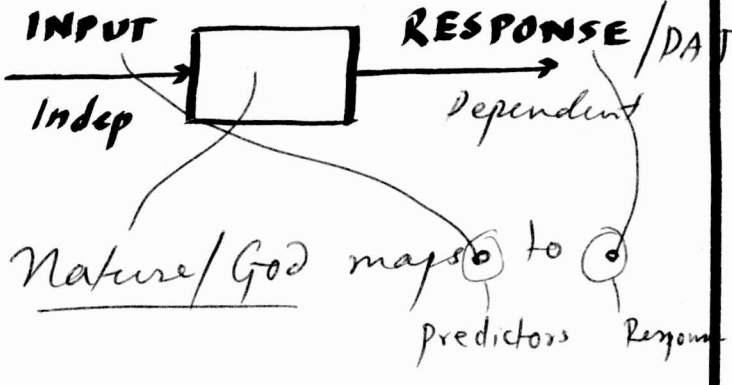
1) Machine-learning (or Data Science) = OPTIMIZATION + STATISTICS

→ In the past, you needed to infer before you predicted. No longer the case!

This course is about DATA MODELING { Generative, Predictive

STATISTICS



There are two cultures of modeling:

① INFERENCE ("Generative") (stats / Biostats)

Look inside the box: HOW is (does the) Nature is associating X → Y?

② PREDICTION ("Predictive"): (ML, CS)

Predict future responses.

Common Task Framework (CTF)

(Predictive Modeling)

① Public Dataset! This is training data

② Competitors: Infer a class prediction rule from data

③ Scoring Referee: Runs the prediction rule against test dataset and scores.

Eg: "The Netflix challenge"

This is pretty commonsensical...

Course Layout

① Learning
- Supervised
  - SVM
  - CV
  - Neural Nets
  - kNN
  - Kernel tech.
- Unsupervised
  - k-means
  - EM

② Linear
- Regression
  - Ridge
  - LASSO
  - PCR
  - PLS
- Classification
  - LDA
  - QDA
  - Naive Bayes
  - Logistic Reg

③ Local Smoothers + Additive Models
- LOESS
- Kernel smoothing
- Spline smoothing
- Backfitting

④ Tree-Based Methods

⑤ Ensemble Methods
- Tree
- Bayesian Model Avg.
- Stacking
- Bagging
- Random forest
- Boosting.

- If model fits perfectly, p-value = 1, "COMPLETE SEPARATION"
- Goodness of Fit $\not\Leftrightarrow$ Predictive Power (Assumptions)
- "Obs. Studies rarely imply causality..."

Est.

- same formula
- Avg Est mean resp for all settings of x^*
- CONF INT
- Uncertainty — β_0, β_1, β_i ...
- Var has same

Pred

- same formula
- Avg Response for one setting of x^*
- PREDINT (WIDER!)
- Uncertain — β_0, β_1, ... AND (new obs!)
- var. is more/diff! new value!

$S_{xx} = \Sigma(x_i - \bar{x})(x_i - \bar{x}) = \Sigma(x_i - \bar{x})^2$

$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$

RESIDUAL STDERR^2 = MSE = Variance

STDDEV = Residual Std Error.

LOGIT / POISSON: DO NOT (TRY TO) Reject H_0 of Good fit!!

DUMMY 2 vars (categorical)
3 values,

With intercept: 2(3-1)

More predictors fit to irreducible Error, yo.

~~Adj~~ Adj R^2 can be negative!

Vars:
- Controlling / "Default"
  - adj for BIAS
- Explanatory (Explan-/est-resp)
- Predictive (Boost pred. power Even if no Explan power) ☺

Scatterplot
- LIN
- CONSVAR
- INDEP

MEAN ZERO
Res Analysis
- CONST VAR
- IND

Res vs Fitted

Normality < QQplot, Hist

- Predicting Forecasting
- Modeling
- Hyp testing

ANOVA

				Mean	F
Bet	tr	$k-1$	SSR	$MSR = \frac{SSR}{k-1}$	bet/within
Within	err	$N-k$	SSE	$MSE = \frac{SSE}{N-k}$	
		$N-1$			

Cooks if Outlier $CD > 1$, $CD > \frac{4}{N}$

Cannot throw away! could be heavy tailed.

$VIF_{x_i} < Max(10, \frac{1}{1-R_i^2})$

Prediction ② uncertainty
- new obs
- param estimate.

How much param values change if X_i is removed

Repair NORM, CONSVAR } transform Y w/ Box Cox
Linear } transform X

$\lambda = 1$	NOPE
$\lambda = 1/2$	$\sqrt{y}$
$\lambda = 0$	$\ln(y)$
$\lambda = -1$	$1/y$

Box-Cox

CHECK FOR UNCORRELATED ERRORS, not ind.

Multiple Reg:
- Stand Resid. vs Var — LINEARITY
- Stand. Res vs Fitted — CV, IND
- QQ plot, Hist — NORM

DONT Check Pred. Var for normality!
→ USING ~~RESIDUALS~~ NOT RESPONSE RESIDUALS.

ANOVA → Hyp test $\beta_0 = \beta_1 = \cdots \beta_n = 0$ } That these all don't do shit

P

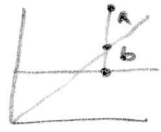
$\frac{N-P-1}{N-1}$

Res. St. Err is STDDEV ⓐ → Square for variance

SIMPLE REG: dofFree = $N-2$ ($\hat{\beta}_0, \hat{\beta}_1 X$)

$N-P$ → things you see in output! ($x_1, x_2, \ldots, x_P$)

$N-(P+1)$



$R^2 = \frac{SSR}{SST}$

$a = SSE$

$b = SSR + SST =$

MAE: Measured unit or competing models $\frac{1}{N}\sum abs(Y - \hat{Y})$

MSE: Like this but less robust to outliers $\frac{1}{N}\sum (\hat{Y} - \hat{Y})^2$

MAPE: $\frac{100}{N} * \sum abs\left(\frac{Y - \hat{Y}}{Y}\right)$

Prediction Interval $P = \frac{\text{Pred. err}}{\text{Pred Value}} = \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$

91% 55%

$\frac{50+45}{55+45}$

30

45

$\frac{6^2}{A} = 14$

$MSE \times (N-K) = SSE$

$(10.13^2)(30-4)$

$0.6518 = 1 - \frac{SSE}{SST}$

$\frac{SSE}{SST} = 1 - 0.6518 = 0.3482$

AB CD
AB BC CD
AC BD
AD

ABCD

Catherine
Catherine

Cloud < lots
nothing sp.

(connection ok?)
A bit
weak here...
sounds like
monolith!

- 118n — ?
- Module federation?
- Resolve conflicts?

(React ~~Libr~~ Libraries...

React Query?

design system
Component library
CLI

State management/
Redux

Maturity
Adventurous

✓ Testing strategies
{ Snapshot tests?
unit tests

$\frac{50}{56} + \frac{25}{\frac{x}{45}} = 0.9$

$\frac{5}{9}$

- Versioning
  └ Semantic versioning

$\frac{25}{45} \times 100$

SPA → K8s?

- AWS? → not too strong here....

Catherine
Catherine

Se
Sexton

Handsome
Amste Tate

AWS ok
but
MF?

"Regression" ⇒ the mean of a resp. variable : $\mu\{Y|X\}$
given explanatory variables

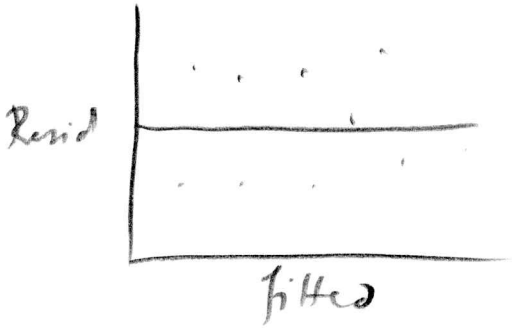
$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation in response}}$

Does not explain Lin!
Must use residual plots!

- Residual = Observed (y) − Predicted ($\hat{y}$)

$e = y - \hat{y}$ $\quad \Sigma e = \mu(e) = 0$
sum mean

- "Normal" ⇒ symmetric



"Linear" Model when

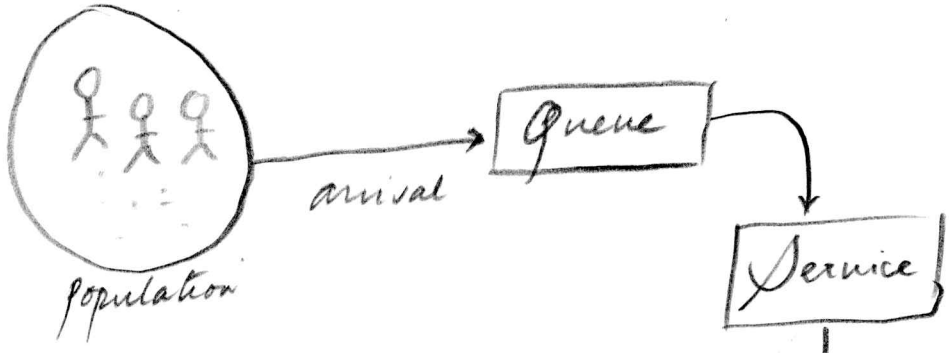
$Y = b_0 + b_1 X_1 + b_2 X_2 \ldots$

Y = Sum of (Either constant OR parameter × predictor var)

① Random around 0 is good

② Some non-linear variance not explained
Homos and biased

③ "Fanned"
⇒ Hetero's! variance increases
↓
Unequal variances



SVM: Scale data!
Some coeffs are near zero and irrelevant.

Standardizing: Scale to mean zero and stddev 1

$\frac{x - \mu}{\sigma}$ $\frac{x_{ij} - \mu_j}{\sigma_j}$

Scaling

Outlier detection

imputation. Good God

Splitting: All about ~~set~~ ~~Fixed~~ Random Effect

→ High-perf. model might get boost from Random Effects.
→ above-average Random Effects.

Train → Valid
Train → Test
Train → Validate
} !! how to rotate!

not ~~in~~ loop across set!

Exp Smoothing Ts for baseline

Change detection

All obs are considered!
look @ expanded form

Supervised: SVM, k-means.

DOE: Not much data avail.

Unsuper: kNN

Classif: SVM, k-means, Regression

Change Detection: CUSUM

Time-Series: Exp. Smooth, ARIMA, Garch.

Model quality

Optimization

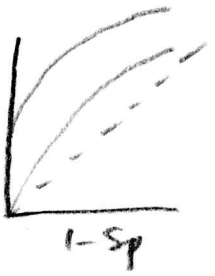
Simulation

Detrending
↳ use for factor based (SVM, Reg)
↳ Reg factor by factor and remove from resp!

Auto regression only works w/ time-series data.

Regression: DESC AND PRED
think about coefficients.
Not PRESC.

Data
- Types
  - Unstruct
  - Struct
    - Numeric
      - Time
      - indepnd
    - Categorical — Binary
- Prep.
  - outliers
  - scaling
  - Matching
  - Heteroscedastic?
    Check w/ Q-Q and Box-Cox the MoFo
  - Detrend
    for use w/ regression, SVM



Classification Error Rate
in <u>bian2 dowawan2</u>.

H_0: $\beta_1, \ldots \beta_p$ are zero

$F(p, n-p-1)$ fstat.

$n-k$, $k-1$ } ANOVA

p, $n-p-1$ } Mult. Reg.

} F stat Stuff.

$Adj\ R^2 = 1 - (n-1)\left[\frac{1-R^2}{n-p-1}\right]$

See 2nd stats in anova, t-test...

Why t- for Mult Reg?
and 3 for Logi!!

ASSUMPTIONS
CANOVA: no kurwa's!

Mein

MLR

$$t\text{-stat} = \frac{\hat{\beta}_1 - \beta^*}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

↑ tstat
↓ p-value } reject H_0

Poisson: null dev: $n-1$ dof
Residual dev: $n-p-1$ dof } Same for Logistic!

200 null = 1
Res = 200 − p − 1

↓
Eg: group one
group two } Count these!!

p-value = 1 − pchisq(dev(model), n−p−1)