Statistical Inference

1. Make a good guess of the parameter
2. Quantify our uncertainty

"What can be learned from the data"

Bayesian View

"Data is given, parameter is random"

"θ drawn from prior, Y drawn from likelihood $f(Y|\theta)$"

- prior with marginal likelihood to get posterior

$$P(\theta=\theta_0 \mid Y=y) = \frac{P(Y=y \mid \theta=\theta_0) \cdot P(\theta=\theta_0)}{P(Y=y)}$$

(likelihood; prior; $P(Y=y)$ ← marginal over all θ)

Point Estimates: Mean, Median, Mode (MAP) of Posterior

MAP = MLE if prior is flat

Interval Estimates: Credible Intervals

Uncertainty

A $1-\alpha$ credible $[L, U]$

$P(\theta$ in between L, U given data$)$

(θ: random; L, U: not random)

"Says something about parameter itself"

Standard Bayes: known from specified prior

Empirical Bayes: parametrize the prior and maximize marginal likelihood $f(Y;K) = \int f(Y|\theta) f(\theta;K) d\theta$ (pushes MLE to hyperparameters K)

* For normal: the smaller the b (varriance), flatter the prior
the closer MAP = MLE and posterior $\sim N(\theta^*, se(MLE))$
OR as $n \to \infty$, posterior shifts towards likelihood / MLE
* flat is not uniformative, because can use Credible Intervals of MLE

Frequentist View: data comes from some true distribution

Maximum Likelihood

"Estimating parametric models"

$l_n(\theta) = \prod f(Y_i, \theta)$

$P(\text{data} \mid \text{theta})$

$MLE = \text{argmax}\, l_n(\theta) = \hat{\theta}_n$

Wald Test "Approximate Normal"

$H_0: \theta^* = \theta_0$, $\hat{\theta}_n \sim N(\theta^*, \hat{se}(\hat{\theta}_n))$

$$\frac{\hat{\theta}_n - \theta_0}{\hat{se}(\hat{\theta}_n)} \to N(0,1)$$

(estimated param; "hypothetical")

Interval Estimates

$\hat{\theta}_n \pm cv \cdot \hat{se}(\hat{\theta}_n)$

"Don't know null distribution"? → Bootstrap

① Simulate $\tilde{\theta}_n$ drawn from sample $\hat{F}_n$ rather than F

② Estimate $\hat{se}(\hat{\theta}_n)$

③ Construct CI

- Wald $\pm 1.96 \hat{se}(\hat{\theta}_n)$
- Pivotal $\alpha/2, 1-\alpha/2$ quantiles of $2(\hat{\theta}_n - \hat{\theta}_n)$
- Quantile: $\alpha/2, 1-\alpha/2$ of $\tilde{\theta}_n$

Estimation: learn some statistic

$Bias_\theta(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$

$Var(\hat{\theta}_n) = E[\hat{\theta}_n^2] - E[\hat{\theta}_n]^2$

$MSE(\hat{\theta}_n) = Bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$

Uncertainty Inference

- Student t
- pvalue: probability of given result or more significant given null is true
- CI: over 95% of intervals will contain true para over new draws of data

Predictive Modeling "Understand relationships between variables"

① Finding functions f s.t. $Y_i \approx f(X_i)$, want $f(x) = E[Y|X=x]$ minimizing $E[Y - f(x)]^2$ SSR

② Let be F a blackbox, focus on predictive inference (avg value of Y among $X=x$)

The Linear Model: $Y = \beta^T X + \varepsilon$ $E[\varepsilon]=0$, becomes parametric when $\varepsilon \sim N(0,\sigma^2)$

unbiased for X with linearly independent columns → $\hat{\beta}_n = (X^TX)^{-1}X^TY$ called Ordinary Least Squares closed form solution to minimze MSE $\|Y-X\beta\|^2$

Residuals $\hat{\varepsilon}_i = \hat{\varepsilon}_i = Y - X\cdot\hat{\beta}_n$, $\hat{\sigma}^2_n = \frac{1}{n(-p)}\|E\|^2$ ← unbiased (Biased), $\hat{Cov}(\hat{\beta}_n) = \hat{\sigma}_n (X^TX)^{-1}$, $se(\hat{\beta}_{n,i}) = \hat{\sigma}_n \sqrt{(X^TX)^{-1}_{ii}}$

Generalized Linear Models

for binary can use LR

$g(z) = \frac{\exp(z)}{1+\exp(z)}$

But use Wald Test Instead of Student T-test for Inference

Predictive Inference (hypothesis of β)

Can Use MLE: Wald $\frac{\hat{\beta}_{n,i} - \beta_i^*}{\hat{se}(\hat{\beta}_{n,i})} \to N(0,1)$, $H_0: \beta_i^* = 0$?

Can Use Student t_i with $n-p$ degrees freedom, "More conservative, larger intervals" (Prefferred)

If ε_i are normal, Student t is exact parametric test

$t_n = \frac{\hat{\mu}_n - \mu_0}{\frac{\hat{\sigma}_n}{\sqrt{n}}}$ $\hat{se}(\hat{\mu}_n) = \frac{\hat{\sigma}_n}{\sqrt{n}}$

Predictive CI's on Outcome

$E[Y|X=x] = \hat{\beta}_n^T x \pm 1.96\sigma \cdot \sqrt{x^T(X^TX)^{-1}x}$

95% CI for Y accounts for conditional var σ

Dangers of Linear Models:

Collinearity: two copies of same variable, can't interpret Beta's

High Dimension: If data $n <$ parameters p will overfit

Heteroskedastic Noise: Violates the assumption $\varepsilon_i \sim N(0,\sigma^2)$ and $se(\hat{\beta}_{ni})$ is wrong

We think significance but there is none because se is too small.

1/20 coefficients will not contain true — Multiple Testing: OLS confidence intervals are not joint CI's and need to use Bonferonni correction, cant consider statistical set simultaneously

Improving the Linear Model (Baseline OLS)

Cross Validation

More folds, lower bias but higher variance

Less folds: higher bias, lower variance

- obtain unbiased estimates of MSE

Prediction error: $\ell^*(\hat{f}) = E[(Y - \hat{f}(x)^2)]$ average squared loss MSE

$R^2 = 1 - \frac{\ell^*(\hat{f})}{Var(Y)}$ percentage of Y variance reduced by x features

For homoskedastic noise $Var(Y|x) = \sigma^2$ ← is constant

so $R^2(\hat{f}^*) = 1 - \frac{\sigma^2}{Var(Y)}$ minimum error we can aspire

Try to improve R^2 in-sample and out of sample simultaneously

Stepwise Regression on AIC → (With Cross Validation)

AIC = maximum log likelihood of model − # of parameters

Bias-Variance Tradeoff: $E[(Y - \hat{f}(x))^2 | X = x]$ ← expectation of prediction error

Methods

Regularization → ~~high bias if too~~ variance

Ridge + Lasso = Elastic Net

Shrinkage

- simple model w/ least variance
- shrink coeff towards zero

Selection + shrinkage

- absolute sum of coeff penalty term $\|B\|_1$
- makes terms 0 · cancaux a bias

→ Bias is reduced when feature space is rich but variance is highly change on new data

→ If Bias is high, model is too simple and variance is low on new data

→ Error is minimized where model complexity balances error

$\hat{\beta}_n^{\lambda ridge} \in \operatorname{argmin} \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \cdot \|\beta\|_2^2$

strength of regularization, high bias for low variance

penalty term = square of coefficient

Prediction Error = Variance σ^2 under homoskedasticity + bias² estimating $\hat{f}^*(x)$ + $Var(\hat{f}(x))$ variance

uncontrolled error irreducible

Causal Inference: causality ≠ association = prediction

Assignment at random? Yes " Average Treatment effect ATE = E[Y(1) - Y(0)]

Assignment not at random

Parametric | Nonparametric | Method 1 | Method 2

A/B testing

- $Y(0), Y(1) \perp\!\!\!\perp T$ (↑ independence)
- Calculate ATE by two sample means
- Any observed difference causal effect or violation of non-interference consistency assumptions
- Use two sample t-test if equal variance
- Randomization ensures no alternate explanations for Correlation except causation (idiosyncrasies of each unit
- Can use Welch's for unequal var

Process of AB Test

1. Split n units randomly into two groups
2. Apply control, treatment
3. Measure outcomes
4. Use t/welch to compare difference

Permutation Test

- Sharp null $H_0: Y(0) = Y(1), TE = 0$
- Non parametric assumption when n is small and heavy tail is significant
- Big $\hat{\Delta}_n$, H_0 is false and small $\hat{\Delta}_n$ $H_0 = 0$, is plausible
- Simulate Δn by random permutations of assignment to treatment
- Get a distribution over $\hat{\Delta}n$ and compare $|\hat{\Delta}n|$ to overall distribution
- Can perform for any statistic medians, KS etc

Process of Permutation

1. Calculate difference delta
2. Simulate $\hat{\Delta}$ over many permutations / random shuffles
3. Calculate p value and show distribution

Ignorability

- Controlling for all alternative explanations of non-causal difference

$Y(t) \perp\!\!\!\perp T|X$

(assignment at random within each $X = x$ subpopulation)

$E[Y(t)|X] = E[Y|T=t, X]$

and ATE =

$E\left[E[Y|T=1,X] - E[Y|T=0,X]\right]$

using OLS

Process

1. Filter data
2. Run OLS
3. Treatment coeff is causal effect

Instrumental Variables

- Add a variable to data to simulate randomness
- Must be ignorable, only affects outcome via treatment and effects treatment

Ignorable - does not share common effects with outcome Y. no confounding factors.

Relevance - instrument Z has effect on Treatment

Exclusion - Z affects outcome Y only through X

$$\tau = \frac{E[Y|Z=1] - E[Y|Z=0]}{P(T=1|Z=1) - P(T=1|Z=0)}$$

If weak → we get large effect

↑ = Intention to treat causal effect

↑ Compliance w/ assigned treatment

causal effect had everyone complied

Two Stage Least Squares

- Empirically verifies the relevance assumption (R^2 determines strength)
1. Predict expected value of treatment based on IV
2. Use new treatment on outcome to remove individual idiosyncracies

Heterogeneous Effects LATE ≠ ATE

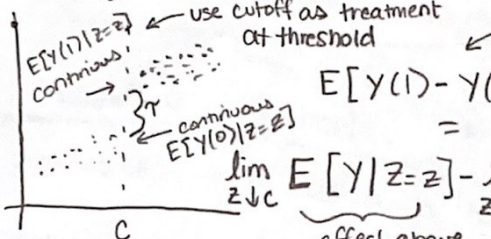
IV only works for compliers

$LATE = E[Y(1) - Y(0) | \text{Complier}]$

• "Closer to gold for causal effects"

Causal Inference Method: Regression Discontinuity Designs

• Analyze effects at natural discontinuity, as if random assignment



Avg Treat Effect at threshold

$$E[Y(1)-Y(0)|Z=c] = \lim_{z \downarrow c} E[Y|Z=z] - \lim_{z \uparrow c} E[Y|Z=z]$$

effect above − effect below

• the size of discontinuity is causal LATE

Parametric in the subpopulation near threshold.

• To estimate use OLS and $I[Z \geq c]$ (treatment by sharp threshold) as variable read off coefficient

• Nonparametrically use KNN sample averages for nearest K-neighbors

• Can extrapolate near cutoff by adding covariates to help reduce error in local extrapolation

---

What if treatment assignment not sharp at c?

Fuzzy RDD = RDD + IV ← nudge towards/against treatment where $I[Z \geq c]$ at cutoff

Step 1: Run RDD to find effect of cutoff on treatment

$\tau_1 = Ex.\ (\ I[\text{classize} \geq 30] \sim I(\text{enrollment} > 40)$

Treatment; RDD IV; $I[Z \geq c]$ Cutoff

Step 2: Run RDD to find effect on outcome

$\tau_2 = Ex.\ E[\text{Mathscore} \mid \text{enrollment} = 40]$ ← $I[Z \geq c]$ Cutoff IV @

Step 3 Get effect of treatment on mathscore

$\tau = \frac{\tau_2}{\tau_1}$ super local avg treatment effect of classize on enrollment

---

Causal Inference Online Decision Making

"Maximize reward of A/B test balancing exploration vs exploitation"

optimal arm → $t^* = \text{argmax}\ E[Y(t)]$ ← reward of arm

$$\text{Regret}(N) = \sum_{n=1}^{N\text{-epochs}} \mu(t^*) - \mu(T_n)$$

$\mu(t^*)$: optimal reward; $\mu(T_n)$: our choice reward

Regret(N) should be sublinear in N, $\frac{\text{Regret}(N)}{N} \to 0$

Multi Armed Bandit Strategies

Goal: $E\,\text{Regret}(N) \leq c' \log(N)$

Greedy: See all arms, pick highest reward for N epochs

ε-Greedy: See all arms, wp 1-ε be greedy, ε pull random

* $\text{Regret}(N) \geq cN$, $\text{Regret}(N) \geq \delta \frac{\varepsilon N(m-1)}{m}$ ← c

K, ε-Greedy: Experiment for K, then be greedy.

$$\text{Regret}(N) \geq \delta \frac{K(m-1)}{m} + \underbrace{\delta p(N-K)}_{\text{constant wp} \geq 0}$$

Fix, N, K Greedy: tune K to get regret N

$$E[\text{Regret}(N)] \leq \Delta \frac{2m}{} \log(Nm) + \Delta$$

• don't know δ or δ^2 so suboptimal

UCB: Create CI for each arm, pick arm that can potentially do best, update CI to not select in future.

Pull arm t w/ max $\hat{\mu}_n(t) + \alpha \sqrt{2\log(n)}\, se(\hat{\mu}(t))$

For binary $\hat{\mu}_n(t) + \alpha \sqrt{\frac{\log(n)}{2n(t)}}$ ← epochs; ← boost if not pulled frequ

$E(\text{Regret}) < c \log(N)$

Thompson Sampling: draw reward from posterior and pretend like true reward

$E(\text{Regret}) \leq c \log(N)$

* posterior auto adjust best guess for reward

Contextual Bandit: Observe context X_n before T_n

$t^*(X) = \text{argmax}\ \mu(t,x)$ $t = 1 \cdots m$

$\mu(t,x) = E[Y(t) \mid X = x]$ ← reward; ← contextualized

X needs to be iid, $\mu(t,x)$ is linear