Lecture 11

Automatic Variable Selection - Recap

Approach 1: greedy algorithms, use F-test for extra sum of squares as a stopping criteria.

Approach 2: penalty function combining SS_{res} and # of predictors

Issue: classical inference does not apply post selection

Next: regularization

$\boxed{\text{Regularization}}$

<u>Ridge regression</u> - was created originally to handle the case when $Z^T Z$ is (nearly) singular

nearly singular.

- The estimate is

$$\tilde{\beta}_\lambda = (Z^T Z + \lambda I)^{-1} Z^T y \qquad \lambda > 0$$

shrinks the estimated LS coeficients. as: $\lambda \to \infty$
$\tilde{\beta} \to 0$
as: $\lambda \to 0$
$\tilde{\beta} \to \hat{\beta}$

Variation: if we do not want to shrink the intercept, we use

$$\tilde{\beta}_\lambda = \left(Z^T Z + \lambda \begin{pmatrix} 0 & \\ & I_{p-1} \end{pmatrix} \right)^{-1} Z^T y \quad , \quad \lambda > 0$$

-

$$\ell(\beta; y, Z, \lambda) = \|y - Z\beta\|^2 + \lambda \|\beta\|^2$$

$$\tilde{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \ \ell(\beta; y, Z, \lambda), \ \lambda > 0$$

in other words, $\|\beta\|^2$ is the penalty for additional parameters

- Advantages:
  - Usually more accurate for predictions
  - More stable when predictors are correlated
- Dissadvantages:
  - Give non-zero values for all predictors
- sarifices unbiassdness for reduces variance

Ridge Regression - Baysian Connection

Suppose

$$y \sim N(Z\beta, \sigma^2)$$

$$\beta \sim N(0, \tau^2 I_p)$$

The posterior dist. of β:

$$f_{\beta|y}(\beta|y) = f_\beta(\beta) f_{y|\beta}(y|\beta) \Big/ f_y(y)$$

$$= \frac{\frac{1}{\tau}(2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2\tau^2}\|\beta\|^2} \times \frac{1}{\sigma}(2\pi)^{-\frac{n}{2}} e^{-\|y - Z\beta\|^2/2\sigma^2}}{f_y(y)}$$

- Maximizing posterior

$\Leftrightarrow$ minimizing $-\log\left(f_\beta(\beta) \times f_{y|\beta}(y|\beta)\right)$

$\Leftrightarrow$ minizing $\frac{1}{2\tau^2}\|\beta\|^2 + \frac{1}{2\sigma^2}\|y - Z\beta\|^2$

$\Leftrightarrow$ minimizing $\|y - Z\beta\| + \frac{\sigma^2}{\tau^2}\|\beta\|^2$

this is the objective function in ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$

also, $\tilde{\beta}$ is the posterior mean: $\tilde{\beta} = E[\beta \mid Y = y, Z]$

mean: $\beta = \mathbb{E}[\beta \mid Y = y, \ldots]$

Principle Components Regression

- Suppose $X \in \mathbb{R}^{n \times d}$, d is large
- if n is not too large compared to d, then we are trying to estimate a model with # parameters close to the # of samples.
- We can reduce the dimension of X to k by taking the k "directions" of highest variance:

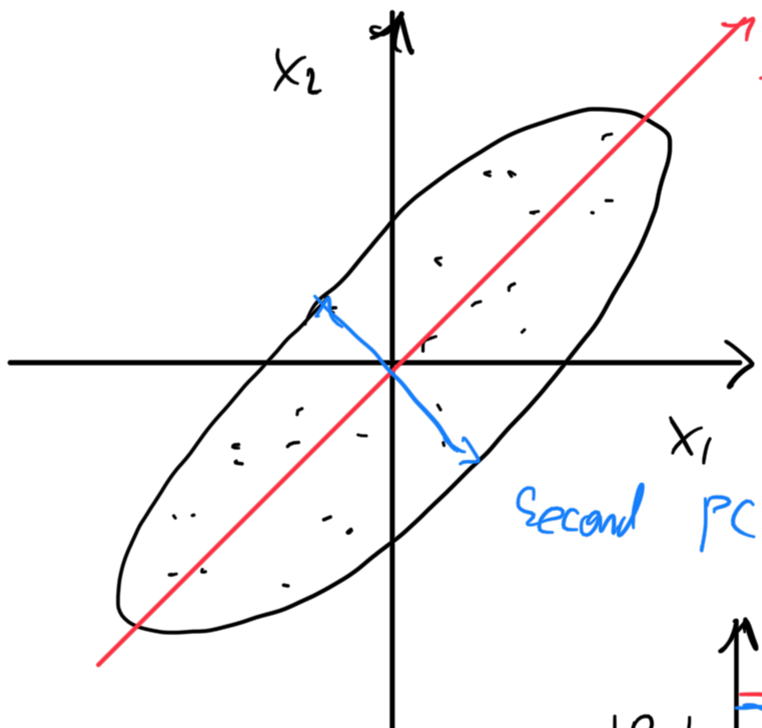
$$\max_{v} \; \text{Var}(X^T v) \quad \text{s.t. } v^T v = I_k$$

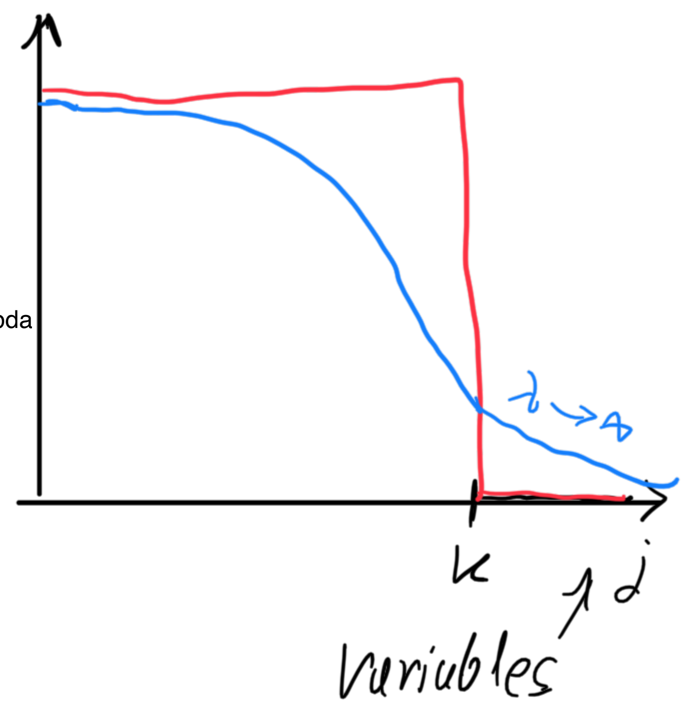
(assuming $\mathbb{E}[X] = X$)

- Advantage: picks out k most important dimensions
- [illegible]

- Issue: may knock-out predictors that may actually be best for predicting y.

PCR vs. Ridge





red line is PCA: for the first K PC's you have the normal value for beta.
Blue line is ridge regression (L2): gets smaller gradually with the value of lambda

L1 Regression / LASSO / Basis Pursuit

- We try to minimize:

$$\ell(\beta; y, Z, \lambda) = \|y - Z\beta\|^2 + \lambda\|\beta\|_1$$

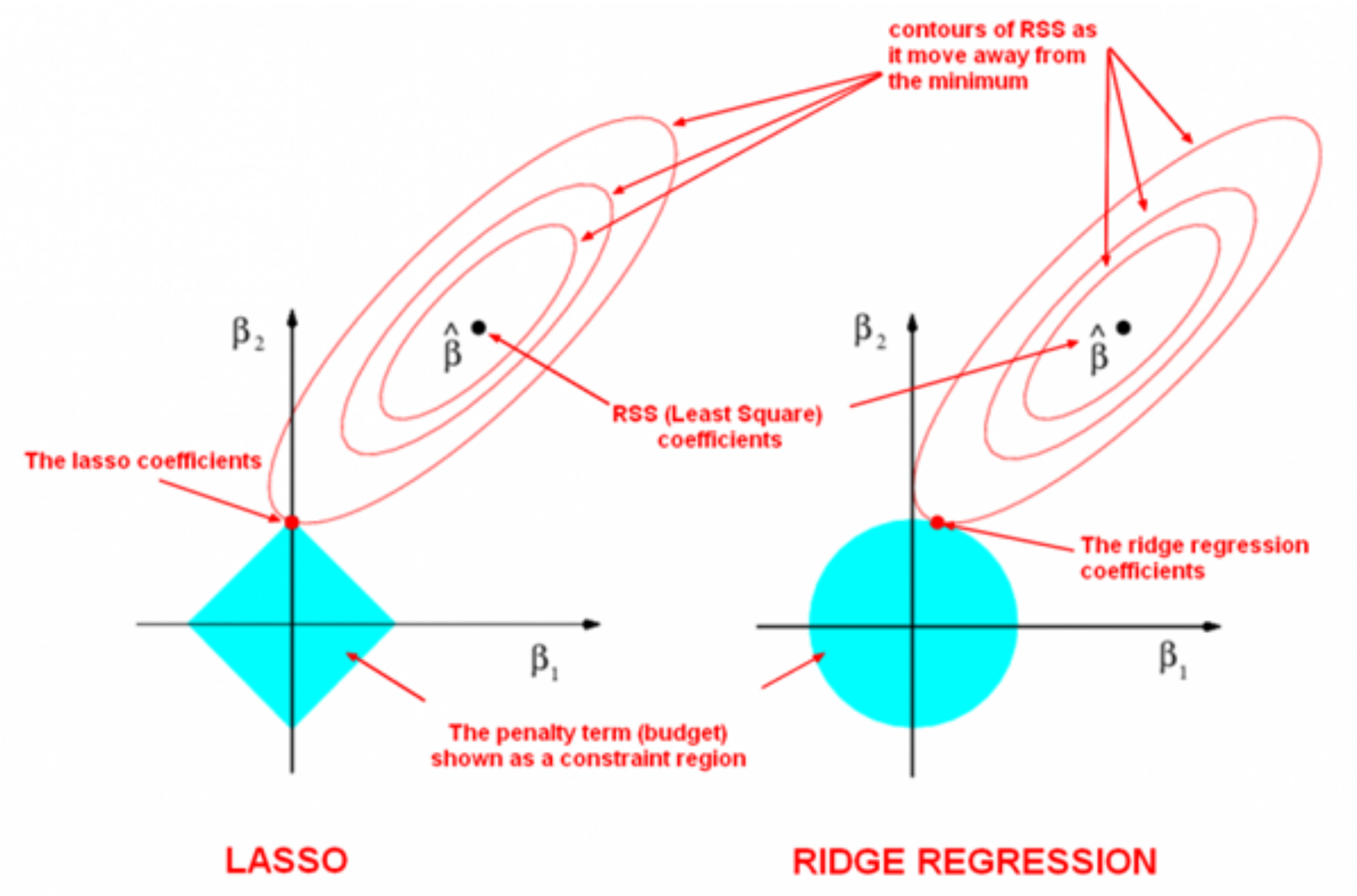
$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

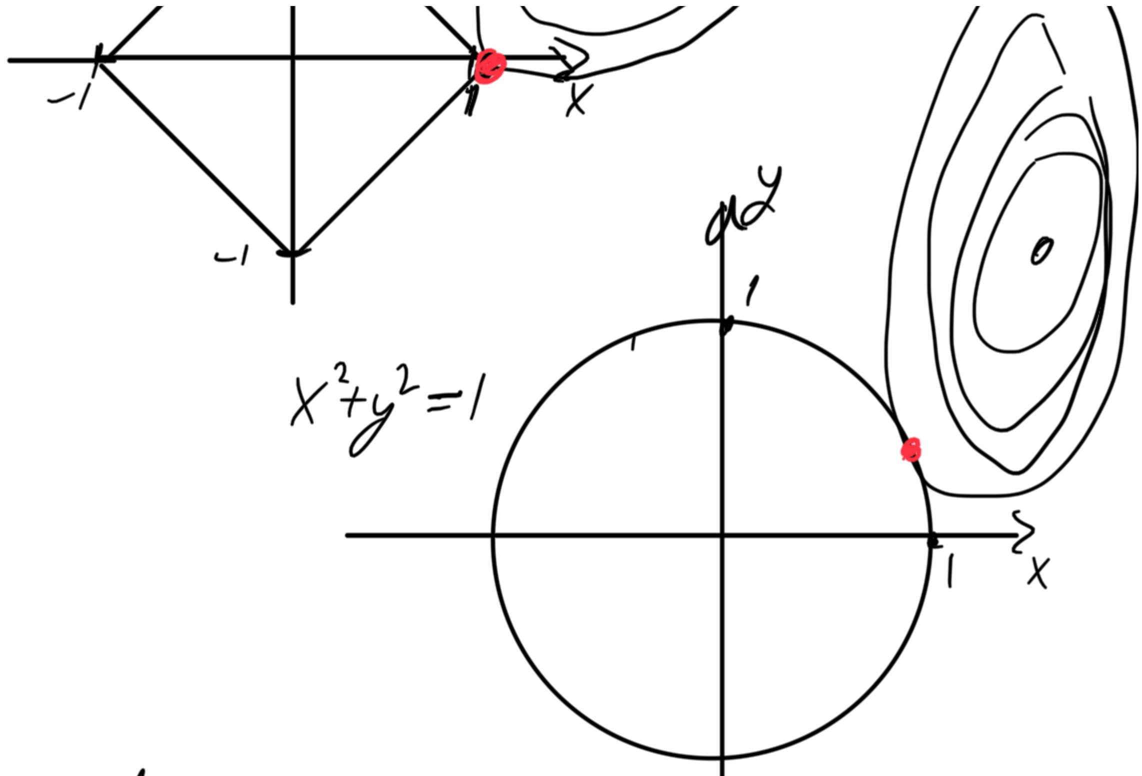
- This has a unique minimum, but no closed-form formula

- Usually, use steepest descent method

- Find λ using CV

- L_1 allows for more exact zero estimates because the edges of the L_1-unit cube are "pointy"


contours of RSS as
it move away from
the minimum
β_2
$\hat{\beta}$
RSS (Least Square)
coefficients
The lasso coefficients
The ridge regression
coefficients
β_1
The penalty term (budget)
shown as a constraint region
LASSO
RIDGE REGRESSION



- sacrifices unbiassedness for less variance.

$\boxed{\text{Violation of Assumptions}}$

Usually we assume $Y \sim N(Z\beta, \sigma^2 I)$ (implies $E[Y] = Z\beta$)

What can go wrong:

- bias $E[Y] \neq Z\beta$

- Non-normality

- Heteroscedasticity: Variances are not common across observations:

$$\mathrm{Var}[Y] = \sigma^2 V \qquad V \neq I_n$$

$\boxed{\text{Bias}}$

Missing Information/Variable

- Suppose $\mathbb{E}[Y] = Z\beta + \tilde{z}\cdot\tilde{\beta}$
  $\tilde{z} \in \mathbb{R}^n, \ \tilde{\beta} \in \mathbb{R}$

- If we regress y on Z:

$$\beta_{OLS} = \underbrace{(Z^T Z)^{-1} Z^T}_{}(\underbrace{Z\beta}_{} + \tilde{z}\cdot\tilde{\beta})$$

$$= \hat{\beta} + (Z^T Z)^{-1} Z^T \tilde{z}\cdot\tilde{\beta}$$

- if $Z^T\tilde{z} = 0$, then we get the usuall $\hat{\beta}$, otherwise $E[\beta_{OLS}] \neq \beta$

Detection:

- We can detect $\tilde{z}$ should be in our model by several methods:

  - Add $\tilde{z}$ to the regression and test for fit using the extra sum of squares

  - Plot $\hat{\varepsilon}_i$ vs $\tilde{z}$ and look for linear relationship

  - Better: plot $\hat{\varepsilon}_i$ vs the residuals of $\tilde{z}$ on Z (added variable plot)

    predict z tilda based on z

    (No need to regress $\hat{\varepsilon}_i$ on Z bc. they are orthogonal)

- In practice, you should check all administrative things that data may depend on: $\hat{\varepsilon}_i$ vs: i, calander time, file order/number, trends, blocking.

Non - Normality

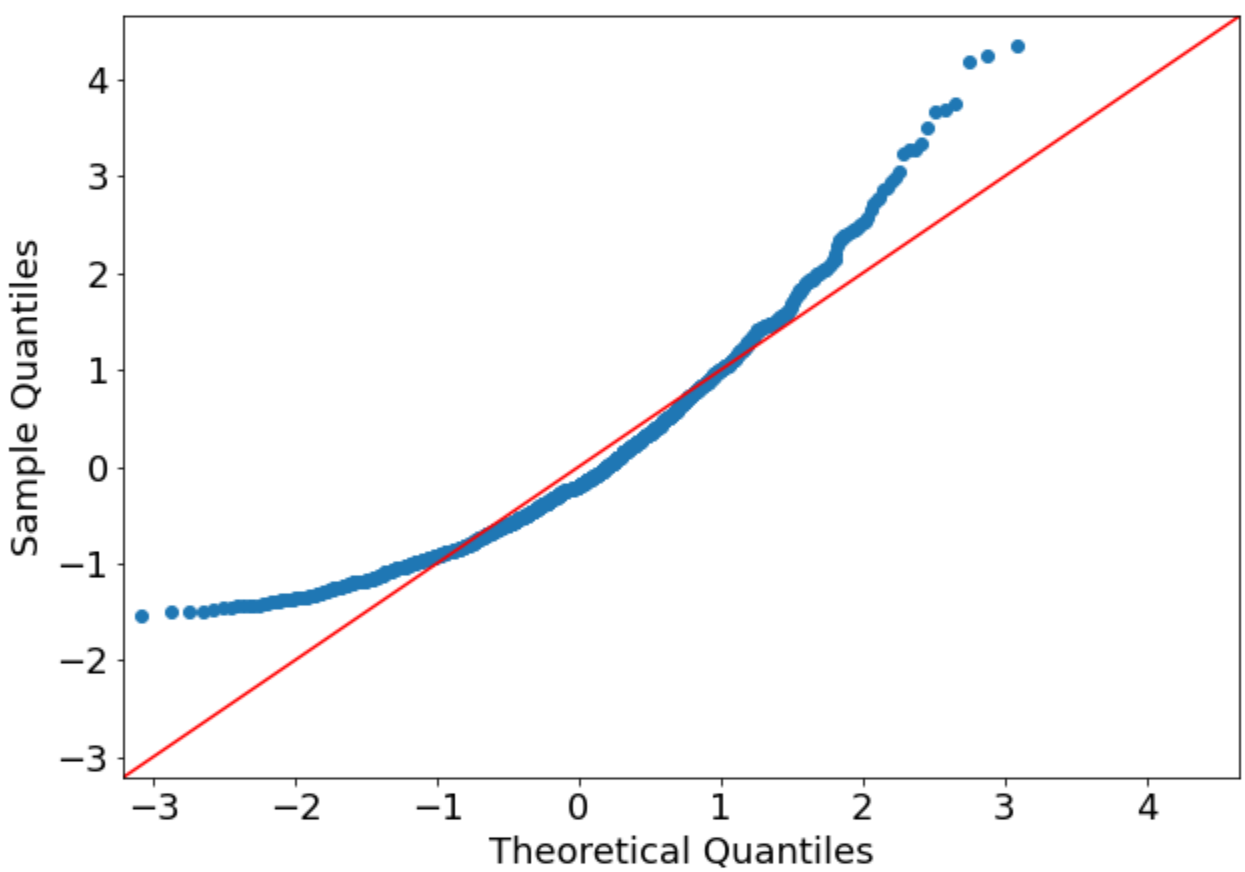
We hope that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
even if $\varepsilon_i \sim (0, \sigma^2)$ that's
ok bc. CLT fixes things.

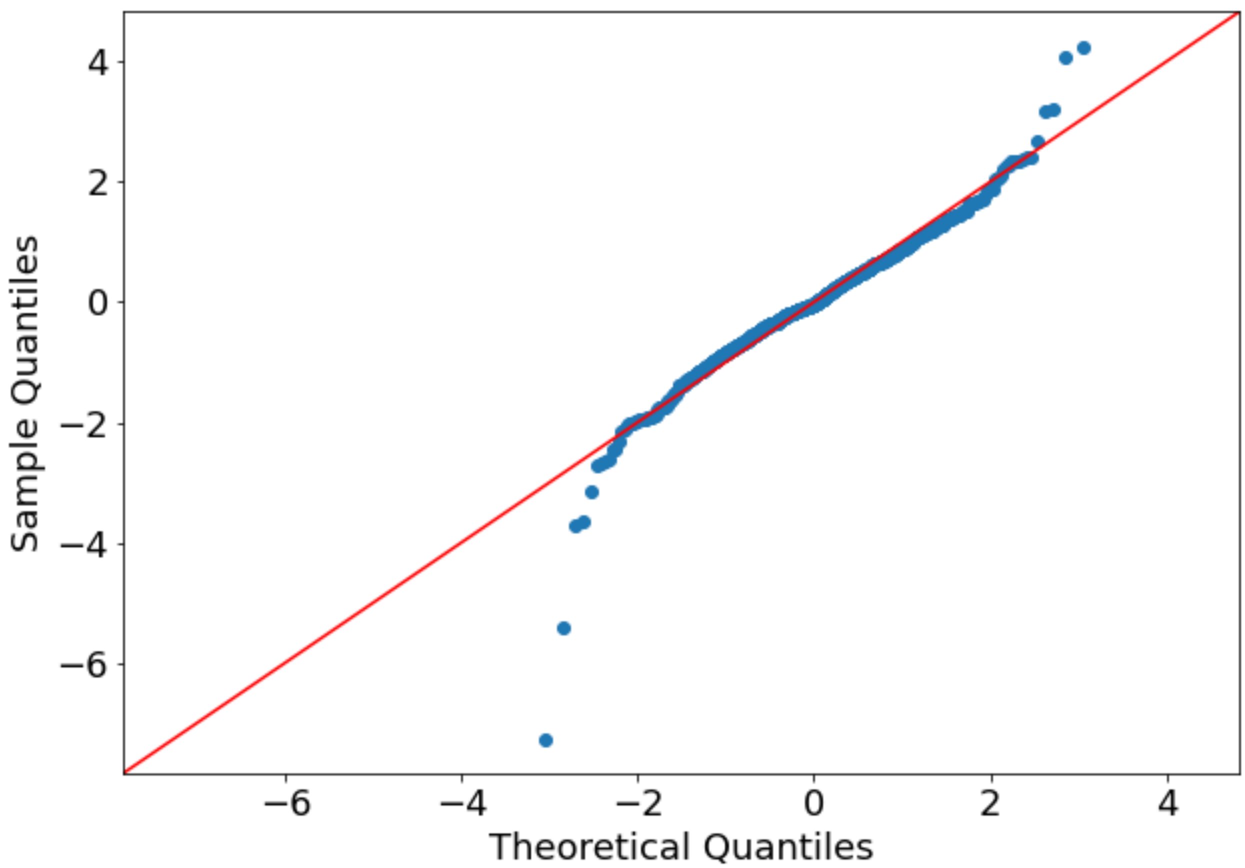
- Conditions for CLT:
  1) $\lambda_{min}(Z^T Z) \to \infty$
     minimal eginvalues
     $Z^T Z$ is a matrix version of n
  2) No z_{ij} is "too large"
  3) ε_i is not heavy tailed

- Detection

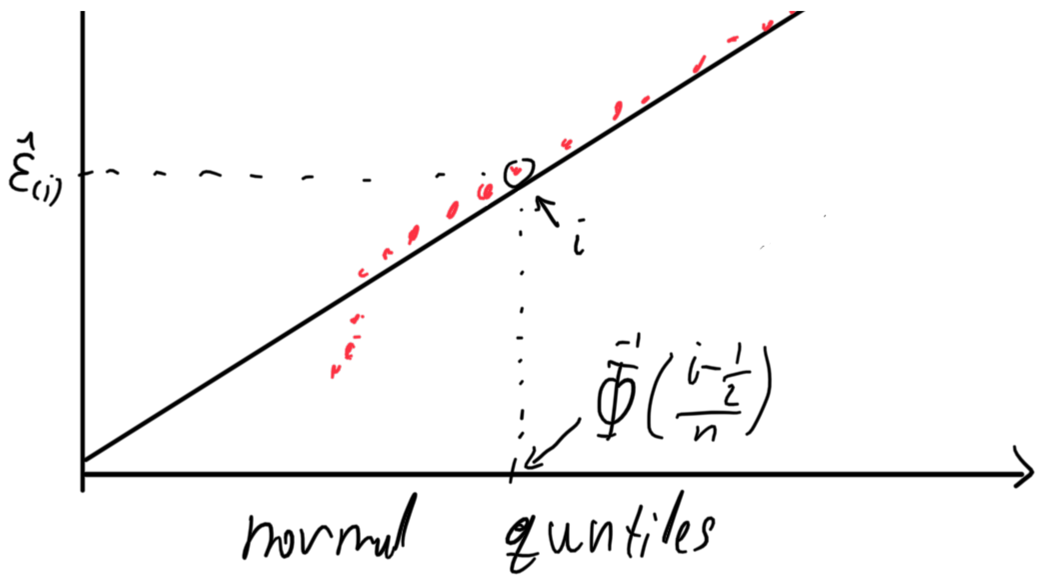
eyeball approach: QQ - Plot:
  1) Sort $\hat{\varepsilon}_i$ so that $\hat{\varepsilon}_{(1)} \le \hat{\varepsilon}_{(2)} \cdots \le \hat{\varepsilon}_{(n)}$
  2) Plot $\hat{\varepsilon}_{(i)}$ vs. $\Phi^{-1}\left(\frac{i - 1/2}{n}\right)$

$$\hat{\varepsilon}_{(1)} \longleftrightarrow \Phi^{-1}\left(\frac{1 - \frac{1}{2}}{n}\right)$$


Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4



4
2
0
−2
−4
−6
Sample Quantiles
−6
−4
−2
0
2
4
Theoretical Quantiles

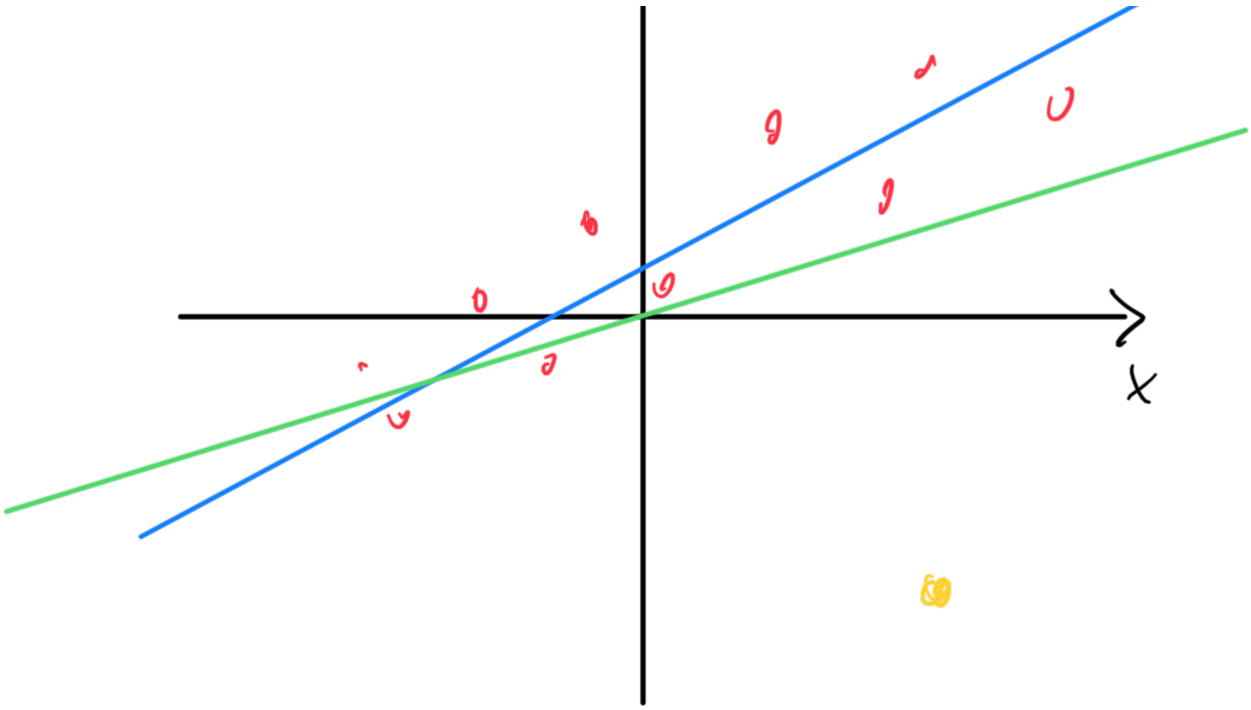


- Ideally the plot follows a striaght line where the slope is σ and the intercept is μ (which is zero with residuals)

- There are many tests for normality of $\hat{\varepsilon}$: Jarque-Bera, Anderson-Darling, Kolmogorov-Smirnoff

- Non-normality is usually an issue only when it comes to outliers

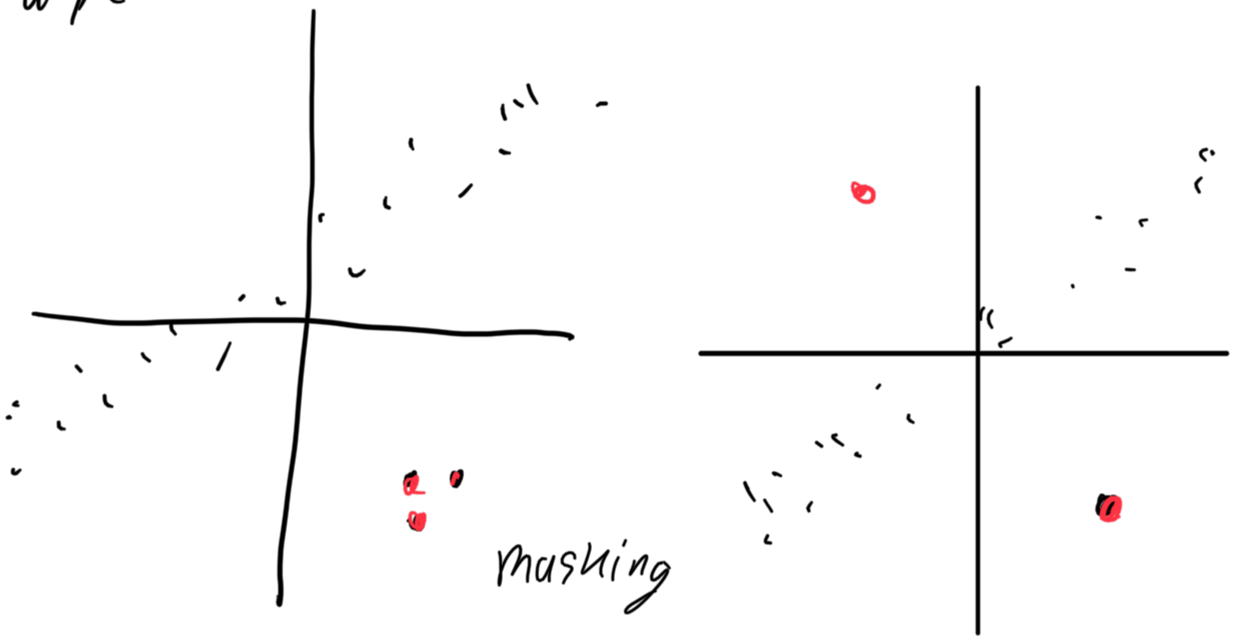
$\boxed{\text{Otliers}}$

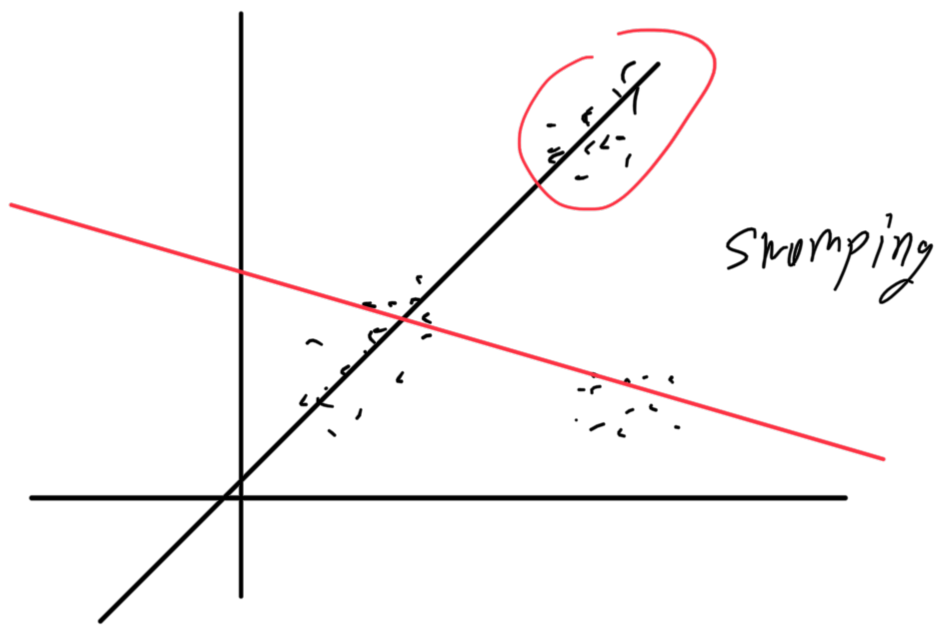


Detection

- A much larger $\frac{|\hat{\varepsilon}_i|}{s}$ than the rest.
- Better: look at $\frac{\hat{\varepsilon}_i^{(i)}}{s^{(i)}}$ (leave-one-out residual)
- Issue: there could be more than 1.
- There exist various methods and heuristics for "auto removal" of outliers. They may fail.

Example:





- Robust Regression methods try to be less sensitive to outliers:

Example 1: L_1 regression:

$$\min_{\beta} \sum_{i=1}^{n} |y_i - z_i^T \beta|$$

Example 2:

Least trimed regression:

Take smallest 80%, say, of squared residuals and fit sum that minimizes those:

$$\boxed{\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{\lfloor 0.8n \rfloor} |\hat{\varepsilon}_{(i)}(\beta)|^2}$$

- The model is robust if less than 20% of the data are outliers

- The bad: difficult to compute (non-convex)