# 228.371 - Statistical Modelling for Engineers and Technologists

## Week 2. Regression
## Fitting Equations To Data

Dr. Daniel Walsh
IIMS 3.07 x 41032
d.c.walsh@massey.ac.nz

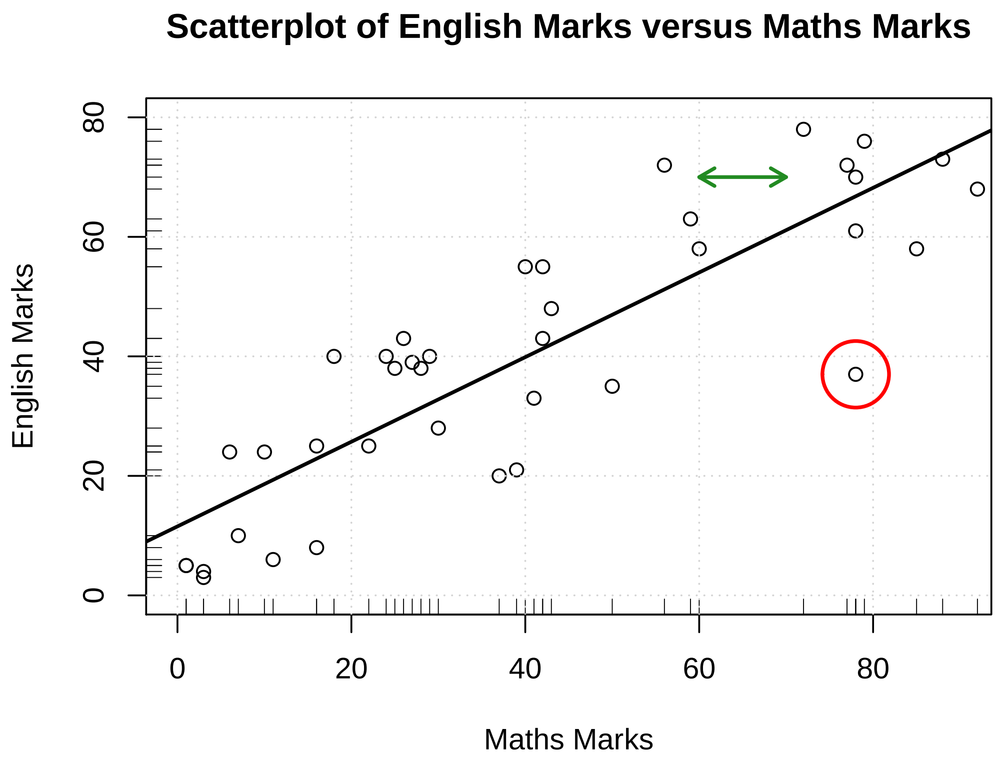
Semester One - 2015

# Scatterplot: Introduction

- Consider quantitative data that are pairs $(x,\ y)$.
- The two variables need not be in the same units, but related in some way.

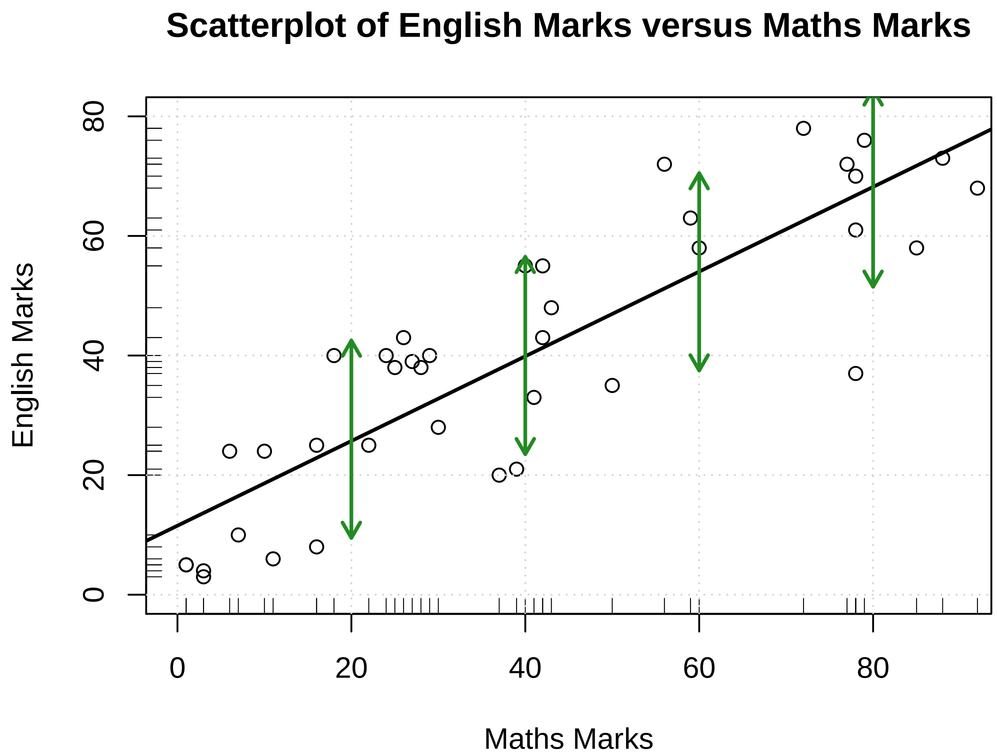
- Plot paired data $(x,\ y)$ - called a scatterplot. `plot(x,y)`
- The basic objective is to see whether a relationship exists between $x$ and $y$.
- Particularly, check for **Trends**, **Gaps**, **Outliers** etc, and whether the relationship is **LINEAR** (straight line fit).

# Scatterplot: `textmarks.txt`



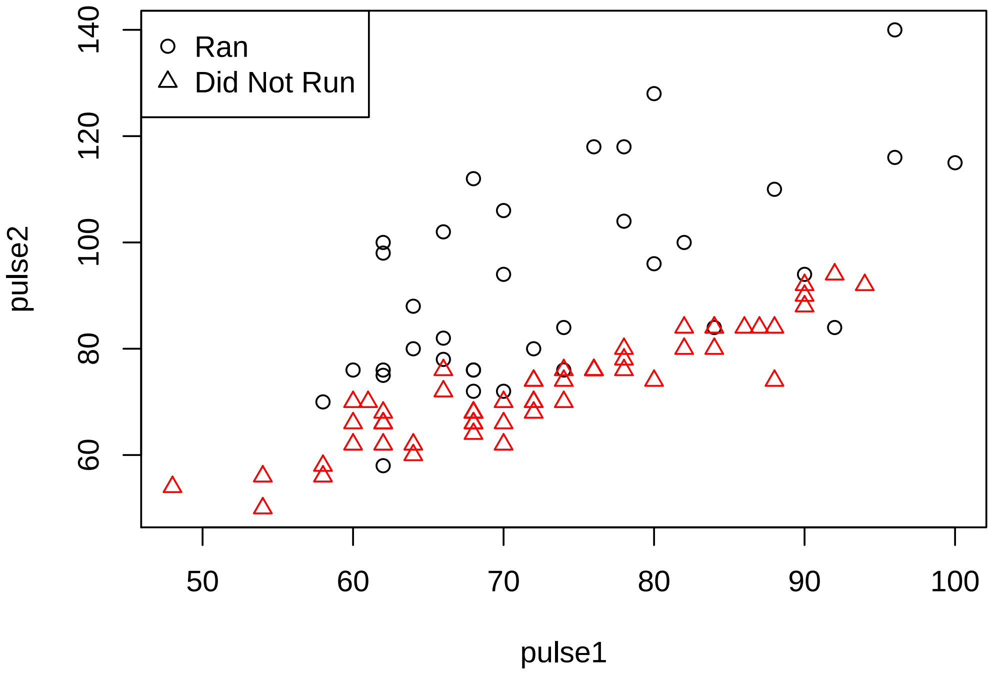

- **Trend** - positive, linear.
- Possible **gap** in Maths marks between 60 and 70.
- **Outliers** - none in marginals, but what about the circled point?

# Scatterplot: `textmarks.txt`




- **Variability in** $y$ is constant as $x$ changes (i.e. vertical scatter about trend).
- Can also use boxplots examine marginal distribution of each variable.

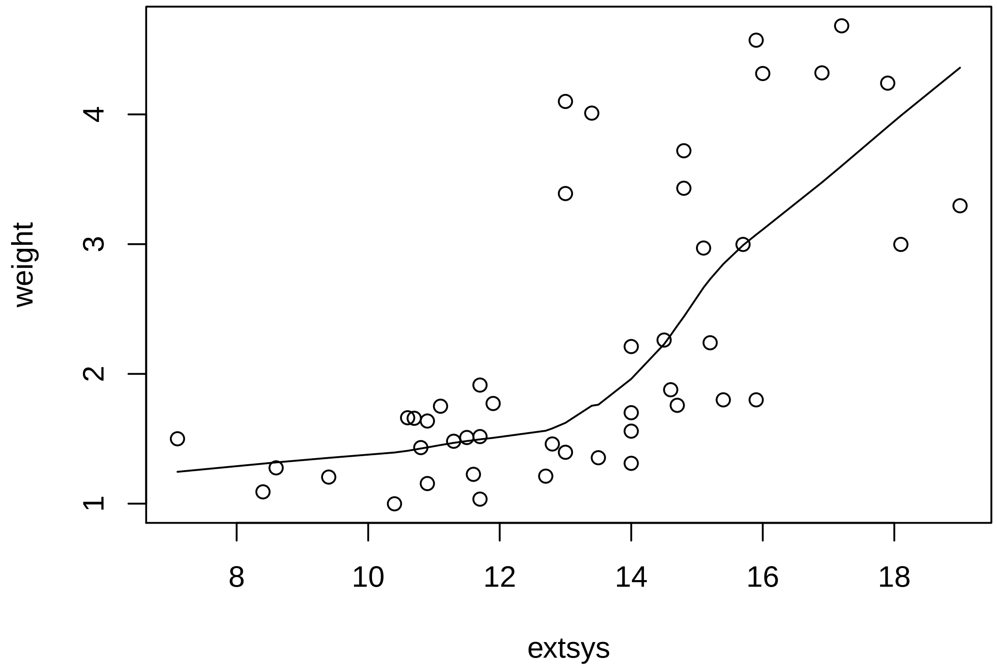
# Displaying Groups: `pulse.txt`




```
pulse <- read.table("Data/pulse.txt", header=TRUE)
attach(pulse)
plot(pulse1, pulse2, pch=ran, col=ran)
legend("topleft", c("Ran", "Did Not Run"), pch=1:2)
```

**Groups:** can use different plotting symbols to show groups.

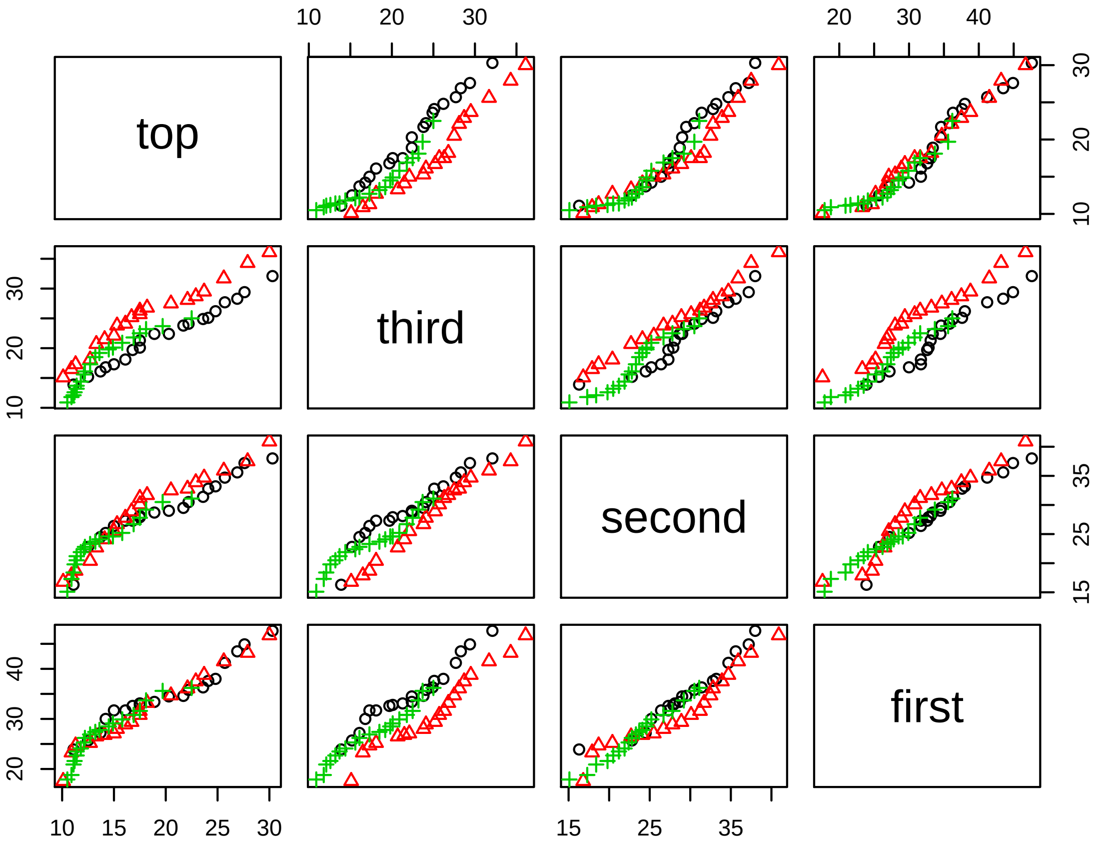
# Lowess smoother: `horsehearts.txt`




```
horsehearts <- read.table("Data/horsehearts.txt", header=TRUE)
attach(horsehearts)

plot(weight ~ extsys)
lines(lowess(weight ~ extsys))
```

**Lowess smoother:** helps to identify trend (line or curve).

# Scatterplot matrix: `pines.txt`




```
pines <- read.table("Data/pines.txt", header=TRUE)
pairs(pines[2:5], col=pines$area, pch=pines$area)
head(pines,1)
```

```
  area  top third second first
1    1 11.1  13.9   16.3  23.9
```

- Shows scatterplots of each pair of variables.

# Correlation coefficient

Correlation coefficient $r_{xy}$ measures the **strength of a linear relationship** between two variables $X$ and $Y$

**Properties:**

$$-1 \leqslant r_{xy} \leqslant 1$$

| | |
|---|---|
| $r_{xy} = 0$ | nonlinear relationship between $X$ and $Y$ |
| $r_{xy} = 1$ | perfect linear relationship between $X$ and $Y$ (points lie on a straight line; sign indicates sign of slope) |

See Study Guide for calculation formula for $r$ and scatterplots for different values of $r$

# Correlation coefficient

```
 cor (pines[2:5])
```

```
             top     third    second     first
top    1.0000000 0.9165563 0.9551619 0.9724647
third  0.9165563 1.0000000 0.9467956 0.9083708
second 0.9551619 0.9467956 1.0000000 0.9669647
first  0.9724647 0.9083708 0.9669647 1.0000000
```

```
 cor.test (pines$top, pines$first)
```

```
Pearson's product-moment correlation

data:  pines$top and pines$first
t = 31.7789, df = 58, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9541512 0.9835248
sample estimates:
      cor
0.9724647
```

# Our Aim

- An **equation** (mathematical model) describing relationship between a **response** variable and one or more **explanatory** variables.
- Fitting the model, estimating the unknown coefficients (parameters) in the model.
- Our model may represent a straight line, or a curved function (using polynomial functions for example).

# Terminology

- $y$ response variable (dependent).
- $x_1, x_2, \ldots x_p$ explanatory variables / predictors / covariates / regressor variables (independent).
- Regression Models.
  - We will only consider $y \sim N(\mu, \sigma^2)$ (Normally/Gaussian distribution - mean,variance).
  - $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ or equivalently
  - $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

# Mechanistic Models

$$V = IR$$

- Ohm's Law is an example of a mechanistic model.
- In a mechanistic model, the form of the relationship is known.
- If $y =$voltage and $x =$current are known, and resistance ($\beta$) is unknown, then:

$$y = \beta x$$

- Fitting the model involves estimating $\beta$.

# Empirical Models

- If there is no pre-conceived notion of the form of the relationship find an **empirical** model.
- Insight into the underlying physical mechanism?
- Predict the response as accurately as possible, e.g. instrument calibration curves.
- Usually try linear model first, then more complicated models.

# Errors

- Response and explanatory variables rarely satisfy a mathematical equation exactly.
- Experimental situation:
  - Measurement errors.
  - Additional unrecorded factors.
- Observational data:
  - Even more unrecorded factors.

# Linear Models

- In a **linear model**, the mean response is **linear in the parameters**, e.g.

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \epsilon \\
y &= \beta_1 \sin(x_1) + \beta_2 \log(x_2/x_3) + \beta_3 x_4 x_5 + \epsilon \\
\log(y) &= \beta_0 + \beta_1 x_1 + \epsilon
\end{aligned}
$$

# Non-linear Models

$$y = \frac{\beta_0 + \beta_1 x_1}{\beta_2 x_2 + \beta_3 x_3} \quad \text{and} \quad y = \beta_0 e^{-\beta_1 x}$$

- The above are examples of non-linear models.
- Parameter estimation is much easier for linear models (often used for empirical models).
- Mechanistic models are often non-linear.
- Try to linearize equation by taking logs, etc.

# Regression Analysis

**Regression** is the tendency of the response variable ($y$) to vary with one or more explanatory variables ($x$).

The **regression equation** describes this relationship mathematically.

**Simple regression:** one explanatory (or predictor) variable.

**Multiple regression:** more than one explanatory (or predictor) variable.

Regression first used by Francis Galton (late 1800's) to describe tendency of tall fathers to have not-so-tall sons ("regression towards the mean").

# Regression Analysis

Simple regression equation:

$$y_i = \mu_{y|x} + \epsilon_i,$$

If a linear relationship holds then

$$\mu_{y|x} = \beta_0 + \beta_1 x_i.$$

And so a fitted model will also be a straight line:

$$\hat{y}_i = b_0 + b_1 x_i$$

$\beta_0$ and $\beta_1$ are **unknown model parameters**, while $b_0$ and $b_1$ are **statistics calculated from the sample data**.

# Regression Analysis

Further assumptions required for inference about model parameters are:

- $y_i$ is **normally distributed**.
- $var[y_i]$ is constant $(= \sigma^2)$, i.e. **does not change with x**.
- $\epsilon_i \sim \text{Normal}(0,\ \sigma^2)$ (independent and identically distributed - iid).

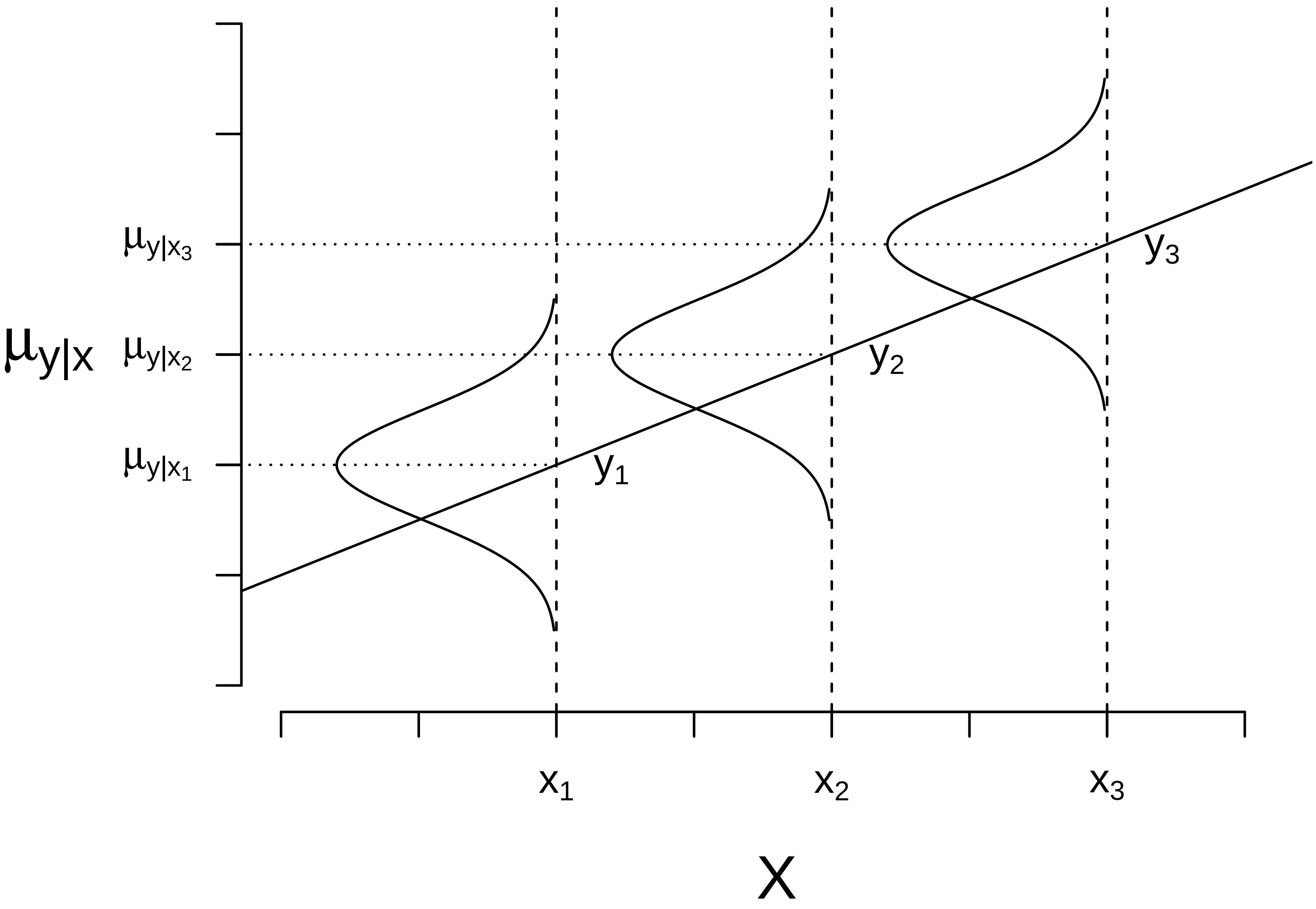
Combining with linearity of regression we can summarise as:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$$

- The prediction errors are $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$.
- In practice, since $\beta_0$ or $\beta_1$, are unknown, so are the errors.
- We estimate them with values $b_0$ and $b_1$ (or $\hat{\beta}_0$ and $\hat{\beta}_1$) that make the prediction errors "as small as possible".

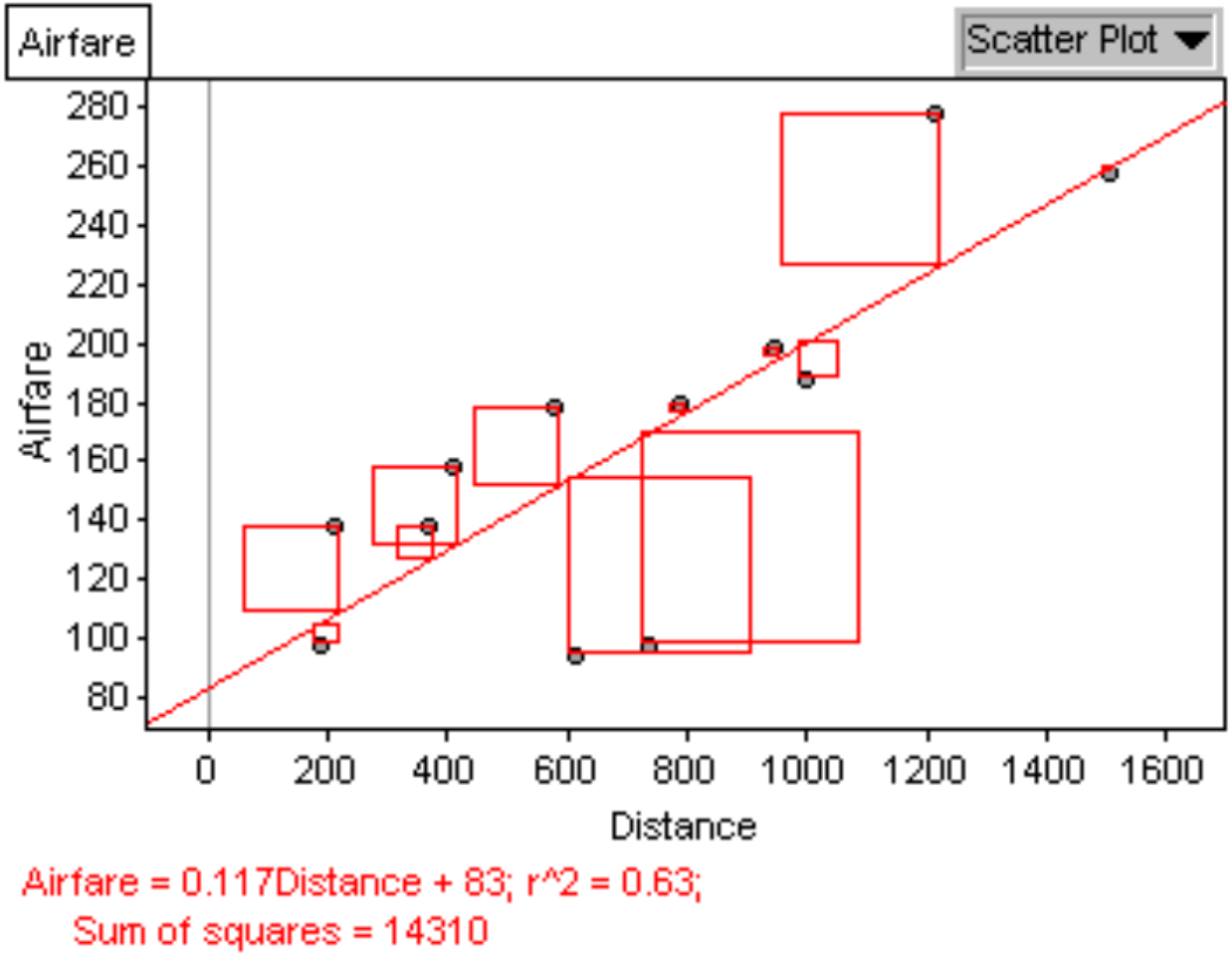
# Regression Analysis

**Graphical depiction of regression assumptions**



# Least-Squares Concept

- **Least squares regression line:** given by those values of $b_0$ and $b_1$ that minimise $\sum_{i=1}^{n} e_i^2$ (sum of the squared residuals).



# Sum of Squared Residuals

- **Total Sum of Squares**

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

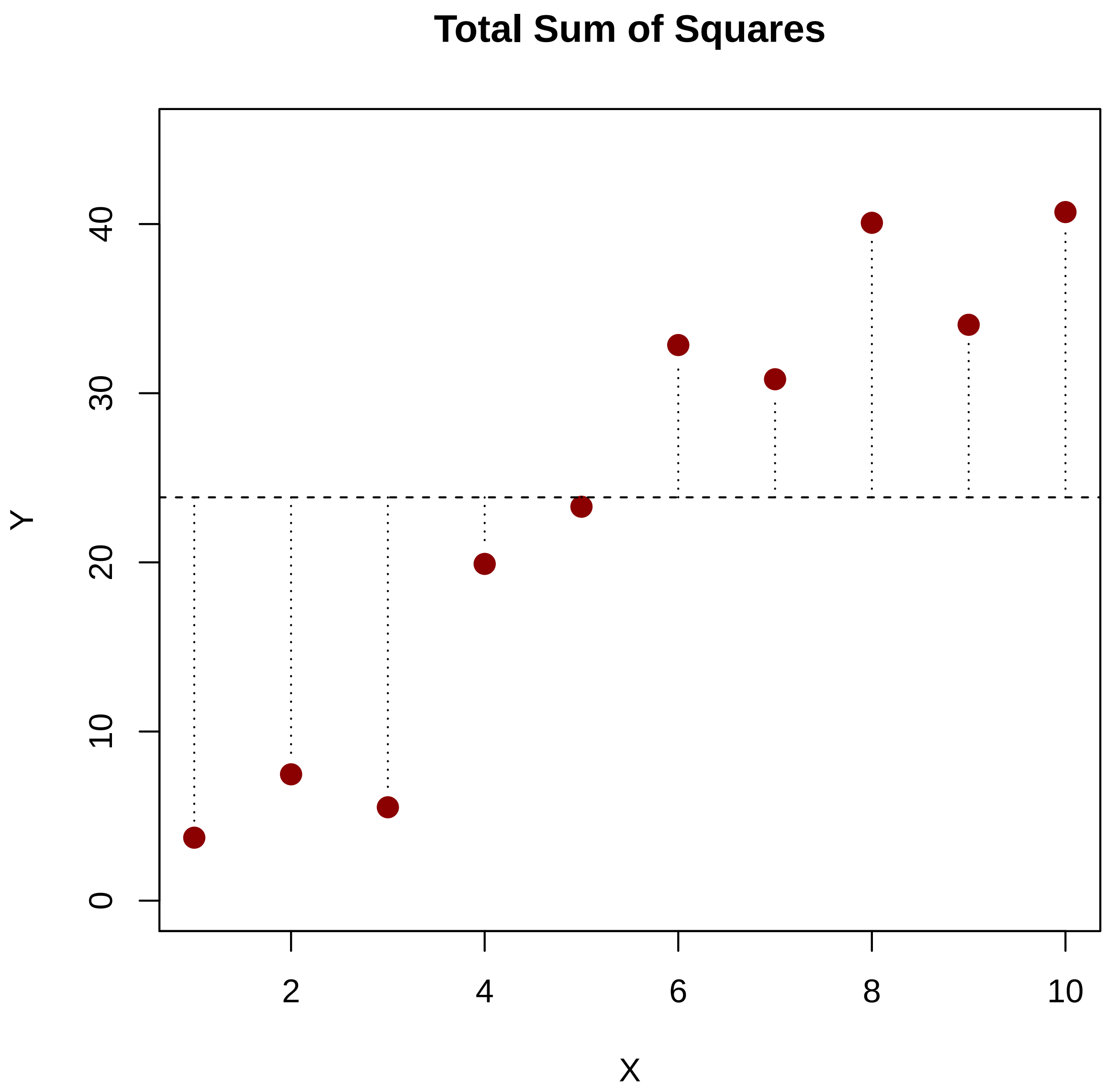
- **Residual of Sum of Squares**

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Regression Sum of Squares**

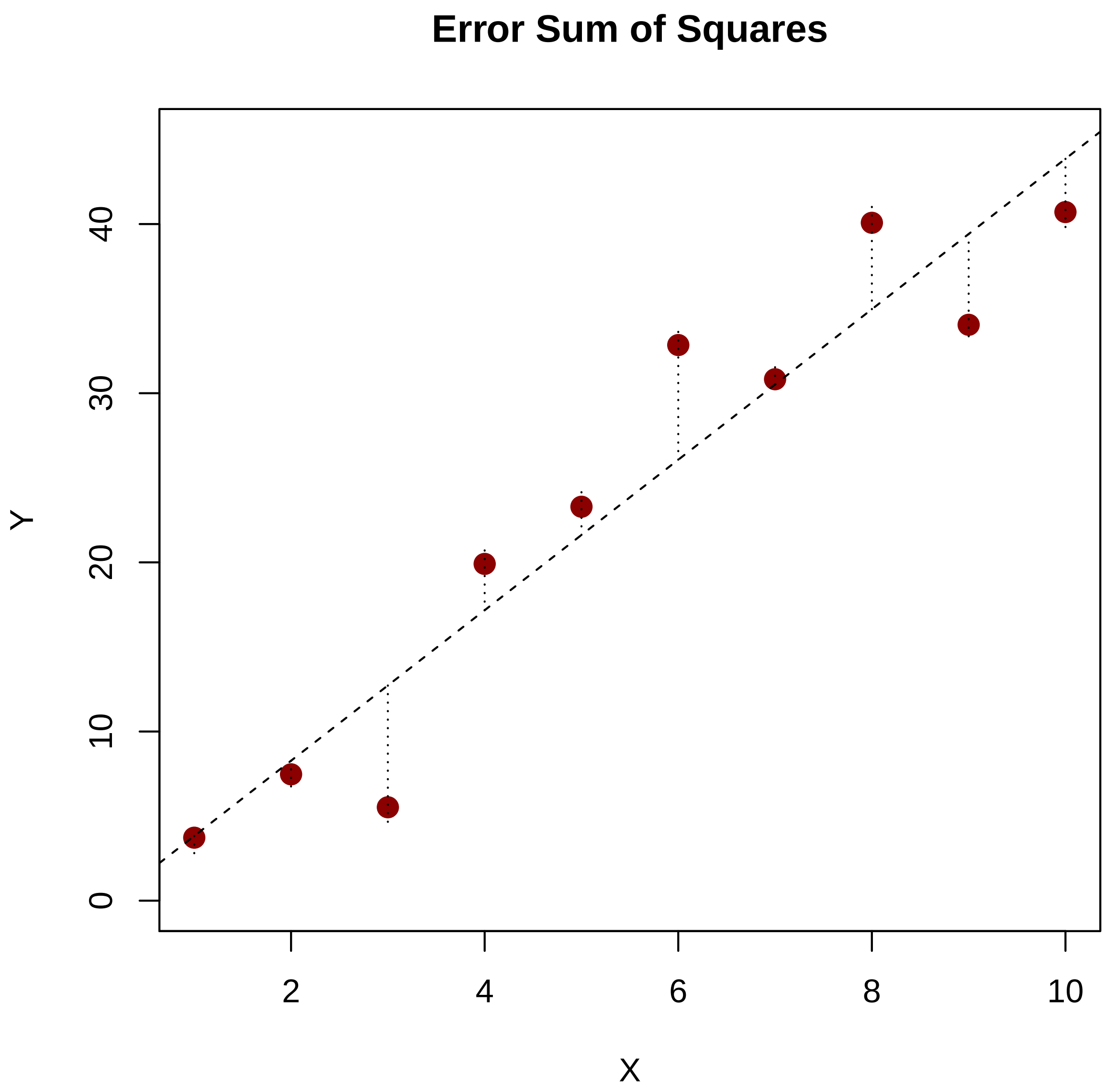
$$SS_{Reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

# Total Sum of Squares



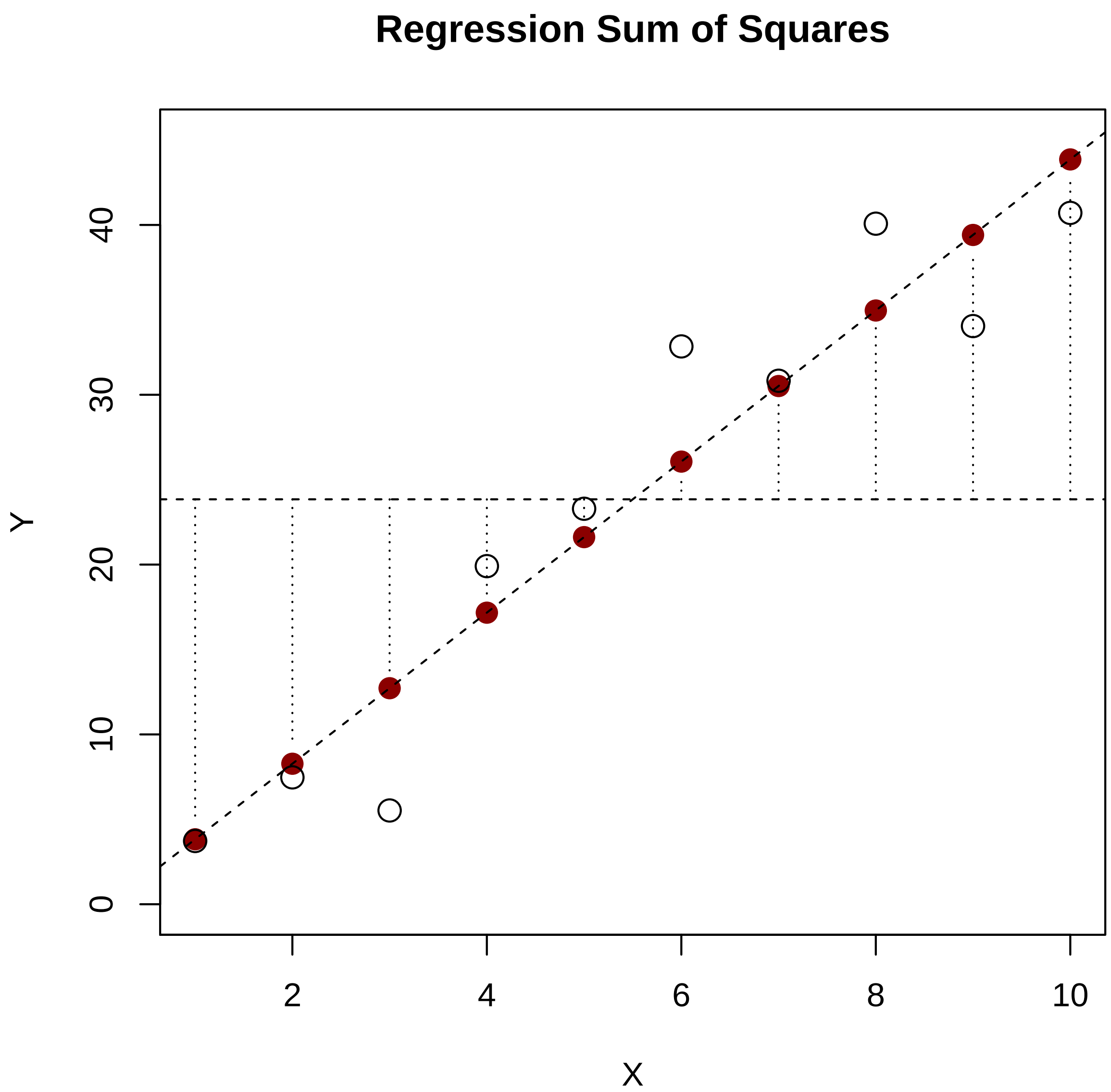

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

# Residual Sum of Squares




$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

# Regression Sum of Squares




$$SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

# Recap: Important Concepts

- **Errors** - $\epsilon_i$, random variables whose values cannot be determined exactly.
- **Fitted Values** - $\hat{y}_i$, which predict $y_i$ from $x_i$ using $b_0$ and $b_1$.
- **Residuals** - $y_i - \hat{y}_i$, which approximate the errors.
- **Least Squares** - estimates $b_0$ and $b_1$ which minimise sum of squared residuals.

# Estimate Formulas

- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$
- $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- You will not need these formulae because the software will do the calculations for you.

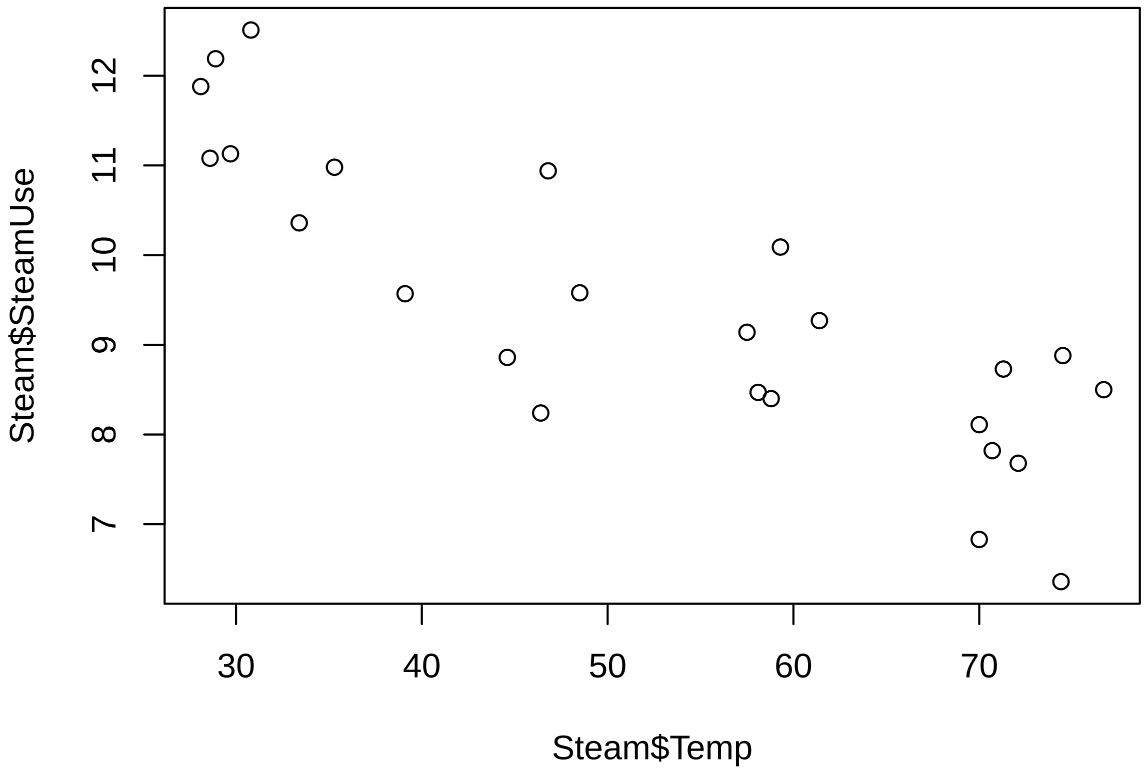
# Example: `Steam.csv`

- **Response:** Monthly steam consumption in chemical plant.
- **Explanatory:** Average operating temperature.

| SteamUse | Storage | Glycerin | Wind | CalDays | OpDays | ColdDays | Temp | Startups |
|---|---|---|---|---|---|---|---|---|
| 10.98 | 5.2 | 0.61 | 7.4 | 31 | 20 | 22 | 35.3 | 4 |
| 11.13 | 5.12 | 0.64 | 8 | 29 | 20 | 25 | 29.7 | 5 |
| 12.51 | 6.19 | 0.78 | 7.4 | 31 | 23 | 17 | 30.8 | 4 |
| 8.4 | 3.89 | 0.49 | 7.5 | 30 | 20 | 22 | 58.8 | 4 |
| 9.27 | 6.28 | 0.84 | 5.5 | 31 | 21 | 0 | 61.4 | 5 |
| 8.73 | 5.76 | 0.74 | 8.9 | 30 | 22 | 0 | 71.3 | 4 |
| 6.36 | 3.45 | 0.42 | 4.1 | 31 | 11 | 0 | 74.4 | 2 |
| 8.5 | 6.57 | 0.87 | 4.1 | 31 | 23 | 0 | 76.7 | 5 |
| 7.82 | 5.69 | 0.75 | 4.1 | 30 | 21 | 0 | 70.7 | 4 |
| 9.14 | 6.14 | 0.76 | 4.5 | 31 | 20 | 0 | 57.5 | 5 |
| 8.24 | 4.84 | 0.65 | 10.3 | 30 | 20 | 11 | 46.4 | 4 |
| 12.19 | 4.88 | 0.62 | 6.9 | 31 | 21 | 12 | 28.9 | 4 |
| 11.88 | 6.03 | 0.79 | 6.6 | 31 | 21 | 25 | 28.1 | 5 |
| 9.57 | 4.55 | 0.6 | 7.3 | 28 | 19 | 18 | 39.1 | 5 |
| 10.94 | 5.71 | 0.7 | 8.1 | 31 | 23 | 5 | 46.8 | 4 |
| 9.58 | 5.67 | 0.74 | 8.4 | 30 | 20 | 7 | 48.5 | 4 |
| 10.09 | 6.72 | 0.85 | 6.1 | 31 | 22 | 0 | 59.3 | 6 |
| 8.11 | 4.95 | 0.67 | 4.9 | 30 | 22 | 0 | 70 | 4 |
| 6.83 | 4.62 | 0.45 | 4.6 | 31 | 11 | 0 | 70 | 3 |
| 8.88 | 6.6 | 0.95 | 3.7 | 31 | 23 | 0 | 74.5 | 4 |
| 7.68 | 5.01 | 0.64 | 4.7 | 30 | 20 | 0 | 72.1 | 4 |
| 8.47 | 5.68 | 0.75 | 5.3 | 31 | 21 | 1 | 58.1 | 6 |
| 8.86 | 5.28 | 0.7 | 6.2 | 30 | 20 | 14 | 44.6 | 4 |
| 10.36 | 5.36 | 0.67 | 6.8 | 31 | 20 | 22 | 33.4 | 4 |
| 11.08 | 5.87 | 0.7 | 7.5 | 31 | 22 | 28 | 28.6 | 5 |

# Example: Steam.csv

```
Steam <- read.csv ("Data/Steam.csv")
plot (Steam$SteamUse ~ Steam$Temp)
```




- First plot data. A straight line is a tentative model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

# Example: Steam.csv

- Fit a linear regression model with the `lm()` command in R.

```
m1 <- lm (SteamUse ~ Temp, data=Steam)
summary (m1)
```

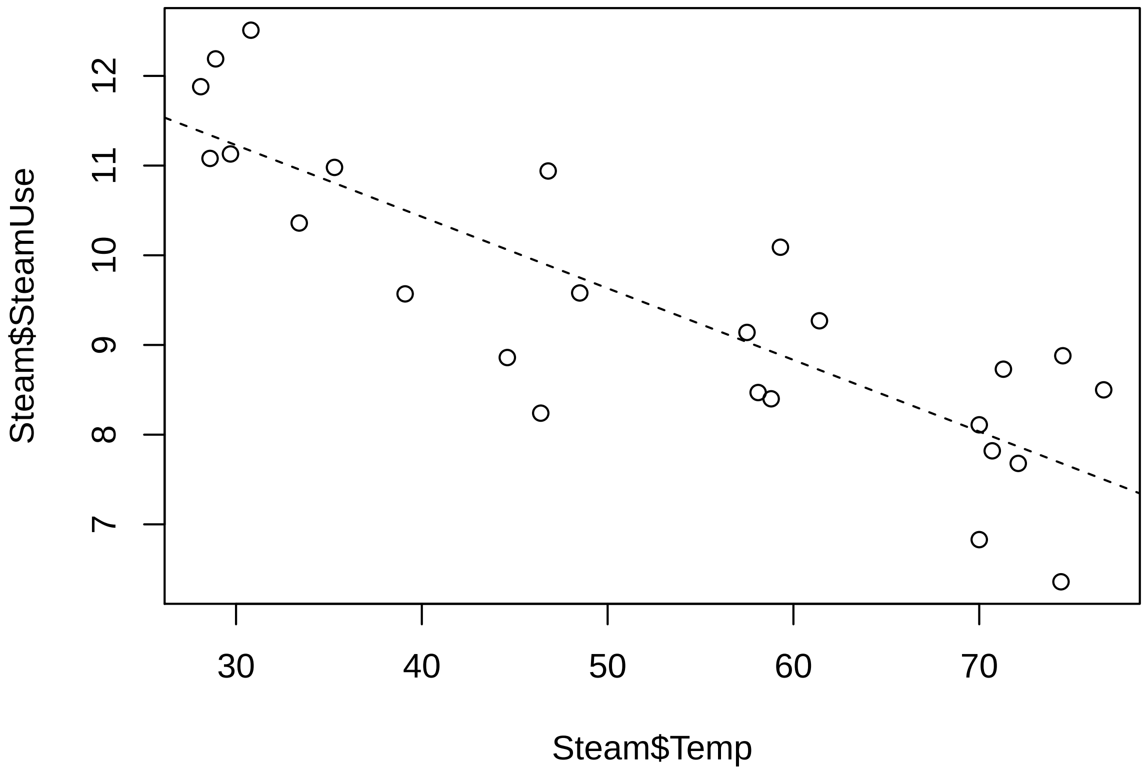
```
Call:
lm(formula = SteamUse ~ Temp, data = Steam)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6789 -0.5291 -0.1221  0.7988  1.3457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
Temp        -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-squared:  0.7144, Adjusted R-squared:  0.702
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07
```

```
coef(m1)
```

```
(Intercept)        Temp
13.62298927 -0.07982869
```

# Example: `Steam.csv`

```
plot (Steam$SteamUse ~ Steam$Temp)
abline (coef (m1), lty=2)
```

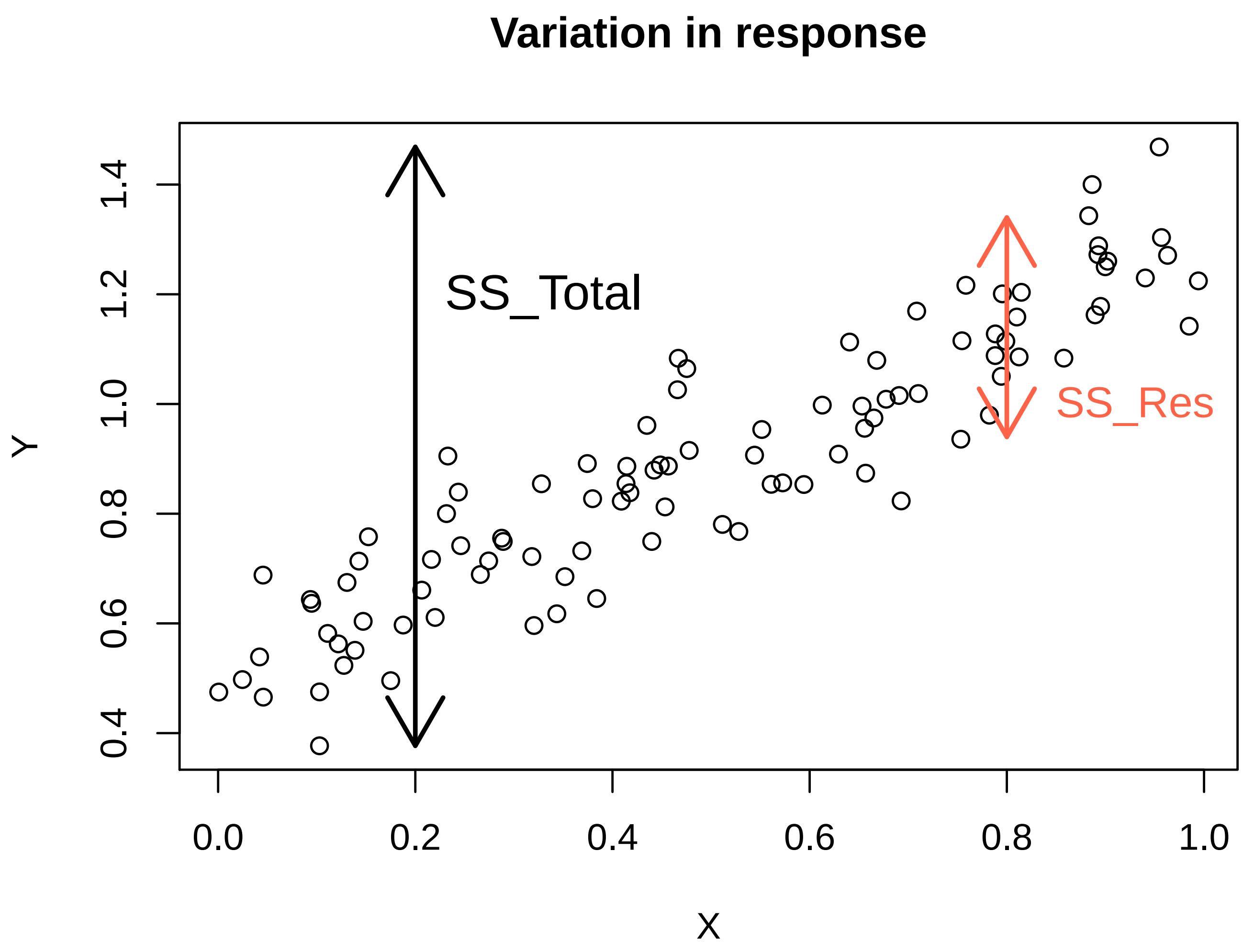



- Negative, moderately strong, linear relationship.

# How good is the fit?

The proportion of variation explained by the fit is called **R-squared** and is given by:

$$R^2 = SS_{Res}/SS_{Total} = 1 \; - \; SS_{Res}/SS_{Total}$$



# How good is the fit?

If all the data lies on a straight line then $SS_{Res} = 0$ and $R^2 = 1$ (or $100\%$) - the fit explains everything.

Good fit means **small** $SS_{Res}$ and **large** $R^2$.

How large? Rule of thumb is that:

$R^2$ **should be at least 0.5 or 50%**

In the `Steam.csv` example the $R^2$ value was 71%.

```
summary(m1)$r.squared
[1] 0.7144375
```

# Residual Analysis

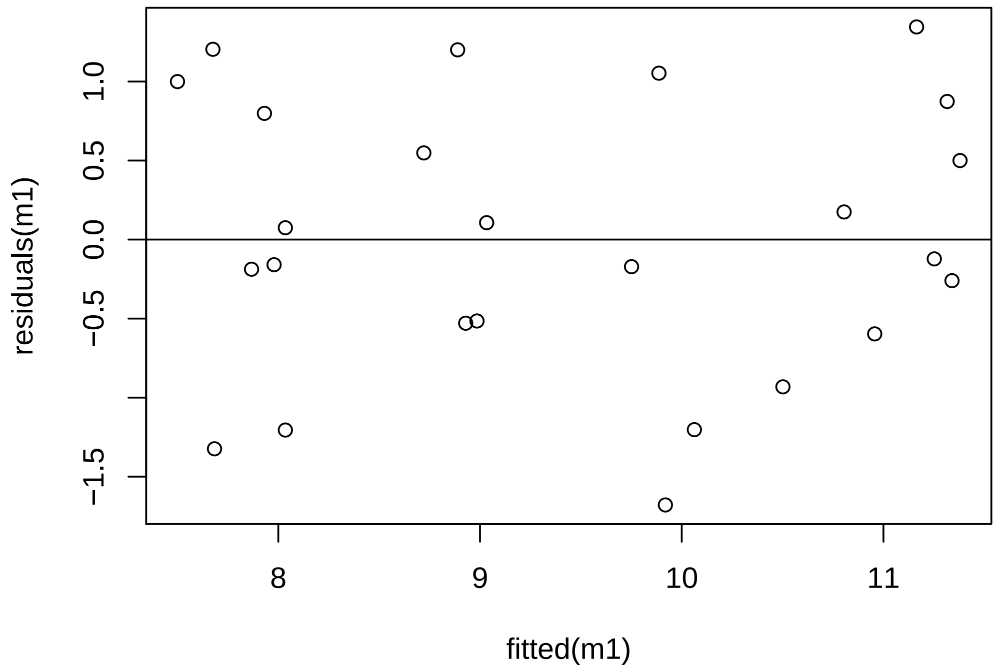
Recall that *residual* $=$ *observed* - *fit*, i.e.

$$e_i = y_i - \hat{y}_i$$

Examining (plotting) residuals shows how well the fit explains the systematic variation in the data

Ideally plot of residuals against fitted values should be **random scatter** about $0$ of **constant size** (horizontal band).

# Residual Analysis




Residual plot for regression line shows random scatter of **constant width**.

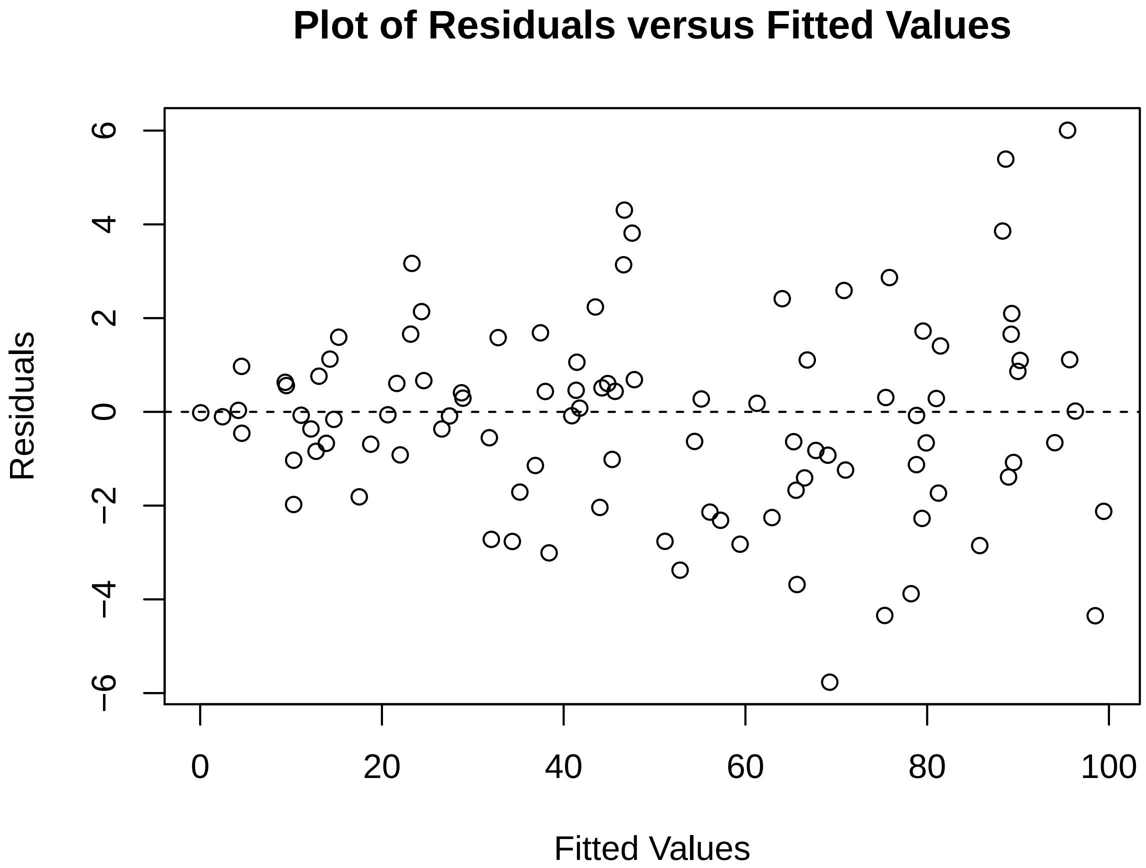
```
plot (residuals(m1) ~ fitted(m1))
abline (h=0)
```

Can also plot with:

```
plot (m1$residuals ~ m1$fitted.values)
```

# Residual Analysis

If the scatter is not pattern less and constant width then the plot may suggest **a transformation of the y variable:**



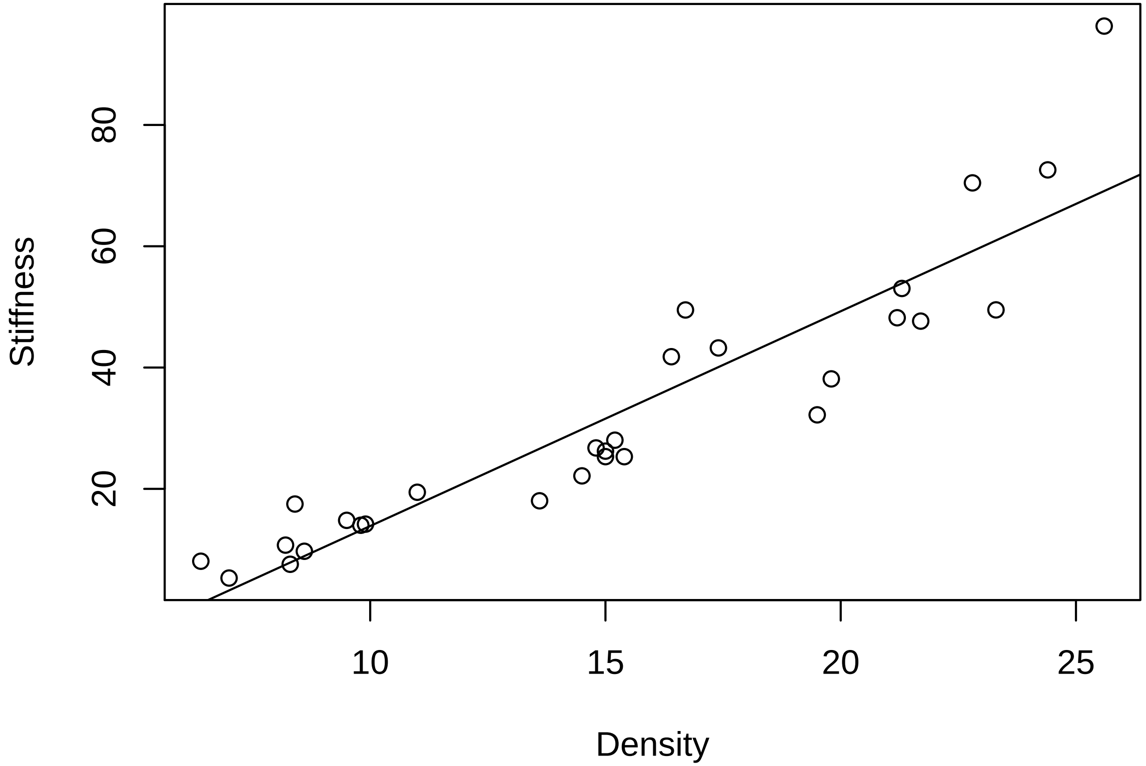

Scatter **increases with fits**. Should try a **shrinking transformation** of y, e.g. log or square root.

# Example: `Particleboard.csv`

- Manufacture of new type of particleboard.
- Attempt to model the relationship between density & stiffness.
- 30 sheets were manufactured and measured.
- Data can be found in `Particleboard.csv`.

# Example: Particleboard.csv

```
Part <- read.csv (file="Data/Particleboard.csv", header=TRUE)
plot (Part$Stiffness ~ Part$Density, xlab="Density", ylab="Stiffness")
m2 <- lm (Stiffness ~ Density, data=Part)
abline (coef(m2))
```



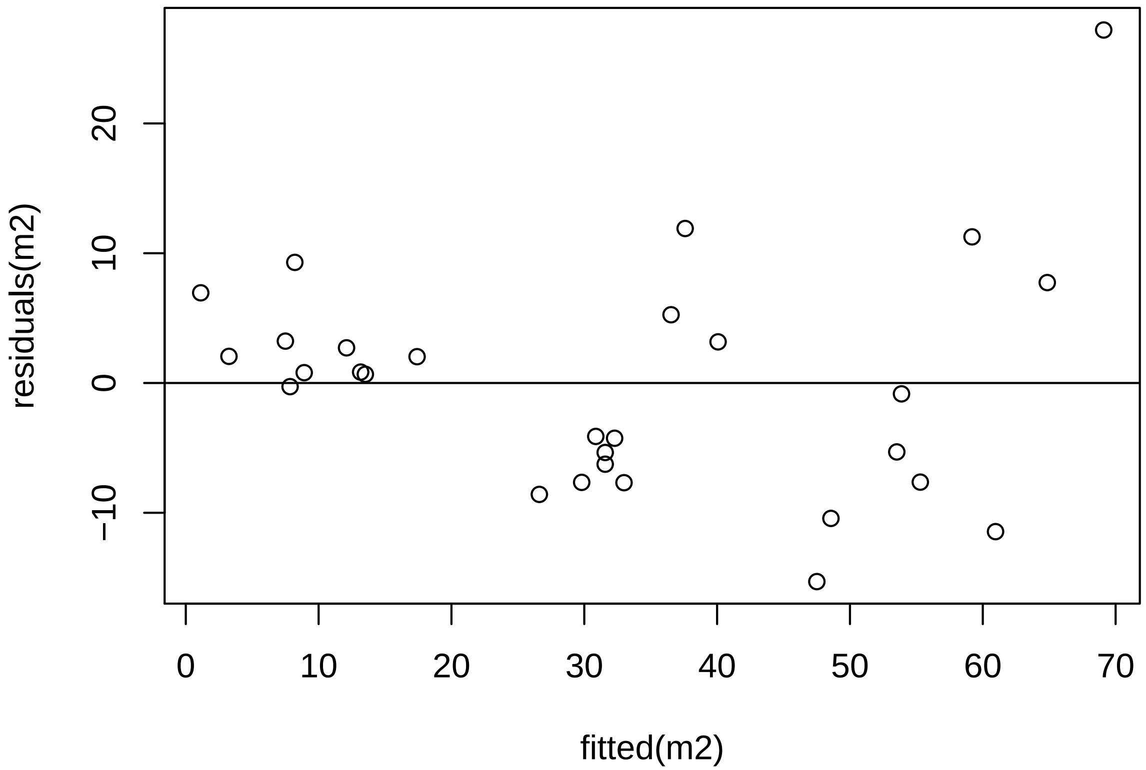

- Strong positive relationship.
- Increasing variance.
- Linear trend? Exponential? Transform?

# Example: `Particleboard.csv`

- No known physical law so we are in need of an empirical model.
- Try $y = \beta_0 + \beta_1 x + \epsilon$.
- Or perhaps $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.
- Or try $\log$ transformation.
- Generally, empirical models are not advised for extrapolation.

# Example: `Particleboard.csv` - Residuals

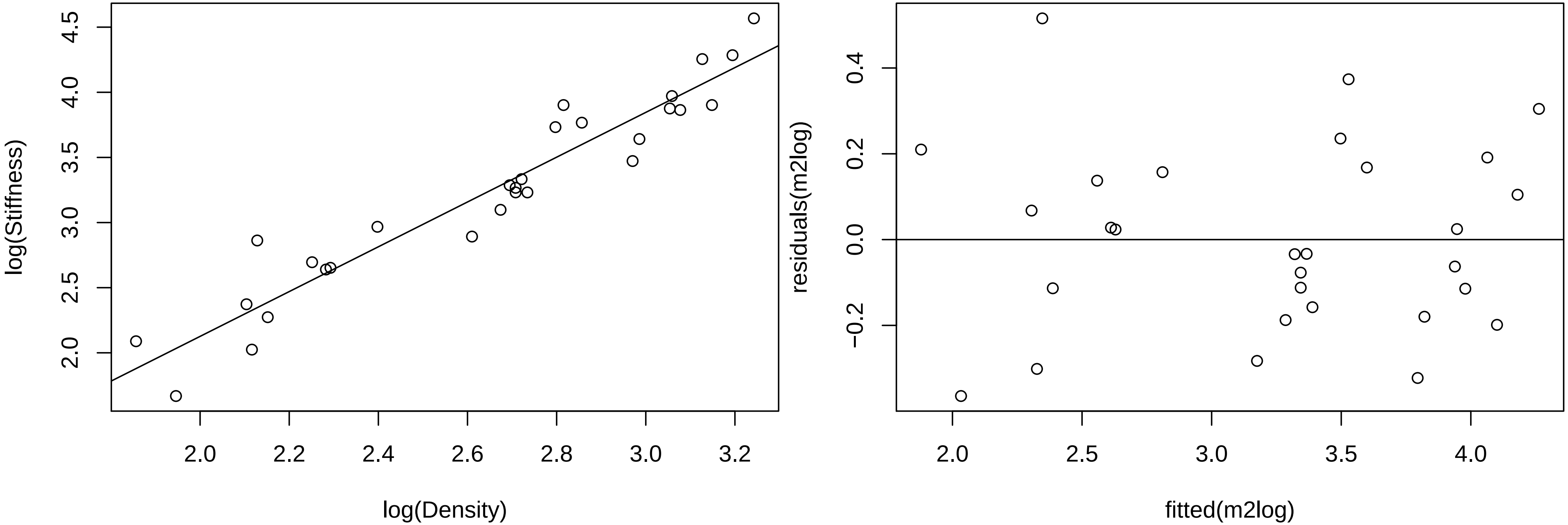
```
plot (residuals(m2) ~ fitted(m2))
abline (h=0)
```




- Curvature in the residuals plot.
- Variability is increasing. Heteroscedasticity.
- Try a transformation.

# Example: `Particleboard.csv` - $\log$ transform

```
plot (log(Part$Stiffness) ~ log(Part$Density),
      xlab="log(Density)", ylab="log(Stiffness)")
m2log <- lm (log(Stiffness) ~ log(Density), data=Part)
abline (coef(m2log))
```




- Here $\log$ is the natural log (base $e = 2.718...$).
- Trend is now linear.
- Variability is constant.

# Example: `Particleboard.csv` - Summary comparisons

```
summary (m2)
Call:
lm(formula = Stiffness ~ Density, data = Part)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2997  -6.2553   0.6735   3.2294  27.2010

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.5338     4.7355  -4.547 0.000103 ***
Density       3.5405     0.2922  12.119 1.98e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.869 on 27 degrees of freedom
Multiple R-squared:  0.8447, Adjusted R-squared:  0.8389
F-statistic: 146.9 on 1 and 27 DF,  p-value: 1.981e-12
```

```
summary (m2log)
Call:
lm(formula = log(Stiffness) ~ log(Density), data = Part)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36461 -0.15759 -0.03319  0.15720  0.51573

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.3130     0.2728  -4.813 5.04e-05 ***
log(Density)   1.7196     0.1020  16.861 7.35e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2206 on 27 degrees of freedom
Multiple R-squared:  0.9133, Adjusted R-squared:  0.9101
F-statistic: 284.3 on 1 and 27 DF,  p-value: 7.345e-16
```

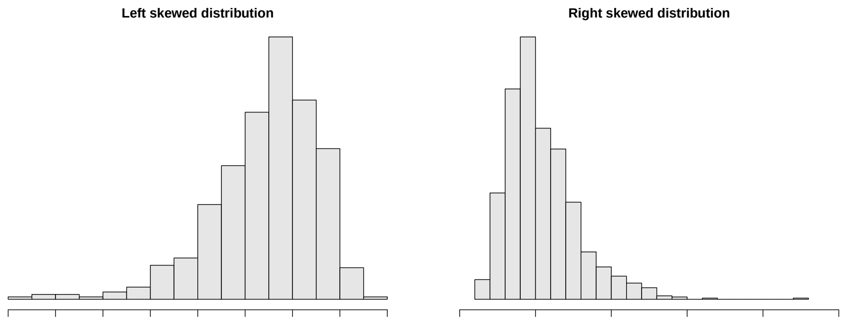
- Transformation has increased $R^2$.
- Residual standard error has decreased.

# Transformations - Fixing Linearity and Normality

Power transformations can either **stretch** large values (good for left skewed data), or **shrink** large values (good for right skewed data).

$$y^{\star} = \begin{cases} \text{sign}(\lambda).\text{y}^{\lambda} & \lambda \neq 0 \\ \log(\text{y}) & \lambda = 0 \end{cases}$$

The sign function, sign$(\lambda)$, is $+1$ if $\lambda > 0$ and -1 if $\lambda < 0$.




| Power, $\lambda$ | Formula | Name | Result |
|---|---|---|---|
| 3 | $y^3$ | cube | stretches large values |
| 2 | $y^2$ | square | |
| 1 | $y$ | raw | |
| 0.5 | $\sqrt{y}$ | square root | shrinks large values |
| 0 | $\log(y)$ | logarithm | |
| $-0.5$ | $-1/\sqrt{y}$ | reciprocal root | |
| $-1$ | $-1/y$ | reciprocal | |

# Testing Coefficients

If the regression coefficient of a variable ($\beta_i$) is zero, then changes in that variable do not affect the response variable.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
Temp        -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The hypotheses tested are: $H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0$
- The test statistic is: $\frac{\texttt{Estimate - 0}}{\texttt{Std. Error}}$.
- The $p$-value (`Pr(>|t|)`) is the probability of getting an estimate as extreme (far away from zero) as we did, given $H_0$ is true, i.e given the coefficient is zero.
- Here the slope coefficient and intercept (not as interesting) are significantly different from zero.

# Interpretation of Parameter Estimates

**Intercept**

- Expected response when $x = 0$.
- *Steam production example:*

```
print (m1)
```

```
Call:
lm(formula = SteamUse ~ Temp, data = Steam)

Coefficients:
(Intercept)         Temp
   13.62299     -0.07983
```

  - Expected steam use at $0^o F$.
  - Estimate steam use at $= 13.6$ lb/month at $0^o F$.

- Many times the intercept is meaningless to interpret.

# Interpretation of Parameter Estimates

**Slope**

- Expected increase in the response when $x$ increases by 1.
- *Steam production example:*
    - Estimate that steam use will increase by $-0.07983$ lb/ month for each increase in average temperature of $1^o F$.
- So we are predicting a decrease in steam use as the average temperature increases which makes sense and matches the plot.

# Interpretation of Parameter Estimates

- $y = \beta_0 + \beta_1 x$
  - If $x$ is increased by one unit, $y$ changes by an *addition* of $\beta_1$ units.
- $\log(y) = \beta_0 + \beta_1 x$
  - If $x$ is increased by one unit, $y$ changes by a *factor* $e^{\beta_1}$ units.
- $\log(y) = \beta_0 + \beta_1 \log(x)$
  - If $x$ multiplied by a factor of 2, $y$ changes by a *factor* of $2^{\beta_1}$.

# $F$-test

- There is an overall test to see if the model is useful in predicting change in the response.
- This is the $F$-test - it is related to the sum of squares.
- For a model with $p$ parameters:

| | | | | | |
|---|---|---|---|---|---|
| Sum of squares : | $SS_{Total}$ | $=$ | $SS_{Reg}$ | $+$ | $SS_{Res}$ |
| $df$ : | $n-1$ | $=$ | $p$ | $+$ | $(n-p-1)$ |
| Mean squares : | $SS_{Total}/(n-1)$ | | $SS_{Reg}/p$ | | $SS_{Res}/(n-p-1)$ |

- The test statistic of the $F$-test is $\frac{MS_{Reg}}{MS_{Res}} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)}$.
- Under the null hypothesis (model not significant) the test statistic follows and $F$ distribution with $p$ and $n-p-1$ degrees of freedom.
- If the $p$-value is $<$ 0.05 say, the model explains significant amount of variation in $y$.

# $F$-test

- For simple linear regression (one explanatory variable), the $F$-test is identical to the $t$-test for the slope ($t^2 = F$).

```
summary(m1)
```

```
Call:
lm(formula = SteamUse ~ Temp, data = Steam)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6789 -0.5291 -0.1221  0.7988  1.3457 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
Temp        -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-squared:  0.7144, Adjusted R-squared:  0.702 
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07
```

- This result does not hold in the case of multiple regression (more than one explanatory variable).

# Prediction and Estimation

- For any value $x_0$, least squares line gives fitted value $b_0 + b_1 x_0$.
- This value estimates two different quantities:

  $\mu_{y|x}$ **mean** response when $x = x_0$.

  $y_0$ **actual** (individual) response when $x = x_0$.
- We use **confidence** intervals to predict **mean** responses.
- We use **prediction** intervals to predict **individual** responses.
- The latter interval will be larger as it includes the uncertainty in the mean value.
- Both intervals get larger as $x_0$ gets further from its mean.
- Avoid extrapolation.

# Prediction and estimation

Exact prediction and confidence intervals can be found using R, e.g. for model m1 when Temp=70:

```
PI8 <- predict (m1, data.frame(Temp=70), interval='prediction')
PI8
```

```
       fit      lwr      upr
1 8.034981 6.119328 9.950634
```

```
CI8 <- predict (m1, data.frame(Temp=70), interval='confidence')
CI8
```

```
       fit      lwr      upr
1 8.034981 7.506674 8.563288
```

# Assessing and Correcting Lack of Fit

- Assumptions:
  - The relationship is linear.
  - Variance of the response is constant.
  - The errors are uncorrelated, particularly serially.
  - Errors are normally distributed.
- We will look at ways in which the validity of these assumptions can be assessed.

# Non-linearity

- Step one - plot the data.
- Step two - plot the errors vs the explanatory variable.
- Patterns in the plots suggest some form of non-linearity.
- Either add polynomial terms (dealt with later), or transform the data.

# Non-Constant Variance

- As the $x$ variable increases, the variability of $y$ also commonly) increases.
- This may be detected in the scatterplot $y$ vs $x$ it will be clearer in the scatterplot of residuals vs fits.
- Some variance stabilising transformations exist.
- If error standard deviation proportional to $x$, try taking logs.

# Correlated Errors

- Most often violated when the observations collected sequentially.
- There may be a variable, not included in the model, which changes over time.
- Plot the residuals vs their “order” and look for unusual runs. (Can’t always do this!)
- Durbin - Watson test.

# Non Normal Errors

- The model errors should be normally distributed however some violation of this assumption will make little difference to conclusions.
- The measured residuals may be non normal for smaller sample sizes.
- Normal probability plots will show problems.

# Useful plots of residuals:

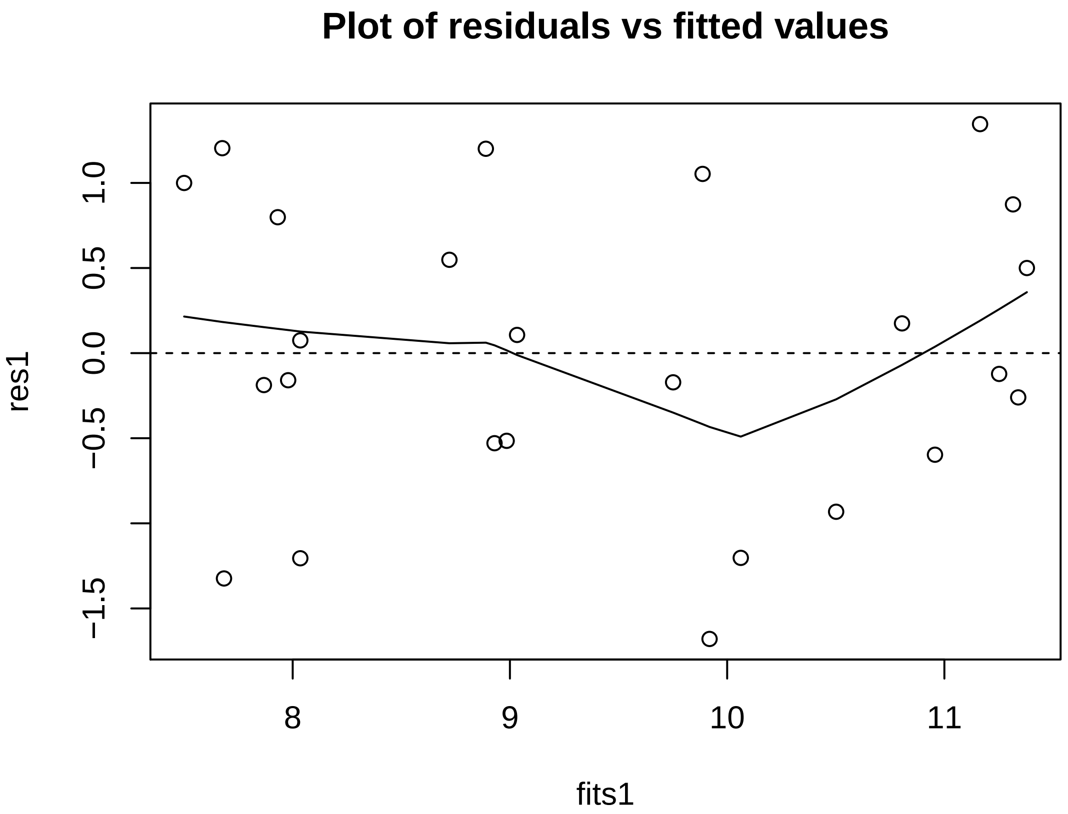
- Residuals vs fitted values.
- Histogram of residuals.
- q-q plot of residuals.
- Plot of residuals vs order of the data.

Can plot separately in `R`, or use `plot.lm()` command which produces similar set of plots.

*Note:* all tests based on $F$, $t$ etc require normality of residuals ($F$ more than $t$).

# Residuals vs Fitted values

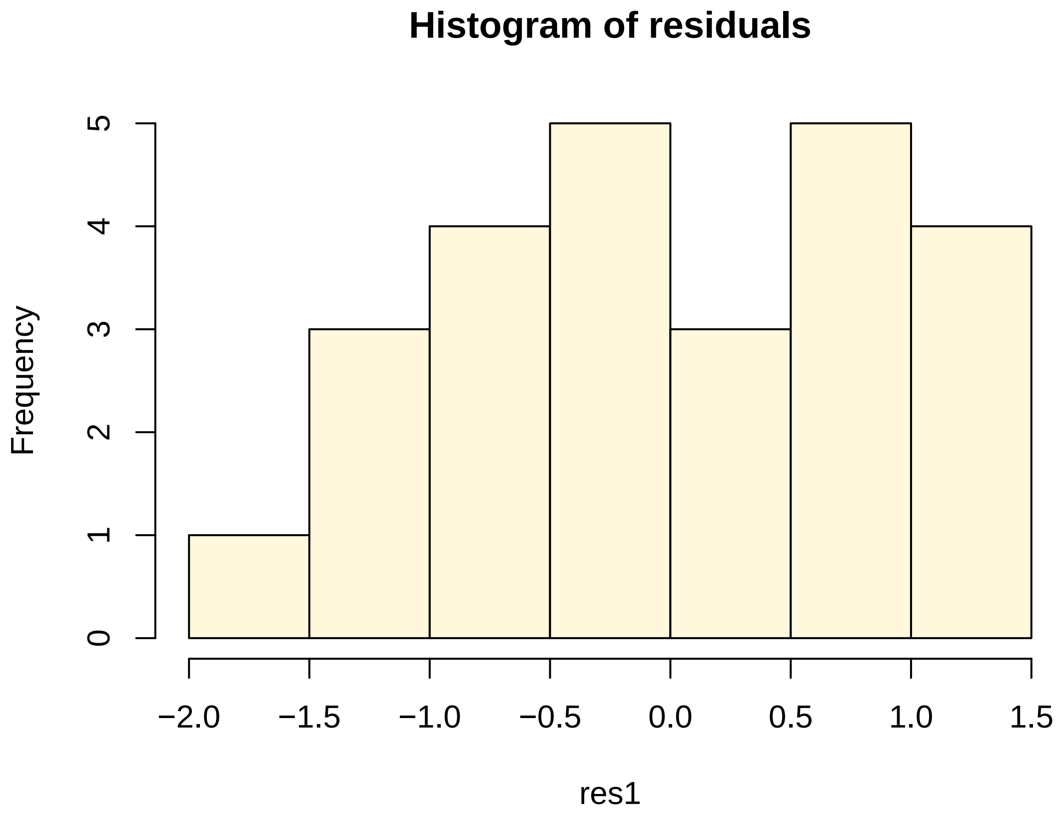
```
res1  <- residuals(m1)
fits1 <- fitted(m1)
plot(fits1,res1, main = "Plot of residuals vs fitted values")
abline(h=0, lty = "dashed")
lines(lowess(res1~fits1))
```




Plot shows random spread of points with constant variance.

# Histogram of Residuals

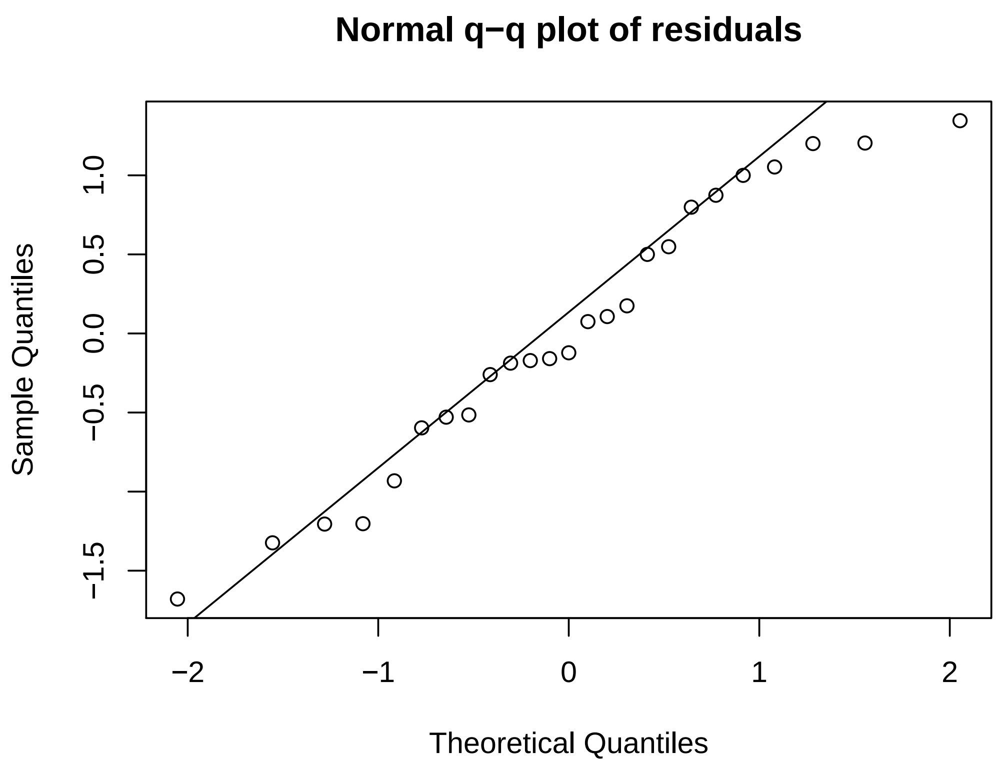
```
hist(res1, col="cornsilk", main="Histogram of residuals")
```




Residuals appear to be **normally distributed**.

# Quantile-quantile plot of residuals

```
qqnorm(res1, main = "Normal q-q plot of residuals")
qqline(res1)
```




Some curvature but **normal assumption is probably ok** (can always test for normality if necessary).

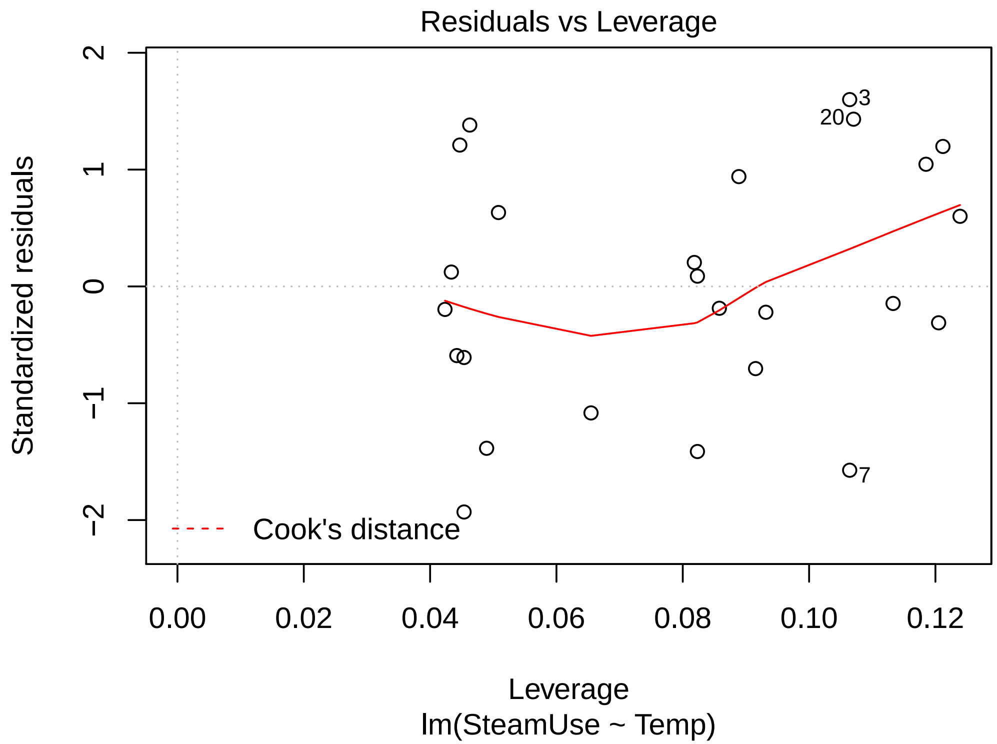
```
shapiro.test(res1)
```

```
Shapiro-Wilk normality test

data:  res1
W = 0.9596, p-value = 0.4064
```

# Identifying Influential Points

Influential or leverage points are observations that, when removed, have a large effect on the regression model and coefficients.

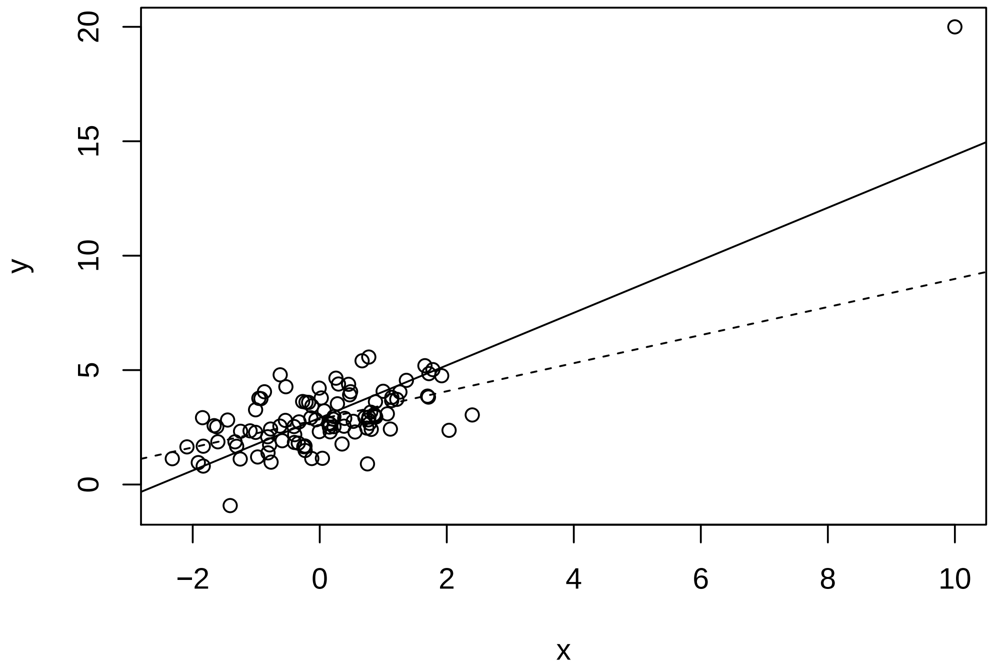
```
plot(m1, which=5)
```

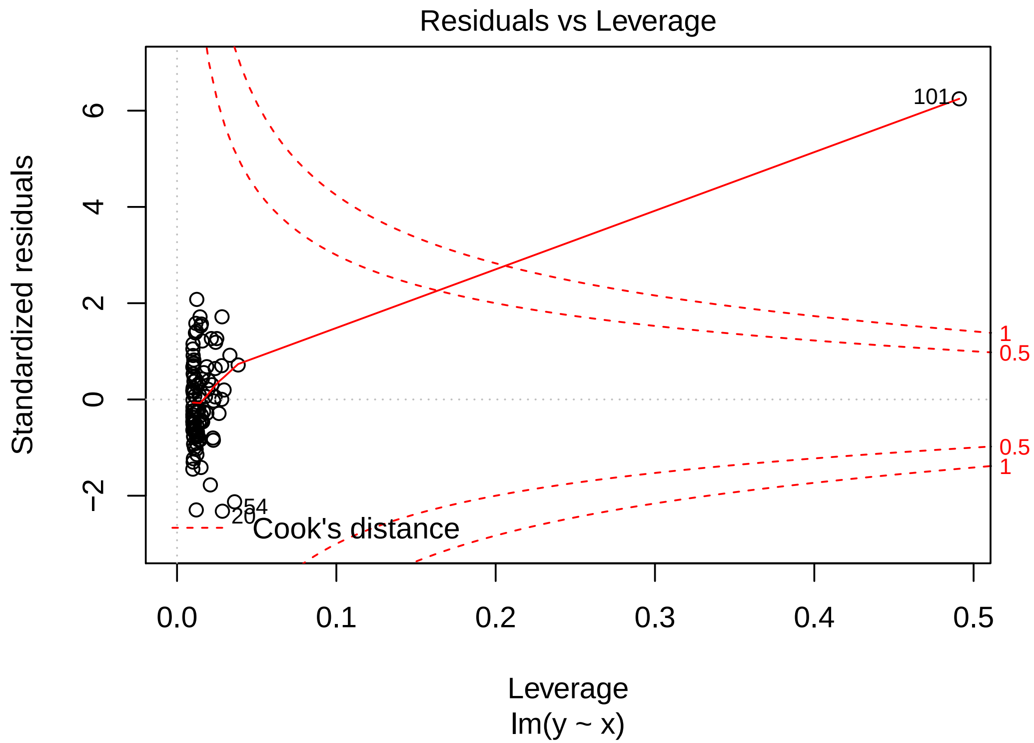



- Need to be wary of over-interpreting regressions with influential points.
- You can end up modelling the outlier.

# Plot of Residuals vs Leverage

```
x0 <- rnorm(100); y0 <- 3 + .5*x0 + rnorm(100)
x <- c(x0, 10); y <- c(y0, 20)
mLev <- lm (y ~ x); mLev0 <- lm (y0 ~ x0)
plot (y ~ x); abline (coef(mLev)); abline (coef(mLev0), lty=2)
plot (mLev, which=5)
```



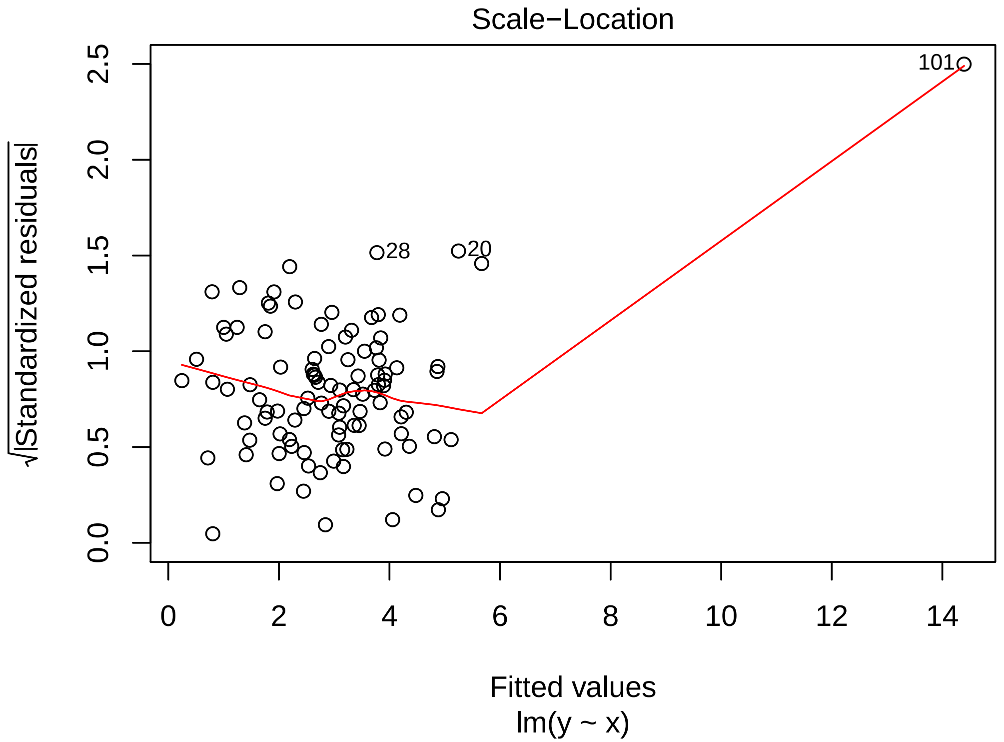
# Leverage and Influence

- Outliers are not necessarily influential.
- High leverage observations are not necessarily influential.
- Influential observations are not necessarily outliers.
- In `R` we find the leverages ourselves. Points with more than twice the average leverage are having undue influence. This is a rule of thumb.
- There are different measures of influence; $h_{ii}$, DFITS, Cook's distance, etc.

# Identifying Outliers

Plots of **standardised residuals** can be identify outliers.

```
plot(mLev, which=3)
```




- Examine outliers especially when they are large in **size** ( $> 3$) or in **number** ( $> 10\%$).

# What to do with Outliers?

- Is there an error in measuring / transcribing?
  - Delete from the data set.
- Is there auxiliary information to describe its differentness?
  - Possible delete from the data set.
- If neither of the above proceed with caution.
- Plot the residuals.

# Residual Plots

The four common residual plot can be plotted easily in one go:

```
par (mfrow=c(2,2))
plot (m1)
```