

Statistics for Linguists

2023-11-22

Power, effect size and linear models

From last class

- sample variance
- t statistic
- (z-statistic)
- p-value
- interval probabilities

For this class

- Power, Type I and II errors
- Models
- Linear models
- ANOVA

P-value

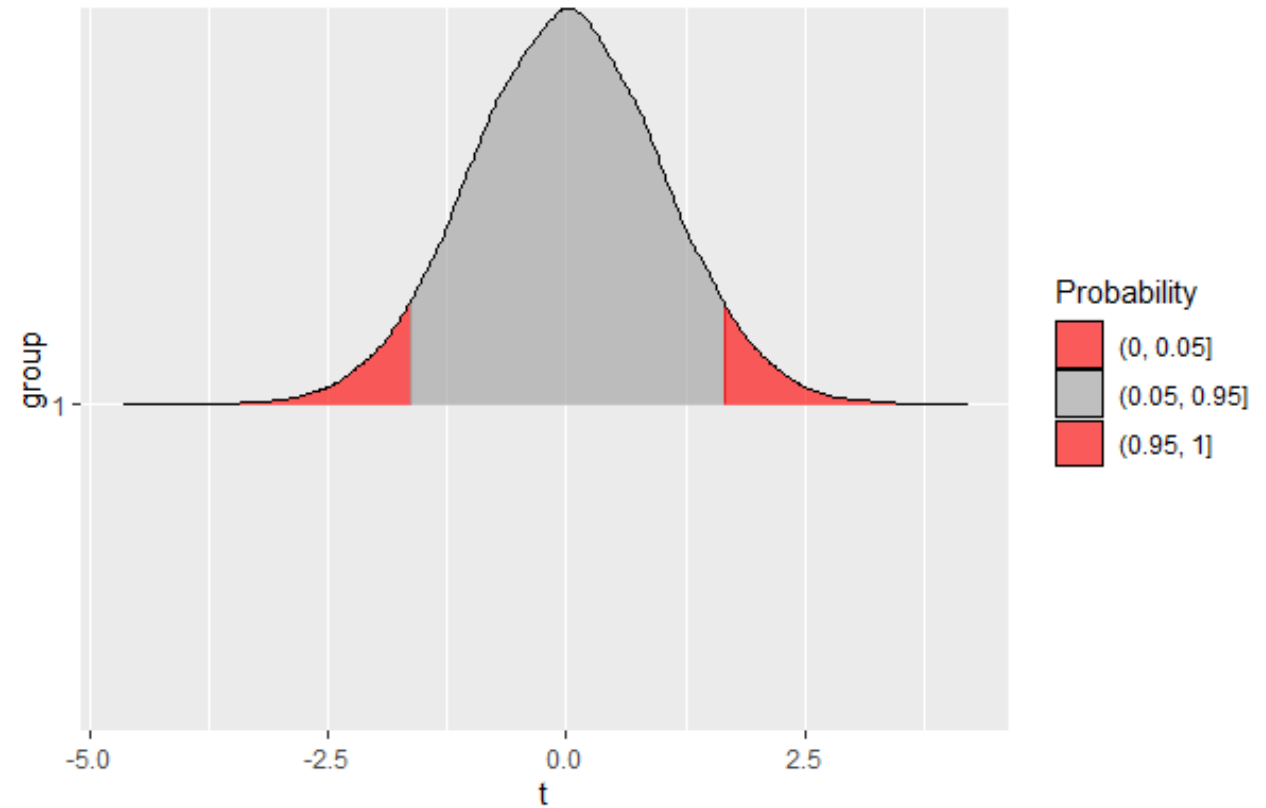
- What is a p-value?
- How is it calculated?
- Why is it calculated this way?

P-value

- What is a p-value?
 - Its an interval probability in relation to a hypothetical distribution that imagines you performed an experiment forever (or gathered data forever).
- How is it calculated?
 - You need a t statistic (z statistic) and a probability distribution (a t distribution if you are doing a t test) -imaginary hypothetical distribution of values we assume if there was no difference between your groups or no relationship between your variables.
- Why is it calculated this way?
 - There's a complicated historical reason for this - but basically you are calculating the probability of your data given a null hypothesis.
 - You are not doing the more intuitive thing of calculating the probability of your hypothesis given your data.

P-value

- Imagine running the test for t over and over again – you would get a distribution of t statistics
- The p-value is a calculation of how likely that t statistic is according to that distribution (called the t distribution)



P values

- A p-value is the probability of receiving a specific statistic (T statistic, F statistic, R-squared) in relation to a hypothetical sampling distribution.
- You can calculate the p-value from the T/F etc. statistic.
- The decision to use a given alpha level (e.g. 0.05) is arbitrary.
- Often, the decision to use one statistic depends on properties of your data (are they continuous, how large is your sample size, are the groups of equal size) – sometimes its just arbitrary though.

P values

- **Type 1 error:** you reject the null hypothesis when its true (false positive).
- **Type 2 error:** you accept the null hypothesis when its false (false negative).

Type II Error & Power

- Failing to reject the null hypothesis when its false
- E.g. you get a p-value below significance level, and you think your study doesn't support the alternative hypothesis
- Power of a test: the probability of not making a Type II Error
- (there's a thing called Power analysis, which asks how larger your sample has to be to detect a particular effect)

Type II Error & Power

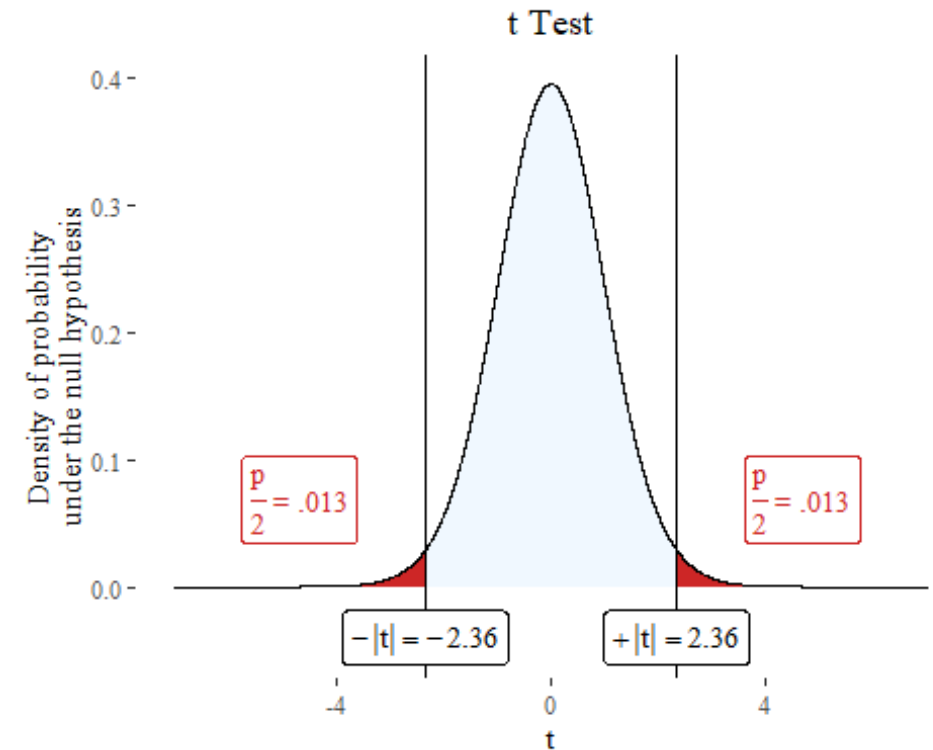
- What affects Power?
- Significance/alpha level (lower level = less power)
- Sample size (larger sample = more power)
- Sample size (larger sample = more power)
- Variability (less variability = more power)
- Magnitude of the effect (larger magnitude = more power)

```

set.seed(1234)
n <- 30
vowel.durations <- rnorm(n, mean = 80, sd=10)
h1 <- c("strong", "weak")
h1 <- sample(h1, n, replace=TRUE, prob=c(0.5, 0.5))
h2 <- c("strong", "weak")
h2 <- sample(h2, n, replace=TRUE, prob=c(0.5, 0.5))
h3 <- c("strong", "weak")
h3 <- sample(h3, n, replace=TRUE, prob=c(0.5, 0.5))
h4 <- c("strong", "weak")
h4 <- sample(h4, n, replace=TRUE, prob=c(0.5, 0.5))
h5 <- c("strong", "weak")
h5 <- sample(h5, n, replace=TRUE, prob=c(0.5, 0.5))
h6 <- c("strong", "weak")
h6 <- sample(h6, n, replace=TRUE, prob=c(0.5, 0.5))
df <- data.frame(vowel.durations, h1,h2,h3,h4,h5,h6)

plotttest(t.test(vowel.durations~h4, data=df))

```



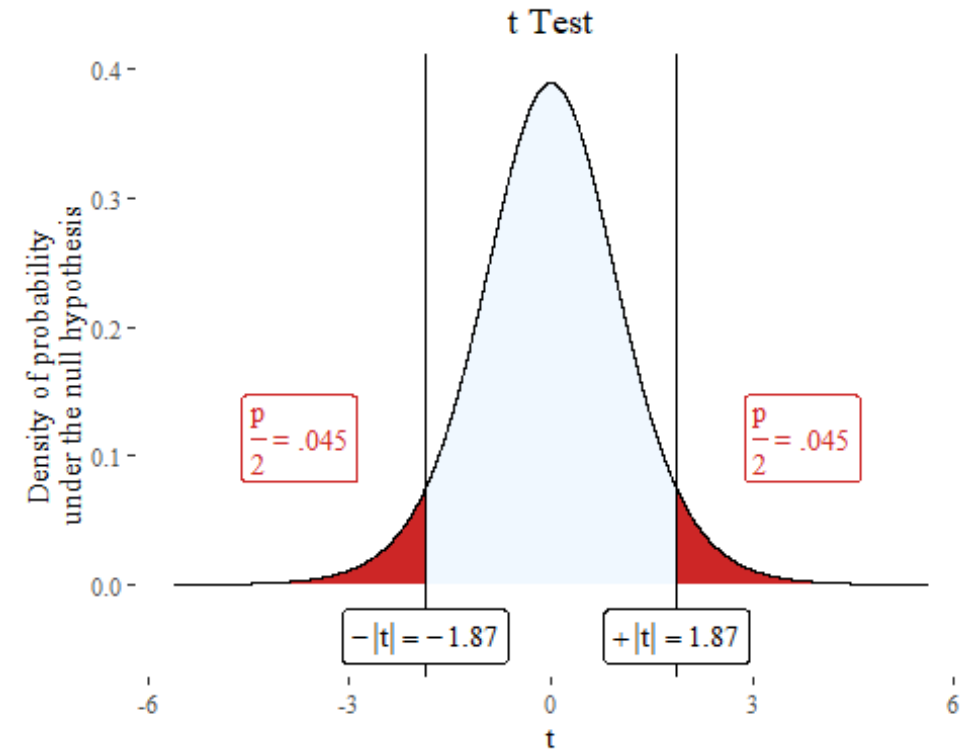
Type I Error & Sensitivity

- Type I Error: Rejection of a true null hypothesis
- Sensitivity: The probability of not making a Type I Error

```

set.seed(123)
n <- 6
strong.vowels <- rnorm(n, mean = 85, sd = 10)
weak.vowels <- rnorm(n, mean = 80, sd = 10)
strong <- rep("yes", times = n, length.out = n)
weak <- rep("no", times = n, length.out = n)
vowel.durations <- list.append(strong.vowels, weak.vowels)
)
prominence <- list.append(strong, weak)
t <- (mean(strong.vowels) - mean(weak.vowels)) / (sqrt((var(strong.vowels)/n) + (var(weak.vowels)/n)))
t
## [1] 1.873338
plotttest(t.test(vowel.durations~prominence), tails = "one")

```



Trying and trying again or **p-hacking**

- If we have enough variables we will eventually find a significant correlation.
- If we sample more from the population we will eventually get a significant value.
- If we do multiple comparisons we should adjust the p value downwards
- In principle (although hard to follow in practice), use of p values should involve a stopping rule stated prior to the study, because if you gather enough data you will always eventually get a significant result.

What to report

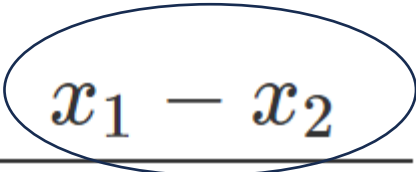
- When you report statistical tests, you wouldn't just report the p-value, rather you report the following:
- p-value
- test statistic
- confidence intervals (more on this later)
- effect size

A note on effect sizes

- We are often more interested in ‘effect sizes’: how large the difference is – note that the p-value conflates the effect size with the sample size
- There are other measures of effect size we’ll get to later.

$$t = \frac{x_1 - x_2}{\sqrt{\frac{V_1}{N_1} + \frac{V_2}{N_2}}}$$

Effect size

A blue oval highlights the numerator of the t-test formula, $x_1 - x_2$. A blue arrow points from the text 'Effect size' to the right side of this oval.

Model

- A “model” is a simplified version of some aspect of the world
- They are cognitive tools for understanding
- Collins & Pinch analogize models with ‘Golems’ from Jewish folklore
- The idea: a useful tool to be called forth for a specific purpose, but that can also be ‘dangerous’ (lead you astray)



<https://upload.wikimedia.org/wikipedia/commons/9/9f/Prague-golem-reproduction.jpg>

Model

Models are stupid

“Models are, by and large, stupid. My point of contention is with the conclusion that stupid models are not useful. Quite the contrary. Stupid models are extremely useful. They are useful because humans are boundedly rational and because language is imprecise. It is often only by formalizing a complex system that we can make progress in understanding it. Formal models should be a necessary component of the behavioral scientist’s toolkit. Models are stupid, and we need more of them.”

Smaldino, Paul E. 2017. Models are stupid and we need more of them. *Computational Social Psychology*. Valalcher, R. (ed.)



<https://upload.wikimedia.org/wikipedia/commons/9/9f/Prague-golem-reproduction.jpg>

Dependent vs. independent variable

- What makes a model 'inferential'
- **Dependent variable:** The variable you are predicting (interesting in explaining)
- **Independent/predictor variable:** The variable that you use to make predictions

Inferential statistics vs. exploratory data analysis

- If you don't have a dependent vs. independent variable distinction you aren't doing inferential stats, you are doing exploratory data analysis.
- (Some would claim: if you don't have a distinction between dependent vs. independent variable before you start modelling, you aren't *really* doing inferential statistics either)
 - Nosek et al. 2019 for discussion

Types of variables

- Types of variables
- Continuous: In principle, indefinitely differentiable
 - coded as 'double' or 'numeric' in R
- Categorical: Discrete categories
 - Coded as 'character' or 'factor' in R
- Binary: Just 0 or 1 (only two options)
- Ranked: Coded in ranked levels

Major 'basic' (most used) types of inferential models

- **Regression:** dependent and predictor variables are both continuous
- **ANOVA:** dependent variable is continuous, predictor variable is categorical
- **Logistic regression:** dependent variable is binary, predictor variable is continuous
- **Chi-squared:** Both variables are categorical

Parametric

- 'Parametric' statistical tests make assumptions about the structure of the data
 - They assume that parameters are normally distributed
 - They assume variance stays the same across all levels / distribution
 - **homoscedastic**
 - They assume errors are normal
 - They assume errors are independent from one another

Parametric vs. nonparametric tests

- Nonparametric tests do not make these assumptions, they just ask about monotonic increase of one variable in relation to another
- Tau statistic
- Pearson's rho
 - I find these more useful if you are just looking at correlations for exploratory purposes (e.g. you have a lot of variables and you want to see how well they are associated one by one)

Some notes

- A lot of statistics classes teach based on flow charts which tell you which tests you should use, but lots of people bash them for teaching students to apply tests without thinking about what makes sense.
 - Clayton 2021; McElreath 2023
- It often makes sense to use more than one modelling technique to see if they converge on the same result – different models might give you different perspectives on the data
 - Page 2018

Clayton, Aubrey. 2021. *Bernoulli's fallacy*. Columbia University Press

McElreath, Richard. 2023. *Statistical Rethinking*. Chapman & Hall.

Page, Scott E. 2018. *The Model Thinker*. Basic Books.

One Variable

Categorical

One-Sample z-test for proportion

To test whether a population proportion is different than some hypothesized value.

Hypothesis Test:

$$H_0: p = p_0$$
$$H_a: p \neq p_0 \text{ or } > <$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$p\text{-value} = \text{normalcdf}(\text{lower}, \text{upper}, 0, 1)$$

Quantitative

One-Sample t-test for mean

To test whether there is a difference between a population mean and some hypothesized value.

Hypothesis Test:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0 \text{ or } > <$$

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$p\text{-value} = \text{tcdf}(\text{lower}, \text{upper}, df)$$

Both Categorical

Two-Sample Proportion

To test whether two population proportions differ.

Hypothesis Test:

$$H_0: p_1 = p_2$$
$$H_a: p_1 \neq p_2 \text{ or } > <$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{Where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$p\text{-value} = \text{normalcdf}(\text{lower}, \text{upper}, 0, 1)$$

Chi-Square

Used to test for a relationship in population between two categorical variables

Hypothesis Test:

H_0 : There is no relationship in population between Var 1 & Var 2
 H_a : There is a relationship in population between Var 1 & Var 2

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$p\text{-value} = \chi^2\text{cdf}(\text{lower}, \text{upper}, df)$$

Two Variables

One of Each

Two-Sample t-test

To test for a difference in two independent population means

Hypothesis Test:

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2 \text{ or } > <$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$p\text{-value} = \text{tcdf}(\text{lower}, \text{upper}, df)$$

ANOVA F-test

To test whether at least one group mean differs from the others

Hypothesis Test:

$H_0: \mu_1 = \mu_2 = \mu_3 \dots \mu_k$
 H_a : at least one μ_i different from others

$$F = \frac{MSB}{MSE}$$

$$p\text{-value} = \text{Fcdf}(\text{lower}, \text{upper}, df_1, df_2)$$

Both Quantitative

Paired t-test

To test whether there is an average difference between two dependent (paired) populations

Hypothesis Test:

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0 \text{ or } > <$$

$$t = \frac{\bar{x}_d}{\frac{s_d}{\sqrt{n}}}$$

$$\text{Where } s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

$$p\text{-value} = \text{tcdf}(\text{lower}, \text{upper}, df)$$

Simple Linear Regression

To test for a linear relationship in population between two quantitative variables

Population Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Overall F-test:

$$F = \frac{MSB}{MSE}$$

$$p\text{-value} = \text{Fcdf}(\text{lower}, \text{upper}, df_1, df_2)$$

Statistic test flow chart

- There's a cool interactive one online

<https://www.statsflowchart.co.uk/>

What is regression?

- T-test is used when the dependent variable is continuous and the independent variable is categorical.
- Chi-squared test can be used when all of your variables are categorical.
- **Regression analysis** can be used when the dependent and independent variables are continuous.
- (Logistic) regression analysis (next week?) is used when the dependent variable is binary or ordinal and the independent variable is continuous.

Regression analyses in linguistics

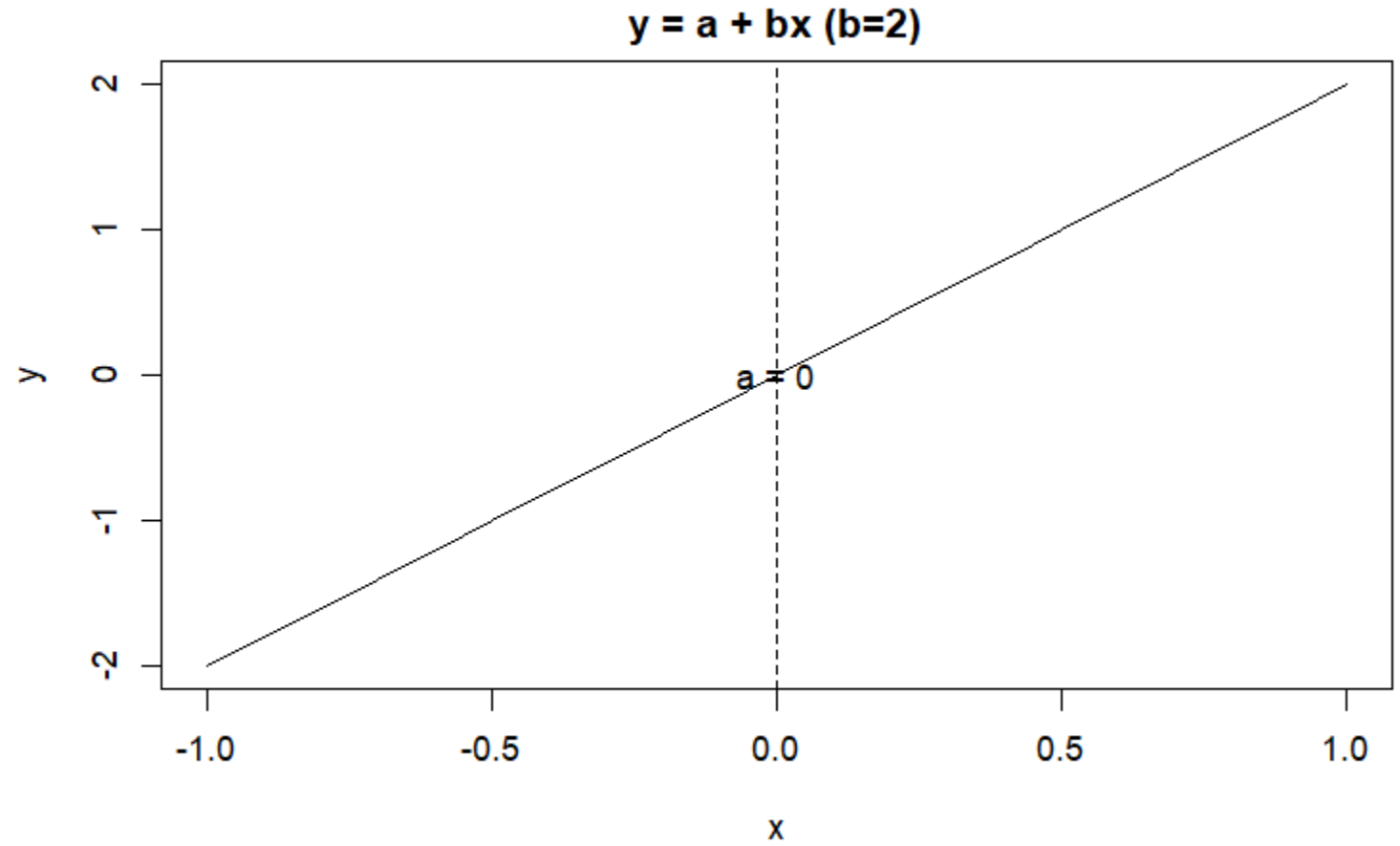
- Morphosyntax: Relationship between phonological complexity and morphological complexity. (Easterday et al. 2021)
- Phonology: The shape and direction of a pitch contour over time.
- Language acquisition: Age against vocabulary acquisition (or any measure of language proficiency)
- Psycholinguistics: Reaction time against frequency of the word.

Regression analysis

- **Model choice:** When you do a regression analysis you have to decide on what the relationship should be.
- What are the different types of relationships two continuous variables can have with one another?

Simple linear model

- a is the intercept
- b is the coefficient
- Y is the dependent variable
- X is the independent variable



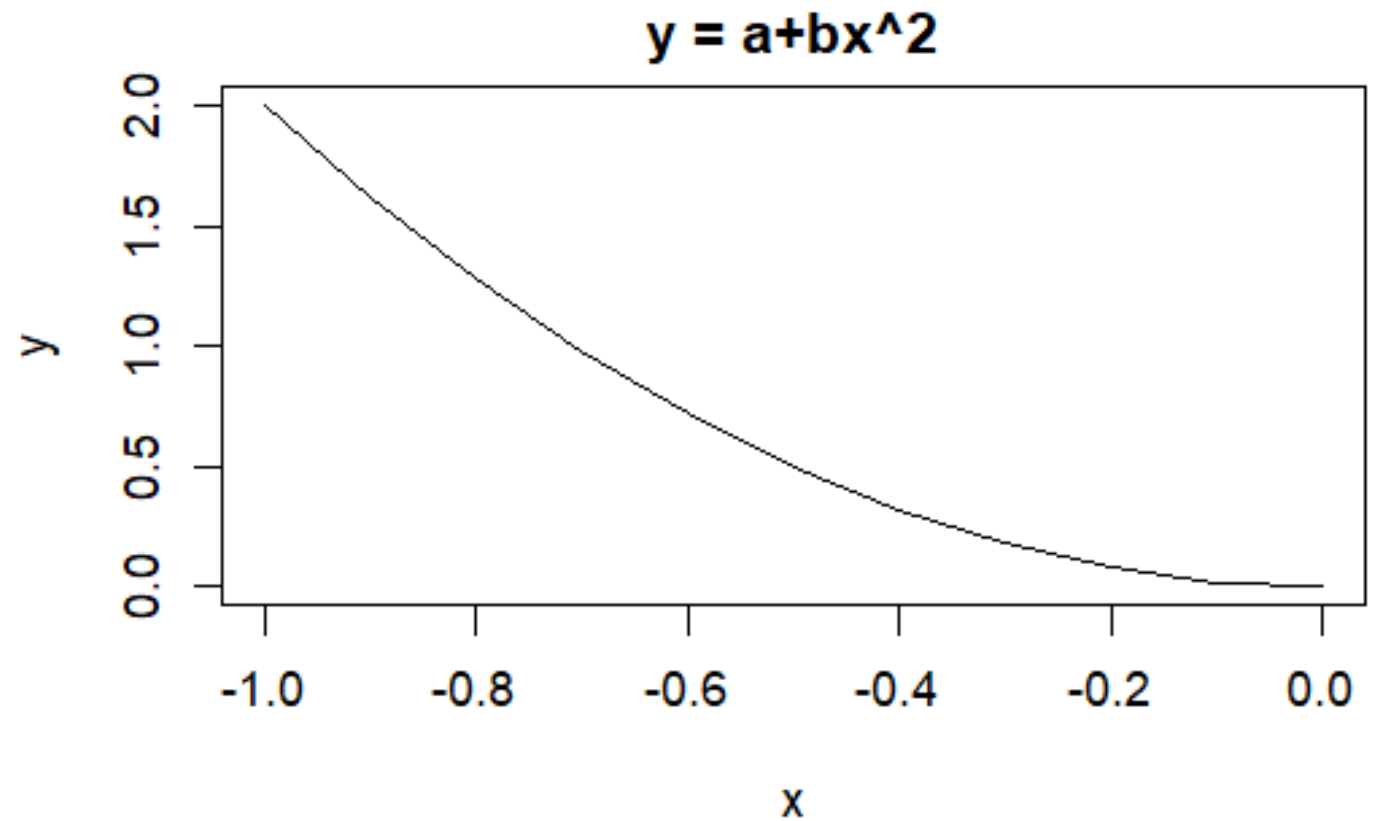
Exponential function

- Will it still be a straight line if it looks like this?

$$y = a + bx^2$$

Exponential curve

$$y = a + bx^2$$

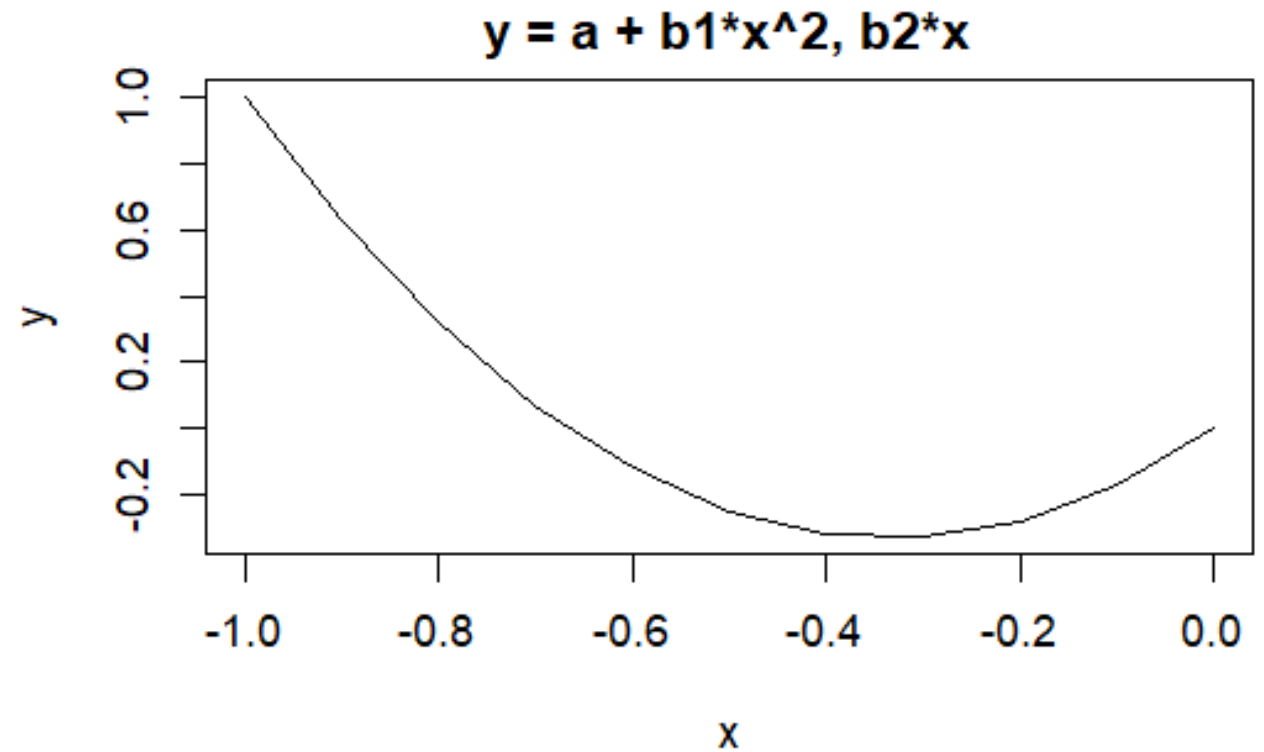


And this?

$$y = a + bx^2$$

Parabolic curve

$$y = a + b_1x^2 + b_2x$$



Model choice

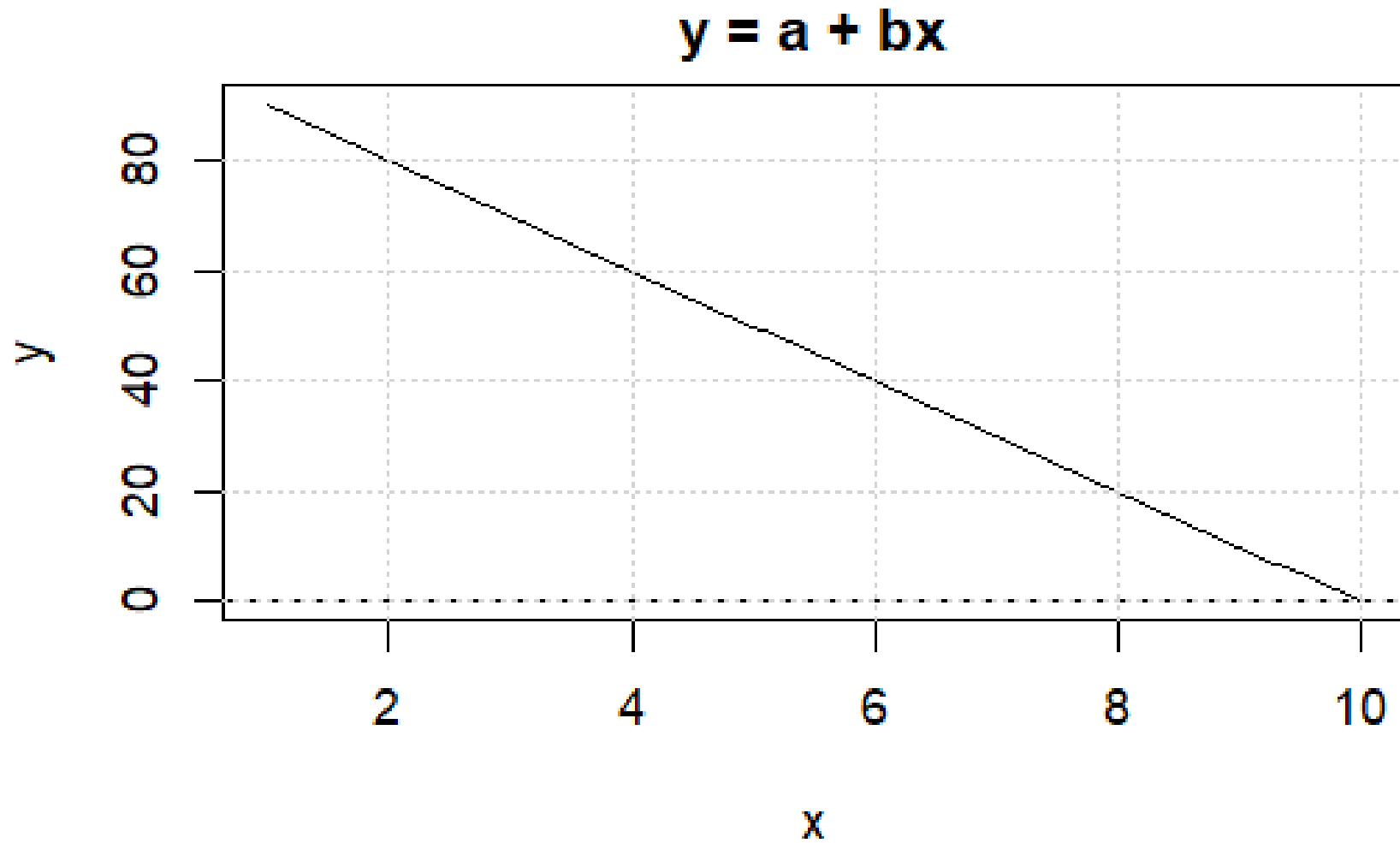
- Part of “model choice” is what type of relationship you think x and y display with one another.
- Note that the interpretation of the coefficient b / β could be very different depending on the type of model you have.

Calculate 'b'

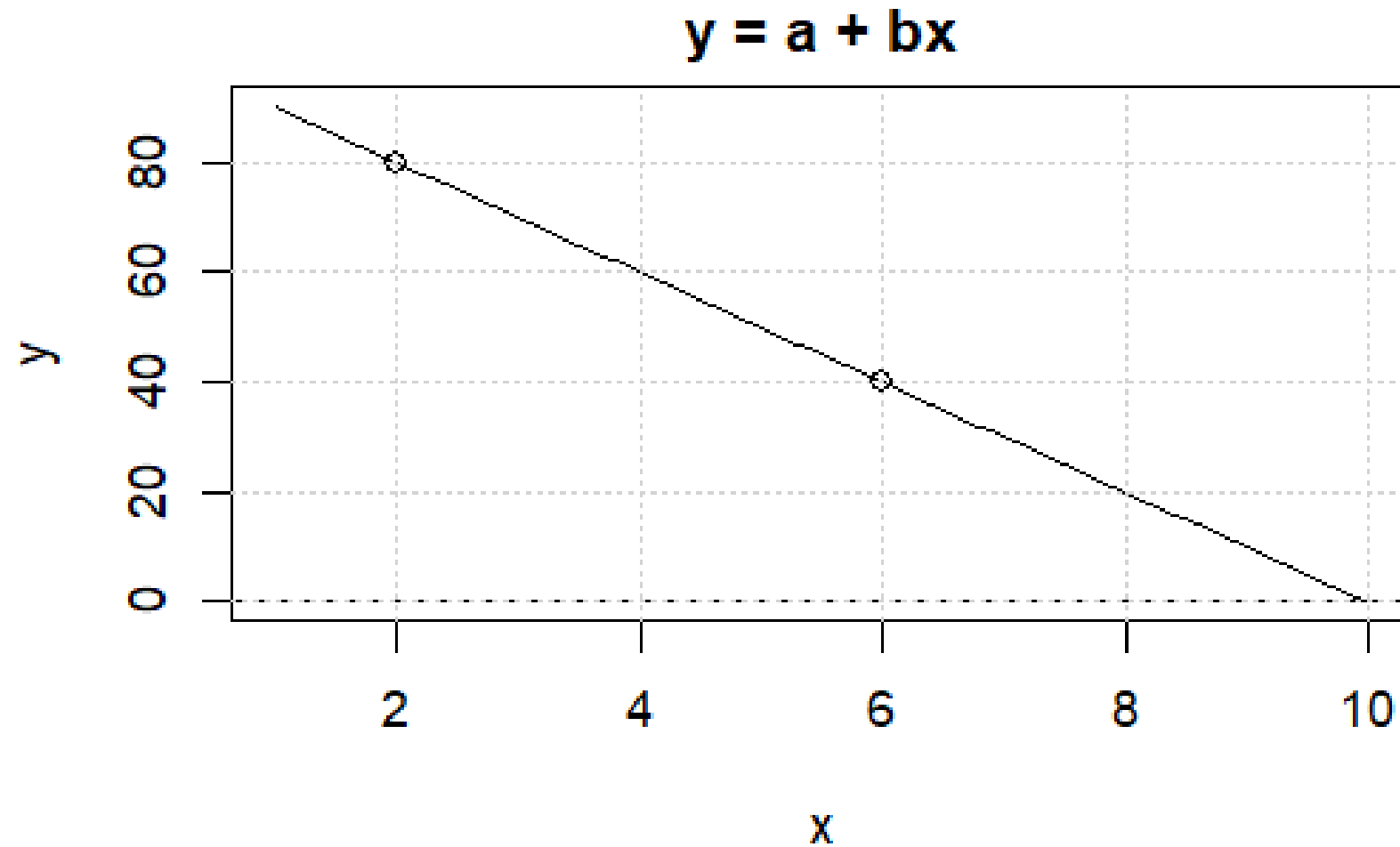
- To understand the values of a linear model, we need to understand first the meaning of b.

$$b = \frac{\text{change in } y}{\text{change in } x}$$

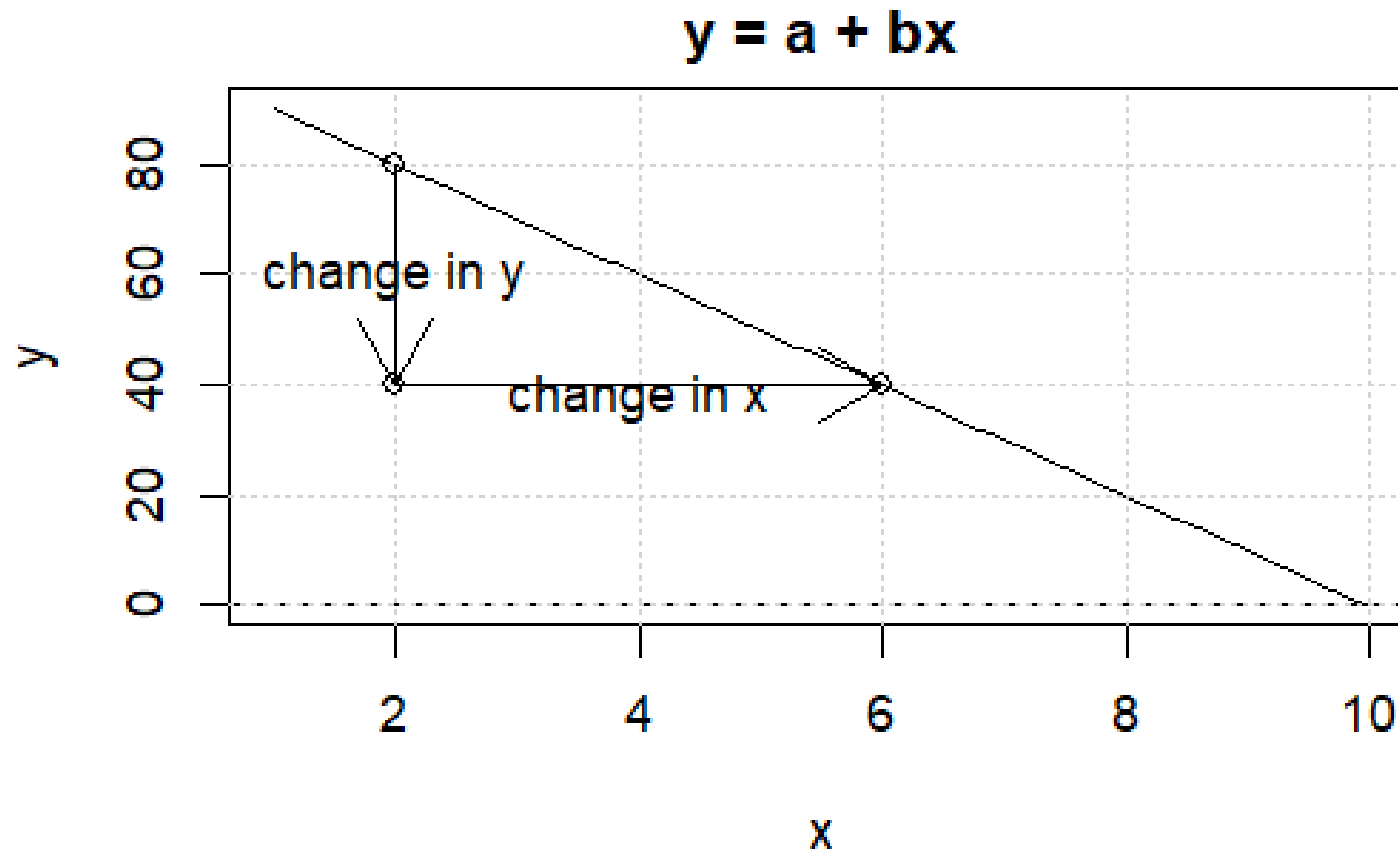
Calculate 'b'



Calculate 'b'



Calculate 'b'



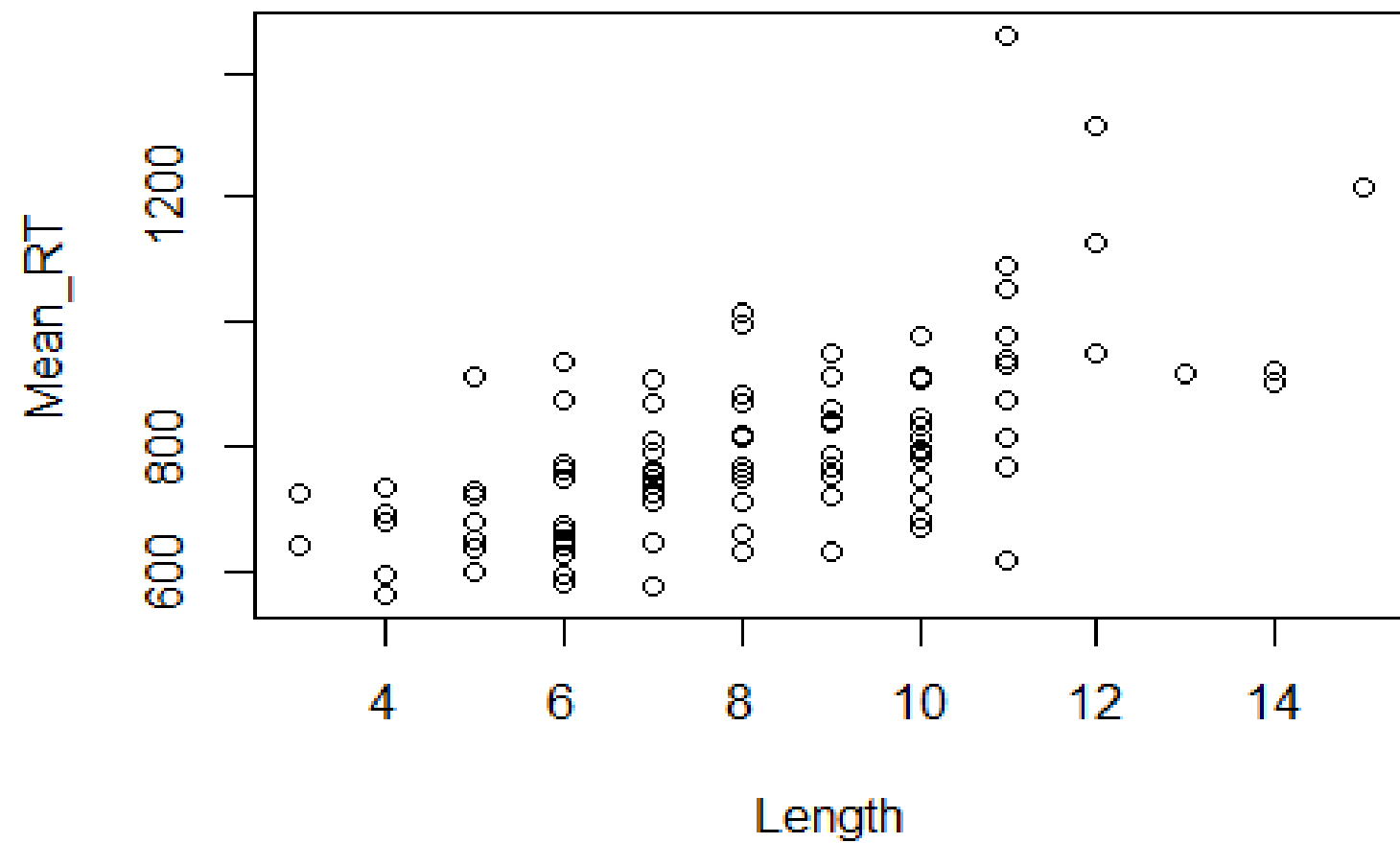
$$b = \frac{-40}{4}$$

Word length and reaction times

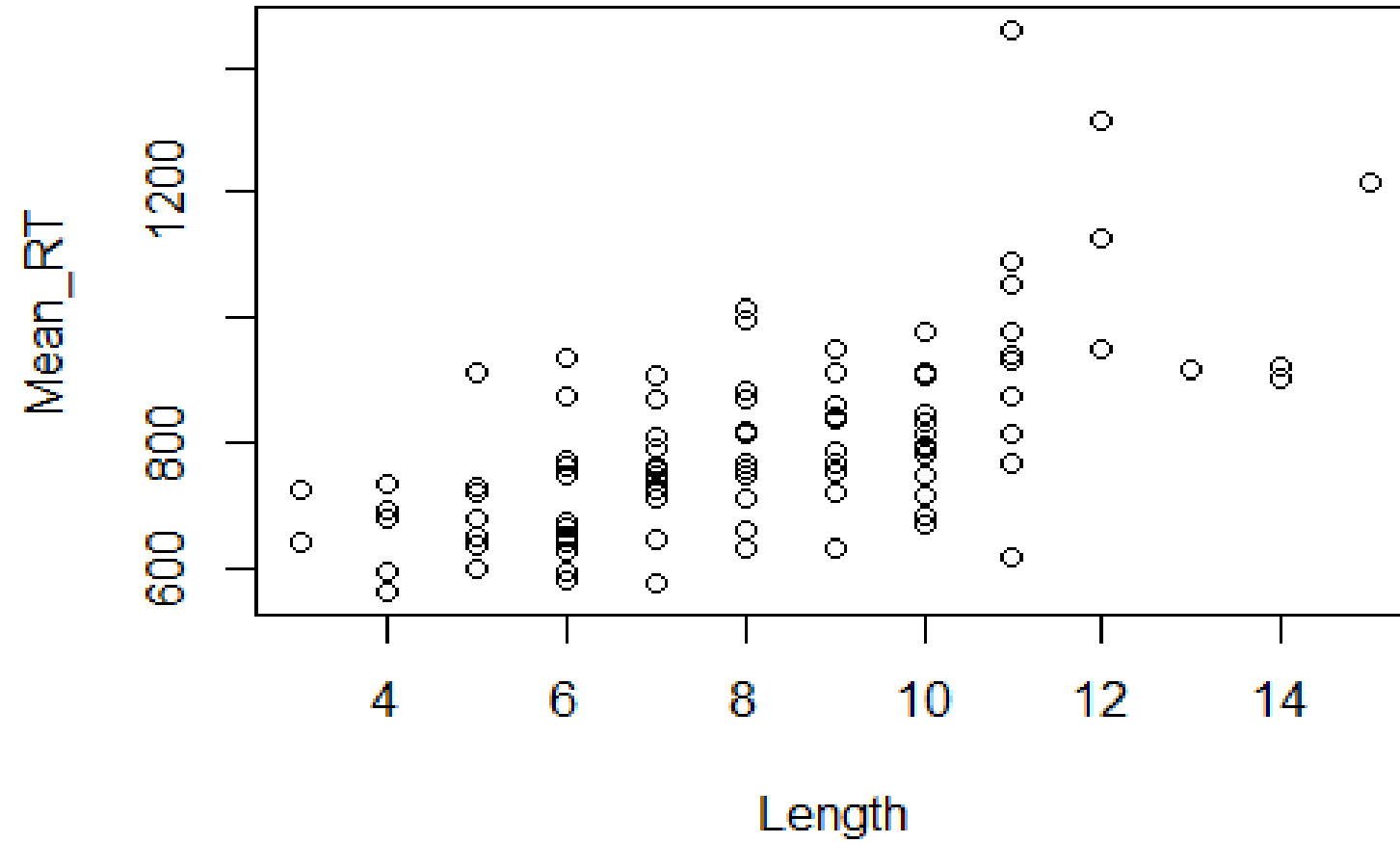
- Download ldt.csv, it is from Rling(), which can be downloaded from Levhsina's github website

Word length and reaction times

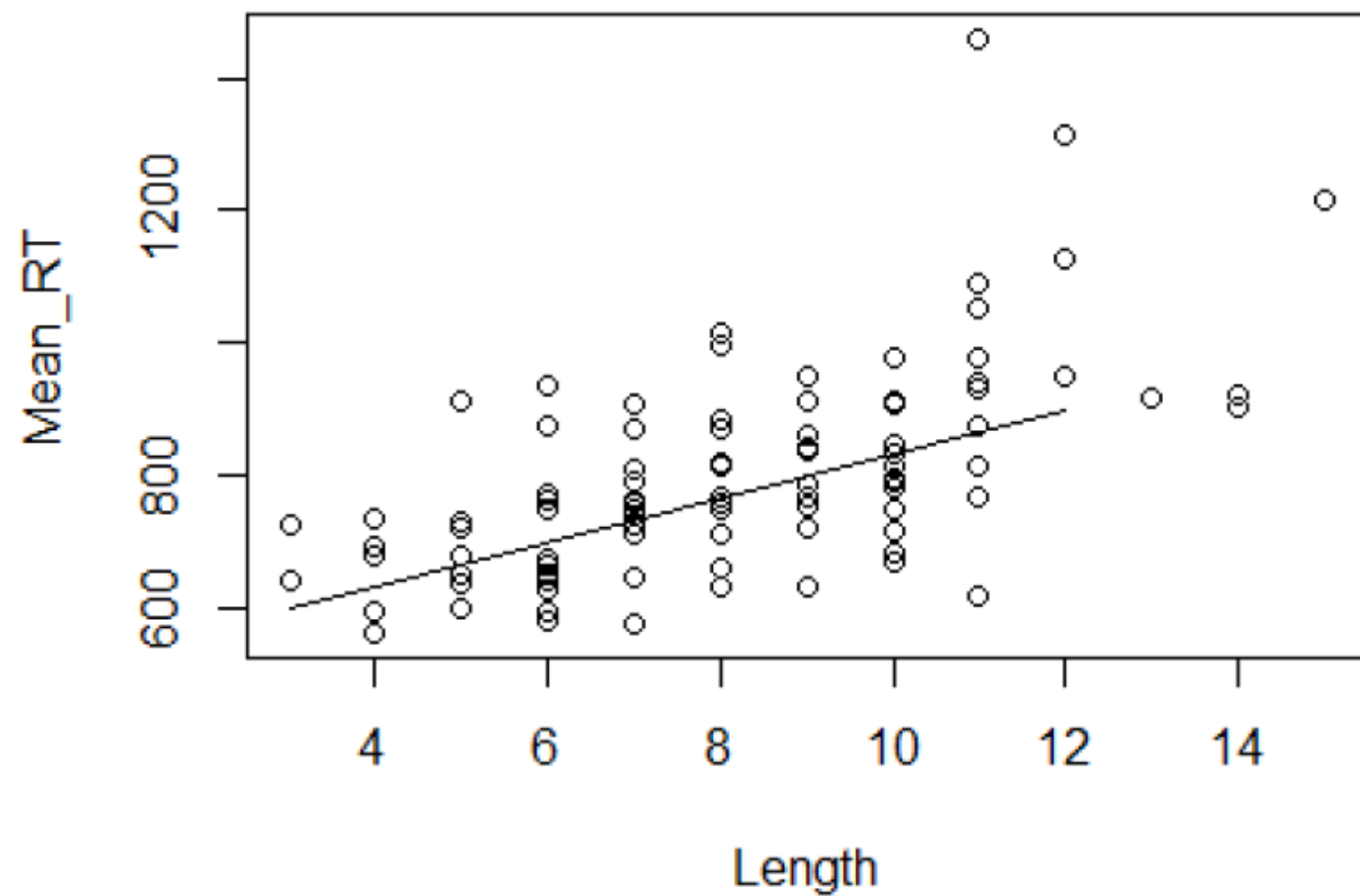
##	Length	Freq	Mean_RT
##	Min. : 3.00	Min. : 0.0	Min. : 564.2
##	1st Qu.: 6.00	1st Qu.: 53.5	1st Qu.: 713.1
##	Median : 8.00	Median : 310.5	Median : 784.9
##	Mean : 8.23	Mean : 3350.3	Mean : 808.3
##	3rd Qu.:10.00	3rd Qu.: 2103.2	3rd Qu.: 905.2
##	Max. :15.00	Max. :75075.0	Max. :1458.8



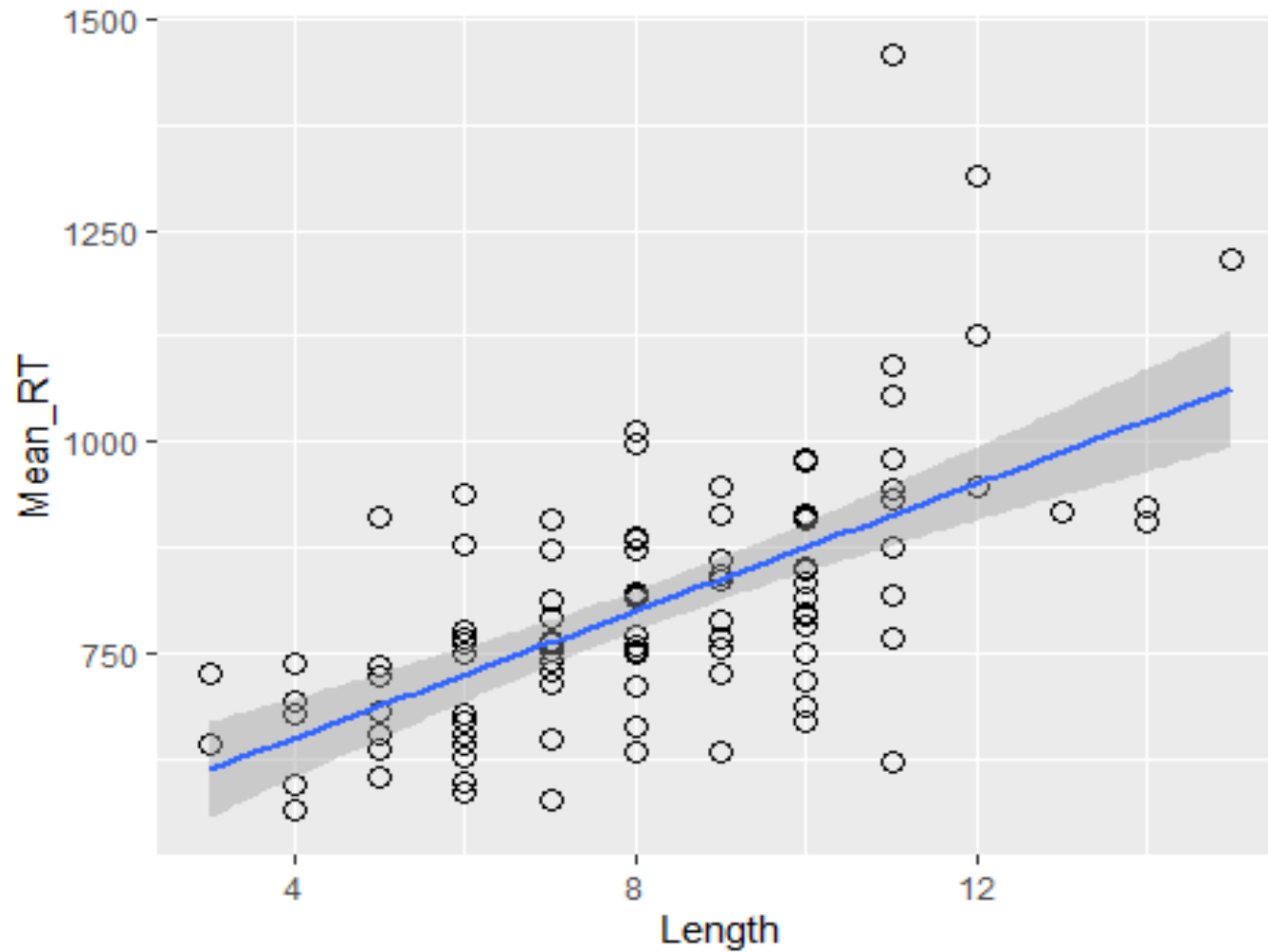
$$b = \frac{dy}{dx} = \frac{900 - 600}{12 - 3} = 33.3$$



$$b = \frac{dy}{dx} = \frac{900 - 600}{12 - 3} = 33.3$$



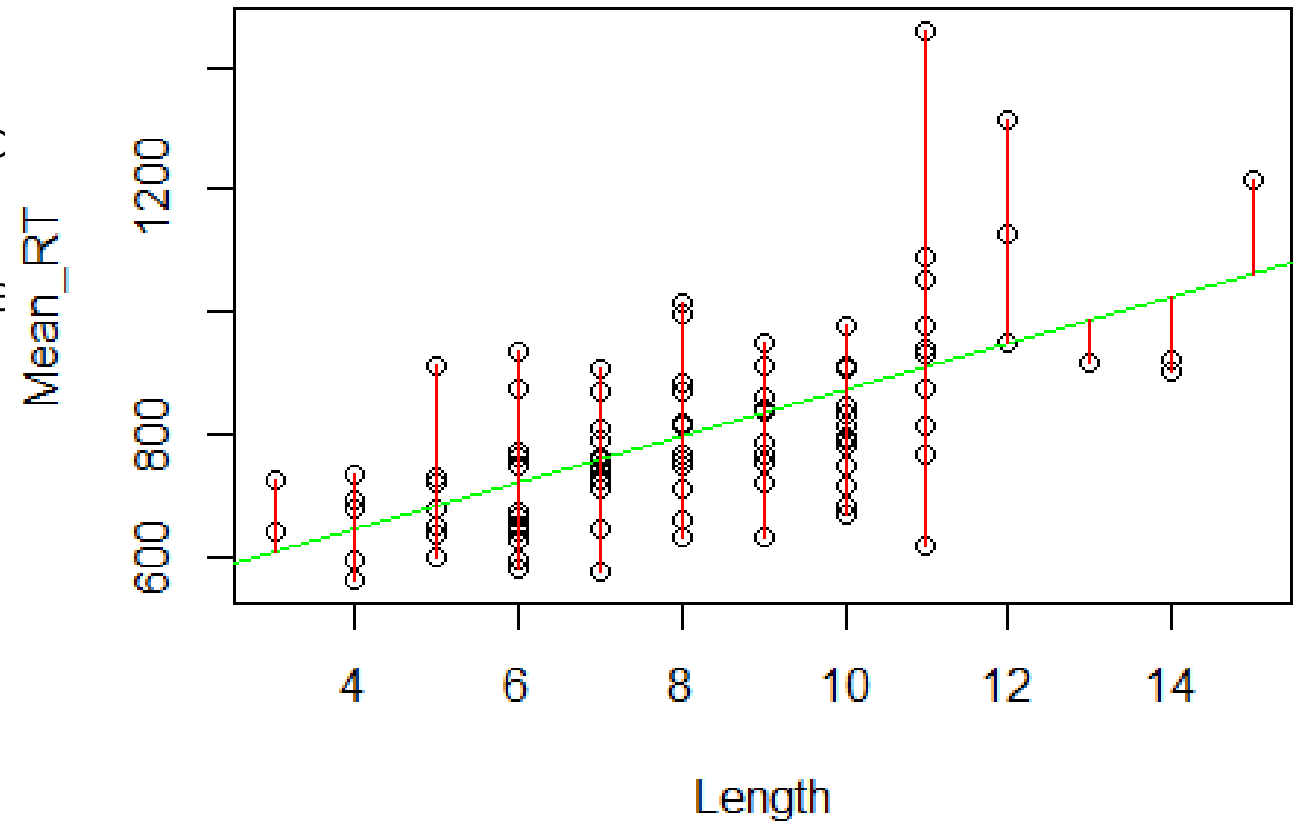
```
lm(Mean_RT~Length)
##
## Call:
## lm(formula = Mean_RT ~ Length)
##
## Coefficients:
## (Intercept)          Length
##      498.44          37.64
```



```
ggplot(ltd, aes(x=Length, y = Mean_RT))+  
geom_point(shape=1, size=3)+stat_smooth(method=lm)
```

Residuals

- A residual is the difference between each data point and the value predicted by the model at the same value of x .
- Some residuals are positive and others are negative.



Residuals

- \hat{y} refers to the predicted value.
- d refers to the residual of a specific data point.

$$d = y - \hat{y}$$

$$d = y - (a + bx)$$

$$d = y - a + bx$$

Residuals

- A regression calculates a line with intercept and slope that make the sum of the residuals equal to 0 (or as close to 0 as possible)

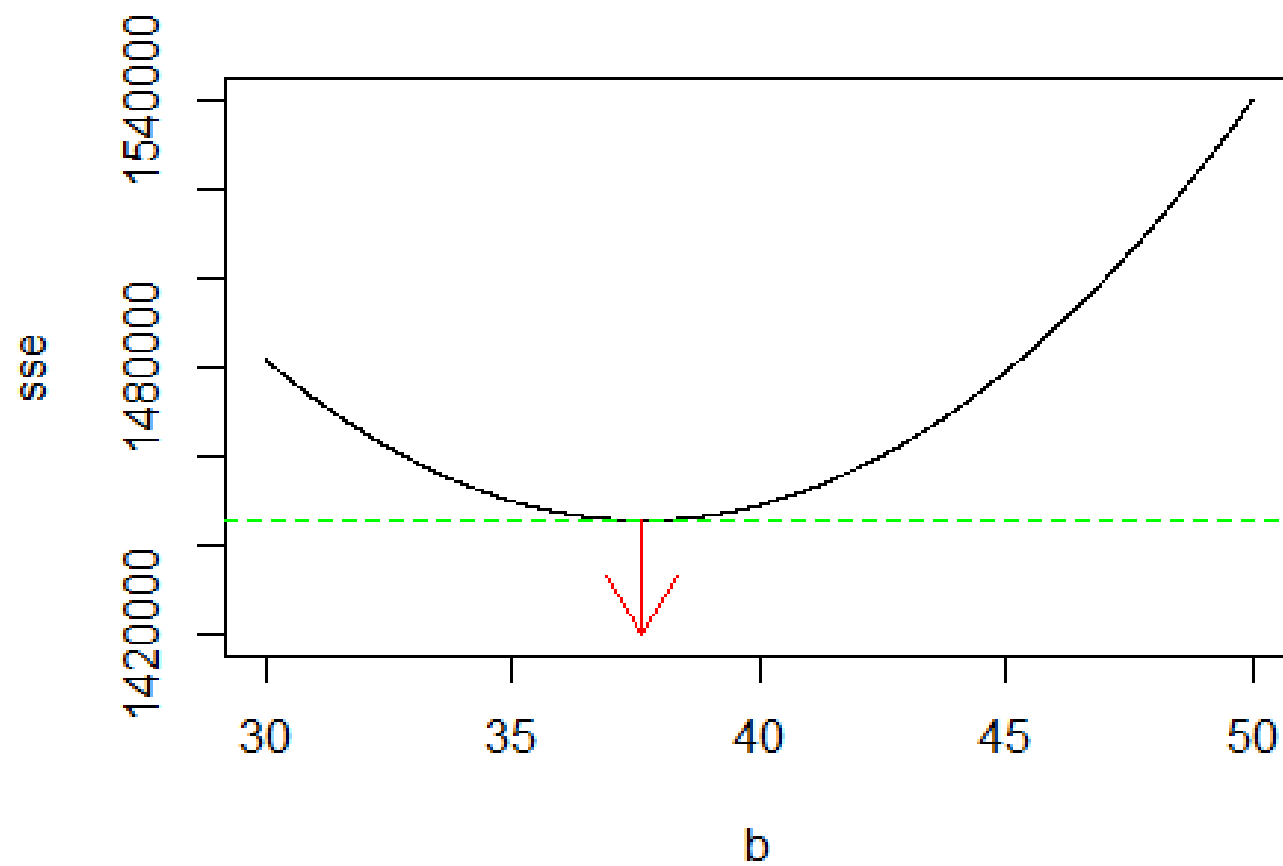
$$\sum d = \sum (y - a - bx) = 0$$

$$SSE = \sum d^2$$

Residuals

- Imagine guessing what the slope.
- After this we change the value of the slope.
- Then we work out the new intercept $a = y - bx - d$
- Then we predict the fitted values of y (reaction time) for the new b .
- After this we work out the residuals
- Then we calculate the SSE (Sum of squares error)
- Then we repeat for values of b until we arrive at the smallest number.

Residuals



Residuals

- We get the same result with `lm()`

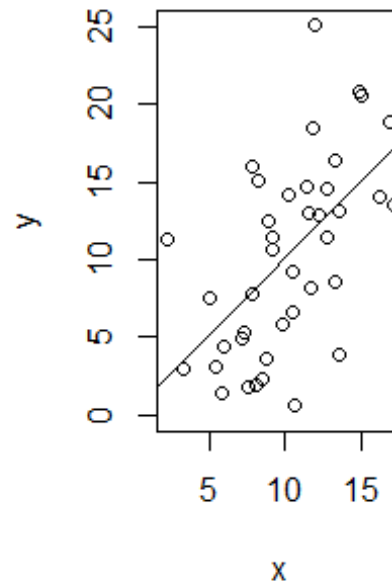
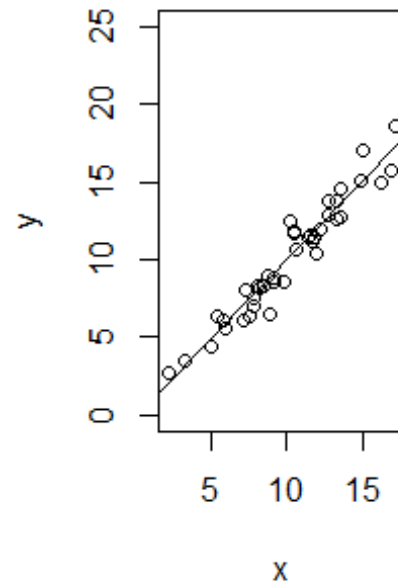
```
lm(Mean_RT~Length)
##
## Call:
## lm(formula = Mean_RT ~ Length)
##
## Coefficients:
## (Intercept)      Length
##      498.44       37.64
```

We stopped here 2023-11-22

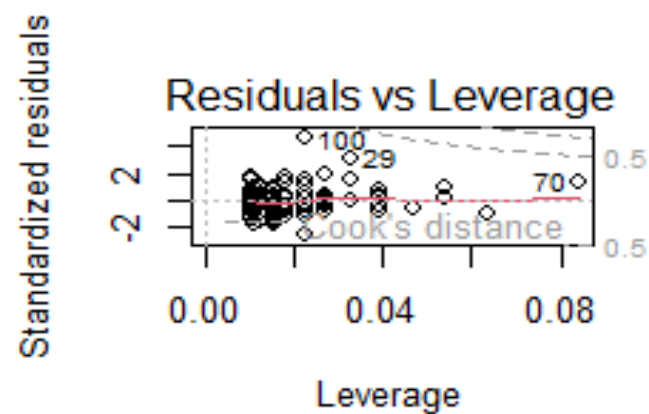
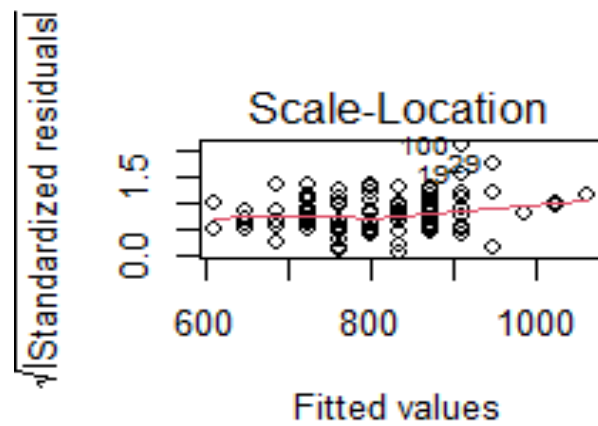
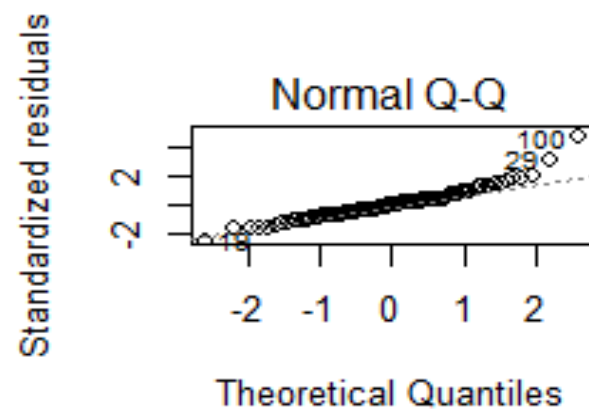
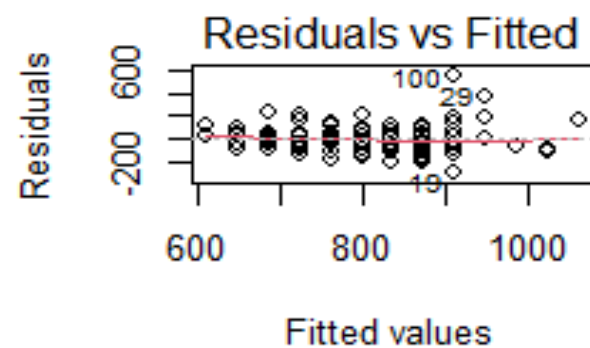
The following slides will be included in next weeks lecture

Model fitting

- Two regression lines can have the same slope and intercept, but be different with respect to their residuals.



Model checking



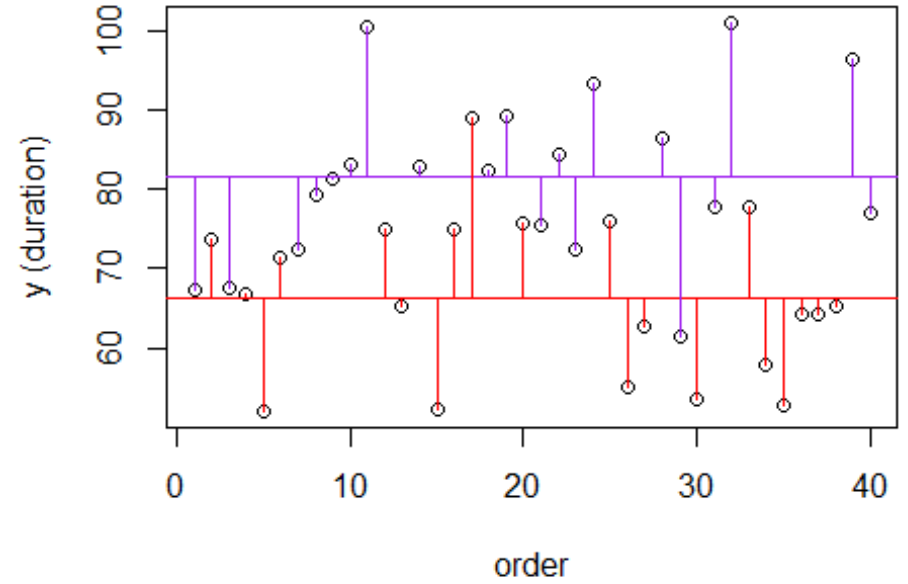
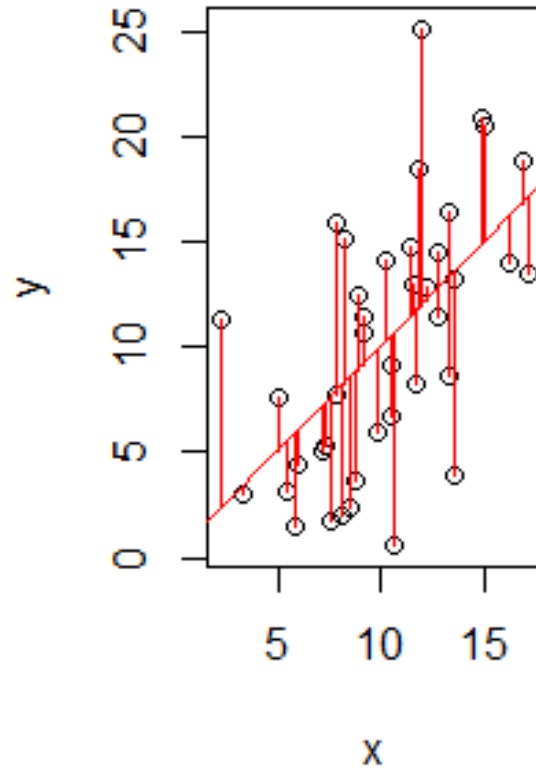
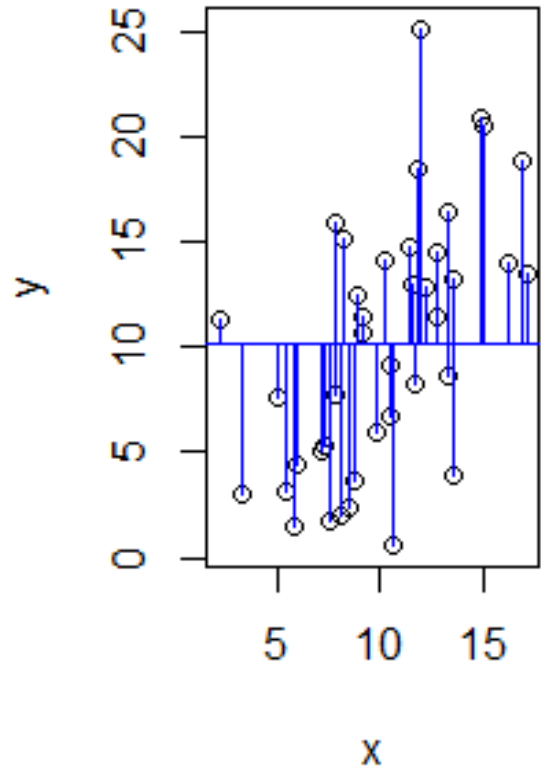
ANOVA

- What is typically referred to as ANOVA (Analysis of variance) often refers to a type of statistical analysis when all the predictor variables categorical. There is some controversy with respect to whether ANOVA can just be understood as a special case of regression analysis. Compare
- Cottingham et al. 2005 ``Regression versus ANOVA'' in Frontiers in Ecology and the Environment
- Gelman 2006 ``Analysis of variance - why it is more important than ever'' in The Annals of Statistics

ANOVA and Regression

- ANOVA has the same mathematical structure as regression, but some regard it as conceptually different,
- ANOVA can also mean the analysis of data into batches and groups (hierarchical modeling), rather than narrowly as the classical ANOVA test.
- For historical reasons the ANOVA is more often used in experimental settings and regression analysis for observational data.
- Sometimes classical ANOVA is used, when regression really should be: the research bins the data of the predictor variable into groups to make it categorical (see Stoll & Gries 2009) in Journal of Language Acquisition.

ANOVA and regression



ANOVA

- ANOVA belongs to family of statistical techniques known as general linear models.
- You are comparing the variances between two groups in relation to the variance of the groups combined.

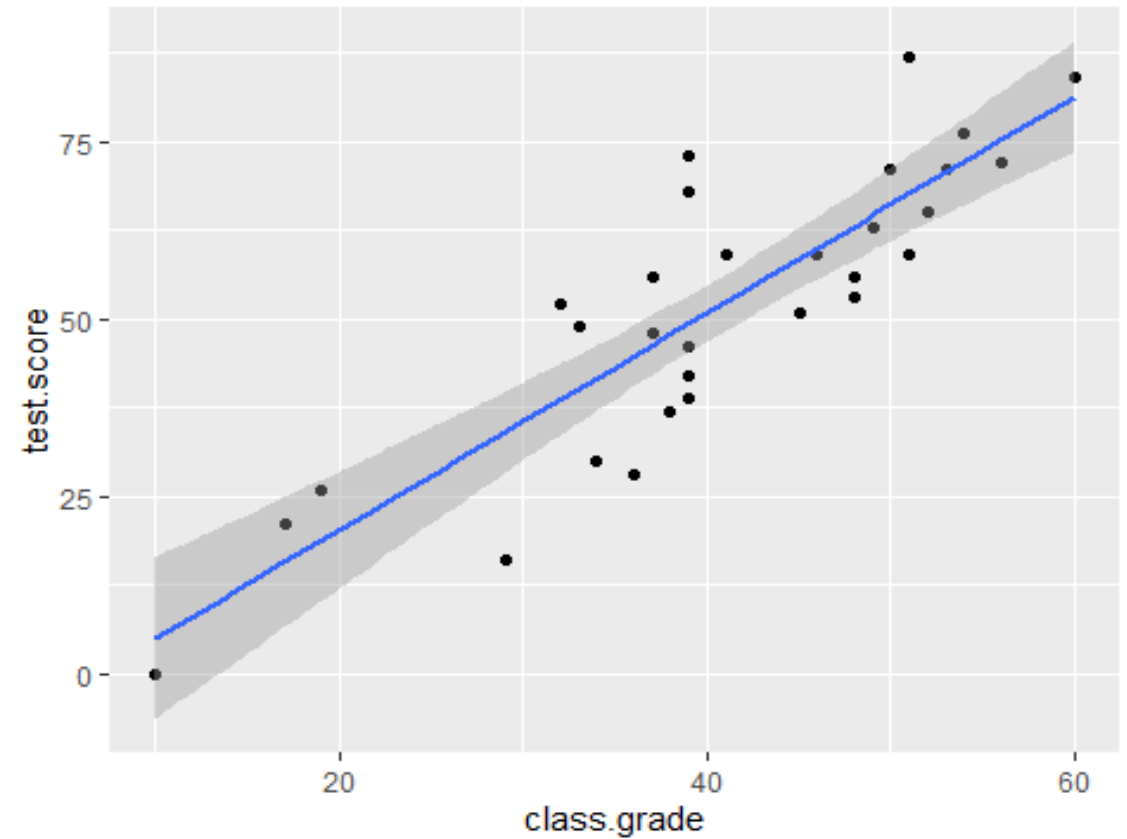
ANOVA

Simulate some values according to a linear relationship / model.

```
set.seed(1)
school <- sample(c("good", "bad"), size=30, replace=30)
school.binary <- ifelse(school == "good", 1, 0)
class.grade <- as.integer(rtruncnorm(a=0, b=100, m = 40, sd=15, n =30))
intercept <- 0
b1 <- 15
b2 <- 1
error <- rnorm(m=0, sd=10, n=30)
test.score <- intercept + b1*school.binary +b2*class.grade + error
test.score <- ifelse(test.score < 0, 0, test.score)
test.score <- ifelse(test.score> 100, 100, test.score)
test.score <- as.integer(test.score)
data <- data.frame(test.score, school, class.grade)
```

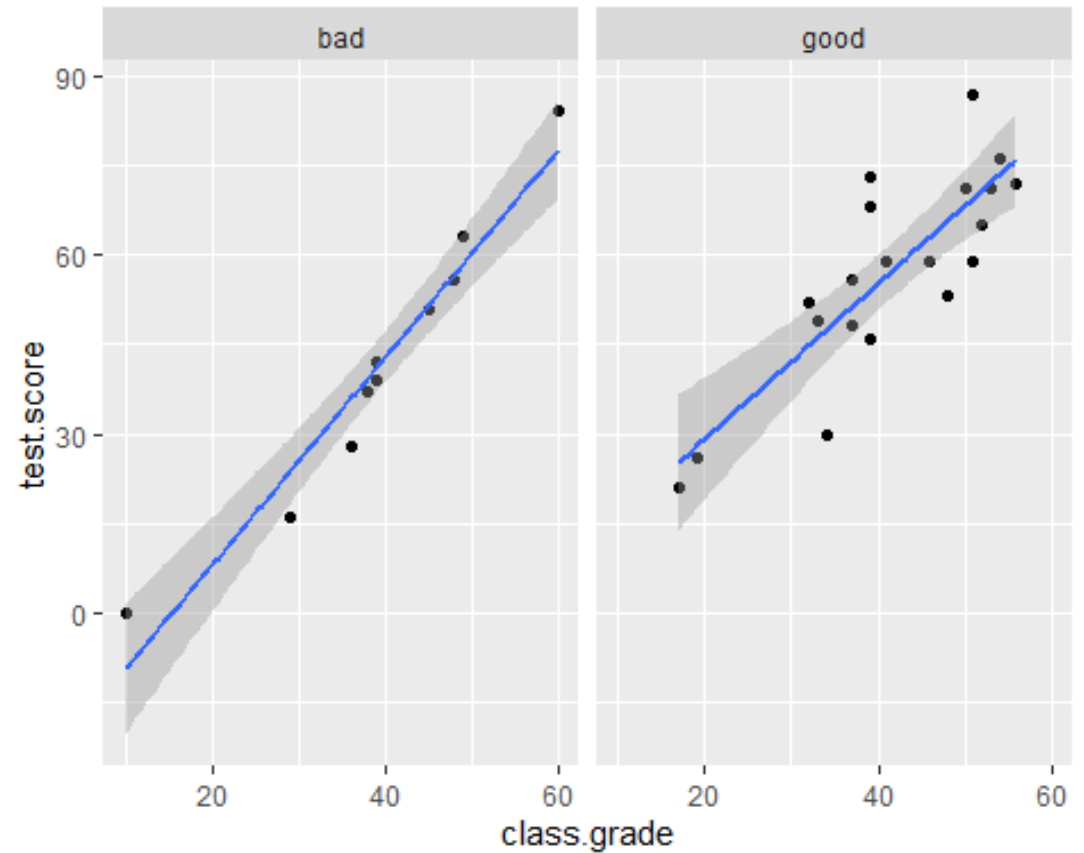
ANOVA

```
ggplot(data, aes(class.grade, test.score)) +  
  geom_point()+  
  geom_smooth(method='lm')
```

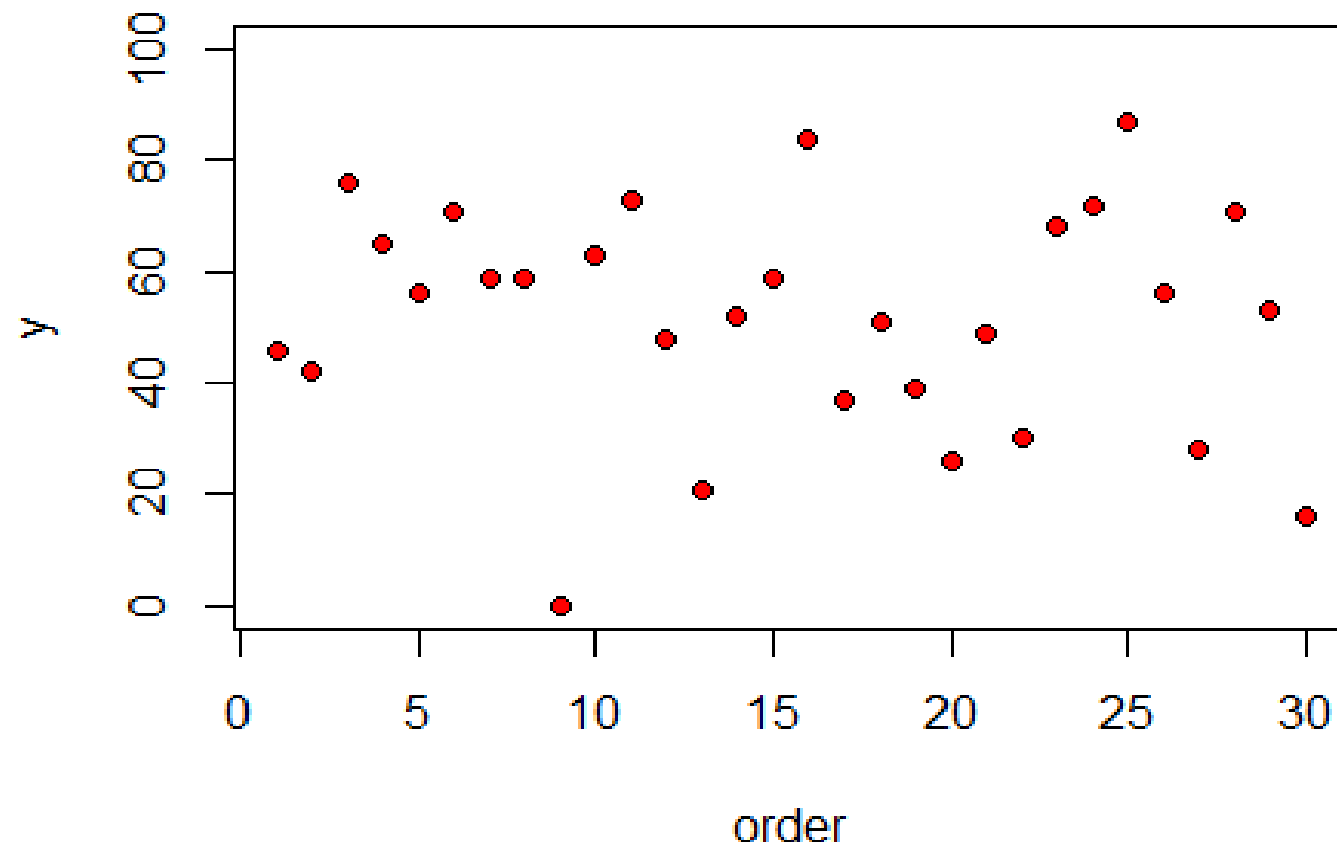


ANOVA

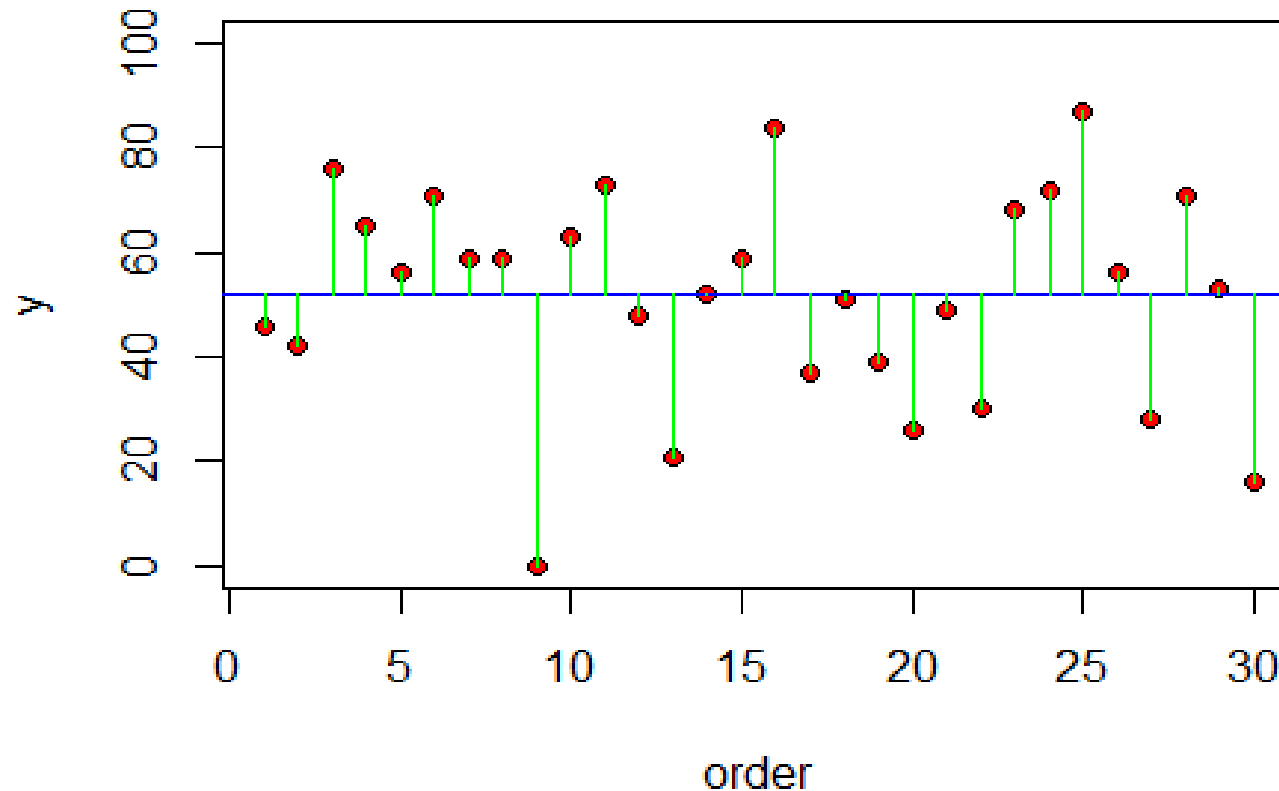
```
ggplot(data,aes(class.grade, test.score)) +  
  geom_point()+  
  geom_smooth(method='lm')+  
  facet_wrap(~school)
```



```
plot(1:30,test.score,ylim=c(0,100),ylab="y",xlab="order",pch=21,bg="red")
```




```
plot(1:30,test.score,ylim=c(0,100),ylab="y",xlab="order",pch=21,bg="red")  
abline(h=mean(test.score),col="blue")  
for(i in 1:100) lines(c(i,i),c(mean(test.score),test.score[i]),col="green")
```

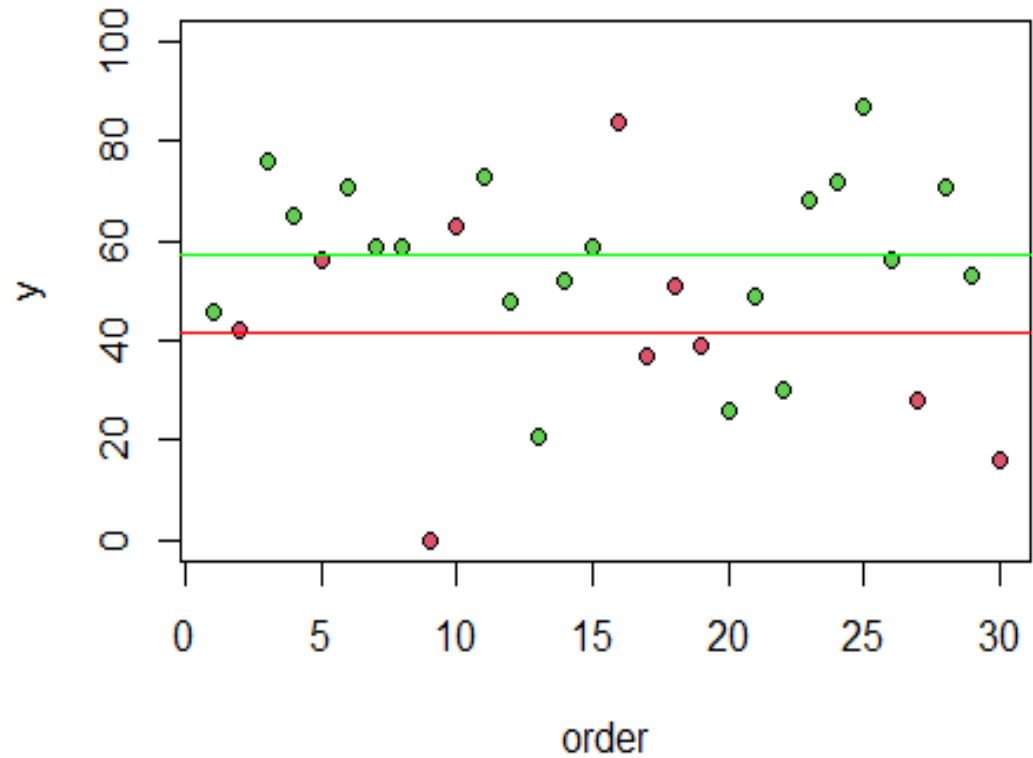


$$SSY = \sum (y - \bar{y})^2$$

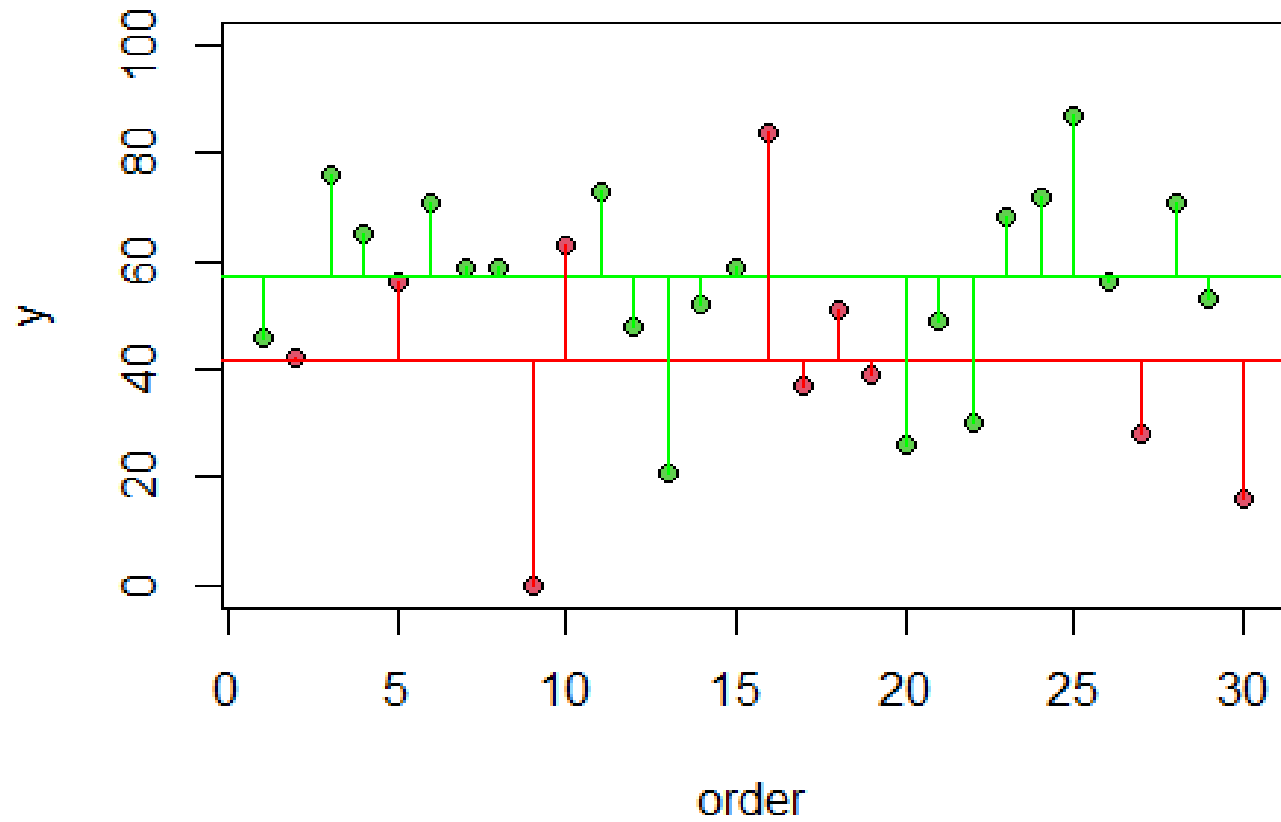
Total sum squares = SSY

**The sum of all the lengths of
all the green lines.**

```
plot(1:30,test.score,ylim=c(0,100),ylab="y",xlab="order",  
pch=21,bg= as.numeric(as.factor(school))+1 )  
abline(h=mean(test.score[school=="good"]),col="green")  
abline(h=mean(test.score[school=="bad"]), col="red")
```



$$SSE = \sum (y - \bar{y}_1)^2 + \sum (y - \bar{y}_2)^2$$



$$SSY = SSG + SSE$$

```
SSEgood <- sum((test.score[school=="good"]-mean(test.score[school=="good"]))^2)
SSEbad <- sum((test.score[school=="bad"]-mean(test.score[school=="bad"]))^2)
SSE = SSEgood + SSEbad
```

```
SSG = SSY - SSE
```