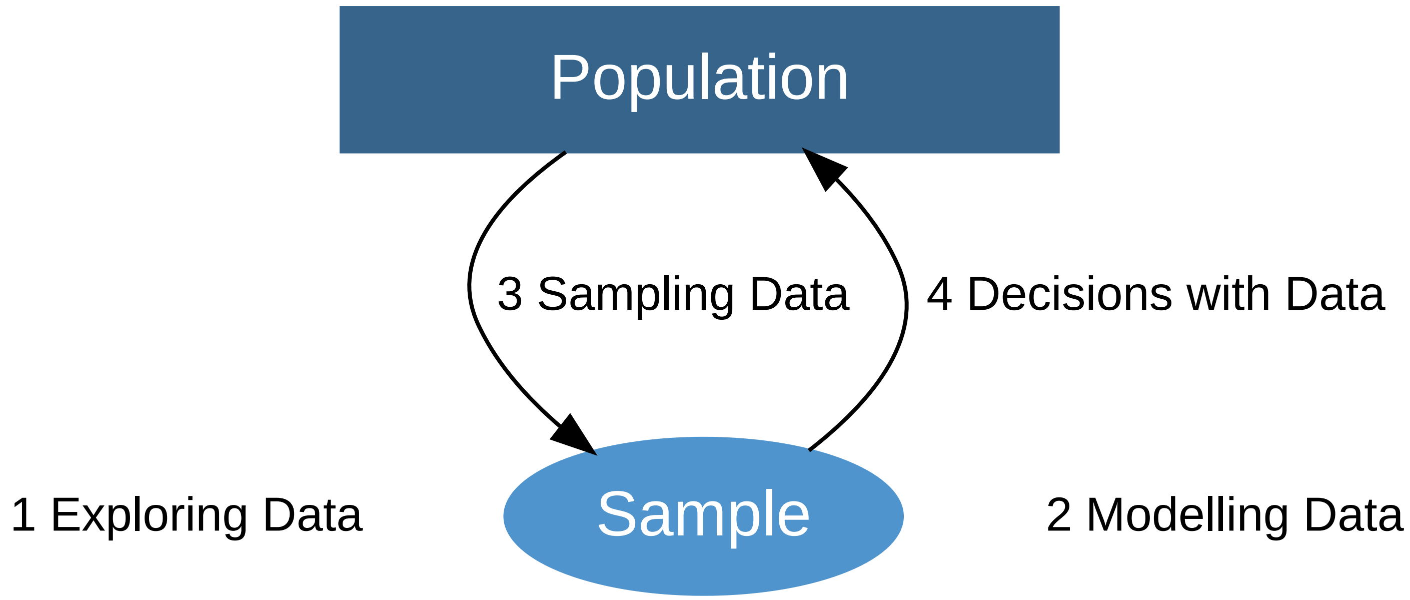
Residual Plot

Modelling Data | Linear Model

Unit Overview



Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?

Residual Plot

Data Story | How is the air quality in North-West Sydney related to Central-East Sydney?

Residuals

Residual plot

Vertical Strips

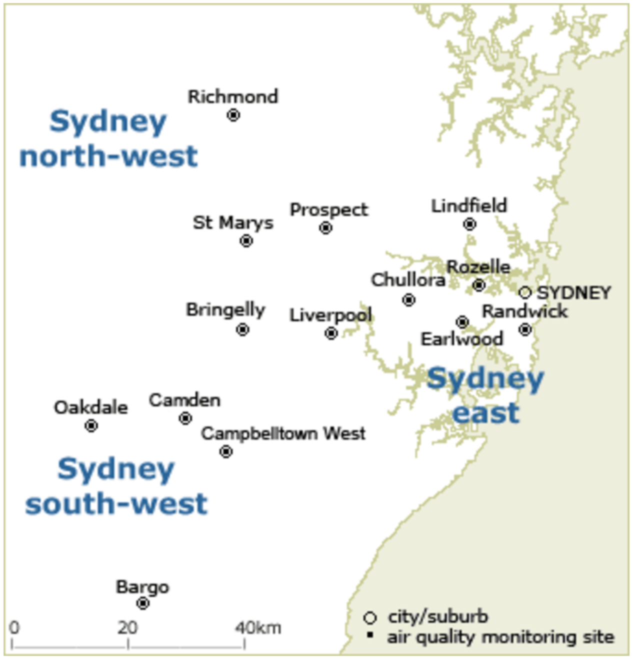
Summary

Data Story

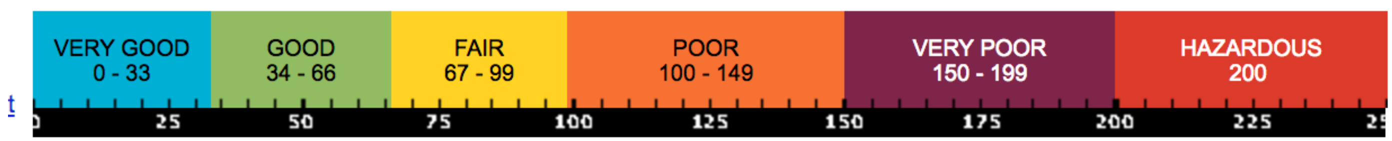
How is the air quality in North-West Sydney related to Central-East Sydney?

AQI data

- The Office of Environment and Heritage (OEH) monitors the air quality of Sydney through has 14 active monitoring sites.



- At each site, data readings are taken for 6 pollutants:
  - Ozone
  - Nitrogen dioxide
  - Visibility
  - Carbon monoxide
  - Sulfur dioxide
  - Particles
- There are combined into the air quality index (AQI).



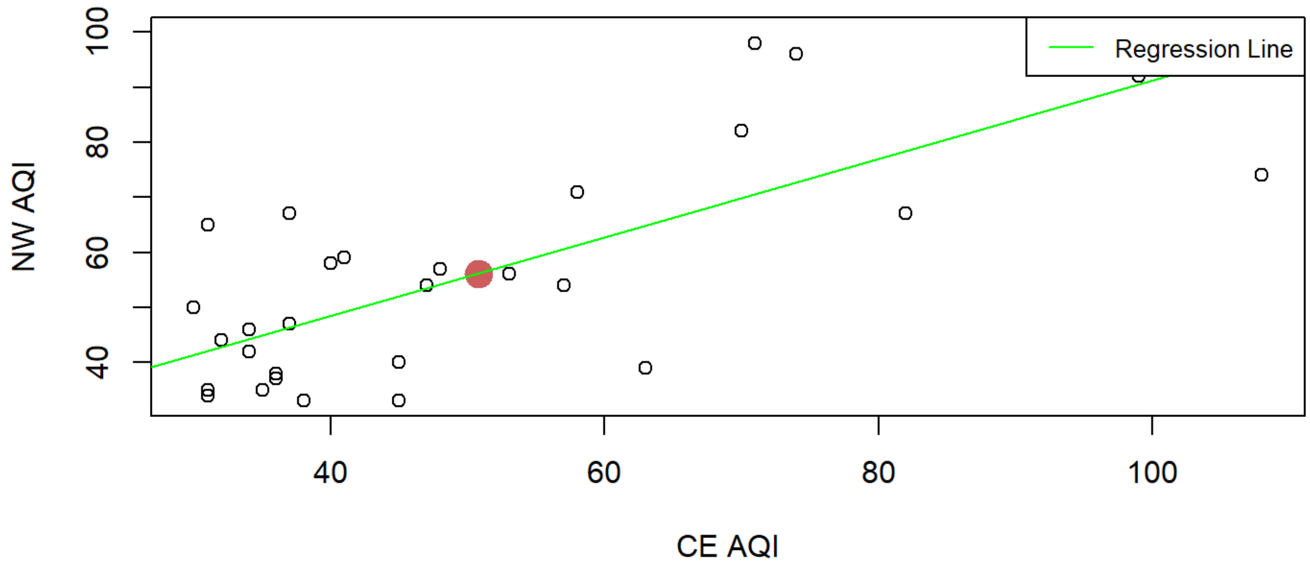
What is the AQI today?

AQI data

- We are considering data for the AQI for July 2015 for two regions:
    - Sydney’s central-east (CE)
    - Sydney’s north-west (NW)

```
data = read.csv("http://www.maths.usyd.edu.au/u/UG/JM/DATA1001/r/current/data/AQI_July2015.csv")
CE = data$SydneyCEAQI
NW = data$SydneyNWAQI
```

Prediction error?



We can now make predictions using the regression line. But what is the prediction **error**?

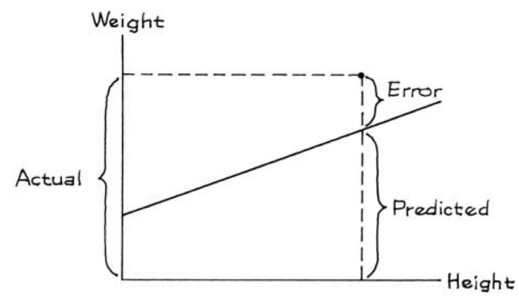
Residuals

Residuals

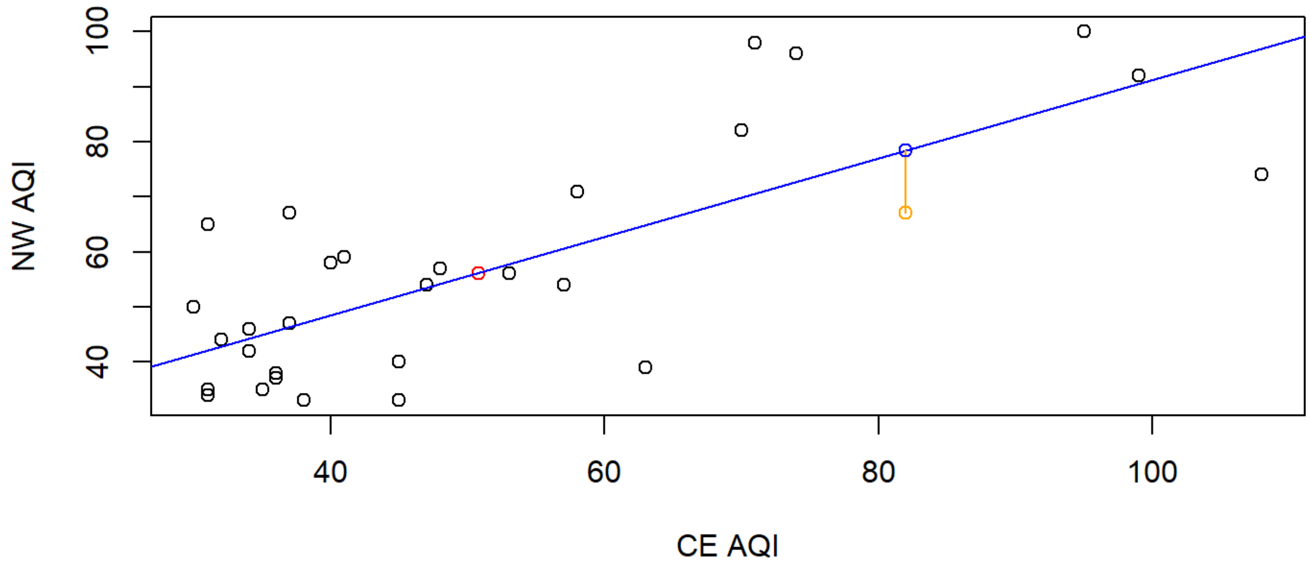
Residual (Prediction error)

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

Figure 2. Prediction error equals vertical distance from the line.



Statistics, Freedman et al p182



When the CE AQI is 82, the actual value of the NW AQI is 67 with predicted value 78.4, so the residual is -11.4.

More formally, a residual is $e_i = y_i - \hat{y}_i$, given the actual value (y_i) and the prediction ($\hat{y}_i$).

```
l = lm(NW ~ CE)
NW[10] - l$fitted.values[10]
```

```
##       10
## -11.41741
```

```
# Or directly
l$residuals[10]
```

```
##       10
## -11.41741
```

We can now calculate individual residuals, but can we summarise **all** the residuals?

The RMS Error

Population RMS error

- The RMS error represents the **average gap** between the points and the regression line.
- It is a clever average of the residuals.
- It is like a “standard deviation for the line”.

$$\text{RMS error}_{pop} = \text{RMS of (gaps from the line)} = \sqrt{\text{mean of (gaps)}^2}$$

More formally, $\text{RMS error}_{pop} = \sqrt{\frac{e_1^2 + e_2^2 + \ldots e_n^2}{n}}$.

```r
res = NW - l$fitted.values
sqrt(mean(res^2))  # RMS error (Pop)
```

```
## [1] 13.22338
```

Is this a big error for AQI?

The RMS error for Baseline Prediction

Population RMS Error for baseline prediction

- Baseline prediction uses $\bar{y}$ for every value of x.
- The RMS error for the baseline method is the standard deviation for y.

$$\text{RMS error}_{pop} = SD_y$$

As usual, there are 2 slightly different formulas for the baseline RMS Error, depending on whether your data is the population or a sample.

```
blres = NW - mean(NW)  # Baseline residuals
sqrt(mean(blres^2))  # RMS Error (Pop)
```

```
## [1] 20.27034
```

```
sd(NW)  # RMS (Sample), as sd() in R is for sample
```

```
## [1] 20.60541
```

```
# Adjustment to make into RMS Error (Pop)
sd(NW) * sqrt((length(CE) - 1)/(length(CE)))
```

```
## [1] 20.27034
```

Quick formula for the RMS Error

Population RMS Error (quick formula)

$$\text{RMS error}_{pop} = \sqrt{1 - r^2}\, SD_y$$

```
sqrt(1 - (cor(CE, NW))^2) * sd(NW)  # RMS Error (Sample)
```

```
## [1] 13.44196
```

```
sqrt(1 - (cor(CE, NW))^2) * sd(NW) * sqrt((length(CE) - 1)/(length(CE)))  # RMS Error (Pop)
```

```
## [1] 13.22338
```

Special Cases

Perfect correlation (line): $r = \pm 1$

RMS error = 0, as the points lie on a line

$r = 0$

RMS error = SD_y, as the regression line is no help in predicting y.

Smallest RMS error

The smallest possible RMS error is for the regression line.

Summary: Populations and Samples

As usual, for simplicity in what follows, we will assume our data is a sample, and use the quick formula for the RMS Error using R.

Summary	In R
Population RMS Error: RMS Error_{pop}	
Sample RMS Error: $\text{RMS Error}_{sample}$	`sqrt(1-(cor(x,y))^2)*sd(y)`

```
sqrt(1 - (cor(CE, NW))^2) * sd(NW)  # RMS Error (Sample)
```

```
## [1] 13.44196
```

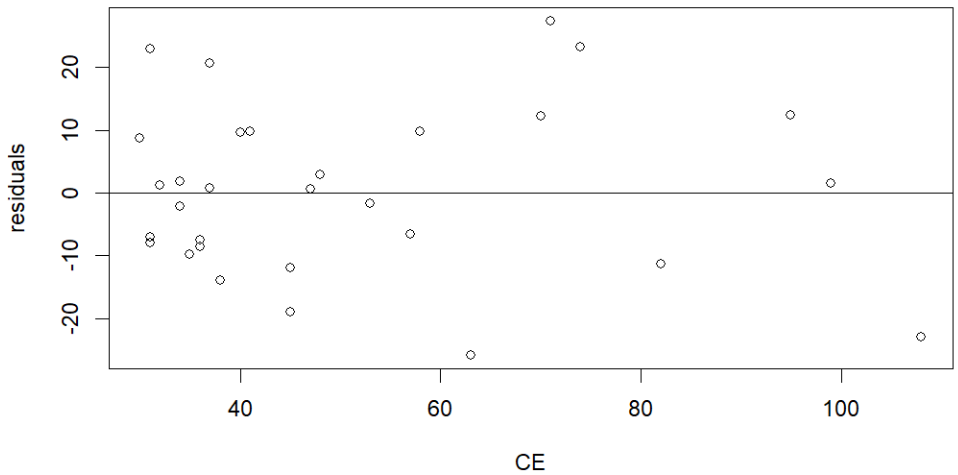
Residual Plot

Residual Plot

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 about 0).

```
plot(CE, l$residuals, ylab = "residuals")
abline(h = 0)
```



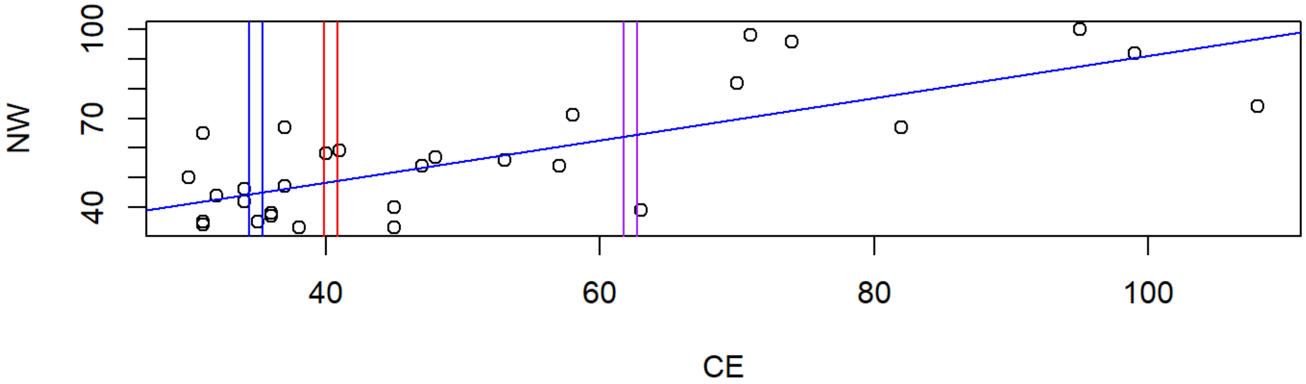
Does this residual plot look random?

Vertical Strips

Vertical Strips

Vertical strips

- If the vertical strips on the scatter plot show equal spread in the y direction, then the data is **homoscedastic**.
  - The RMS Error can be used as a measure of spread for individual strips.
- If the vertical strips don't show equal spread in the y direction, then the data is **heteroscedastic**.
  - The RMS Error can't be used as a mesaure of spread for individual strips.



Is the AQI data homoscedastic?

The Normal distribution within strips

Normal distribution within vertical strips

- If the data is homoscedastic, then we can use the Normal approximation within the vertical strips.
- We consider the y values within the strip as a new data set y^* with

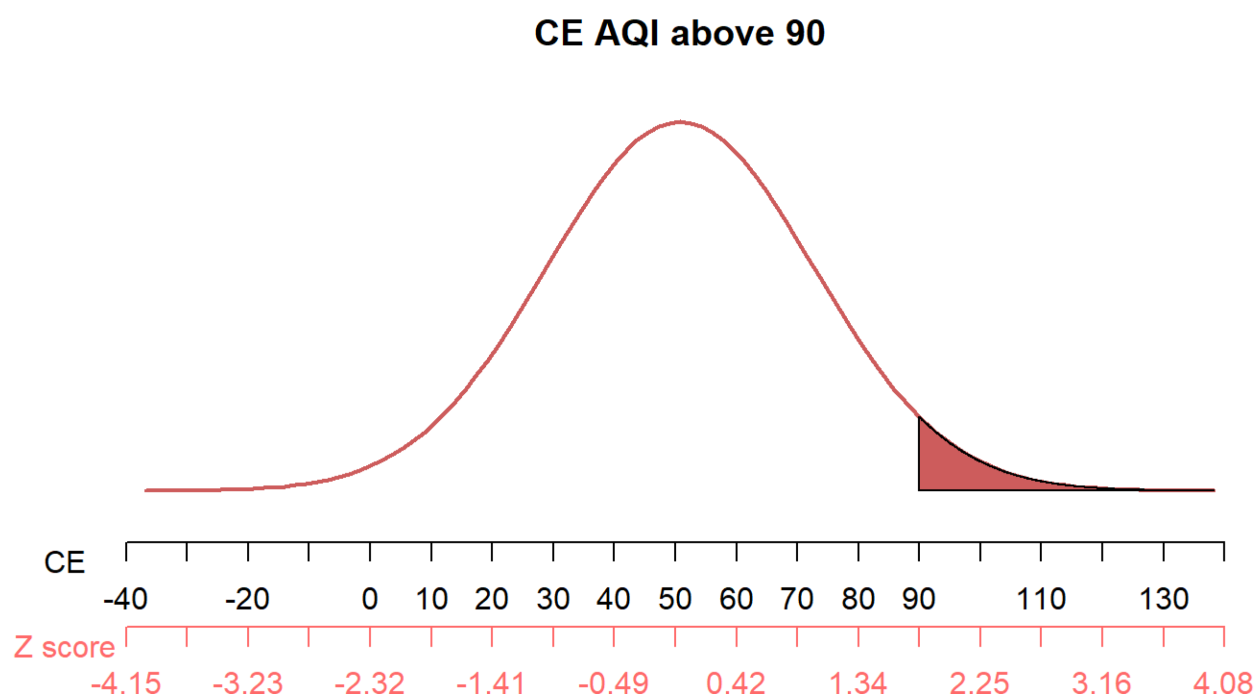
$$\bar{y}^* = \bar{y} + z_x r SD_y$$

$$SD_y^* \approx \text{RMS Error}$$

where z_x is the z-score for the strip.

Example 1

On what percentage of days, was the CE AQI above 90?



- The mean of CE is 50.77 and the SD is 21.88.

```
# Mean & SD
c(mean(CE), sd(CE))
```

```
## [1] 50.77419 21.87953
```

- The z score is $z_x = \frac{90-50.77}{21.88}$ = 1.79.
- So percentage of days in which the CE AQI was above 90 is approximately 0.04.

```
pnorm(1.79, lower.tail = F)
```

```
## [1] 0.03672696
```

```r
# Actual data
length(which((CE > 90)))/length(CE)
```

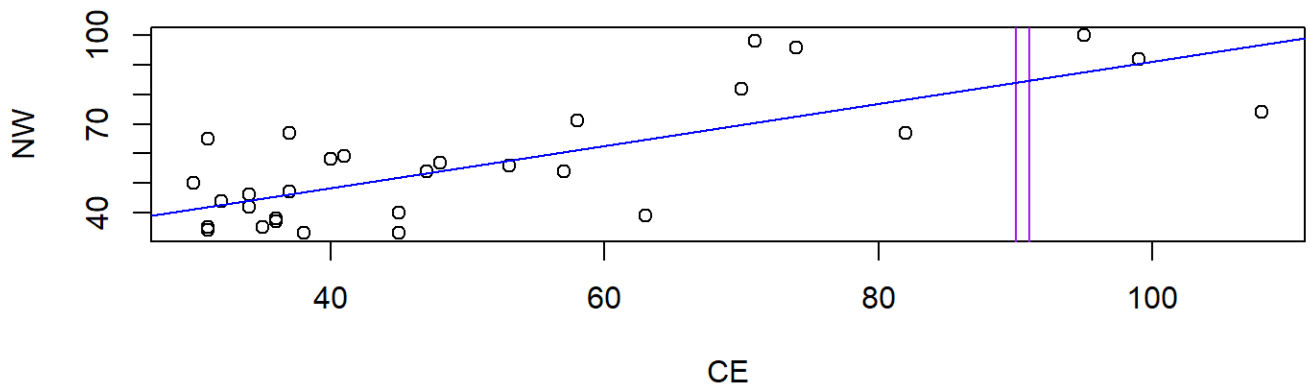
```
## [1] 0.09677419
```

```r
# Normal approximation
z = (90 - mean(CE))/sd(CE)
1 - pnorm(z)
```

```
## [1] 0.0365018
```

Example 2 (vertical strip)

Of the days where the CE AQI was equal to 90, what percentage of days had the NW AQI above 95?



Consider the vertical strip where CE is equal 90: y^*.

$$\bar{y}^* = \bar{y} + z_x r SD_y = 56.12903 + 1.79 \times 0.757917 \times 20.6 = 84.07646$$

$$SD_y^* \approx RMS = 13.44196$$

```
c(mean(NW), sd(NW), cor(CE, NW))
```

```
## [1] 56.129032 20.605407  0.757917
```

Note we are using value for $\text{RMS Error}_{sample}$ here.

Now work out the z score in this vertical strip.

- The z score is $z = \frac{95 - 84.07646}{13.44196} = 0.8126485$.
- So percentage of days which had NW above 90 is approximately 0.21.

```
pnorm(0.8126485, lower.tail = F)
```

```
## [1] 0.2082098
```

Summary

- For a Regression line, the **residuals** are the gaps between the **actual value** and the **prediction**.
- The prediction error for the Regression line is the RMS of the gaps from the line.

$$\text{RMS Error}_{pop} = \sqrt{1 - r^2}\,\text{SD}_y$$

- For baseline prediction

$$\text{RMS Error}_{pop} = \text{SD}_y$$

- The residual plot is a diagnostic for seeing whether a linear model was appropriate - if it is random, then linear model seems appropriate.

- If the vertical strips on the scatter plot show **equal spread** in the y direction, then the data is **homoscedastic**, and the RMS can be used as a measure of spread for individual strips.
- If the data is homoscedastic, then we can use the **Normal approximation** within the vertical strips.

Key Words

prediction error, residuals, RMS Error, baseline prediction, residual plot, vertical strips, homoscedastic, heteroscedastic