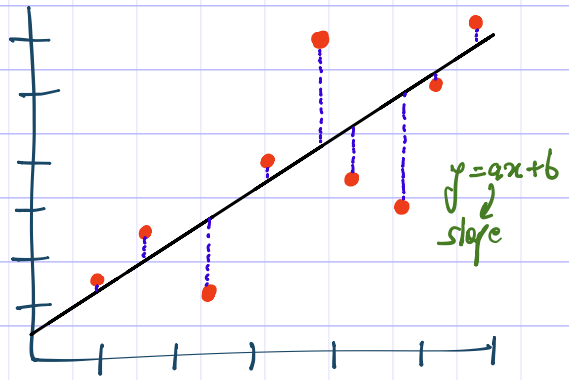
Linear Regression

→ fitting a line to data
→ Least squares
→ Linear Regression

Concepts:

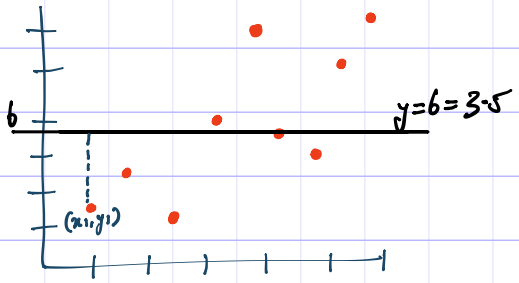
#1 we want to minimize the distance between the observed values and the line



#2 we do this by taking the derivative and finding where it is equal to 0

#3 The final line minimizes the sum of squares (It gives the "least squares") between it & the real data.

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



The distance between the line & the first data point is $b-y_1$, the second data point is $b-y_2$

Total distance $= (b-y_1)+(b-y_2)+(b-y_3)+(b-y_4)+\cdots$

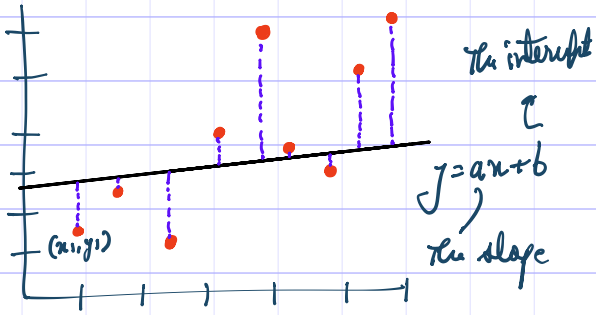
* we can take the absolute values $|(b-y_1)|$ but it will make the math tricky.

So they ended up squaring each term. Squaring ensures that each term is positive.

$(b-y_1)^2+(b-y_2)^2+(b-y_3)^2+\cdots$

$= \sum (y-b)^2$

* This is our measure of how well this line fits the data.
* It's called the sum of squared residuals
  → difference between the real data & the line
* If we rotate the line a little bit, the sum of squared residuals = 18.72 (better than before)



* we want to find the optimal values of (a) & (b) so that we minimize the sum of squared residuals.

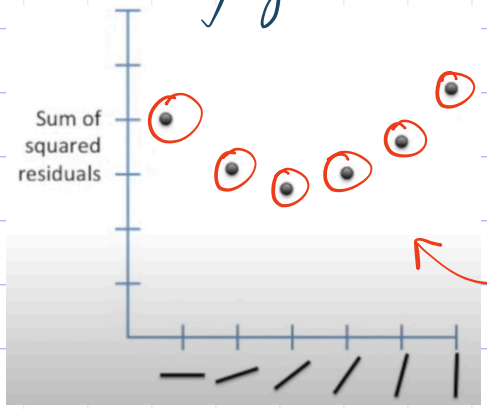
sum of squared residuals $= \sum ((ax_1+b)-y_1)^2+((ax_2+b)-y_2)^2+\cdots$

(ax_1+b) → value of the line at x_1
y_1 → observed value at x_1

Calculating the distance between the line & the observed value

Least Squares: we want the line that will give us the smallest sum of squares, this method of finding the best values of (a) & (b) is called "Least Squares"

If we plotted the sum of squared residuals vs. each rotation, we'd get something just like this:



we take derivative of this function

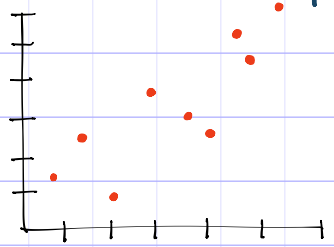
To find the optimal rotation for the line the derivative tells us the slope of the function at every point.

The slope at the best point is zero

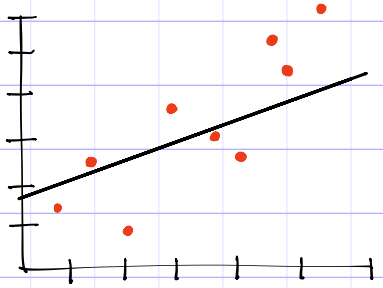
different rotations are just different values of (a) (slope) & (b) the intercept.

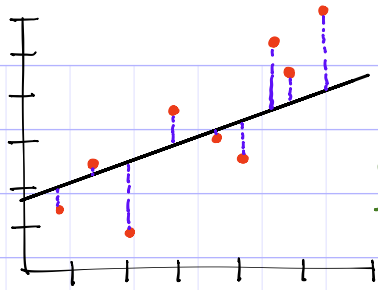
Main Ideas behind Linear Regression

1. use least squares to fit a line to the data
2. Calculate R^2
3. Calculate a p-value for R^2



draw a line through the data





residual
distance from a line to a data point is called a "residual"

measure the distance of the data from the line, square each distance & add them up.

$y = 0.1 + 0.78x$

Since, the slope is not 0 it means that knowing a mouse's weight will help us make a guess about that mouse's size.

Calculating R^2 is the first step in determining how good that guess will be.

1. Calculate the average mouse size
2. Now, sum the squared residuals
3. find SS(mean) "sum of squares around the mean"

$SS(mean) = (data - mean)^2$

$\text{variation around the mean} = \frac{(data - mean)^2}{n}$ ← sample size

$Var(mean) = \frac{SS(mean)}{n}$

↓ average sum of squares per mouse

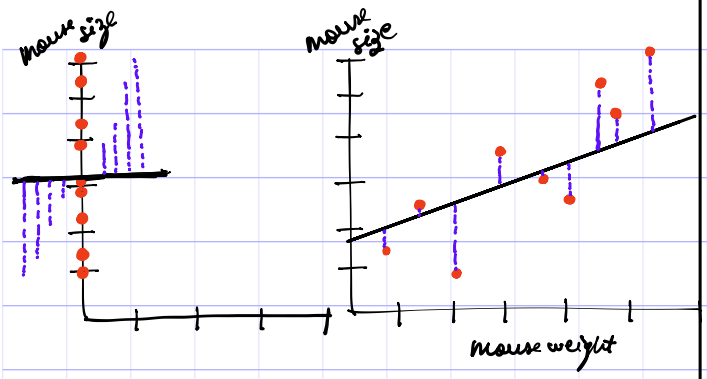
SS(fit) ⇒ sum of squares around the least squares fit

$SS(fit) = (data - line)^2$

$Var(fit) = \frac{(data - line)^2}{n}$

$\text{distance between the data point & the line squared & added} = \frac{SS(fit)}{n}$

Variance (Something) = Sum of squares / the no. of those things



R^2 tells us how much of the variation in mouse size can be explained by taking mouse weight into account.

→ heavier mice are bigger

→ lighter mice are smaller

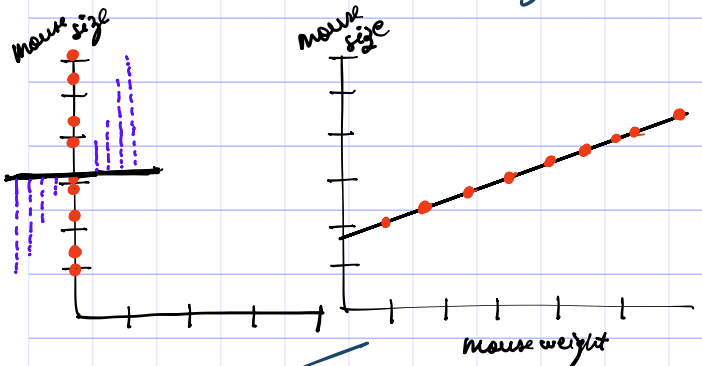
$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6 = 60\%$$

60% reduction in variance when we take the mouse weight into account

OR

mouse weight explains 60% of the variance in mouse size



In this case knowing mouse weight means you can make a perfect prediction of mouse size.

Var(mean) = 11.1 Var(fit) = 0

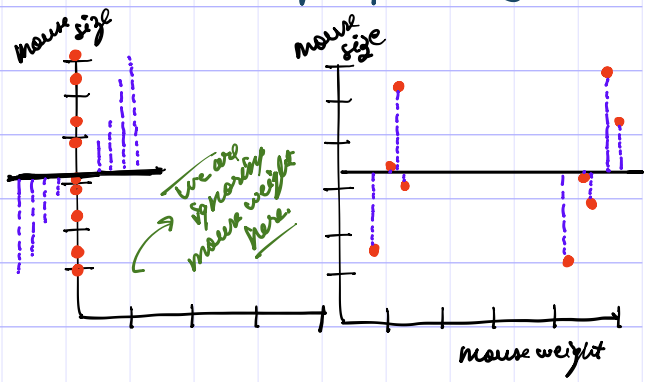
$$R^2 = \frac{\text{V(mean)} - \text{Var(fit)}}{\text{V(mean)}} = 100\%$$

mouse weight "explains" 100% of the variation in mouse size.

Additionally, knowing mouse weight doesn't help us predict mouse size.

→ heavy mouse could be large or small

→ light mouse could be large or small (with equal probability)



Var(mean) = 11.1 Var(fit) = 11.1

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

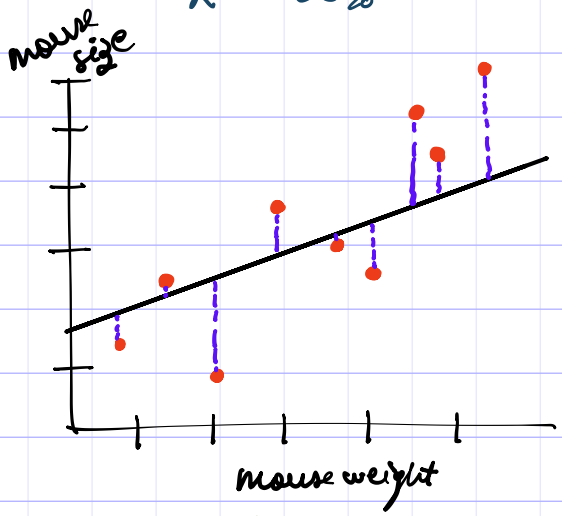
$$\Rightarrow 0\%$$

mouse weight doesn't "explain" any of the variation around the mean.

we applied R^2 to a simple equation of line

$$y = 0.1 + 0.78x$$

$$R^2 = 60\%$$



Meaning # 60% of the variation in mouse size could be explained by mouse weight.

measure, square & sum the distance from the data to the mean

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

measure, square & sum the distance from the data to the complicated equation

more complicated Example

we wanted to know if mouse weight & tail length did a good job predicting the length of the mouse's body.

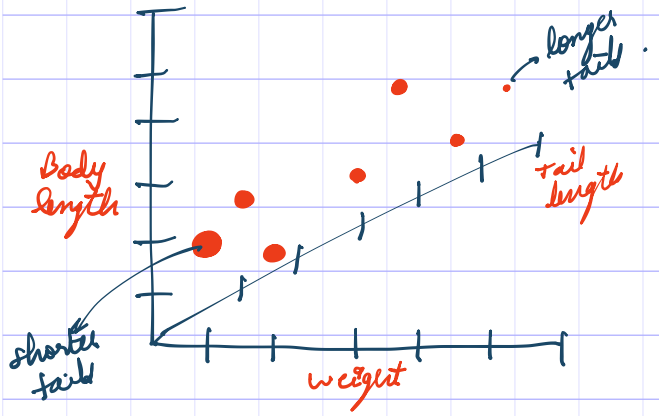
3-D graph

multivariable model

we measure a bunch of mice

weight	tail length	Body length
3.5	2.9	3.1
1.3	2.1	2.8
5.9	4.1	6.1
4.8	3.2	3.8
...	...	...

we want to know, how well weight and tail length ... predict body length

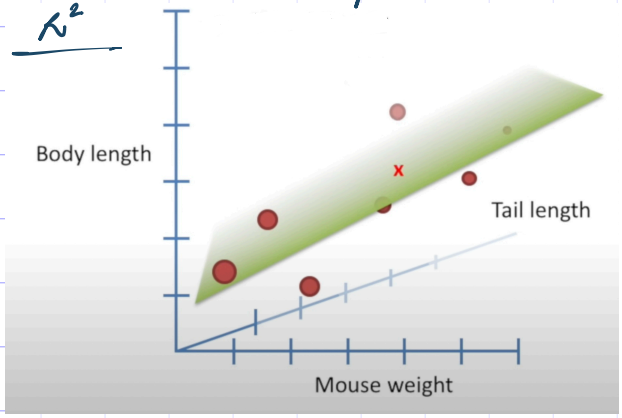


since we have 3-D graph we will fit plane.

$y = 0.1 + 0.7x + 0.5z$

- y intercept (Tail length & mouse weight equal to 0) → body length
- x → weight
- z → Tail length

Just like before, we can measure the residuals, square them & then add them up to calculate R^2



Consider equation → $y = 0.1 + 0.7x + 0z$

→ If the tail length is useless & doesn't make SS(fit) smaller then least-squares will ignore it by making that parameter = 0

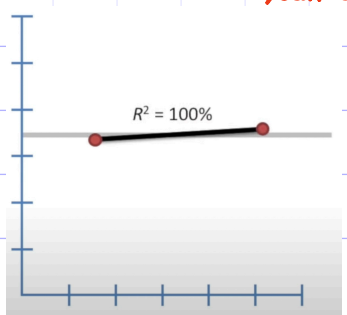
→ plugging the tail length into the equation would have no effect on predicting the mouse size.

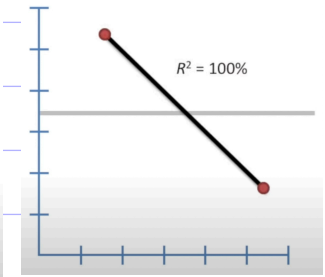
This means, equations with more parameters will never make SS(fit) worse than equations with fewer parameters.

mouse size = 0.3 + mouse weight + flip a coin + fav color + astrological sign.

The more parameters we add to the equation, the more opportunities we have for random events to reduce SS(fit) & result in a better R^2

"adjusted R^2" → scales R^2 by the number of parameters

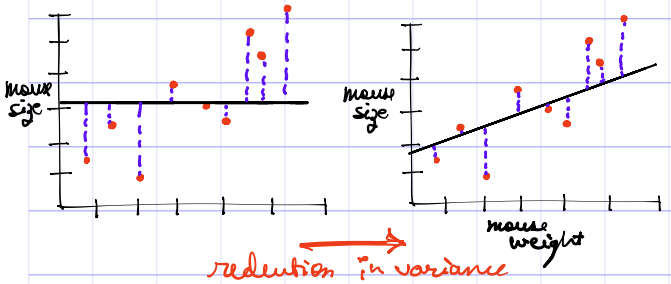




100% is a great number, but any two random points will give us the same thing.

We need a way to determine if the R^2 value is statistically significant.

We need a p-value



reduction in variance

$$R^2 = \frac{Var(mean) - Var(fit)}{Var(mean)}$$

$$= \frac{Var(mouse\ size) - Var(\text{after taking weight into account})}{Var(mouse\ size)}$$

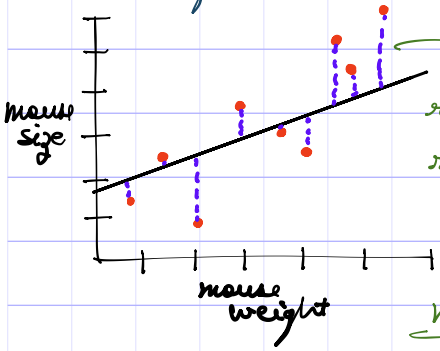
R^2 = Variation in mouse size explained by weight / Variation in mouse size without taking W into account

The p value of R^2 comes from something called "F"

F = Variation in mouse size explained by weight / Variation in mouse size not explained by Weight

Numerators for R^2 & F are same →

Variation in mouse size explained by Weight which is the reduction in variance when we take weight into Account



→ These dotted lines (residuals) represent the variation that remains after fitting the line.

↓ This is the variation not explained by weight

F = Variation in mouse size explained by weight / Variation in mouse size not explained by weight

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$ numerators are same

$$F = \frac{(SS(mean) - SS(fit)) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

or

$$F = R^2 * \text{degree of freedom}$$

$$DOF = \frac{n - p_{fit}}{p_{fit} - p_{mean}}$$

p_{fit} → the no. of parameters in the fit line

fit line → $y = mx + b$.

2 parameters

p_{mean} → the no. of parameters in the mean line

$y = b$ → 1 parameter

$p_{fit} - p_{mean} = 2 - 1 = 1$

$SS(mean) - SS(fit) / (p_{fit} - p_{mean})$

↓ Variance explained by extra parameter

Why divide SS(fit) by $(n - p_{fit})$ instead of just n ??

more parameters you have in equation, the more data you need to estimate them

Ex: you need two parameters to estimate a line, but you need 3 parameters to estimate a plane

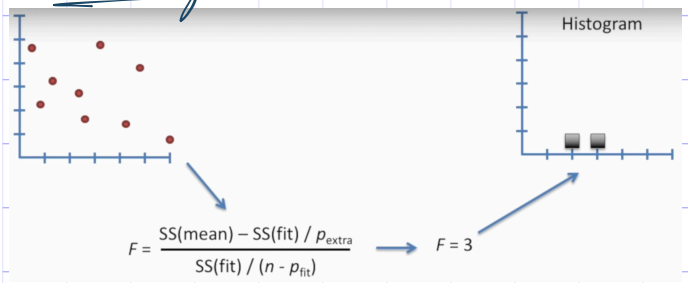
If the fit is good:

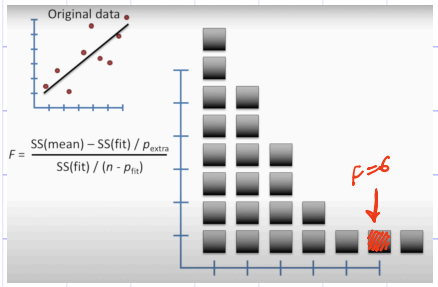
F = variation explained by extra parameters in the "fit" / variation not explained by extra parameters in the "fit"

→ large number / small number

F = really large numbers

find the values of F for different set of random data & create a histogram.





The p-value is
numbes of more extreme values
all the values

Summary (linear Regression)

given some data that you think are related...

1) Quantifies the relationship in the data (this is R^2)
→ this needs to be large

2) Determine how reliable that relationship is (this is the p-value that we calculate with F)
→ this needs to be small

You need both to have an interesting result

linear Regression in R

```
mouse.data <- data.frame(
weight = c(...), size = c(...))
mouse.data
plot(mouse.data$weight, mouse.data$size)
mouse.regression <-
lm(size ~ weight, data = mouse.data)
```

y values — x values

$y = ax + b$ → b: y-intercept; a: slope

size = y-intercept + slope × weight

```
> mouse.regression <- lm(size ~weight, data=mouse.data)
> summary(mouse.regression)

Call:
lm(formula = size ~ weight, data = mouse.data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5482 -0.8037  0.1186  0.6186  1.8852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5813     0.9647   0.603   0.5658
weight        0.7778     0.2334   3.332   0.0126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.19 on 7 degrees of freedom
Multiple R-squared:  0.6133,    Adjusted R-squared:  0.558
F-statistic:  11.1 on 1 and 7 DF,  p-value: 0.01256
```

original call to the lm function

summary of the residuals (close to zero)

weight gives us the reliable estimate of size

→ least squares estimates for the fitted line

size = y-intercept + slope × weight

0.5813 value of the intercept

size = 0.5813 + slope × weight

0.7778 value for the slope

size = 0.5813 + 0.7778 × weight

we want the p-value to be < 0.05 for statistically significant.