

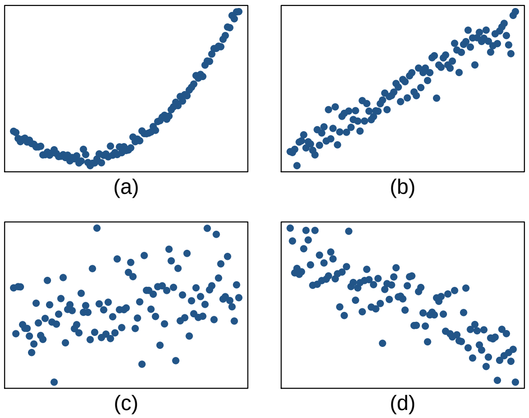
Which of the following is the best guess for the correlation between % in poverty and % single mother household?



(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

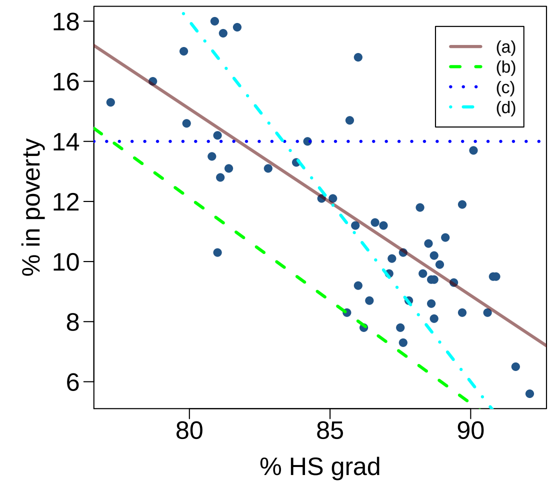
Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?

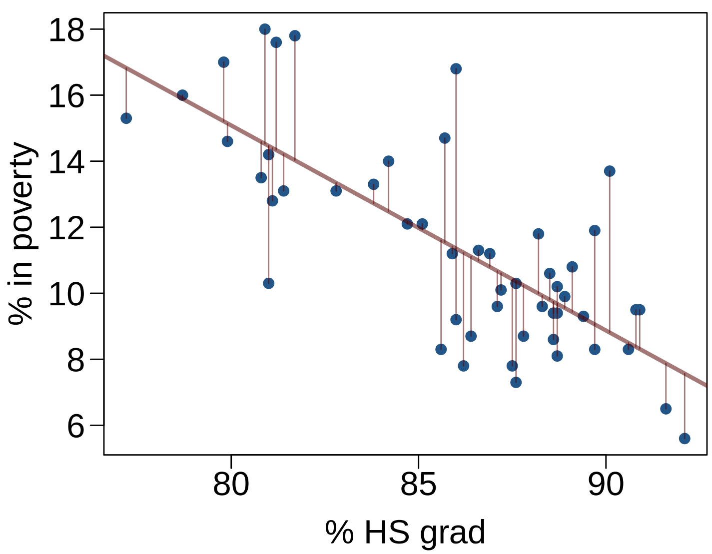


Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?



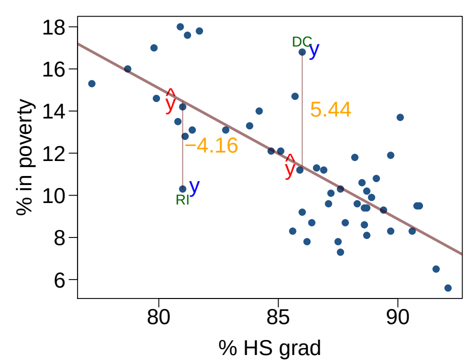
Quantifying best fit



Residuals

A residual is the difference between the observed and predicted y.

$$y_i = \hat{y}_i + e_i \quad \Rightarrow \quad e_i = y_i - \hat{y}_i$$



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
     $$|e_1| + |e_2| + \cdots + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – *least squares*
     $$e_1^2 + e_2^2 + \cdots + e_n^2$$
- Why least squares?
  1. Most commonly used
  2. Square is a nicer function than absolute value
  3. In many applications, a residual twice as large as another is more than twice as bad

The least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

predicted y

intercept

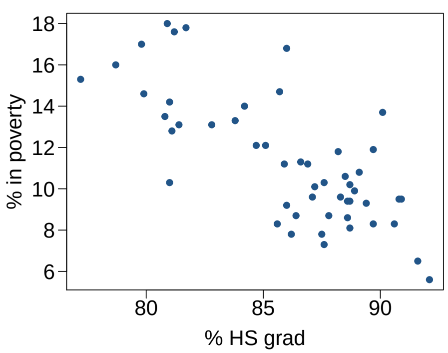
slope

predictor variable

Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Slope

The slope of the regression line is calculated as

$$b_1 = \frac{s_y}{s_x} R$$

In context...

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

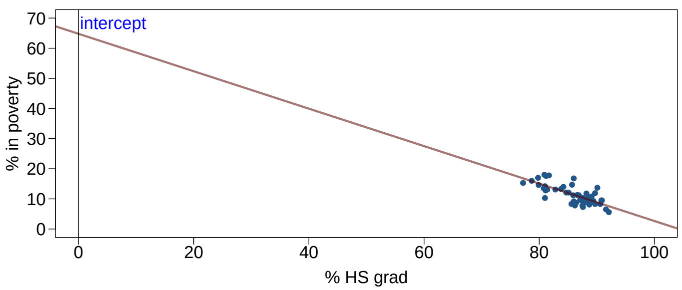
Interpretation

For each % point increase in HS graduate rate, we would *expect* the % living in poverty to decrease *on average* by 0.62% points.

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line must pass through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x}$$



$$b_0 = 11.35 - (-0.62) \times 86.01 = 64.68$$

Interpreting Intercepts

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

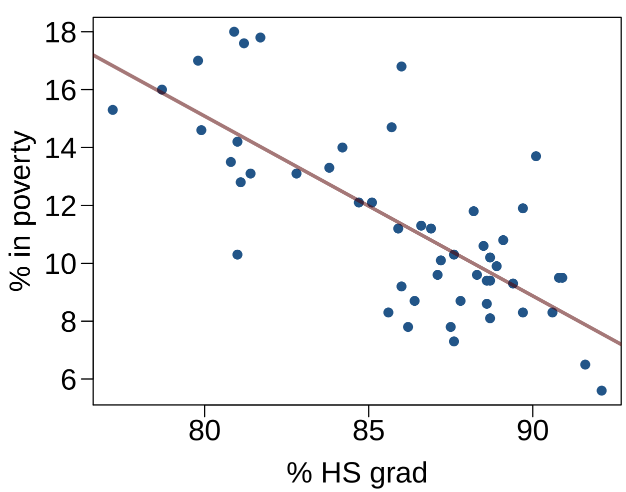
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

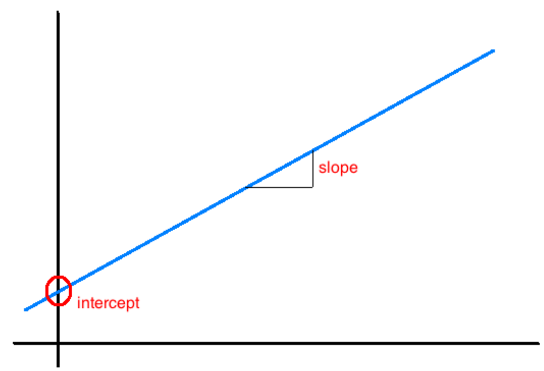
Regression line

$$[\widehat{\%\ in\ poverty}] = 64.68 - 0.62\ [\%\ HS\ grad]$$



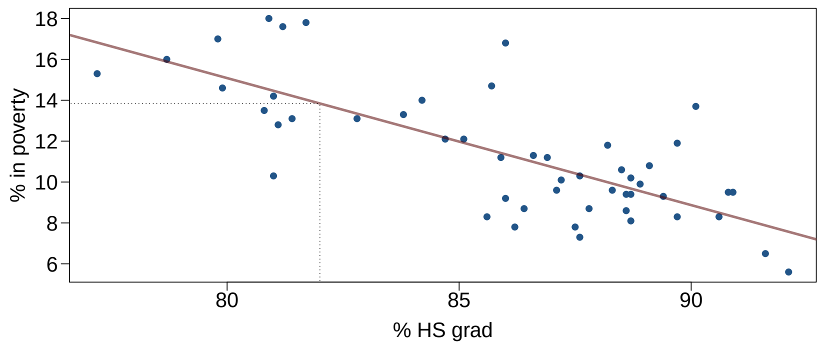
Interpretation of slope and intercept

- *Intercept:* When $x = 0$, y is expected to equal *the intercept*.
- *Slope:* For each *unit* increase in x, y is expected to *increase/decrease* on average by *the slope*.



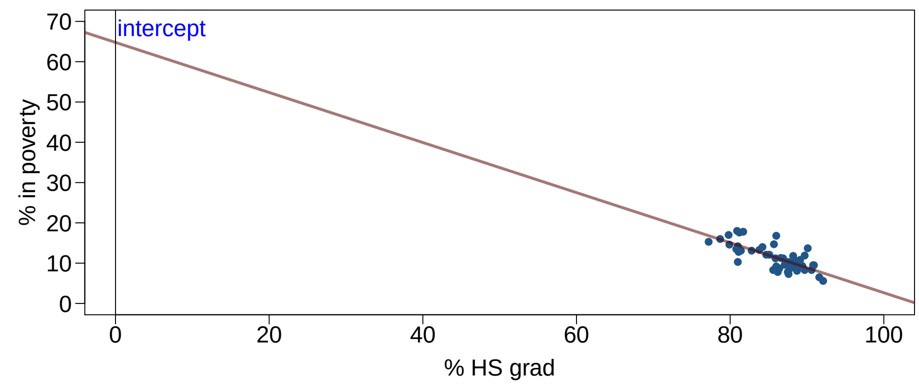
Prediction

- Using the linear model to predict the value of the response variable for a given value of the predictor variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value - we'll talk about this next time.

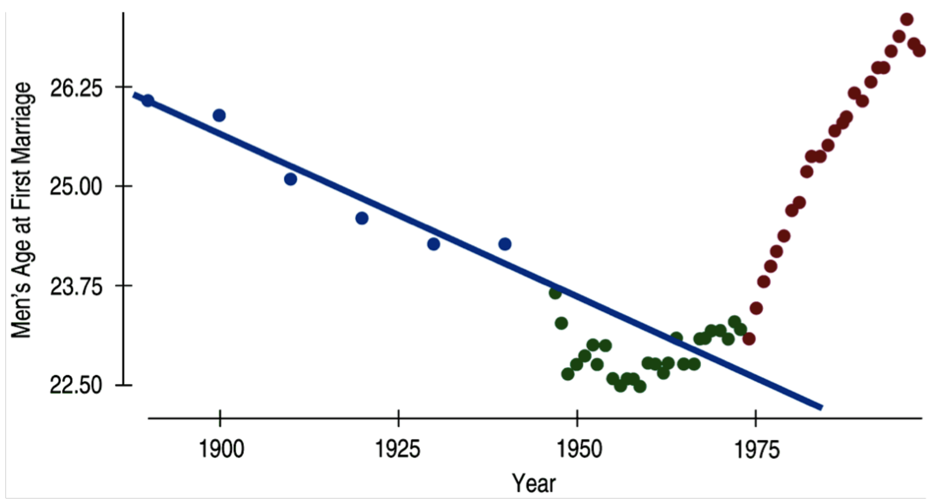


Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation



Examples of extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend | Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.

Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times

Examples of extrapolation

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

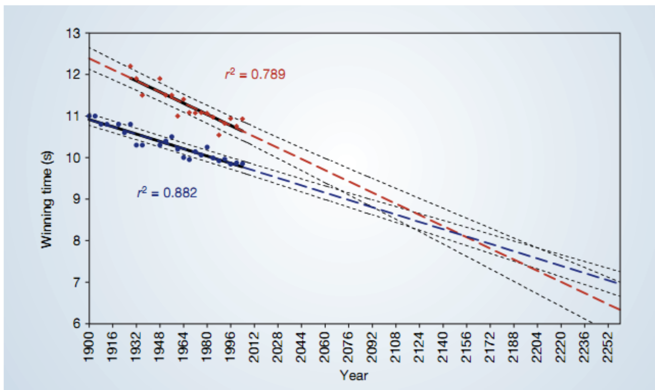


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women) … sect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable (y) is explained by the predictor variables (x).
- The remainder of the variability is "unexplained".
- Sometimes refered to as the coefficient of determination.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.5625$.

Interpretation of R^2

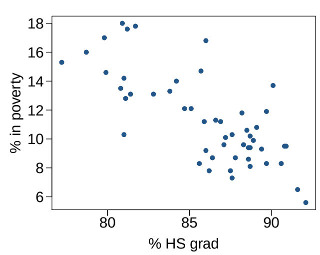
Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.5625$?

(a) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

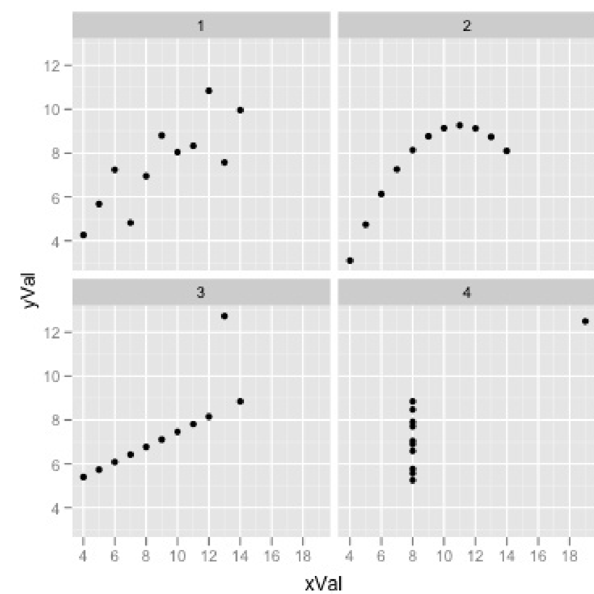
(b) 56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 75% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Anscombe's Quartet



Anscombe's Quartet - Data

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	0.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

All four datasets have the same regression line:

$$y = 3 + 0.5x$$