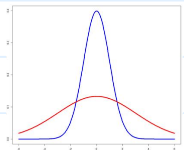
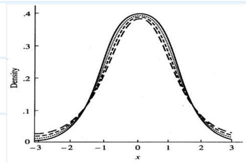
Robust regression

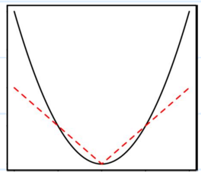
- Recall: $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, usually assume error $\boldsymbol{\varepsilon}$ is Normal $\Rightarrow$ ordinary least square (OLS) approach best. **Q**: what if error not Normally distributed?
- Recall: particular concern when errors not Normal $\Rightarrow$ long-tailed error
  $\Rightarrow$ large errors are expected to appear more often
  - 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:
    - "leave-out-one" nature in outlier tests
    - not statistically efficient for the estimation of $\boldsymbol{\beta}$
  - Two ways of handling outliers or large errors:
    (a) change data, keep model (b) keep data, change model
- Statistical modeling: $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, where error $\boldsymbol{\varepsilon}$ can be modeled as
  - $\boldsymbol{\varepsilon} \sim$ a mixture distribution, e.g.,
    $\boldsymbol{\varepsilon} \sim \pi\, N(0, \sigma^2) + (1-\pi)\, N(0, c\,\sigma^2)$, $0<\pi<1$ and $c>1$
  - $\boldsymbol{\varepsilon} \sim \sigma\, t_d$ distribution with a small d
  - $\boldsymbol{\varepsilon} \sim$ any distribution with median=0

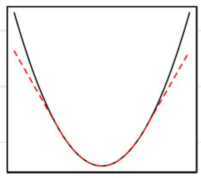


t_5 (long dash), t_{10} (short dash), t_{30} (dot), $N(0,1)$ (solid)

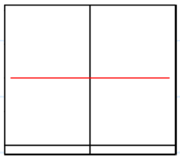
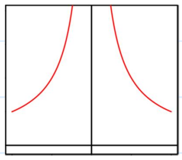
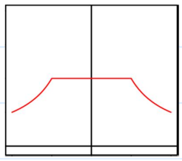


- alternative approach: robust regression (observations are weighted unequally)
  - M-estimators: find $\boldsymbol{\beta}$ to minimize $\Sigma_i\, \rho((y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma)$
    - choice of ρ :
      - $\rho(z) = z^2$ is OLS
      - $\rho(z) = |z|$ is called least absolute deviations (LAD) regression
      - Huber method: $\rho(z) = z^2$, if $|z| \le c$, and $2c|z| - c^2$, if $|z|>c$.
        It's a compromise between OLS and LAD.
        [**Q**: how to pick c? suggestion: $c \in [1, 2]$]
      - many other choices, such as Tukey's biweight, Hampel, ...

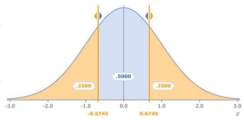




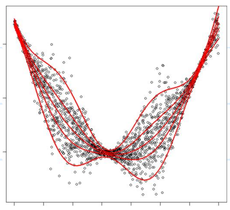
- - compute M-estimates (related to iteratively re-weighted least square, IRWLS):
      - for LS with weights, estimate $\boldsymbol{\beta}$ by solving $(\boldsymbol{X}^T\boldsymbol{WX})\boldsymbol{\beta}=\boldsymbol{X}^T\boldsymbol{WY}$
        $\Rightarrow \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{X\beta})=\boldsymbol{0}$, i.e., $\Sigma_i\, w_i x_{ij}\, (y_i - \Sigma_k\, x_{ik}\beta_k) = 0$ for all $j=1,\ldots,p$
      - for robust estimates, differentiating the M-estimate criterion w.r.t. β_j and setting to zero, we get:
        $\Sigma_i\, \rho'((y_i - \Sigma_k\, x_{ik}\beta_k)/\sigma)\, x_{ij} = 0$ for all $j=1,\ldots,p$
      - let $u_i = (y_i - \Sigma_k\, x_{ik}\beta_k)/\sigma$, we get $\Sigma_i\, (\rho'(u_i)/u_i)\, x_{ij}\, (y_i - \Sigma_k\, x_{ik}\beta_k) = 0$
        $\Rightarrow$ set $w_i = \rho'(u_i)/u_i$, can use WLS to estimate $\boldsymbol{\beta}$
        $\Rightarrow$ but, u_i depends on the residuals $\hat{\varepsilon}_i \leftarrow \hat{\boldsymbol{\beta}} \leftarrow w_i \leftarrow u_i \leftarrow \hat{\varepsilon}_i \leftarrow \ \ldots$
        $\Rightarrow$ IRWLS

- weights for various ρ [note: larger residuals cause smaller weights in robust method]
  - OLS: $w(u)=2$ is a constant
  - LAD: $w(u)=1/|u|$ --- note the asymptote at 0, it may make a weighting approach difficult
  - 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 $\boldsymbol{\beta}$, say OLS
  (2) compute residuals $\hat{\varepsilon}_i$
  (3) compute u_i, may use median $|\hat{\varepsilon}_i - \text{median}(\hat{\varepsilon}_i)|$ /0.6745 to estimate σ
  (4) compute $w_i = \rho'(u_i)/u_i$
  (5) do WLS to get a new estimate of $\boldsymbol{\beta}$,
  then go to step (2) until converge



- ➢ resistant regression (more resistant to outliers than M-estimators):
  - least trimmed squares (LTS):
    find $\boldsymbol{\beta}$ to minimize $\sum_{i=1}^{q} |y_i - \mathbf{x}_i^T\boldsymbol{\beta}|^2_{(i)}$,
    where (i) indicates sorting, and $q<n$ [$q \approx (n+p+1)/2$ is recommended]
  - least median of squares (LMS): find $\boldsymbol{\beta}$ to minimize median $|y_i - \mathbf{x}_i^T\boldsymbol{\beta}|^2$



  - $\boldsymbol{\beta}$ 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)
- ➢ quantile regression

- **Q**: Why not always use robust estimates?
  - ➢ if errors are (close to) normally distributed, robust estimators are less efficient
  - ➢ very little distribution theory for robust estimator: can estimate $\boldsymbol{\beta}$ and (possibly) their standard errors, but, methodology and software for inference, such as testing, is not easy to come by. [⇒ may try bootstrap method]
  - ➢ recommendation: use robust estimates as a check on OLS estimates. If they are close, use OLS theory. If not, try to find out why.
- 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.

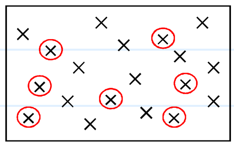
❖ **Reading**: Faraway (1st ed.), 6.4, ❖ **Further reading**: D&S, chapter 25

Incomplete data

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

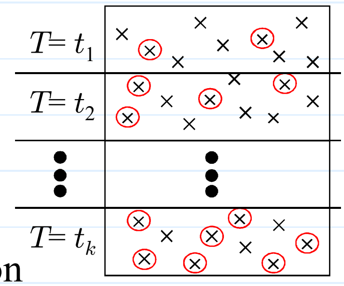
- ask why the data are missing, i.e., *what is the missing mechanism?*
  - missing completely at random (MCAR): missing probability is the same for all cases ⇒ non-informative missing

Sampled cases



  - missing at random (MAR): missing probability is not constant, but depends on a *known* mechanism, say some observed variables T ⇒ non-informative missing if T are included in the model
  - missing not at random (MNAR): missing probability is not constant, and depends on some *unknown* mechanism ⇒ informative missing, e.g.:
    - People having something to hide are typically less likely to provide information
    - Patients drop out a drug study more often when they feel treatment is not working

Sampled cases



    MNAR data require special assumptions and modeling [see Little and Rubin, 2019] ⇒ Analyses without considering the information in missingness may cause biased conclusion.

- some fix-up methods for non-informative missing
  - approach 1: deletion, i.e., ignore and delete cases with missing value ⇒ no bias but lose information. It is OK if % of missing data is small.

- approach 2: single imputation (SI), i.e., fill-in or impute a missing value, e.g.,
  - replace missing value by average of predictor, often causing a bias of $\boldsymbol{\beta}$ toward $\boldsymbol{0}$.
  - use a regression model to predict x_i using other predictors,

    ⇒ how much trouble to take in building these models?

    ⇒ may be difficult with multiple missing values

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

    ⇒ **Q**: Is inference valid after estimating the coefficients?
  - A SI value tends to be less variable than the missing value because the imputed value does not include the error variation.
- approach 3: multiple imputation (MI), i.e., impute a missing value m times by multiple draws from predictive distribution
  - MI re-includes error variation, which reflects uncertainty about imputed values and yields valid estimates of standard errors.
  - 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 jth imputed result, $j = 1, \ldots, m$.
    - The combined estimate of β_i is: $\hat{\beta}_i = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{ij}$

- The combined standard errors s_i of $\widehat{\beta}_i$ is given by:

$$s_i^2 = \frac{1}{m}\sum_{j=1}^{m} s_{ij}^2 + \left(1+\frac{1}{m}\right)\mathsf{var}(\widehat{\beta}_i),$$

where $\mathsf{var}(\widehat{\beta}_i)$ is the (unbiased) sample variance over the imputed $\hat{\beta}_{ij}$'s .

➢ approach 4: maximum likelihood method

Assuming complete data $D = (D_{obs}, D_{mis})$, 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} :

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

- the EM algorithm

But,

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

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