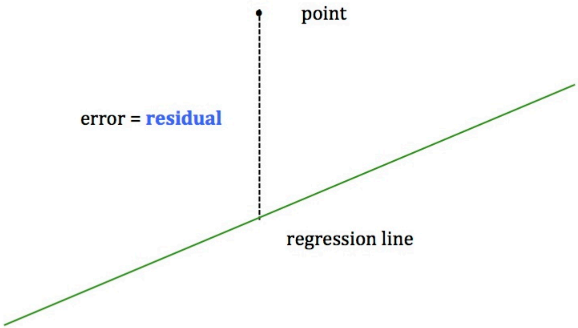
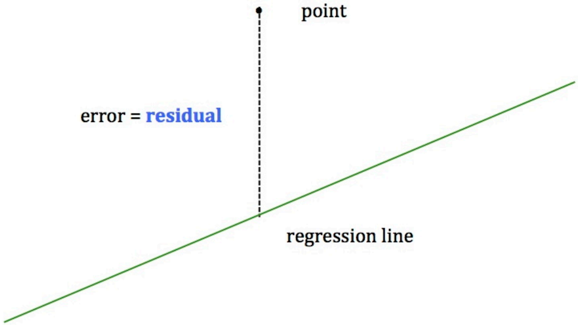
Residuals

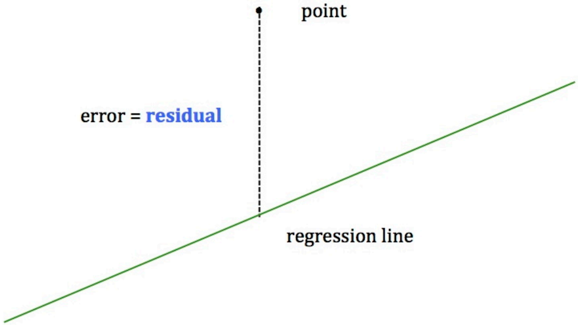


Residuals



r.m.s. error of regression = r.m.s. of residuals

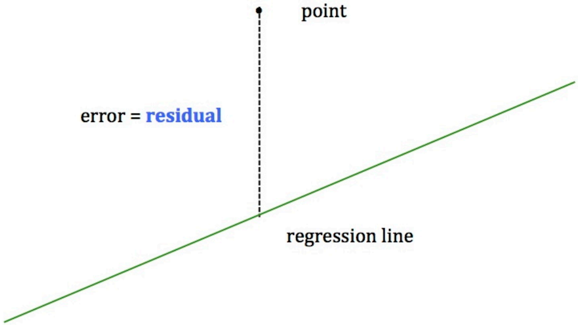
Residuals



r.m.s. error of regression = r.m.s. of residuals

$$= \sqrt{1 - r^2} \cdot \textbf{SD of } y$$

Residuals



r.m.s. error of regression = r.m.s. of residuals

$$= \sqrt{1 - r^2} \cdot \textbf{SD of } y$$

no matter what the shape of the scatter diagram

Comparing two estimates

Estimating chick weight

egg size:	average 23.12 mm	SD = 0.45 mm	
chick weight:	average 6.43 gm	SD = 0.4 gm	$r = 0.6$

Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm SD $= 0.45$ mm

chick weight: average 6.43 gm SD $= 0.4$ gm $r = 0.6$

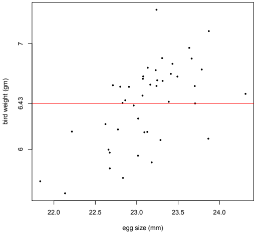
based on no information

Comparing two estimates

Estimating chick weight

egg size:	average 23.12 mm	SD = 0.45 mm	
chick weight:	average 6.43 gm	SD = 0.4 gm	$r = 0.6$

based on no information



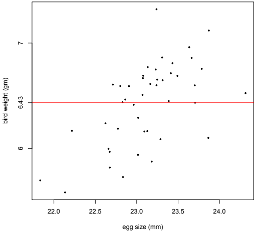
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm   SD $= 0.45$ mm

chick weight: average 6.43 gm   SD $= 0.4$ gm $r = 0.6$

based on no information

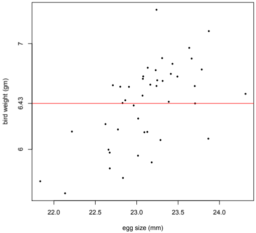


r.m.s. error

Estimating chick weight

egg size: average 23.12 mm SD $= 0.45$ mm
chick weight: average 6.43 gm SD $= 0.4$ gm $r = 0.6$

based on no information



r.m.s. error
$=$ r.m.s. of deviations

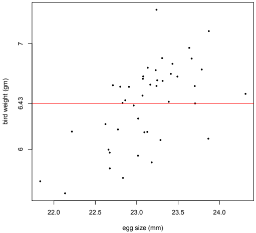
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm SD $= 0.45$ mm

chick weight: average 6.43 gm SD $= 0.4$ gm $r = 0.6$

based on no information



r.m.s. error

$=$ r.m.s. of deviations

$= 0.4$ gm

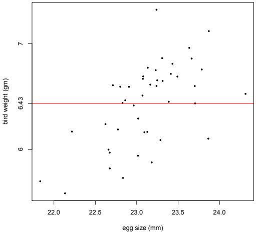
Estimating chick weight

egg size: average 23.12 mm SD $= 0.45$ mm

chick weight: average 6.43 gm SD $= 0.4$ gm $r = 0.6$

based on no information

based on egg size



r.m.s. error
$=$ r.m.s. of deviations
$= 0.4$ gm

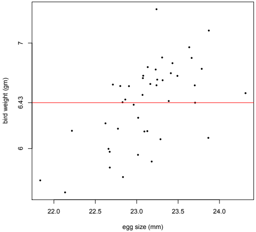
Comparing two estimates

Estimating chick weight

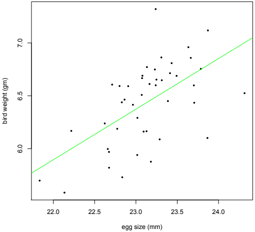
egg size: average 23.12 mm SD = 0.45 mm

chick weight: average 6.43 gm SD = 0.4 gm $r = 0.6$

based on no information



based on egg size



r.m.s. error

= r.m.s. of deviations

= 0.4 gm

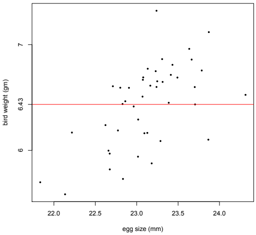
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm SD $= 0.45$ mm

chick weight: average 6.43 gm SD $= 0.4$ gm $r = 0.6$

based on no information

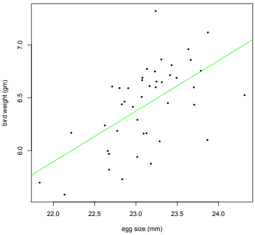


r.m.s. error

$=$ r.m.s. of deviations

$= 0.4$ gm

based on egg size



r.m.s. error

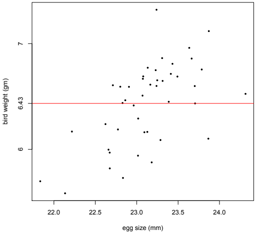
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm   SD = 0.45 mm

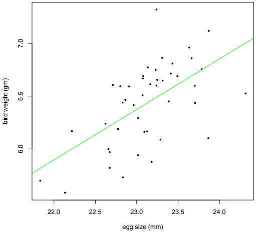
chick weight: average 6.43 gm   SD = 0.4 gm $r = 0.6$

based on no information



r.m.s. error
= r.m.s. of deviations
= 0.4 gm

based on egg size



r.m.s. error
= r.m.s. of residuals

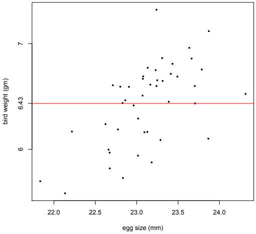
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm SD = 0.45 mm

chick weight: average 6.43 gm SD = 0.4 gm $r = 0.6$

based on no information

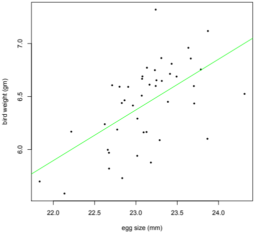


r.m.s. error

= r.m.s. of deviations

= 0.4 gm

based on egg size



r.m.s. error

= r.m.s. of residuals

$= \sqrt{1 - 0.6^2} \times 0.4$ gm

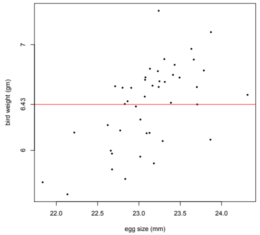
Comparing two estimates

Estimating chick weight

egg size: average 23.12 mm SD = 0.45 mm

chick weight: average 6.43 gm SD = 0.4 gm $r = 0.6$

based on no information

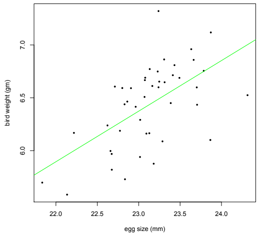


r.m.s. error

= r.m.s. of deviations

= 0.4 gm

based on egg size



r.m.s. error

= r.m.s. of residuals

$= \sqrt{1 - 0.6^2} \times 0.4 \text{ gm} = 0.32 \text{ gm}$

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot$ SD of y

- $r = 1$ or -1:

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

- $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error.

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot$ SD of y

- $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

• $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

• $r = 0$:

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

• $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

• $r = 0$: No linear association; regression is the same as using the average.

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

• $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

• $r = 0$: No linear association; regression is the same as using the average. Formula says r.m.s. error of regression $=$ SD of y when $r = 0$.

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot$ SD of y

- $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

- $r = 0$: No linear association; regression is the same as using the average. Formula says r.m.s. error of regression $=$ SD of y when $r = 0$.

- All other r:

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1-r^2} \cdot \text{SD of } y$

• $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

• $r = 0$: No linear association; regression is the same as using the average. Formula says r.m.s. error of regression $=$ SD of y when $r = 0$.

• All other r: Regression is not perfect, but better than using the average.

Formula gets the basics right

Rough size of error in regression

r.m.s. error of regression $= \sqrt{1 - r^2} \cdot \text{SD of } y$

• $r = 1$ or -1: Scatter is a perfect straight line; regression makes no error. Formula says r.m.s. error of regression $= 0$ when $r = 1$ or -1.

• $r = 0$: No linear association; regression is the same as using the average. Formula says r.m.s. error of regression $=$ SD of y when $r = 0$.

• All other r: Regression is not perfect, but better than using the average. Formula says r.m.s. error of regression $=$ fraction $\times$ SD of y

Some useful analogies

one variable **two variables**

Some useful analogies

one variable	**two variables**
• normal curve	

Some useful analogies

one variable	**two variables**
• normal curve	football shaped scatter diagram

Some useful analogies

one variable	**two variables**
• normal curve	football shaped scatter diagram
• average	

Some useful analogies

one variable	**two variables**
• normal curve	football shaped scatter diagram
• average	regression line

Some useful analogies

one variable	**two variables**
• normal curve	football shaped scatter diagram
• average	regression line
• SD	

Some useful analogies

one variable	**two variables**
• normal curve	football shaped scatter diagram
• average	regression line
• SD	r.m.s. error of regression

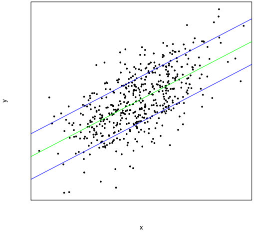
Using the analogies

Using the analogies

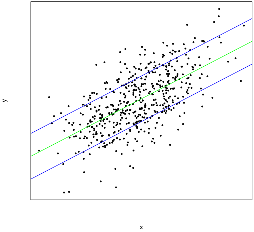
Football shaped scatter diagram

Using the analogies

Football shaped scatter diagram

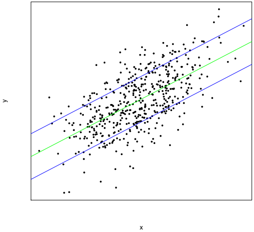


Football shaped scatter diagram



regression line $\pm$ 1 r.m.s. error
contains about 68% of the points

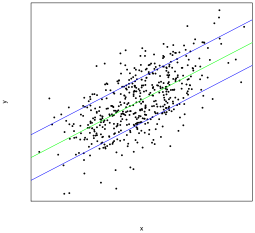
Football shaped scatter diagram

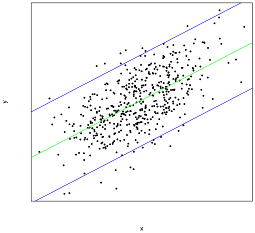


regression line $\pm$ 1 r.m.s. error contains about 68% of the points

For about 68% of the points, the regression estimate is correct to within 1 r.m.s. error.

Football shaped scatter diagram



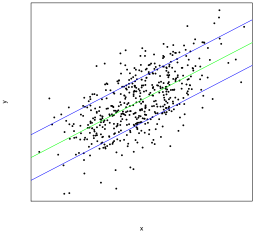


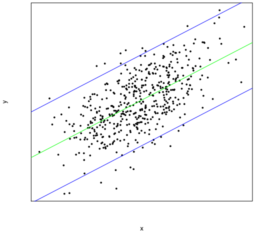
regression line $\pm$ 1 r.m.s. error contains about 68% of the points

For about 68% of the points, the regression estimate is correct to within 1 r.m.s. error.

Using the analogies

Football shaped scatter diagram



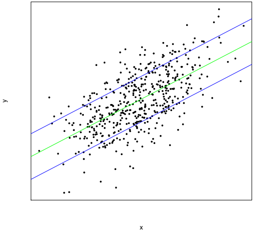


regression line $\pm$ 1 r.m.s. error contains about 68% of the points

regression line $\pm$ 2 r.m.s. errors contains about 95% of the points

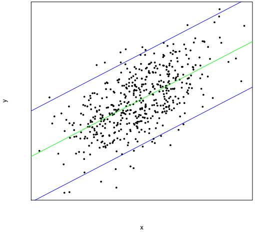
For about 68% of the points, the regression estimate is correct to within 1 r.m.s. error.

Football shaped scatter diagram



regression line $\pm$ 1 r.m.s. error contains about 68% of the points

For about 68% of the points, the regression estimate is correct to within 1 r.m.s. error.



regression line $\pm$ 2 r.m.s. errors contains about 95% of the points

For about 95% of the points, the regression estimate is correct to within 2 r.m.s. errors.

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\qquad r = 0.5$

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\quad r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated,

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\qquad r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds
- 40 pounds is how many r.m.s. errors?

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\qquad r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds
- 40 pounds is how many r.m.s. errors? $40/17.32 = 2.31$ r.m.s. errors

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds
- 40 pounds is how many r.m.s. errors? $40/17.32 = 2.31$ r.m.s. errors

Answer: Area between -2.31 and 2.31 under the standard normal curve

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\quad r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds
- 40 pounds is how many r.m.s. errors? $40/17.32 = 2.31$ r.m.s. errors

Answer: Area between -2.31 and 2.31 under the standard normal curve $= 97.91\%$ That is, **roughly 98%.**

A numerical example

Heights: average 67 inches, SD 3 inches
Weights: average 160 pounds, SD 20 pounds
scatter diagram is roughly football shaped $\qquad r = 0.5$

For what percent of the people, roughly, is the regression estimate of weight based on height correct to within 40 pounds?

y: variable being estimated, so y is weight

- r.m.s. error of regression $= \sqrt{1 - 0.5^2} \cdot 20$ pounds $= 17.32$ pounds
- 40 pounds is how many r.m.s. errors? $40/17.32 = 2.31$ r.m.s. errors

Answer: Area between -2.31 and 2.31 under the standard normal curve $= 97.91\%$ That is, **roughly 98%.**

Same statement: For **roughly 2%** of the people, the regression estimate of weight based on height is **off by more than 40 pounds.**

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

- **The average of these weights is roughly the regression estimate:**

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

- **The average of these weights is roughly the regression estimate:**
$(72 - 67)/3 = 1.67$; $0.5 \times 1.67 = 0.835$; $0.835 \times 20 + 160 =$ **176.7 lb**

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

- **The average of these weights is roughly the regression estimate:**
$(72 - 67)/3 = 1.67$; $0.5 \times 1.67 = 0.835$; $0.835 \times 20 + 160 =$ **176.7 lb**

- **Roughly what is the SD of these weights?**

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

• **The average of these weights is roughly the regression estimate:**
$(72 - 67)/3 = 1.67$; $0.5 \times 1.67 = 0.835$; $0.835 \times 20 + 160 =$ **176.7 lb**

• **Roughly what is the SD of these weights?**
In a football shaped scatter diagram, the vertical spread is about the same throughout.

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

• **The average of these weights is roughly the regression estimate:**
$(72 - 67)/3 = 1.67$; $0.5 \times 1.67 = 0.835$; $0.835 \times 20 + 160 =$ **176.7 lb**

• **Roughly what is the SD of these weights?**
In a football shaped scatter diagram, the vertical spread is about the same throughout. **So the r.m.s. error of regression is the rough size of the deviations from average in any vertical strip.**

The average and SD of a vertical strip

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb   r.m.s. error = 17.32 lb

Vertical strip: weights of people who are 72 inches tall

• **The average of these weights is roughly the regression estimate:**
$(72 - 67)/3 = 1.67$; $0.5 \times 1.67 = 0.835$; $0.835 \times 20 + 160 =$ **176.7 lb**

• **Roughly what is the SD of these weights?**
In a football shaped scatter diagram, the vertical spread is about the same throughout. **So the r.m.s. error of regression is the rough size of the deviations from average in any vertical strip.**

The SD of the weights of the people who are 72 inches tall is about 17.32 lb.

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$

Heights: average 67 in, SD 3 in

Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb
- distribution is roughly normal: consequence of bivariate normal scatter

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb
- distribution is roughly normal: consequence of bivariate normal scatter

Of the people who are 72 inches tall, about what percent weigh more than 200 lb?

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb
- distribution is roughly normal: consequence of bivariate normal scatter

Of the people who are 72 inches tall, about what percent weigh more than 200 lb?

$z = (200 - 176.7)/17.32 = 1.34$

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb
- distribution is roughly normal: consequence of bivariate normal scatter

Of the people who are 72 inches tall, about what percent weigh more than 200 lb?

$z = (200 - 176.7)/17.32 = 1.34$

Answer: All the area to the right of 1.34 on the standard normal curve

Vertical strips and the normal curve

Estimating weight based on height: football shaped scatter; $r = 0.5$
Heights: average 67 in, SD 3 in
Weights: average 160 lb, SD 20 lb

Weights of people who are 72 inches tall

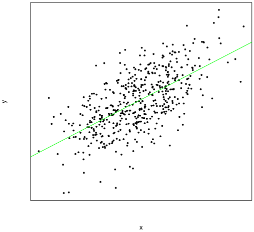
- average is roughly the regression estimate: 176.7 lb
- SD is roughly the r.m.s. error: 17.32 lb
- distribution is roughly normal: consequence of bivariate normal scatter

Of the people who are 72 inches tall, about what percent weigh more than 200 lb?
$z = (200 - 176.7)/17.32 = 1.34$

Answer: All the area to the right of 1.34 on the standard normal curve
$=$ **9%**, roughly

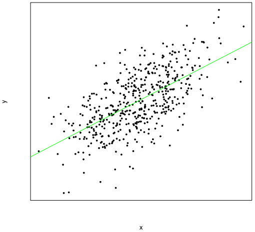
Plotting the residuals

Scatter and regression line

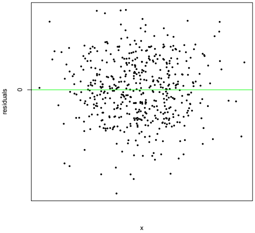


Plotting the residuals

Scatter and regression line

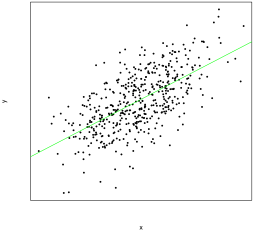


Residual plot

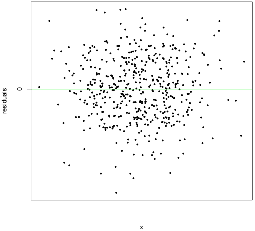


Plotting the residuals

Scatter and regression line

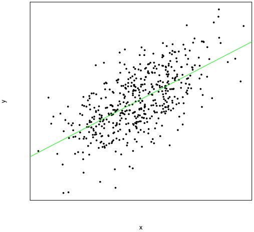


Residual plot

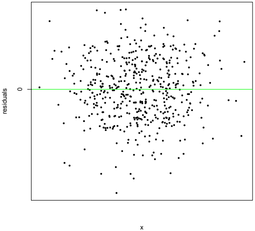


Math facts: Regardless of the shape of the scatter diagram,

Scatter and regression line



Residual plot

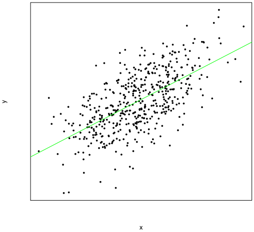


Math facts: Regardless of the shape of the scatter diagram,

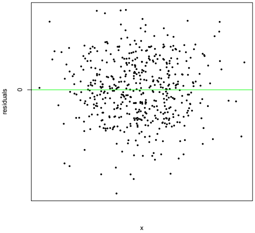
- the average of the residuals is always 0;

Plotting the residuals

Scatter and regression line



Residual plot

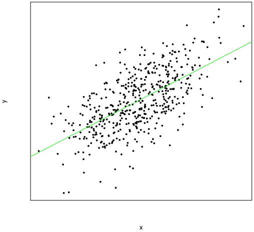


Math facts: Regardless of the shape of the scatter diagram,

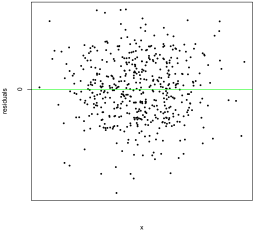
- the average of the residuals is always 0;
- there is no linear association between the residuals and x.

Plotting the residuals

Scatter and regression line



Residual plot

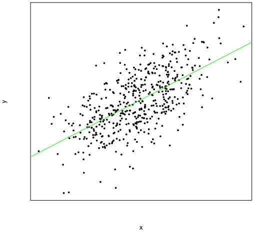


Math facts: Regardless of the shape of the scatter diagram,

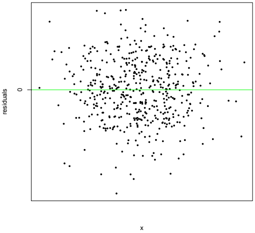
- the average of the residuals is always 0;
- there is no linear association between the residuals and x.

The residual plot cannot show any trend or linear relation.

Scatter and regression line



Residual plot



Math facts: Regardless of the shape of the scatter diagram,

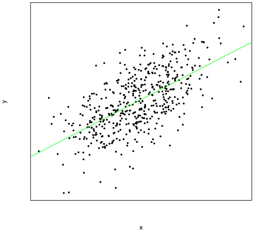
- the average of the residuals is always 0;
- there is no linear association between the residuals and x.

The residual plot cannot show any trend or linear relation.

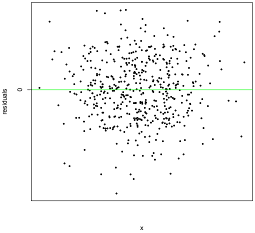
Good regression:

Plotting the residuals

Scatter and regression line



Residual plot



Math facts: Regardless of the shape of the scatter diagram,

- the average of the residuals is always 0;
- there is no linear association between the residuals and x.

The residual plot cannot show any trend or linear relation.

Good regression:

Residual plot looks like a formless blob around the horizontal axis.

The residual plot as a diagnostic tool

The residual plot as a diagnostic tool

average weight (in pounds) of women of different heights (in inches)

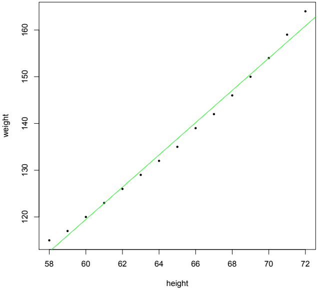
The residual plot as a diagnostic tool

average weight (in pounds) of women of different heights (in inches)
$r = 0.995$

The residual plot as a diagnostic tool

average weight (in pounds) of women of different heights (in inches)

$r = 0.995$



The residual plot as a diagnostic tool

average weight (in pounds) of women of different heights (in inches)

$r = 0.995$