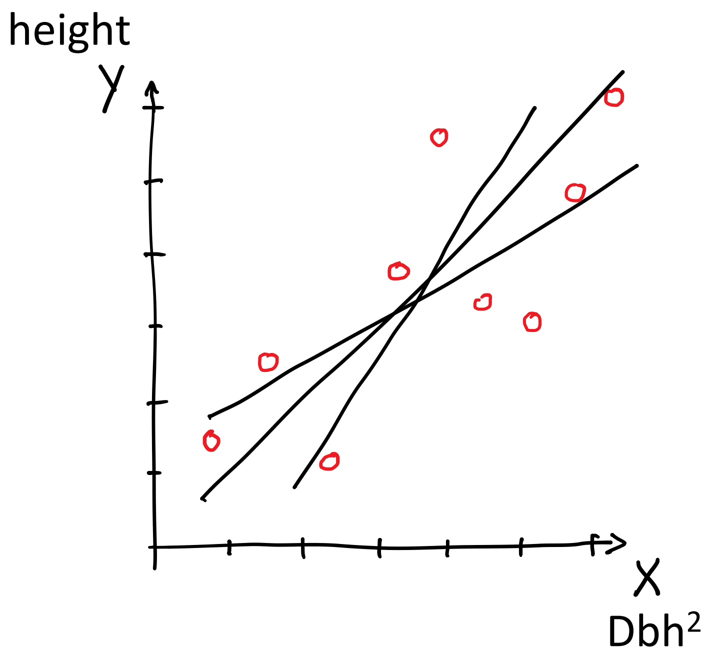
3. Fitting equations

Fitting a line to data

aka least squares

aka Linear regression

We add a line to the data to see what the trend is.
But which line fits the data best?



line equation:

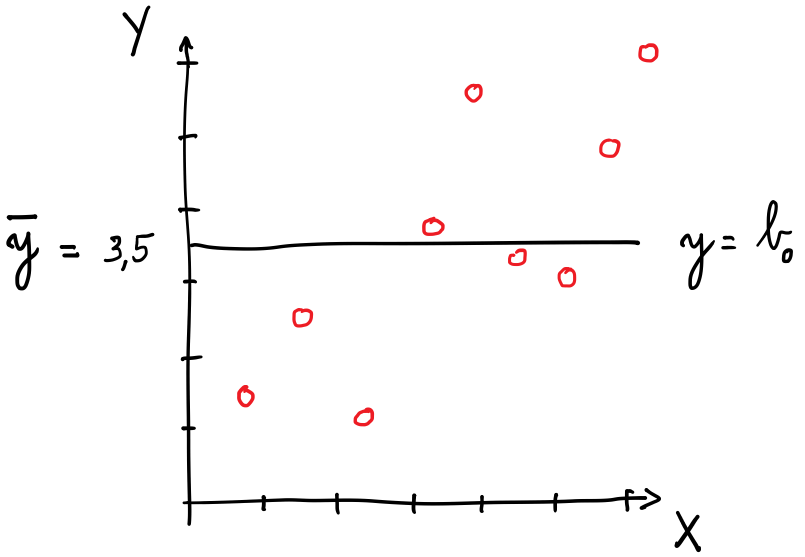
$$y = b_0 + b_1 \times x$$

Y intercept

slope

A **horizontal** line that cuts through the **average Y-value corresponds to "NO RELATIONSHIP BETWEEN X AND Y".**

Probably the worst fit here, but it is a **starting point to find the optimal line to fit our data**.



line equation:

$$y = b_0 + b_1 \times x$$

Y intercept

slope

$$\begin{cases} b_0 = \bar{y} = 3{,}5 \\ b_1 = 0 \end{cases}$$

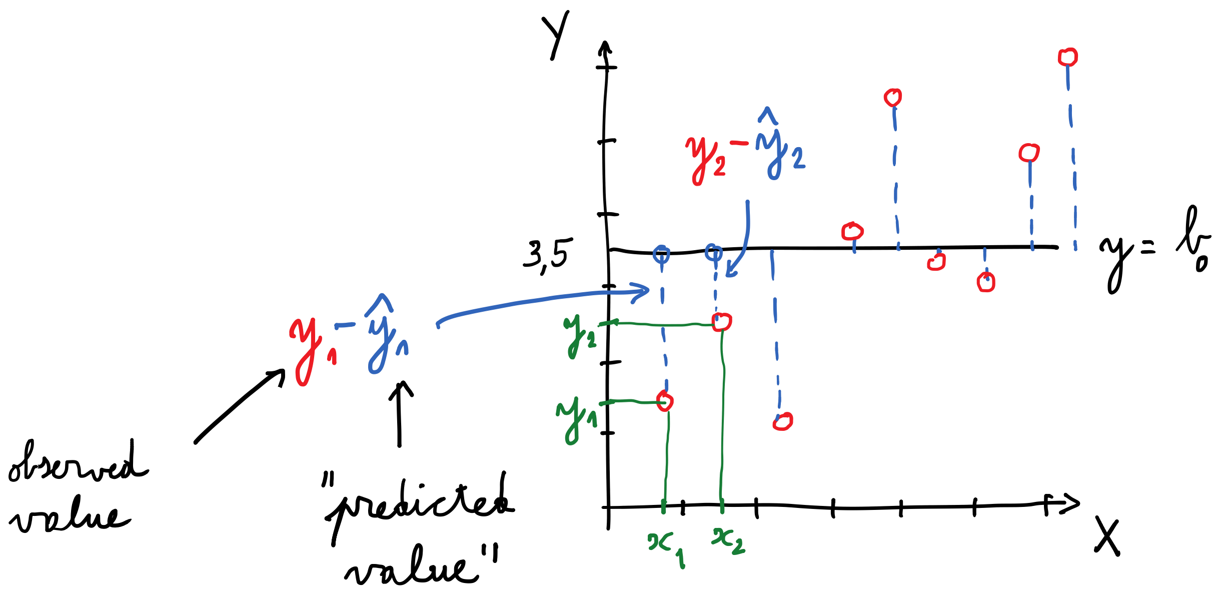
We can measure how well this line fits the data by seeing how close it is to the data points.

- We add up the distance of each point to the line.
- Squaring ensures that each term is positive.

OBSERVED y-value at x_1

$$(\boxed{y_1} - \boxed{\hat{y}_1})^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2 + (y_5 - \hat{y}_5)^2 + (y_6 - \hat{y}_6)^2 + (y_7 - \hat{y}_7)^2 + (y_8 - \hat{y}_8)^2 + (y_9 - \hat{y}_9)^2$$

PREDICTED y-value at x_1



We can measure how well this line fits the data by seeing how close it is to the data points.

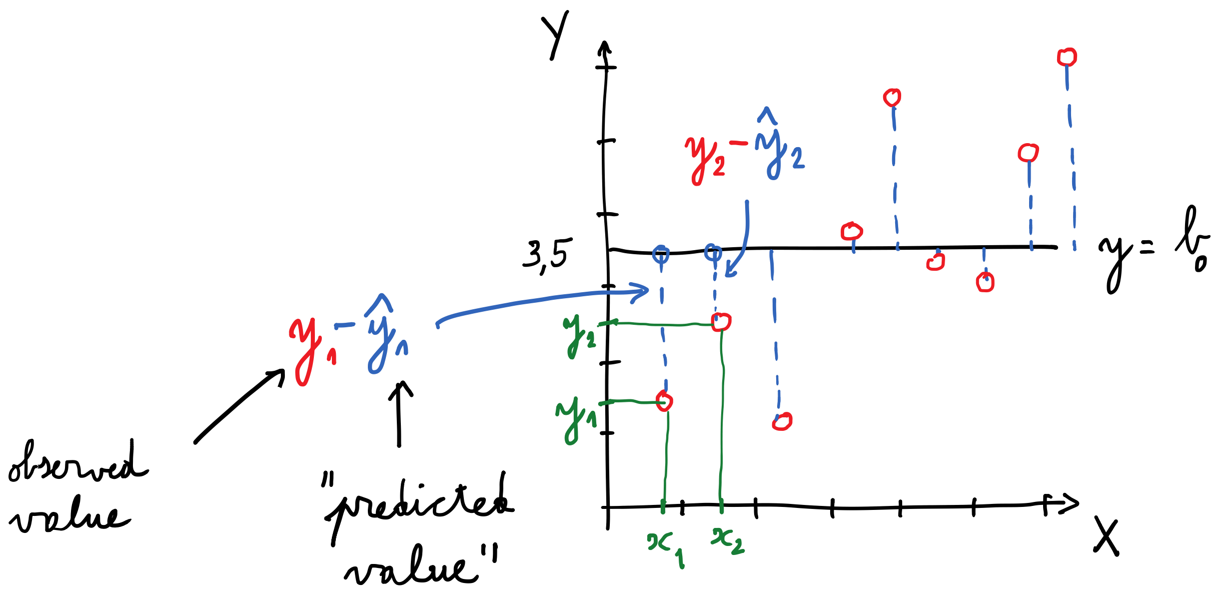
- We add up the distance of each point to the line.
- Squaring ensures that each term is positive.

OBSERVED y-value at x_1

$$(\boxed{y_1} - \boxed{\hat{y}_1})^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2+(y_6-\hat{y}_6)^2+(y_7-\hat{y}_7)^2+(y_8-\hat{y}_8)^2+(y_9-\hat{y}_9)^2$$

PREDICTED y-value at x_1

$$= (y_1 - b_0 + b_1 \times x_1)^2+(y_2 - b_0 + b_1 \times x_2)^2+(y_3 - b_0 + b_1 \times x_3)^2+(y_4 - b_0 + b_1 \times x_4)^2+(y_5 - b_0 + b_1 \times x_5)^2$$
$$+(y_6 - b_0 + b_1 \times x_6)^2+(y_7 - b_0 + b_1 \times x_7)^2+(y_8 - b_0 + b_1 \times x_8)^2+(y_9 - b_0 + b_1 \times x_9)^2$$



$$\begin{cases} b_0 = \bar{y} = 3{,}5 \\ b_1 = 0 \end{cases}$$

We can measure how well this line fits the data by seeing how close it is to the data points.

- We add up the distance of each point to the line.
- Squaring ensures that each term is positive.

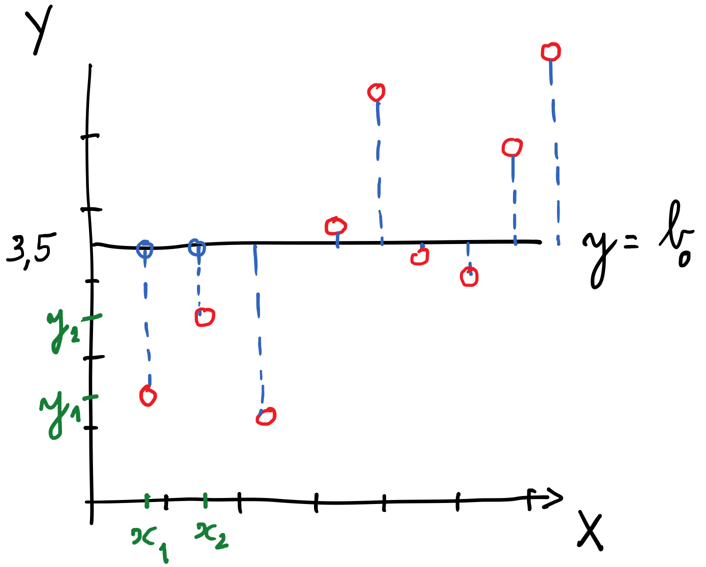
$$(y_1 - \hat{y}_1)^2+(y_2 - \hat{y}_2)^2+(y_3 - \hat{y}_3)^2+(y_4 - \hat{y}_4)^2+(y_5 - \hat{y}_5)^2+(y_6 - \hat{y}_6)^2+(y_7 - \hat{y}_7)^2+(y_8 - \hat{y}_8)^2+(y_9 - \hat{y}_9)^2$$

$$= (y_1 - b_0 + \cancel{b_1 \times x_1})^2+(y_2 - b_0 + \cancel{b_1 \times x_2})^2+(y_3 - b_0 + \cancel{b_1 \times x_3})^2+(y_4 - b_0 + \cancel{b_1 \times x_4})^2+(y_5 - b_0 + \cancel{b_1 \times x_5})^2$$
$$+(y_6 - b_0 + \cancel{b_1 \times x_6})^2+(y_7 - b_0 + \cancel{b_1 \times x_7})^2+(y_8 - b_0 + \cancel{b_1 \times x_8})^2+(y_9 - b_0 + \cancel{b_1 \times x_9})^2$$

$$= (y_1 - b_0)^2+(y_2 - b_0)^2+(y_3 - b_0)^2+(y_4 - b_0)^2+(y_5 - b_0)^2+(y_6 - b_0)^2+(y_7 - b_0)^2+(y_8 - b_0)^2+(y_9 - b_0)^2$$

$$= 24.62$$

Sum of squared errors (SSE)
= measure of how well the line fits the data.



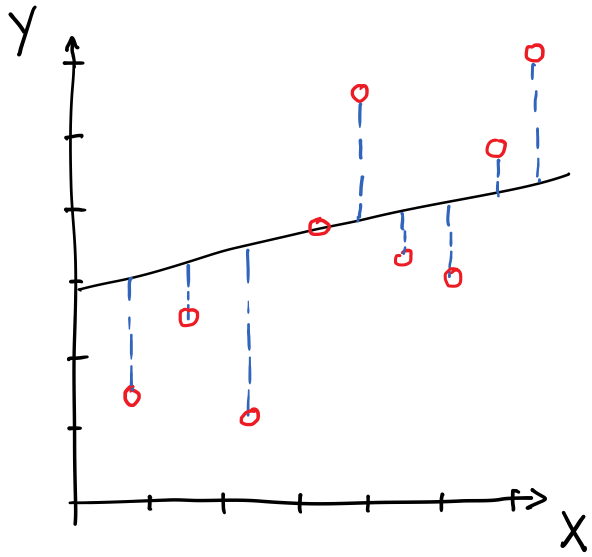
$$\begin{cases} b_0 = \bar{y} = 3,5 \\ b_1 = 0 \end{cases}$$

We want to find the optimal values for “b_0” and “b_1”, so that we minimize the sum of squared errors.

<u>= LEAST SQUARES method</u>

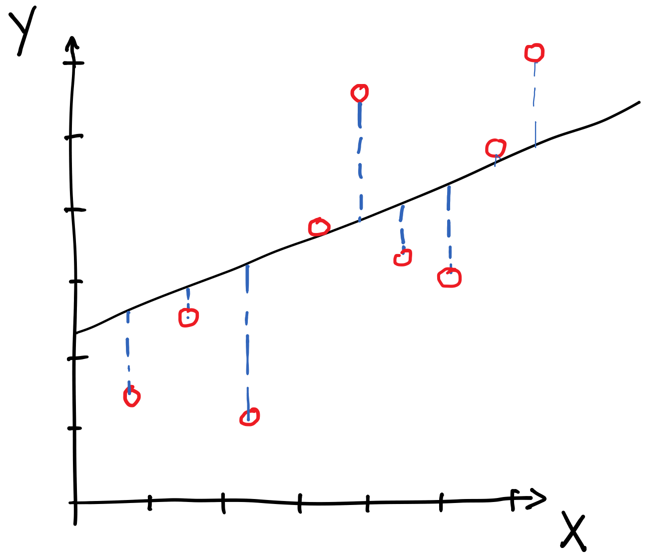
If we rotate the line a little bit :

The sum of squared errors = 18.72.
The fit gets **better**.



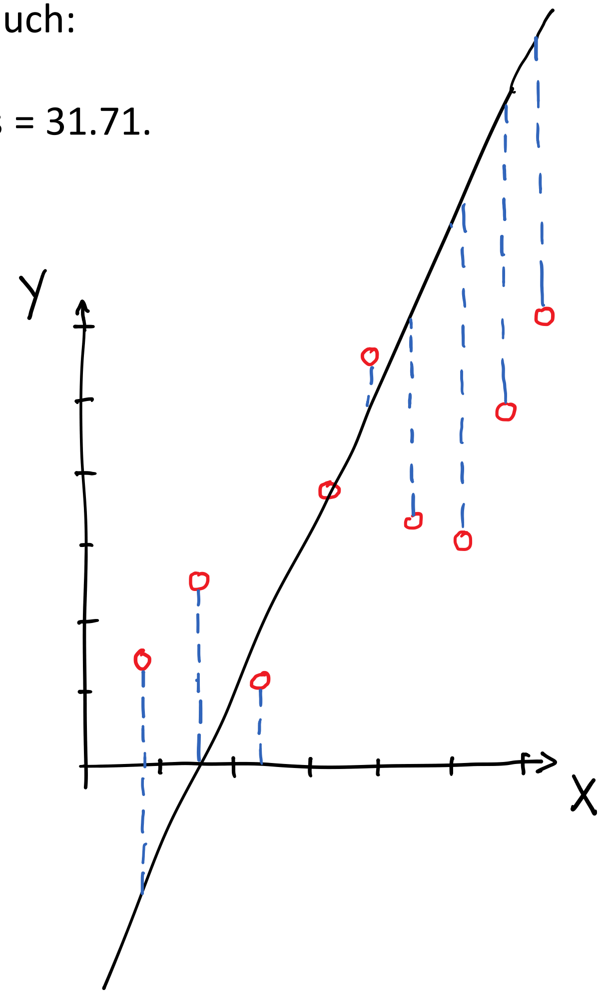
If we rotate the line a little more :

The sum of squared errors = 14.05.
The fit gets **even better**.

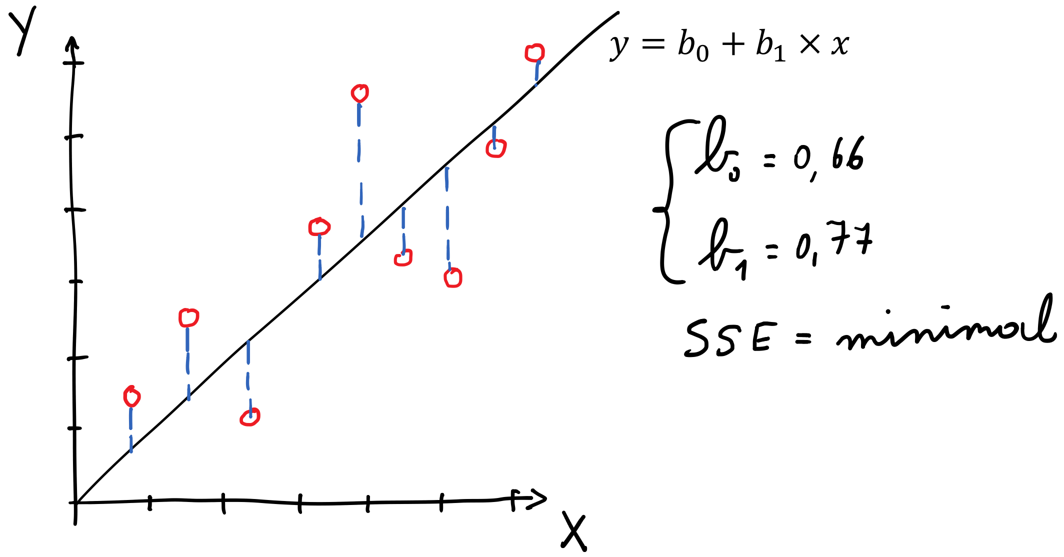


If we rotate the line too much:

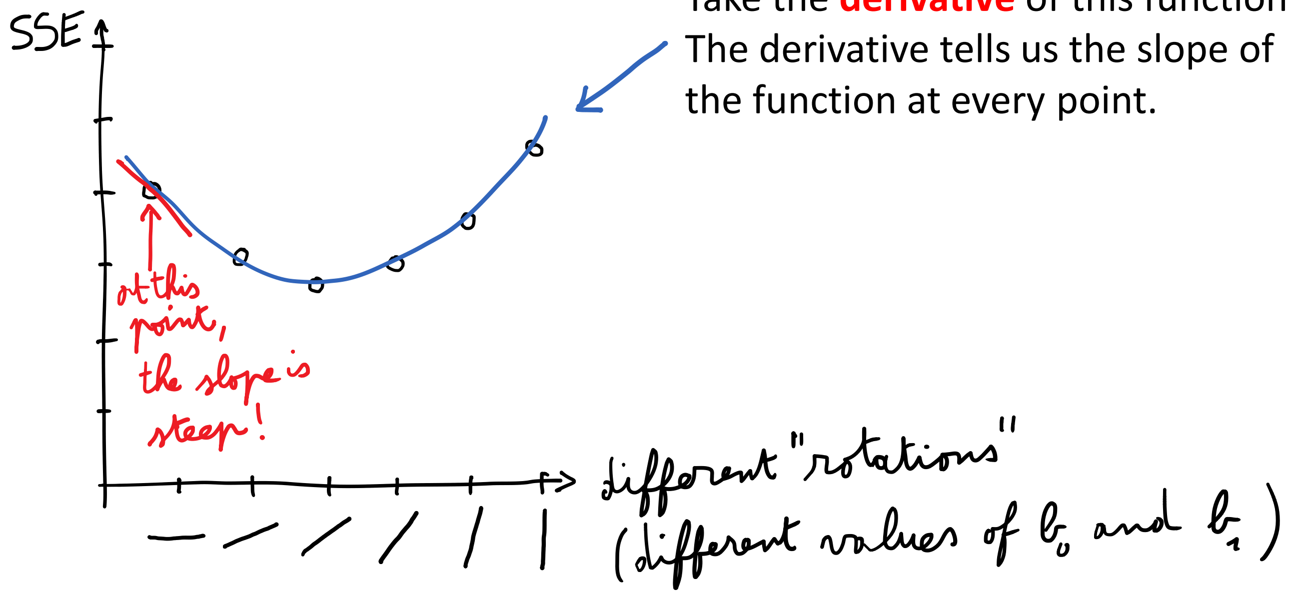
The sum of squared errors = 31.71.
The fit gets **worse**.



How do we find the optimal values for "b_0" and "b_1", so that we minimize the sum of squared errors (SSE)?



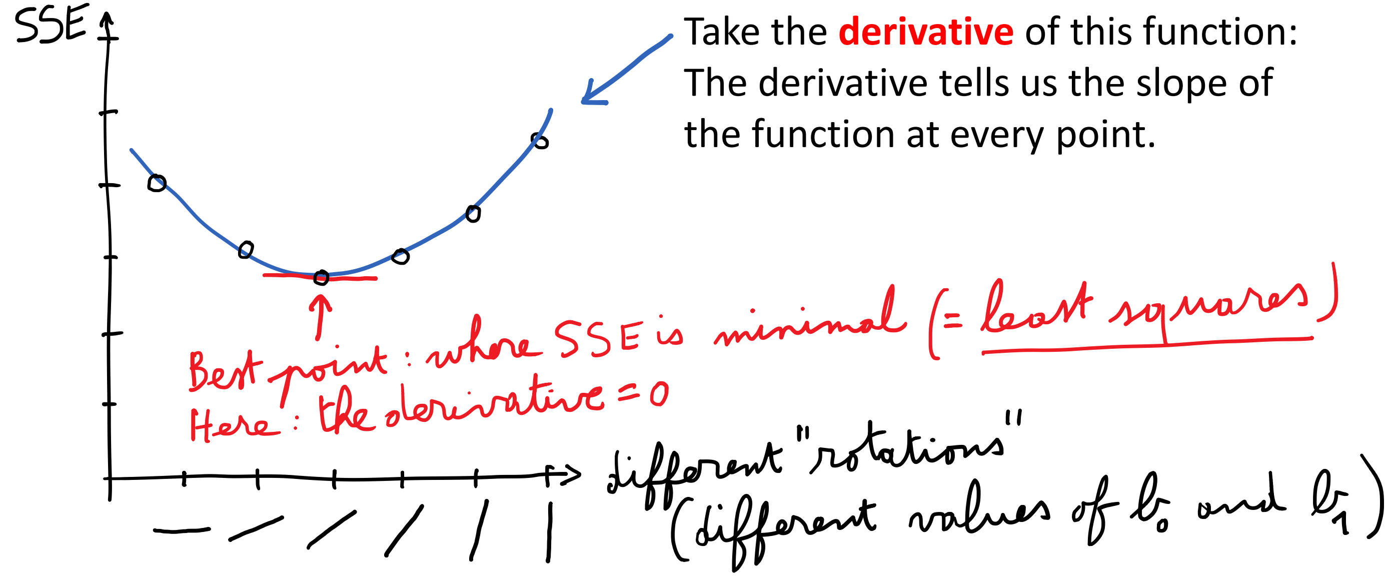
How do we find the **optimal rotation** of the line?



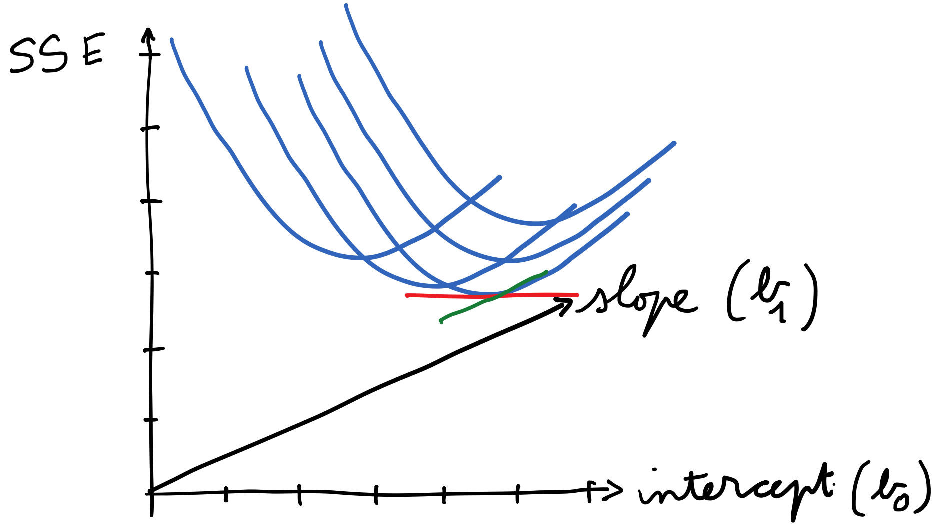
Take the **derivative** of this function:
The derivative tells us the slope of the function at every point.

How do we find the **optimal rotation** of the line?

Take the **derivative** of this function:
The derivative tells us the slope of the function at every point.



We can use a 3D graph to show how different values for intercept (b_0) and slope (b_1) result in different SSE.



1) Take partial derivatives of the SSE function
    - with respect to the intercept (b_0) and
    - with respect to the slope (b_1)

2) Set them equal to zero

3) and solve: deduct what are the values for b_0 and b_1, when the partial derivatives of SSE are zero (= when SSE is minimal).

Take **partial derivatives** with respect to *b0 and b1,* set them equal to zero and solve.

$$\frac{\partial SSE}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))$$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - b_1 \sum_{i=1}^{n} x_i$$

$$0 = \sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i$$

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

$$\frac{\partial SSE}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - (b_0 + b_1 x_i))$$

$$0 = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} b_0 x_i - b_1 \sum_{i=1}^{n} x_i^2$$

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} b_0 x_i$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} b_0 x_i}{\sum_{i=1}^{n} x_i^2}$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} (\bar{y} - b_1 \bar{x}) x_i}{\sum_{i=1}^{n} x_i^2}$$

With some further manipulations:

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s^2_{xy}(n-1)}{s_x^2(n-1)} = \frac{SPxy}{SSx}}$$

SPxy refers to the corrected ***sum of cross products for x and y***
SSx refers to the corrected ***sum of squares for x***

4. Simple Linear Regression

Idea is:

- variable of interest (dependent variable) y_i *; hard to* measure
- m "easy to measure" variables (predictor/ independent) that are related to the variable of interest, labeled x_{1i} , x_{2i},.....x_{mi}
- measure y_i, x_{1i},.....x_{mi} for a sample of *n items*
- use this sample to estimate an equation that relates y_i *(dependent variable) to* x_{1i},..x_{mi} *(independent or* predictor variables)
- once equation is fitted, one can then just measure the *x's, and get an estimate of y without measuring it*

o also can examine relationships between variables

Examples:

Sample of 9 trees (experimental units), in which height, Dbh (diameter at 1.3 m above ground in cm), and volume are measured:

- **Dependent variable Y** = Height (hard to measure)
- **Explanatory (independent) variable X** = Dbh (easy to measure; squared for a linear equation)

OR

- **Dependent variable Y** = Volume – hard to measure
- **Explanatory (independent) variables X_1** = Dbh and **X_2** = height

Types of equations

Simple Linear Equation:

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$$

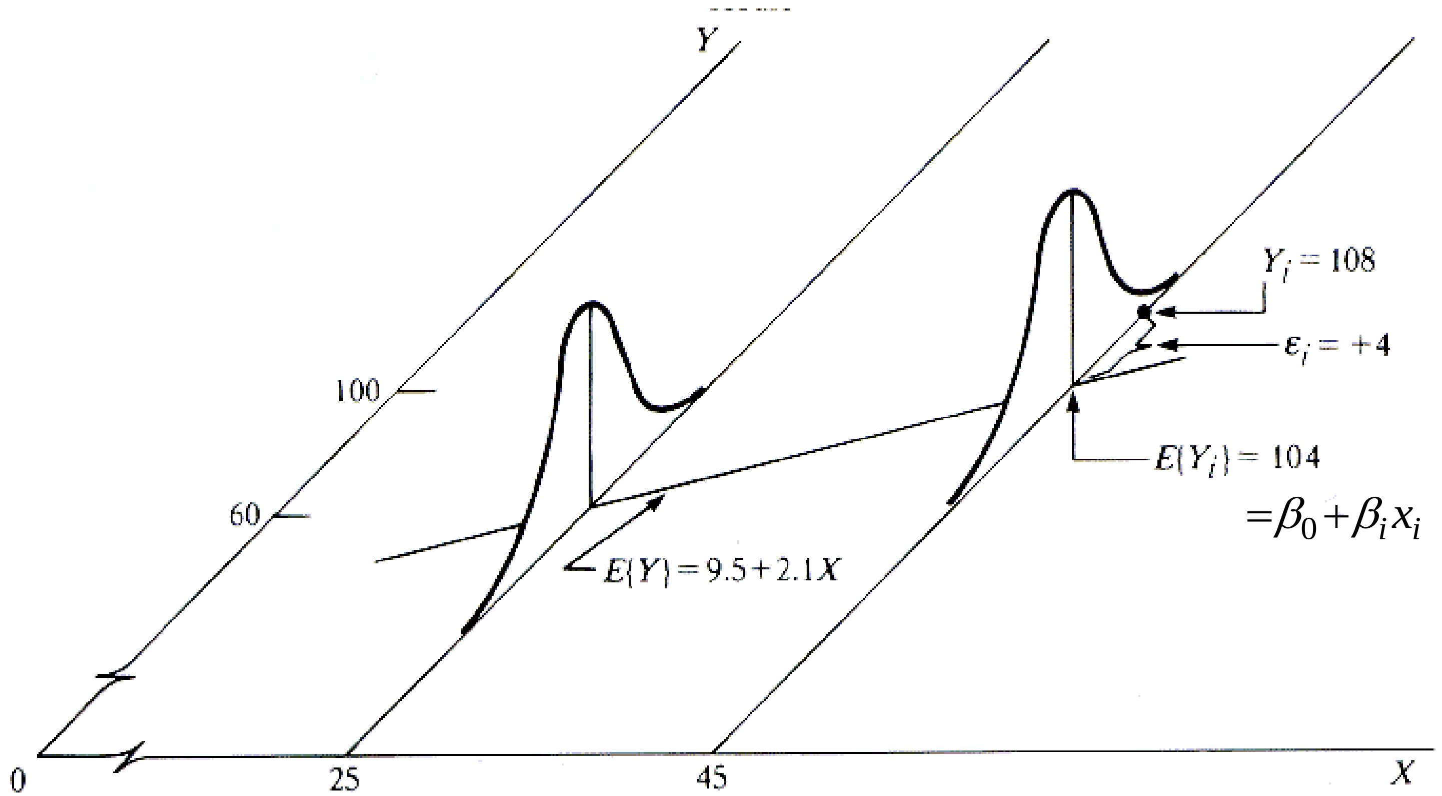
Multiple Linear Equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$$

Nonlinear Equation: takes many forms, for example:

$$y_i = \beta_0 + \beta_1 x_{1i}^{\beta_2} x_{2i}^{\beta_3} + \varepsilon_i$$

Simple Linear Regression (SLR)



The regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
implies that y_i comes from Normal probability distributions with

$$\text{mean} = E\{y_i\} = \beta_0 + \beta_1 x_i$$

and variances $\sigma^2 = \sigma^2\{y_i\} = \sigma^2\{\varepsilon_i\} =$ the same for all levels of x.

Simple Linear Regression (SLR)

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

$$E\{\varepsilon_i\} = 0$$

$$\Leftrightarrow E\{y_i\} = \beta_0 + \beta_1 x_i$$

E{}:
Read: "the expected value of …. is….."
Think: "the mean of it's sampling distribution is…."

$$\Leftrightarrow \text{For } x = x_i,$$

$$y_i \text{ comes from a probability distribution,}$$

$$\text{whose mean is } E\{y_i\} = \beta_0 + \beta_1 x_i$$

Objective:

Find estimates of β_0, β_1, β_2 ... β_m *such that the sum of* squared differences between measured y_i *and predicted* y_i *(usually labeled as* $\hat{y}_i$ *, values on the line or surface) is* the smallest (*minimize the sum of squared errors, called* least squared error).

OR

Find estimates of β_0, β_1, β_2 ... β_m *such that the likelihood* (probability) of getting these *y values is the largest* (*maximize the likelihood).*

Finding the minimum of sum of squared errors is often easier. In some cases, they lead to the same estimates of parameters.

Simple Linear Regression (SLR)

Population: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Sample: $y_i = b_0 + b_1 x_i + e_i \quad \hat{y}_i = b_0 + b_1 x_i \qquad e_i = y_i - \hat{y}_i$

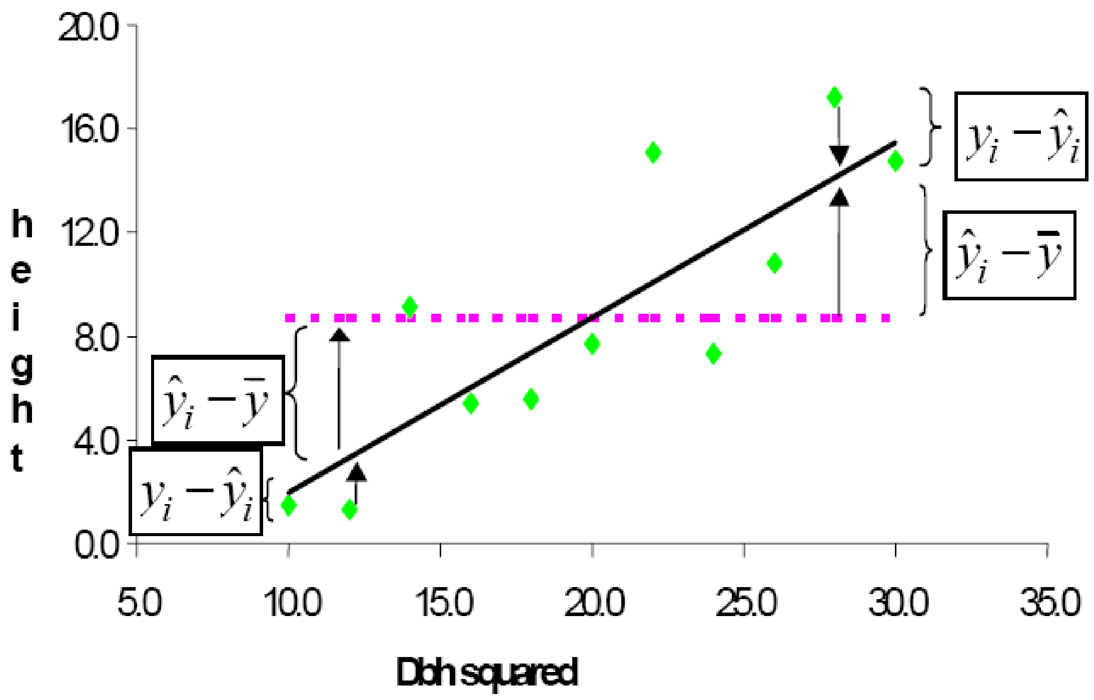
b_0 *is an estimate of* β_0 *[intercept]*

b_1 *is an estimate of* β_1 *[slope]*

$\hat{y}_i$ *is the predicted y; an estimate of the average for y for a* particular x value

e_i *is an estimate of* ε_i*, called the error or the residual;* represents the variation in the dependent variable (the *y)* which is not accounted for by predictor variable (the *x).*

Example: Tree Height (m) – hard to measure; Dbh (diameter at 1.3 m above ground in cm) – easy to measure – use Dbh squared for a linear equation



$y_i - \bar{y}$ Difference between measured y and the mean of y

$y_i - \hat{y}_i$ Difference between measured y and predicted y

$\hat{y}_i - \bar{y} = (y_i - \bar{y}) - (y_i - \hat{y}_i)$ Difference between predicted y and mean of y

Simple Linear Regression (SLR)

Find b_o *(intercept;* y_i *when* $x_i = 0$*) and* b_1 *(slope) so that*

$SSE=\Sigma e_i^2$ *(sum of squared errors over all n sample* observations) is the smallest (least squares solution)

- The variables do not have to be in the same units. Coefficients will change with different units of measure.
- Given estimates of b_o *and* b_1*, we can get an estimate of* the dependent variable (the *y) for ANY value of the x,* within the ranges of *x's represented in the original data.*

Least squares solution for SLR

Find the set of estimated parameters (coefficients) that minimize sum of squared errors

$$\min(SSE) = \min\left(\sum_{i=1}^{n} e_i^2\right) = \min\left(\sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right)^2\right)$$

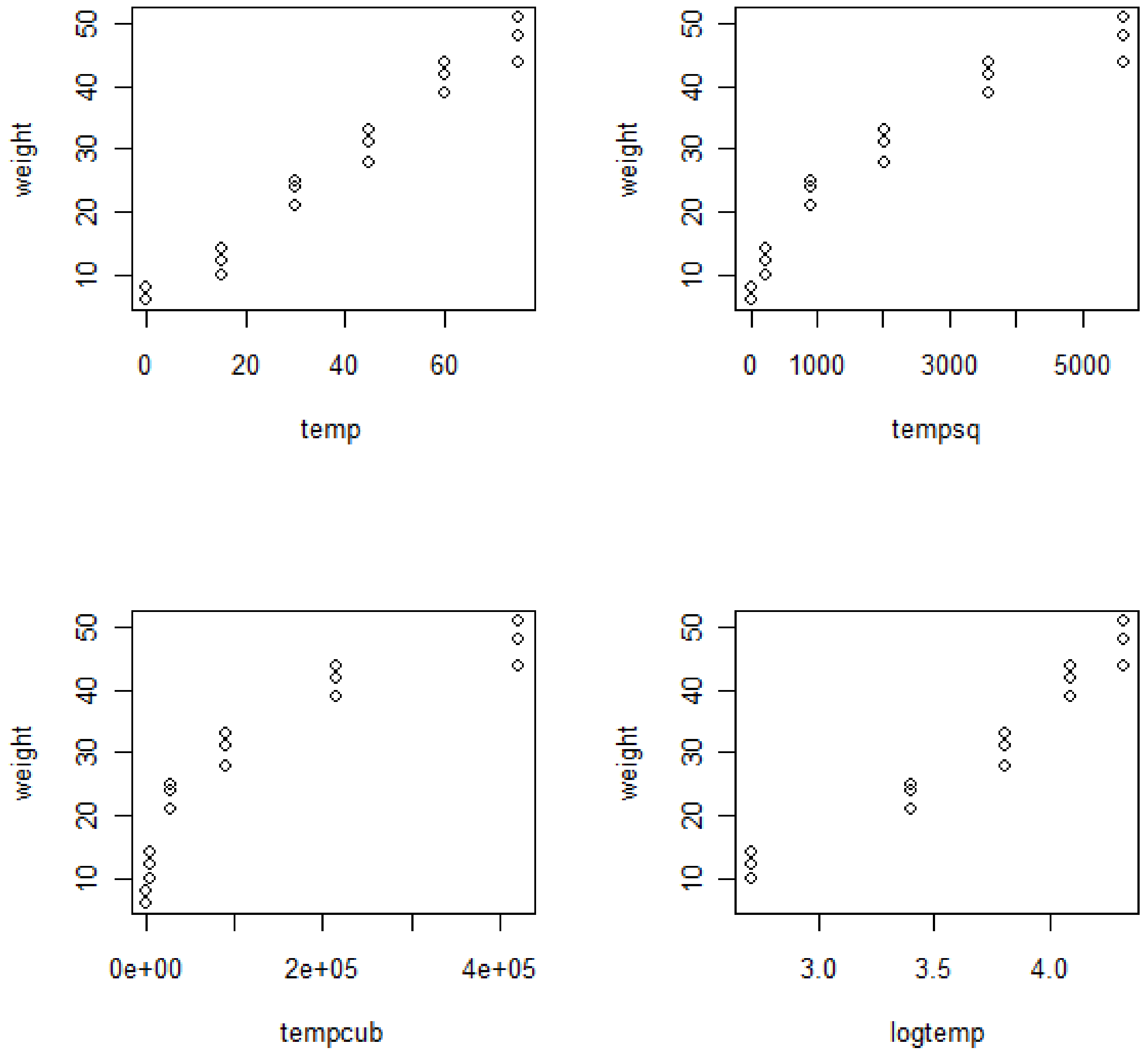
SLR example

Temperature (x)	Weight (y)	Weight (y)	Weight (y)
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

Observation	temp	weight
1	0	8
2	0	6
3	0	8
4	15	12
5	15	10
6	15	14
7	30	25
8	30	21

Et cetera…

Script 3_SLR.R


weight
10
20
30
40
50
0
20
40
60
temp
weight
10
20
30
40
50
0
1000
3000
5000
tempsq
weight
10
20
30
40
50
0e+00
2e+05
4e+05
tempcub
weight
10
20
30
40
50
3.0
3.5
4.0
logtemp

Estimate slope and intercept

Obs.	temp	weight	x-diff	x-diff. sq.
1	0	8	-37.50	1406.25
2	0	6	-37.50	1406.25
3	0	8	-37.50	1406.25
4	15	12	-22.50	506.25

Et cetera

mean	37.5	27.11

SSX=11,812.5 SSY=3,911.8 SPXY=6,705.0

$$b_1 = \frac{SPxy}{SSx} \qquad b_0 = \bar{y} - b_1 \times \bar{x}$$

b1:	0.567619
b0:	5.825397

NOTE: calculate b1 first, since this is needed to calculate b0.

calculate residuals for the equation and the sum of squared error (SSE)

Obs.	weight	y-pred	residual	residual sq.
1	8	5.83	2.17	4.73
2	6	5.83	0.17	0.03
3	8	5.83	2.17	4.73
4	12	14.34	-2.34	5.47

Et cetera

SSE:	105.89

SLR in R

```
> model <- lm(weight~temp)
> model

Call:
lm(formula = weight ~ temp)

Coefficients:
(Intercept)         temp
     5.8254       0.5676
```

- The output from lm, here model, is a linear model *object*, also called an lm object.

Script 3_SLR.R

- In R you can get a description of most objects when using the summary() function:

```
> summary(model)

Call:
lm(formula = weight ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3968 -1.6111 -0.0825  2.1389  4.1175 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.82540    1.07497   5.419 5.68e-05 ***
x            0.56762    0.02367  23.980 5.73e-14 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.573 on 16 degrees of freedom
Multiple R-squared: 0.9729,	Adjusted R-squared: 0.9712 
F-statistic: 575.1 on 1 and 16 DF,  p-value: 5.732e-14
```

- The linear model object, model, in this example, contains several attributes (sub-entries) that you can access using the names() function:

```
> names(model)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> model$coefficients
(Intercept)           x
   5.825397    0.567619
> model$residuals
         1          2          3          4          5          6
7
 2.1746032  0.1746032  2.1746032 -2.3396825 -4.3396825 -0.3396825
2.1460317
         8          9         10         11         12         13
14
-1.8539683  1.1460317 -0.3682540  1.6317460 -3.3682540  4.1174603 -
0.8825397
        15         16         17         18
 2.1174603 -0.3968254  2.6031746 -4.3968254
```

- However, there is a preferred way to access most of the common attributes. These are called accessor functions.

Item type	Sub-entry in linear model	Preferred way to access it
Model coefficients	model$coefficients	coefficients(model)
Residuals	model$residuals	residuals(model)
Predicted outputs, $\hat{y}$	model$fitted.values	fitted(model)

```
> coefficients(model)
(Intercept)           x
   5.825397    0.567619
```

Questions:

1. Are the assumptions of simple linear regression met? Evidence?
    - **If assumptions were not met, we would have to make some transformations and start over again!**
2. If so, interpret if this is a good equation based on goodness of fit measures.
3. Is the regression significant?

Assumptions of SLR

Once coefficients are obtained, we must **check the assumptions** of SLR. Assumptions must be met to:

- assess goodness of fit (i.e., how well the regression line fits the sample data)
- test significance of the regression and other hypotheses
- calculate confidence intervals and test hypothesis for the true coefficients (population)
- calculate confidence intervals for mean predicted *y* value given an *x value (i.e. for the predicted y* given a particular value of the *x)*

Need good estimates (unbiased or at least consistent) of the standard errors of coefficients and a known probability distribution to test hypotheses and calculate confidence intervals.

Checking **assumptions** using residual Plots

Assumptions of :

1. a linear relationship between the *y and the x;*
2. equal variance of errors;
3. independence of errors (independent observations); and
4. Normal distribution of errors

can be visually checked by using **RESIDUAL PLOTS**

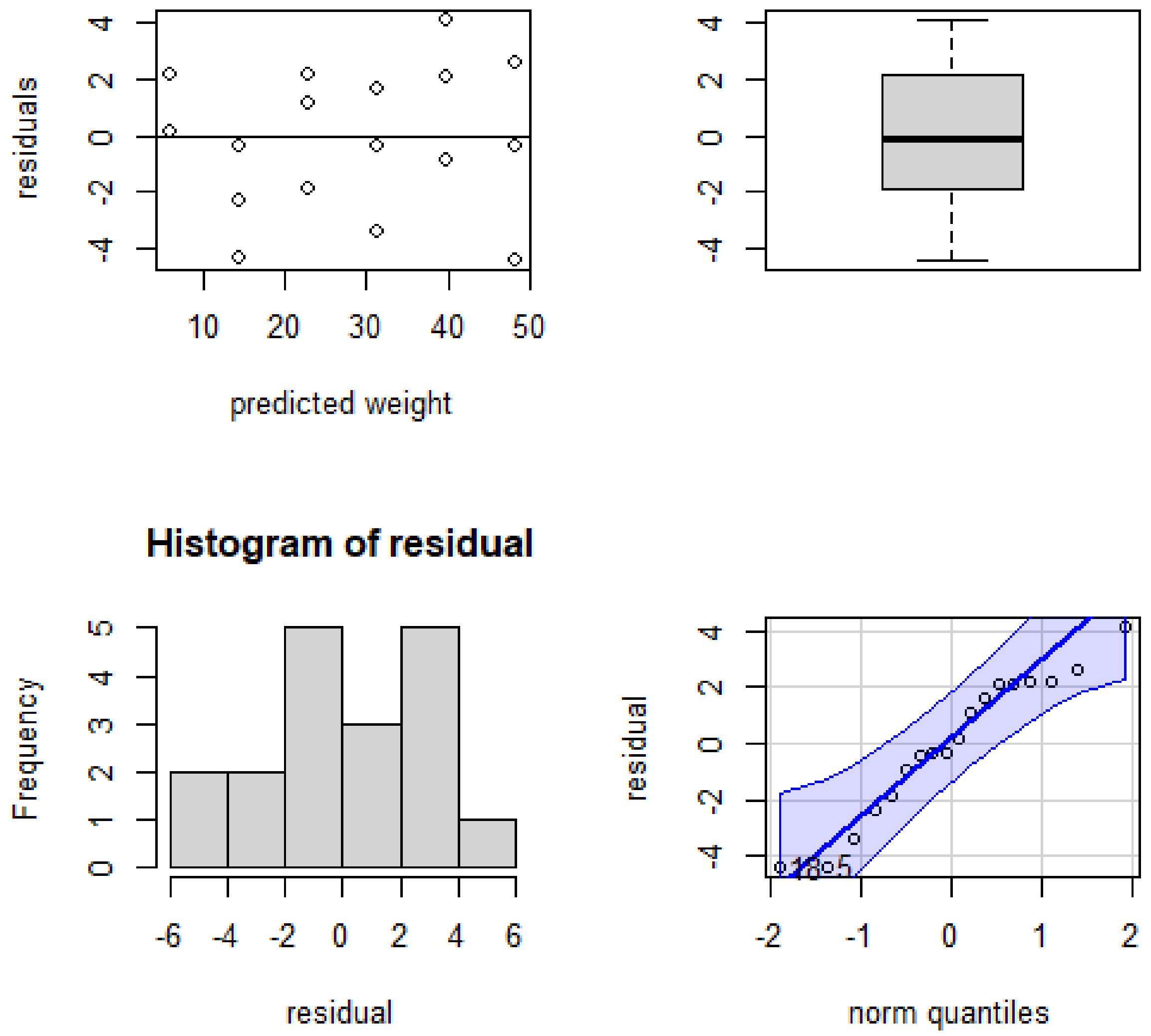
A residual plot shows the residual (i.e., y_i - $\hat{y}_i$ *) as the y-axis* and the predicted value ($\hat{y}_i$) as the x-axis.

Residual plots can also indicate unusual points (outliers) that may be measurement errors, transcription errors, etc.

Residual plots to test departures from SLR model

- Non-independence of errors
  - ✓Sequence plot of residuals (against time or adjacent geographic areas)
  - => An important variable has been omitted from the model
- Non-normality of errors
  - ✓ Distribution plots of the residuals: boxplot (histogram, dot plot, stem-and-leaf plot)
  - ✓ **Normal probability plot** of the residuals: ordered residuals are plotted against their expected value under normality

Residual plots



Are the assumptions met?

Checking if the residuals are normally distributed

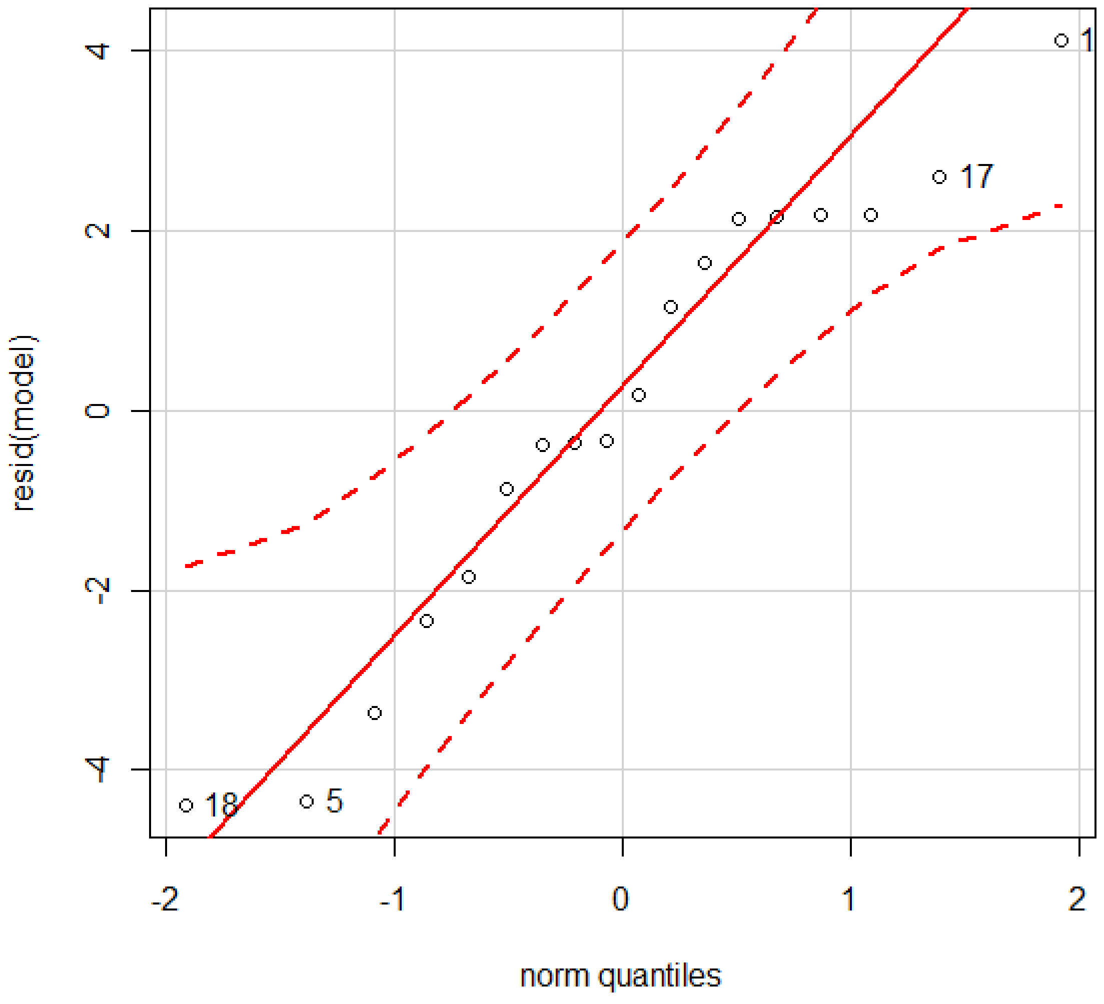
- The qqPlot(...) function will check that the residuals are normally distributed. You *first need to install* and load the car library though:

Non-linearity → non-normality of the residuals or non-constant variance (heteroscedasticity)

Use the mouse to click on the outliers and ID them

Right-click to stop adding points

```
> library(car)
> qqPlot(residuals(model), id.method = "identify")
[1] "18" "5"  "17" "13"
```


resid(model)
norm quantiles
4
2
0
-2
-4
-2
-1
0
1
2
1
17
18
5

Are the assumptions met?

Checking if the residuals are normally distributed

- **Shapiro-Wilk Normality Test**

H0: Residuals are Normal

Ha: Residuals are not Normal

```
> shapiro.test(residuals(model))

        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.9435, p-value = 0.3325
```

Residuals

- Unknown true error $\varepsilon_i = y_i - E\{y_i\}$

  Assumed to be
  - independent
  - Normally distributed with
    - mean 0 and
    - constant variance $\sigma^2\{\varepsilon_i\} = \sigma^2\{y_i\}$ (the same for all levels of x)
- **Residuals = observed error**: $e_i = y_i - \hat{y}_i$

  Should reflect the properties of ε_i *if the model is appropriate*

Properties of residuals

- Mean of n residuals e_i for a SLR model = 0

$$\bar{e} = \frac{\sum e_i}{n} = 0$$

=> no information about expected value of true errors

- Variance of n residuals e_i

  = MSE = “Error Mean Square” or “Residual Mean Square”

  n-2 degrees of freedom because β_0 and β_1 had to be estimated

$$s^2 = MSE = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2}$$

$$E\{MSE\} = \sigma^2 \text{ (means MSE is an unbiased estimator of } \sigma^2\text{)}$$

$$\mathrm{s} = \sqrt{\mathrm{MSE}}$$

- In R you can get a description of most objects when using the summary() function:

```
> summary(model)

Call:
lm(formula = weight ~ temp)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3968 -1.6111 -0.0825  2.1389  4.1175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.82540    1.07497   5.419 5.68e-05 ***
x            0.56762    0.02367  23.980 5.73e-14 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.573 on 16 degrees of freedom
Multiple R-squared: 0.9729,     Adjusted R-squared: 0.9712
F-statistic: 575.1 on 1 and 16 DF,  p-value: 5.732e-14
```

= SE_E
= root MSE

Semistudentized residuals

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

but $\sqrt{MSE}$ is an estimate of the standard deviation σ of the error terms ε_{i}

It is only an approximat ion of the standard deviation of the residuals e_i

Hence, we call the statistic e_i^* a *semistudentized residual*

Both studentize d and semistuden tized residuals can be very useful in identifyin g outlying observatio ns.

Residual plots to test departures from SLR model

- Non-linearity
  - ✓Residual plot against predictor x_i

  Assumption not met ⇔ curved line

  The regression line does not fit the data well; biased estimates of coefficients and standard errors of coefficients
- Non-consistency of error variance (=**heteroscedasticity**)
  - ✓ Residual plot against predictor x_i
  - ✓ Absolute Residual plot against predictor x_i
- Presence of Outliers
  - ✓Semi-studentized residuals:
    - how many standard deviations is each observation from the fitted value
    - Rule of thumb: Discard when $\left|\frac{e}{\sqrt{MSE}}\right| > 4$

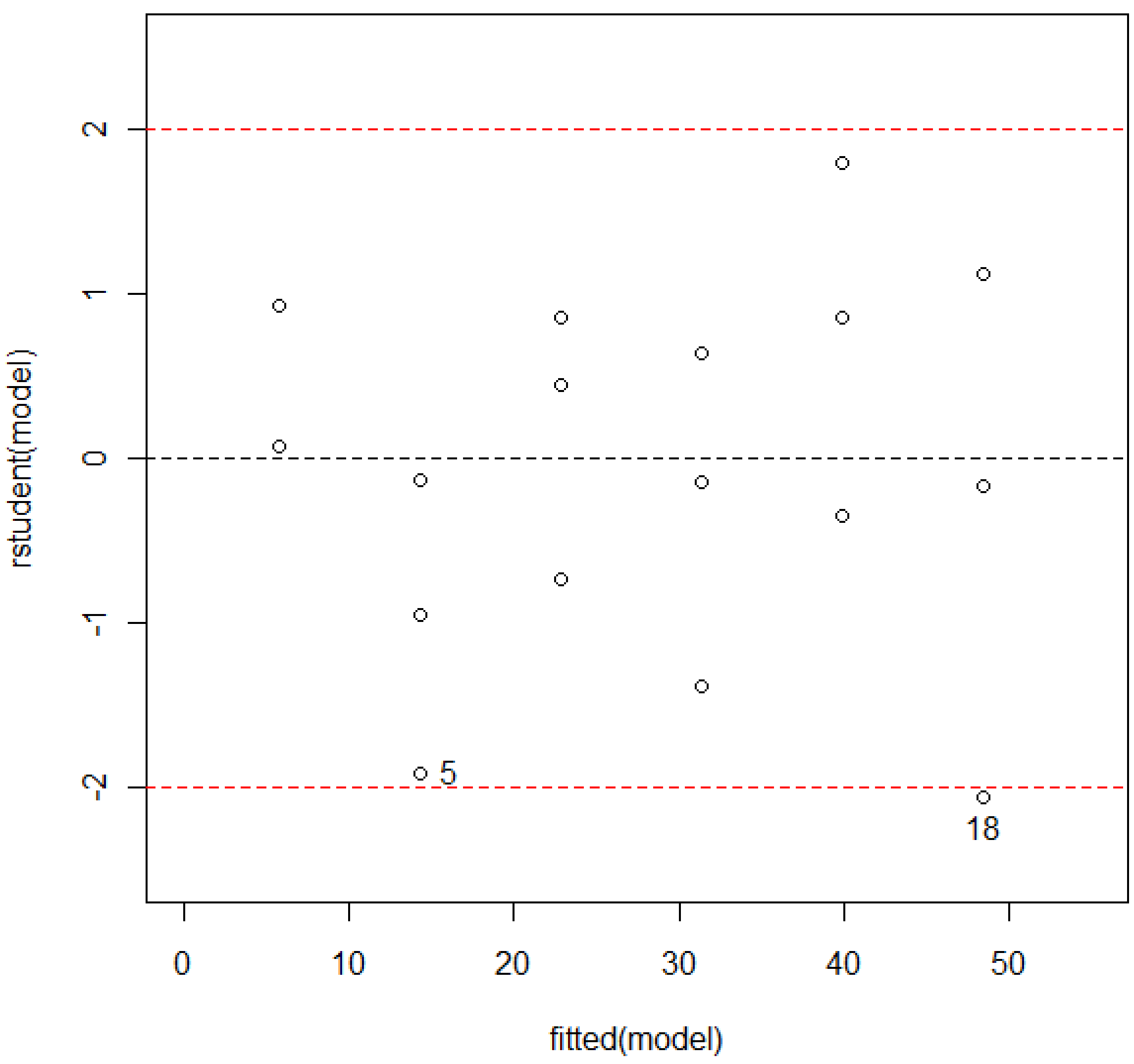
Are the assumptions met?

Plot Studentized Residuals against the fitted values

- Curvature
- non-constant variance (heteroscedasticity)
- outliers

The interval 0 ±2 contains 95% of the data (1 in 20 observations will naturally lie outside these limits). Observation 18 lies outside the limits and should be investigated.

```
> plot(fitted(model), rstudent(model), ylim=c(-2.5,2.5),
xlim=c(0,55))
> abline(h=0, lty=2)
> abline(h=c(-2,2), lty=2, col="red")
> identify(rstudent(model)~fitted(model))
[1]  5 18
```



rstudent(model)
2
1
0
-1
-2
5
18
0
10
20
30
40
50
fitted(model)

Are the assumptions met?

Plot Studentized residuals (Discrepancy) against Leverage (hat-values)

These are the two components of **Cook's Distance**, a statistic that reports the overall impact upon the parameter estimates of the observations

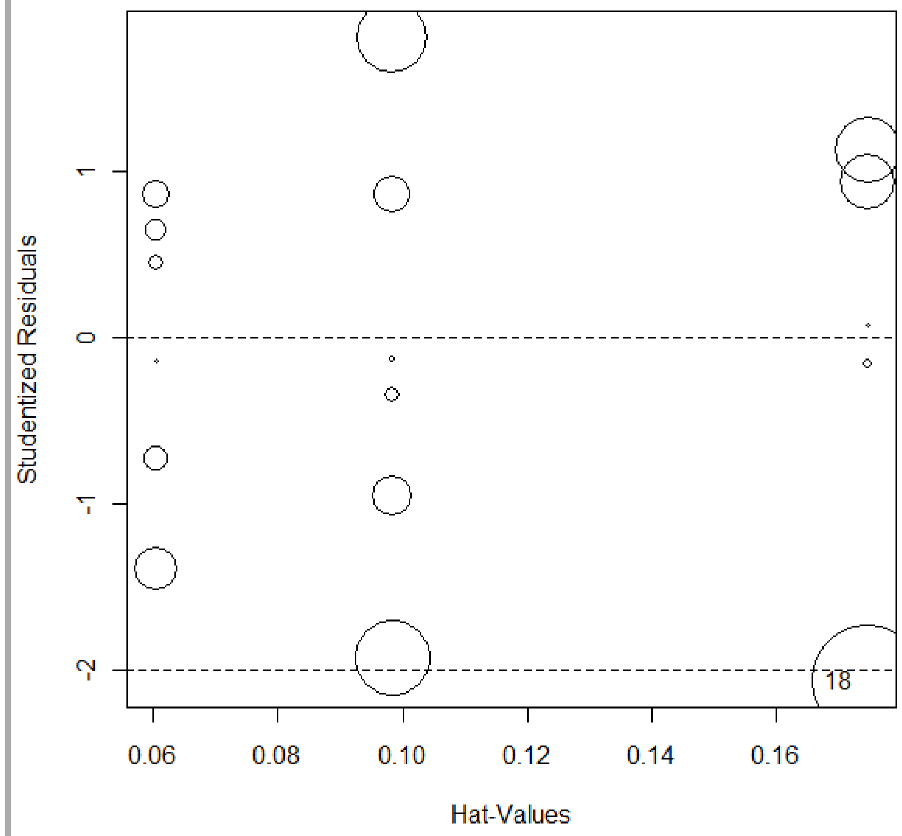
- Leverage: outlying values of the **independent** variables
- **Discrepancy** is measured by standardized residuals –> outliers

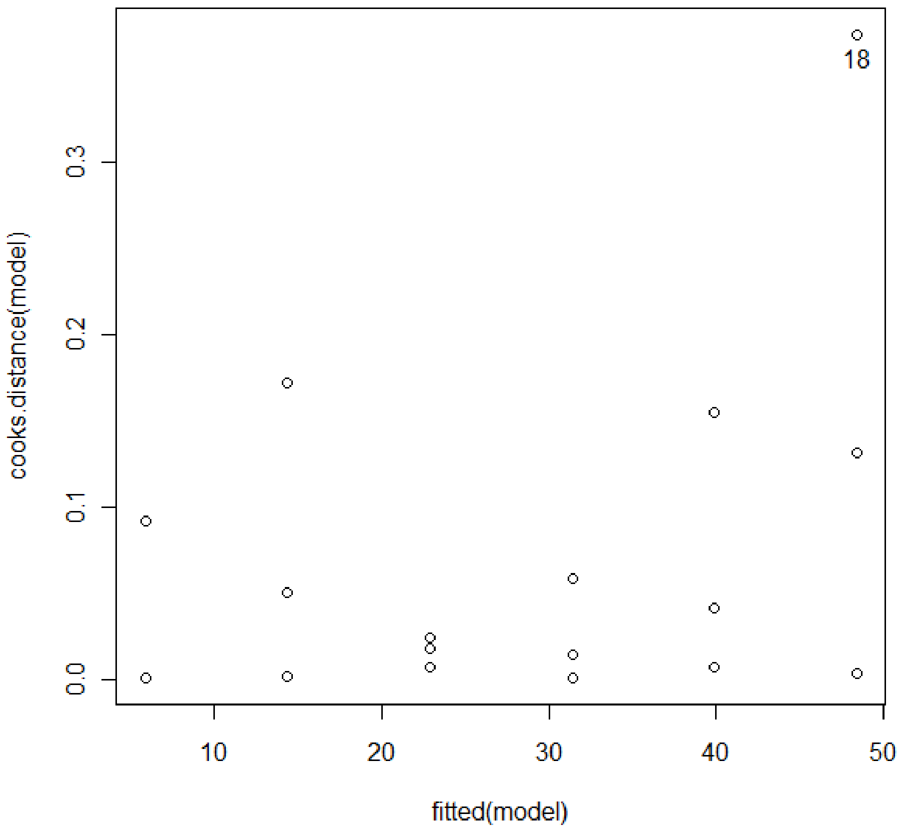
Plot Cook's Distance against the fitted values

Losely described as **leverage x discrepancy**.

Rule of thumb: Cook's Distances greater that 1 should attract our attention.

```
> influencePlot(model, id.method="identify")# areas of the
circles proportional to Cook's distances
StudRes       Hat     CookD
18 -2.063973 0.1746032 0.6118089
> plot(fitted(model), cooks.distance(model))
> cutoff <- 1
> abline(h=cutoff, lty=2, col="red")
> identify(fitted(model), cooks.distance(model))
[1] 18
```



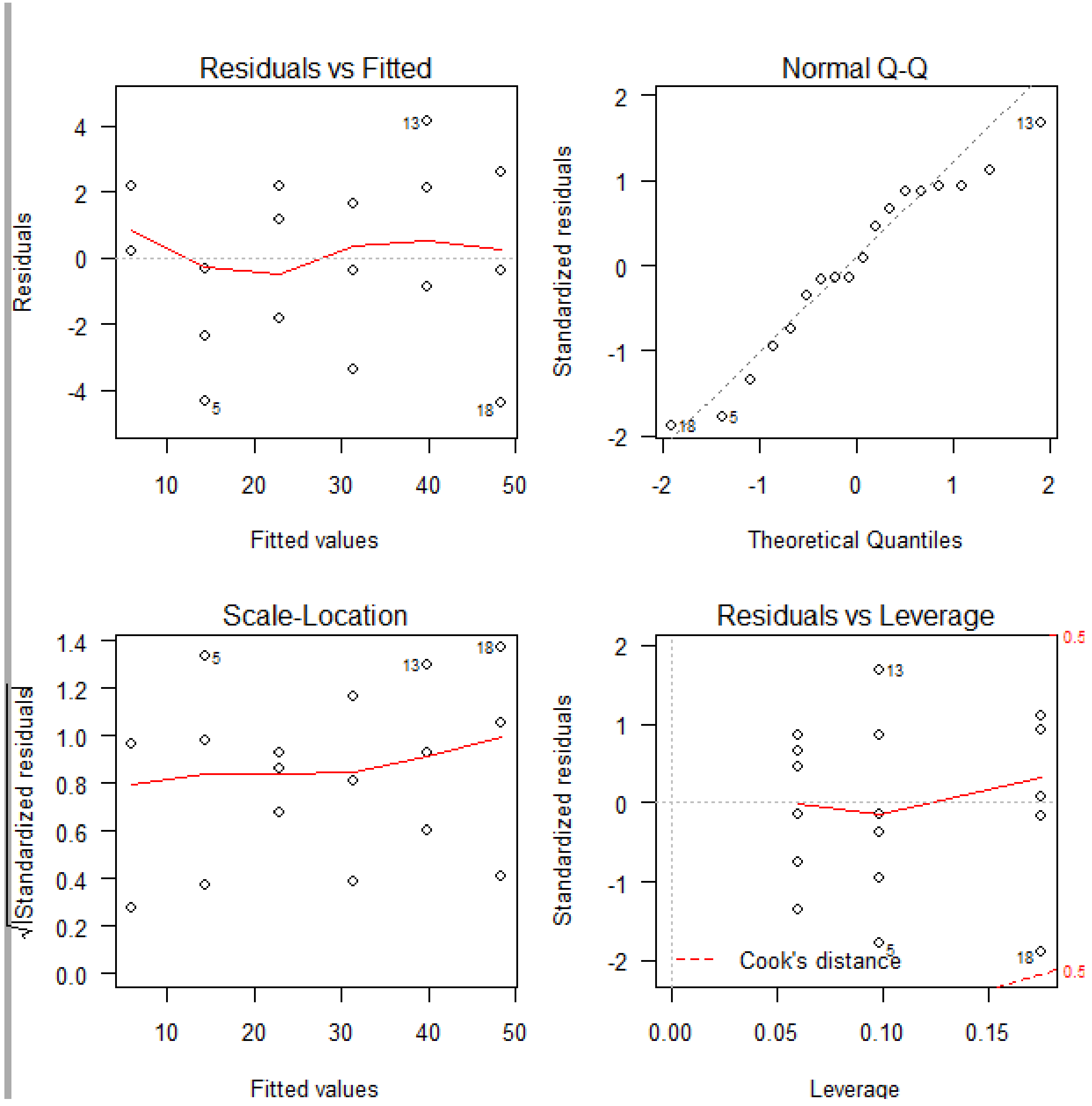


Hat-values
= leverage
= 'weighted' distance between x_i and the mean of x_i's
= between 0 and 1

The plot() function produces four graphs when applied to a linear model (lm).

1. **Residuals against the fitted values**, with a smooth red curve superimposed.
2. **Quantile plot of the standardized residuals against the normal distribution**.
3. **Square root of the standardized residuals against the fitted values**, along with a smooth red line.
4. **Leverage of the observations against the standardized¸residuals**. Contours of these distances (isoCooks?) at 0.5 and 1.0 are added to the graph to assist interpretation.

```
> opar <- par(mfrow = c(2, 2), mar = c(4, 4, 3, 1), las =
1)
> plot(model)
> par(opar)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Removing outliers and rebuilding the model

After investigation of the points, we may decide to remove point 5 and 18 and rebuild the model:

```
> remove = -c(5, 18)
> remove
[1]  -5 -18
> model.rebuild <- lm(model, subset=remove)
```

Measurements and sampling assumptions

The remaining assumptions are based on the measurements and collection of the sampling data.

5. *The x values are measured without error (i.e., the x values are fixed).*

   This can only be known if the process of collecting the data is known. For example, if tree diameters are very precisely measured, there will be little error. If this assumption is not met, the estimated coefficients (slopes and intercept) and their variances will be biased, since the x values are varying.

Measurements and sampling assumptions

The remaining assumptions are based on the measurements and collection of the sampling data.

6. *The y values are randomly selected for value of the x variables (i.e., for each x value, a list of all possible y values is made, and some are randomly selected).*

   For many biological problems, the observations will be gathered using simple random sampling or systematic sampling (grid across the land area). This does not strictly meet this assumption. Also, more complex sampling design such as multistage sampling (sampling large units and sampling smaller units within the large units), this assumption is not met. If the equation is “correct”, then this does not cause problems. If not, the estimated equation will be biased.

Questions:

1. Are the assumptions of simple linear regression met? Evidence?
   - **If assumptions were not met, we would have to make some transformations and start over again!**
2. If so, interpret if this is a good equation based on **goodness of fit** measures.
3. Is the regression significant?

Measures of Goodness of Fit

How well does the regression fit the sample data?

- For simple linear regression, a graph of the original data with the fitted line marked on the graph indicates how well the line fits the data [not possible with MLR]
- Two measures commonly used to describe the degree of linear association: **coefficient of determination (r^2)** and **standard error of the estimate(SE_E)**.

To calculate r^2 and SE_E

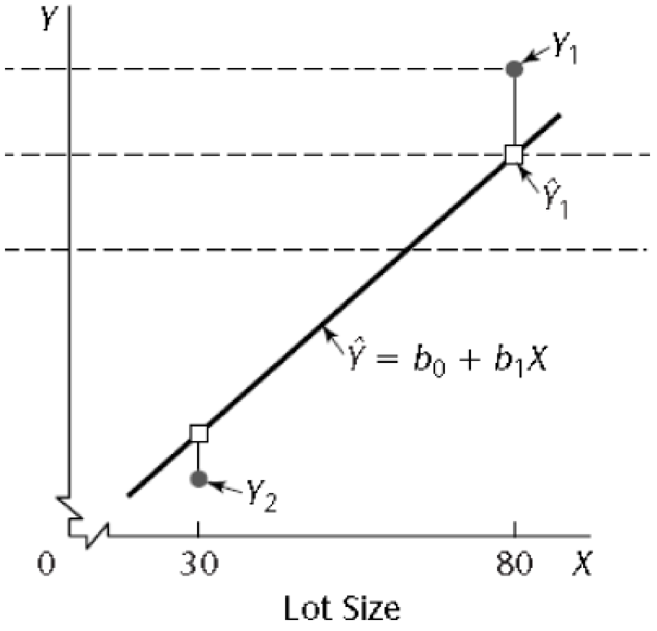
First calculate SSE, SSy and SSreg:

- ***SSE*** (this is what was minimized):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

The sum of squared differences between the measured and estimated *y's.*

If all Y_i observations fall on the regression line ⇔ SSE = 0



calculate residuals for the equation and the sum of squared error (SSE)

Obs.	weight	y-pred	residual	residual sq.
1	8	5.83	2.17	4.73
2	6	5.83	0.17	0.03
3	8	5.83	2.17	4.73
4	12	14.34	-2.34	5.47

Et cetera

SSE:	105.89

To calculate r^2 and SE_E

- ***SSy***: the sum of squares for *y:*

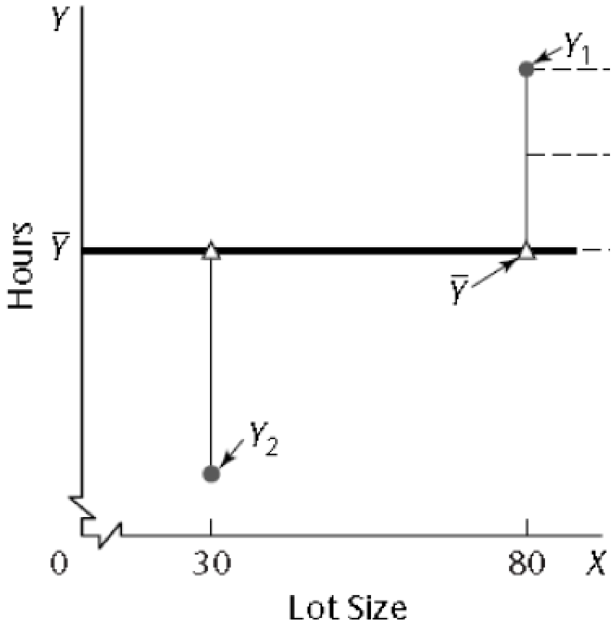
$$SSy = \sum_{i=1}^{n} (y_i - \bar{y})^2 = s_y^{\ 2}(n-1)$$

The sum of squared difference between the measured y and the mean of y-measures.

NOT taking predictor variable into account

If all Y_i observations are equal ⇔ SSy = 0

NOTE: In some texts, this is called the **sum of squares total (SSTO)**.



To calculate r^2 and SE_E

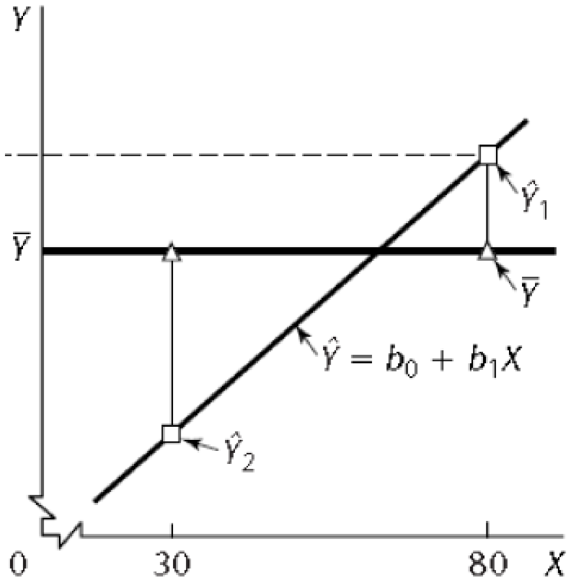
- ***SSreg***: the sum of squares regression:

$$SSreg = \sum_{i=1}^{n} \left(\bar{y} - \hat{y}_i\right)^2 = b_1 SPxy = SSy - SSE$$

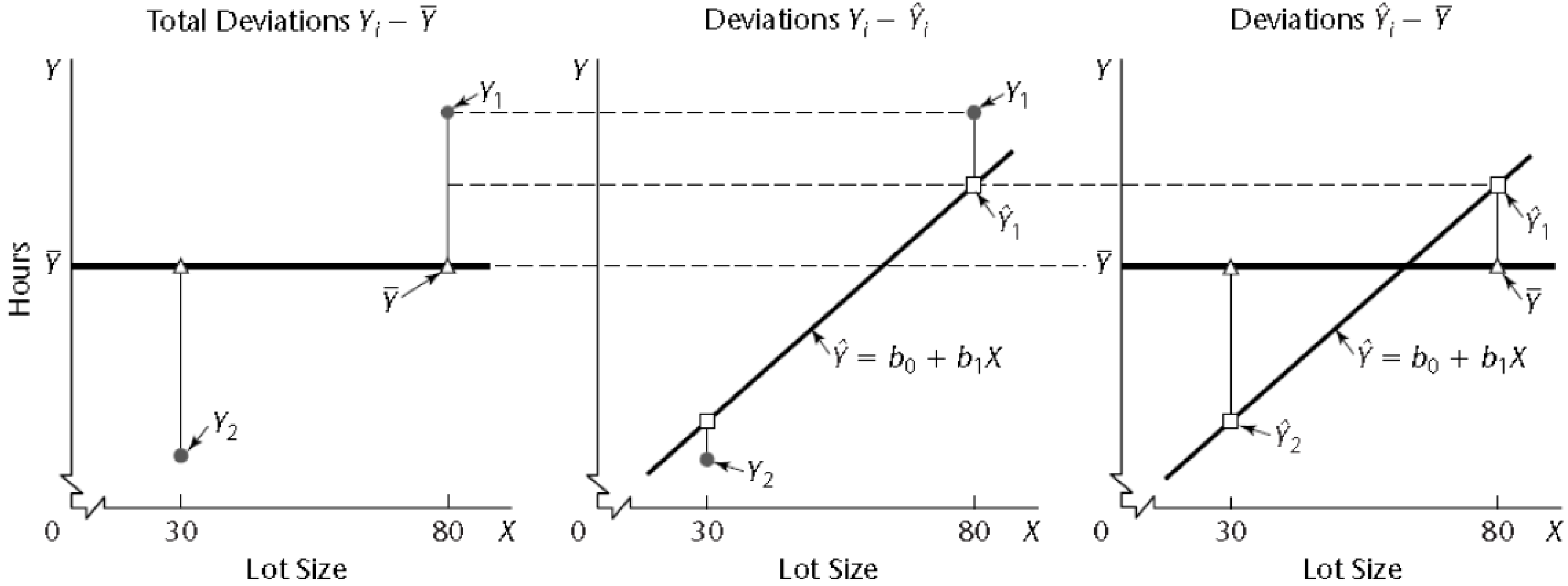
The sum of squared differences between the mean of y-measures and the predicted *y's from the fitted equation.*

= the sum of squares for *y – the sum of squared* errors.

➔ Part of variability of Y_i which is accounted for by the regression line



Partitioning of total deviations



$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Deviation of fitted regression value around mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around fitted regression line}}$$

Remarkable property

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$$

Or *SSy = SSreg + SSe*

Remarkable property

- (this can be proven:)

Proof: $$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$$

$$\begin{aligned}(Y_i - \bar{Y})^2 &= \sum[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum[(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\ &= \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\end{aligned}$$

but

$$\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum \hat{Y}_i(Y_i - \hat{Y}_i) - \sum \bar{Y}(Y_i - \hat{Y}_i) = 0$$

Breakdown of Degrees of Freedom

- SSY (=SSTO)
  - 1 linear constraint due to the calculation and inclusion of the mean (equivalently: because sum must be 0)
    - n-1 degrees of freedom
- SSE
  - 2 linear constraints arising from the estimation of β_0 and β_1
    - n-2 degrees of freedom
- SSR
  - All fitted values are calculated from the same regression line: Two degrees of freedom in the regression parameters, one is lost due to linear constraint
    - 1 degree of freedom

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$$

Remarkable: n – 1 = (n - 2) + 1

r^2

$$r^2 = \frac{SSy - SSE}{SSy} = 1 - \frac{SSE}{SSy} = \frac{SSreg}{SSy}$$

- SSE, SSY are based on *y's used in the equation* - will not be in original units if *y was transformed*
- **r^2 = coefficient of determination**; proportion of variance of *y, accounted for by the regression using x*
- Is the square of the correlation between *x and y*
- 0 (very poor – horizontal surface representing no relationship between y and x's) to 1 (perfect fit – surface passes through the data)

r^2

between 0 and 1

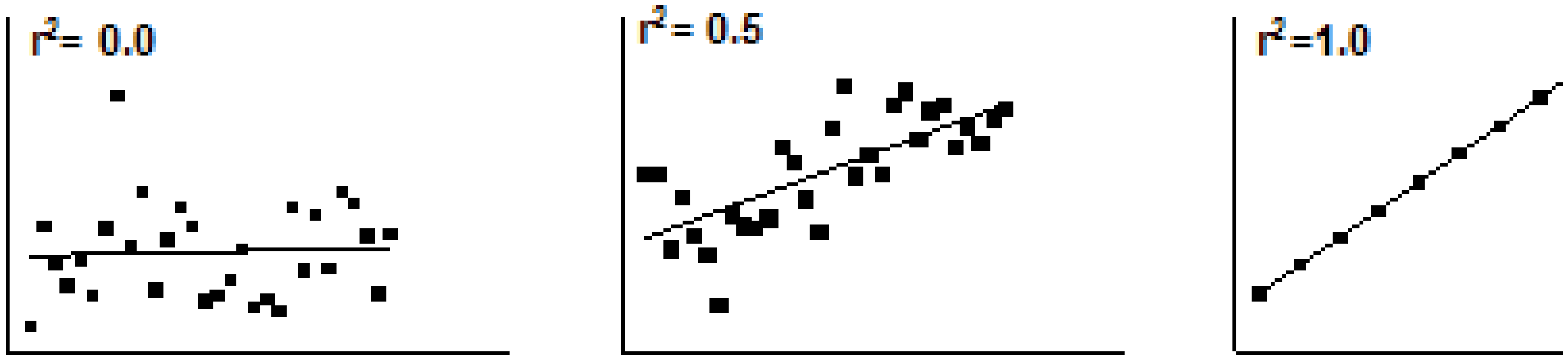
no units

$r^2 = 0$ means that knowing X does not help you predict Y.

There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values.

$R^2 = 1$ means that knowing X lets you predict Y perfectly

All points lie exactly on a straight line with no scatter..



SE_E = Standard Error of the Estimate

$$SE_E = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

= root of the mean square error (MSE)

- SSE is based on *y's used in the equation – will not* be in original units if *y was transformed*
- SE_E - standard error of the estimate; in same units as y
- Under normality of the errors:
  - ±1 $SE_E \cong$ 68% of sample observations
  - ±2 $SE_E \cong$ 95% of sample observations
  - Want low SE_E

- In R you can get a description of most objects when using the summary() function:

```
> summary(model)

Call:
lm(formula = weight ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3968 -1.6111 -0.0825  2.1389  4.1175 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.82540    1.07497   5.419 5.68e-05 ***
x            0.56762    0.02367  23.980 5.73e-14 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.573 on 16 degrees of freedom
Multiple R-squared: 0.9729,     Adjusted R-squared: 0.9712 
F-statistic: 575.1 on 1 and 16 DF,  p-value: 5.732e-14
```

= SE_E
= root MSE

= r^2

y-variable was transformed

- Can calculate estimates of r^2 and SE_E for the original y-variable unit, in order to compare to r^2 and SE_E of other equations where the *y* was not transformed.

- Estimated r^2: **I^2 (Fit Index)**

$$I^2 = 1 - SSE/SSY$$

  - where SSE, SSY are in original units. NOTE must "back-transform" the predicted *y's to calculate the* SSE in original units.
  - Does not have the same properties as r^2, however:
    - it can be less than 0
    - it is not the square of the correlation between the y (in original units) and the *x used in the* equation.

y-variable was transformed

- Estimated standard error of the estimate (SE_E') *:*

$$SE_E' = \sqrt{\frac{SSE(original\ units)}{n-2}}$$

  - SE_E' - standard error of the estimate ; in same units as original units for the dependent variable
  - want low SE_E'

Questions:

1. Are the assumptions of simple linear regression met? Evidence?
   - **If assumptions were not met, we would have to make some transformations and start over again!**
2. If so, interpret if this is a good equation based on **goodness of fit** measures.
3. Is the regression significant?

Testing whether the Regression is Significant

- Does knowledge of x improve the estimate of the mean of y?
- Or is it a flat surface, which means we should just use the mean of y as an estimate of y for any x?

Mean Square (MS)

= Sum of Squares divided by it’s associated degrees of freedom

- MSE = SSE/ (*n-2):*

Called the **Mean squared error**, as would be the average of the squared error if we divided by *n.*

Instead, we divide by *n-2. Why? The degrees of freedom* are *n-2; n observations with two statistics estimated from* these, *b0 and b1*

Under the assumptions of SLR, is an unbiased estimate of the true variance of the error terms (error variance)

- MSR = SSR/1:

Called the **Mean Square Regression**

Degrees of Freedom=1

Under the assumptions of SLR, this is an estimate of the error variance PLUS a term of variance explained by the regression using *x.*

Mean Square (MS)

The Mean Squares are NOT additive

$$\frac{SSy}{n-1} \neq \frac{SSreg}{1} + \frac{SSE}{n-2}$$

$$\neq MSR + MSE$$

ANOVA table for simple lin. regression

Source of Variation	SS	df	MS	E{MS}
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	MSR = SSR/1	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	n-2	MSE = SSE/(n-2)	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	n-1		

Expected Mean Squares

(statistical theory provides these results:)

$$E\{MSE\} = \sigma^2$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \overline{X})^2$$

$$= \sigma^2 + SSreg$$

➔mean of sampling distribution of MSE = σ^2

regardless of linear relationship X~Y

i.e. regardless of $\beta_1 = 0$ or $\beta_1 \neq 0$

➔mean of sampling distribution of MSR is also σ^2 <u>if</u> $\beta_1 = 0$

OR: When $\beta_1 = 0$, the sampling distributions of MSR and MSE tend to be the same

Regression significant?

H0: Regression is not significant

H1: Regression is significant

Same as:

H0: $\beta_1 = 0$ [true slope is zero meaning no relationship with x]

H1: $\beta_1 \neq 0$ [slope is positive or negative, not zero]

This can be tested using an F-test, or with a t-test since we are only testing one coefficient (more on this later)

Analysis of Variance approach

- When $\beta_1 = 0$ [true slope is zero, no relationship with x],
  MSE and MSR sampling distributions are located identically
  MSE and MSR will tend to be of the same order of magnitude

- When $\beta_1 \neq 0$ [true slope is positive or negative, not zero],
  MSR tends to be larger than MSE

- **F-test of $\beta_1 = 0$ *vs.* $\beta_1 \neq 0$**

F-test for $\beta_1 = 0$ *vs.* $\beta_1 \neq 0$

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

ANOVA test statistic

$$F^* = \frac{MSR}{MSE}$$

Essentially the ratio of explained and unexplained variance
-> F-test

- Large values for F^* support H_a
- Values for F^* near 1 support H_0

→Upper tail test

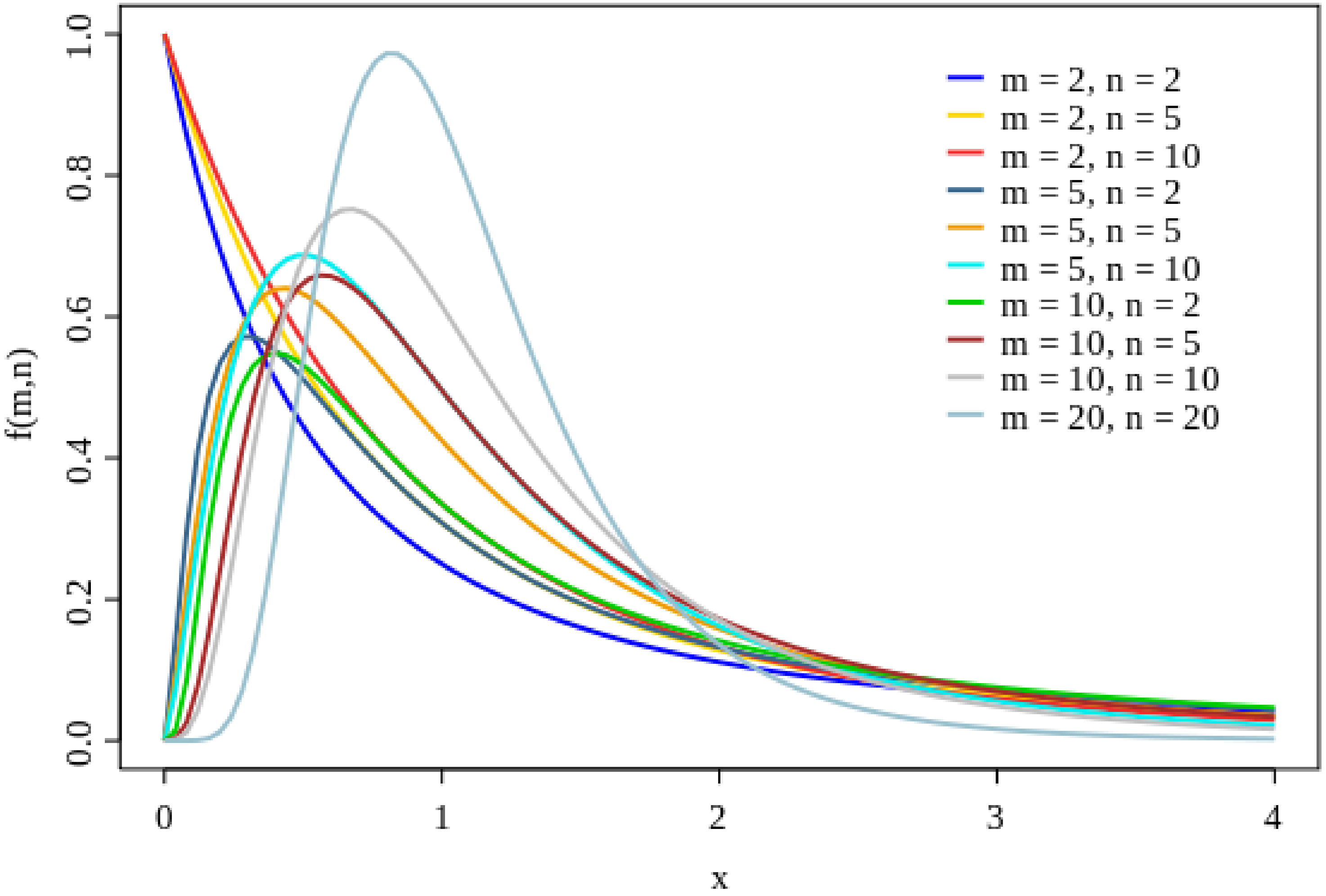
If H_0 holds true, (it can be shown that) F^* follows the $F(1, n-2)$ distribution

F-distribution

- Is the ratio of 2 variables that each have a χ^2-distribution eg. The ratio of 2 sample variances for variables that are each normally distributed.
- Is a family of curves based on the d.f. of numerator and d.f. of denominator
- Never negative
- **Mean is approximately 1**
- Need the percentile, and two degrees of freedom (one for the numerator and one for the denominator)

➢ Each alpha value involves a separate table

F-distribution



Hypothesis test decision rule

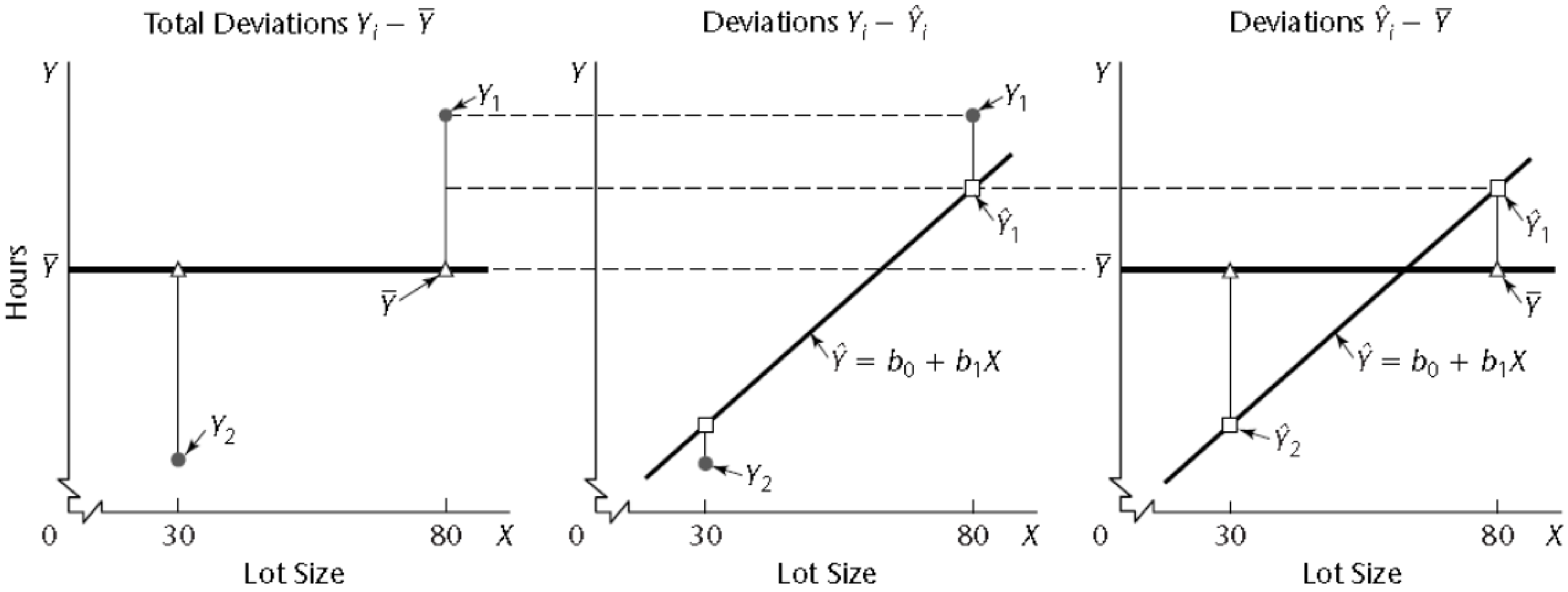
Since F^* is distributed as $F(1,n-2)$ when H_0 holds, the decision rule to follow when the risk of a Type I error is to be controlled at α is:

If $F^* \leq F(1-\alpha; 1, n-2)$, conclude H_0
If $F^* > F(1-\alpha; 1, n-2)$ conclude H_a

If the F for the fitted equation is larger than the critical F-value (from the table) , we reject H0 (not likely true). The regression is significant, in that the true slope is likely not equal to zero

Does this make sense?

- When is MSR/MSE large?



Information for the F-test is often shown as an Analysis of Variance Table:

Source	df	SS	MS	F	p-value
Regression	*1*	*SSreg*	*MSreg*= *SSreg*/1	*F*= *MSreg*/*MSE*	Prob F> $F_{(1,n\text{-}2,1\text{-}\alpha)}$
Residual	n-2	*SSE*	*MSE*= *SSE*/(n-2)		
Total	n-1	*SSy*			

The ANOVA approach for testing the significance of the regression

```
> anova(model)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
temp       1 3805.9  3805.9  575.06 5.732e-14 ***
Residuals 16  105.9     6.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

Hypothesis test for true Slope β_1

Sampling distribution of b_1

= different values of b_1 obtained with repeated sampling for fixed levels of predictor variable x

= Normal distribution (because β_1 is a linear combination of observations y_i)

$$\text{mean} = E\{b_1\} = \beta_1$$

$$\text{variance} = \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2}$$

Estimated Standard Error for Slope β_1

Estimator of $\sigma^2\{b_1\}$:

$$s_{b_1}^2 = \frac{MSE}{\sum (x_i - \bar{x})^2}$$

Estimator of $\sigma\{b_1\}$:

$$s_{b_1} = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$$

Studentized statistic:

$$\frac{b_1 - \beta_1}{s_{b_1}} \text{ is distribute d as } t(n-2)$$

1-α confidence interval for β_1 :

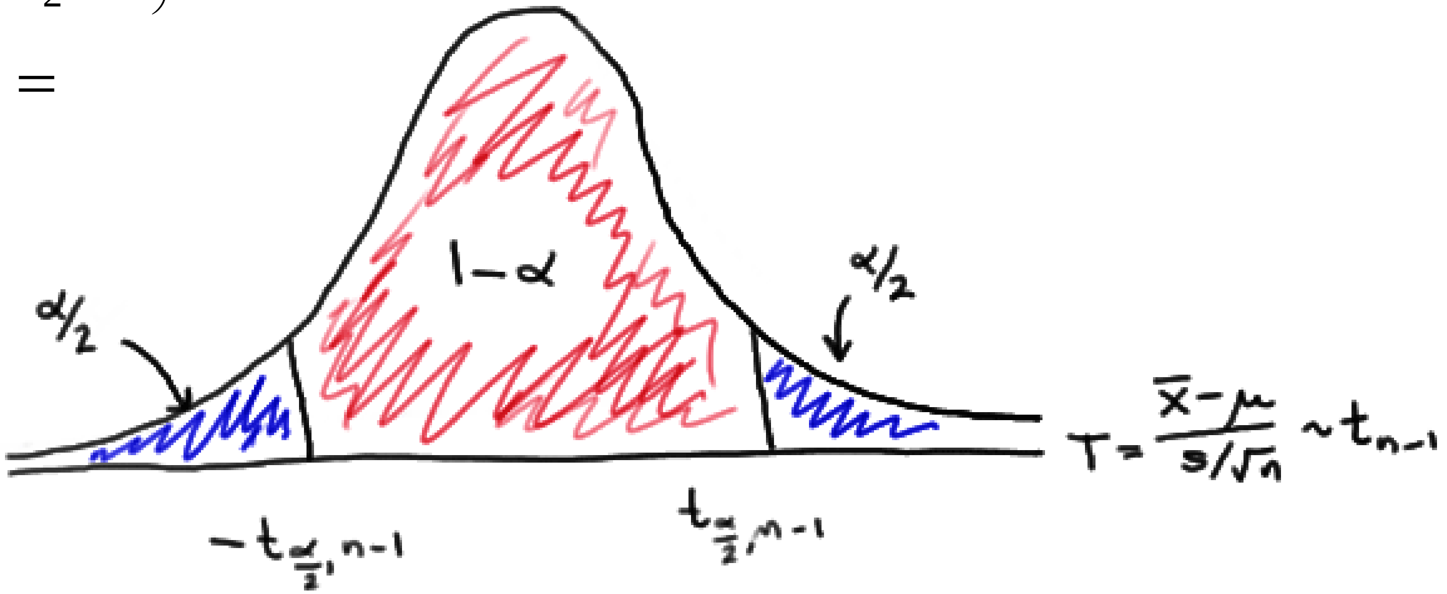
$$P\{t_{\left(\frac{\alpha}{2};n-2\right)} \le \frac{b_1-\beta_1}{s_{b_1}} \le t_{\left(1-\frac{\alpha}{2};n-2\right)}\} = 1-\alpha$$

$$\text{or } P\{-t_{\left(1-\frac{\alpha}{2};n-2\right)} \le \frac{b_1-\beta_1}{s_{b_1}} \le t_{\left(1-\frac{\alpha}{2};n-2\right)}\} = 1-\alpha$$

$$\Leftrightarrow P\{b_1 - t_{\left(1-\frac{\alpha}{2};n-2\right)} s_{b_1} \le \beta_1 \le b_1 + t_{\left(1-\frac{\alpha}{2};n-2\right)} s_{b_1}\} = 1-\alpha$$

$$\Rightarrow 1-\alpha \text{ confidence interval for } \beta_1 =$$

$$b_1 \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} s_{b_1}$$



Hypothesis test for true Slope β_1

1-α confidence interval for β_1 :

$$b_1 \pm \mathrm{t}_{\left(1-\frac{\alpha}{2};n-2\right)} s_{b_1}$$

Hypothesis tests for true slope:

$$\mathrm{H0}: \beta_1 = \mathrm{c}$$

$$\mathrm{Ha}: \beta_1 \neq \mathrm{c}$$

$$\text{If H0 is true, then test statistic}: \frac{b_1 - c}{s_{b_1}} \text{ distributed as } \mathrm{t}_{\left(1-\frac{\alpha}{2};n-2\right)}$$

Reject H0 if $|t| > t_c$

Example: t-test for true Slope β_1

- Test whether or not there is a linear association between x and y, with the risk of a type I error at α = 0.05

$$\text{H0}: \beta_1 = 0$$

$$\text{Ha}: \beta_1 \neq 0$$

$$\text{test statistic}: t^* = \frac{b_1}{s_{b_1}} \text{ distributed as } t_{\left(1-\frac{\alpha}{2};n-2\right)}$$

$$\text{If } |t^*| \leq t(1-\alpha/2; n-2), \text{ conclude H0}$$

$$\text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conclude Ha}$$

This is an alternative for the F-test (ANOVA approach)

Estimated Standard Error for Intercept β_0

Sampling distribution of b_0

= different values of b_0 obtained with repeated sampling for fixed levels of predictor variable x

= Normal distribution (because β_0 is a linear combination of observations y_i)

$$\text{mean} = E\{b_0\} = \beta_0$$

$$\text{variance} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum \left(x_i - \bar{x} \right)^2} \right]$$

Estimated Standard Error for Intercept β_0

Estimator of $\sigma^2\{b_0\}$:

$$s_{b_0}^2 = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]$$

Estimator of $\sigma\{b_0\}$:

$$s_{b_0} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}$$

Studentized statistic:

$$\frac{b_0 - \beta_0}{s_{b_0}} \text{ is distributed as } t(n-2)$$

Estimated Standard Error for Intercept β_0

1-α confidence interval for β_0 :

$$b_0 + t_{\left(1-\frac{\alpha}{2};n-2\right)} s_{b_0}$$

Hypothesis tests for true intercept:

$$\text{H0}: \beta_0 = \text{c}$$

$$\text{Ha}: \beta_0 \neq \text{c}$$

$$\text{If H0 is true, then test statistic} : \frac{b_0 - c}{s_{b_0}} \text{ distributed as } t_{\left(1-\frac{\alpha}{2};n-2\right)}$$

Reject H0 is $|t| > t_c$

calculate 95% confidence intervals for b0 and b1

- For 95% confidence intervals for b0 and b1, would also need estimated standard errors:

$$s_{b_0} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{SSx}\right)} = \sqrt{6.62 \times \left(\frac{1}{18} + \frac{37.5^2}{11812.50}\right)} = 1.075$$

$$s_{b_1} = \sqrt{\frac{MSE}{SSx}} = \sqrt{\frac{6.62}{11812.50}} = 0.0237$$

- The t-value for 16 degrees of freedom and the 0.975 percentile is 2.12

$$b_0 \pm t_{1-\alpha/2,n-2} \times s_{b_0}$$

For β_o: $$5.825 \pm 2.120 \times 1.075$$

$$b_1 \pm t_{1-\alpha/2,n-2} \times s_{b_1}$$

For β_1: $$0.568 \pm 2.120 \times 0.0237$$

	Est. Coeff	St. Error
For b0:	5.825396825	1.074973559
For b1:	0.567619048	0.023670139

CI:	b0	b1
t(0.975,16)	2.12	2.12
lower	3.54645288	0.517438353
upper	8.104340771	0.617799742

Question:
Could the real intercept be equal to 0?

Is the regression significant?

Calculating confidence intervals for the model parameters

```
> confint(model, level=0.95)
                2.5 %    97.5 %
(Intercept) 3.5465547 8.1042390
temp        0.5174406 0.6177975
```

- In R you can get a description of most objects when using the summary() function:

```
> summary(model)

Call:
lm(formula = weight ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3968 -1.6111 -0.0825  2.1389  4.1175 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.82540    1.07497   5.419 5.68e-05 ***
x            0.56762    0.02367  23.980 5.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.573 on 16 degrees of freedom
Multiple R-squared: 0.9729,     Adjusted R-squared: 0.9712 
F-statistic: 575.1 on 1 and 16 DF,  p-value: 5.732e-14
```

$\mathrm{H0}: \beta_0 = 0$

$\mathrm{Ha}: \beta_0 \neq 0$

$\text{t-statistic} : \dfrac{5.82540 - 0}{1.07499}$

if $|t| > t_{critical}$, we reject H0

this is the same as $\mathrm{Pv} < \alpha$

Selecting among alternative models

Process to Fit an Equation using Least Squares

Steps:

1. Sample data are needed, on which the dependent variable and all explanatory (independent) variables are measured.
2. Make any transformations that are needed to meet the most critical assumption: The relationship between *y* and *x is linear.*

   Example: volume = $\beta_0 + \beta_1\ dbh^2$ *may be linear whereas* volume versus dbh is not. Use y_i = *volume* , $x_i = dbh^2$.
3. Fit the equation to minimize the sum of squared error.
4. Check Assumptions. If not met, go back to Step 2.
5. If assumptions are met, then interpret the results.
   - Is the regression significant?
   - What is the r^2? What is the SE_E?
   - Plot the fitted equation over the plot of *y versus x.*

For a number of models, select based on:

1. Meeting assumptions: If an equation does not meet the assumption of a linear relationship, it is not a candidate model
2. Compare the fit statistics. Select higher r^2 (or I^2), and lower SE_E (or SE_E')
3. Reject any models where the regression is not significant, since this model is no better than just using the mean of *y as the predicted value.*
4. Select a model that is biologically tractable. A simpler model is generally preferred, unless there are practical/biological reasons to select the more complex model

Confidence interval for the true mean of y given a particular x value

For the mean of all possible y-values given a particular value of x ($\mu_y|x_h$):

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$

where

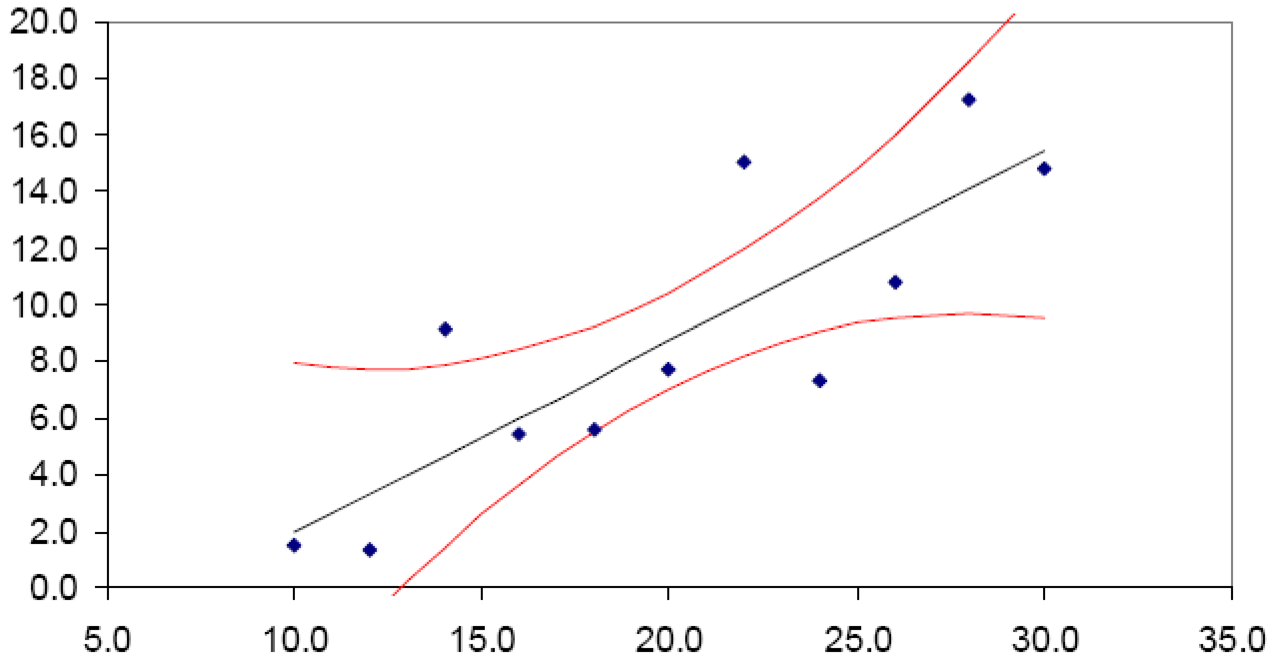
$$\hat{y} \mid x_h = b_0 + b_1 x_h$$

"Given x = …. , what is the estimated **average** y value (predicted value) and a 95% confidence interval for this estimate? "

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

Confidence bands

Plot of the confidence intervals for the mean of y for several x-values will appear as:



Confidence Interval for 1 or more y-values given a particular x value

For one possible new y-value given a particular value of x:

$$\hat{y}_{(new)} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}(new)|x_h}$$

Where

$$\hat{y}_{(new)} \mid x_h = b_0 + b_1 x_h$$

"Given x = …. , what is the estimated corresponding y **for any new observation**, and a 95% confidence interval for this estimate? "

$$s_{\hat{y}(new)|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

Confidence Interval for 1 or more y-values given a particular x value

For the average of *g* new possible y-values given a particular value of x:

$$\hat{y}_{(new)} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}(newg)|x_h}$$

where

$$\hat{y}_{(new)} \mid x_h = b_0 + b_1 x_h$$

$$s_{\hat{y}(new\,g)|x_h} = \sqrt{MSE\left(\frac{1}{g} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

"Given x = …. , what is the estimated **average** out of **g** new y values and a 95% confidence interval for this estimate? "

Given a temperature of 22, what is the estimated **average** weight (predicted value) and a 95% confidence interval for this estimate?

$$\hat{y} \mid x_h = b_0 + b_1 x_h$$

$$\hat{y} \mid (x_h = 22) = 5.825 + 0.568 \times 22 = 18.313$$

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

$$s_{\hat{y}|x_h} = \sqrt{6.62 \times \left(\frac{1}{18} + \frac{(22 - 37.5)^2}{11812.50}\right)} = 0.709$$

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$

$$18.313 - 2.12 \times 0.709 = 16.810$$

$$18.313 + 2.12 \times 0.709 = 19.816$$

- Given a temperature of 22, what is the estimated weight for any new observation, and a 95% confidence interval for this estimate?

$$\hat{y} \mid x_h = b_0 + b_1 x_h$$

$$\hat{y} \mid (x_h = 22) = 5.825 + 0.568 \times 22 = 18.313$$

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

$$s_{\hat{y}|x_h} = \sqrt{6.62 \times \left(1 + \frac{1}{18} + \frac{(22 - 37.5)^2}{11812.50}\right)} = 2.669$$

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$

$$18.313 - 2.12 \times 2.669 = 12.66$$

$$18.313 + 2.12 \times 2.669 = 23.97$$

Using a linear model with new *x*-values

- predict(model, ...) .With no additional options, will return the model training predictions, the same output as fitted(model).

```
> predict(model)
        1         2         3         4         5         6 
7         8 
 5.825397  5.825397  5.825397 14.339683 14.339683 14.339683 
22.853968 22.853968 
        9        10        11        12        13        14 
15        16 
22.853968 31.368254 31.368254 31.368254 39.882540 39.882540 
39.882540 48.396825 
       17        18 
48.396825 48.396825 
```

- But we must first create a prediction data set to use the model on new *x* data. The key point is that the column name in this new data frame *must be the same* one that was used to build the model (i.e. temp).

Using a linear model with new *x*-values

- But we must first create a prediction data set to use the model on new *x* data. The key point is that the column name in this new data frame *must be the same* one that was used to build the model (i.e. temp).

```
> temp.new <- data.frame(temp = c(22, 32, 64))
> temp.new
  temp
1   22
2   32
3   64
```

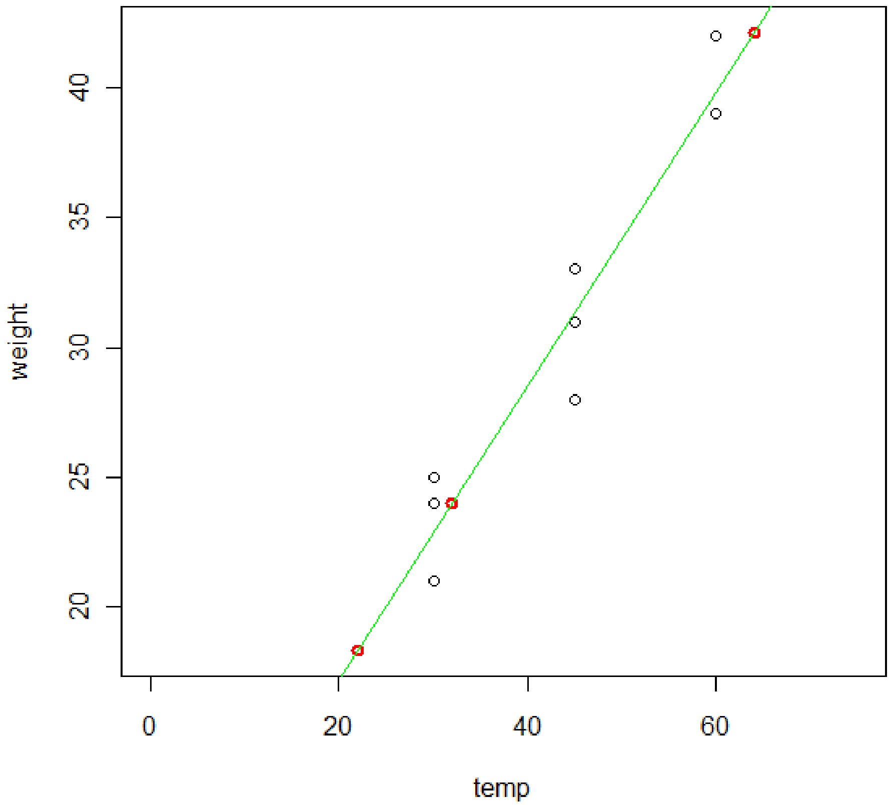
- Now use this new data frame in the predict(model, ...) function as the newdata argument:

```
> y.hat.new <- predict(model, newdata=temp.new)
> y.hat.new
       1        2        3
18.31302 23.98921 42.15302
```

Using a linear model with new *x*-values

- Let's visualize this: these predictions are shown in red, and the least squares line in green.

```
> plot(temp, weight, ylim=range(y.hat.new))
> points(temp.new$temp, y.hat.new, col="red", lwd=2)
> abline(model, col="green")
```



Using a linear model with new *x*-values

- Given a new set of observations for x, what are the estimated **average** weights and the 95% confidence intervals for these estimates:

```
> y.hat.new.PI <- predict(model, newdata=temp.new,
interval="confidence", level=0.95)
> y.hat.new.PI
       fit      lwr      upr
1 18.31302 16.81059 19.81544
2 23.98921 22.67447 25.30394
3 42.15302 40.30355 44.00248
```

- What are the prediction intervals for the new dataset?

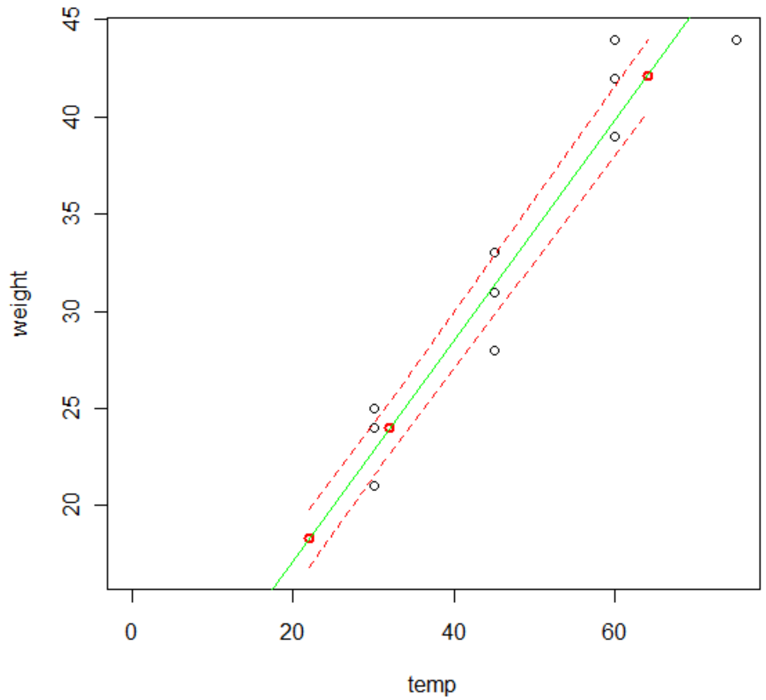
```
> y.hat.new.PI <- predict(model, newdata=temp.new,
interval="prediction", level=0.95)
> y.hat.new.PI
       fit      lwr      upr
1 18.31302 12.65619 23.96984
2 23.98921 18.37931 29.59910
3 42.15302 36.39429 47.91174
```

Setting intervals specifies computation of confidence or prediction (tolerance) intervals, sometimes referred to as ***narrow vs. wide intervals****.*

Using a linear model with new *x*-values

- now add the prediction interval to our visualization. Notice the expected quadratic curvature.

```
> plot(temp, weight, ylim=range(y.hat.new.PI))
> points(temp.new$temp, y.hat.new.PI[,1], col="red", lwd=2)
> abline(model, col="green")
> lines(temp.new$temp, y.hat.new.PI[,2], col="red", lty=2)
> lines(temp.new$temp, y.hat.new.PI[,3], col="red", lty=2)
```



Measure of predictive ability

- Currently, it is common practice to assess the predictive ability of (multivariate) models by comparing predictions with reference values for a test set. From the squared deviations, a root mean squared error of prediction (**RMSEP**) is calculated as

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

where *n* denotes the size of the test set, and $\hat{y}_i$ and y_i are the prediction and reference value for sample *i*, respectively.

- Compare RMSEP to SE_E of the model

RMSEP: weaknesses

- A **constant** measure for prediction uncertainty that cannot lead to prediction intervals with correct coverage probabilities (say 95%).
- A crucial assumption is that the reference values are sufficiently precise; this is certainly not always true - often the prediction is even better than the reference value.
- The intrinsically high variability of the RMSEP estimate requires large test sets, which is wasteful.

Testing a linear model

- Construct an x-vector input and a y-vector response both with 200 observations. Use 150 observation to build the model, then use the remaining 50 to test the model.

```
> input <- rnorm(200, mean=50, sd=12)
> response <- 0.7*input + 50 + rnorm(200, sd=10)
> # Create index vectors that indicate observations for
building and testing:
> build.index = seq(1, 150)
> test.index  = seq(151, 200)
> # Build the model:
> model <- lm(response ~ input, subset=build.index)
```

Testing a linear model

```
> summary(model)

Call:
lm(formula = response ~ input, subset = build.index)

Residuals:
    Min      1Q  Median      3Q     Max
-22.848  -7.161   1.395   7.072  28.037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.4901     3.8402  13.929  < 2e-16 ***
input         0.6163     0.0746   8.261 7.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 10.2 on 148 degrees of freedom
Multiple R-squared: 0.3156,     Adjusted R-squared: 0.311
F-statistic: 68.25 on 1 and 148 DF,  p-value: 7.474e-14
```

Testing a linear model

```
> # Test model. Create data frame from the rest of the
"input" x-variable.
> x.new <- data.frame(input = input[test.index])
> y.hat.new.PI <- predict(model, newdata=x.new,
interval="prediction", level=0.95)
> # Get the actual y-values from the testing data
> y.actual = response[test.index]
> errors <- y.hat.new.PI[,1] - y.actual
> ##using RMSEP
> # Calculate RMSEP, and compare to model's standard error,
and residuals.
> RMSEP <- sqrt(mean(errors^2))
> RMSEP
[1] 9.451627
> summary(residuals(model))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-31.5400  -5.8590   0.3865   0.0000   6.2870  28.9700
```

Testing a linear model

```
> ##do the reference values lie in the 95% CI of
predictions from new values?
> plot(input, response, ylim=range(y.hat.new.PI))
> abline(model, col="green")
> lines(x.new$input, y.hat.new.PI[,2], col="red", lty=2)
> lines(x.new$input, y.hat.new.PI[,3], col="red", lty=2)
> points(x.new$input, y.actual, col="red", lwd=2)
```