Session 2 – relationships between variables, regression

Review: random variables? $Y, X$ function/statistical model

$$Y = f(X)$$

"Correlation does not equal causation"

$\boxed{\text{Correlation}}$ : any statistical relationship between random variables

types of correlation: linear, nonlinear

$\boxed{\text{Causal relationship}}$ : relationship between two variables where cause & effect can be established. Mathematically rigorous definition of probabalistic causality is complex: sea Pearl, 1999

Intuition 1: cause proceeds effect

Intuition 2: causes are correlated with effect

Intuition 3: to most accurately evaluate cause & effect we have to experimentally manipulate cause.

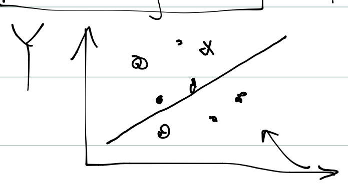
$\boxed{\text{confounding}}$ : A relationship between two variables due to the presence of a third variable.

Examples of cause & effect: from paper/smoking

Discussion: is it possible to obtain data when cause & effects are not correlated even if the underlying causal relationship is <u>REAL</u>?

$\boxed{\text{Linear Model}}$ : when the relationship at stakes can be expressed as $Y = \beta X$ $\beta$ is called "regression coefficient"

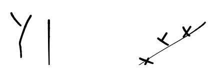
$\boxed{\text{Linear regression}}$ : fits a line to a bunch of numbers. eg.



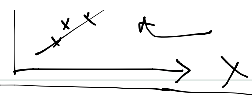

large $R^2$

"$R^2$" is a measure of how scattered the points are.

also known as % variance explained.




small $R^2$



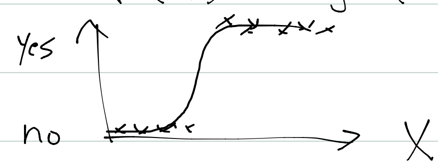

**multiple linear regression** : fits a line to multiple variables

i.e. $Y = \beta_1 X_1 + \beta_2 X_2 + \ldots$ etc this is also called "adjusting"

Demo 1 : using R-statistical package to do regression.

**logistic regression** : regression involving a binary variable

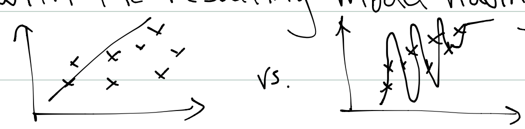
odds of $(Y) = \text{logit}(\beta X) = \frac{1}{1+\exp(-\beta x)}$




p-values can be obtained for the rejecting hypothesis $\beta = 0$

can also do logistic regression on multiple variables.

**"overfitting"**: picking a complex model to reduce error, but with the resulting model having little validity. Example:




$R^2$ increases as # of variables ↑

**Occam's razor:** simpler model is preferred.

Demo #2: adding terms to regression models.

Multiple regression addresses the problems of confounding to a certain extent, but not fully.

Final take home message: Complex, multivariate relationships often require large datasets & complex fitting procedures (i.e. Machine learning)

Exercise: in class demo of how regression works.

How to control for confounding?

(1) Stratification ⟶ divide the original data into subgroups ⟹ analyze

(2) regression based $\longrightarrow$ $Y = \beta X + \beta' X'$ causal confounding

if $\beta$ is still significant $\longrightarrow$ $\beta$ is significant "adjusting for $X'$"

(3) more complex strategies : matching, etc.