Day 1

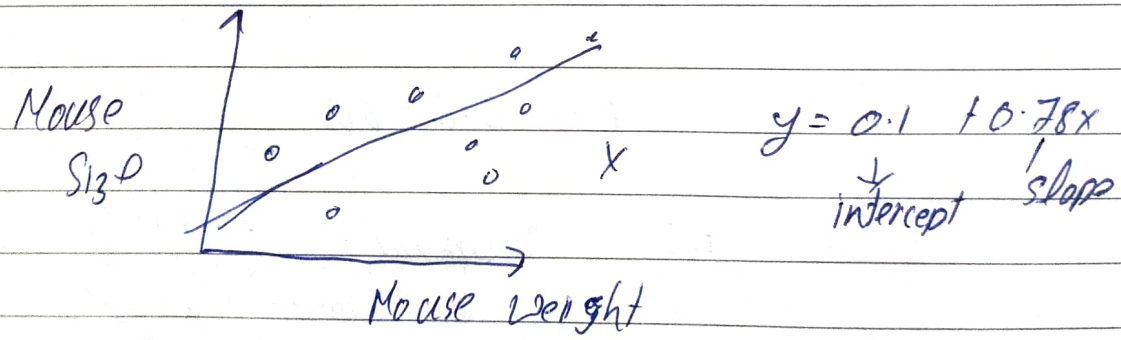
1.1 Linear Models

Linear Regression

the Main Ideas!
1) Use Least-squares to fit a line to the data
2) Calculate R^2
3) Calculate a p-value for R^2

Distance from a line to data is called residual

First draw a line then find the residuals and square it and have the sum



Calculating R^2 gives the probability of getting the guess right or not

To get R^2

1) Calculate avg mouse size
2) Draw a line in avg
3) Sum up the squared residuals
   This is called SS (mean)

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

$$\text{Variation around the mean} = \frac{(data - mean)^2}{n}$$

4) SS(fit) Sum of squares around the least-square fit basically the line

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

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

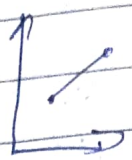
$$= \frac{11.1 - 4.4}{11.1}$$

$$R^2 = 0.6 = 60\%$$

$$R2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

Mouse weight explains 60% of mouse size

When there is only 2 Points



SS (mean) = a value but SS(fit) = 0

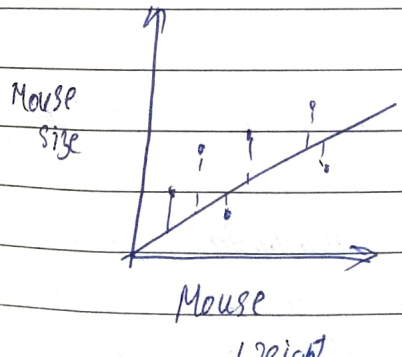
So R^2 = 100%.

which means if you have only two pts the R^2 = 100%

So to ~~get~~ determine the r squared value is statistically significant. We need a p value

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

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



F= The variation in mouse size explained by weight / The variation in mouse size not explained by weight

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

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

P_{fit} is the number of parameters in the fit line

the line is given by $y = y\text{intercept} + \text{slope}$ (2 para)

$P_{fit} = 2$

P_{mean} is the no of parameters in mean line value

$y = y\text{intercept}$

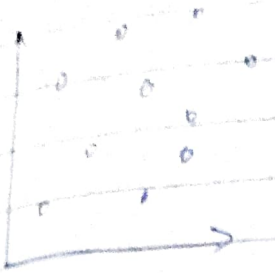
$P_{mean} = 1$

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

Intuitively, the more parameters you have in your equation, the more data you need to estimate them. For example, you only need two pts to estimate a line, but you need 3 points to estimate a plane

if the fit is good

$F = \frac{\text{large no}}{\text{small no}}$ the F is a large value



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

$F = 2$

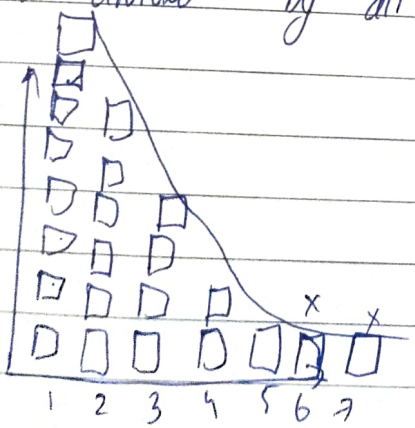
Then generate another data

then you get new value of $F = 3$

The plot all the F's in histogram

the find the F value with the orginal data
lets say $F = 6$

the p value is number of more extreme values divided by all the values



The decrease of distribution

lets take a pvalue with $(n - p_{fit}) = 10$ $(p_{mean} - p_{fit}) = 1$
$(n - p_{fit}) = 6$

Here the 10 gets taper off faster

this means the p-value will be smaller when there are more parameters sample relative to the no of parameter in the fit eqn

Linear regression

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

2) Determine how reliable that relationship is (this is the pvalue that we calculate with F) This need to be small

4th

import pandas as pd

pd.read_csv(path)

Building your Model

Steps

Define : What type of model

Fit : Capture patterns from provided data

Predict

Evaluate

Melbourne_model = DecisionTreeRegressor (random_state=1)

melbourne_model.fit (x, y)