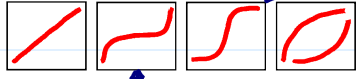


estimation/testing/inference insensible to something (e.g. extreme value, outlier, variation, ...) **Robust regression** *still good for error distribution with light tails*

- Recall: $Y=X\beta+\epsilon$, usually assume error ϵ is Normal \Rightarrow ordinary least square (OLS) approach best. **Q:** what if error not Normally distributed? 
- Recall: particular concern when errors not Normal \Rightarrow long-tailed error \Rightarrow large errors are expected to appear more often

\Rightarrow large errors are expected to appear more often *causing problem for large errors, which would influence the fit more seriously*

\Rightarrow OLS not necessary best when large errors exist (**Q:** why? $RSS = \sum (y_i - x_i^T \beta)^2$)

Previous approach: check and remove observations with large residuals, i.e., regard them as outliers, use OLS after removing them \Rightarrow not effective when there are many outliers because:

difficult to detect outlier cluster

“leave-out-one” nature in outlier tests \Rightarrow not statistically efficient for the estimation of β

Two ways of handling outliers or large errors: *robust regression*

(a) change data, keep model (b) keep data, change model

treating the appearance of large errors as “normal” condition in the model

Statistical modeling: $Y=X\beta+\epsilon$, where error ϵ can be modeled as

- $\epsilon \sim$ a mixture distribution, e.g., $I = \begin{cases} 1, & \pi \\ 0, & 1-\pi \end{cases} \begin{cases} \epsilon | I=1 \sim N(0, \sigma^2) \\ \epsilon | I=0 \sim N(0, c\sigma^2) \end{cases}$
- $\epsilon \sim \pi N(0, \sigma^2) + (1-\pi) N(0, c\sigma^2)$, $0 < \pi < 1$ and $c > 1$
- $\epsilon \sim \sigma t_d$ distribution with a small d \Rightarrow $d=1$, $t_1 =$ Cauchy $\Rightarrow \epsilon \sim t_d \Rightarrow E(\epsilon^k)$ not exist for $k \geq d$ due to heavy tails
- $\epsilon \sim$ any distribution with median=0 \Rightarrow nonparametric approach, infinitely many parameters

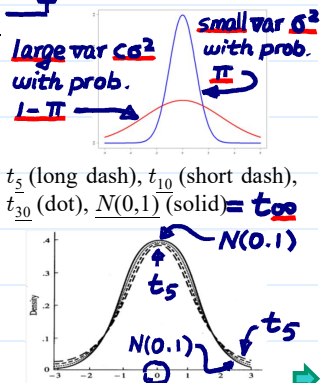
small var σ^2 with prob. π
large var $c\sigma^2$ with prob. $1-\pi$

parameter

pdf symmetric about 0

nonparametric approach, infinitely many parameters

t_5 (long dash), t_{10} (short dash), t_{30} (dot), $N(0,1)$ (solid) = too



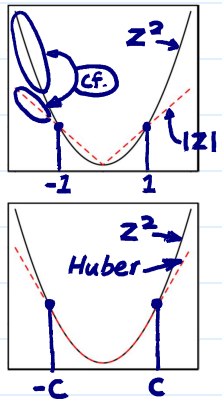
alternative approach: robust regression (observations are weighted unequally) p. 10-2

M-estimators: find β to minimize $\sum_i \rho((y_i - x_i^T \beta)/\sigma)$ *standard deviation of ϵ parameter*

- choice of ρ : *loss function*
- $\rho(z) = z^2$ is OLS $\Rightarrow \hat{\epsilon}_i(\beta)$ *standardized error: ϵ_i/σ*
- $\rho(z) = |z|$ is called least absolute deviations (LAD) regression $\Rightarrow \sigma$ is irrelevant in minimization if $\rho'(t \cdot z) = g(t) \rho'(z)$
- Huber method: $\rho(z) = z^2$, if $|z| \leq c$, and $2c|z| - c^2$, if $|z| > c$.

It's a compromise between OLS and LAD.

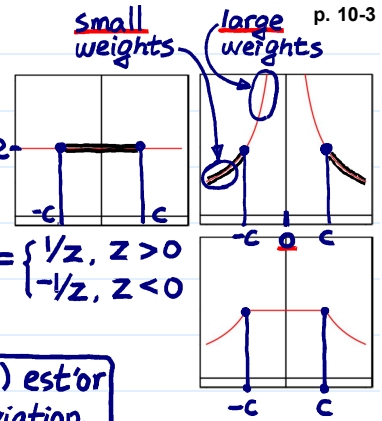
Q: how to pick c ? suggestion: $c \in [1, 2]$ $\Rightarrow c=1, |z| > 1 \Leftrightarrow |\hat{\epsilon}_i(\beta)| > \sigma$
 $\Rightarrow c=2, |z| > 2 \Leftrightarrow |\hat{\epsilon}_i(\beta)| > 2\sigma$



- compute M-estimates (related to iteratively re-weighted least square, IRWLS): *LNp. 6-5*
- for LS with weights, estimate β by solving $(X^T W X) \beta = X^T W Y$
- $\Rightarrow X^T W (Y - X \beta) = 0$, i.e., $\sum_i w_i x_{ij} (y_i - \sum_k x_{ik} \beta_k) = 0$ for all $j=1, \dots, p$
- for robust estimates, differentiating the M-estimate criterion w.r.t. β_j and setting to zero, we get: $\sum_i \rho'((y_i - \sum_k x_{ik} \beta_k)/\sigma) x_{ij} = 0$ for all $j=1, \dots, p$
- let $u_i = (y_i - \sum_k x_{ik} \beta_k)/\sigma$, we get $\sum_i (\rho'(u_i)/u_i) x_{ij} (y_i - \sum_k x_{ik} \beta_k) = 0$
- \Rightarrow set $w_i = \rho'(u_i)/u_i$, can use WLS to estimate β (if w_i 's/ u_i 's known)
- \Rightarrow but, u_i depends on the residuals $\hat{\epsilon}_i \leftarrow \hat{\beta} \xleftarrow{WLS} w_i \leftarrow u_i \leftarrow \hat{\epsilon}_i \leftarrow \dots$
- \Rightarrow IRWLS \Rightarrow can be used to estimate σ

WLS estimator:
 $\hat{\beta} = (X^T W X)^{-1} X^T W Y = (X^T W^* W^* X)^{-1} X^T W^* W^* Y$
obtained from the normal equation (LNp 3-5) of WLS:
 $(X^T W^* W^* X) \beta = X^T W^* W^* Y$

- weights for various ρ [note: larger residuals cause smaller weights in robust method] $\rho(z) = z^2 \Rightarrow \frac{\rho'(z)}{z} = \frac{2z}{z} = 2$
 - OLS: $w(u) = 2$ is a constant \Rightarrow all $\hat{\epsilon}_i(\beta)$'s equally weighted
 - LAD: $w(u) = 1/|u|$ --- note the asymptote at 0, it may make a weighting approach difficult $\rho(z) = |z| \Rightarrow \frac{\rho'(z)}{z} = \begin{cases} 1/z, & z > 0 \\ -1/z, & z < 0 \end{cases}$
 - Huber: $w(u) = 2$, if $|z| \leq c$, and $2c/|u|$, if $|z| > c$.



procedure: IRWLS for M-estimator

- (1) start with any estimate of β , say OLS
- (2) compute residuals $\hat{\epsilon}_i$

a robust (why?) est'or of standard deviation, but non-linear

75% quantile of $N(0, \sigma^2)$ is 0.6745 σ

s.e. $(\hat{\beta}_j) = \hat{\sigma} \sqrt{(x^T W x)^{-1}_{jj}}$

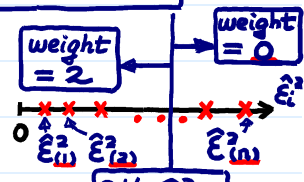
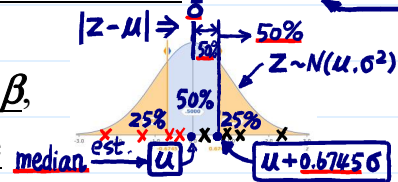
(3) compute u_i , may use median $|\hat{\epsilon}_i - \text{median}(\hat{\epsilon}_i)| / 0.6745$ to estimate σ

(4) compute $w_i = \rho'(u_i) / u_i$

(5) do WLS to get a new estimate of β

then go to step (2) until converge

get final $\hat{\beta}, W, \hat{\sigma}$



resistant regression (more resistant to outliers than M-estimators):

M-est'or least trimmed squares (LTS): only use $\hat{\epsilon}_{(1)}^2, \dots, \hat{\epsilon}_{(g)}^2$

Use order statistics $\hat{\epsilon}_{(1)}^2, \hat{\epsilon}_{(2)}^2, \dots, \hat{\epsilon}_{(n)}^2$ of $\hat{\epsilon}_i^2(\beta)$'s

find β to minimize $\sum_{i=1}^g |y_i - x_i^T \beta|^2$

LTS: P: where (i) indicates sorting, and $g < n$ [$g \approx (n+p+1)/2$ is recommended]

not continuous

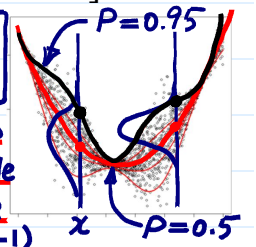
least median of squares (LMS): find β to minimize median $|y_i - x_i^T \beta|^2$

β estimated by S-estimation method. [see Rousseeuw and Leroy, 1987]

resistant regression will do well even if a substantial proportion of data is "bad" (see an example in Lab)

a choice of small $g < n$ in LTS & the use of median in LMS

- quantile regression
 1. directly model $\text{median}(y_x) = x^T \beta$ & p th-quantile $(y_x) = x^T \beta_p$
 2. for normal errors, $x^T \beta = E(y_x) = \text{median}(y_x)$ & all quantile



- Q: Why not always use robust estimates? lines are parallel under constant variance assumption (check graph in Lnp. 4-1)
- if errors are (close to) normally distributed, robust estimators are less efficient

very little distribution theory for robust estimator: can estimate β and (possibly) their standard errors, but, methodology and software for inference, such as testing, is not easy to come by. [\Rightarrow may try bootstrap method] treating empirical cdf of residuals as the true error distribution

recommendation: use robust estimates as a check on OLS estimates. If they are close, use OLS theory. If not, try to find out why. \Rightarrow can use robust estimate in the detection of outliers (large residuals)

Note: robust estimators provide protection against long-tailed errors, but they cannot overcome problems with non-constant variance or curvature in the mean of residuals.

MSE = $\text{Var} + \text{Bias}^2$ OLS $\xrightarrow{\text{Gauss-Markov condition}}$ BLUE \rightarrow unbiased \leftarrow ridge est'or (biased) \Rightarrow overcome strong collinearity
 linear combination of Y \leftarrow robust est'or (non-linear) \Rightarrow overcome large errors

Reading: Faraway (1st ed.), 6.4, Further reading: D&S, chapter 25
 censored data, e.g., only know $y > c$, but not observe exact y (data partially known), often seen in lifetime data. Incomplete data $\left\{ \begin{array}{l} \text{missed} \\ \text{truncated} \\ \text{censored} \end{array} \right\}$ information lost

Some values of some cases are missing. Q: When this happened, what can be done?
 find them --- may not be possible \leftarrow values unknown

ask why the data are missing, i.e., what is the missing mechanism?

$M_i | (x_i, y_i)$ missing completely at random (MCAR): missing probability M_i : missing indicator (r.v.)
 have same p is the same for all cases \Rightarrow non-informative missing M_i 's indep Bernoulli (P_i)

$(X, Y) \perp M$ missing at random (MAR): missing probability is not constant, but depends on a known mechanism, say some observed variables $T \Rightarrow$ non-informative missing if T are included in the model

like SRS in LNP 5-6-7

missing not contain useful information in studying (X, Y)

missing not at random (MNAR): missing probability is not constant, and depends on some unknown mechanism \Rightarrow informative missing, e.g.:

e.g., T not observed

$(X, Y) | M=0 \sim (X, Y) \Rightarrow Y | X, M=0 \sim Y | X$ (LM)

population SRS \downarrow sample \downarrow observed data

Same "pattern" \Rightarrow MCAR \downarrow observed data

containing useful information of (X, Y) in missing mechanism

- People having something to hide are T typically less likely to provide information
- Patients drop out a drug study more often when they feel treatment is not working T

$T=t_1$ P_i 's $= g(t_1)$

$T=t_2$ P_i 's $= g(t_2)$

$T=t_k$ P_i 's $= g(t_k)$

P_i 's different But, $P_i = g(t_i)$

MNAR data require special assumptions and modeling [see Little and Rubin, 2019]

wrong $Y | X$ distribution \Rightarrow Analyses without considering the information in missingness may cause biased conclusion.

$(X, Y) | T, M=0 \sim (X, Y) | T \Rightarrow Y | X, T, M=0 \sim Y | X, T$ (LM with T)

$(X, Y) \perp M$ given $T \not\Rightarrow (X, Y) \perp M$

- some fix-up methods for non-informative missing
- approach 1: deletion, i.e., ignore and delete cases with missing value $(X, Y) | M=0 \times (X, Y) \Rightarrow$ no bias but lose information. It is OK if % of missing data is small. $Y | X, M=0 \times Y | X$

approach 2: single imputation (SI), i.e., fill-in or impute a missing value, e.g.,

- replace missing value by average of predictor, estimate $E(X_i)$: unconditional mean of i th predictor often causing a bias of β toward 0. Recall measurement error in predictor (LNP 9-7-8)
- use a regression model to predict x_i using other predictors, $E[X_i | X_{(-i)}]$: conditional mean of i th predictor given the other ones

\Rightarrow how much trouble to take in building these models?

\Rightarrow may be difficult with multiple missing values

\Rightarrow cause some bias, but filled-in case will have lower leverage

only consider imputing missing predictors, not imputing response.

$\hat{\beta}$ \rightarrow $X_1 \sim X_2 + \dots + X_p$

$E[X_i | X_{(-i)}] \rightarrow E(X_i)$ when the collinearity between X_i & $X_{(-i)}$ is weak (e.g., low VIF_i, LNP 9-9)

\Rightarrow Q: Is inference valid after estimating the coefficients? testing / C.I. \rightarrow $s.e.(\hat{\beta})$ underestimated, d.f. = ?

less influence on the fit

A SI value tends to be less variable than the missing value because the imputed value does not include the error variation. might consider Bootstrap

approach 3: multiple imputation (MI), i.e., impute a missing value m times by multiple draws from predictive distribution $X_i | X_{(-i)}$ dist. X_i dist.

if m small, inferior than SI generate pseudo-random numbers

Why can $m > 1$?

MI re-includes error variation, which reflects uncertainty about imputed values and yields valid estimates of standard errors.

$E(X_2) | X_1 = x_1$ dist.

$E(X_2)$

- MI may better mitigate bias
- Let $\hat{\beta}_{ij}$ and s_{ij} be the estimate and standard error of the coefficient β_i of x_i for the j th imputed result, $j = 1, \dots, m$.
 - The combined estimate of β_i is: $\hat{\beta}_i = \frac{1}{m} \sum_{j=1}^m \hat{\beta}_{ij}$

predict mean \xrightarrow{sf} predict future response $E(Y | X) \leftarrow$ Same value $\rightarrow Y | X$

predict \xrightarrow{sf} predict $E[X_i | X_{(-i)}] \leftarrow$ Same value $\rightarrow X_i | X_{(-i)}$

predict $E(x_i) \xrightarrow{cf}$ predict X_i

average within-imputation variation (due to ϵ in LM) \square The combined standard errors s_i of $\hat{\beta}_i$ is given by: $s_i^2 = \frac{1}{m} \sum_{j=1}^m s_{ij}^2 + \left(1 + \frac{1}{m}\right) \text{var}(\hat{\beta}_i)$, where $\text{var}(\hat{\beta}_i)$ is the (unbiased) sample variance over the imputed $\hat{\beta}_{ij}$'s.

between-imputation variation (due to random imputation)

adjustment for finite (m) imputations based on a Bayesian approach

➤ approach 4: maximum likelihood method

Assuming complete data $D = (D_{\text{obs}}, D_{\text{mis}})$, both observed and missing, are from a family of distribution with parameters θ , say multivariate normal, then it is possible to compute maximum likelihood estimates using:

- (if available) the likelihood of θ based on D_{obs} :

Integration might be difficult or impossible

$$\mathcal{L}(\theta | D_{\text{obs}}) = \int f_D(D_{\text{obs}}, D_{\text{mis}} | \theta) dD_{\text{mis}}$$

joint pdf/pmf of complete data D

$$dF_{D_{\text{mis}}}(D_{\text{mis}} | D_{\text{obs}}; \hat{\theta})$$

Note. not impute D_{mis} , but a function (i.e., l) of D_{mis}

- the EM algorithm
- But,
- the distribution assumption might not be tenable
 - tests, inferences, and diagnostics are not easy to come by

Why? compare with imputation using $E_{X_i | \hat{\beta}, X_{(-i)}}(X_i)$ in LNp.10-6

E-step (Expectation, but it's actually an imputation):
 calculate expected complete-data loglikelihood
 $Q(\theta | \hat{\theta}, D_{\text{obs}})$
 $\equiv E_{D_{\text{mis}} | \hat{\theta}, D_{\text{obs}}} [l(\theta | D_{\text{obs}}, D_{\text{mis}})]$

M-step (maximization):
 maximize $Q(\theta | \hat{\theta}, D_{\text{obs}})$ w.r.t. θ

❖ Reading: Faraway (1st ed.), chapter 12; W, 5.6