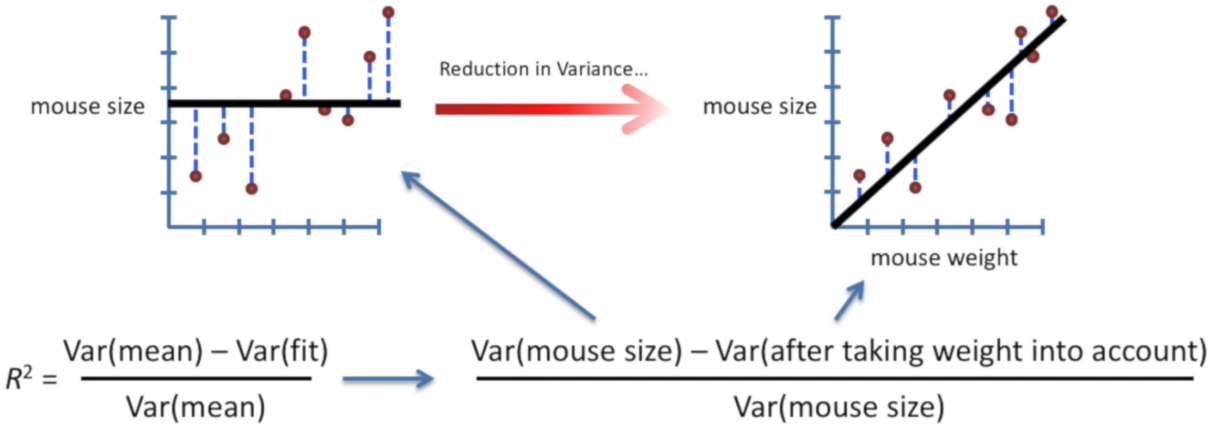
Regression

Linear Regression

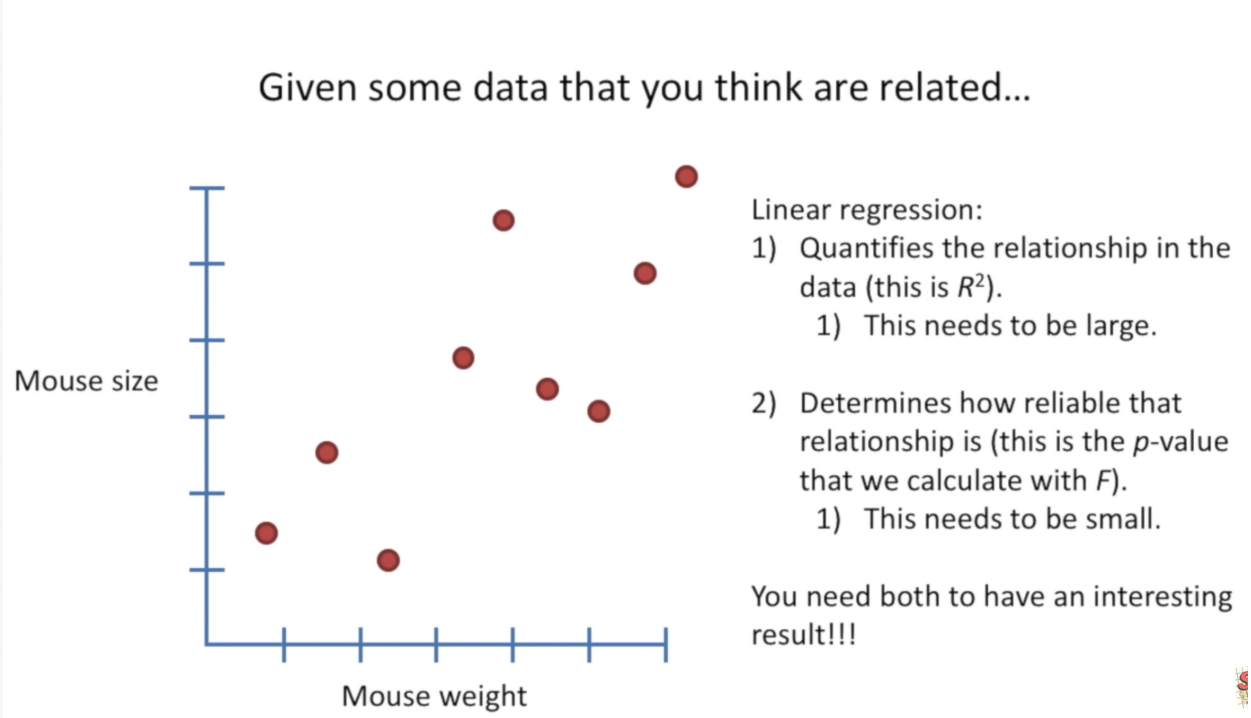
原理是得到所有点到某一条线的距离合最短的线，最终目的是做**prediction**。

这个距离叫residual。

How good is the prediction？ 用R^2. Eg. R^2=0.6 , means mouse weight explains **60% of the variation** in mouse size.



$$R^2 = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size without taking weight into account}}$$



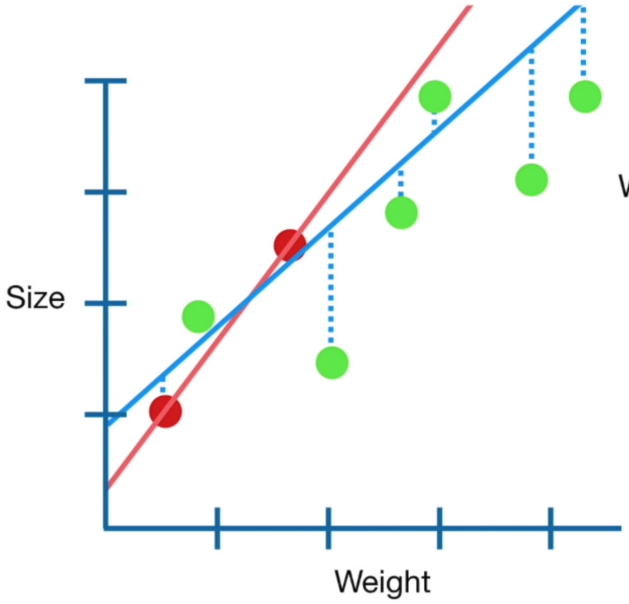
Regularization Regression 1 (Ridge Regression)

With a little bit of penalty, we can have better prediction for long run **by making the predictions less sensitive to the Training Data**

- 最终目的是要min variance
- 用**Ridge Regression** 主要是因为sample size 小的话，容易导致 poor **Least Squares** estimates that result in terrible machine learning predictions.

Ridge Regression 可以用

- liner (continuous variable)
- discrete variable (eg. normal VS high fat)
- logistic regression



When the sample sizes are relatively small, then **Ridge Regression** can improve predictions made from new data (i.e. reduce **Variance**) by making the predictions less sensitive to the **Training Data**.

Regularization Regression 2 (Lasso Regression)

- 跟Ridge Regression很像，区别是：
  - 当variable很多很杂很没用的时候，Lasso可以去掉没用的，让结果更易懂易读
  - 当variable的关联性都很强，很有用的时候，Ridge 的结果更好

Ridge Regression is very similar to… ⟶ **the sum of the squared residuals + λ × the slope²**

…Lasso Regression ⟶ **the sum of the squared residuals + λ × |the slope|**

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

~~+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**~~

But the big difference is that **Lasso Regression** can exclude useless variables from equations.

Regularization Regression (Elastic Net Regression)

Ridge Regression 和Lasso Regression的合体

And by doing so, gets the best of both words, plus it does a better job dealing with correlated parameters.

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X| + \lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

XGboost

(skip for entry student for now)