

Intro and regressions



Probability and Statistics

COMS10011
Dr. Anne Roudaut
csxar@bristol.ac.uk

https://github.com/coms10011/2019_20

who am i?



most resent work



CHI 2018

Engage with CHI

Montréal, Canada

April 21-26

chi2018.acm.org

Human Computer Interaction (HCI)::

a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers

**experimental
psychology** **design**



comp. science

experimental psychology::

the branch of psychology concerned with the scientific investigation of the responses of individuals to stimuli in controlled situations



e.g. bandwagon effect (one of our many cognitive biases)





promised a 2nd marshmallow if resist to eat
the 1st one until lady comes back (20mn)

what is the link with statistics?

like in many fields, statistics is the main tool to **analyse, demonstrate, evaluate or predict**

let's start with
an example

imagine you are designing a graphical interface for a new application on a laptop

how big should the buttons/icons be?

Fitts' law ::

the time required to **acquire a target** of size w at distance d can be described as $T = a + b \log (1 + d/w)$

smaller bin = harder and further = harder



Fitts' law ::

the time required to **acquire a target** of size w at distance d can be described as $T = a + b \log (1 + d/w)$

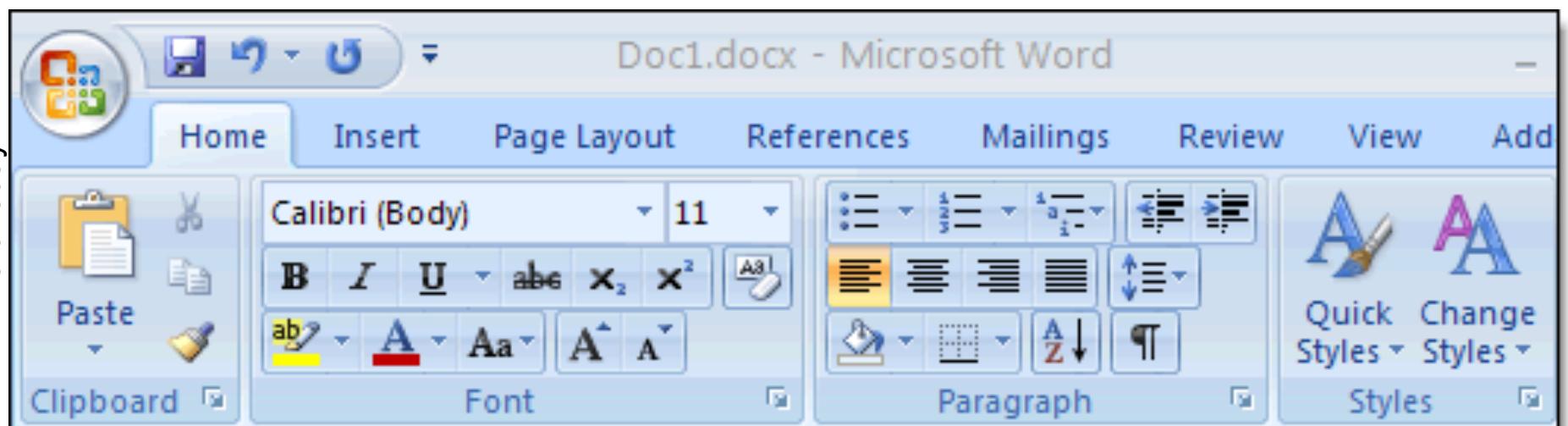
$$T = \underline{a + b} I_{\text{ndex}} D_{\text{ifficulty}}$$


(depends on input device)

Word 2000



Word today



e.g. reason why we have ribbons in Word now

Fitts' law :: $T = a + b \log_2 ID$

time required to **acquire a target**

but where does this equation come from?

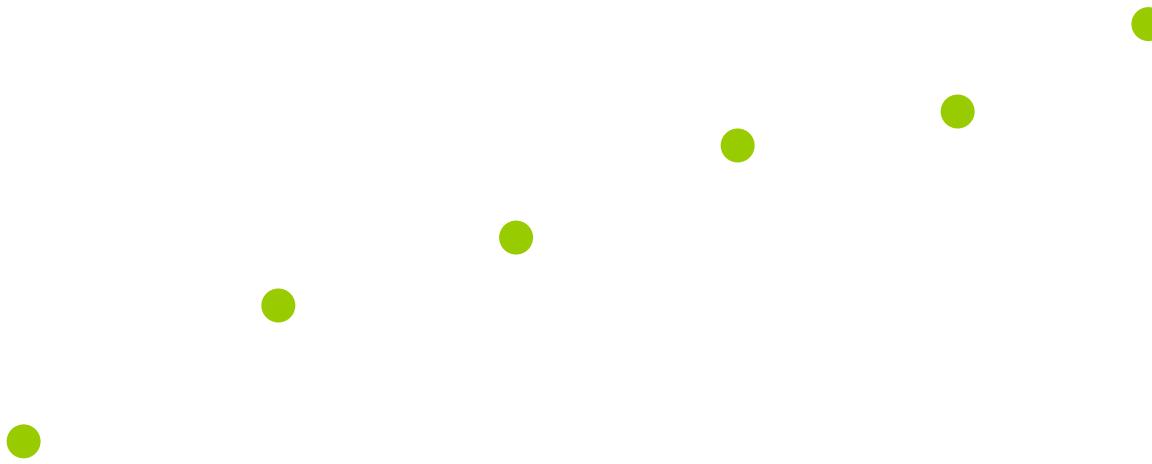
Trial [16] of 210

+



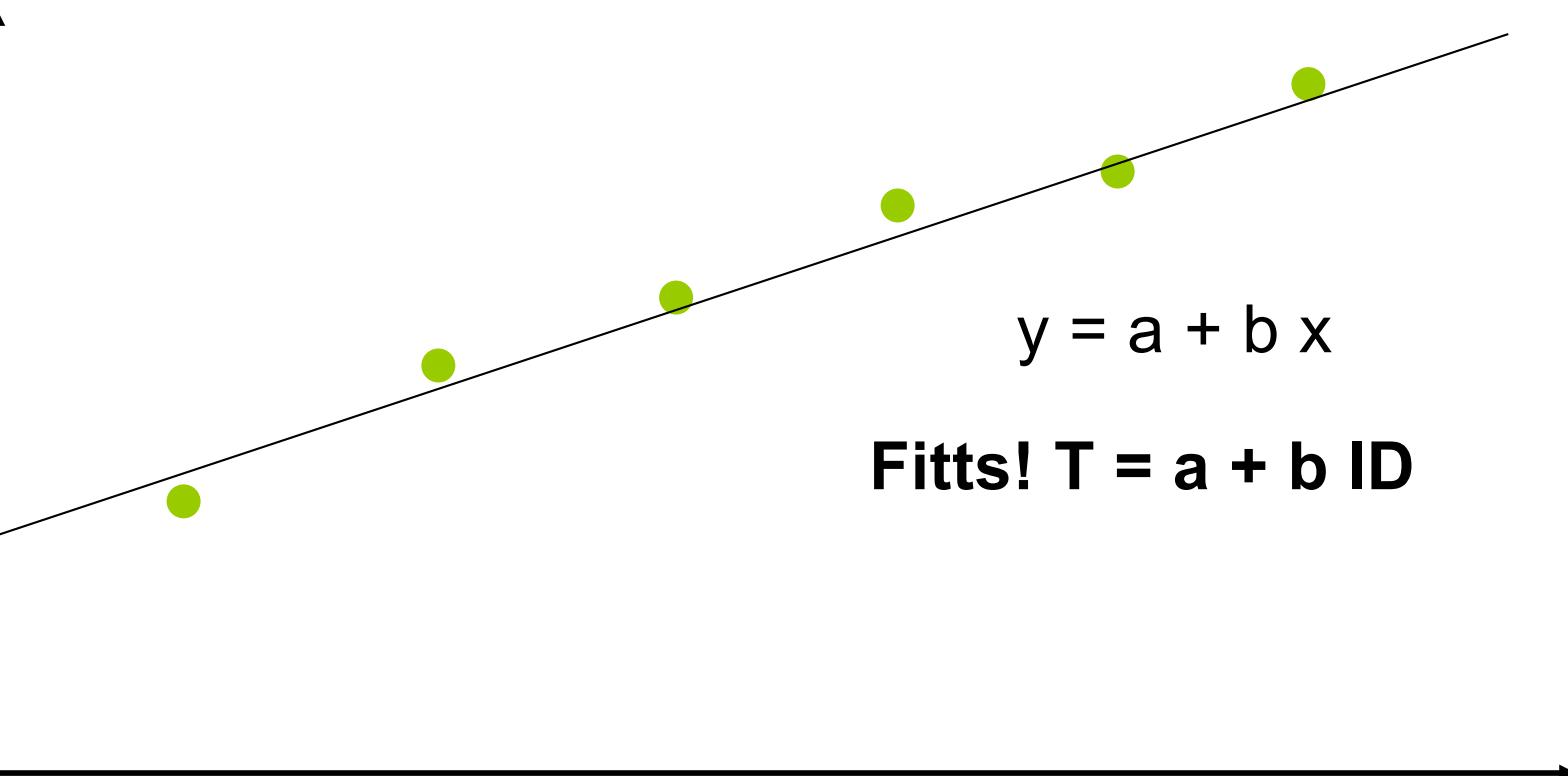
let's run an experiment and ask one participant to click on targets of different IDs

T time
(ms)



ID index of
difficulty

T time
(ms)



this a regression line

ID index of
difficulty

regression ::

a technique for determining the statistical relationship between two or more variables where a change in a **dependent variable** is associated with, and depends on, a change in one or more **independent variables**

arguably the most basic technique for **machine learning**

quick terminology of regressions

T time
(ms)



$$\hat{y} = a + b x$$

ID index of
difficulty

T time
(ms)



$$\hat{y} = a + b x$$

independent variable

ID index of
difficulty

T time
(ms)

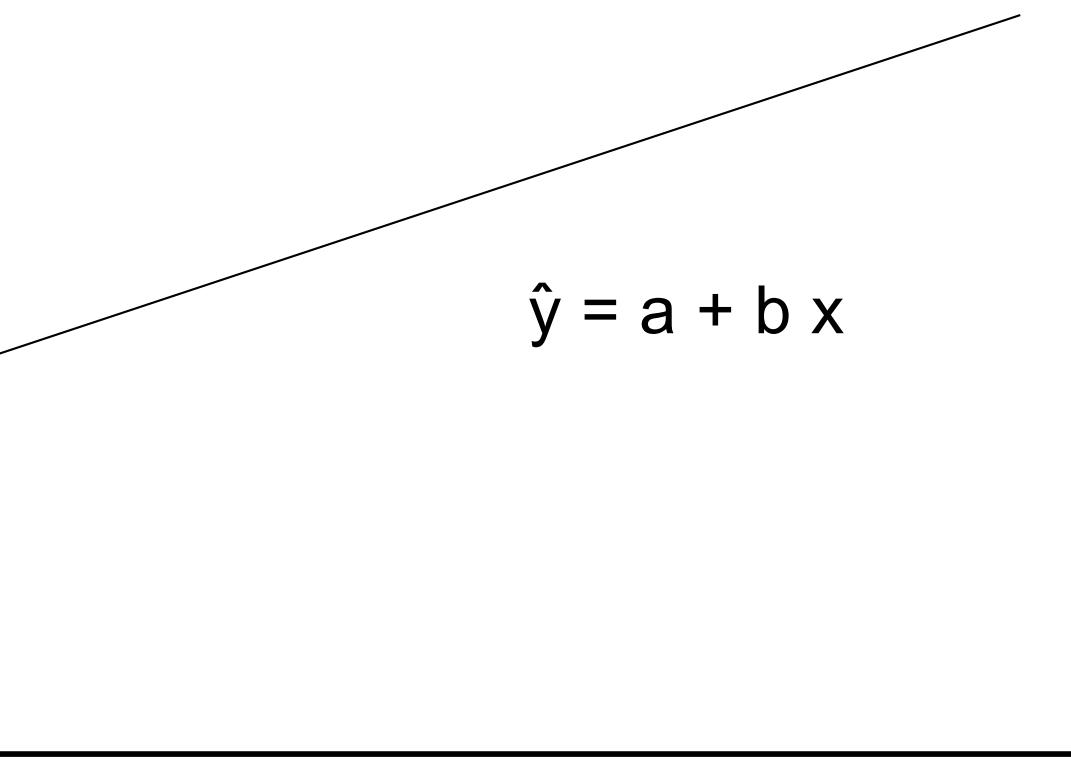
dependent variable



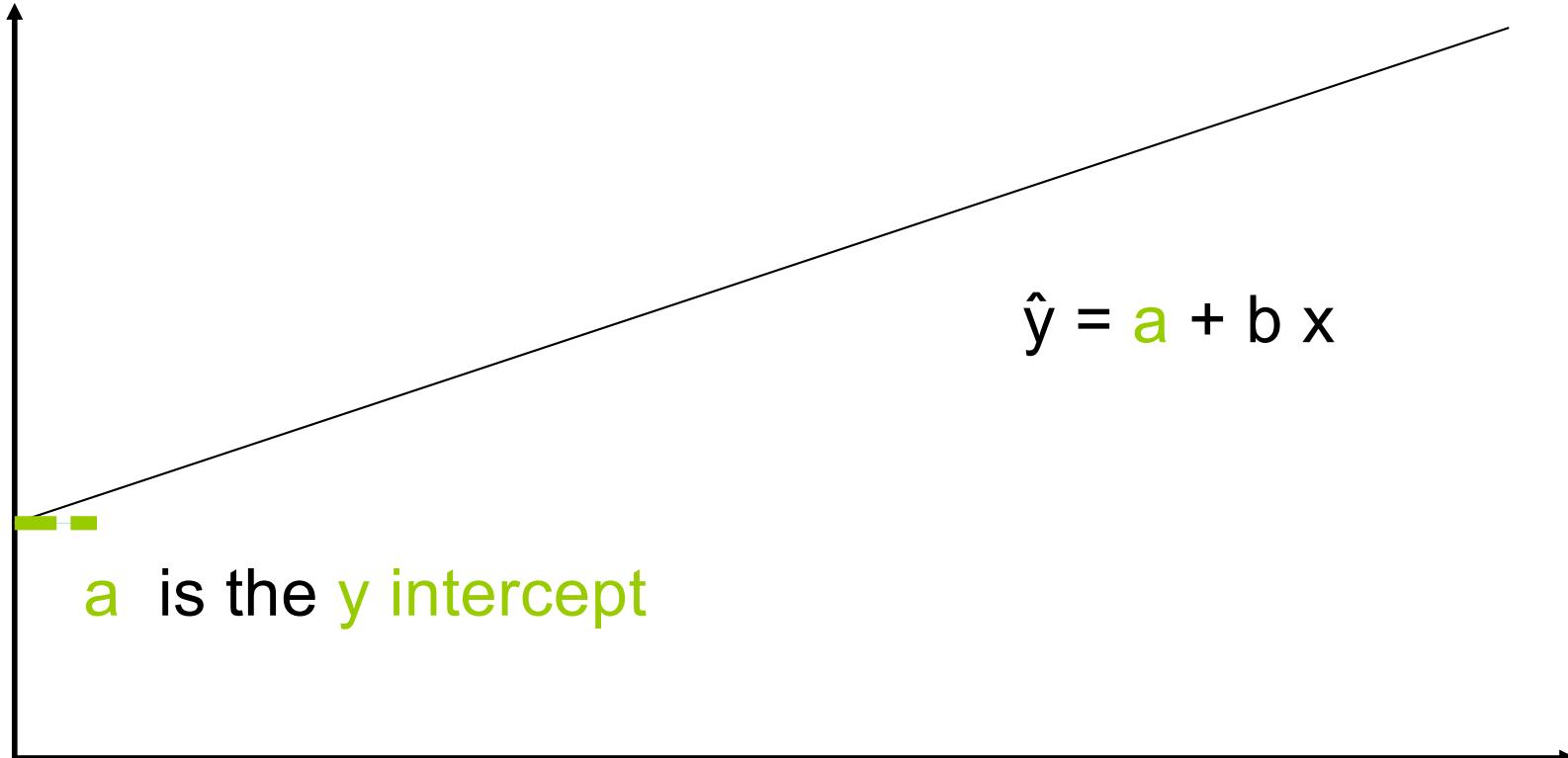
independent variable

$$\hat{y} = a + b x$$

ID index of
difficulty



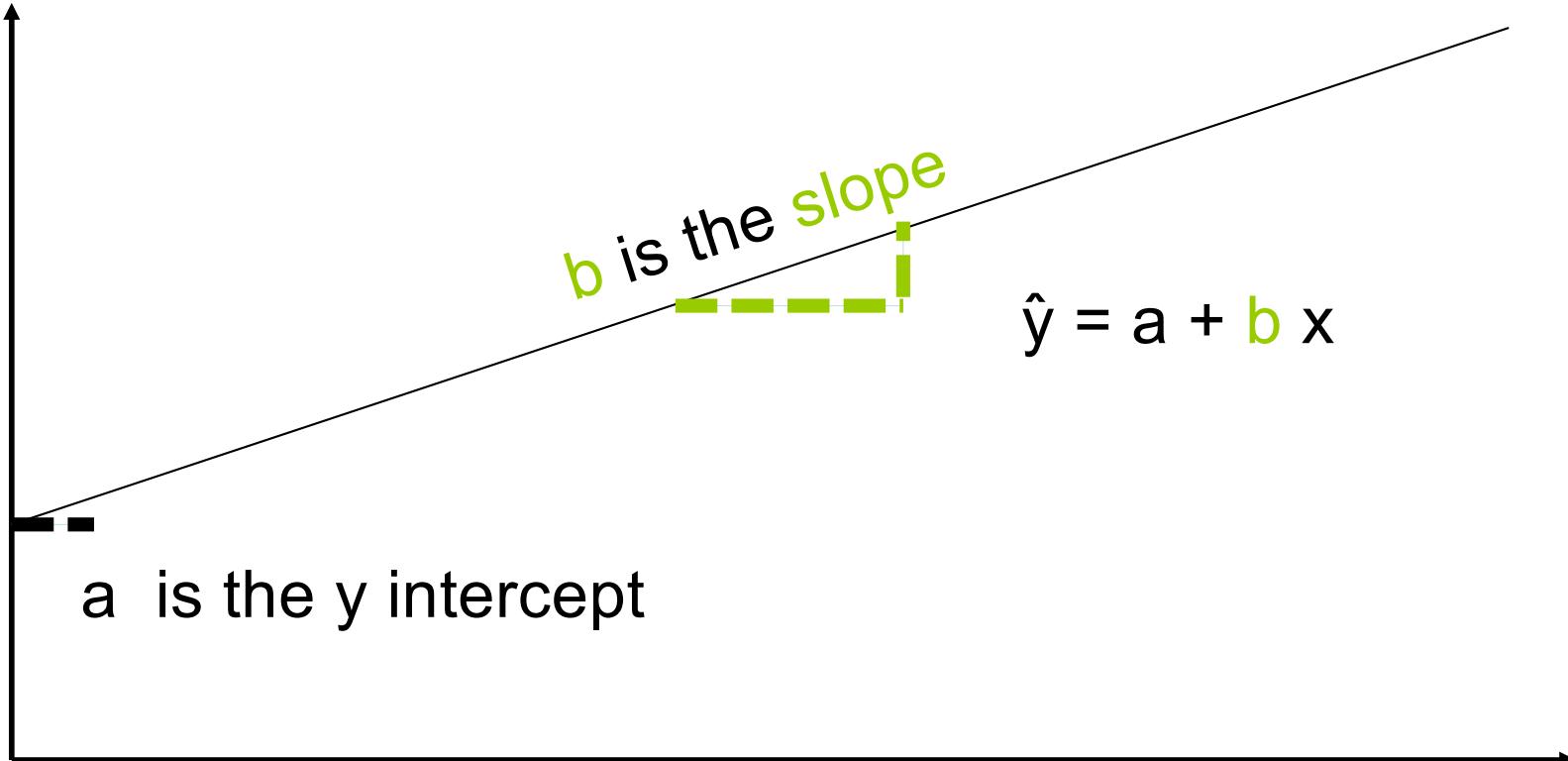
T time
(ms)



a is the y intercept

ID index of
difficulty

T time
(ms)



a is the y intercept

ID index of
difficulty

T time
(ms)

residual (deviation)

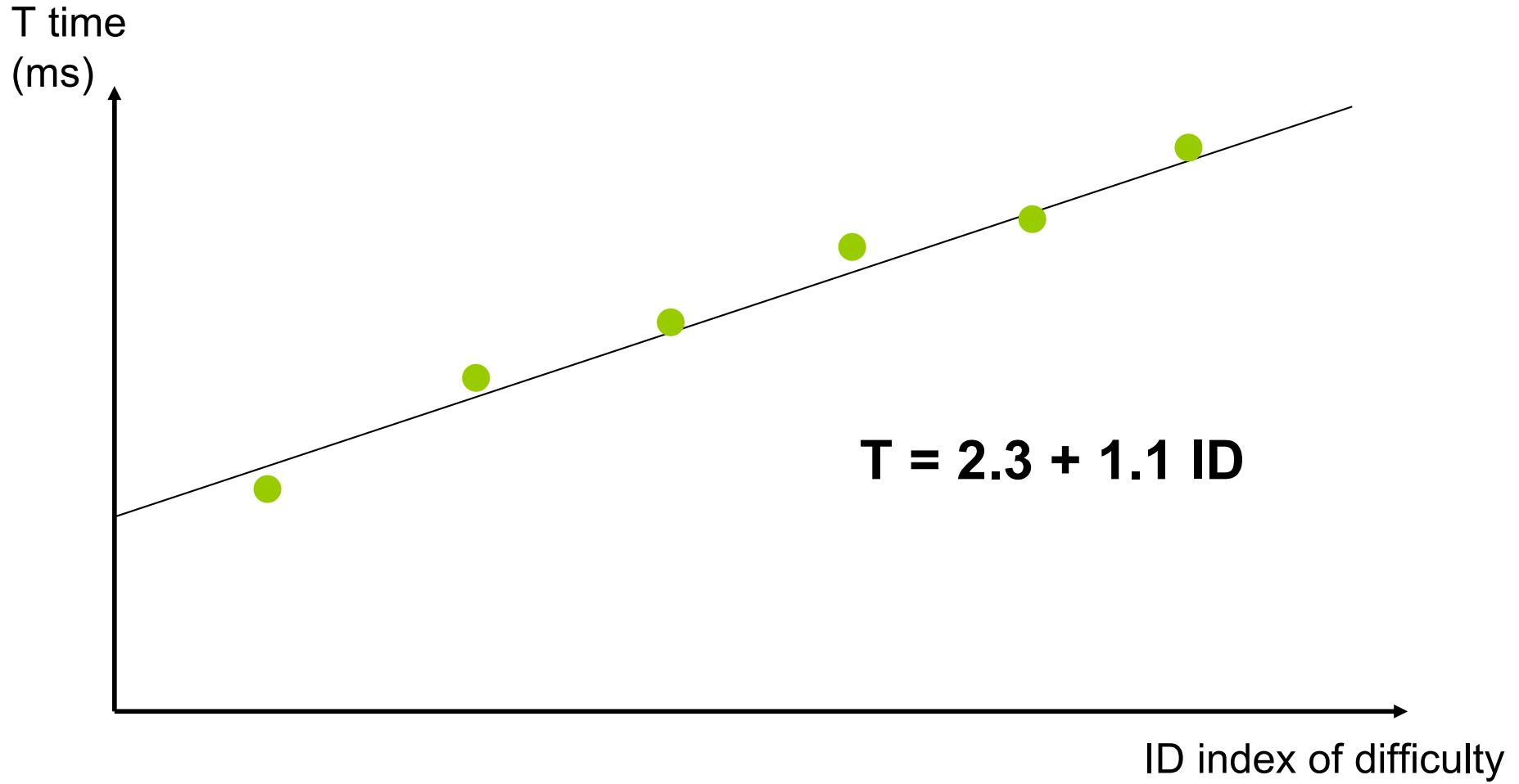
b is the slope

$$\hat{y} = a + b x$$

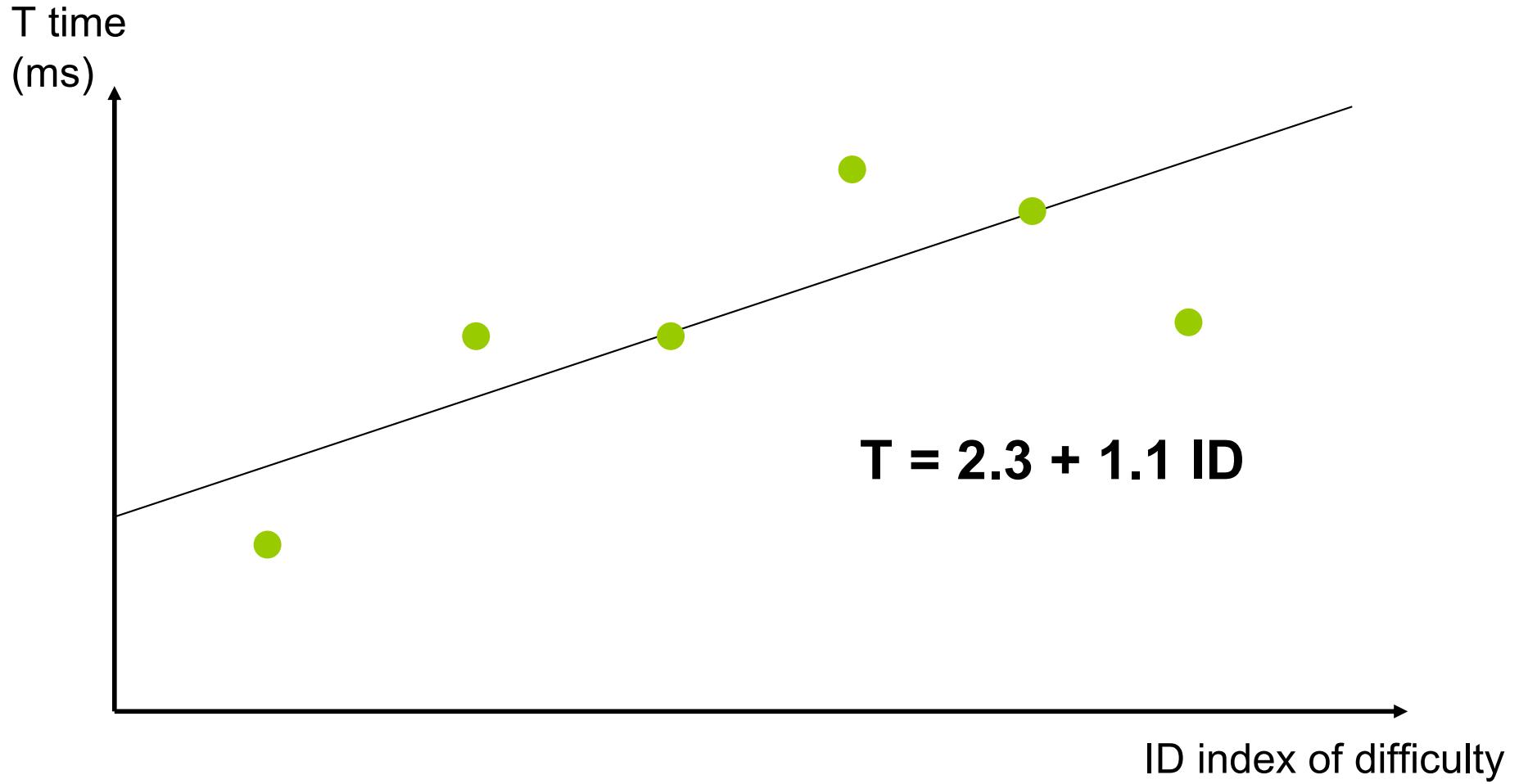
a is the y intercept

ID index of
difficulty

goodness
of fit



how can you be sure this line is a good fit to our data?



how can you be sure this line is a good fit to our data? what about now?

<brainstorming with your neighbor>

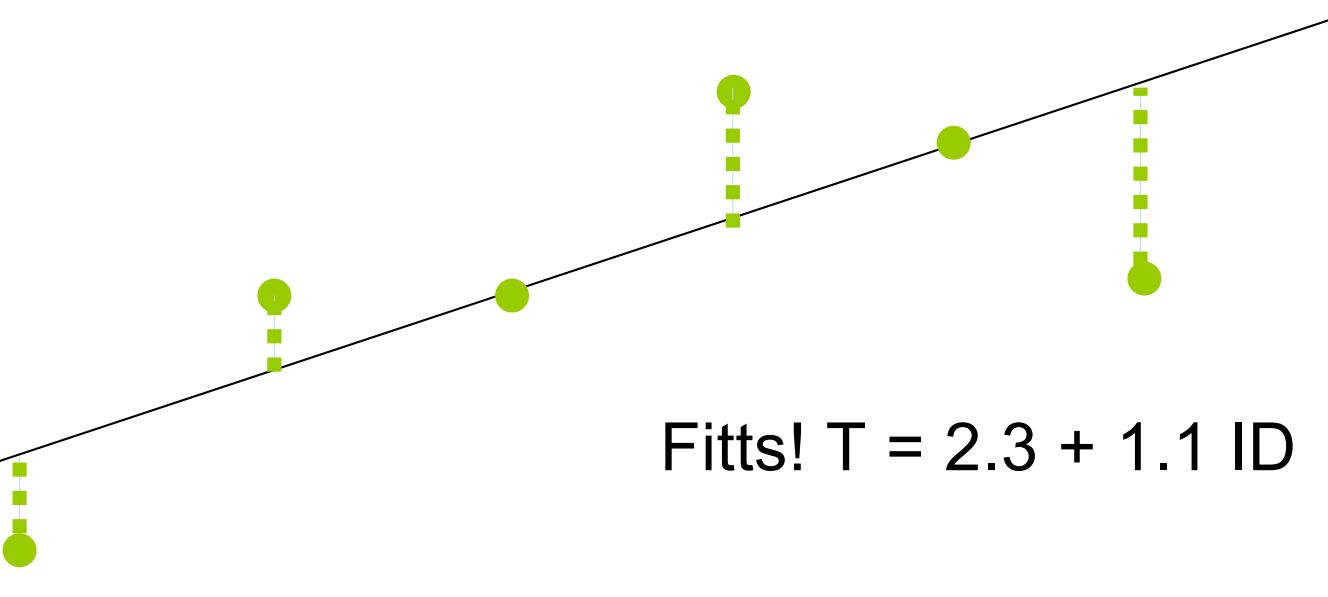
we can compute the **goodness of fit** with several methods

e.g. standard error of the estimate
or R squared

standard error of the estimate



T time
(ms)



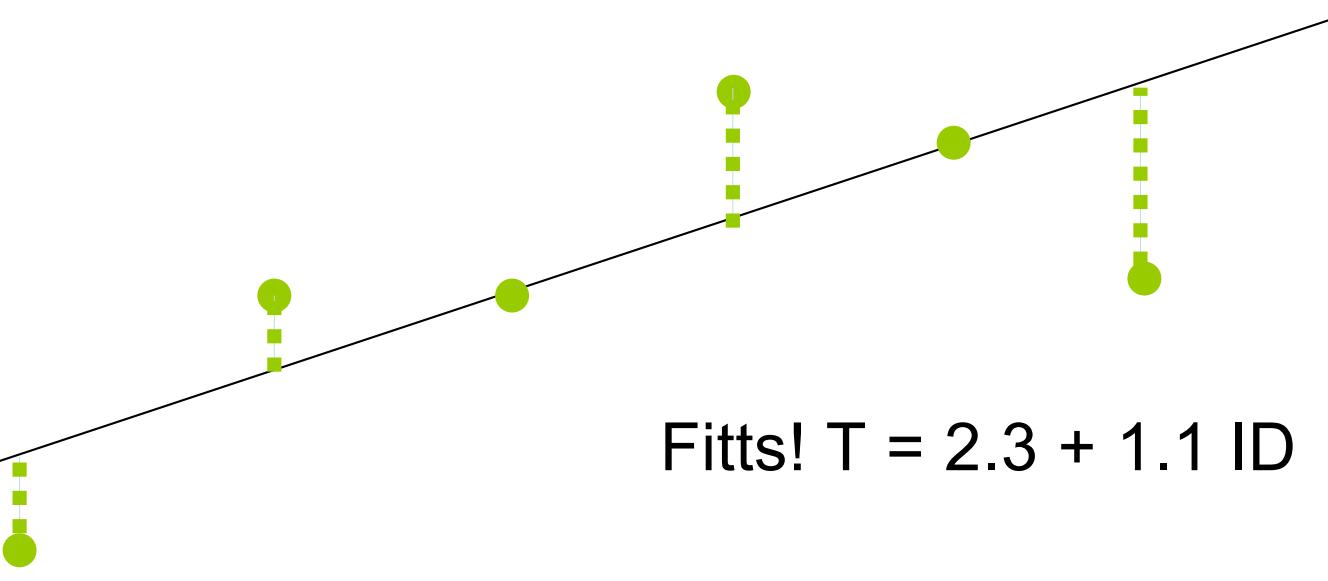
$$\sqrt{\frac{\sum (\text{error})^2}{(\text{sample size}-2)}}$$

ID index of
difficulty

standard error of the estimate



T time
(ms)



$$\text{Fitts! } T = 2.3 + 1.1 \text{ ID}$$

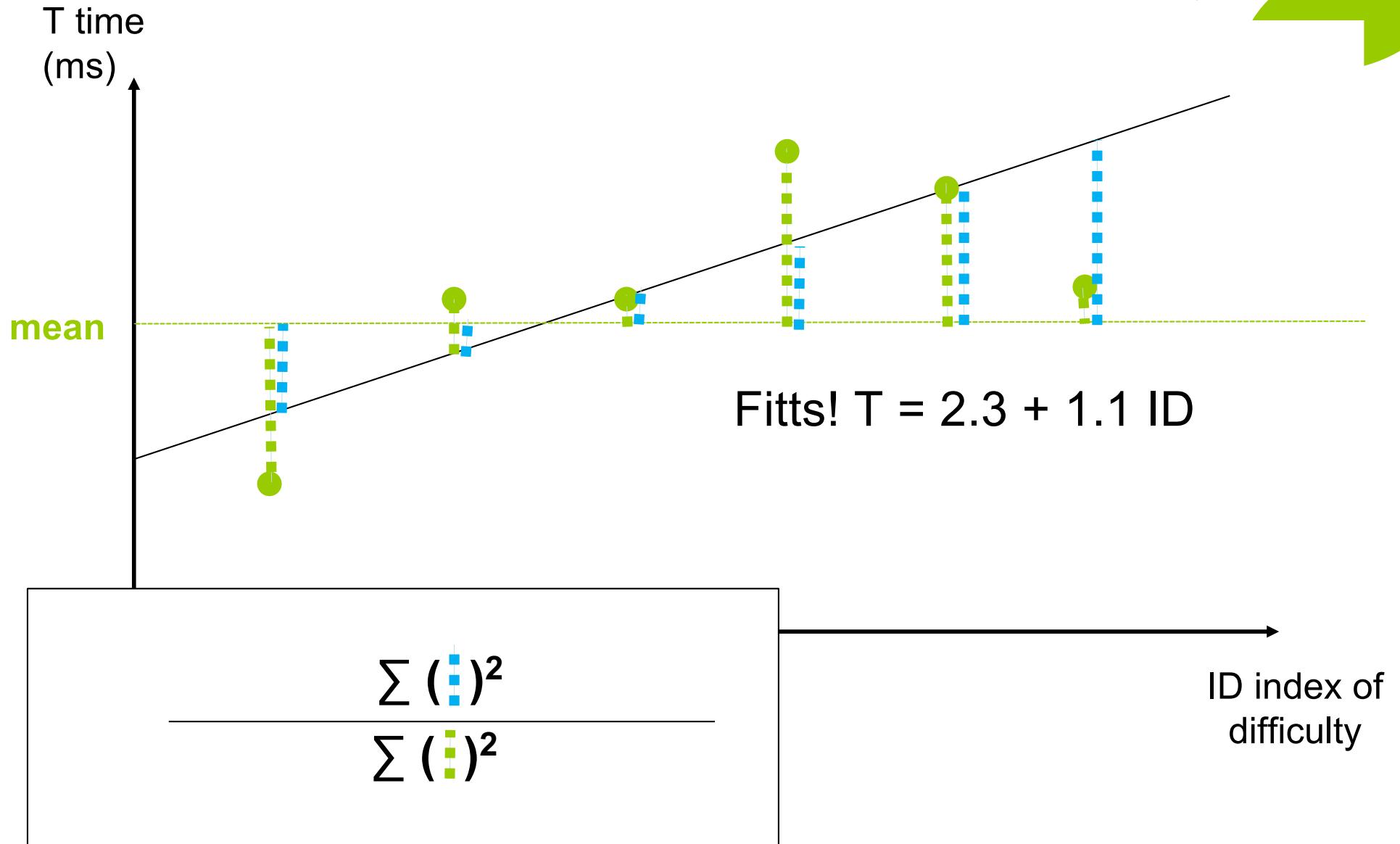
$$\sqrt{\frac{\sum (\text{estimated } \hat{y} - \text{actual } y)^2}{(\text{sample size}-2)}}$$

ID index of
difficulty

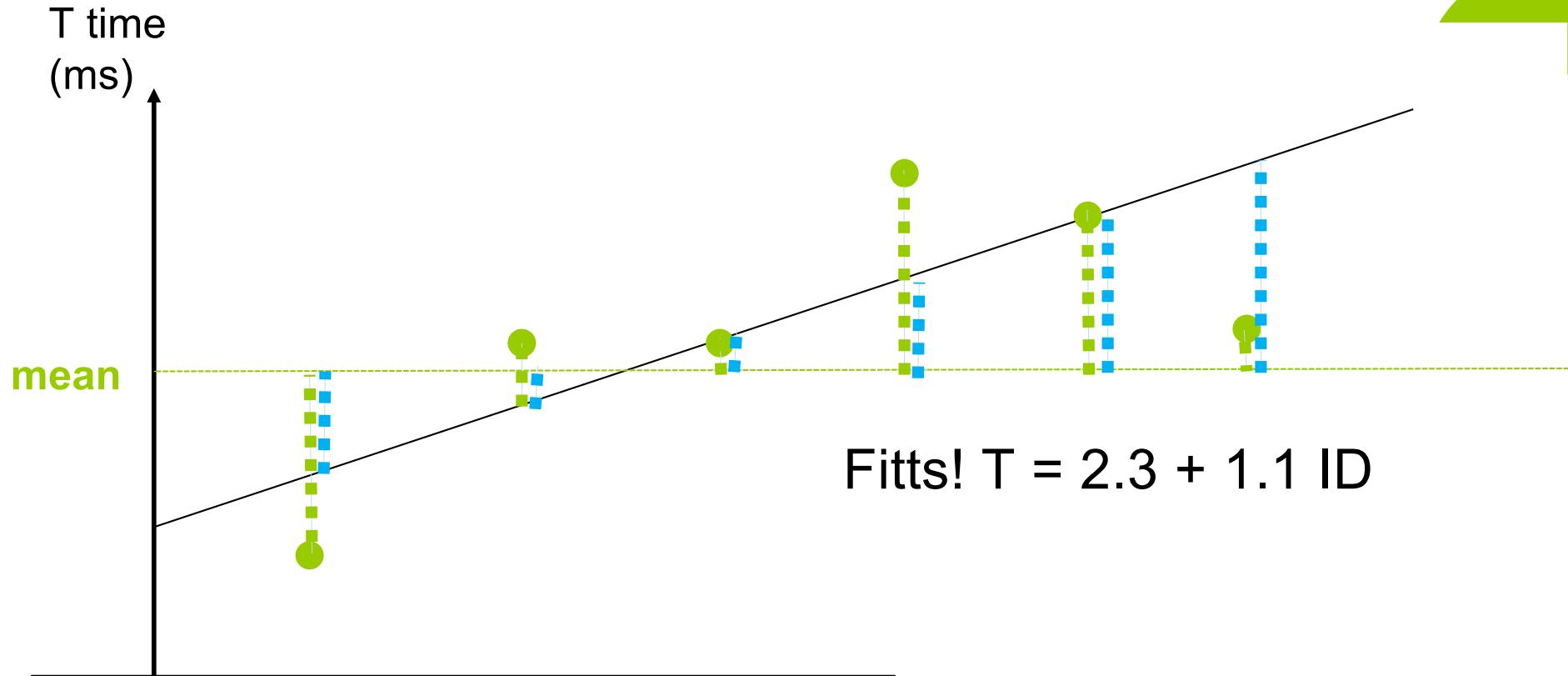
also called degree of freedom

S gives a standard error in the metric of the data (the less the better)

R squared



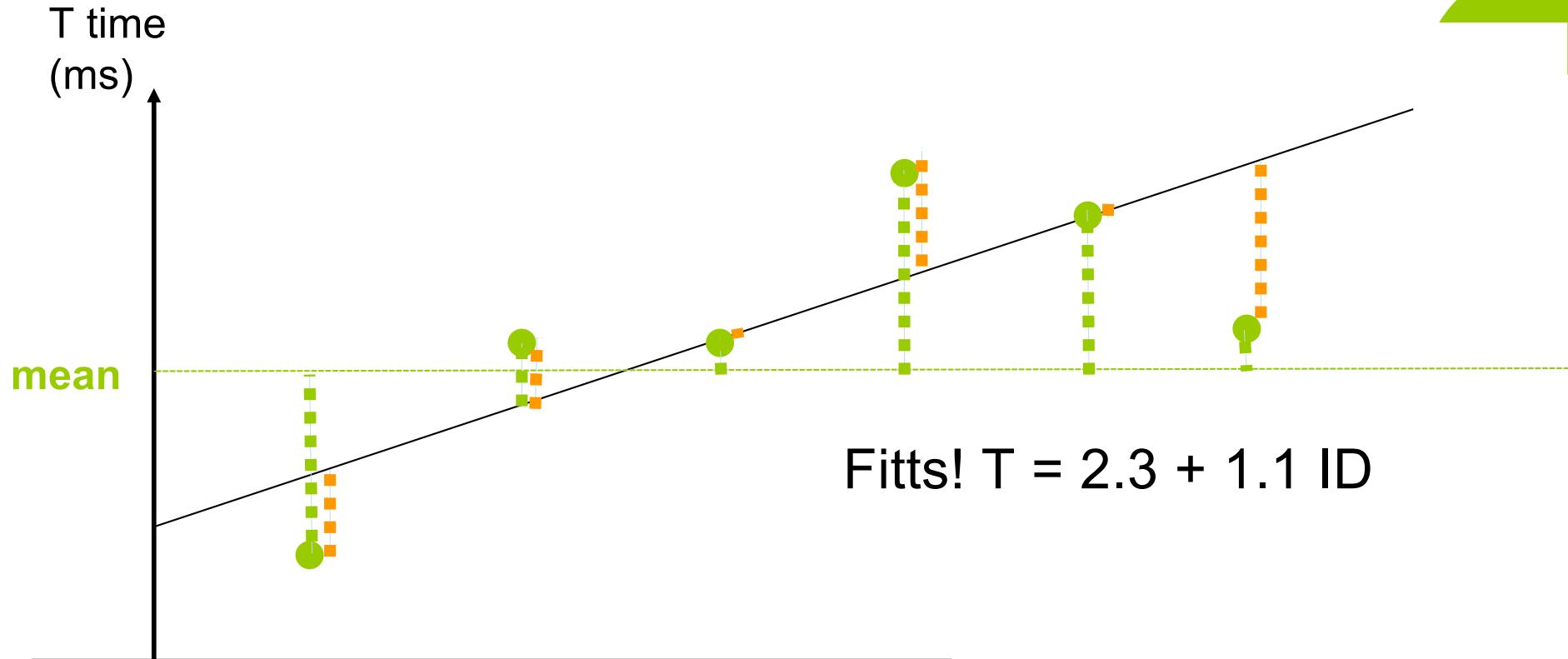
R squared



$$\frac{\sum (\text{estimated } \hat{y} - \text{mean } y)^2}{\sum (\text{actual } y - \text{mean } y)^2}$$

this formula works if your line is computed by the least square regression method (the one used by R)

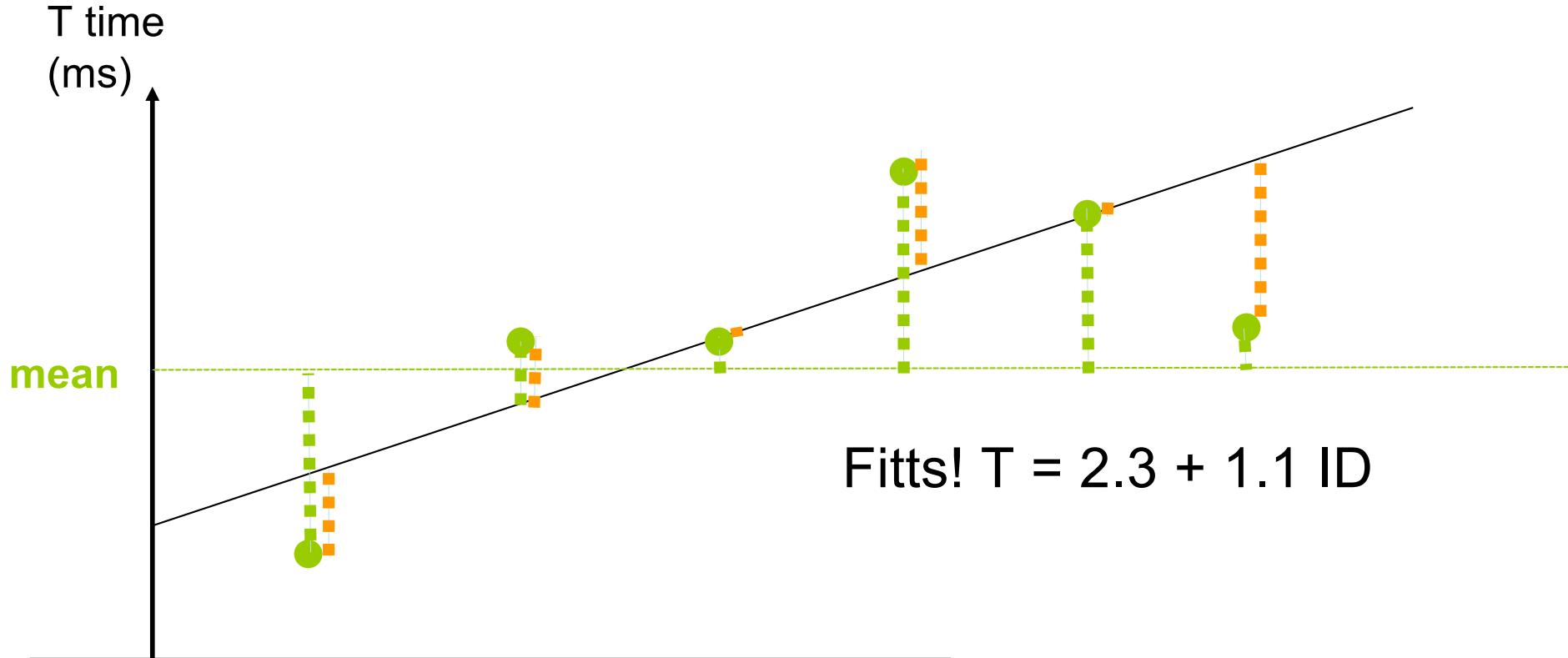
R squared



$$1 - \frac{\sum (\text{actual } y - \text{estimated } \hat{y})^2}{\sum (\text{actual } y - \text{mean } y)^2}$$

you could also find it
in a this format which
is **more genetic**

R squared



$$1 - \frac{\sum (\text{actual } y - \text{estimated } \hat{y})^2}{\sum (\text{actual } y - \text{mean } y)^2}$$

you could also find it
in a this format which
is **more genetic**

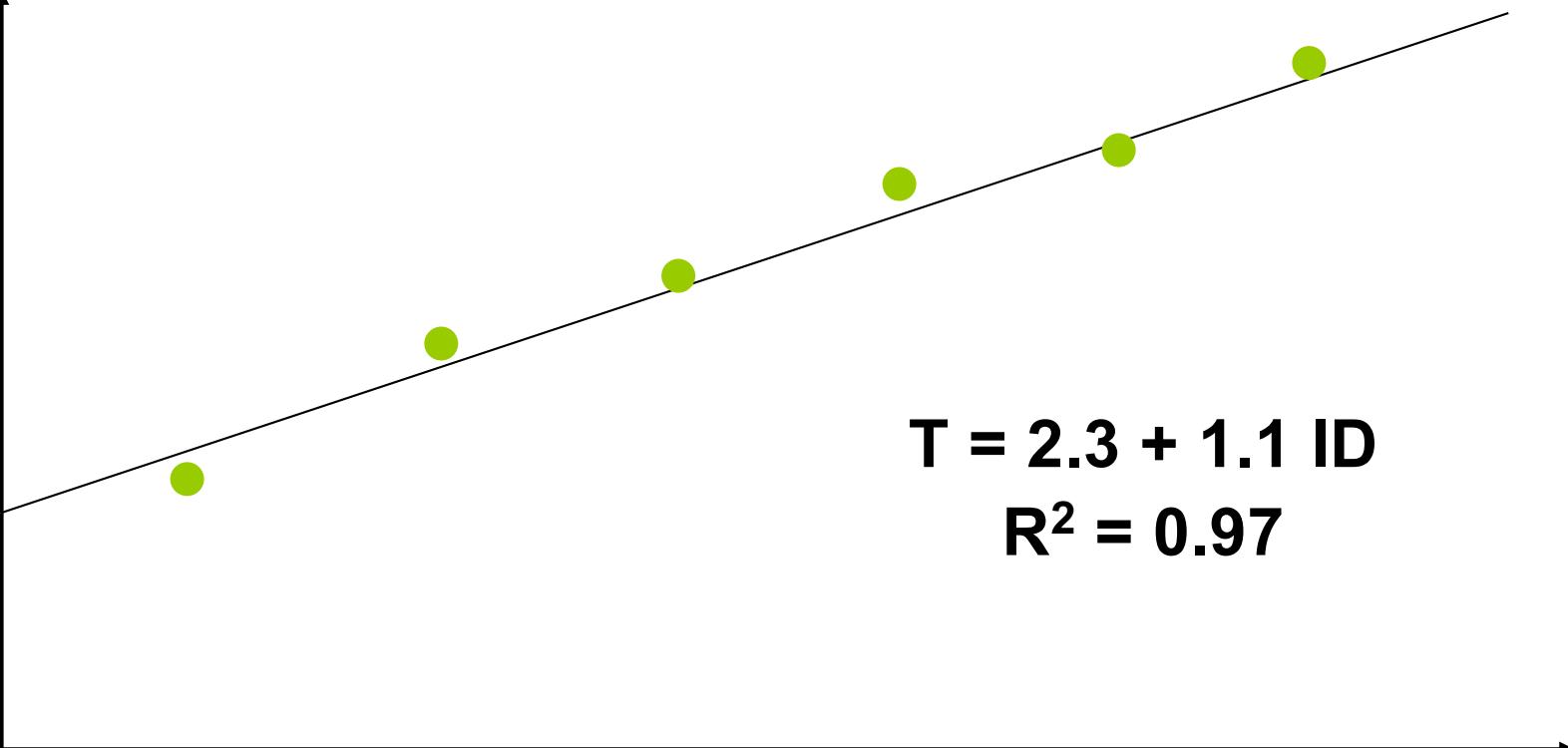
R^2 gives a percentage and 100% means perfect fit
(>70% is better)

T time
(ms)



$$T = 2.3 + 1.1 \text{ ID}$$
$$R^2 = 0.97$$

ID index of
difficulty



to gain additional confidence we repeat

we gain trust in a model if it fits the data with little error when

1. it is verified with a lot of data
2. it holds across very different people
3. it is verified in independent studies...



The information capacity of the human motor system in controlling the amplitude of movement.

PM Fitts - Journal of experimental psychology, 1954 - psycnet.apa.org

Reports of 3 experiments testing the hypothesis that the average duration of responses is directly proportional to the minimum average amount of information per response. The results show that the rate of performance is approximately constant over a wide range of movement amplitude and tolerance limits. This supports the thesis that" the performance capacity of the human motor system plus its associated visual and proprioceptive feedback mechanisms, when measured in information units, is relatively constant over a considerable ...

☆ 59 Cited by 7707 Related articles All 18 versions Web of Science: 3367

Fitts's paper probably most cited in HCI,
studies done and redone many times

practically



[vpn-user-244-044:~ nenisea]\$ R

Install at
<https://www.r-project.org/>

```
> print ("hello world!")
[1] "hello world!"
> █
```

we will be using **R** and I will try to give you
as much as possible of examples

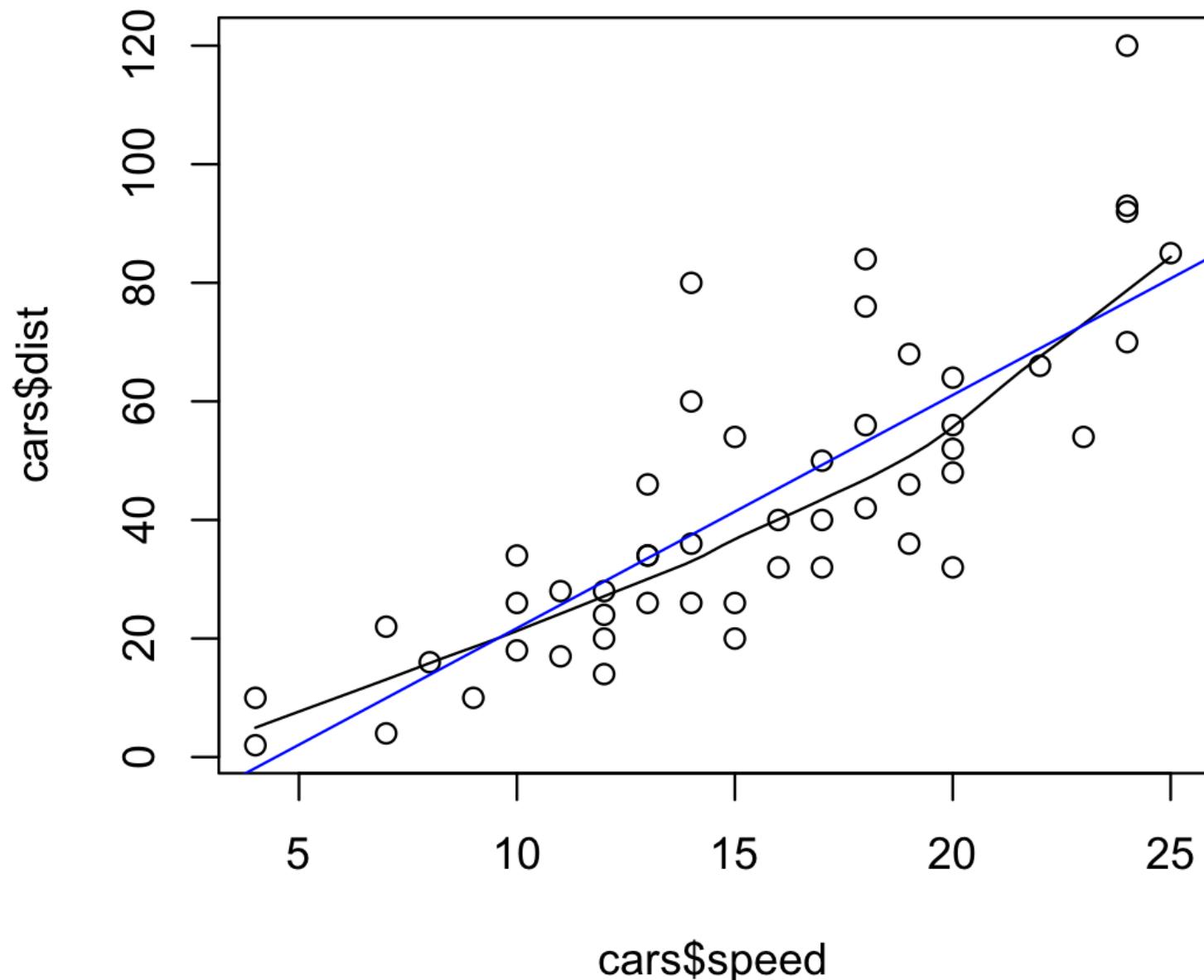


in your terminal

```
head(cars) # cars is a table that already comes with R and  
contain 50 observations of speed and distance in two rows  
  
scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~  
Speed")  
  
linearMod <- lm(dist ~ speed, data=cars) # build linear  
regression model  
  
abline(linearMod, col="blue") # draw the regression line  
  
summary(linearMod) # goodness of fit
```



Dist ~ Speed





Call:

```
lm(formula = dist ~ speed, data = cars)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.069	-9.525	-2.272	9.215	43.201

$$\text{dist} = -17.5791 + 3.9324 * \text{speed}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.5791	6.7584	-2.601	0.0123 *
speed	3.9324	0.4155	9.464	1.49e-12 ***

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

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 39.57 on 1 and 48 DF, p-value: 1.49e-12



Call:

```
lm(formula = dist ~ speed, data = cars)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.069	-9.525	-2.272	9.215	43.201

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.5791	6.7584	-2.601	0.0123 *
speed	3.9324	0.4155	9.464	1.49e-12 ***

also note this
pvalue<0.01
pvalue <0

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

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

(will explain next week)

usages of
regressions



Shop by category ▾

All

[Back to search results](#) | Listed in category: Art > Art from Dealers & Resellers > Prints

FREE SHIPPING



Pablo Picasso MARIE THERESE WALTER Estate

Item condition:

"Mint"

Quantity:

1

4 available / 241 sold

Price: US \$39.99

[Buy It Now](#)

[Add to cart](#)

285 watchers

[Add to watch list](#)

predicting ebay's online auction prices using functional data analysis

Mouse over image to zoom

Experienced
Seller

Hassle-free
Returns

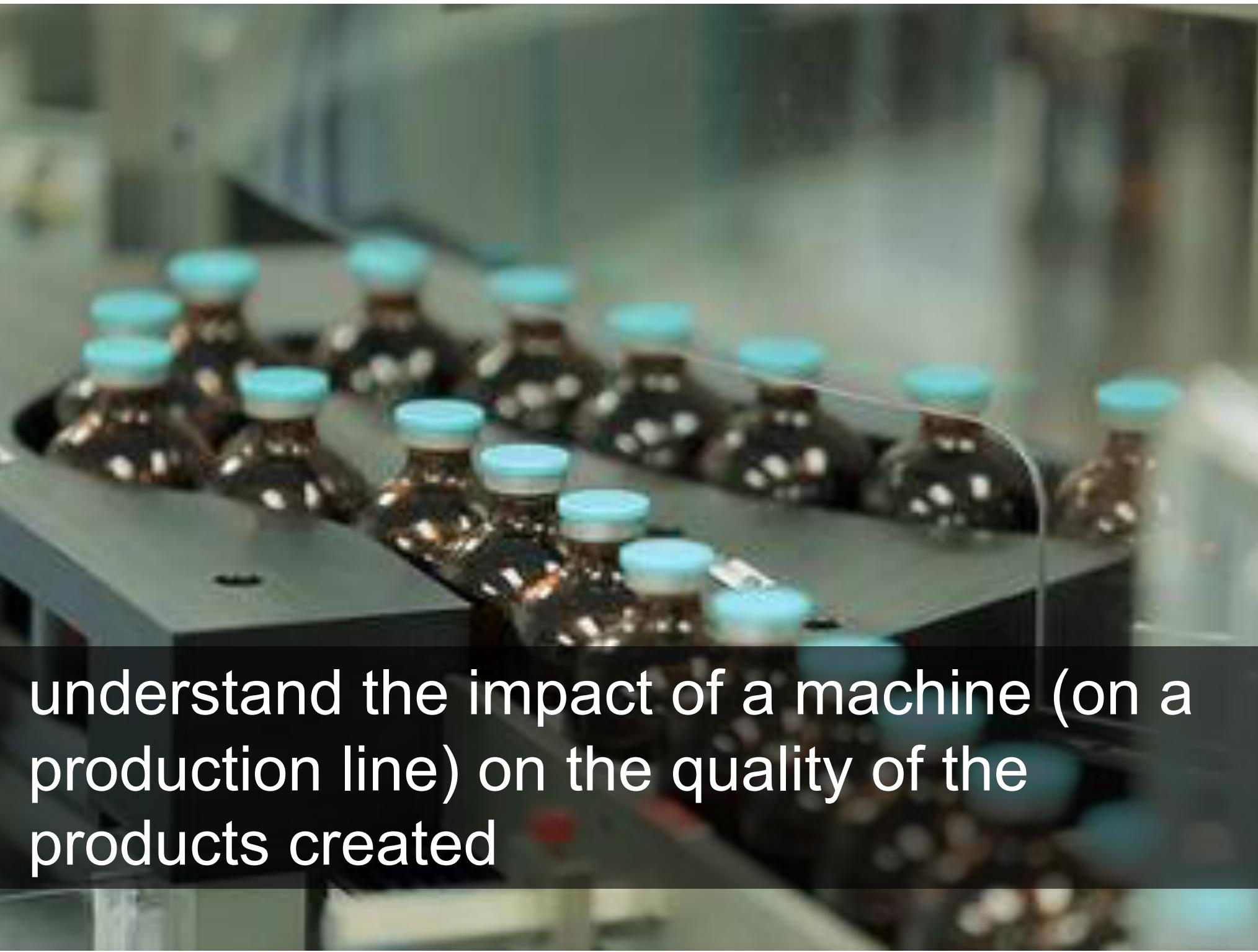
Free
Shipping

\$ Have one to sell? Sell it yourself

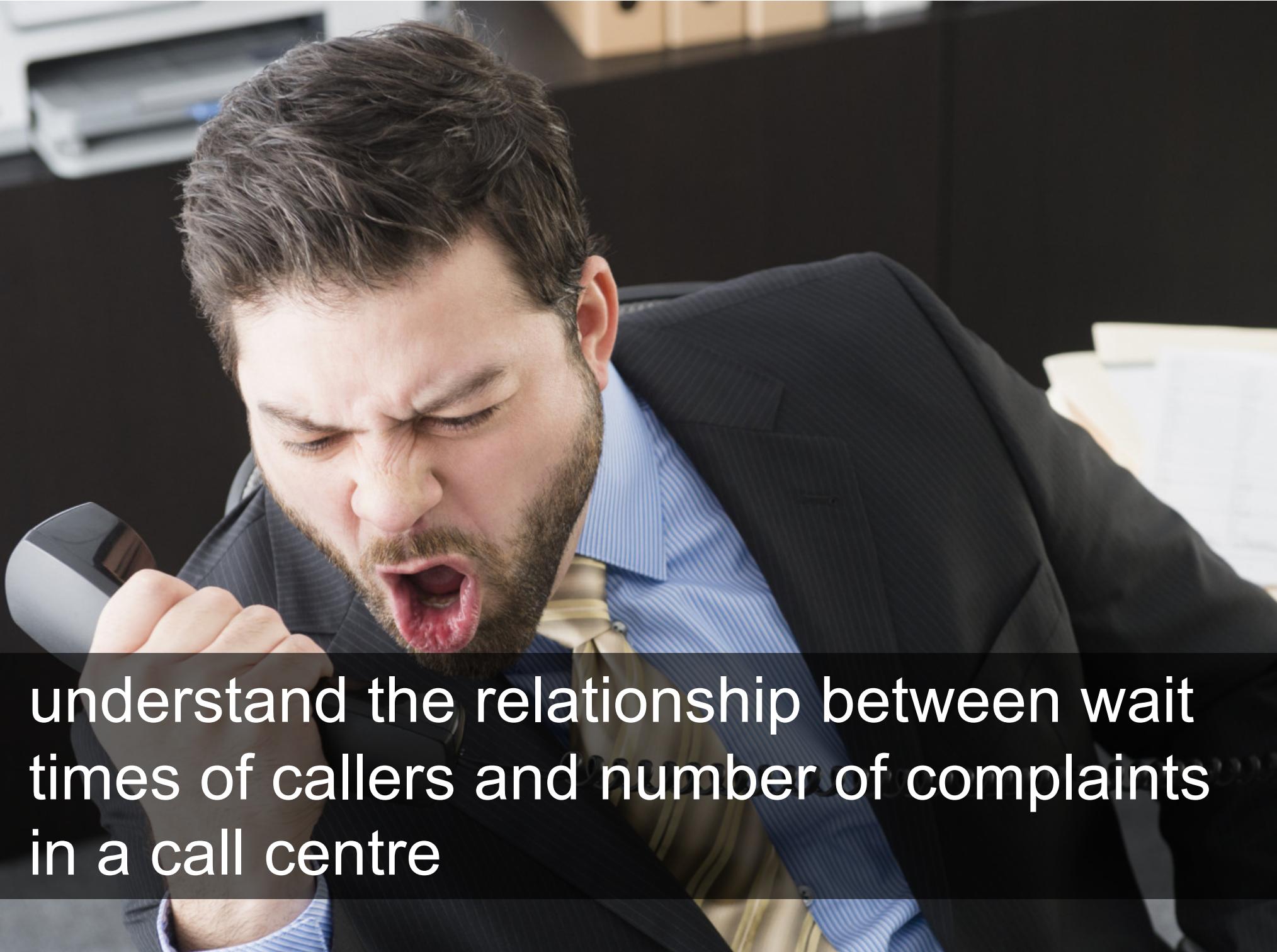
BillMeLater: New customers get \$15 back on 1st purchase
Subject to credit approval. See terms



predicting the number of passerby who will pass in front of a public display and use the data for choosing advertisement prices



understand the impact of a machine (on a production line) on the quality of the products created

A close-up photograph of a man with dark hair and a beard, wearing a dark pinstripe suit jacket, a light blue striped shirt, and a gold tie. He is shouting into a white telephone receiver held in his right hand. His mouth is wide open, and his eyes are squinted. The background is blurred, showing what appears to be an office environment with papers and files.

understand the relationship between wait
times of callers and number of complaints
in a call centre



retail store wants to extend shopping hours
to increase sales, but regression indicates
that increase in revenue not sufficient to
support rise in operating expenses

you can also fit a curve =
polynomial fitting

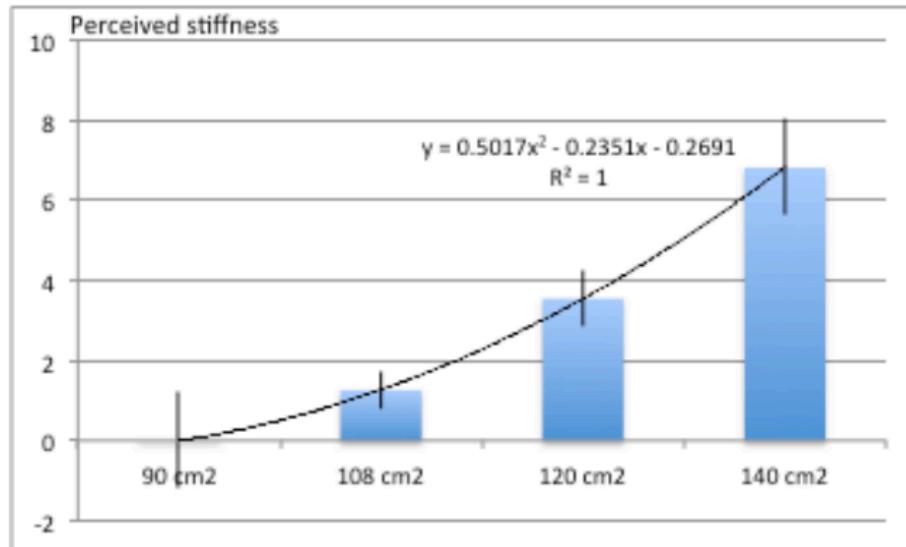


Figure 11. Bradley-Terry-Luce model output as well as a polynomial regression.

Our results are illustrated in Figure 11. We observed a clear distinction between the perceived stiffness of the 4 patches, the size of the patch increasing the perceived stiffness. In particular A is the least restrictive, followed by B, then C and then D is the most restrictive. We found that each paired comparison was significant ($p < 0.0125$). This thus allows us to compare the different patches and conclude that D is the most efficient patch. We also performed a polynomial regression on our data and found a very accurate fit: $y=0.5017x^2-0.2351x-0.2691$ ($R^2=1$). This suggests a quadratic correlation between the area of the patch and the perceived stiffness, which allows us to imagine bigger patches in order to restrict movements of the knee, which would require more stiffness. Of course further investigations need to be done to confirm this.

betwe
the ai
jam th

Prelin

The g
lab. A
the co
potent

First c

One p
hands
sugge
sugge

player
the u
“defrc

where
player
for ter
movei
our id

We al
becor
that th
in two
impro
particl
implej

Patch



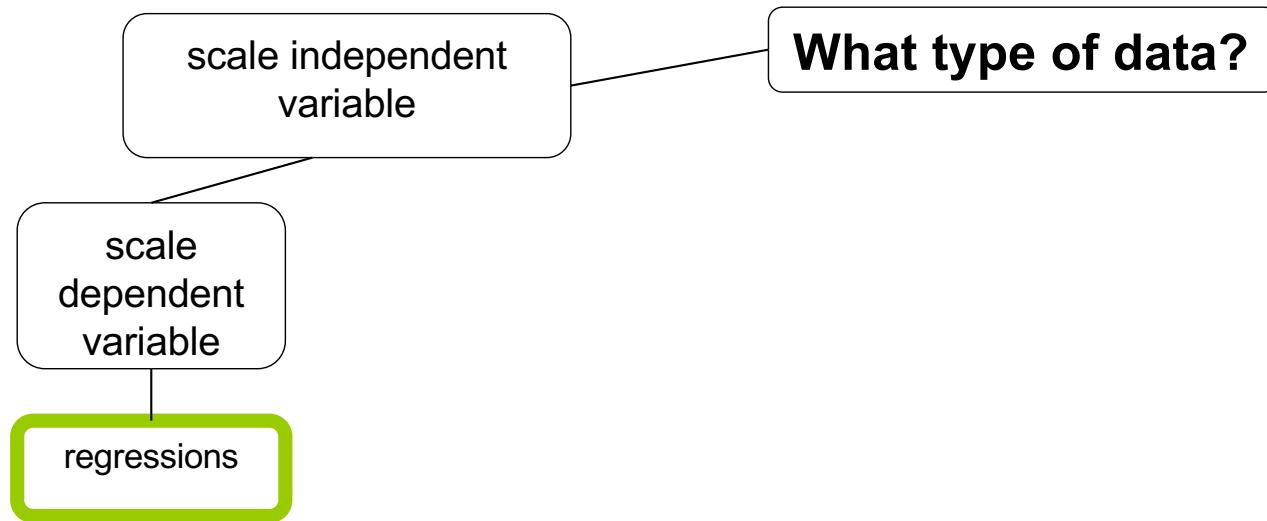
polynomial regression model

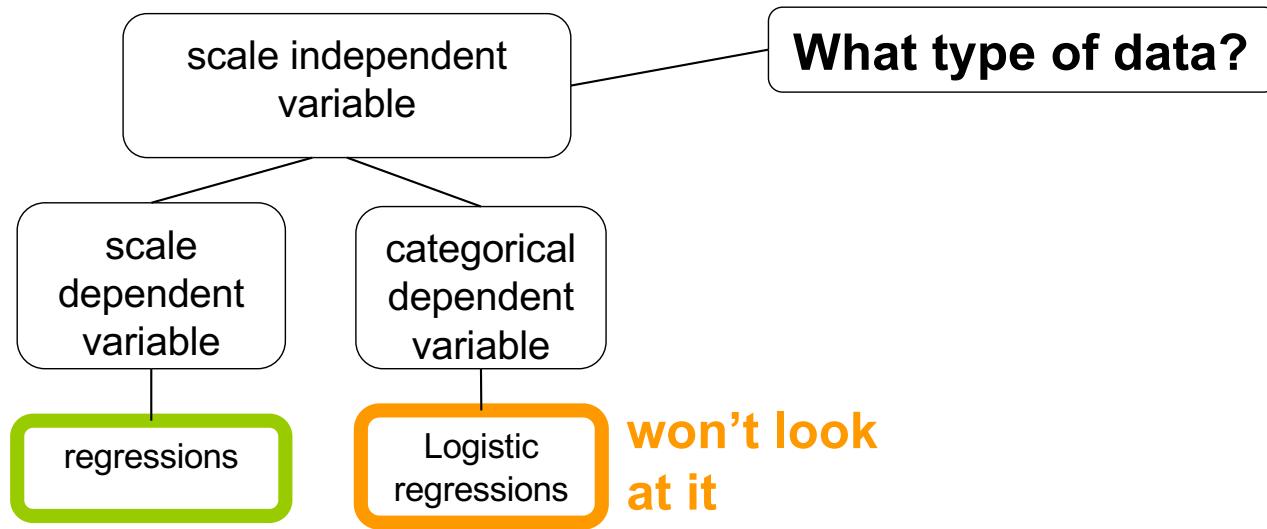
```
Mod2 <- lm(dist~poly(speed,2,raw=TRUE), data=cars)
```

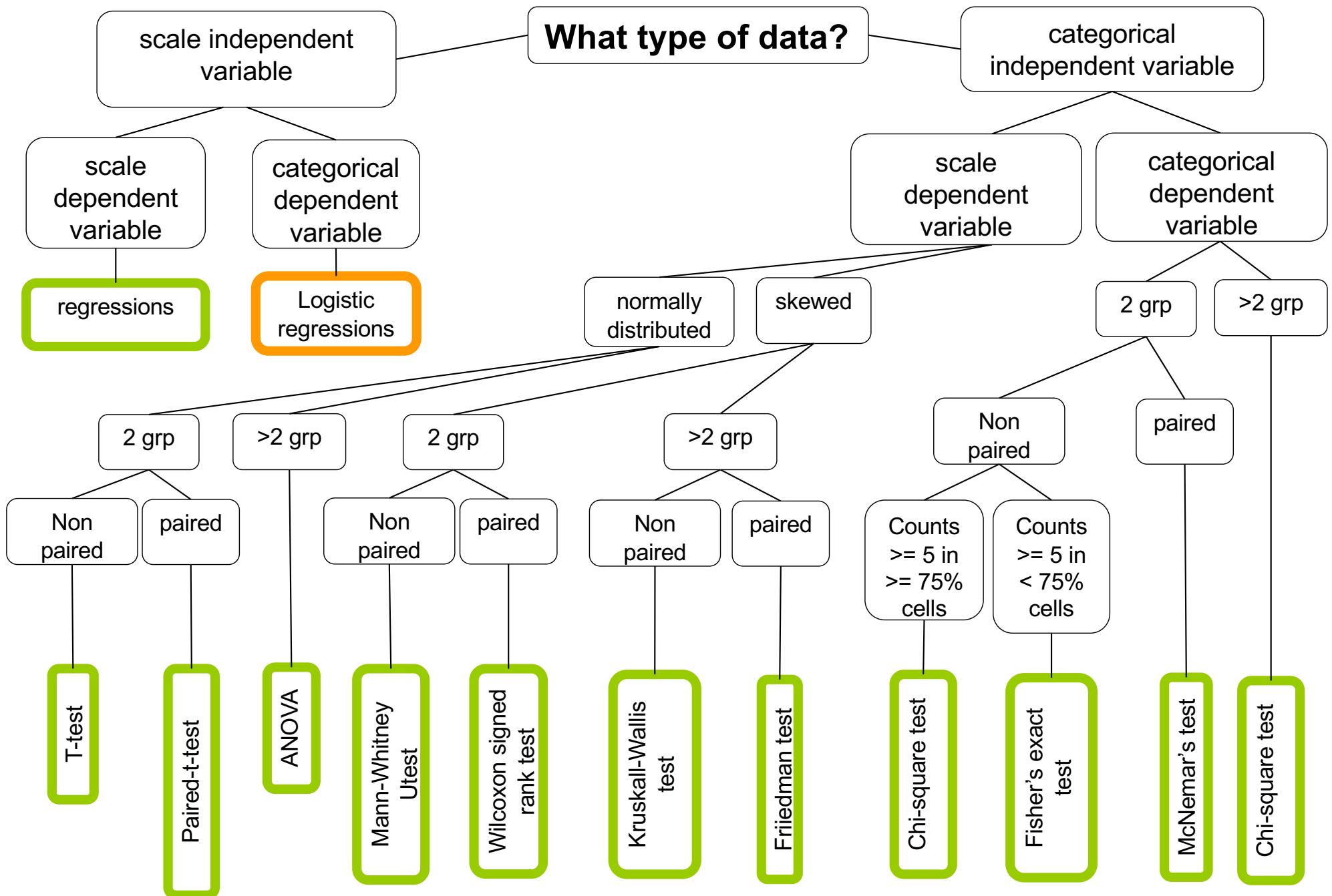
```
Mod3 <- lm(dist~poly(speed,3,raw=TRUE), data=cars)
```

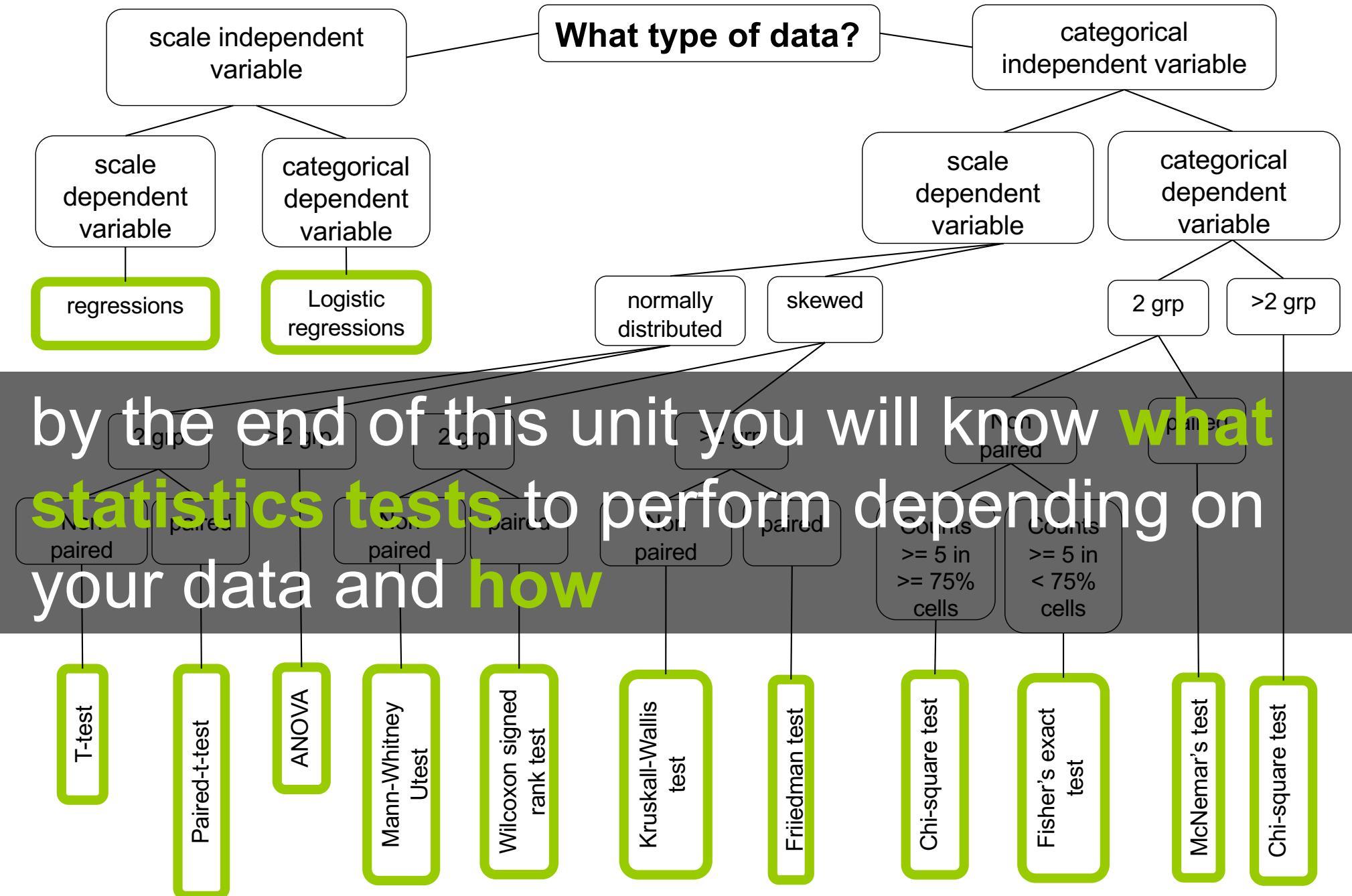
```
Mod4 <- lm(dist~poly(speed,4,raw=TRUE), data=cars)
```

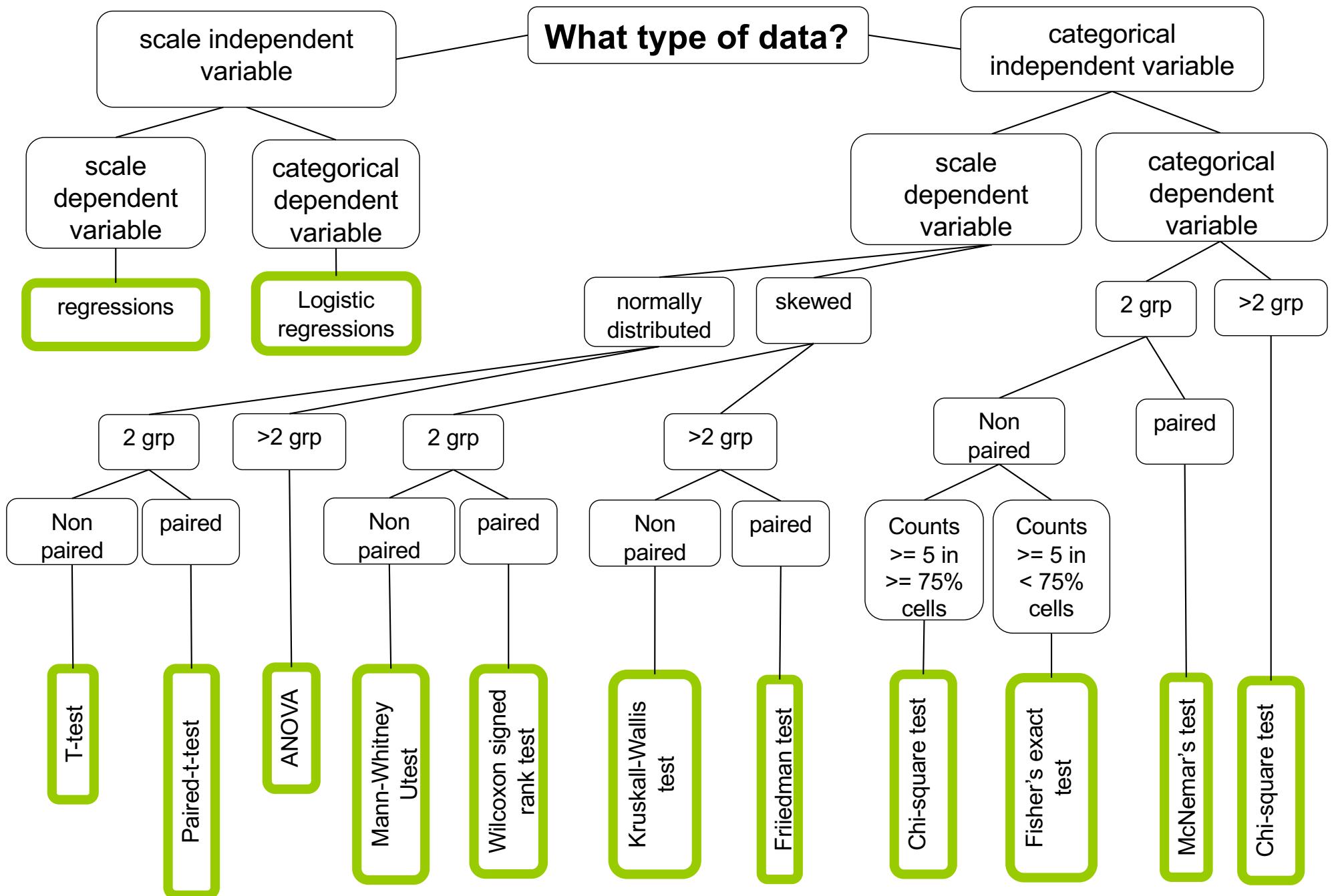
summary











1. Linear regression
2. Hypothesis testing, comparing things
3. Experimental design part 1
4. Experimental design part 2
5. T-test and ANOVA
6. Pre-requisite to ANOVA
7. Non-parametric tests
8. Categorical data: Chi-square
9. Sample size, power and effect size
10. P-hacking and alternatives tests

unit menu

resources

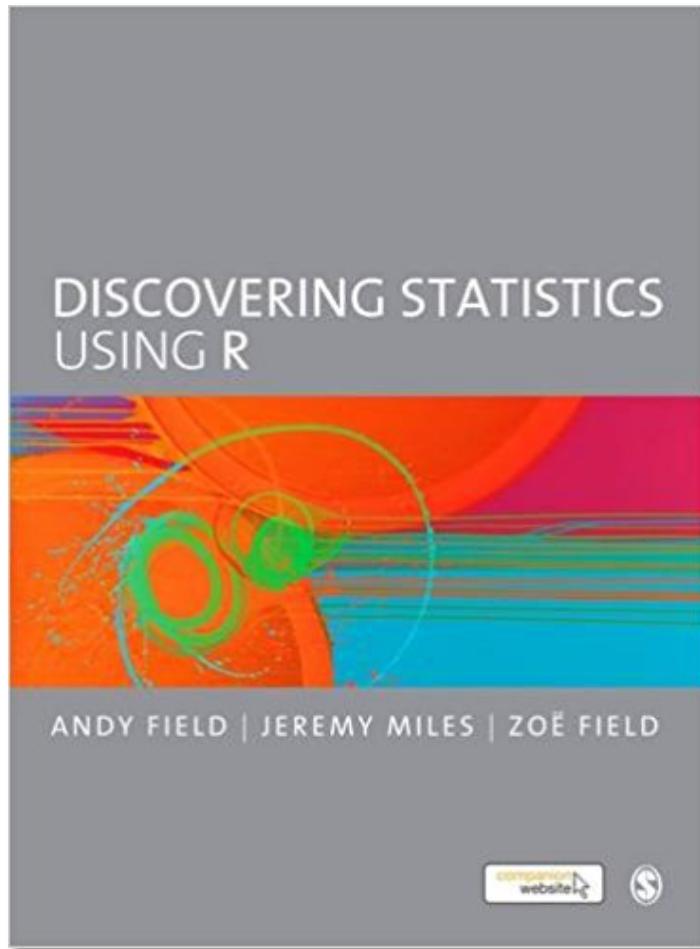
CHOOSING THE CORRECT STATISTICAL TEST IN SAS, STATA, SPSS AND R

The following table shows general guidelines for choosing a statistical analysis. We emphasize that these are general guidelines and should not be construed as hard and fast rules. Usually your data could be analyzed among them based on the number of dependent variables (sometimes referred to as outcome variables), the nature of your independent variables (sometimes referred to as predictors). You also want to consider normally distributed (see [What is the difference between categorical, ordinal and interval variables?](#) for more information on this). The table then shows one or more statistical tests commonly used given these types of data.

Number of Dependent Variables		Nature of Independent Variables	Nature of Dependent Variable(s)/Test(s)				How to			
			SAS	Stata	SPSS	R	SAS	Stata	SPSS	R
1	0 IVs (1 population)									
		Interval & normal	one-sample t-test							
		ordinal or interval	one-sample median	SAS	Stata	SPSS	R			
		categorical (2 categories)	binomial test	SAS	Stata	SPSS	R			
		categorical	Chi-square goodness-of-fit	SAS	Stata	SPSS	R			
	1 IV with 2 levels (Independent groups)	Interval & normal	2 Independent sample t-test	SAS	Stata	SPSS	R			
		ordinal or interval	Wilcoxon-Mann Whitney test	SAS	Stata	SPSS	R			
		categorical	Chi-square test	SAS	Stata	SPSS	R			
			Fisher's exact test	SAS	Stata	SPSS	R			
		Interval & normal	one-way ANOVA	SAS	Stata	SPSS	R			
		ordinal or interval	Kruskal Wallis	SAS	Stata	SPSS	R			
		categorical	Chi-square test	SAS	Stata	SPSS	R			
	1 IV with 2 or more levels (Independent groups)	Interval & normal	paired t-test	SAS	Stata	SPSS	R			
		ordinal or interval	Wilcoxon signed ranks test	SAS	Stata	SPSS	R			
		categorical	McNemar	SAS	Stata	SPSS	R			
	1 IV with 2 levels (dependent/matched groups)	Interval & normal	one-way repeated measures ANOVA	SAS	Stata	SPSS	R			
		ordinal or interval	Friedman test	SAS	Stata	SPSS	R			
		categorical	repeated measures logistic regression	SAS	Stata	SPSS	R			
	1 IV with 2 or more levels (dependent/matched groups)	Interval & normal	factorial ANOVA	SAS	Stata	SPSS	R			
		ordinal or interval	ordered logistic regression	SAS	Stata	SPSS	R			
		categorical	factorial logistic regression	SAS	Stata	SPSS	R			
	1 or more IVs (Independent groups)	Interval & normal	correlation	SAS	Stata	SPSS	R			
		interval & normal	simple linear regression	SAS	Stata	SPSS	R			
		ordinal or interval	non-parametric correlation	SAS	Stata	SPSS	R			
		categorical	simple logistic regression	SAS	Stata	SPSS	R			
	1+ IV	Interval & normal	multiple regression	SAS	Stata	SPSS	R			
			analysis of covariance	SAS	Stata	SPSS	R			
		categorical	multiple logistic regression	SAS	Stata	SPSS	R			
			discriminant analysis	SAS	Stata	SPSS	R			
2+	1 IV with 2 or more levels (Independent groups)	Interval & normal	one-way MANOVA	SAS	Stata	SPSS	R			
	2+	Interval & normal	multivariate multiple linear regression	SAS	Stata	SPSS	R			
	0	Interval & normal	factor analysis	SAS	Stata	SPSS	R			
2 sets of 2+	0	Interval & normal	canonical correlation	SAS	Stata	SPSS	R			

This page was adapted from [Choosing the Correct Statistic](#) developed by James D. Leeper, Ph.D. We thank Professor Leeper for permission to adapt and distribute this page from our site.

<https://stats.idre.ucla.edu/other/mult-pkg/whatstat/>



the **text book** I am using and a suggestion of **YouTube video channel**

videos on regressions

<https://www.youtube.com/watch?v=WWqE7YHR4Jc>

<https://www.youtube.com/playlist?list=PLF596A4043DBAE9C>

1. Explain what is a linear regression
2. Give the terminology of a regression line
3. Give the two formulas of goodness of fit
4. Be able to compute the two formulas given a few observations

take away

end