Correlation and Linear Model

Modelling Data | Linear Model

STAT5002

The University of Sydney

Mar 2025

THE UNIVERSITY OF SYDNEY

Last lecture

Correlation

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

Linear model

- Regression Line
- Prediction

Code for plotting Pearson's data (you need the `UsingR` package):

```r
# install.packages('UsingR')
suppressMessages(library(UsingR))
library(UsingR)  # Loads another collection of datasets
data(father.son)  # This is Pearson's data.
data = father.son
x = data$fheight  # fathers' heights
y = data$sheight  # sons' heights
# scatter plot
plot(x, y, xlim = c(58, 80), ylim = c(58, 80), xaxt = "n", yaxt = "n", xaxs = "i",
    yaxs = "i", main = "Pearson's data", xlab = "Father height (inches)", ylab = "Son height (inches)")
# Adjust the gap between label and plot
axp = seq(58, 80, by = 2)
axis(1, at = axp, labels = axp)
axis(2, at = axp, labels = axp)
```

The (Pearson's) 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**.
- You don't have to remember the formula of r but we need to know how it is defined
    - ➠ the **mean** of the **product** of the variables in **standard units**.

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

Quick calculation in R using `cor()`.

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

```
[1] 0.5013383
```

Invariant properties

Shift and scale invariant

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

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

```
[1] 0.5013383
```

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

```
[1] 0.5013383
```

Symmetry (commutative)

The correlation coefficient is not affected by interchanging the variables.

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

```
[1] 0.5013383
```

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

```
[1] 0.5013383
```

Interpretations of r values

- The correlation coefficient r always takes values between -1 and 1 (inclusive).
- 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.

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: it minimises the **sum of squares** of the **residuals**.

In R:

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

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

Coefficients:
(Intercept)            x
    33.8866       0.5141
```

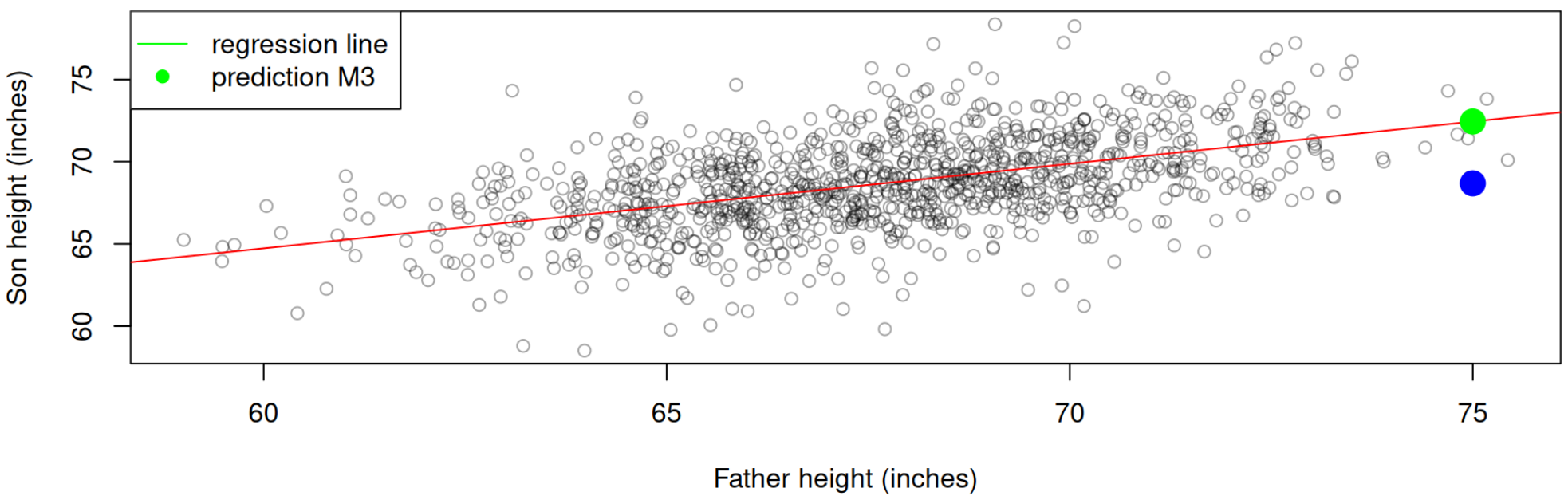
```r
model$coeff
```

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

Prediction using the Regression line

For new born (son), the father is 75 inches tall, how can we predict the son's height?

- Using 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.

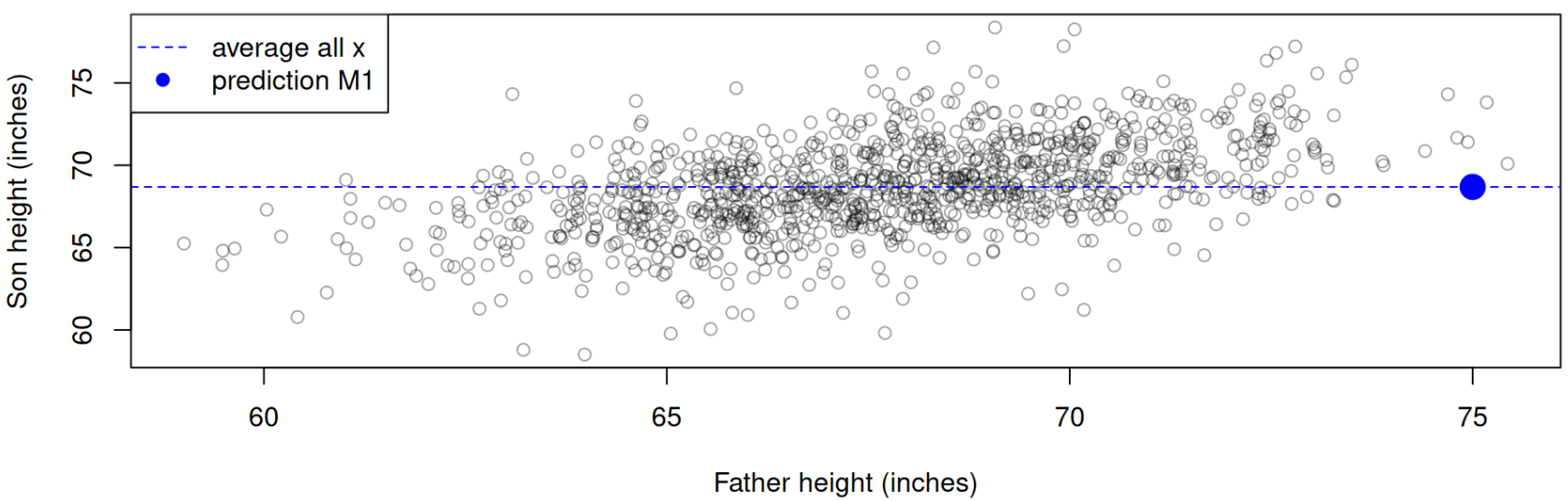


Baseline prediction

- 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.
- Later used for assessing the linear model as the reference model.

```
1 mean(y)
```

```
[1] 68.68407
```



Can we also use Y to predict X?

We can predict Y from X or X from Y, depending on what fits the context. However, we **cannot** just simply rearrange the equation

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

unless $r = \pm 1$ (so 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	$a = \bar{y} - b\bar{x}$	$\tilde{a} = \bar{x} - \tilde{b}\bar{y}$

Today

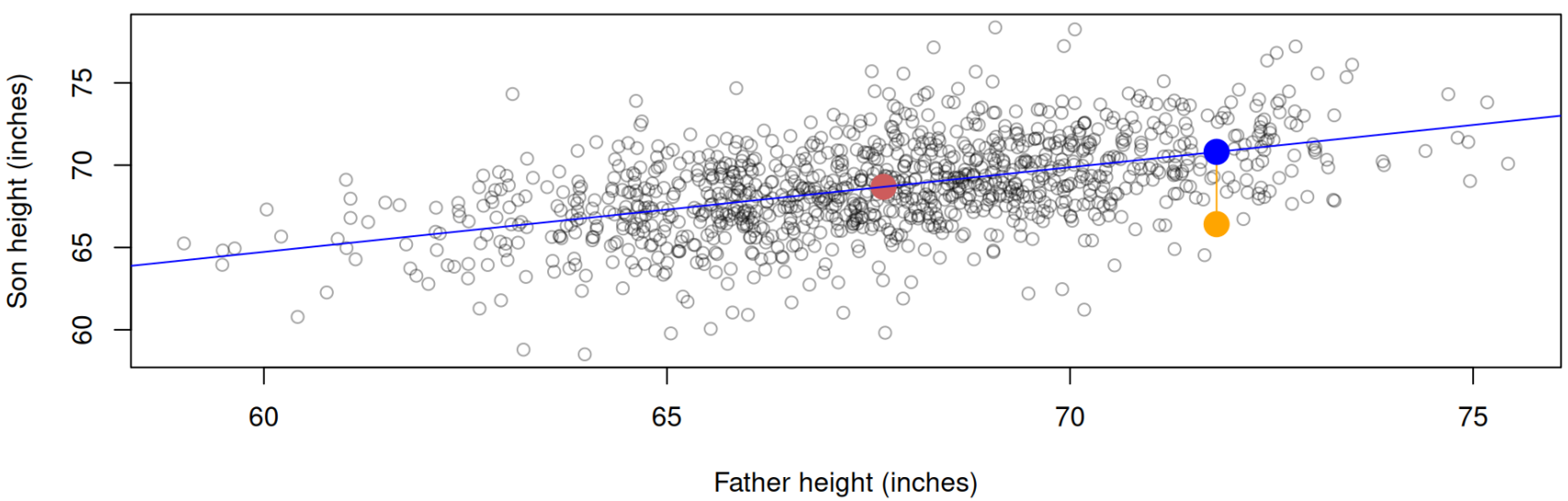
- Residuals and properties
- Coefficient of determination
- Diagnostics of model fit
- Case study

Residuals and properties

Residuals

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

- A **residual** (prediction error) 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 (39-th data point), 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 ($\hat{y}_i = a + bx_i$), a residual is

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

R code for obtaining residuals

```r
l = lm(y ~ x)
x[39]
```

```
[1] 71.81791
```

```r
y[39] - l$fitted.values[39]
```

```
       39
-4.390582
```

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

```
       39
-4.390582
```

Optimal linear model

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 for all possible lines
- see lecture notes for derivations (not for assessment)

Average of residual is zero

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

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

can be expressed as

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

Thus, **the sum and the mean of residual are zero.**

Summary of residual

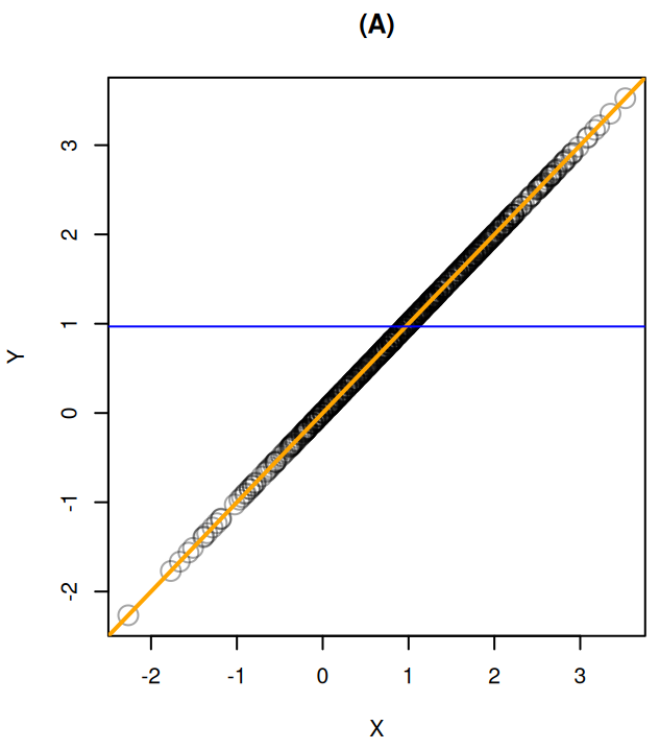
Feature	Regression Line $y \sim x$ ($y = a + bx$)
Connects	$(\bar{x}, \bar{y})$ to $(\bar{x} + \mathrm{SD}_x, \bar{y} + r\mathrm{SD}_y)$
Slope (b)	$r\frac{\mathrm{SD}_y}{\mathrm{SD}_x}$
Intercept (a)	$\bar{y} - b\bar{x}$
Residual	$e_i = y_i - a - bx_i$

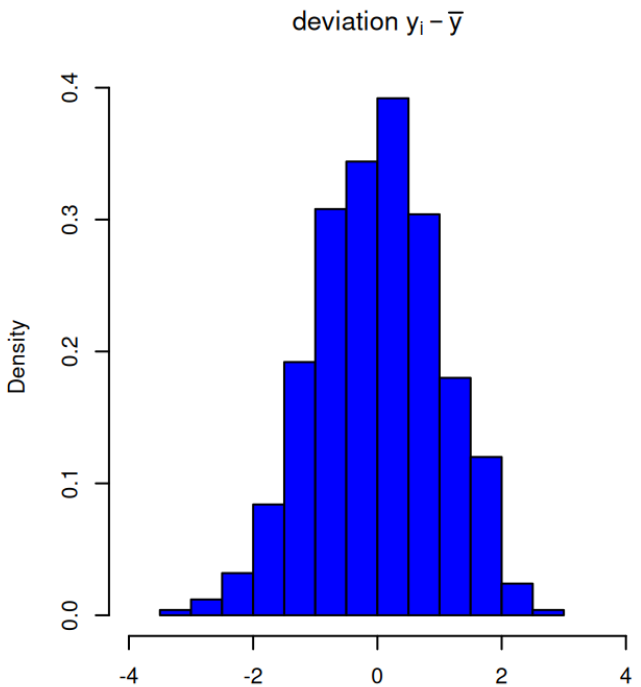
- $y = a + bx$ is the best line that minimises the sum of squared residuals $\sum_{i=1}^{n} e_i^2$.
- The average residual of the regression line is zero: $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.

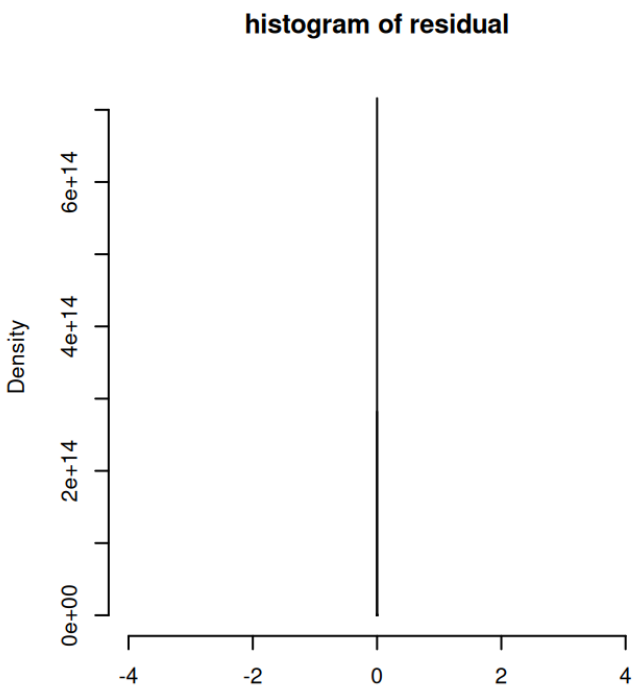
Coefficient of determination

How much variability in y can be explained by linear model?

Scenario (A)

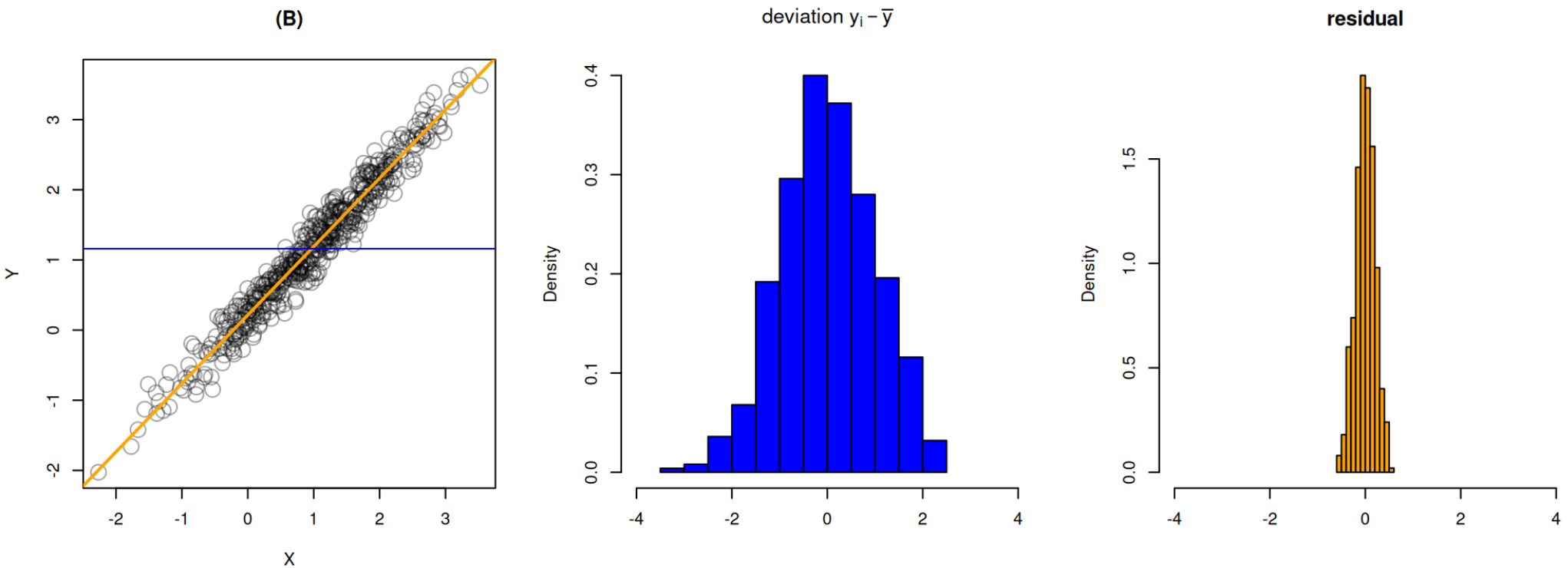






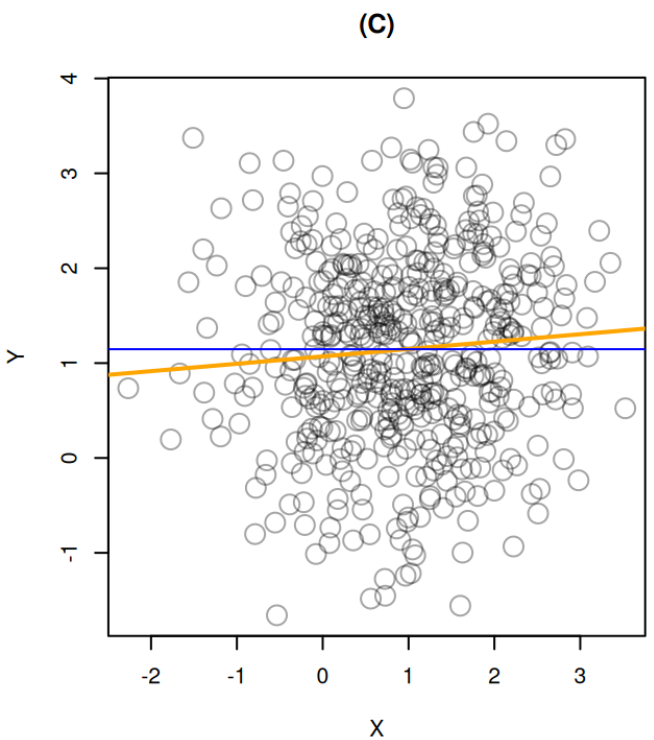
Baseline prediction/deviations in y, Regression line/residuals

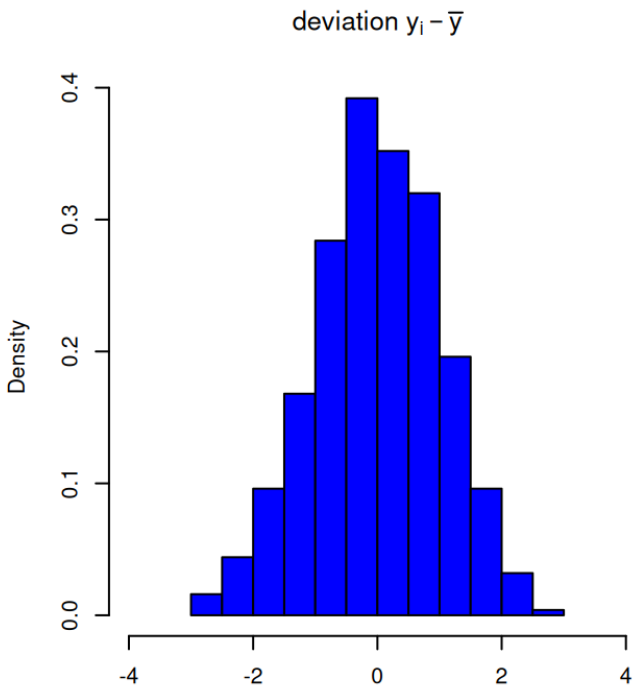
Scenario (B)

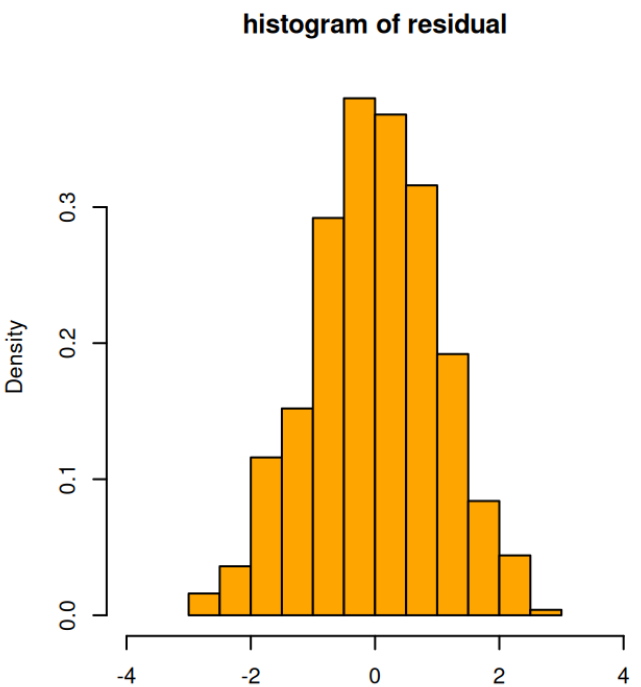


Baseline prediction/deviations in y, Regression line/residuals

Scenario (C)







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?

$$\frac{1}{n-1}\text{SSE} = \text{SD}(e)^2$$

  gives the sample variance of the residuals as the sample mean of e_i is zero.

- In those three scenarios
  - In (A) SSE = 0, and there is no unexplained variation
  - unexplained variation is small for (B)
  - unexplained variation is 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)

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

  measures variation in y left unexplained by the baseline prediction.

- Note that

$$\frac{1}{n-1}\mathrm{SST} = \mathrm{SD}(y)^2$$

  gives the sample variance of the dependent variable.

- 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

- $\frac{\text{SSE}}{\text{SST}}$ is the proportion of total variation that cannot be explained by the simple linear regression model compared to the baseline prediction
- So, the coefficient of determination given by

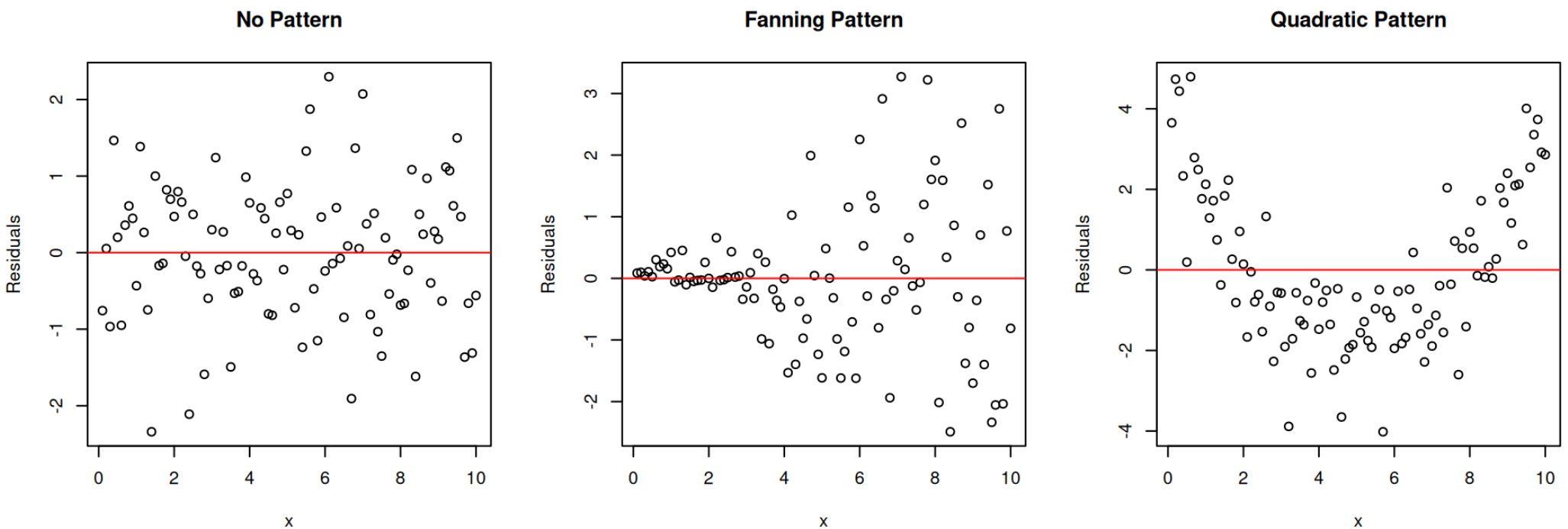
$$1 - \frac{\text{SSE}}{\text{SST}} = r^2$$

- It is exactly the **squared correlation coefficient** (a number between 0 and 1) giving the proportion of observed y variation explained by the model
  - ➠ This can be verified (not for assessment, see lecture notes for details)
- The higher the value of r^2, the more successful is the simple linear regression model in explaining variation in the depednent variable y.
- Note that if SSE = 0 as in case (A), then $r^2 = 1$.

Diagnostics

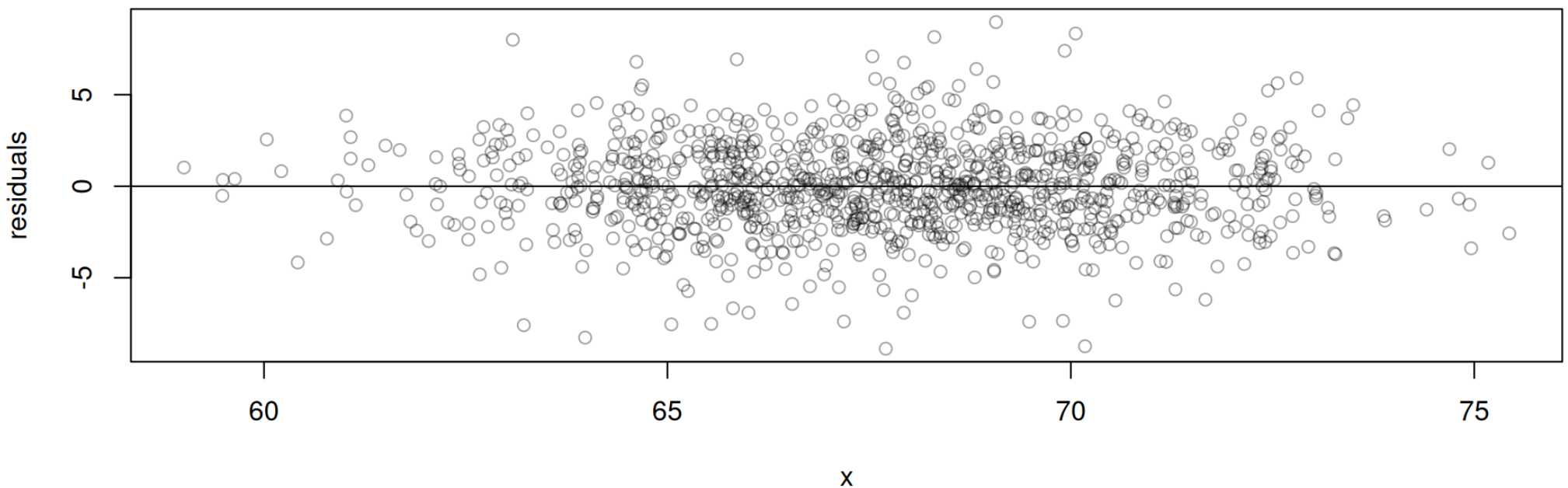
Residual plot

- A residual plot graphs the residuals vs the independent variable x.
- We should always 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.



Does the residual plot of the Pearson's data look random?

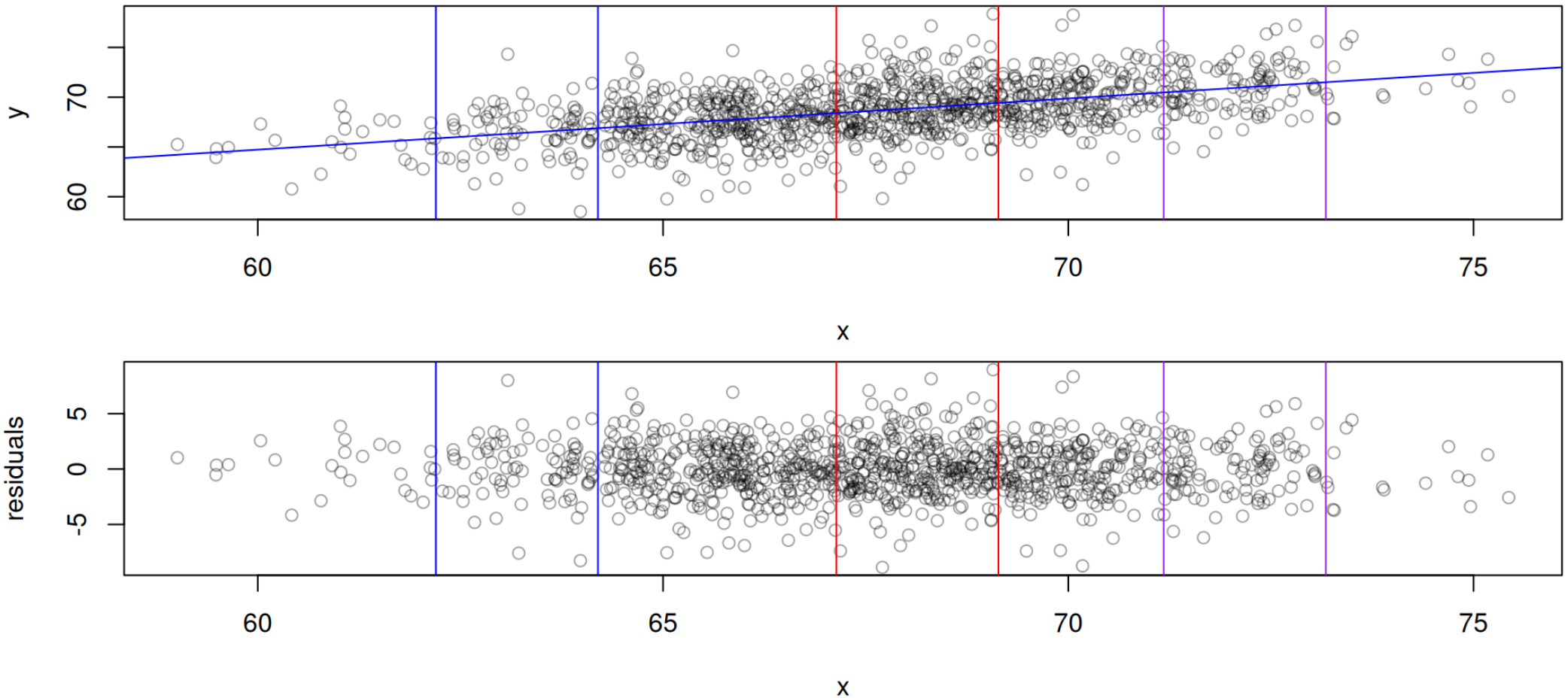
```r
plot(x, l$residuals, ylab = "residuals", xlab = "x", col = adjustcolor("black", alpha.f = 0.35))
abline(h = 0, main = "Pearson's data")
```



Homoscedasticity and Heteroscedasticity

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.
- We can visually check this using **vertical strips**



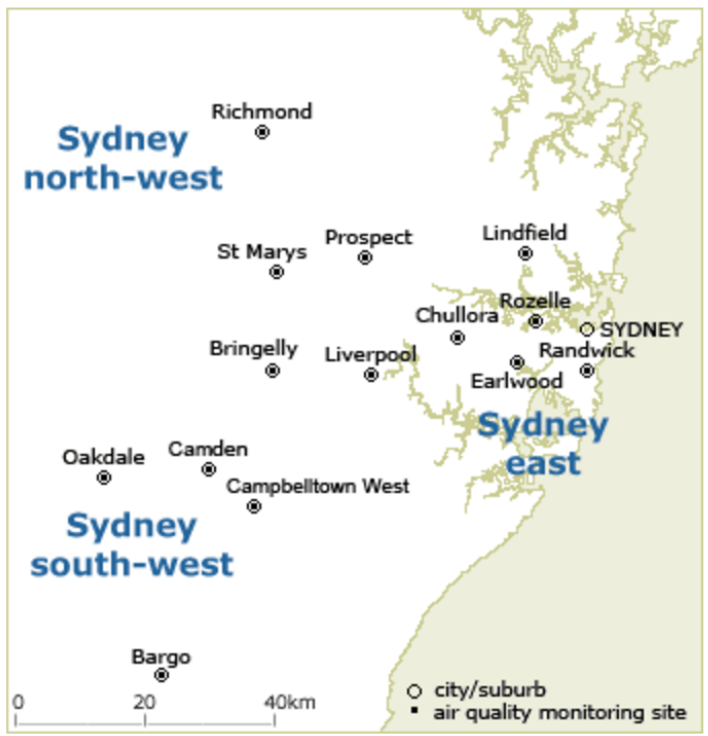
Note

Is the Pearson's height data homoscedastic?

Case study

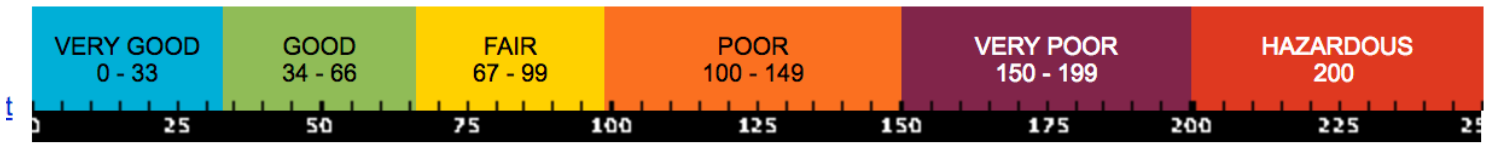
Air quality index (AQI) data

The Office of Environment and Heritage has 14 active sites to monitor the air quality of Sydney.



- At each site, data readings are taken for 6 variables:
  - ➠ Ozone (O_3), Nitrogen dioxide (NO_2), Visibility, Carbon monoxide (CO), Sulfur dioxide (SO_2), and Particles ($PM_{2.5}$, PM_{10}, etc.)

- They are combined into a single air quality index (AQI).



- Who is the AQI useful for?
  - Environmental scientists studying changes in air quality.
  - Potential renters and home-buyers.
- We will consider the **data** for July 2015 for two regions:
  - Sydney's central-east (CE)
  - Sydney's north-west (NW)

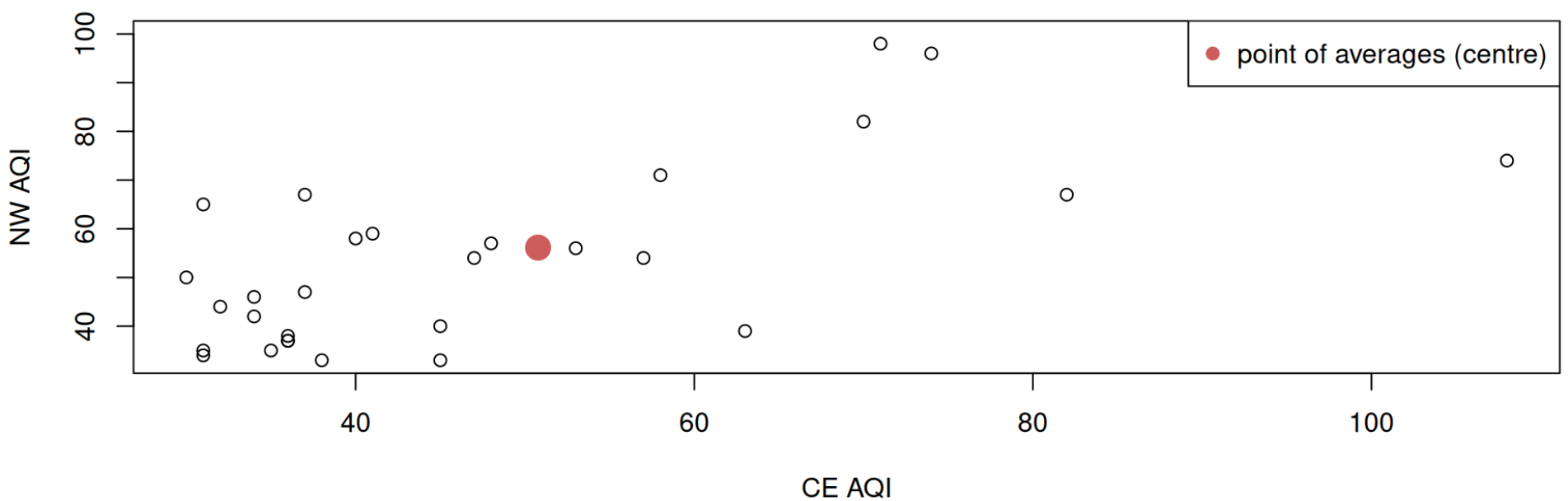
```
1 data = read.csv("data/AQI_July2015.csv")
```

```
1 head(data)
```

```
        Date SydneyCEAQI SydneyNWAQI
1 01/07/2015          99          92
2 02/07/2015          32          44
3 03/07/2015          70          82
4 04/07/2015          74          96
5 05/07/2015          95         100
6 06/07/2015          71          98
```

Scatter plot

```r
CE = data$SydneyCEAQI
NW = data$SydneyNWAQI
plot(CE, NW, xlab = "CE AQI", ylab = "NW AQI")
points(mean(CE), mean(NW), col = "indianred", pch = 19, cex = 2)  # point of averages (centre)
legend("topright", c("point of averages (centre)"), col = "indianred", pch = 19)
```



It appears that it is reasonable to assume that `CE` and `NW` has a linear association.

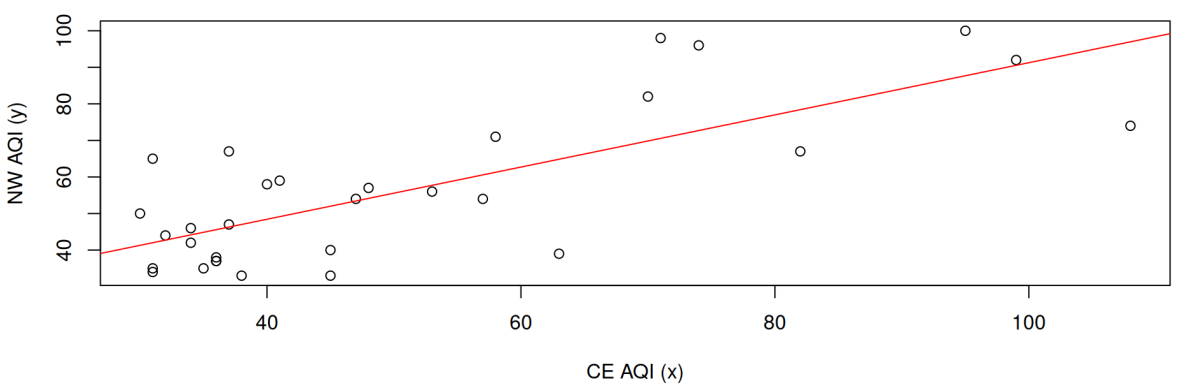
Correlation coefficient and regression line

```r
cor(CE, NW)
```

```
[1] 0.757917
```

What does the size of this correlation coefficient suggest about the data?

```r
model = lm(NW ~ CE)
plot(CE, NW, xlab = "CE AQI (x)", ylab = "NW AQI (y)")
abline(model, col = "red")
```



How much variability of `NW` can be explained by the linear model, compared to the baseline prediction?

```r
round(cor(CE, NW)^2 * 100, 2)
```

```
[1] 57.44
```

Prediction

Suppose the AQI at `NW` is not available on a particular day, how do we estimate its value using the AQI value at `CE` on that day (given the AQI at `CE` is 36)?

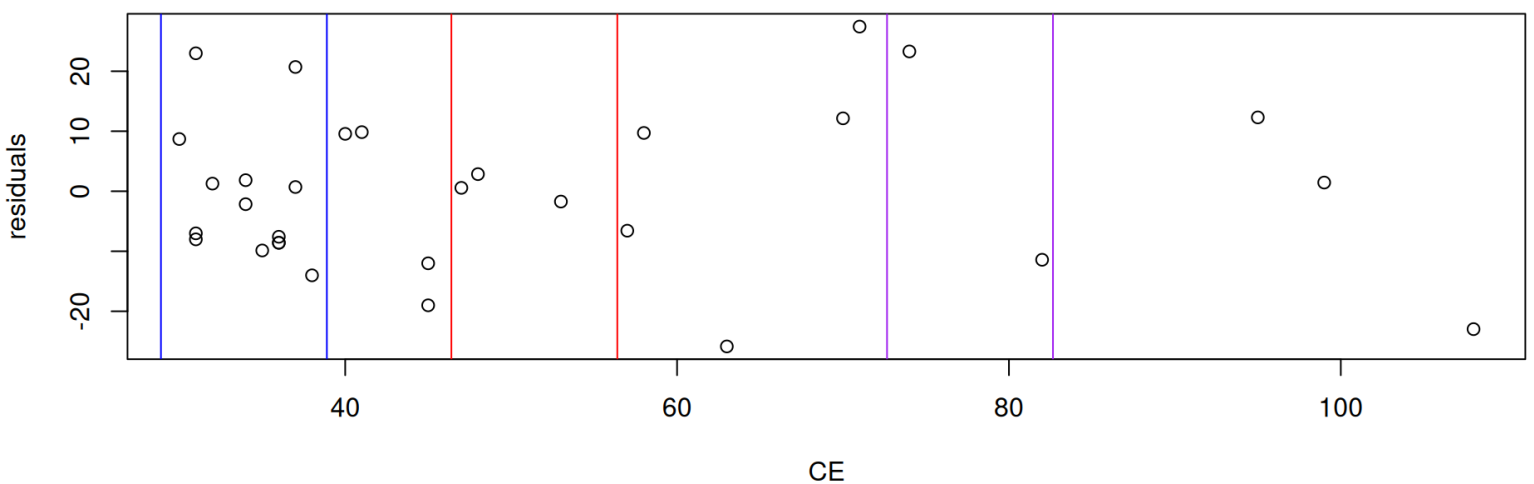
```
1 model$coefficient
```

```
(Intercept)          CE
 19.8873954   0.7137806
```

- For a `CE` reading of 36, we would predict the `NW` air quality to be $y = 19.8874 + 0.7138 \times 36 \approx 45.6$ (round to 1 decimal point).

Model check (residual plot)

```r
plot(CE, model$residuals, ylab = "residuals")
abline(v = mean(CE) - 1 * sd(CE), col = "blue")
abline(v = mean(CE) - 1 * sd(CE) + 10, col = "blue")
abline(v = mean(CE) - 0.2 * sd(CE), col = "red")
abline(v = mean(CE) - 0.2 * sd(CE) + 10, col = "red")
abline(v = mean(CE) + 1 * sd(CE), col = "purple")
abline(v = mean(CE) + 1 * sd(CE) + 10, col = "purple")
```



- Does the residual plot look random?
- Is the data/residual homoscedastic?

Additional warning

Association is not causation!

- Correlation measures association.
- Association does not necessarily mean causation.
  - ➠ A change of air quality at `CE` may not change the air quality at `NW`, and vice versa.
- Both variables may be simultaneously influenced by a other factors (confounder).
- We can use the linear model for prediction, but we need to be very careful in using it to explain causation.