# Correlation and Linear Model

Modelling Data | Linear Model

STAT5002

*The University of Sydney*

Mar 2025

THE UNIVERSITY OF SYDNEY

CRICOS 00026A TEQSA PRV12057

# Data Modelling

## Topic 3: Normal Curve

What is the Normal Curve? And what does it have to do with sample mean?

## Topic 4: Linear Model

How can we describe the relationship between two variables? When is a linear model appropriate?

# Outline

## Correlation

- Bivariate data & scatter plot
- Correlation coefficient
- Properties and warnings

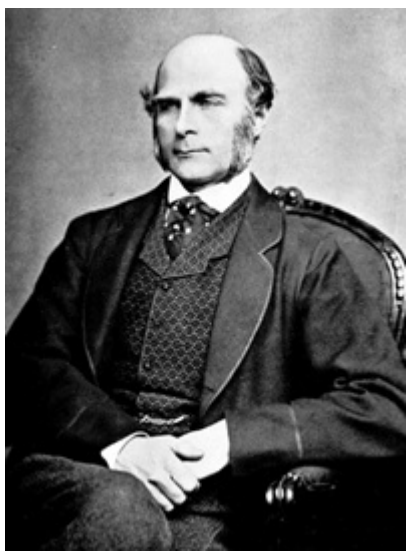
## Linear model

- Regression Line
- Prediction
- Residuals and properties
- Coefficient of determination
- Diagnostics of model fit

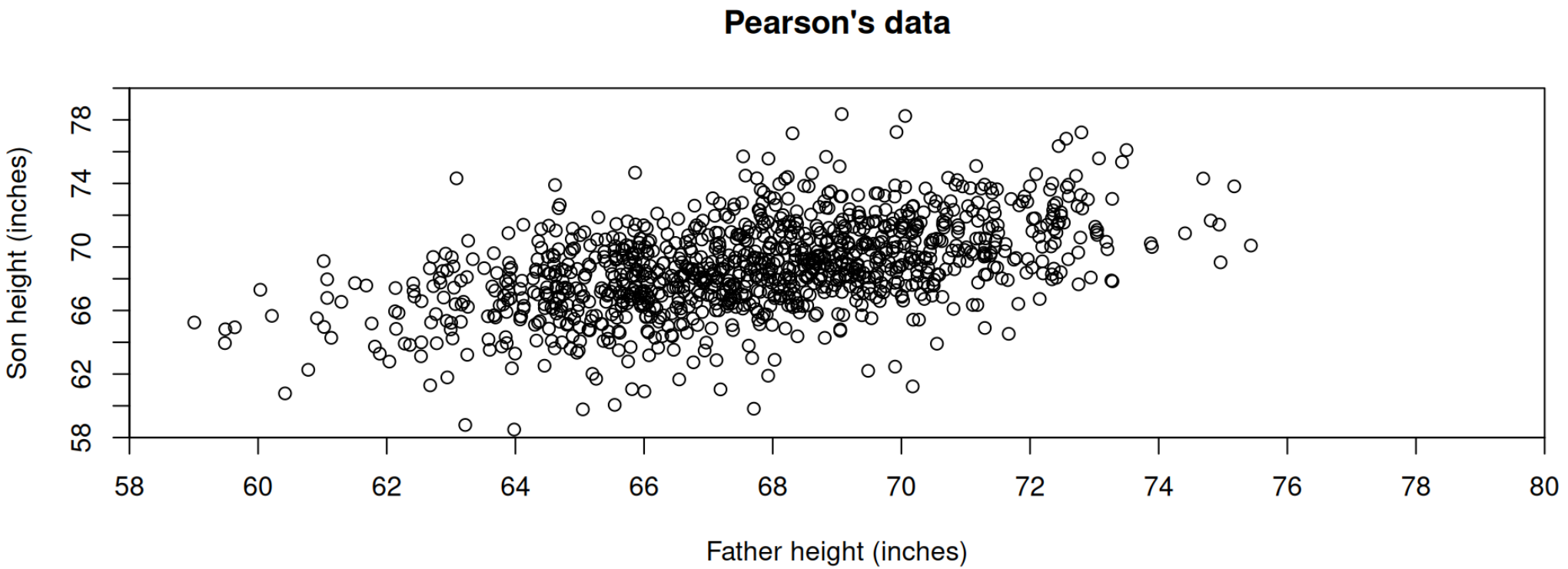
# Scatter plots

# History

- Sir Francis Galton (England, 1822–1911) studied the degree to which children resemble their parents (and wrote travel books on “wild countries”!)
- Galton’s work was continued by his student Karl Pearson (England, 1857–1936). Pearson measured the heights of 1,078 fathers and their sons at maturity.



# Pearson’s plot of heights (scatter plot)




- Plotting the pairs of heights creates a cloud of points.
- Generally, taller fathers tend to have taller sons.

# Statistical Thinking

Why do we care whether there is an association between two variables (here: height of father and son)?

- The association is interesting on its own.
- Association between two variables can be used for prediction, i.e, use outcome in one variable to predict the outcome in another variable.
- How can we quantify a possible association?

# Correlation coefficient

# Bivariate data

Bivariate data involves a **pair** of variables. We are interested in the relationship between the two variables. Can one variable be used to predict the other?

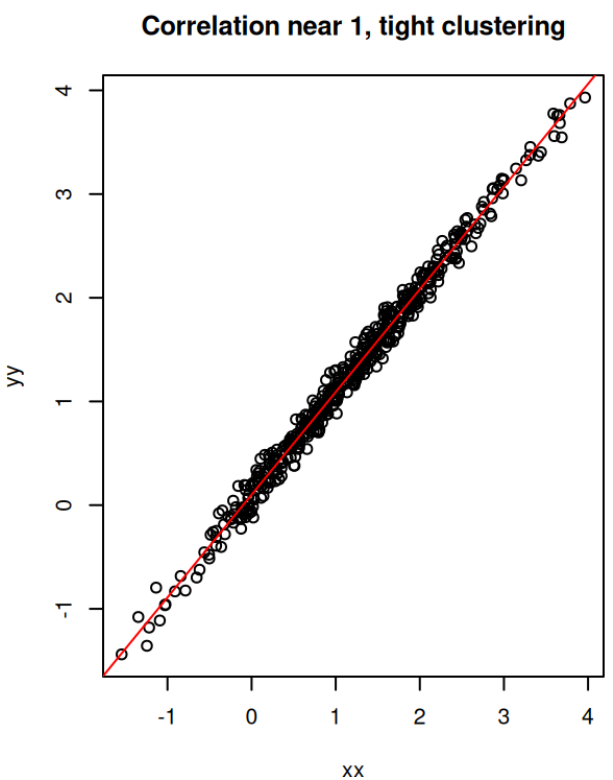
- Formally, we have $(x_i, y_i)$ for $i = 1, 2, \ldots, n$.
- $X$ and $Y$ can have the same role
- $X$ and $Y$ may have different roles: for example, $X$ can be an **independent** variable (or explanatory variable, predictor or regressor) which we use to explain or predict $Y$, the **dependent** variable (or response variable).

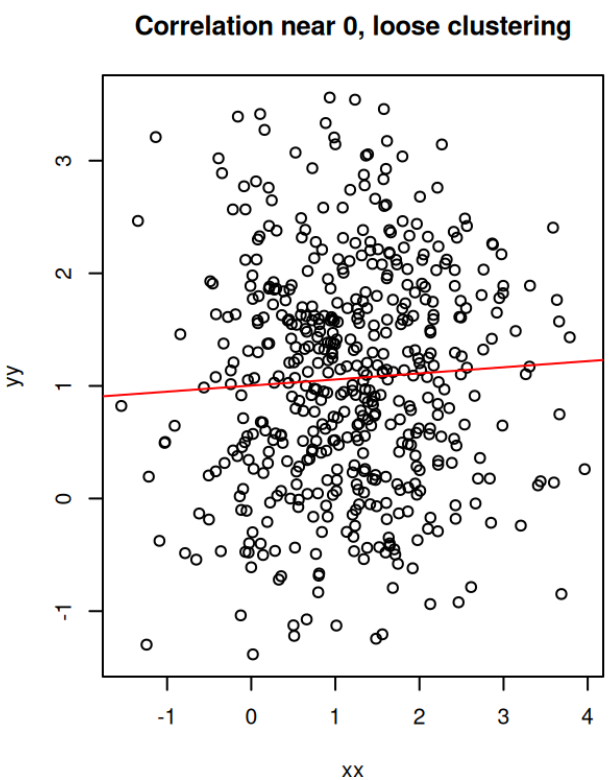
# How can we summarise bivariate data?

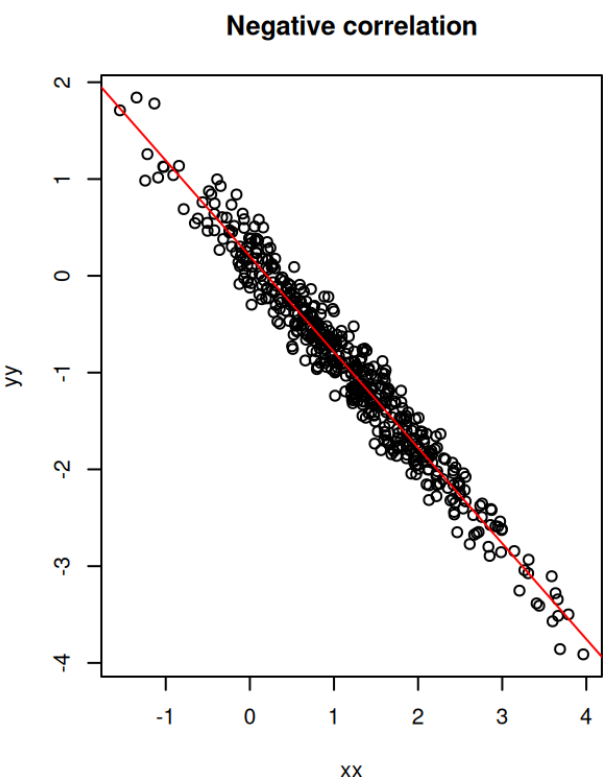
Bivariate data (or a scatter plot) can be summarised by the following **five** numerical summaries:

- Sample mean and sample SD of $X$ ($\bar{x}$, $\mathrm{SD}_x$)
- Sample mean and sample SD of $Y$ ($\bar{y}$, $\mathrm{SD}_y$)
- Correlation coefficient ($r$).

# Association between the two variables




- All clouds have the **same centre and horizontal and vertical spread**.
- However they have **different spread** around a line (linear association). How do we measure this?

# The correlation coefficient

## The (Pearson) correlation coefficient (r)

- A numerical summary measures of how points are spread around the line.
- It indicates both the sign and strength of the **linear association**.
- It is defined as the mean of the product of the variables in **standard units**.

Recall that

$$\text{standard unit} = \frac{\text{data point} - \text{mean}}{SD}$$

Using sample SD, we divide by $n-1$ in the average:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{SD_{sample}(X)} \frac{(y_i-\bar{y})}{SD_{sample}(Y)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

which simplifies to $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

# Obtaining r using the population SD

The same correlation coefficient $r$ can be obtained using the population SD as well (dividing by $n$ in the average).

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{SD_{pop}(X)}\frac{(y_i-\bar{y})}{SD_{pop}(Y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

which also simplifies to $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

**Quick calculation in R** using `cor()`.

```
1 cor(x, y)
```

```
[1] 0.5013383
```

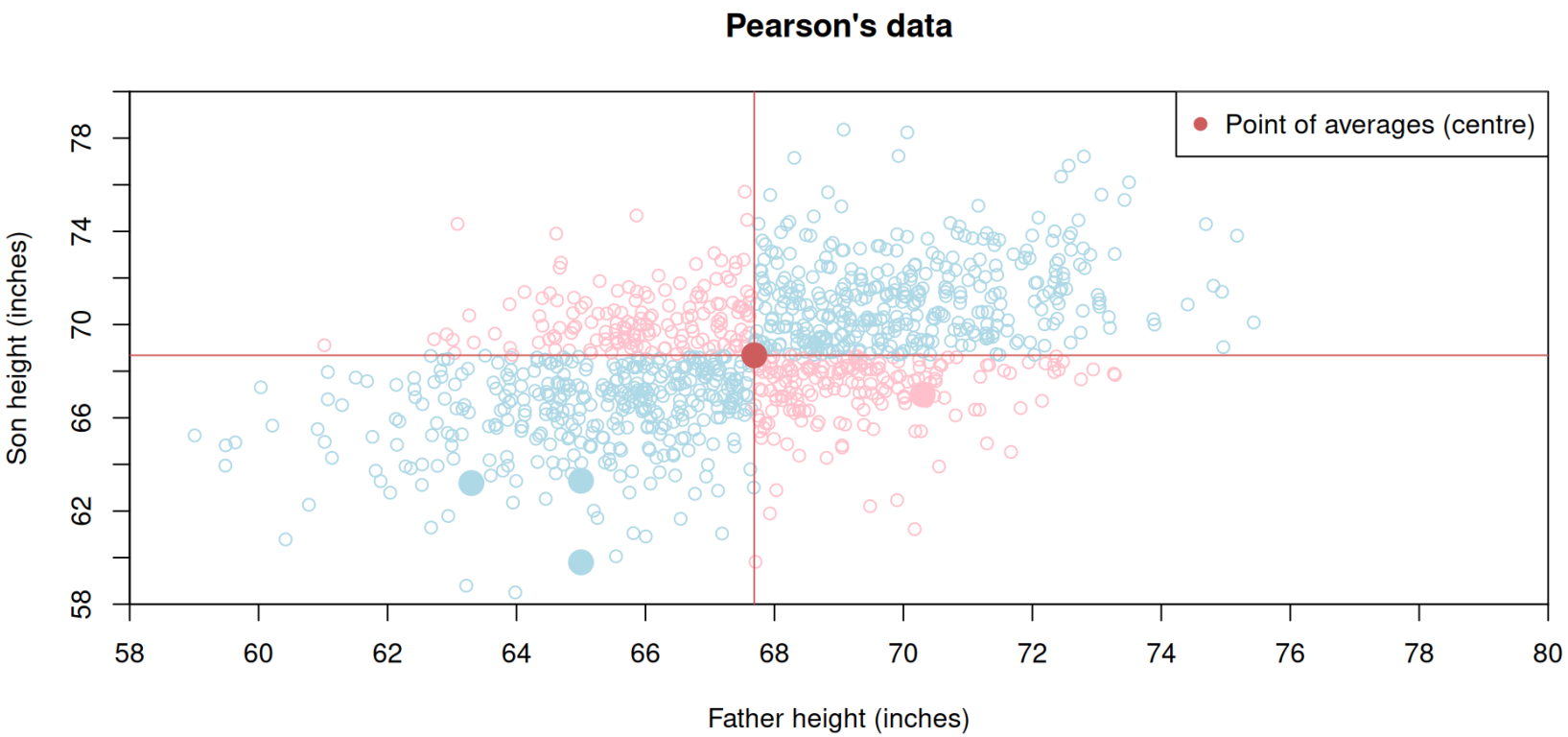
# How does $r$ measure association?

Here, for illustration, we round data to 1 decimal place to make calculations simpler.

| $x$ (father's heights) | $y$ (son's heights) | standard units | standard units | product | quadrant |
|---|---|---|---|---|---|
| | | $\frac{x-67.7}{2.7}$ | $\frac{y-68.7}{2.8}$ | $\left(\frac{x-67.7}{2.7}\right)\left(\frac{y-68.7}{2.8}\right)$ | |
| 65.0 | 59.8 | -0.96 | -3.16 | 3.04 | lower left |
| 63.3 | 63.2 | -1.62 | -1.94 | 3.14 | lower left |
| 65.0 | 63.3 | -1.00 | -1.90 | 1.89 | lower left |
| 70.3 | 67.0 | 0.95 | -0.59 | -0.57 | lower right |
| ⋮ | | | | | |
| | | | | mean=+0.5 | |

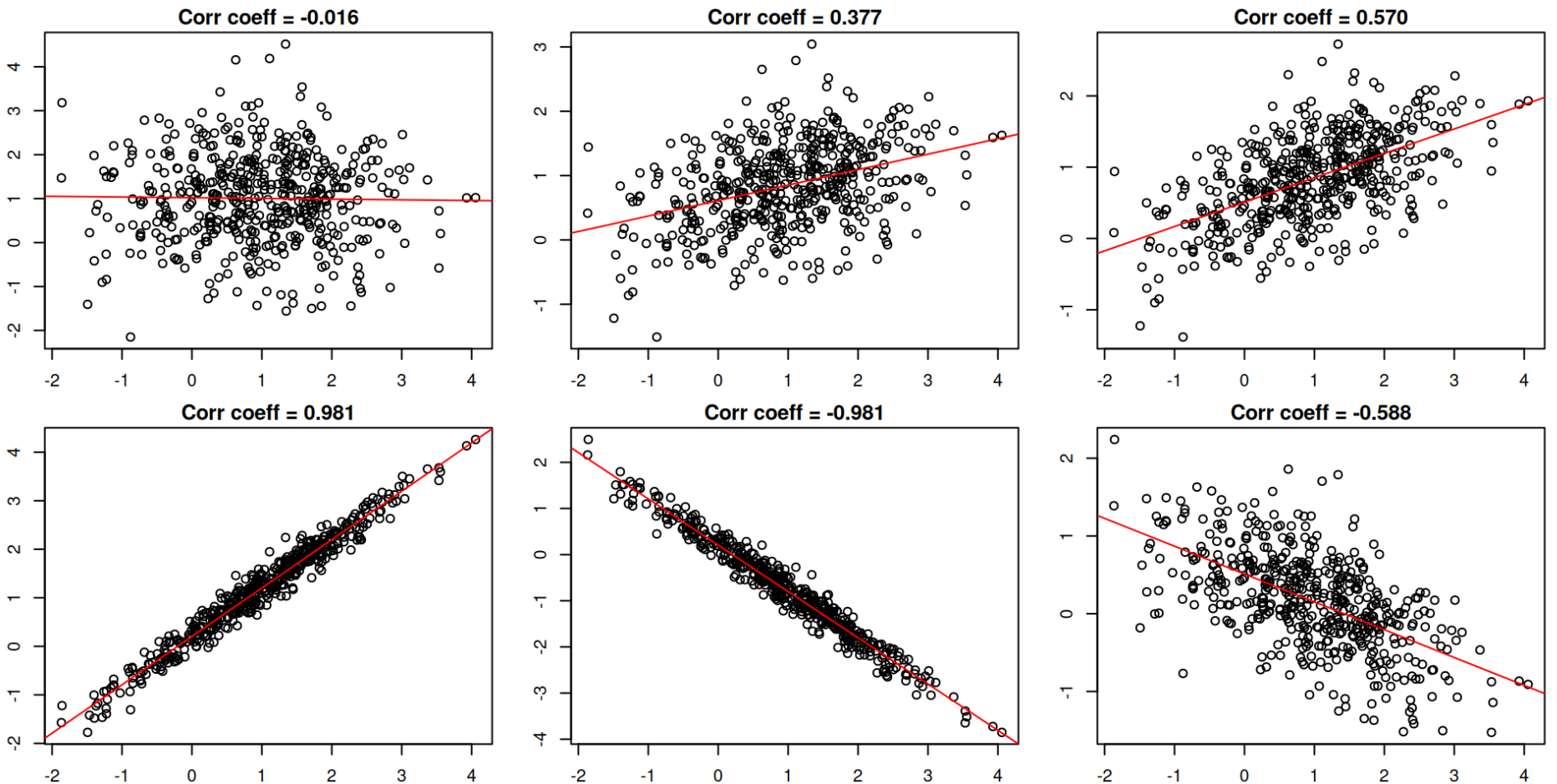
We divide the scatter plot into 4 quadrants, at the point of averages (centre).

- In the upper right and lower left quadrants, products of standard units are (+)
- In the upper left and lower right quadrants, products of standard units are (-)




- A majority of points in the upper right (+) and lower left quadrants (+) will be indicated by a positive $r$
- A majority of points in the upper left (-) and lower right quadrants (-) will be indicated by a negative $r$

# More examples



# Properties and warnings

# Interpretations of r values

- The correlation coefficient $r$ always takes values between -1 and 1 (inclusive).
  - ➠ This can be shown using the definition of $r$ and the Cauchy-Schwarz inequality (only for your information).
- If $r$ is positive: the cloud slopes up.
- If $r$ is negative: the cloud slopes down.
- $r = 0$ implies no linear dependency between two variables.
- As $r$ gets closer to $\pm 1$: the points cluster more tightly around the line.

# Invariant properties

## Shift and scale invariant

The correlation coefficient is shift and scale invariant. Why? **Shifting and scaling do not change the standard unit.**

```
1 cor(x, y)
```

```
[1] 0.5013383
```

```
1 cor(0.2 * x + 3, 3 * y - 1)
```

```
[1] 0.5013383
```

## Symmetry (commutative)

The correlation coefficient is not affected by interchanging the variables.

```
1 cor(x, y)
```

```
[1] 0.5013383
```

```
1 cor(y, x)
```

```
[1] 0.5013383
```

# Warning 1:

## Wrong interpretations of correlation coefficient

### Mistake:

$r = 0.8$ means that 80% of the points are tightly closed around the line.

### Mistake:

$r = 0.8$ means that the points are twice as tightly closed as $r = 0.4$.

> **Note**
>
> $r = 0.8$ suggests a stronger association between variables compared to the case $r = 0.4$ BUT does not suggest the data points are twice as tight.

# Warning 2:

## Outliers can overly influence the correlation coefficient

Suppose there was an extra unusual reading of (100,50).

```
f1 = c(data$fheight, 100)  # Add an extra point to data
s1 = c(data$sheight, 50)
```

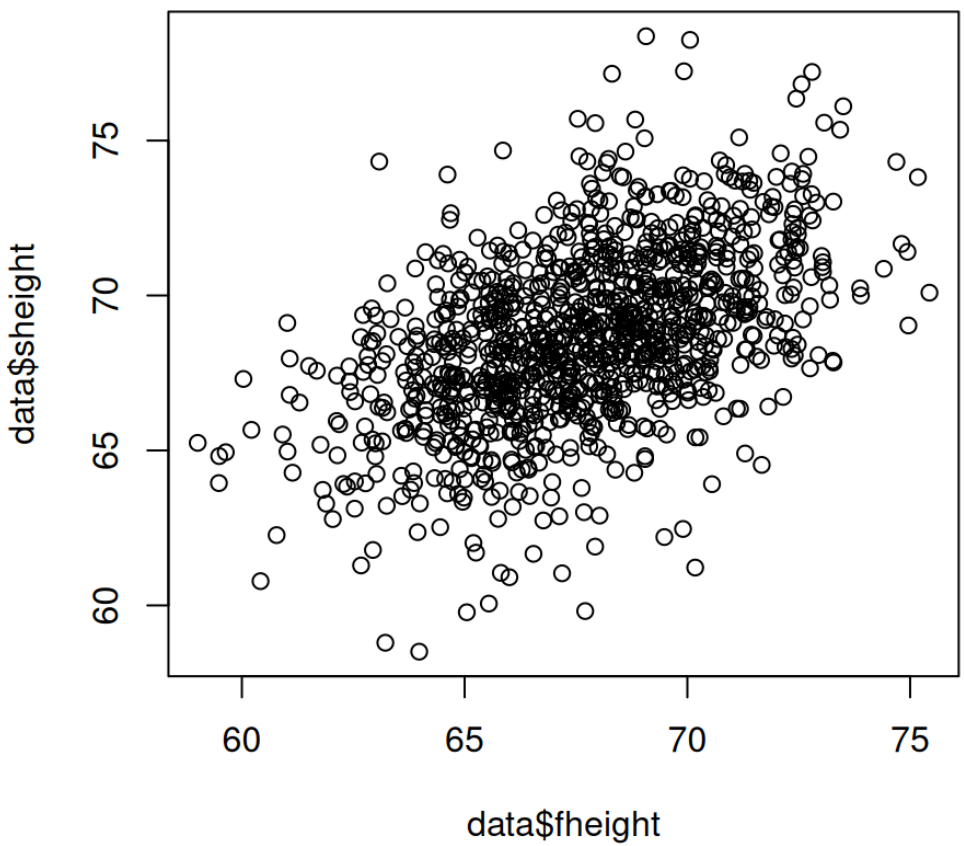
```
cor(data$fheight, data$sheight)
```

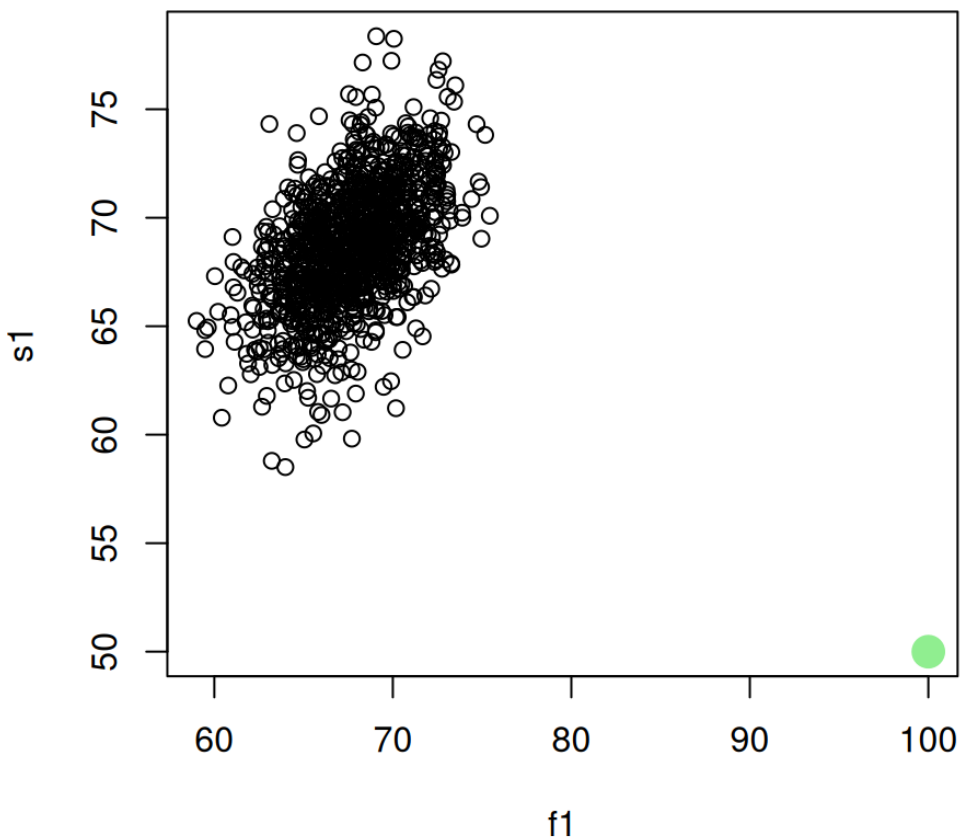
```
[1] 0.5013383
```

```
cor(f1, s1)
```

```
[1] 0.3956794
```

```r
par(mfrow = c(1, 2))
plot(data$fheight, data$sheight)
plot(f1, s1)
points(100, 50, col = "lightgreen", pch = 19, cex = 2)
```



# Warning 3:

## Nonlinear association can’t be detected by the correlation coefficient

What interpretation mistake could be made in the following data set?

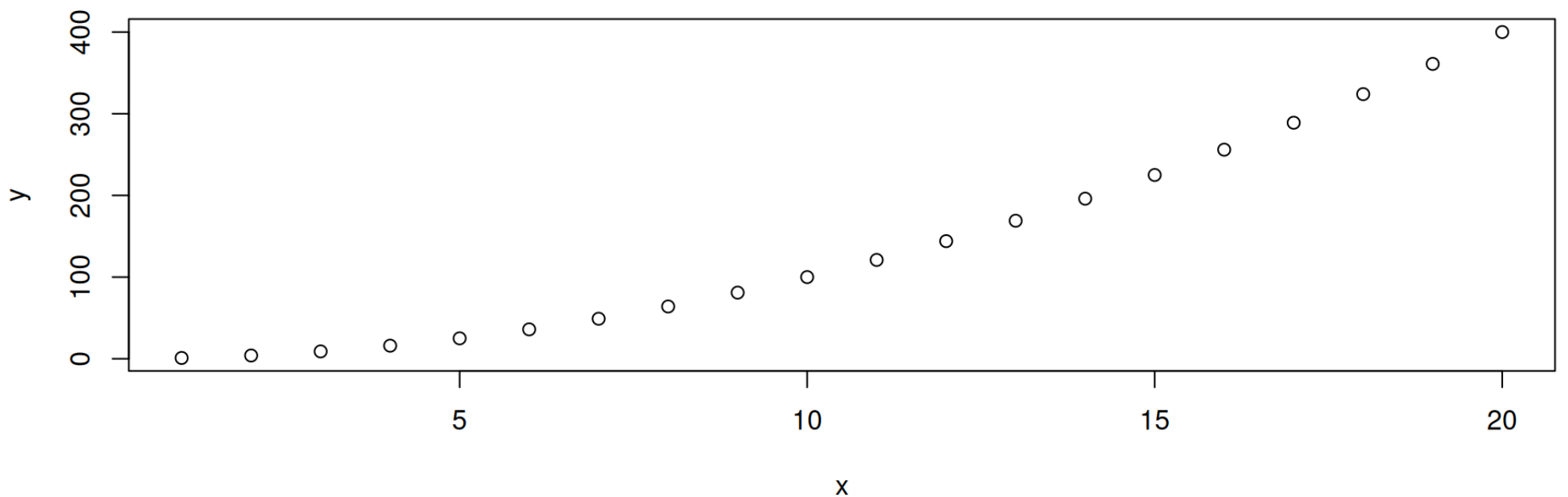
```
x = c(1:20)
y = x^2
cor(x, y)
```

```
[1] 0.9713482
```

Based on the correlation coefficient, the points should cluster very tightly around the line sloping up.

## But look at the scatter plots.

```
1 plot(x, y)
```




This data should be modelled by a quadratic curve, not a line.

**We should always use correlation coefficient together with the scatter plot.**

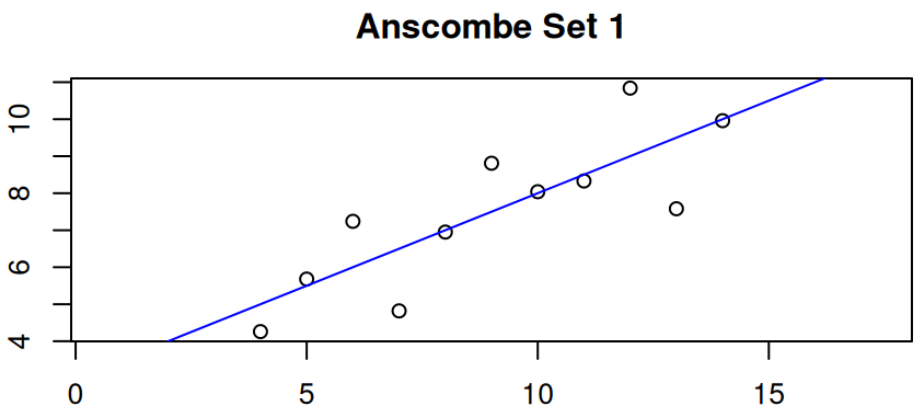
# Warning 4:

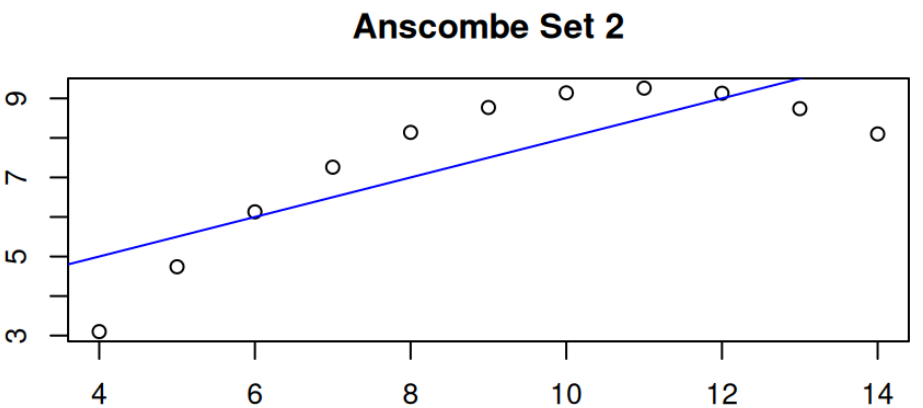
## The same correlation coefficient can arise from very different data

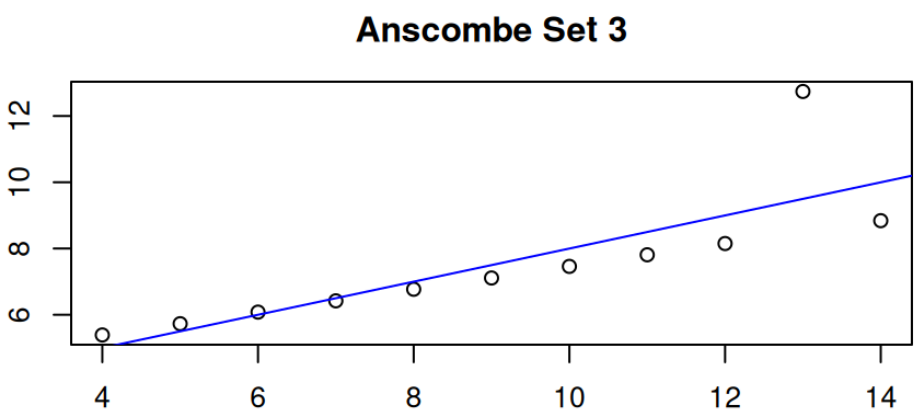
The following data sets (Anscombes Quartet) have the **same** $\bar{x}, SD_x, \bar{y}, SD_y$, and also the **same** value of $r$.

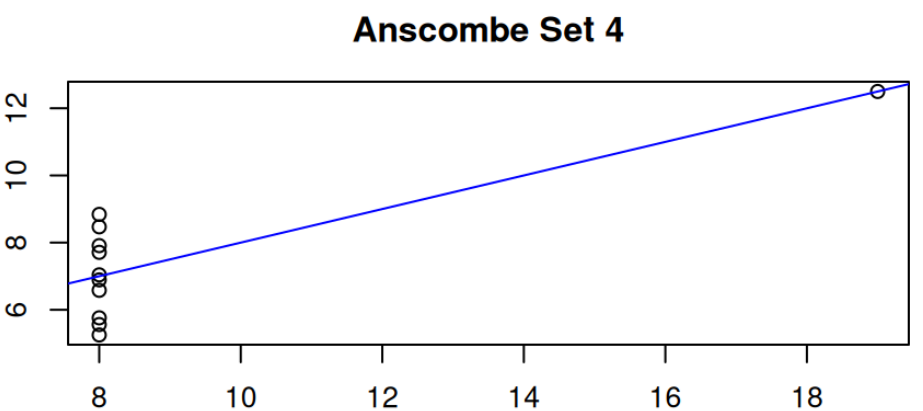
```
x_mean: 9 9 9 9
x_sd: 3.316625 3.316625 3.316625 3.316625
y_mean: 7.500909 7.500909 7.5 7.500909
y_sd: 2.031568 2.031657 2.030424 2.030579
r: 0.8164205 0.8162365 0.8162867 0.8165214
```

But look at the scatter plots.



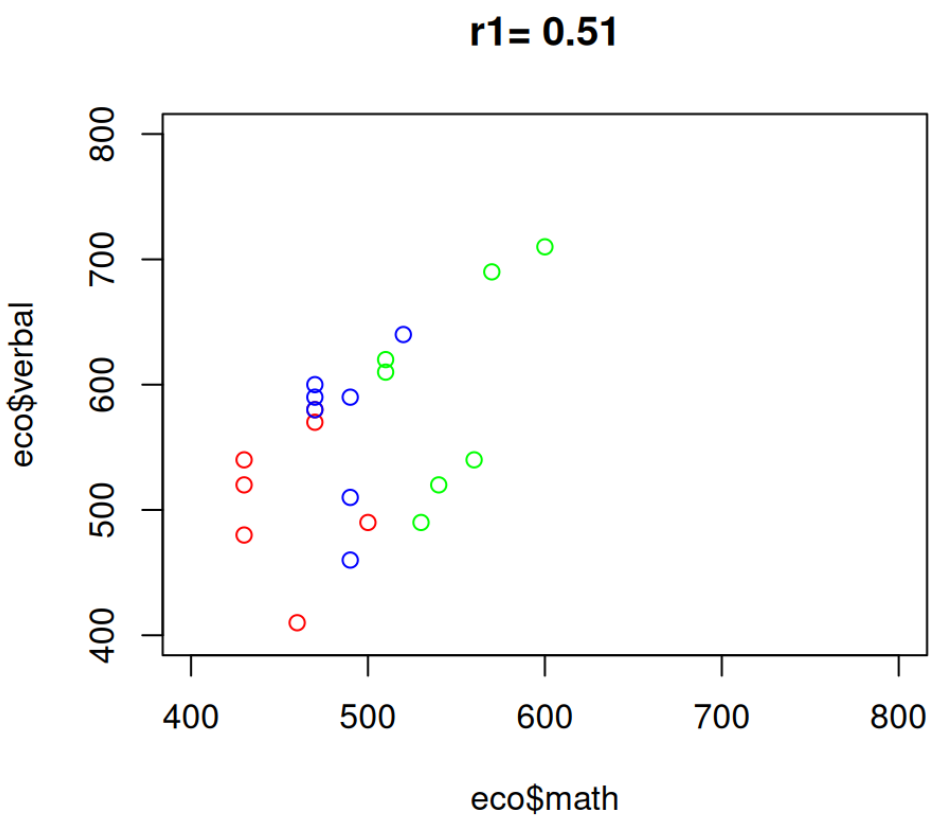



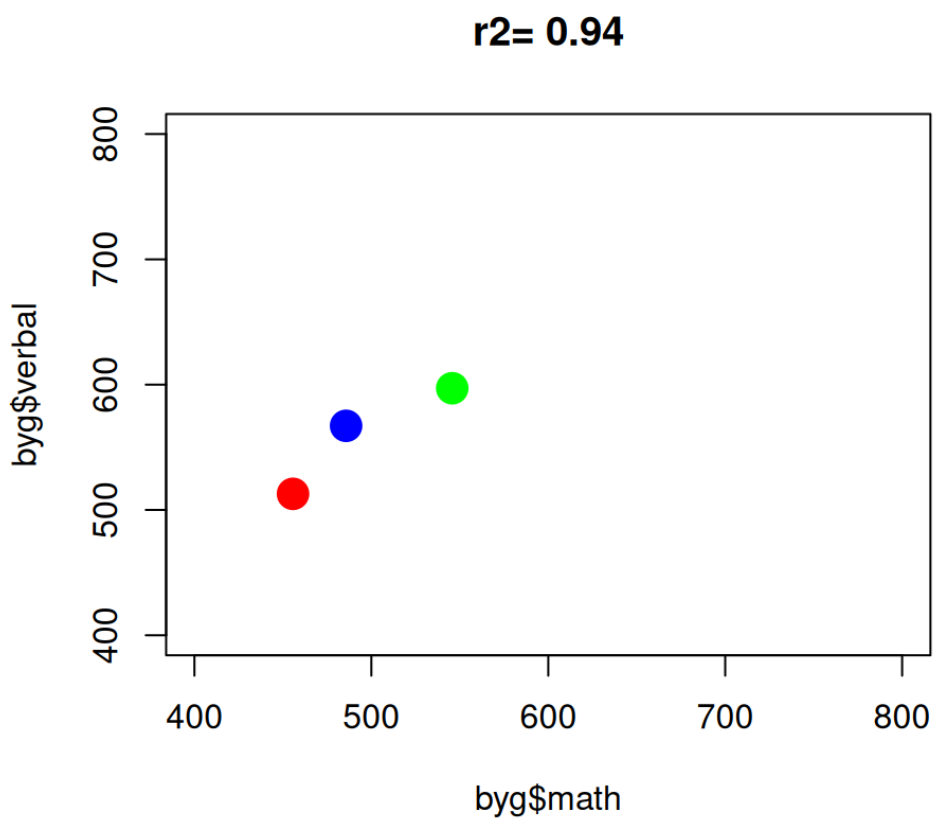
# Warning 5 (not for assessment):

## Ecological correlation tends to inflate the correlation coefficient

- An **ecological correlation** is the correlation between two variables that are group means.
- For example, if we recorded the heights of fathers and sons in many communities, and then calculated the average for each community.
- Correlations at the group level (ecological correlations) can be much higher than those at the individual level.
- See Freedman et al, Statistics p148-149.

# Example




- The 1st plot has all 3 sets of data combined: correlation = 0.51 (not very strong).
- The 2nd plot has the averages of the 3 data sets: correlation = 0.94 (very strong).

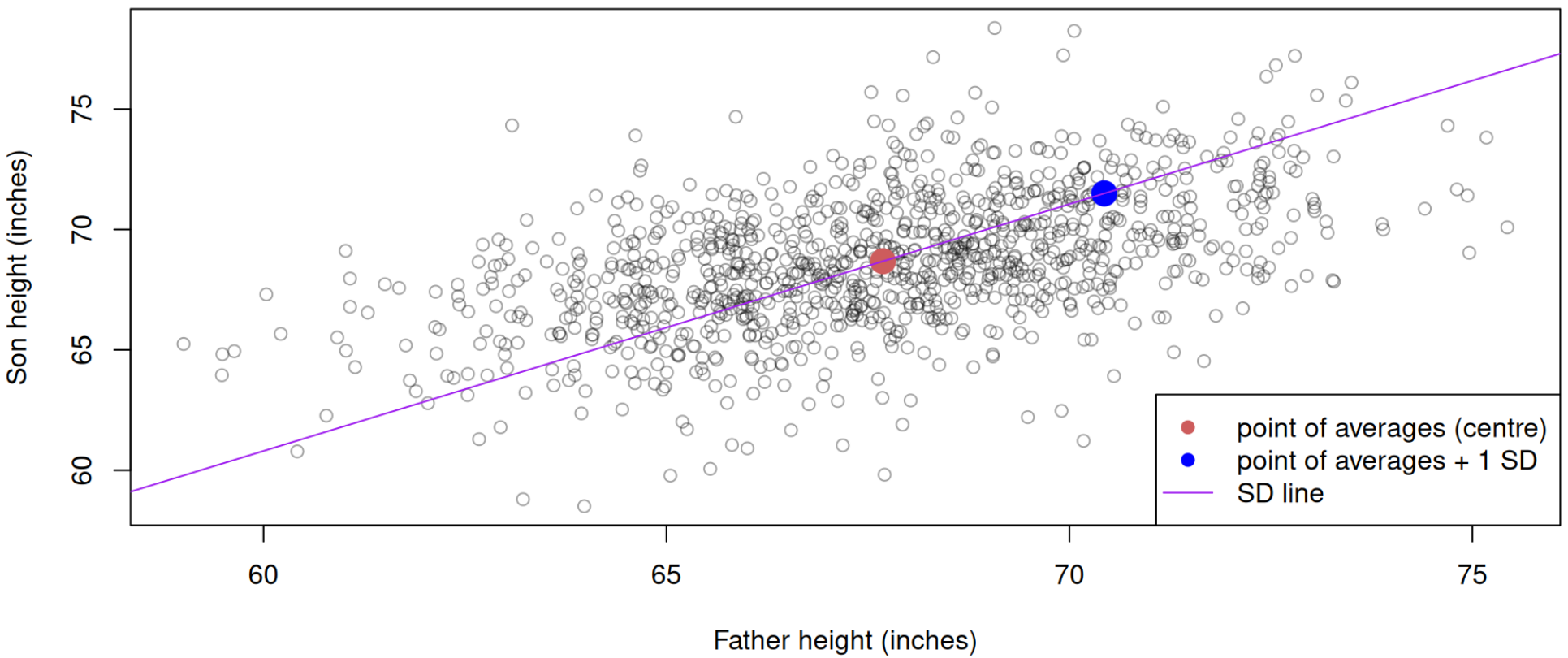
# Regression line

# Pearson’s plot of heights

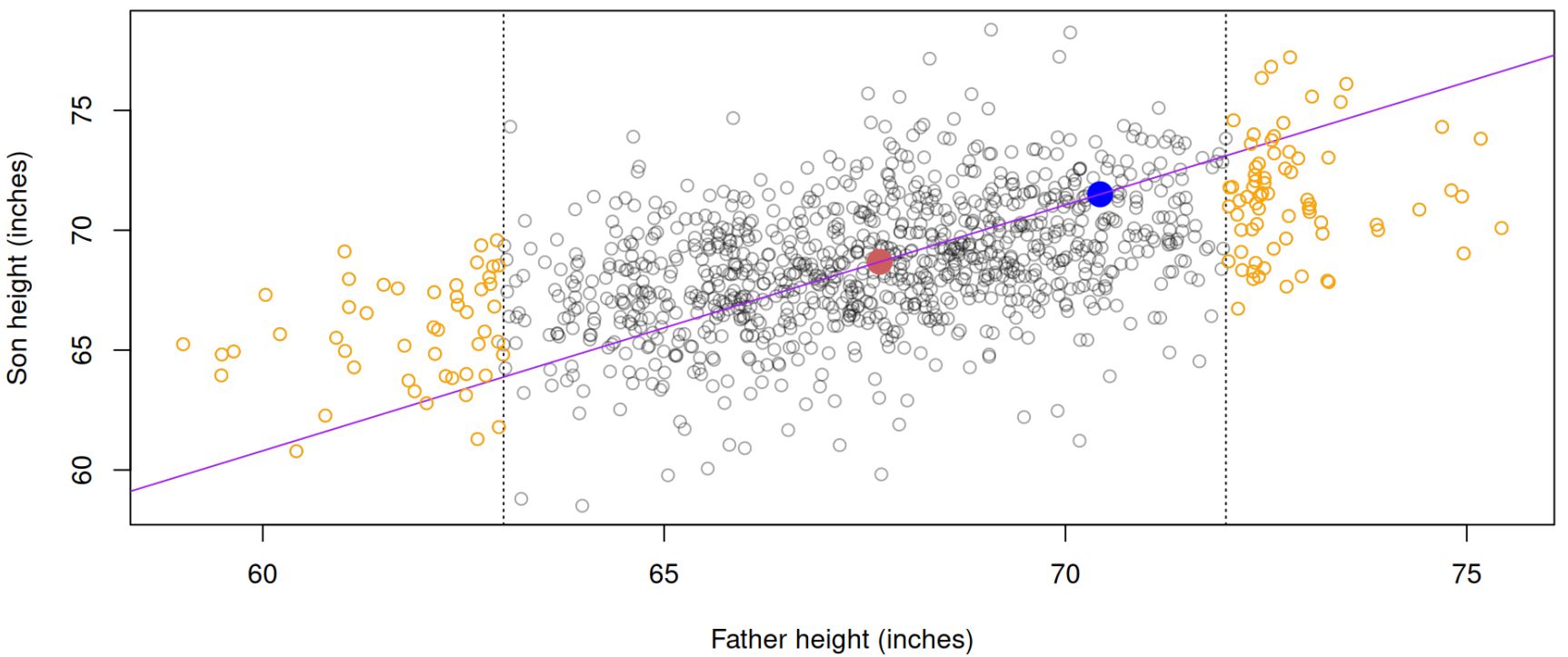

- How can we summarise the data with a line?
- How do we find the **optimal** line?

# 1st option: SD line (not so good)

- The **SD line** might look like a good candidate as it connects the point of averages $(\bar{x}, \bar{y})$ to $(\bar{x} + \mathrm{SD}_x, \bar{y} + \mathrm{SD}_y)$ (for this data with positive correlation).



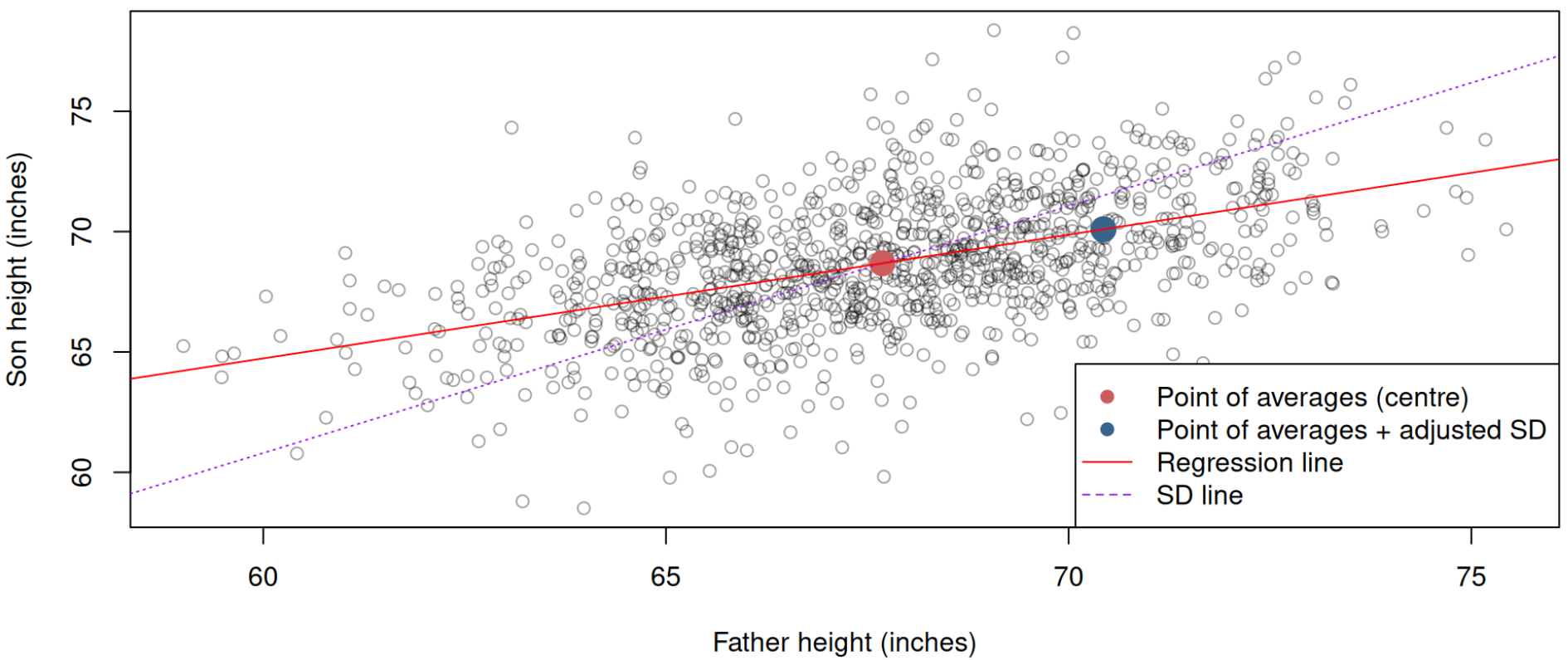
Note how it underestimates (LHS) and overestimates (RHS) at the extremes.



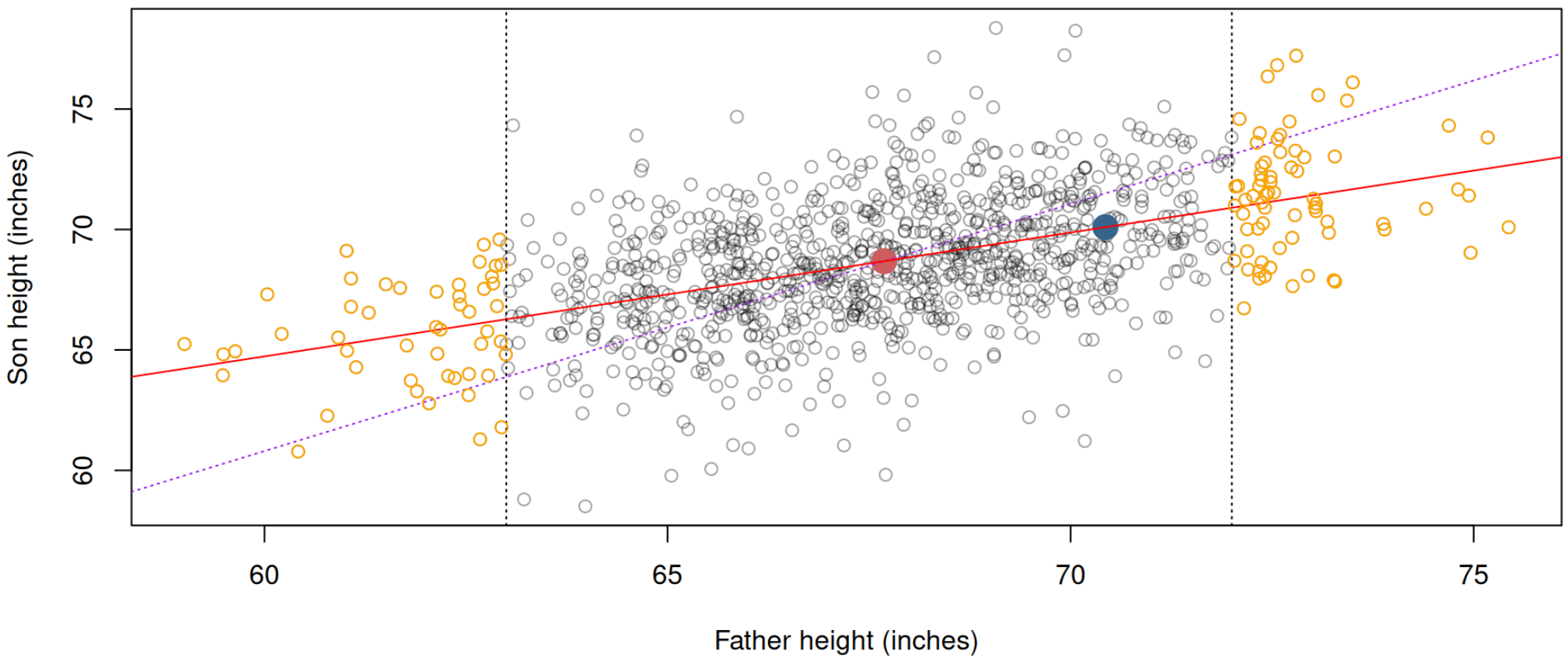

- Recall that $X, Y$ can have the same mean and SD but very different correlation coefficient.
- The above model does not use the correlation coefficient, so it is insensitive to the amount of clustering around the line.
- How to quantify the quality of the fitted line so we can define the **optimal** line?

# Best option: regression line

- To describe the scatter plot, we need to use **all five** summaries: $\bar{x}$, $\bar{y}$, $SD_x$, $SD_y$ and $r$.
- The **regression line** connects $(\bar{x}, \bar{y})$ to $(\bar{x} + \mathrm{SD}_x, \bar{y} + r\mathrm{SD}_y)$



Note the improvement at the extremes.



# Summary of regression line

| Feature | Regression Line $y \sim x$ ($y = a + bx$) |
|---|---|
| Connects | $(\bar{x}, \bar{y})$ to $(\bar{x} + \mathrm{SD}_x, \bar{y} + r\mathrm{SD}_y)$ |
| Slope (b) | $r\frac{\mathrm{SD}_y}{\mathrm{SD}_x}$ |
| Intercept (a) | $\bar{y} - b\bar{x}$ |

**Optimality:** We can derive the regression line using calculus, by minimising the **sum of squares** of the **residuals**.

# In R

```r
lm(y ~ x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    33.8866       0.5141
```

```r
model = lm(y ~ x)
model$coeff
```

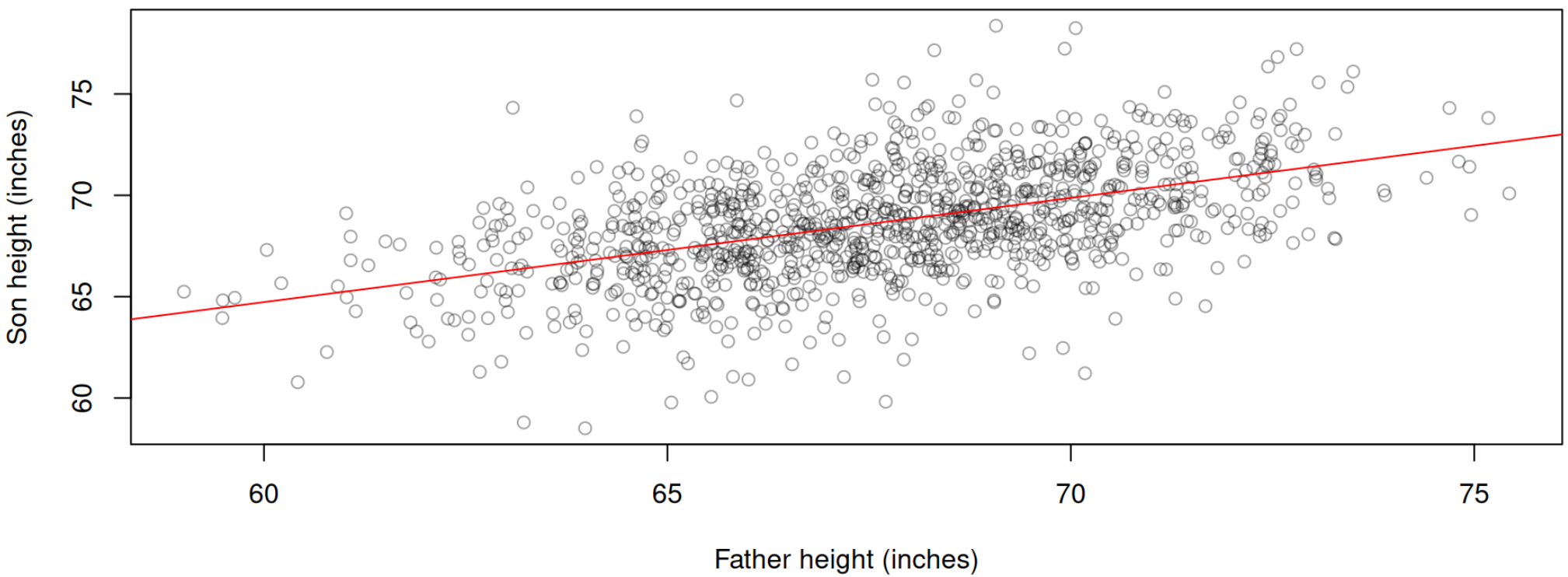
```
(Intercept)           x
  33.886604    0.514093
```

So for $x$ = father height and $y$ = son height, the regression line is

$$y = 33.886604 + 0.514093x$$

# Plotting the regression line

```r
plot(x, y, xlab = "Father height (inches)", ylab = "Son height (inches)", col = adjustcolor("black",
    alpha.f = 0.35))
abline(lm(y ~ x), col = "red")
```



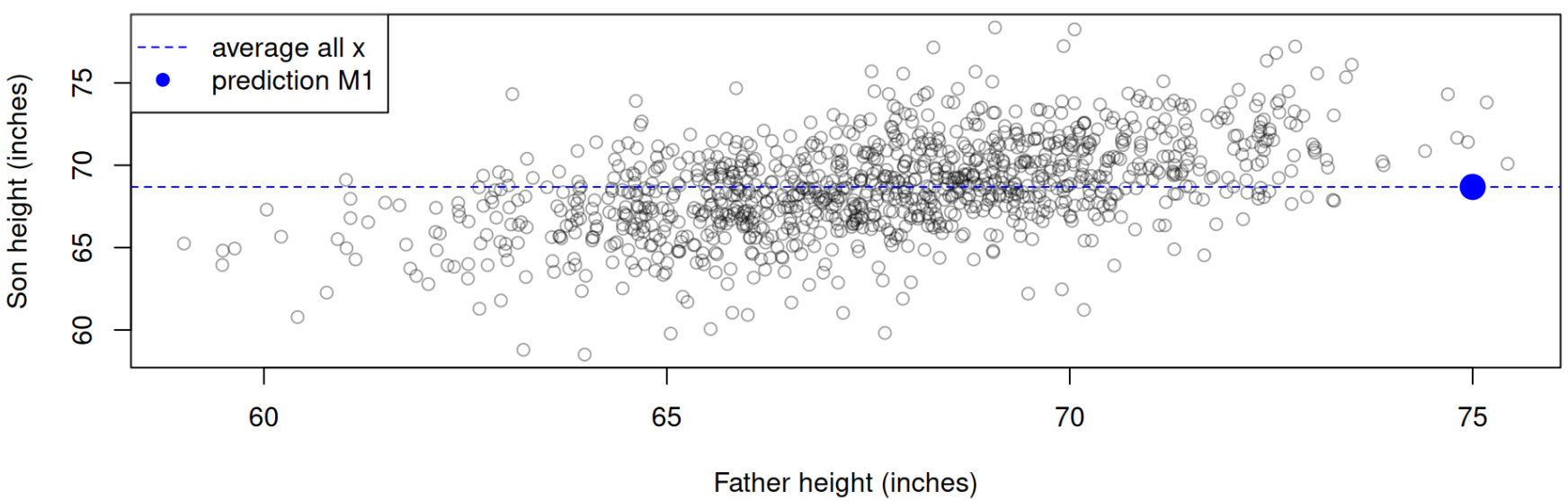
# Prediction

# Baseline prediction

- For new born (son), the father is 75 inches tall, how can we predict the son's height?
- If you don't use the information of the independent variable $x$ at all, a basic prediction of $y$ would be the **average** of $y$ for **all** the $x$ values in the data.
- So for any father's height, we could predict the son's height to be 68.68.

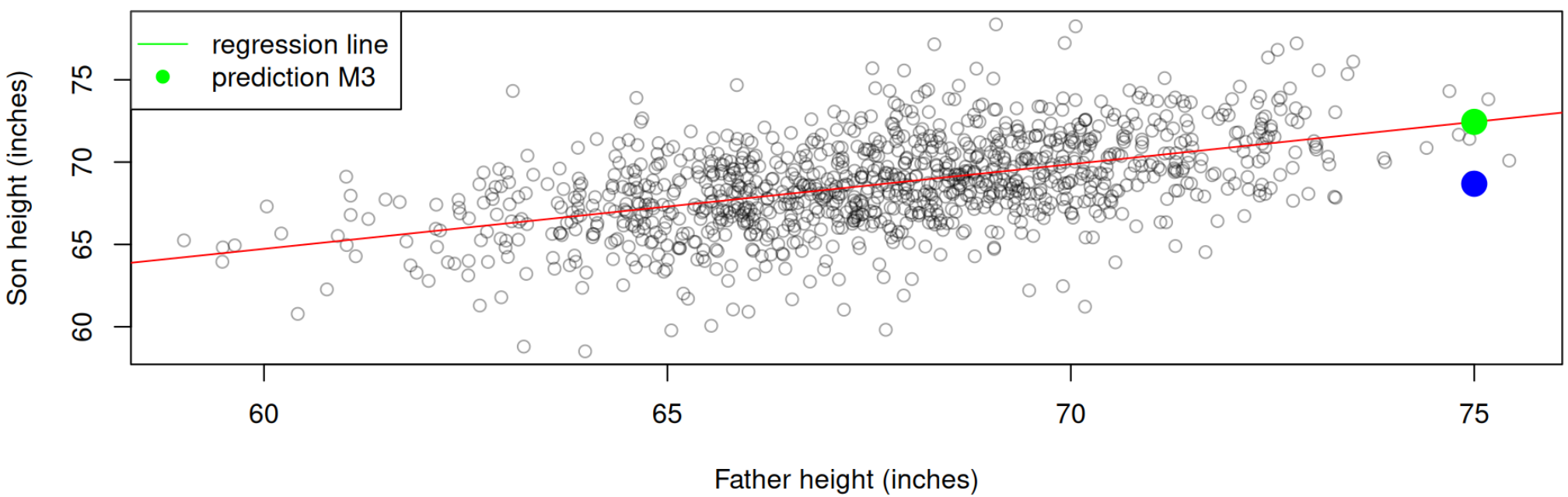
```
1 mean(y)
```

```
[1] 68.68407
```



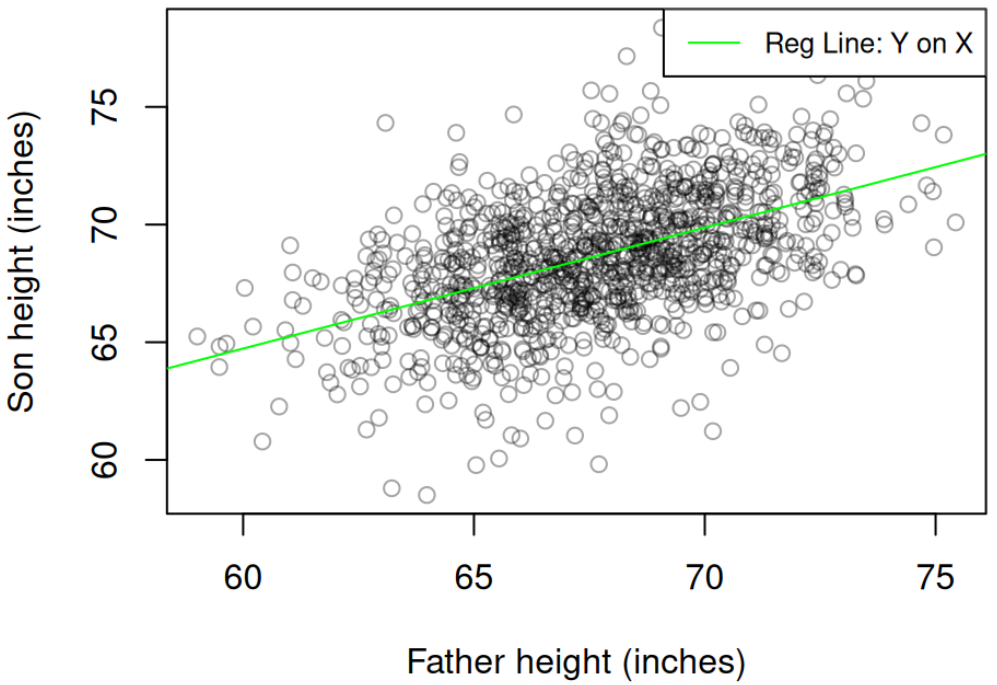
# The Regression line

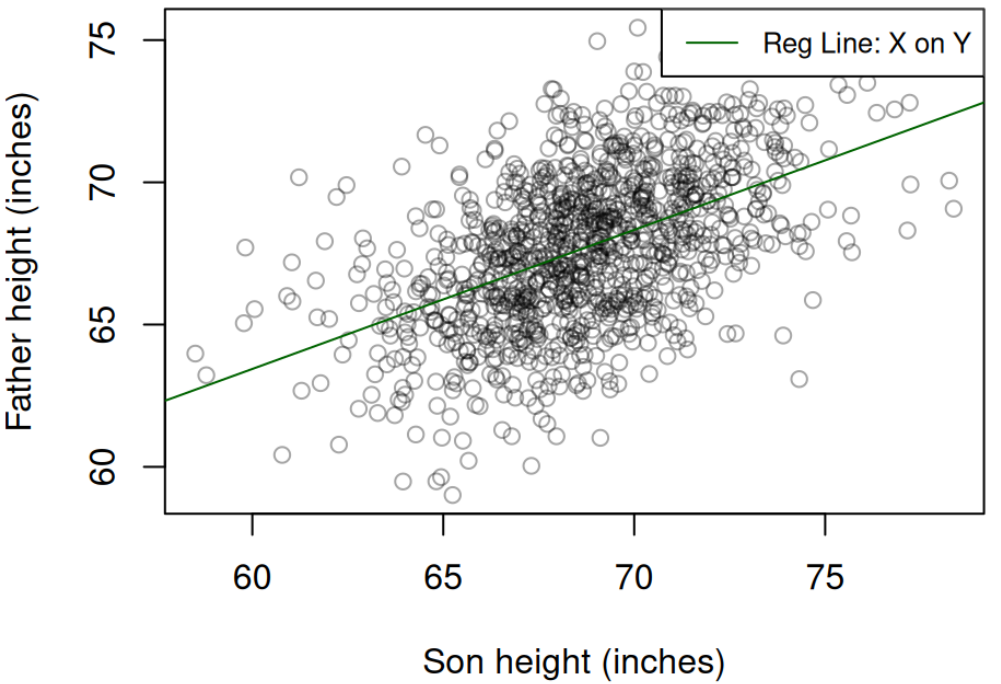
- A better prediction is based on the regression line $y = \text{slope} \times x + \text{intercept}$
- For the height data: $y = 33.886604 + 0.514093x$
- So for any father's height 75, we could predict the son's height to be 72.44.



# Can we also use Y to predict X?

We can predict $Y$ from $X$ or $X$ from $Y$, depending on what fits the context.



# Beware!

- Can we just simply rearrange the equation?

$$(y = a + bx) \implies (x = -\tfrac{a}{b} + \tfrac{1}{b}y)$$

- The answer is NO unless $r = \pm 1$ (data clustered along the line).
- We need to **refit** the model.

| Feature | Regression Line $y \sim x$ $(y = a + bx)$ | Regression Line $x \sim y$ $(x = \tilde{a} + \tilde{b}y)$ |
|---|---|---|
| Connects | $(\bar{x}, \bar{y})$ to $(\bar{x} + \mathrm{SD}_x, \bar{y} + r\mathrm{SD}_y)$ | $(\bar{y}, \bar{x})$ to $(\bar{y} + \mathrm{SD}_y, \bar{x} + r\mathrm{SD}_x)$ |
| Slope | $b = r\frac{\mathrm{SD}_y}{\mathrm{SD}_x}$ | $\tilde{b} = r\frac{\mathrm{SD}_x}{\mathrm{SD}_y}$ |
| Intercept | $a = \bar{y} - b\bar{x}$ | $\tilde{a} = \bar{x} - \tilde{b}\bar{y}$ |

# Rearranging the equation leads to different coefficients

With $x$ as the independent variable and $y$ as the dependent varialbe:

```
1 lm(y ~ x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    33.8866       0.5141
```

With $y$ as the independent variable and $x$ as the dependent varialbe:

```
1 lm(x ~ y)
```

```
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
    34.1075       0.4889
```