Explicit Attributes

Examples

Counts:

$$\sum_{u \in P} I_A(y_u) \quad , \quad I_A = \begin{cases} 1, & \ldots \\ 0, & \ldots \end{cases}$$

Mean attribute:

$$\frac{\sum_{u \in P} y_i}{N}$$

Proportion:

$$a(P) = \frac{1}{N} \sum_{u \in P} I_A(y_u)$$

$\Big\}$ Location attr.

Variance:

$$a(P) = \frac{1}{N} \sum_{u \in P} (y_u - \bar{y})^2$$

Standard dev.

$$a(P) = \sqrt{\text{variance}}$$

$\Big\}$ Spread attr.

Order stat: $y_{(i)} \Leftarrow$ i^{th} smallest variable in pop.

Min: $y_{(1)}$, Max: $y_{(N)}$

Midrange:

$$a(P) = \frac{1}{2} \left[y_{(1)} + y_{(N)} \right]$$

Median:

$$a(P) = \begin{cases} y_{\left(\frac{n+1}{2}\right)} , & n \text{ is odd} \\ \dfrac{y_{(n/2)} + y_{(n/2+1)}}{2} , & n \text{ is even} \end{cases}$$

$\Big\}$ Order location attr.

IQR: $Q_3 - Q_1$

MAD: $a(P) = \underset{u \in P}{\text{median}} \left| y_u - \underset{u \in P}{\text{median}}\, y_u \right|$

$\Big\}$ Order spread attr.

Invariance & Equivariance

Location invariant: $a(\ldots y_i + b \ldots) = a(P)$

" equivariant: $a(\ldots y_i + b \ldots) = a(P) + b$

Scale invariant: $a(\dots m y_i + b \dots) = a(P)$

" equivariant: $a(\dots m y_i + b \dots) = m\,a(P)$

Location-scale invariant: both location & scale invariant

" equivariant: " equivariant

Replication invariant: $a(P^k) = a(P)$

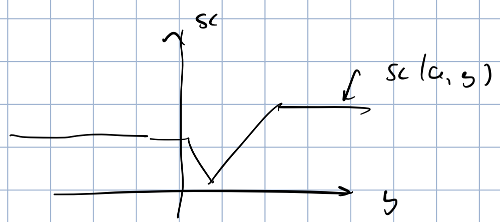
" equivariant: $a(P^k) = k \cdot a(P)$

Influence, Sensitivity, Breakdown

Influence: $\Delta(a, u) = \underbrace{a(y_1, \dots, y_n)}_{\text{w/ unit } u} - \underbrace{a(y_1, \dots, y_{u-1}, y_{u+1}, \dots, y_n)}_{\text{w/out unit } u}$ ⇐ Remove

Sensitivity curve:

$$sc(a, y) = N\left(a(y_1, \dots, y_n, y) - a(y_1, \dots, y_n)\right)$$ ⇐ Add



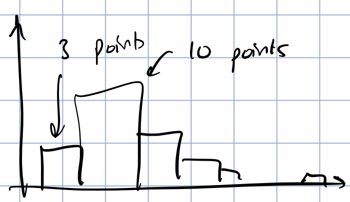
- Perform piecewise cts. if original attribute is piecewise or depends on value of y

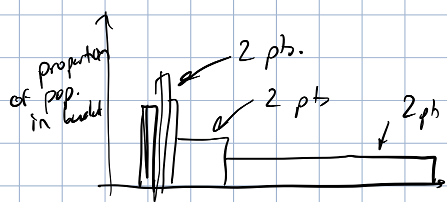
Breakdown: proportion of datapoints set to ∞ for attribute $\to \infty$

↳ High proportion ⇒ robust

Graphical attributes

Histogram w/ equal bin width





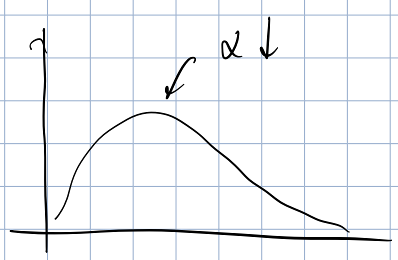
⇓

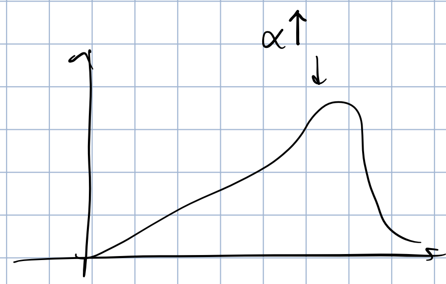
Sturges rule: # of bins $= \lceil \log_2(N) \rceil + 1$

Power Transformations

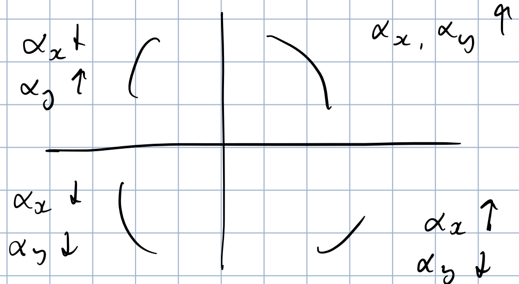
$$T_\alpha(y) = \begin{cases} y^\alpha, & \alpha > 0 \\ \log(y), & \alpha = 0 \\ -(y^\alpha), & \alpha < 0 \end{cases}$$ ⇒ Monotonic

Histogram bump rule:





Scatter plot bump rule:



Order, Rank, Quantiles

Rank: if $y_i = y_{(k)}$ (k^{th} smallest), then $r_i = k$

Quantile: $p_u = \frac{r_u}{N}$

↳ $Q_y(p)$ is p^{th} quantile $= y_{(N \times p)}$

Median: $Q_y(1/2)$

Mid range: $\dfrac{Q_y(1/N) + Q_y(1)}{2}$

IMPLICIT ATTRIBUTES

Examples

Least squares: $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{u \in P} \underbrace{\rho(\theta; u)}$

↳ $\rho(\theta; u) = (y_u - \theta)^2$

Weighted least squares:

$$\rho(\theta; u) = w_u (y_u - \theta)^2$$

Least absolute deviations:

$$\rho(\theta; u) = |y_u - \theta|$$

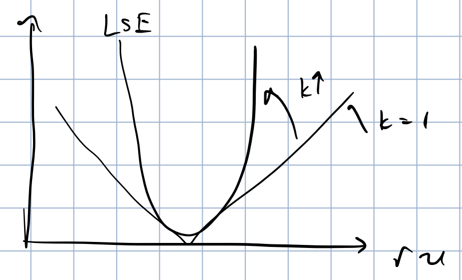
Linear regression:

$$\rho(\theta; u) = \{y_u - (\underbrace{\alpha + \beta}_{\theta} x_u)\}^2$$

Robust regression

Modify $\rho(\theta; u)$ s.t. large residuals have lower weight

Huber loss: $\rho_k(r_u) = \begin{cases} r_u^2/2, & |r_u| \le k \\ k|r_u| - \frac{k^2}{2}, & |r_u| > k \end{cases}$



Least absolute deviations:

$$\rho_k(r_u) = |r_u|$$

Gradient Descent

Algo:

1. $i = 0$, initialize $\hat{\theta}_i = \ldots$
2. Loop:
   a) $g_i = \nabla \rho(\theta; p)|_{\theta = \theta_i}$
   b) $d_i = g_i / \|g_i\|_2$
   c) $\lambda_i = \underset{\lambda > 0}{\operatorname{argmin}}\ \rho(\theta - \lambda d_i)$
   d) $\hat{\theta}_{i+1} = \theta_i - \lambda_i d_i$
   e) Check convergence → Converged? Return; → $i = i+1$, continue loop
3. Return $\hat{\theta} = \theta_i$

Newton's Method

Objective: Find θ s.t. $\Psi(\theta; P) = \sum_{u \in P} \Psi(\theta; u) = \vec{0}$

Algo:

1. Initialize: $i \leftarrow 0$, $\hat{\theta}_0$
2. Loop
   a) Update: $\hat{\theta}_{i+1} = \hat{\theta}_i - \dfrac{\Psi(\hat{\theta}_i; P)}{\Psi'(\hat{\theta}_i; P)}$
   b) Check convergence. Exit if $\hat{\theta}_{i+1} \approx \hat{\theta}_i$

IRLS

Objective: Find $\hat{\theta} = (\alpha, \beta)$ that minimizes $\sum \rho(y_u - \alpha - \beta(x_u - \bar{x}))$

Also:

1. Init: $i \leftarrow 0$, $\hat{\theta}_0 = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}$

2. Loop

   a) Get residual & weights

   $$r_u = y_u - \hat{y}_u = y_u - \underbrace{[1 \quad x_u]}_{z_u'} \hat{\theta}_i$$

   $$w_u = \frac{\rho'(r_u)}{r_u}$$

   b) Solve WLS problem

   $$\sum_{u \in P} w_u r_u z_u = 0 \quad \rightarrow \hat{\theta}_{i+1}$$

   c) Check convergence of $\hat{\theta}_i$ & $\hat{\theta}_{i+1}$. Early exit

Newton Raphson

Goal: $\theta \in \mathbb{R}^n$ s.t. $\psi(\theta; \rho) = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_k \end{bmatrix} = \vec{0}$

Also:

1. Init: $i \leftarrow 0$, $\hat{\theta}_0$

2. Loop

   $$\frac{\partial \psi}{\partial \theta} = \begin{bmatrix} \frac{\partial \psi_1}{\partial \theta_1} & \cdots & \frac{\partial \psi_n}{\partial \theta_1} \\ \vdots & & \\ \frac{\partial \psi_k}{\partial \theta_n} & \cdots & \frac{\partial \psi_k}{\partial \theta_n} \end{bmatrix}$$

   a) Update iterate:

   $$\hat{\theta}_{i+1} = \hat{\theta}_i - [\psi'(\hat{\theta}_i; \rho)]^{-1} \psi(\hat{\theta}_i; \rho)$$

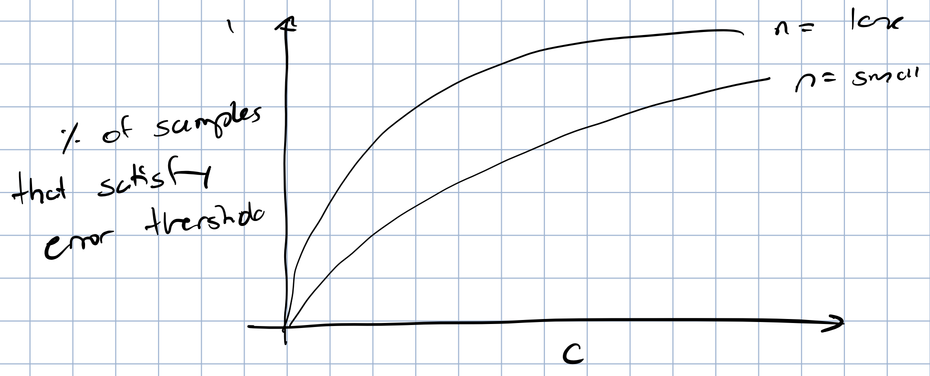
   b) Check convergence & early exit.

SAMPLES

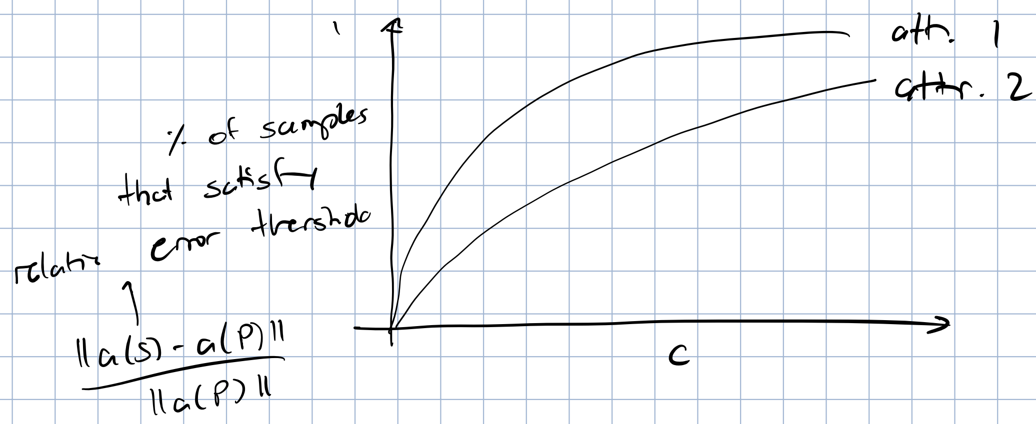
Sample error:

$$a(S) - a(P)$$

Consistency



Across attr. :



Sampling bias

$$E[a(S) - a(P)] = E[a(S)] - a(P)$$

Sampling variance

$$Var[a(S)] = E[a(S)^2] - E[a(S)]^2$$

MSE

$$MSE = Var(a(S)) + Bias(a(S))^2$$

Sampling Mechanism

W/out replacement:

$$P(x) = 1/N, \quad P(x \mid k) = \frac{1}{N-(k-1)}, \quad P(S_n) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \ldots \cdot \frac{1}{N-(n-1)}$$

$$P(S) = \frac{1}{\binom{N}{n}}$$

w/ replacement

$$P(u) = 1/N, \quad P(u \mid k) = \frac{1}{N} \quad P(S_n) = \left(\frac{1}{N}\right)^n$$

$$P(s) = \frac{\frac{n!}{n_1! \cdots n_k!}}{N^n}$$

no duplication

Same as w/ replacement, but sample size $\neq n$ all the time

Unit inclusion probabilities

$$D_u = \begin{cases} 1 & \text{if } u \in S \\ 0 & \text{if } u \notin S \end{cases}$$

$\longrightarrow E[D_u] = \pi_u$

$\longrightarrow Var[D_u] = \pi_u - \pi_u^2$

$\longrightarrow Cov[D_u, D_v] = \pi_{uv} - \pi_u \pi_v$

w/out replacement

$$\pi_u = \frac{1 \cdot \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \quad \pi_{uv} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

w/ replacement

$$\pi_u = 1 - P(u \notin S) = 1 - \left(\frac{N-1}{N}\right)^n$$

$$\pi_{uv} = 1 - [P(u \notin S) + P(v \notin S) + P(u, v \notin S)] = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$$

Horvitz - Thompson Estimator

$$\dot{a}(P) = \sum_{u \in P} y_u \longrightarrow \hat{a}(P) = \sum_{u \in S} \frac{y_u}{\pi_u} = \sum_{u \in S} \frac{y_u}{\pi_u} D_u$$

$$Var[\tilde{a}_{HT}(S)] = \sum_{u \in P} \sum_{v \in P} (\pi_{uv} - \pi_u \pi_v) \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}$$

$$\hat{Var}[\hat{a}_{HT}(s)] = \sum_{u \in P} \sum_{v \in P} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}}\right) \frac{y_u y_v}{\pi_u \pi_v}$$

Function: $a(P) = f\left(\sum_{n \in P}(y_n)\right) \longrightarrow \hat{a}(P) = f(\tilde{a}_{HT}(S))$

INFERENCE

Errors

- Study
- Sample
- Measurement $\rightarrow$ Operator, Instrument, method

Anatomy of Sig. Test

① H_0: P_1 & P_2 from same pop.

② Discrepancy measure

③ Observed discrepancy

④ P-val:

$$Pr(D \geq d_{obs} \mid H_0) \approx \frac{1}{M} \sum_{i=1}^{M} I[D(P_{1,i}, P_{2,i}) \geq d_{obs}]$$

Errors:

- Type I: reject H_0 but H_0 true
- Type II: accept H_0 but H_0 false

Confidence Intervals

① Finite variance:

$$Var[\bar{Y}] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

② Bootstrap C.I.

$$SE = \hat{SD}[\tilde{a}(S)] = \sqrt{\frac{\sum_{i=1}^{B}(a(S_b^*) - \bar{a}^*)^2}{B-1}} \qquad \bar{a}^* = \frac{1}{B}\sum_{i=1}^{B} a(S_b^*)$$

A: Naive normal

$$C.I. = a(S) \pm \underset{N}{c} \times \underbrace{\hat{SD}[\tilde{a}(S)]}_{\text{bootstrap}}$$

B: Percentile

Take $p/2^{th}$ & $(1-p/2)^{th}$ percentile of bootstrap distr.

C: Bootstrap-t

Critical value from distr. of $z_b = \dfrac{a(S_b^*) - a(S)}{\widehat{SD}_* [\hat{a}(S_b^*)]}$ ← 2nd bootstrap

PREDICTION

APSE

$$APSE(P, \hat{\mu}_s) = \frac{1}{N} \sum_{u \in P} (y_u - \hat{\mu}_s(x_u))^2$$

Across multiple samples

$$APSE(P, \tilde{\mu}) = \frac{1}{M} \sum_{j=1}^{M} APSE(P, \hat{\mu}_{s_j})$$

$$= \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (y_u - \tau(x_u)) \rightarrow Ave_x (Var[Y|x])$$

$$\bar{\hat{\mu}}(x) = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{s_j}(x)$$

$$+ \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (\hat{\mu}_{s_j}(x_u) - \bar{\hat{\mu}}(x_u))^2 \longrightarrow Var[\tilde{\mu}]$$

$$+ \frac{1}{N} \sum_{u \in P} (\bar{\hat{\mu}}(x_u) - \tau(x_u))^2 \longrightarrow Bias^2[\tilde{\mu}]$$