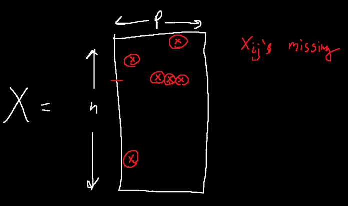
Missingness: when entries of X are not present



If there's missing in X, you can't actually use any of the Algorithms we discussed in this entire class to build a predictive model g. So we need to deal with this very common situation somehow.

The statistical literature talks about three types of "missing data mechanisms" (MDMs) in an effort to model why entries go missing:
(1) Missing Completely at Random (MCAR)
(2) Missing at Random (MAR)
(3) Not Missing at Random (NMAR)
Let M_j denote the Bernoulli rv that the jth feature went missing.

MDM	$P(M_j \mid X_{j,miss}, X_{-j,miss}, X_{-j,obs}, U, \gamma)$
MCAR	$= P(M_j \mid \gamma)$ e.g. data corruption
MAR	$= P(M_j \mid X_{-j,miss}, X_{-j}, \gamma)$ e.g. old person taking a survey
NMAR	does not simplify

$X_{j,miss}$: This is the value of X_j which we don't see

$X_{-j,obs}$: These are the values of the other features (not j) we see

$X_{-j,miss}$: These are the values of the other features (not j) we don't see because their values are missing as well.

U : Features not in the matrix X (unobserved features)

γ : Other constants that are indpendent of all X's and U's.

Two strategies for dealing with missingness

(I) Delete all observations where there's missingness (called "listwise deletion"). This is okay if the # of missing observations << n. Then it's just trivial. But if it's nontrival fraction of n then there are two downsides:
  a) estimation error increases => predictive performance suffers
     If data is MCAR, this is the only downside
  b) the model won't generalize as well as there is "selection bias" in the training data (e.g. only young people). Thus future predictions will be more likely to be extrapolations and thus predictive performance sufferes even more. This occurs in the MAR/NMAR cases.

(II) Imputation: guess / predict the missing values. Of course there are a ton of ways to do this (prediction is a big field). E.g.
  a) Just use xbar_j for all xj missing if metric is continuous or use Mode[xj] if metric is categorical.
     This strategy is not great. It's the g_0 of predictions and you can induce the same selection bias.
  b) You can build an entire predictive model and treat it like every other problem we've seen in the whole class.

Recommendation 1: use IIb via the algorithm "MissForest". MissForest runs a RF model to predict each xj with missingness by using the other features. It does this iteratively until convergence.

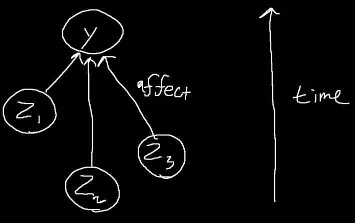
$$y = f(x_1, \ldots, x_p) + \delta$$

$$y = f(x_1, \ldots, x_p, m_1, \ldots, m_p) + \delta_{less}$$

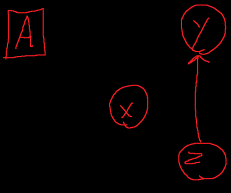
If we include binary features that record if original went missing, that is part of the signal and it captures some of the ignorance.

Recommendation 2: use missingness dummies as new features in your training data. For categorical features, sometimes you don't impute, but you add missing as another level.

Basic Causality. Recall y = t(z_1, ..., z_t)

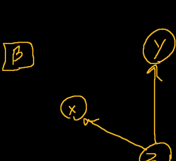


Where are the x's in this picture? They can be anywhere e.g.



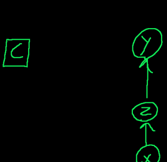
x does not cause y
x does not correlate to y

Bad model! n doesn't matter



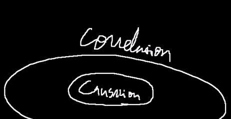
x does not cause y
x correlates to y

I could get a good model here depending on the relationship of x and z



x causes y
x correlates to y

I could get a good model here too.



(1) Correlation does not imply causation but... sometimes it does.
(2) You cannot be causal without being correlated.

In situation A, could you observe in a dataset of size n a high correlation? Sure... there can be high correlations just be chance. These are called "spurious correlations".

If a correlation is non-spurious, that means there is causation somewhere. If the correlation is non-existent, it may mean it's too weak to detect.

A good but fuzzy definition of "x causes y" is that if the value of x is manipulated then the value of y changes.