Correlation and Linear Model

Modelling Data | Linear Model

STAT5002

The University of Sydney

Feb 2025

THE UNIVERSITY OF SYDNEY

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

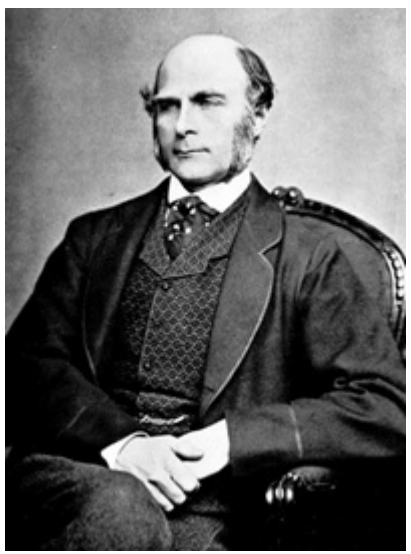
Linear model

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

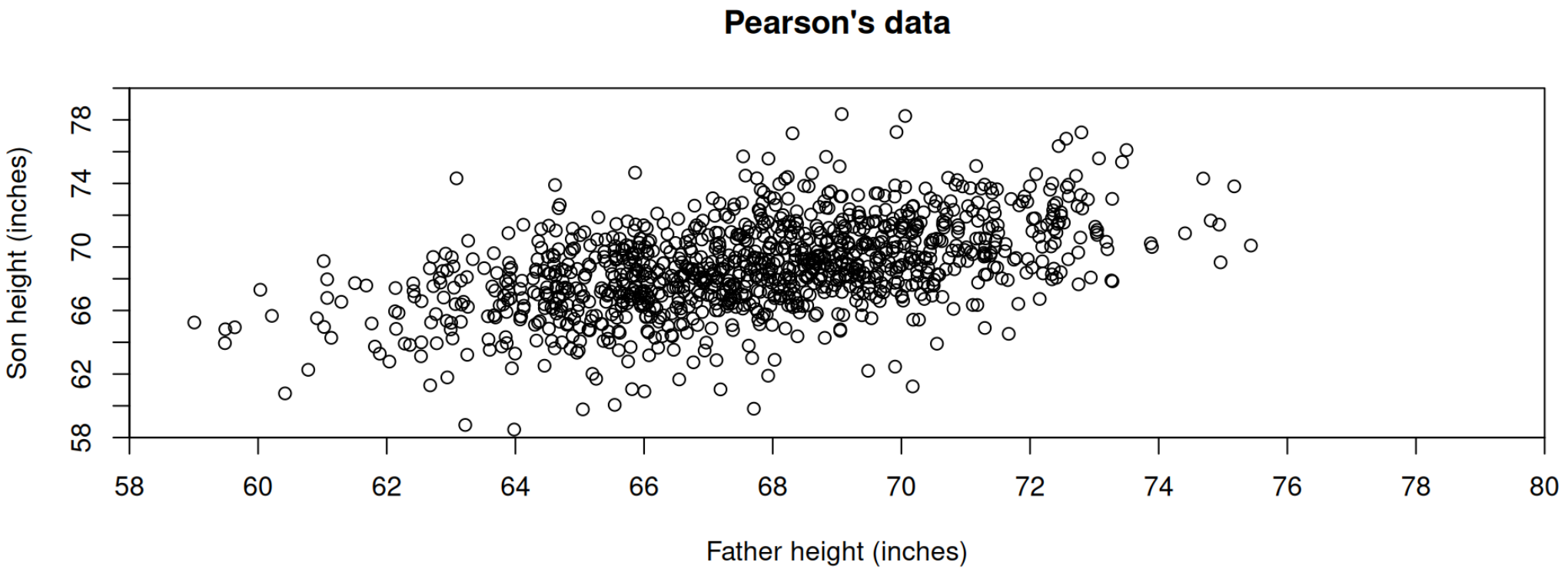
Scatter plots

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?

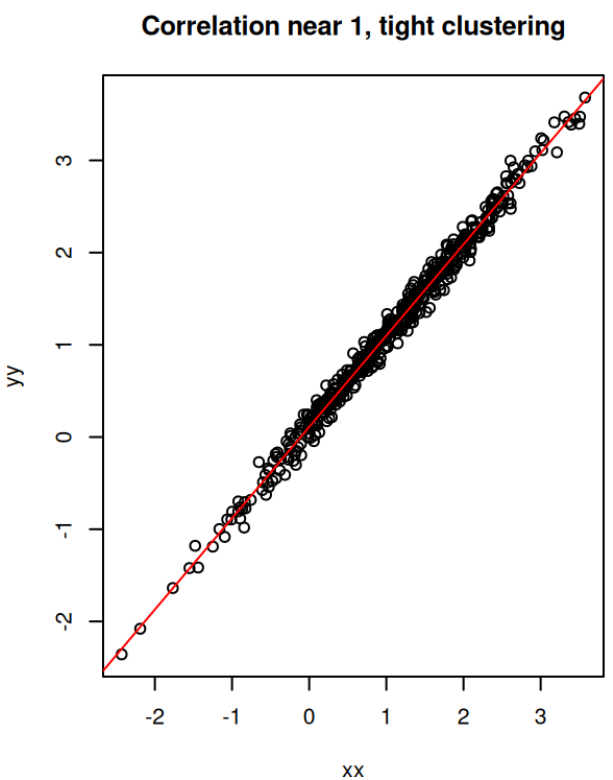
- Formally, we have (x_i, y_i) for $i = 1, 2, \ldots, n$.
- X and Y can have the same role
- X and Y may have different roles: for example, X can be an **independent** variable (or explanatory variable, predictor or regressor) which we use to explain or predict Y, the **dependent** variable (or response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

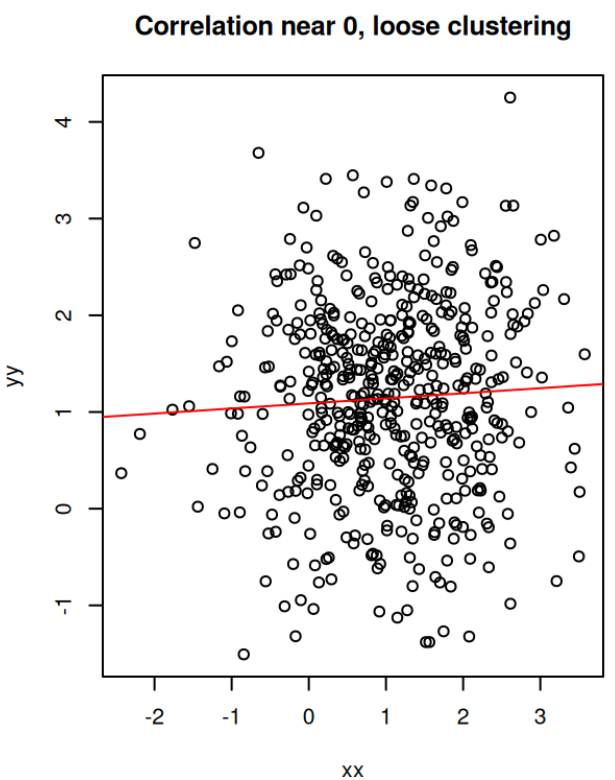
How can we summarise bivariate data?

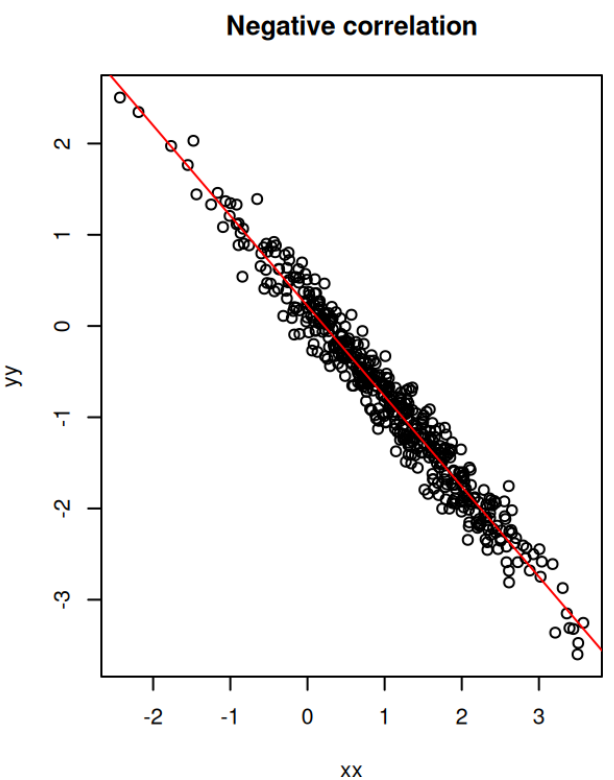
Bivariate data (or a scatter plot) can be summarised by the following **five** numerical summaries:

- Sample mean and sample SD of X ($\bar{x}$, SD_x)
- Sample mean and sample SD of Y ($\bar{y}$, 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
```

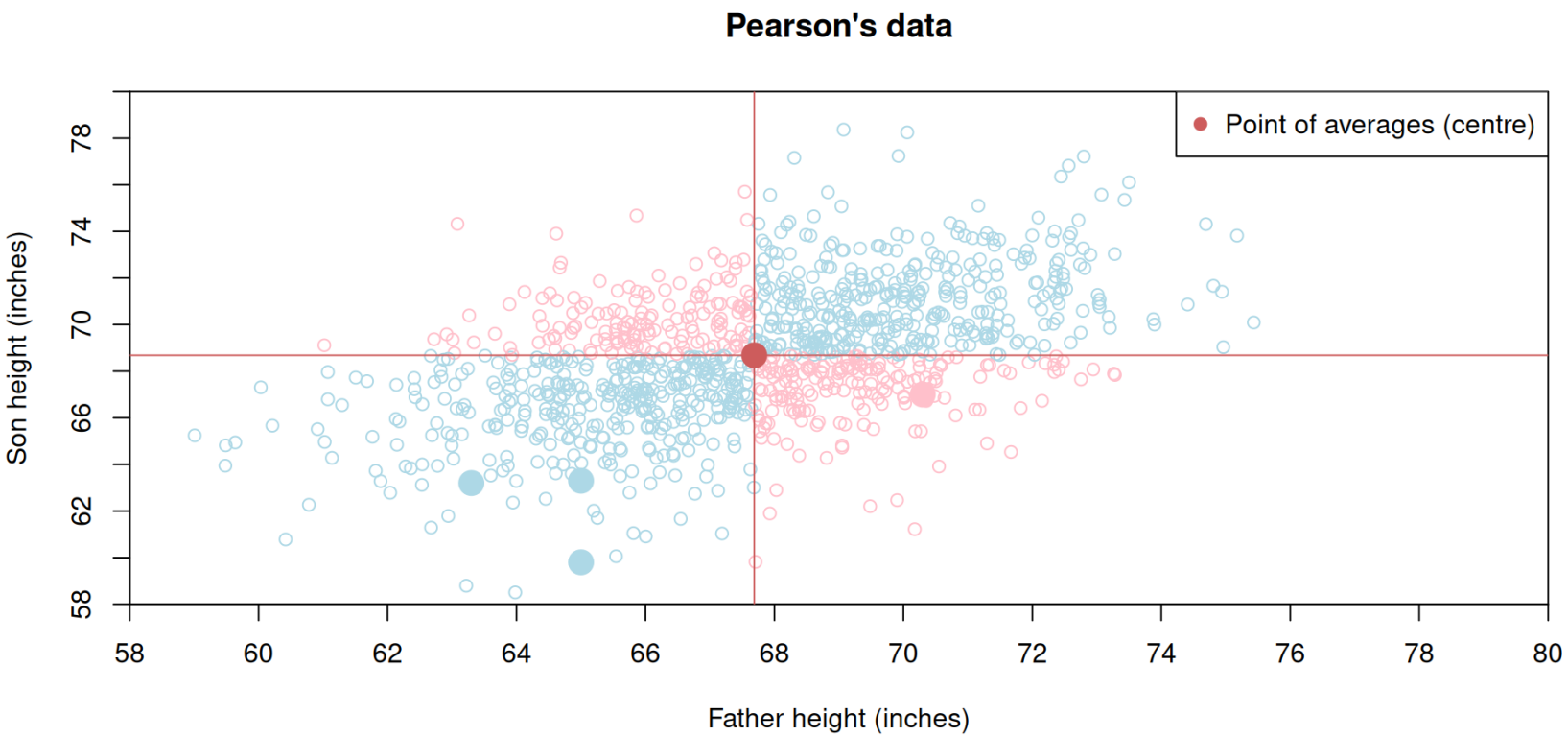
How does r measure association?

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	

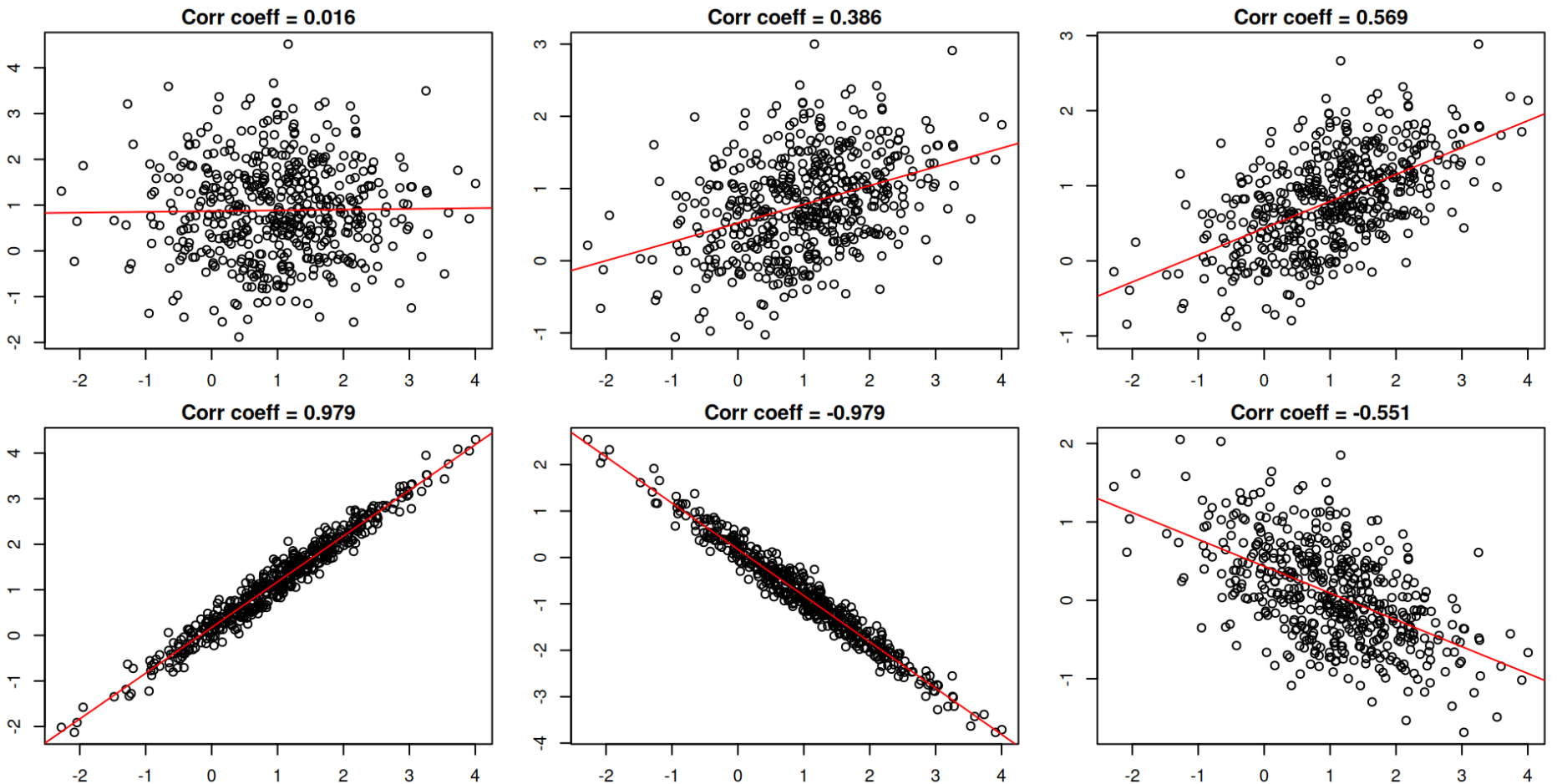
We divide the scatter plot into 4 quadrants, at the point of averages (centre).

- 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 ± 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)
```

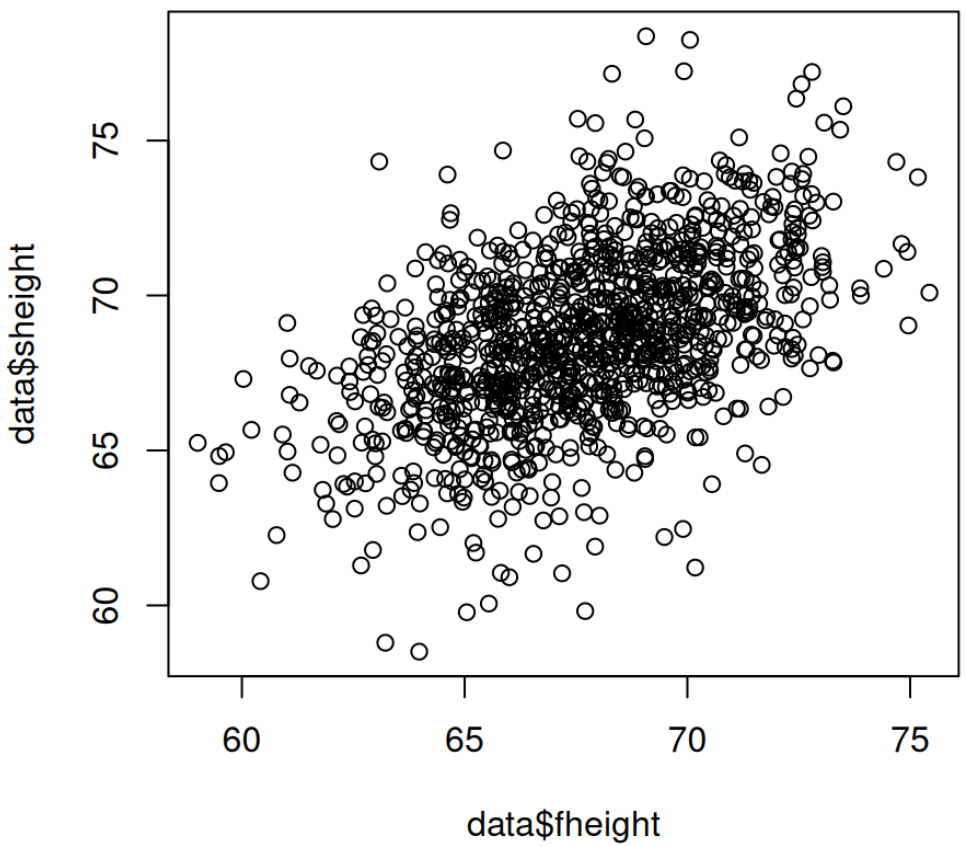
```
cor(data$fheight, data$sheight)
```

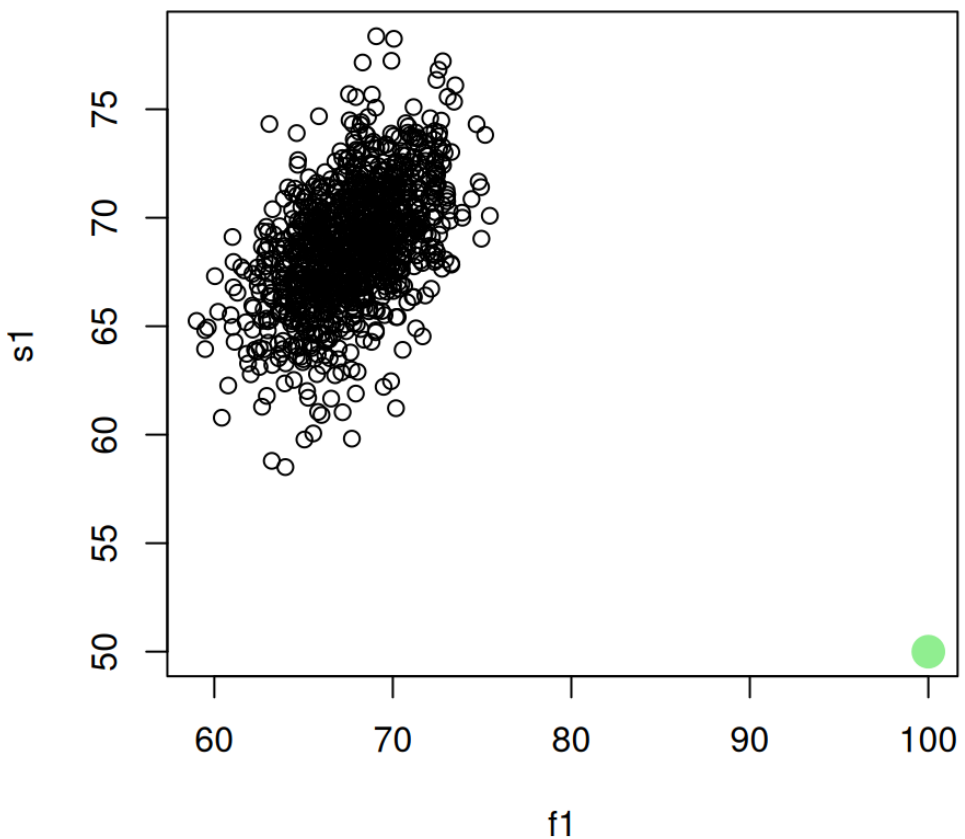
```
[1] 0.5013383
```

```
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?

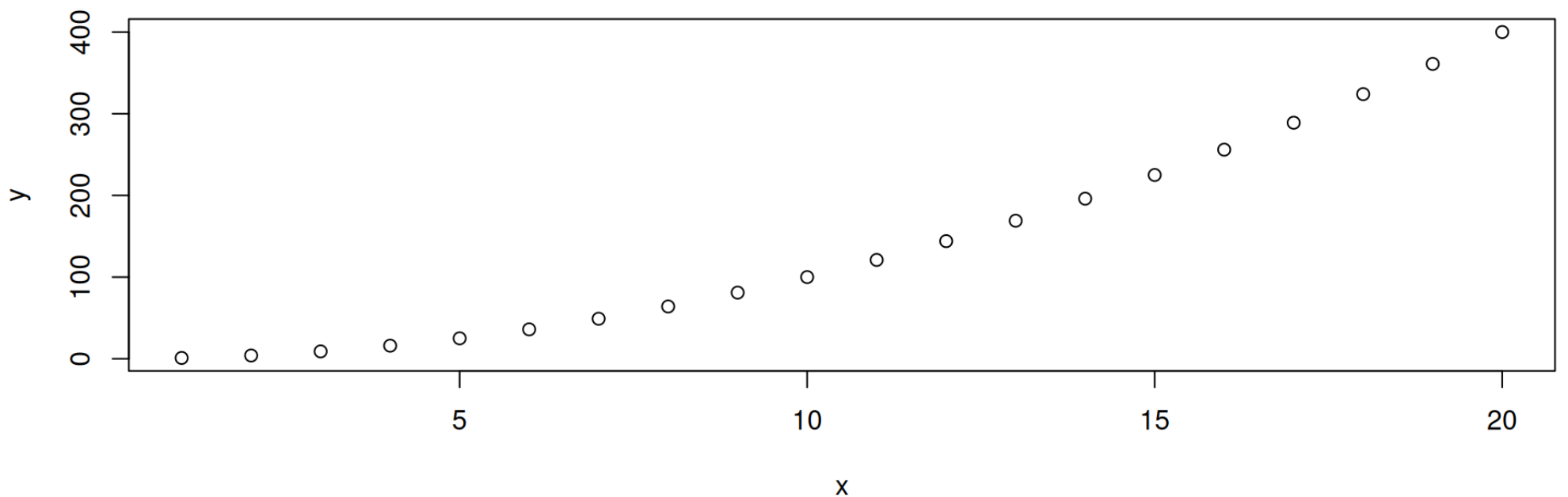
```
x = c(1:20)
y = x^2
cor(x, y)
```

```
[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.

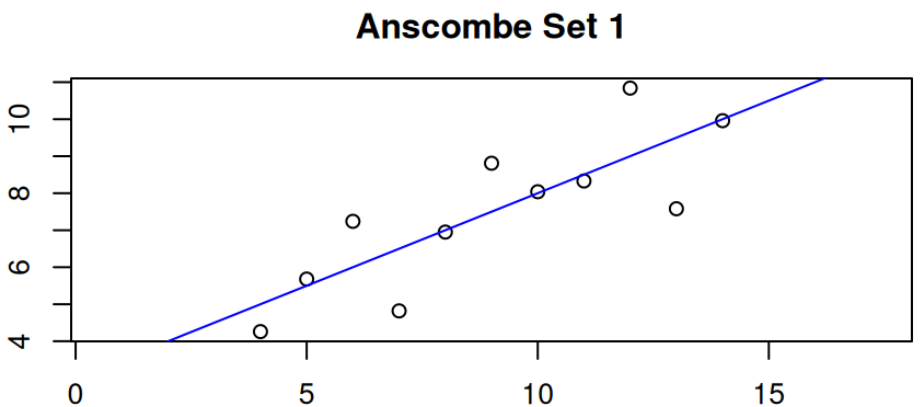
Warning 4:

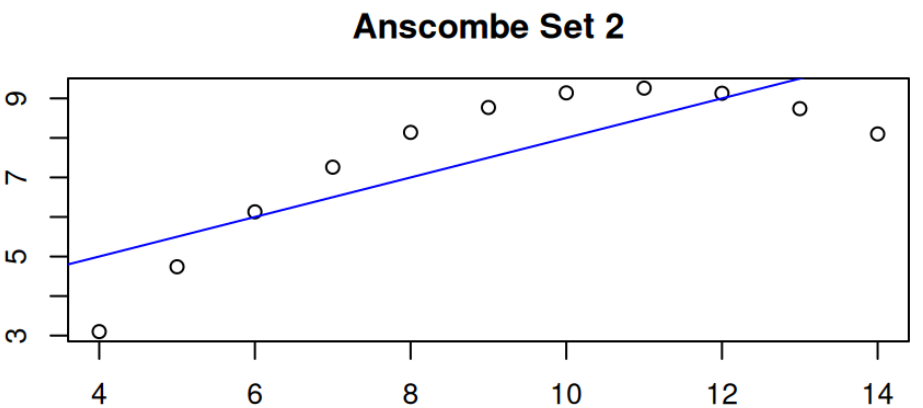
The same correlation coefficient can arise from very different data

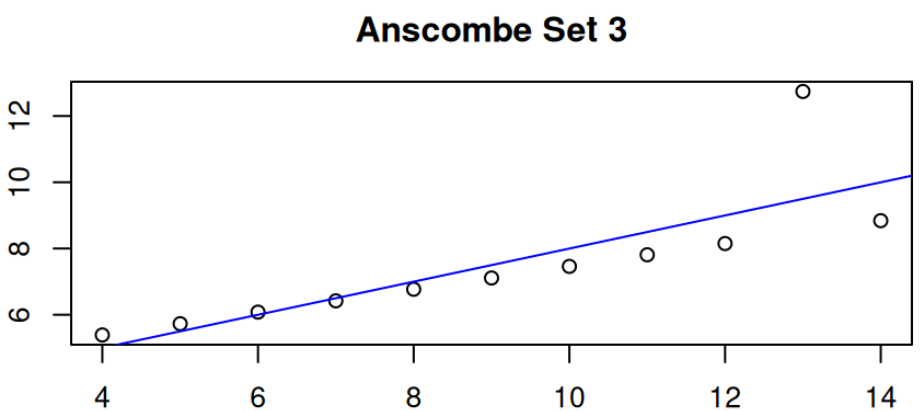
The following data sets (Anscombes Quartet) have the **same** $\bar{x}, SD_x, \bar{y}, SD_y$, and also the **same** value of r.

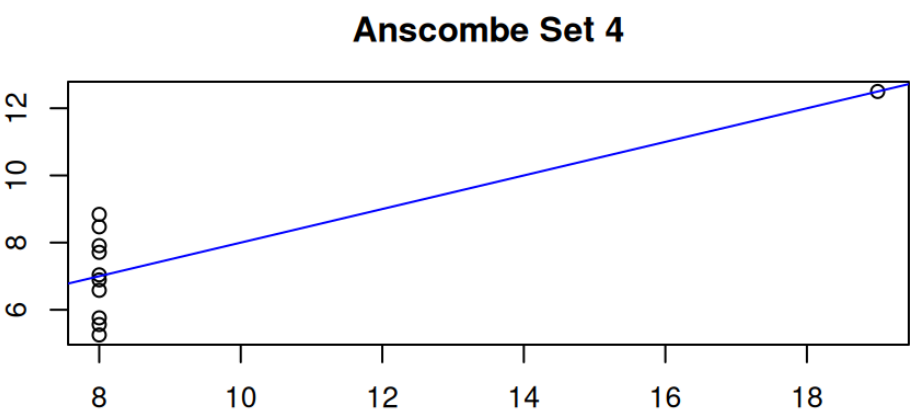
```
x_mean: 9 9 9 9
x_sd: 3.316625 3.316625 3.316625 3.316625
y_mean: 7.500909 7.500909 7.5 7.500909
y_sd: 2.031568 2.031657 2.030424 2.030579
r: 0.8164205 0.8162365 0.8162867 0.8165214
```

But look at the scatter plots.







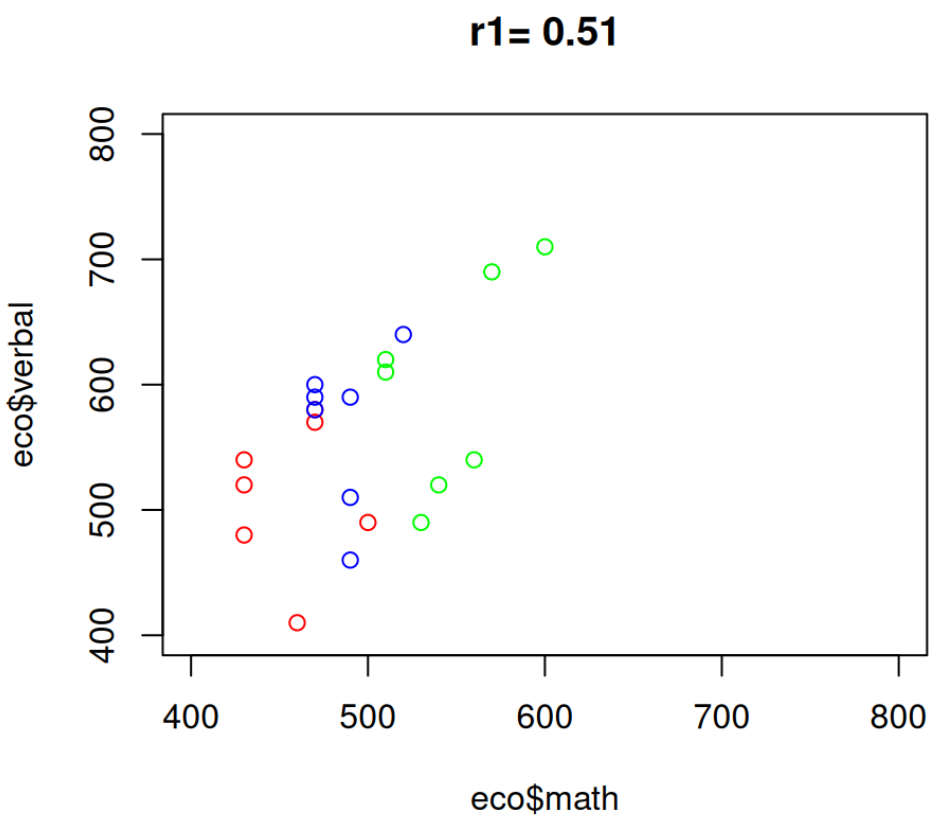


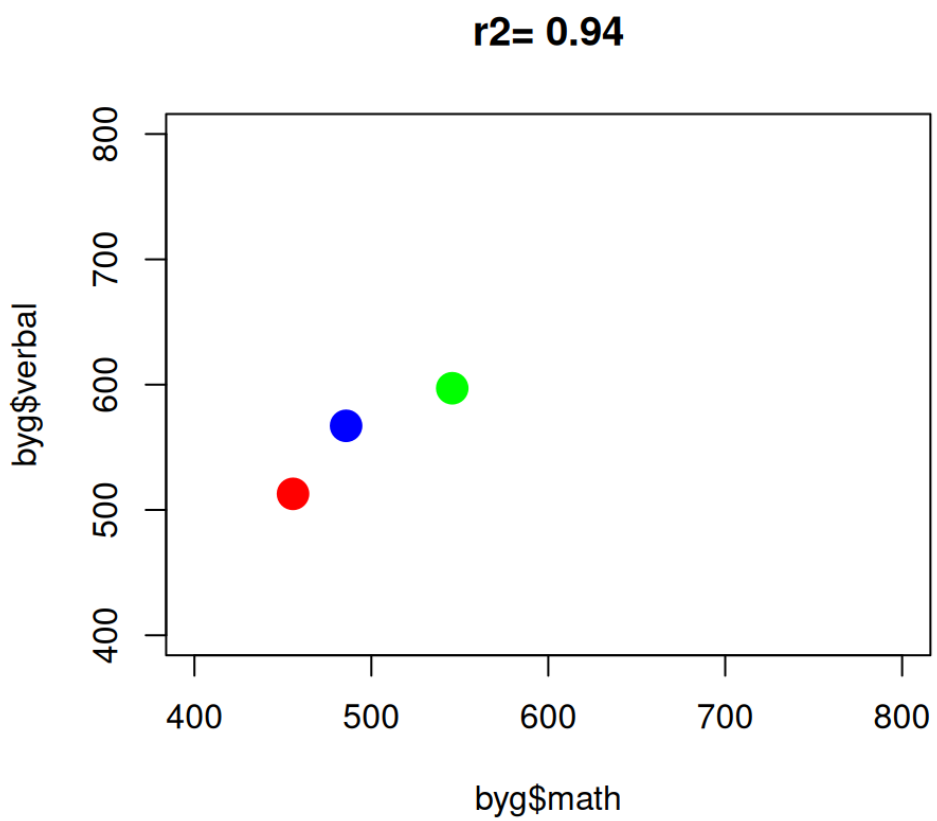
Warning 5 (not for assessment):

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).

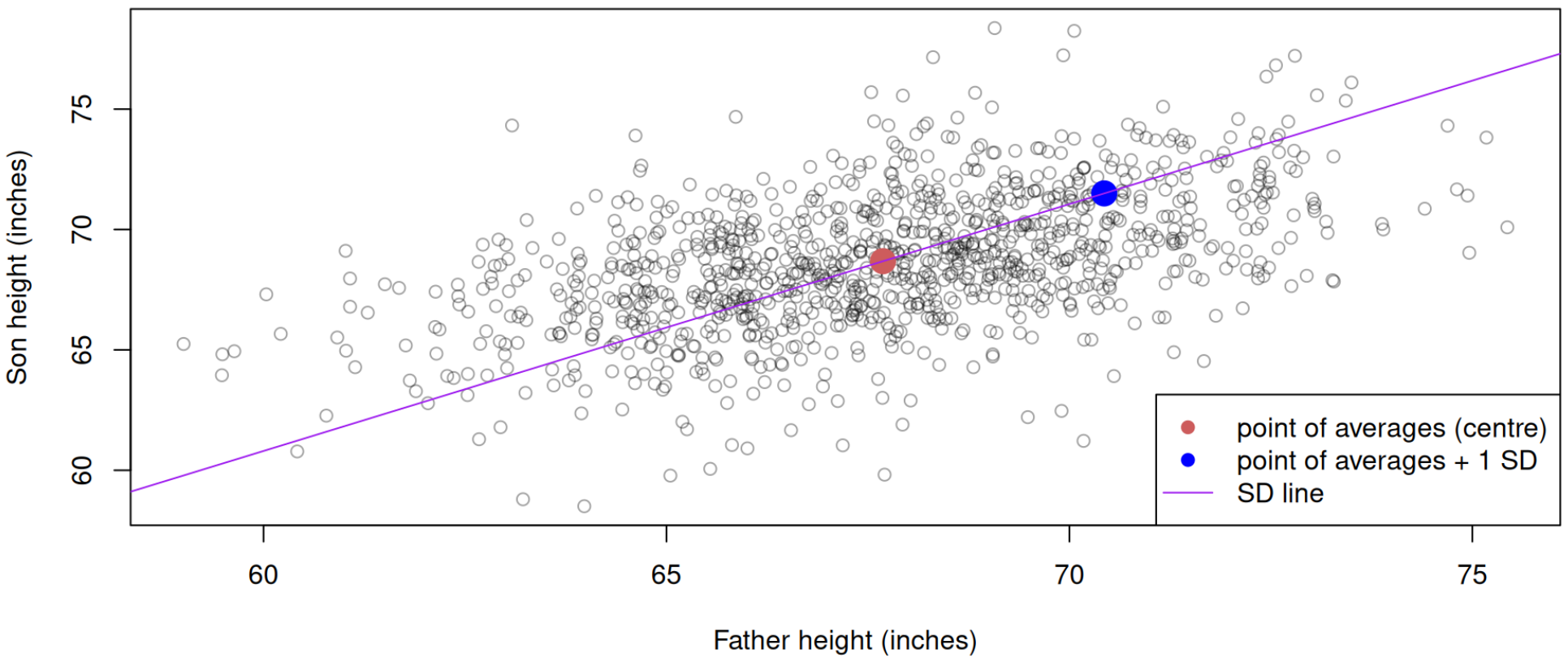
Regression line

Pearson’s plot of heights

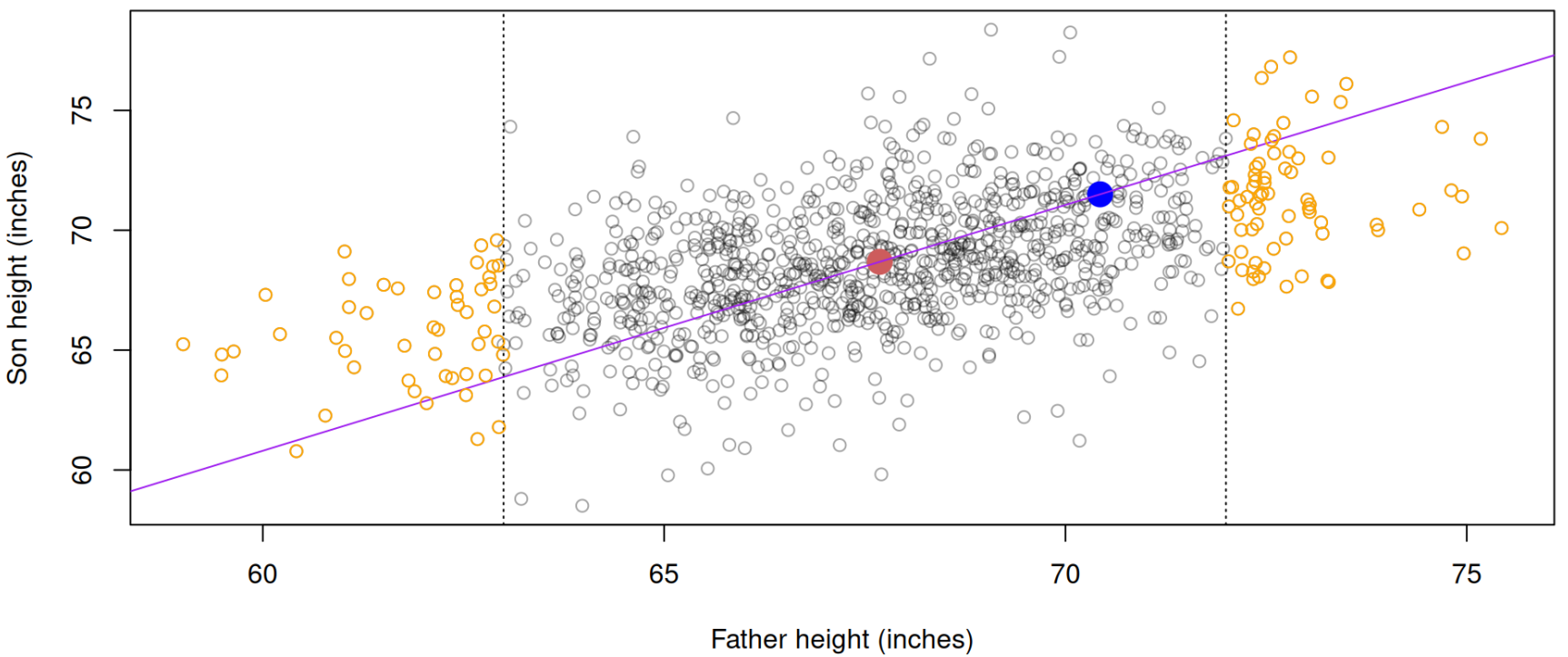
- How can we summarise the data with a line?
- How do we find the **optimal** line?

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).



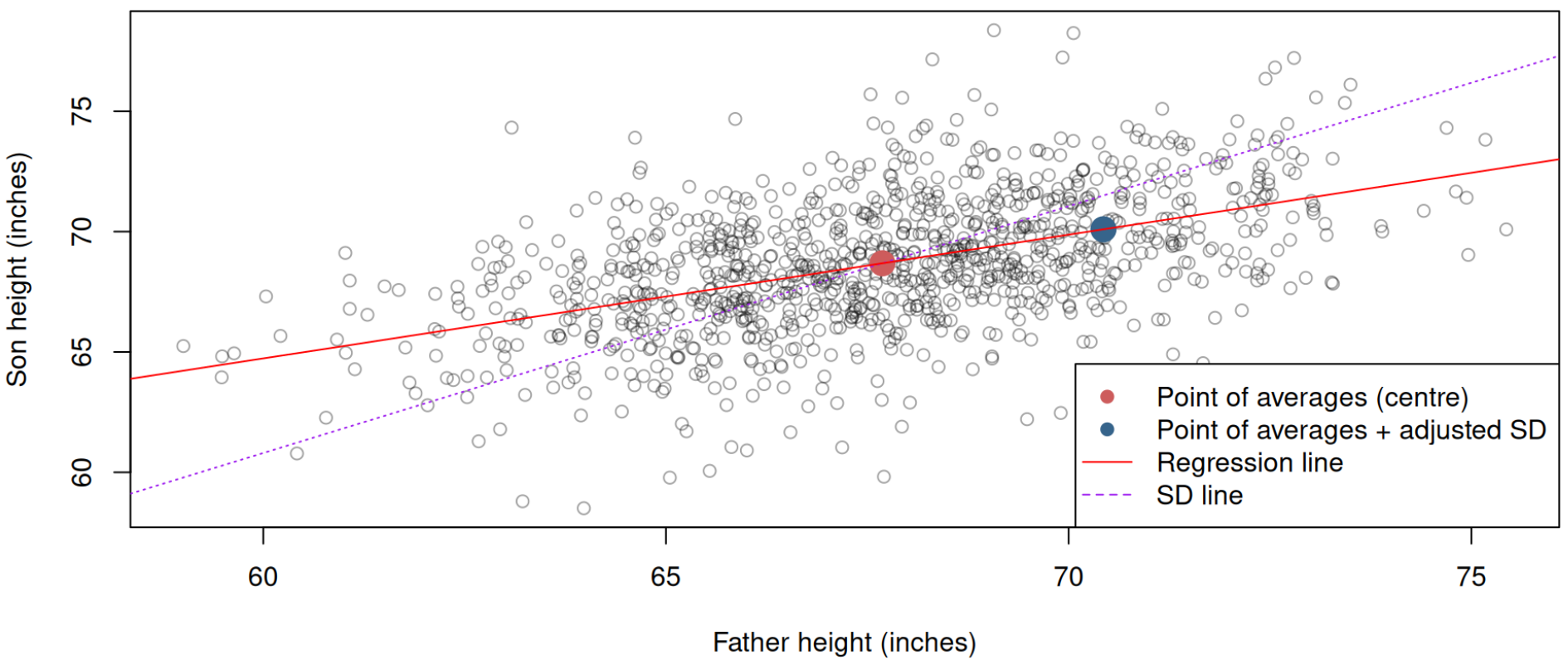
Note how it underestimates (LHS) and overestimates (RHS) at the extremes.



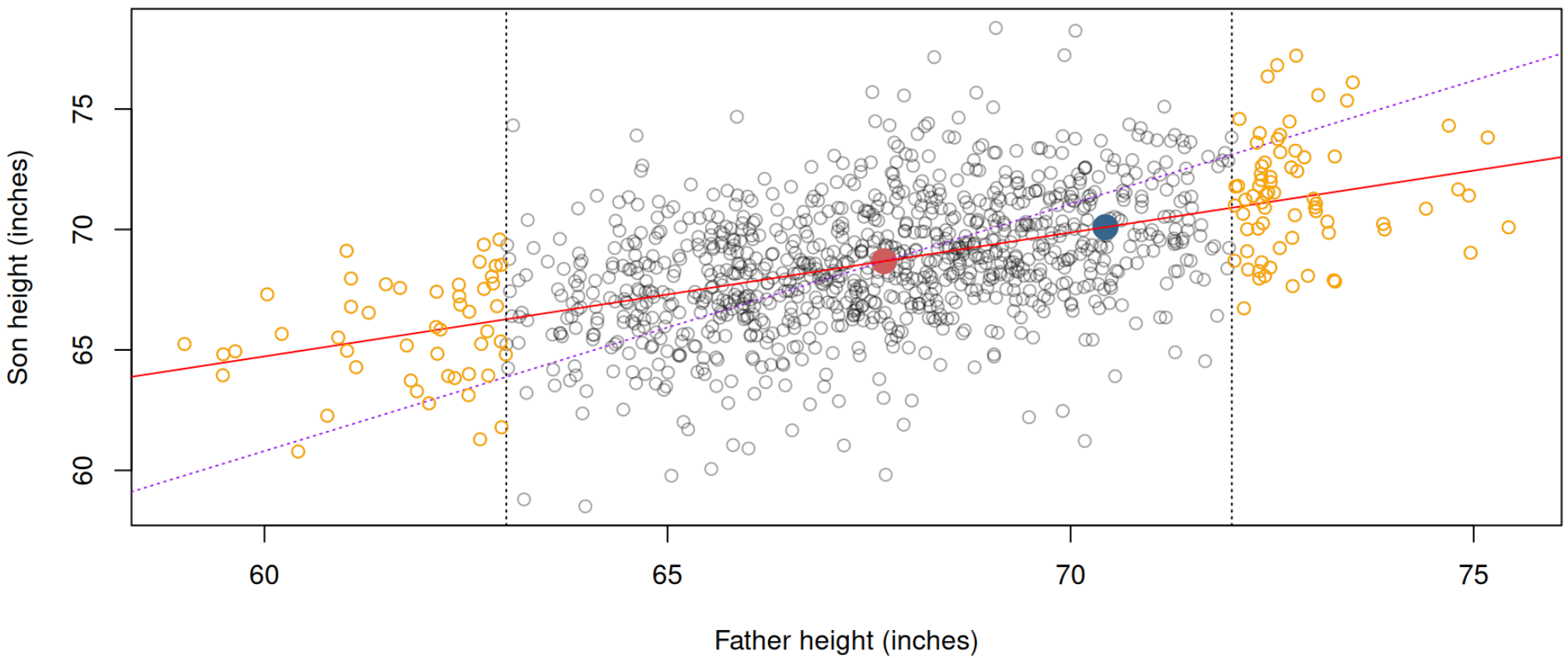
- Recall that X, Y can have the same mean and SD but very different correlation coefficient.
- The above model does not use the correlation coefficient, so it is insensitive to the amount of clustering around the line.
- How to quantify the quality of the fitted line so we can define the **optimal**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

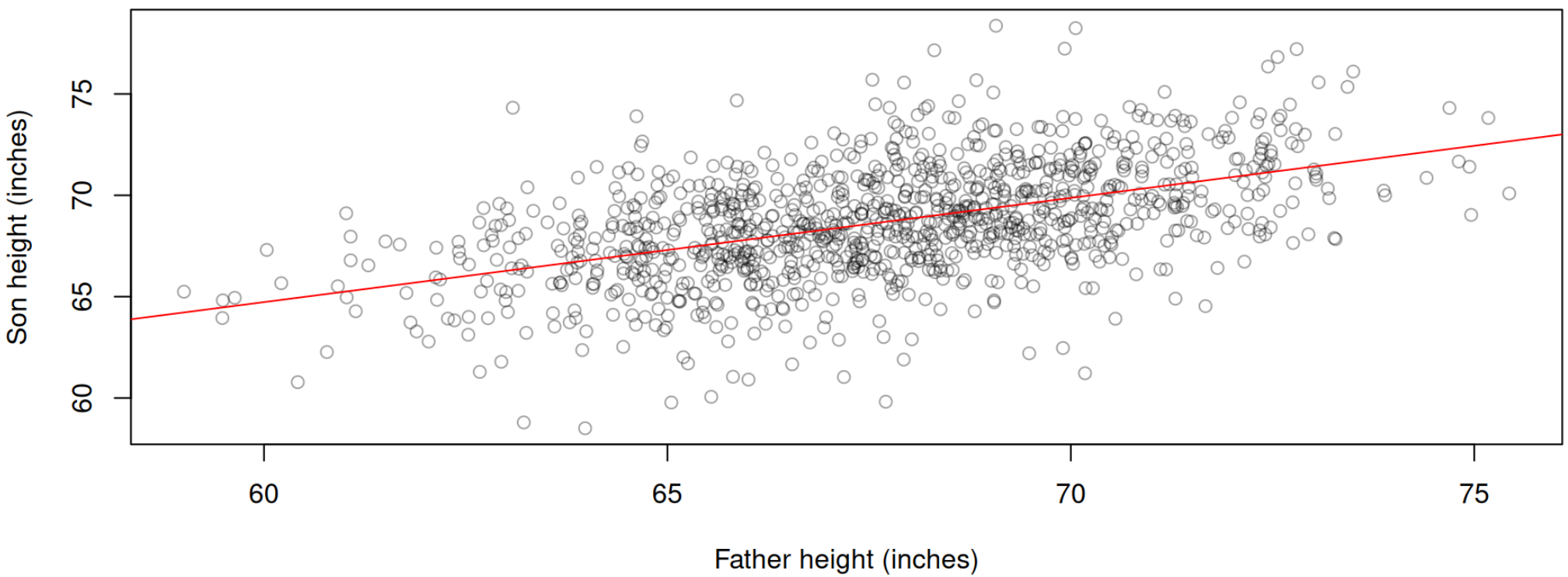
```
(Intercept)           x
  33.886604    0.514093
```

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")
```



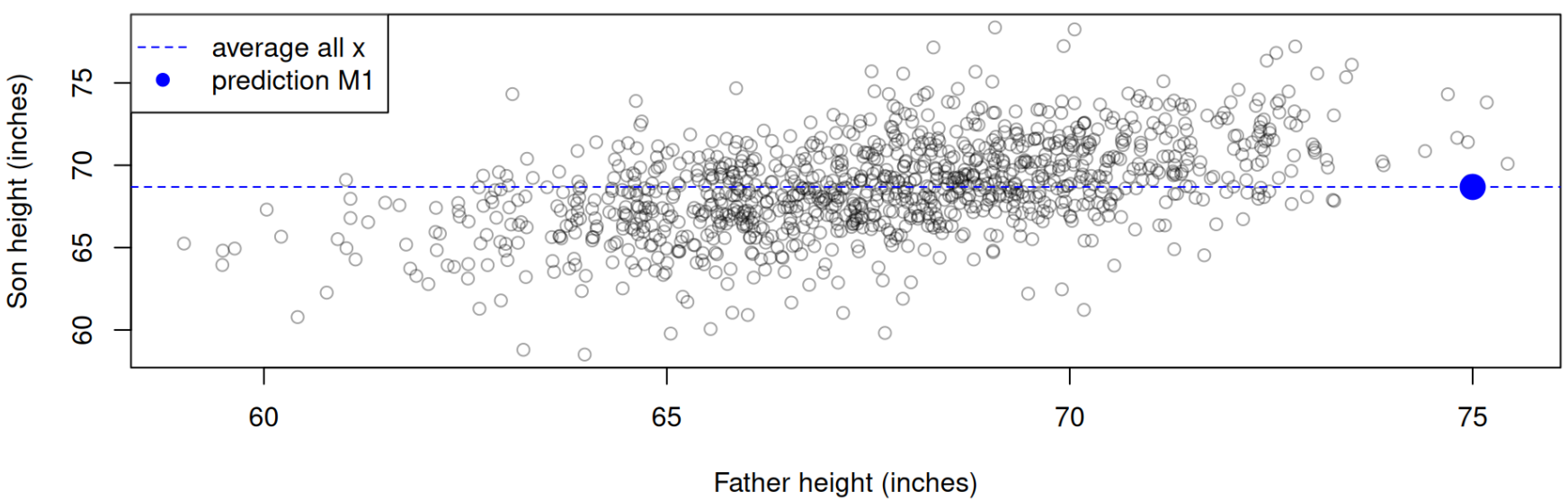
Prediction

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.

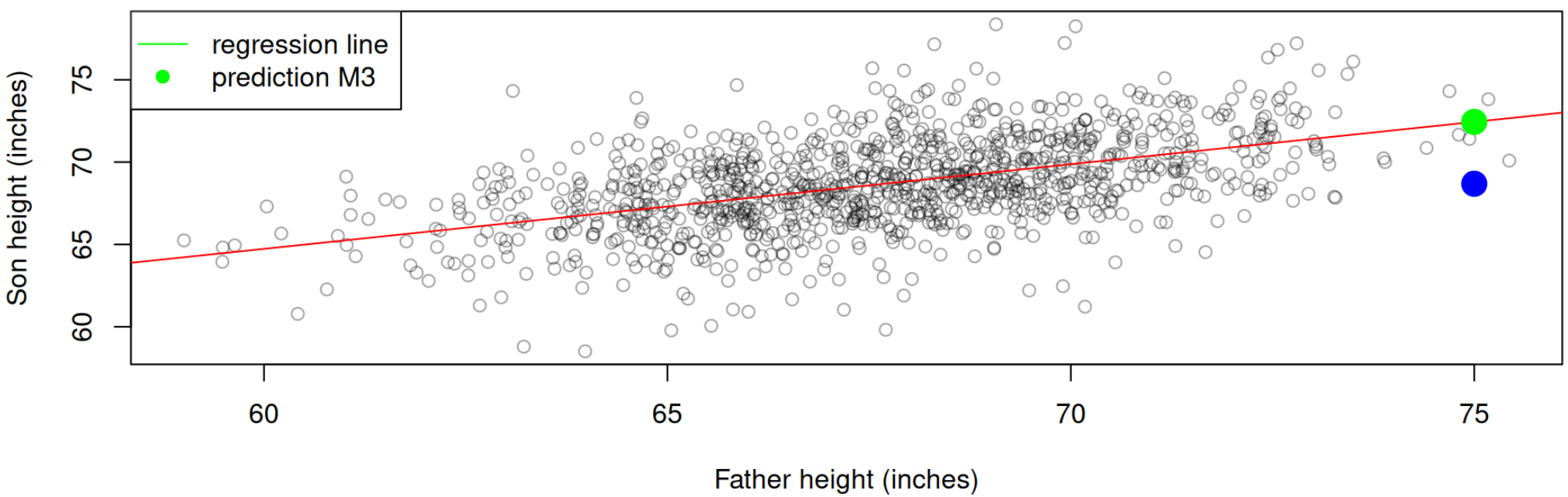
```
1 mean(y)
```

```
[1] 68.68407
```



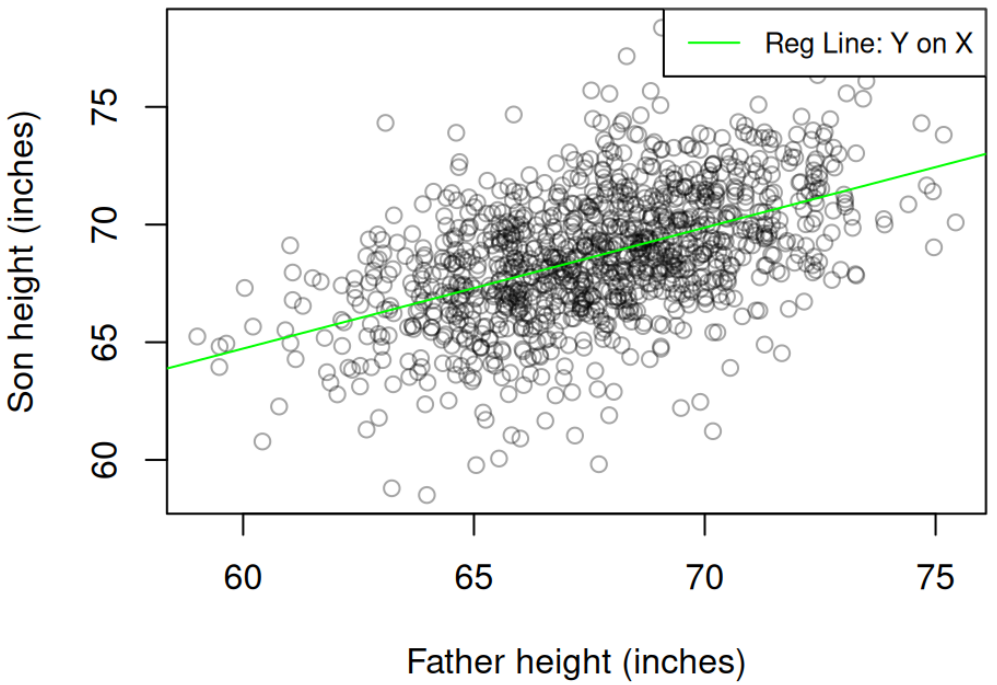
The Regression line

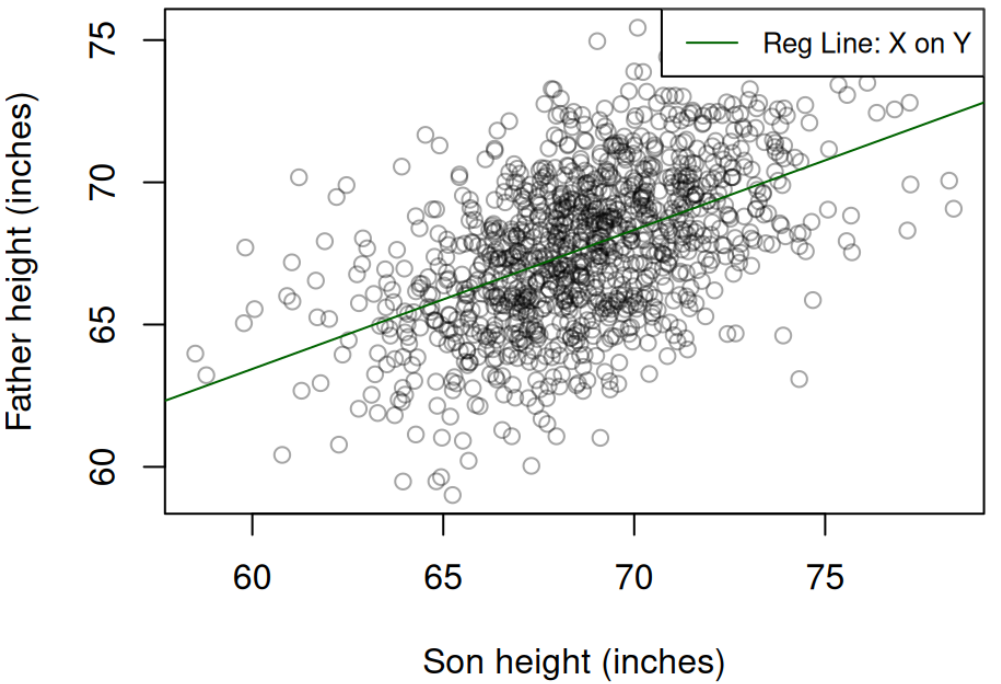
- A better prediction is based on the regression line $y = \text{slope} \times x + \text{intercept}$
- For the height data: $y = 33.886604 + 0.514093x$
- So for any father's height 75, we could predict the son's height to be 72.44.



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

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

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

Coefficients:
(Intercept)            x
    33.8866       0.5141
```

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

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

Coefficients:
(Intercept)            y
    34.1075       0.4889
```

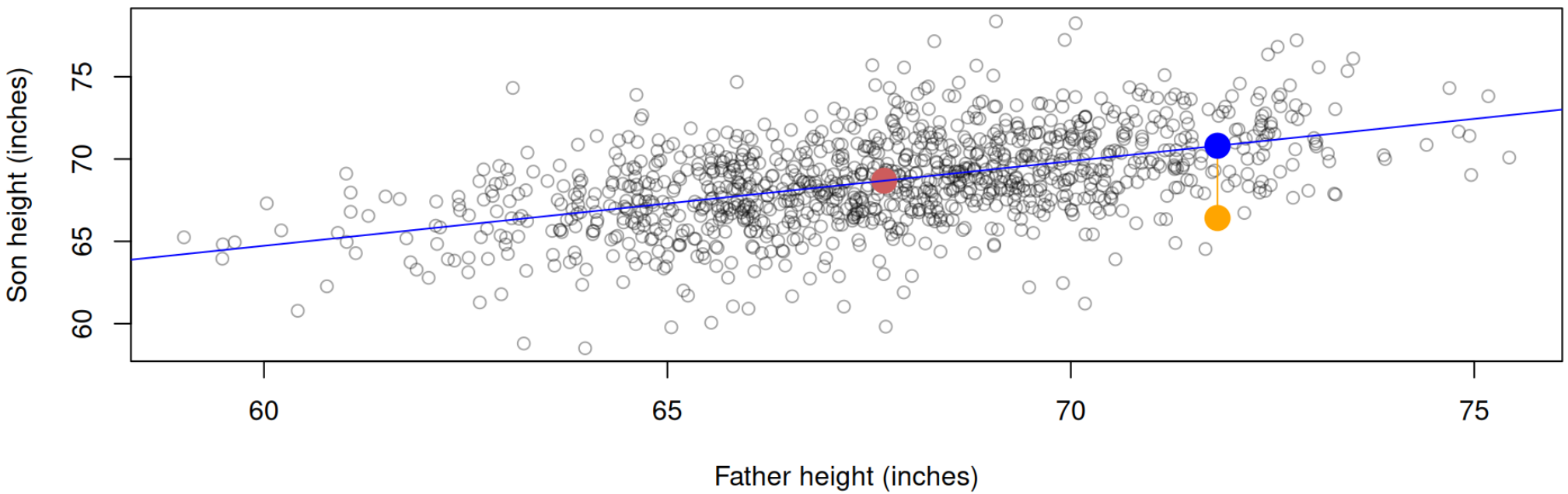
Residuals and properties

Residuals

We can now make predictions using the regression line. But we have some prediction **error**.

Residual (prediction error)

- A **residual** is the vertical distance of a point above or below the regression line.
- A residual represents the error between the actual value and the prediction.



When the father's height is 71.82, the **actual value** of the son's height is 66.42 with **predicted value** 70.81, so the residual is -4.39.

Formally, given the actual value (y_i) and the prediction ($\hat{y}_i$), a residual is

$$e_i(a, b) = y_i - \hat{y}_i = y_i - (\underbrace{a}_{\text{intercept}} + \underbrace{b}_{\text{slope}} x_i).$$

```
1 l = lm(y ~ x)
2 y[39] - l$fitted.values[39]
```

```
      39
-4.390582
```

```
1 l$residuals[39]  # Or directly
```

```
      39
-4.390582
```

The regression line is the **best** (optimal) linear model - it provides the best fit to the data as the sum of the squared residuals $\sum_{i=1}^{n} e_i(a, b)^2$ is as small as it can be.

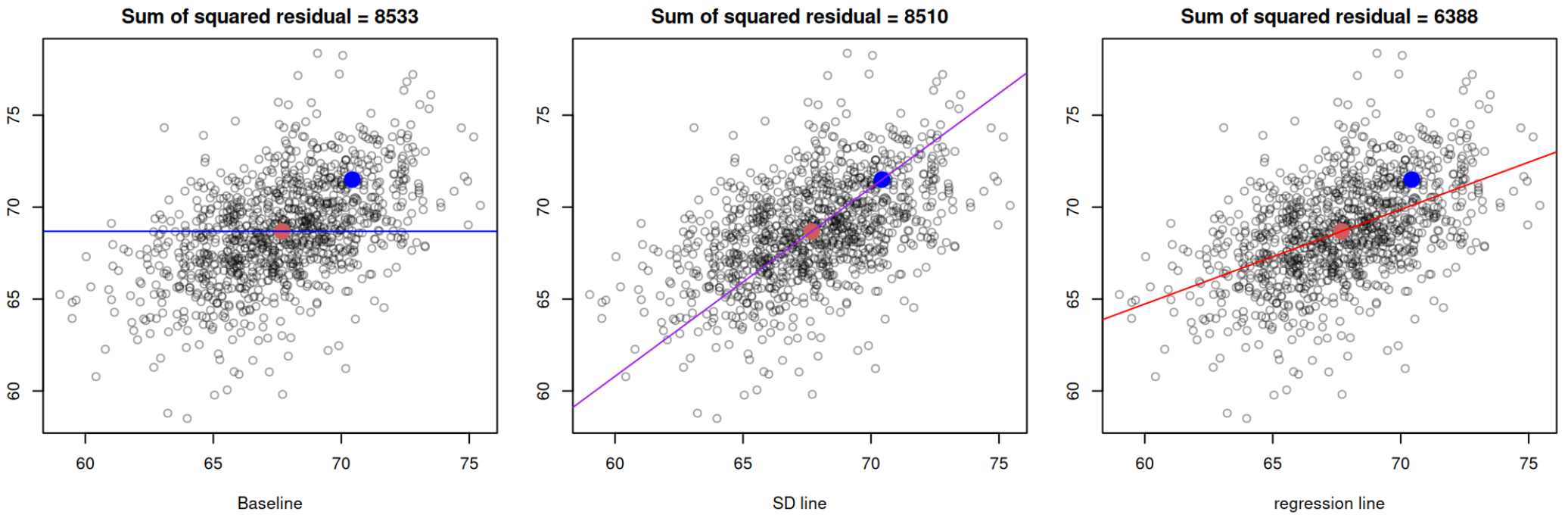
Optimality of regression line

- We first consider a general line $y = \alpha + \beta x$ with intercept α and slope β.
- Given the data set $\{x_i, y_i\}, i = 1, \ldots, n$, a pair of variables (α, β) for defining a line, the residual is

$$e_i(\alpha, \beta) = y_i - (\alpha + \beta x_i).$$

so that the sum of squared residuals becomes

$$f(\alpha, \beta) = \sum_{i=1}^{n} e_i(\alpha, \beta)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$



- Our goal is to find the intercept a and the slope b that minimises $f(\alpha, \beta)$:

$$f(a, b) \leq f(\alpha, \beta) \quad \text{for all} \quad \alpha, \beta$$

- The following derivation of optimality is not for examination

How to find such a minimiser (a, b)? It needs to be a stationary point of the function f such that

$$\frac{\partial f}{\partial \alpha}(a, b) = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) \ = 0$$

and

$$\frac{\partial f}{\partial \beta}(a, b) = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0.$$

We use the **first equation** to find the **intercept**, $\frac{\partial f}{\partial \alpha}(a, b) = 0$ is equivalent to

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

Dividing both sides by n, this gives

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x},$$

which leads to $\boldsymbol{a = \bar{y} - b\bar{x}}$.

We can find the **slope** by substituting $a = \bar{y} - b\bar{x}$ into the **second equation**. This way, $\frac{\partial f}{\partial \beta}(a, b) = 0$ becomes

$$\sum_{i=1}^{n}[y_i - (\bar{y} - b\bar{x}) - bx_i]x_i = 0.$$

After rearrangement,

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = b\sum_{i=1}^{n}(x_i - \bar{x})x_i.$$

Because the sum of deviations is zero (topic 3 in week 3), we have $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, and hence

$$\sum_{i=1}^{n}(y_i - \bar{y})\bar{x} = 0 \quad \text{and} \quad \sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = 0.$$

as $\bar{x}$ is a constant for all i.

$$LHS = \left(\sum_{i=1}^{n}(y_i - \bar{y})x_i\right) - \left(\sum_{i=1}^{n}(y_i - \bar{y})\bar{x}\right) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

$$RHS = b\left(\sum_{i=1}^{n}(x_i - \bar{x})x_i\right) - b\left(\sum_{i=1}^{n}(x_i - \bar{x})\bar{x}\right) = b\sum_{i=1}^{n}(x_i - \bar{x})^2$$

By solving the second equation $\frac{\partial f}{\partial \beta}(a, b) = 0$, the slope is

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Recall that

- $SD_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $SD_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

This gives exactly $b = r\frac{SD_y}{SD_x}$ as we claimed in the definition of the regression line. So that we know the regression line is indeed the best among all lines (linear functions) in the sense of sum of squared residuals.

Average of residual is zero

Given the regression line $y = a + bx$, where $a = \bar{y} - b\bar{x}$, the sum of residual

$$\sum_{i=1}^{n} e_i(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i) = \sum_{i=1}^{n} (y_i - (\bar{y} - b\bar{x}) - bx_i)$$

can be expressed as

$$\sum_{i=1}^{n} (y_i - \bar{y}) - b \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Thus, **the mean (average) of residual is zero.**

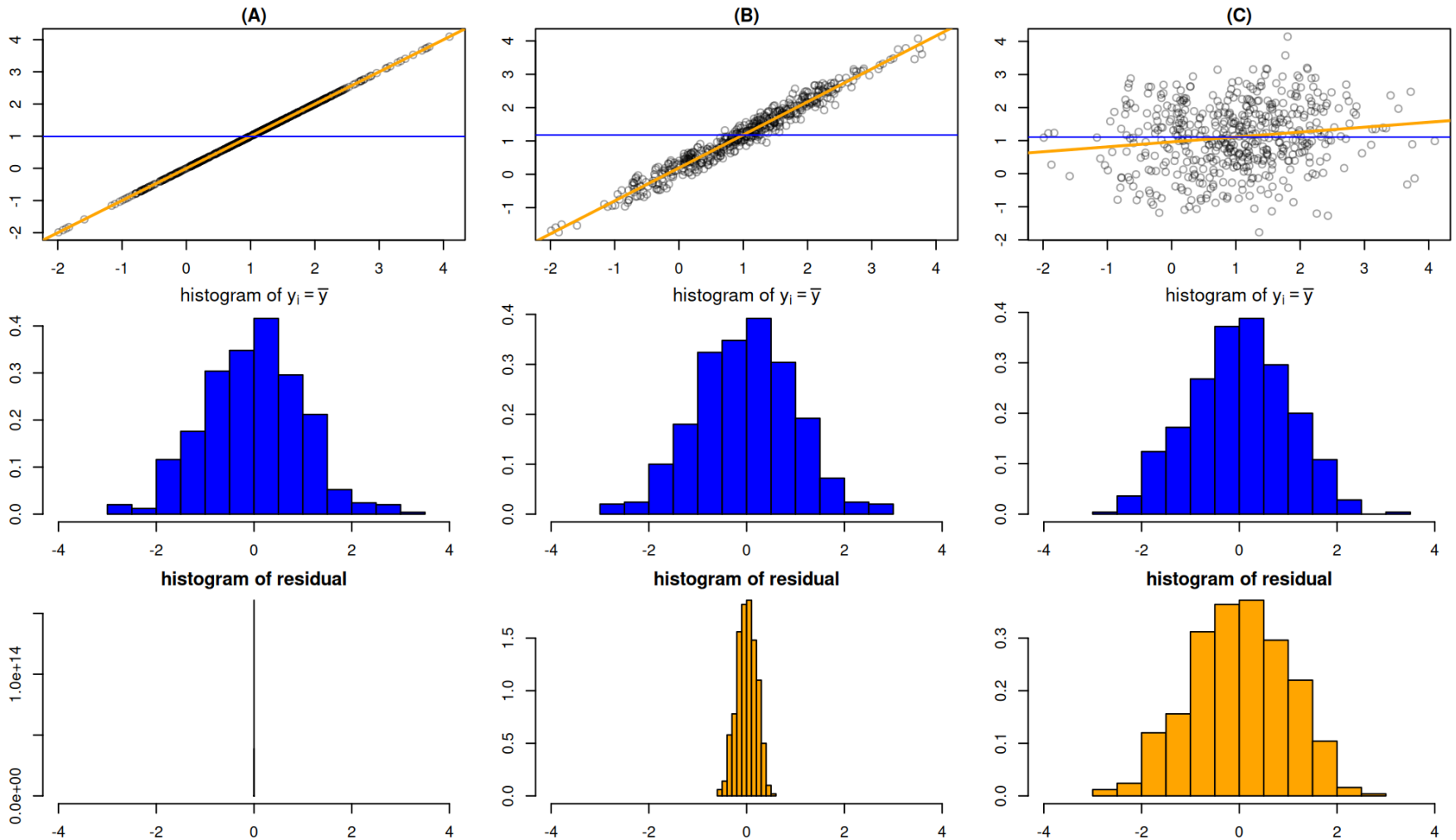
Summary of residual

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)	$\bar{y} - b\bar{x}$
Residual	$e_i = y_i - a - bx_i$

- $y = a + bx$ is the best line that minimises the sum of squared residuals $\sum_{i=1}^{n} e_i^2$.
- The average residual of the regression line is zero: $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.

Coefficient of determination

How much variability of data y can be explained by the linear model?



Baseline prediction/deviations in y, Regression line/residuals

Explaining variations

- The sum of squared residuals (or SSE for sum of squared errors)

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

  measures **variation in y left unexplained by the regression line**.

- Why?

$$\frac{1}{n-1}\text{SSE} = \text{SD}_e^2$$

  as the sample mean of e_i is zero.

- In (A) SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for (B), and large for (C).

- A quantitative measure of **the total amount of variation in observed y values** is given by the total sum of squares (sum of squared deviations about sample mean)

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  measures variation in y left unexplained by the baseline prediction.
- SST $\geq$ SSE.
- Why? The regression is optimal for sum of squared errors, so SSE (regression line) cannot be worse than SST (baseline).

Coefficient of determination

The ratio $\frac{\text{SSE}}{\text{SST}}$ is the proportion of total variation that cannot be explained by the simple linear regression model, and the coefficient of determination is

$$1 - \frac{\text{SSE}}{\text{SST}} = r^2$$

which is the **squared correlation coefficient** (a number between 0 and 1) giving the proportion of observed y variation explained by the model.

- The higher the value of r^2, the more successful is the simple linear regression model in explaining y variation.
- Note that if SSE = 0 as in case (A), then $r^2 = 1$.
- This can be verified using a, b, SDs, and r (see next lab)

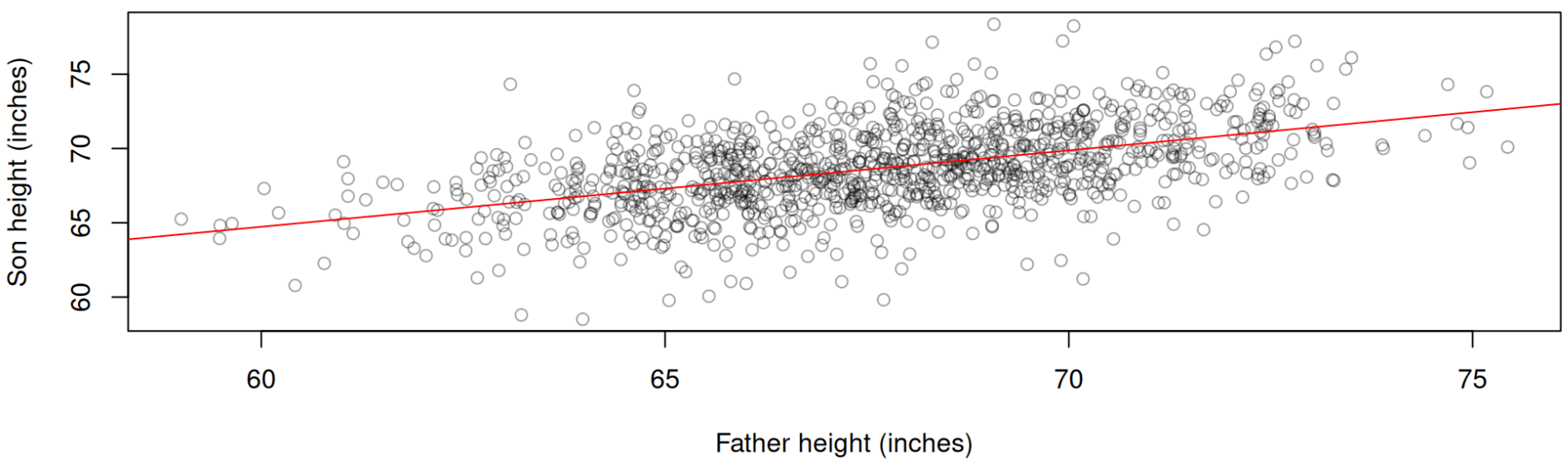
Example

```r
cor(x, y)^2  # quick way
```

```
[1] 0.2513401
```

```r
lm.fit ← lm(y ~ x)
SSE = sum(lm.fit$residuals^2)
SST = sum((y - mean(y))^2)
r2 = 1 - SSE/SST
```

```
[1] 0.2513401
```



The coefficient of determination for Pearson's height data is 0.25, about 25% of the variations in son's height can be explained by the regression line.

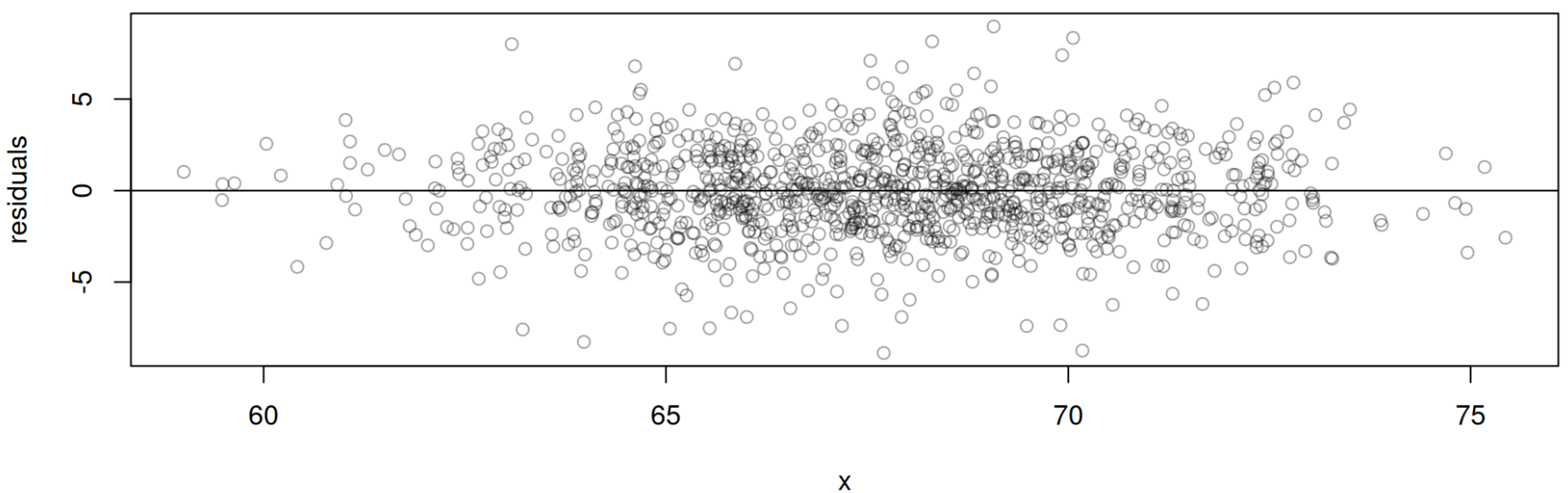
Diagnostics

Residual Plot

- A residual plot graphs the residuals vs x.
- If the linear fit is appropriate for the data, it should show no pattern (random points around 0).
- By checking the patterns of the residuals, the residual plot is a diagnostic plot to check the appropriateness of a linear model.

Residual plot

```r
plot(x, l$residuals, ylab = "residuals", col = adjustcolor("black", alpha.f = 0.35))
abline(h = 0)
```



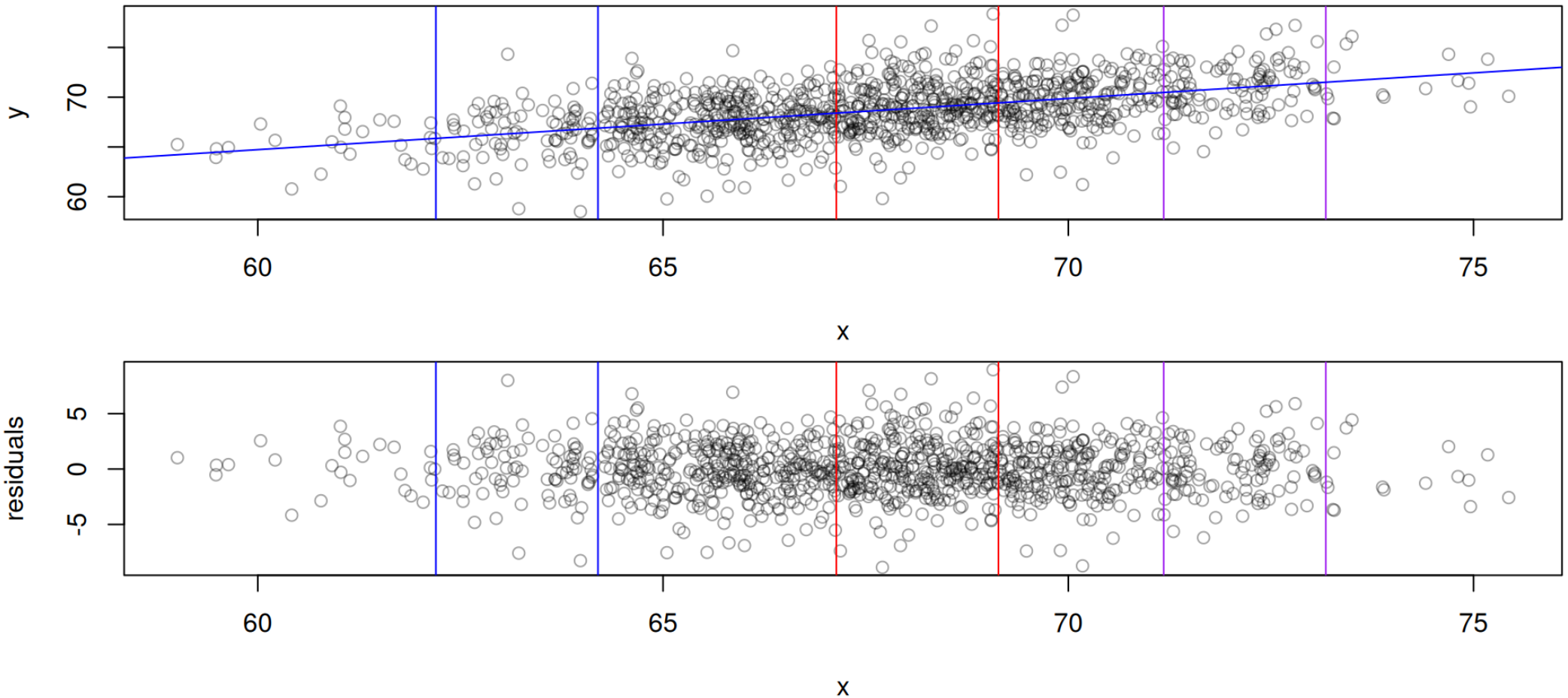
> **Note**
>
> Does this residual plot look random?

Homoscedasticity and Heteroscedasticity

Vertical strips

In linear models and regression analysis generally, we need to check the homogeneity of the spread of the response variable (or the residuals). We can divide the scatter plot or the residual plot into vertical strips.

- If the vertical strips on the scatter plot show equal spread in the y direction, then the data is **homoscedastic**.
  - ➟ The regression line could be used for predictions.
- If the vertical strips don't show equal spread in the y direction, then the data is **heteroscedastic**.
  - ➟ The regression line should not be used for predictions.

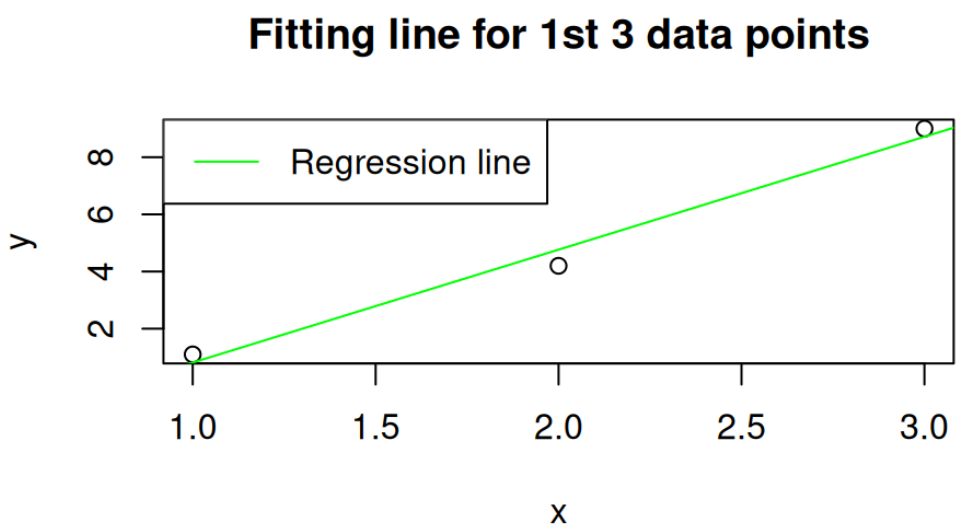


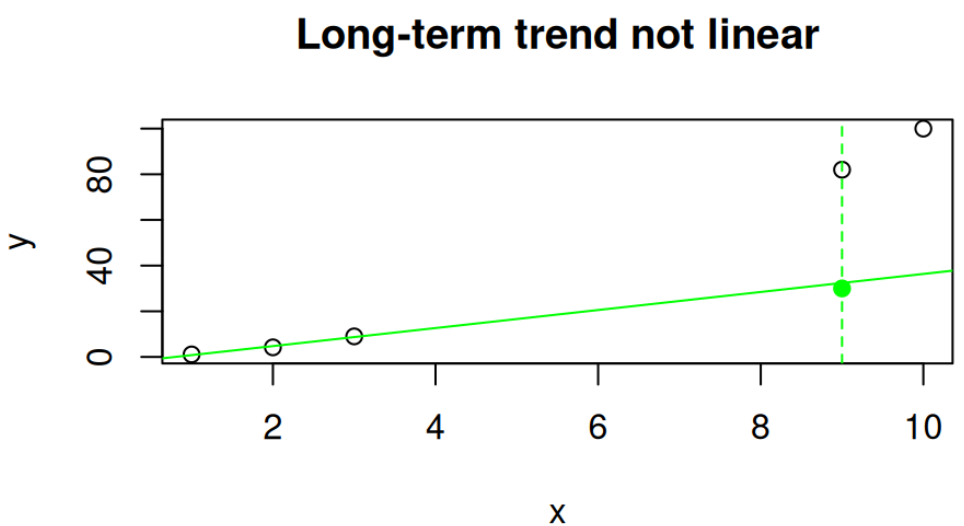
Note

Is the Pearson's height data homoscedastic?

Common mistake 1: Extrapolating

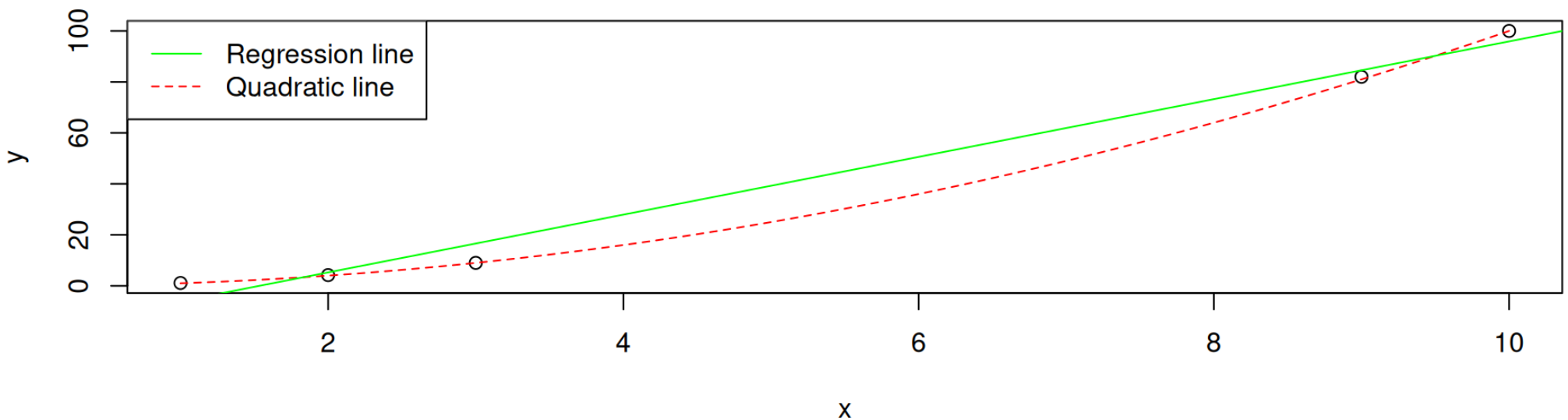
If we make a prediction from an x value that is not within the range of the data, then that prediction can be completely **unreliable**.





Common mistake 2: Not checking the scatter plot

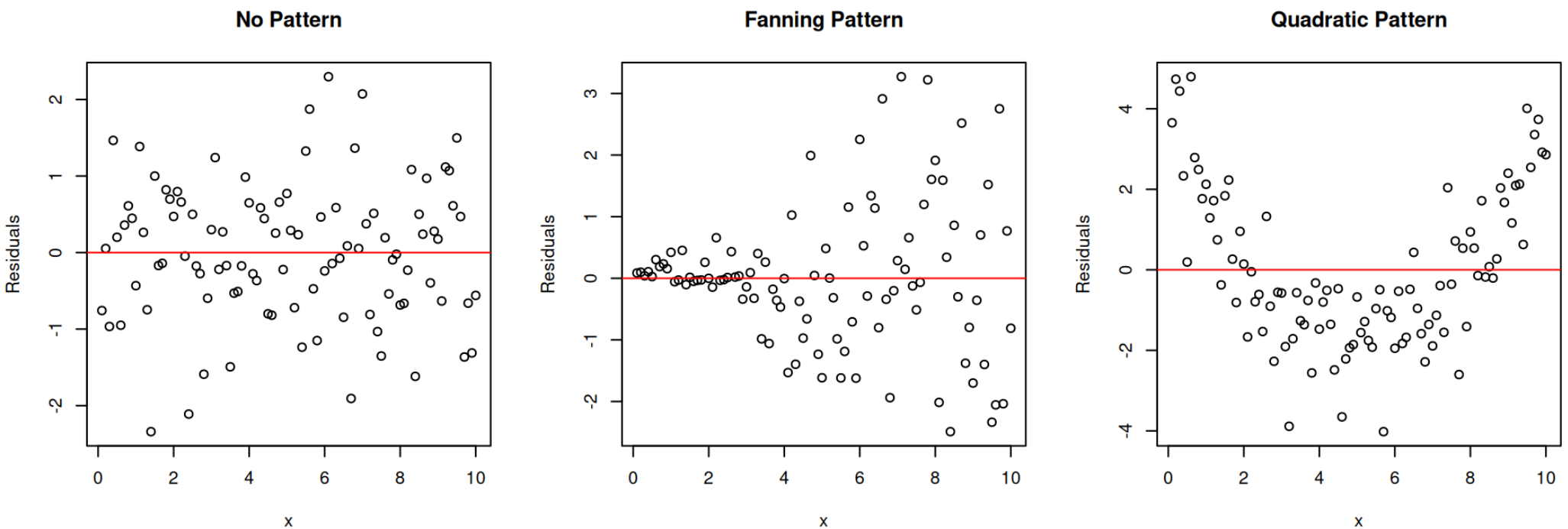
- We can have a high correlation coefficient and then fit a regression line, but the data may not even be linear!
- So always check the scatter plot first!



Note: Even though the correlation coefficient is high $r \approx 0.99$, a quadratic model is more appropriate than a linear model.

Common mistake 3. Not checking the residual plot

- You should also check the residual plot
- This detects any pattern that has not been captured by fitting a linear model.
- If the linear model is appropriate, the residual plot should be a random scatter of points.



Summary

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