Math 1710 Class 20

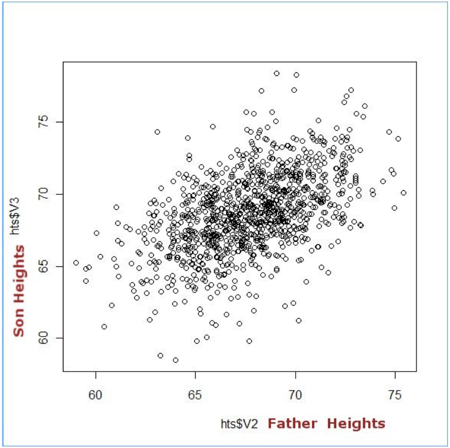
Association, Correlation, Regression

Dr. Back

Oct. 14, 2009

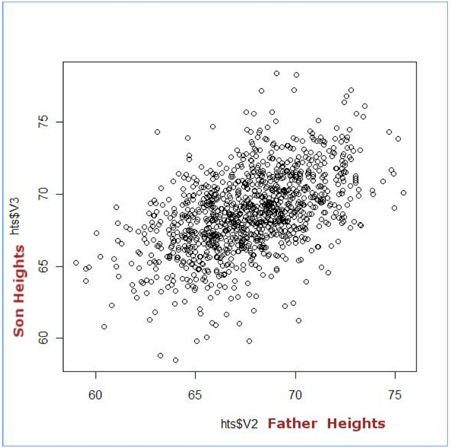
Son's Heights from Their Fathers

Galton's Original 1886 Data



Son's Heights from Their Fathers

Galton's Original 1886 Data



If you know a father's height, what can you say about his son's?

Son's Heights from Their Fathers

If you know a father's height, what can you say about his son's?

> hts	V2	V3
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	61.13723	64.28113
6	63.02254	64.24221
7	65.37053	64.08231
8	64.72398	63.99574
9	66.06509	64.61338
10	66.96738	63.97944
...	...	...

Son's Heights from Their Fathers

If you know a father's height, what can you say about his son's?

>summary(hts$V3) *(Sons)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
58.51	66.93	68.62	68.68	70.47	78.36

>summary(hts$V2) *(Fathers)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
59.01	65.79	67.77	67.69	69.60	75.43

Son's Heights from Their Fathers

If you know a father's height, what can you say about his son's?

>summary(hts$V3) *(Sons)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
58.51	66.93	68.62	68.68	70.47	78.36

>summary(hts$V2) *(Fathers)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
59.01	65.79	67.77	67.69	69.60	75.43

> sd(hts$V3)
[1] 2.814702
> sd(hts$V2)
[1] 2.744868

Son's Heights from Their Fathers

If you know a father's height, what can you say about his son's?

>summary(hts$V3) *(Sons)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
58.51	66.93	68.62	68.68	70.47	78.36

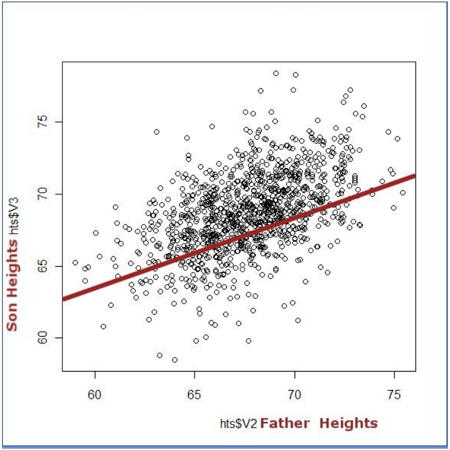
>summary(hts$V2) *(Fathers)*

Min.	*1stQu.*	*Median*	*Mean*	*3rdQu.*	*Max.*
59.01	65.79	67.77	67.69	69.60	75.43

```
> cor(hts$V3 , hts$V2)
[1] 0.5013383
```

Son's Heights from Their Fathers

Galton Data with Line of Regression



Son's Heights from Their Fathers

Suppose $y = \text{height}_{\text{son}}$ were linearly related to $x = \text{height}_{\text{father}}$ by $y = \beta_1 x + \beta_0$.

Son's Heights from Their Fathers

Suppose $y = \text{height}_{\text{son}}$ were linearly related to $x = \text{height}_{\text{father}}$ by $y = \beta_1 x + \beta_0$.

We'd then have $\bar{y} = \beta_1 \bar{x} + \beta_0$ as well and

$$(y - \bar{y}) = \beta_1 (x - \bar{x}).$$

Son's Heights from Their Fathers

Suppose $y = \text{height}_{\text{son}}$ were linearly related to $x = \text{height}_{\text{father}}$ by $y = \beta_1 x + \beta_0$.

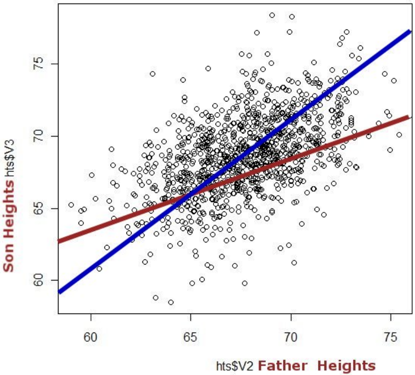
We'd then have $\bar{y} = \beta_1 \bar{x} + \beta_0$ as well and

$$(y - \bar{y}) = \beta_1 (x - \bar{x}).$$

Furthermore, the slope β_1 would be the ratio of standard deviations:

$$\beta_1 = \frac{s_{\text{son}}}{s_{\text{father}}}$$

Galton Data with line of slope $\frac{s_{\text{son}}}{s_{\text{father}}}$ added in blue.



Son's Heights from Their Fathers

The fact that the red "best fitting line" (termed the line of regression) actually has less slope than the blue line is one form of Galton's

"regression to the mean."

Tall fathers tend to have tall sons, but the sons are typically not as extreme in their tallness as their fathers were.

Son's Heights from Their Fathers

The fact that the red "best fitting line" (termed the line of regression) actually has less slope than the blue line is one form of Galton's

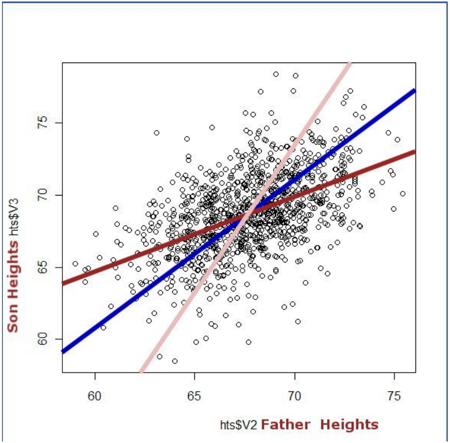
"regression to the mean."

Tall fathers tend to have tall sons, but the sons are typically not as extreme in their tallness as their fathers were.

In modern terms, we see two usages of the word "regression" here.

Son's Heights from Their Fathers

Galton Data with Other Line of Regression in pink:
(predict father's height from son's)



Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Scatterplot: Plot x horizontally, y vertically.

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
If one variable is potentially *explanatory* for the *response* of the other, choose the explanatory variable as x.

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Association is what we care about.
If we know the x value of a point, does it tell us something about the likely y value?

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$

Association is what we care about.

If we know the x value of a point, does it tell us something about the likely y value?

Correlation (a number) and regression (a line) are just techniques to study association.

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Principal Aspects of Association:

- *Direction:*
- *Strength:*
- *Form:*

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Principal Aspects of Association:

- *Direction:* positive or negative
- *Strength:*
- *Form:*

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Principal Aspects of Association:

- *Direction:* positive or negative
- *Strength:*
- *Form:*

Negative means as x increases, y *generally* decreases.

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Principal Aspects of Association:

- *Direction:*
- *Strength:* strong, moderate, or weak
- *Form:*

Association

Given: paired data $(x_1, y_1), \ldots, (x_n, y_n)$
Principal Aspects of Association:

- *Direction:*
- *Strength:*
- *Form:* linear, curved, or clustered

Association Examples

Math 1710
Class 20

V2u

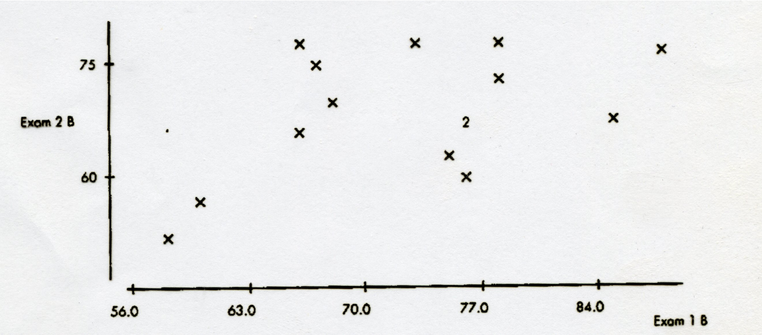
Last Time

Graphs and Association

Correlation

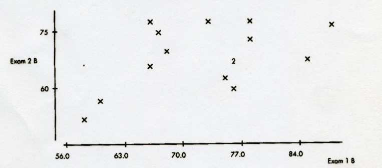
Regression

Example b



Association Examples

Example b



My call:

- *Direction:* positive
- *Strength:* moderate
- *Form:* curved

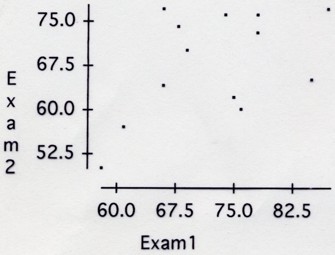
Association Examples

Example b using Data Desk

Pearson Product-Moment Correlation

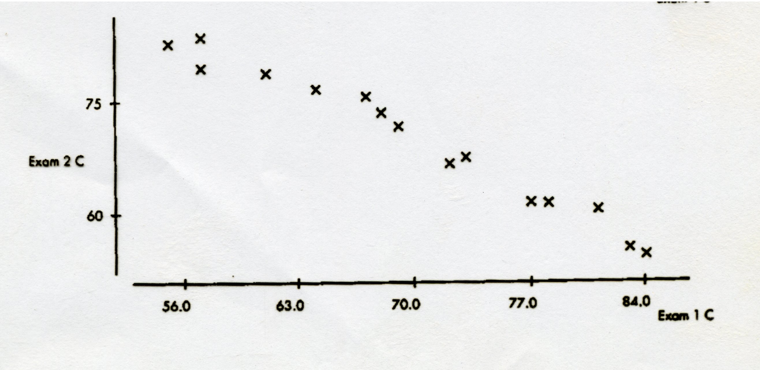
No Selector

	Exam2	Exam1
Exam2	1.000	
Exam1	0.496	1.000



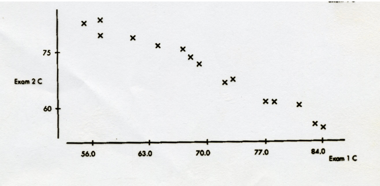
Association Examples

Example c



Association Examples

Example c

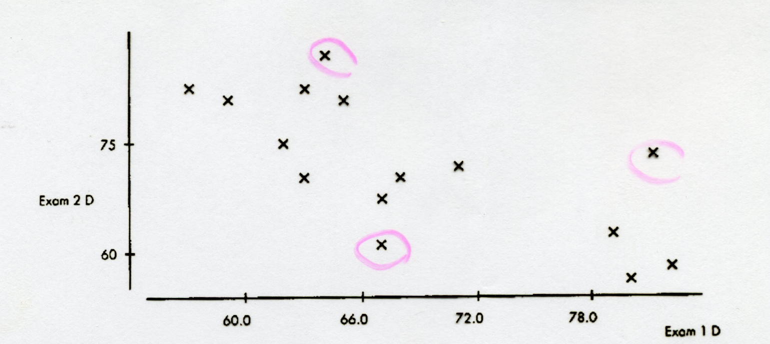


My call:

- *Direction:* negative
- *Strength:* strong
- *Form:* linear

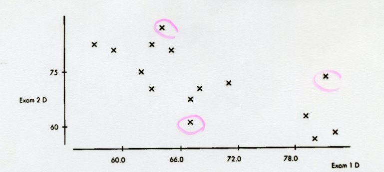
Association Examples

Example d



Association Examples

Example d



My call:

- *Direction:* negative
- *Strength:* moderate
- *Form:* (perhaps some outliers)

Association Examples

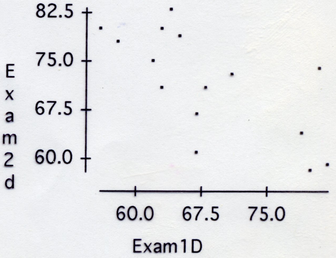
Example d using Data Desk

Pearson Product-Moment Correlation

No Selector

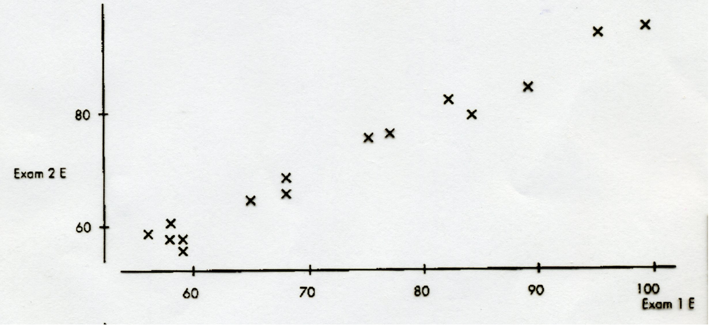
18 total cases of which 3 are missing

	Exam2d	Exam1D
Exam2d	1.000	
Exam1D	-0.694	1.000



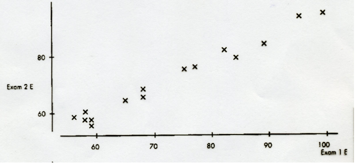
Association Examples

Example e



Association Examples

Example e

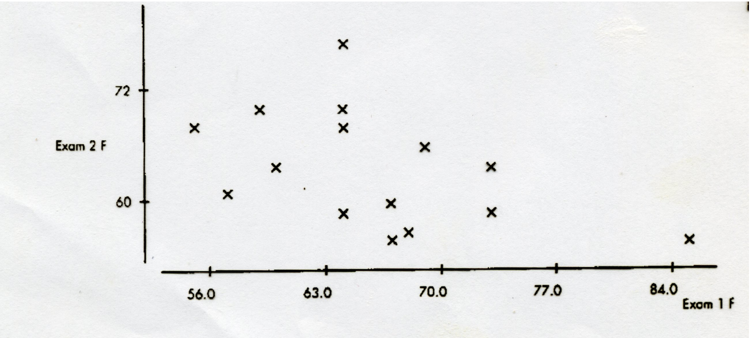


My call:

- *Direction:* positive
- *Strength:* strong
- *Form:* linear

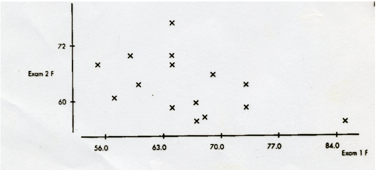
Association Examples

Example f



Association Examples

Example f



My call:

- *Direction:* negative
- *Strength:* weak
- *Form:* curved, maybe an outlier

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Correlation Properties

$$r = \frac{1}{n-1}\Sigma\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Positive association means as x increases, so does y
(And similarly when x decreases.)

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Positive association means as x increases, so does y
(And similarly when x decreases.)
So for pos. association, most terms in the sum are either $(+) \cdot (+)$ or $(-) \cdot (-)$.

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Positive association means as x increases, so does y
(And similarly when x decreases.)
So for pos. association, most terms in the sum are either $(+) \cdot (+)$ or $(-) \cdot (-)$.
Thus with pos. association, r tends to be positive.

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

pos r $\longleftrightarrow$ pos. association

neg. r $\longleftrightarrow$ neg. association

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$-1 <= r <= 1$ ($= \pm 1$ only for perfect linear association)

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$-1 <= r <= 1$ ($= \pm 1$ only for perfect linear association)

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$-1 <= r <= 1$ ($= \pm 1$ only for perfect linear association)
To see $= \pm 1$ for perfect linear association
$y = \beta_1 x + \beta_0$ means $s_y = \beta_1 s_x$ and $\bar{y} = \beta_1 \bar{x} + \beta_0$.

Correlation Properties

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

To see $= \pm 1$ for perfect linear association
$y = \beta_1 x + \beta_0$ means $s_y = \beta_1 s_x$ and $\bar{y} = \beta_1 \bar{x} + \beta_0$.

$$\begin{aligned} r &= \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\beta_1\left(\frac{x_i-\bar{x}}{|\beta_1| s_x}\right) \\ &= \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)^2\frac{\beta_1}{|\beta_1|} \\ &= \frac{\beta_1}{|\beta_1|} \end{aligned}$$

where the last line used the definition of the variance.

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

If you've studied vectors, the fact $-1 <= r <= 1$ comes from the same mathematics which explains why

$$\cos\theta = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|}$$

has right hand side between -1 and $+1$.

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

r is unchanged if x and y are exchanged.

Correlation Properties

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

invariant under rescaling

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

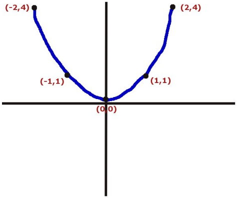
Curved association and r=0 are consistent!

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Curved association and r=0 are consistent!
Five points along $y = x^2$. ($\bar{x} = 0$ and $\bar{y} = 2$.)

x	y	(x-0)	(y-2)	(x-0)(y-2)
0	0	0	-2	0
1	1	1	-1	-1
-1	1	-1	-1	1
2	4	2	2	4
-2	4	-2	2	-4

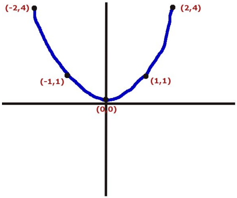


Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Curved association and r=0 are consistent!
Five points along $y = x^2$. ($\bar{x} = 0$ and $\bar{y} = 2$.)

x	y	(x-0)	(y-2)	(x-0)(y-2)
0	0	0	-2	0
1	1	1	-1	-1
-1	1	-1	-1	1
2	4	2	2	4
-2	4	-2	2	-4



r is *exactly* zero even though the association is very strong.

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

r is strongly affected by outliers.

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

samples from independent RV's $\Rightarrow \ r \sim 0$

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

X,Y indep std normal RV's ; set $Y^* = \rho X + \sqrt{1-\rho^2} Y$ Then (X, Y^*) will tend to generate data with $r \sim \rho$. (e.g. $\rho = .99$ $=> \frac{\sqrt{(1-\rho^2)}}{\rho} = .14$!)

What does Best Fitting Mean?

Given any point (x_i, y_i)

What does Best Fitting Mean?

Given any point (x_i, y_i)
and any line $y = c_0 + c_1 x$

What does Best Fitting Mean?

Given any point (x_i, y_i)
and any line $y = c_0 + c_1 x$
we can use the line to get a predicted value

$$\hat{y}_i = c_0 + c_1 x_i$$

What does Best Fitting Mean?

Given any point (x_i, y_i)

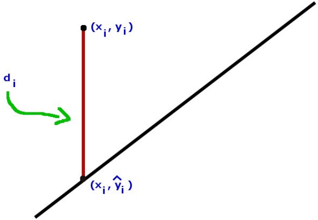
and any line $y = c_0 + c_1 x$

we can use the line to get a predicted value

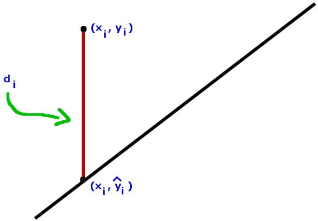
$$\hat{y}_i = c_0 + c_1 x_i$$

and define the residual d_i by

$$d_i = y_i - \hat{y}_i.$$

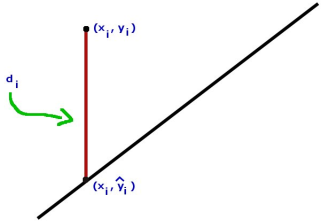


What does Best Fitting Mean?



Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

What does Best Fitting Mean?

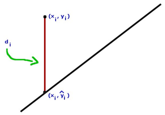


Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

Answer: We'll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

What does Best Fitting Mean?



Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

Answer: We'll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

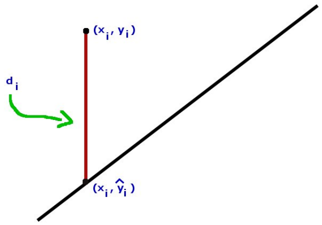
i.e.: *Slope is r in standard deviation units.*

And

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

i.e.: *The point $(\bar{x}, \bar{y})$ lies on the line.*

What does Best Fitting Mean?



Strictly speaking, the word residual refers to this vertical distance d_i JUST in the case that the line is the "answer line"

$$\hat{y} = b_0 + b_1 x.$$