

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

Regularization

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

uniquely singular.

= The estimate is

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

shrinks the estimated LS coefficients. as: $\lambda \rightarrow \infty$

$$\tilde{\beta} \rightarrow 0$$

$$\text{as: } \lambda \rightarrow 0$$

$$\tilde{\beta} \rightarrow \hat{\beta}$$

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

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

x

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

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

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

- Advantages:

- Usually more accurate for predictions
- More stable when predictors are correlated

- Disadvantages:

- Give non-zero values for all predictors
- sacrifices unbiasedness for reduces variance

Ridge Regression - Bayesian Connection

Suppose

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

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

The posterior dist. of β :

$$\begin{aligned} f_{\beta|y}(\beta|y) &= f_{\beta}(\beta) f_{y|\beta}(y|\beta) / f_y(y) \\ &= \frac{1}{\tau} (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2\tau^2} \|\beta\|^2} \times \frac{1}{\sigma} (2\pi)^{-\frac{n}{2}} e^{-\frac{\|y - Z\beta\|^2}{2\sigma^2}} \\ &\quad \underbrace{\hspace{10em}}_{f_y(y)} \end{aligned}$$

- Maximizing posterior

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

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

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

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

also, $\tilde{\beta}$ is the posterior mean.

$$\tilde{\beta} = \text{E}[\beta|y] = \dots$$

$$p = \mathbb{E}[\|p\|^2] - 2\langle p, \mu \rangle$$

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:

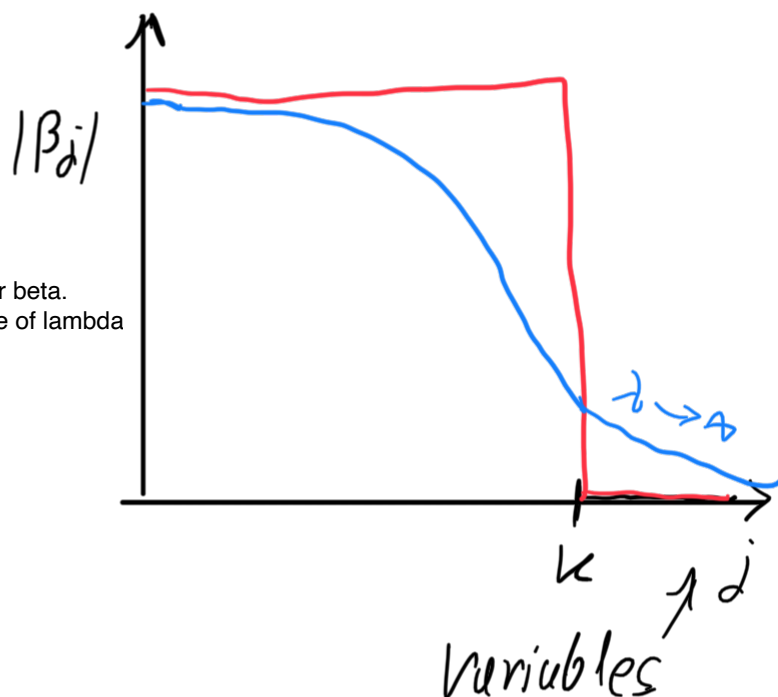
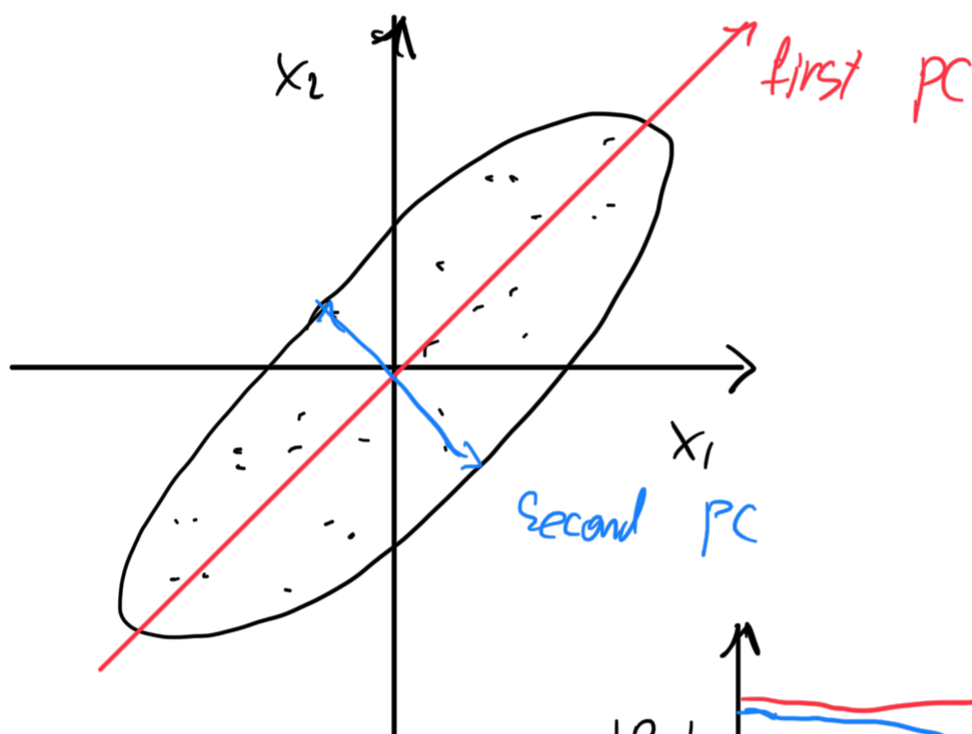
$$\max_{\mathcal{V}} \text{Var}(X^T \mathcal{V}) \quad \text{s.t. } \mathcal{V}^T \mathcal{V} = I_k$$

Assuming $\mathbb{E}(X) = 0$

- Advantage: picks out k most important dimensions

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

PCR vs. Bridge



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:

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

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

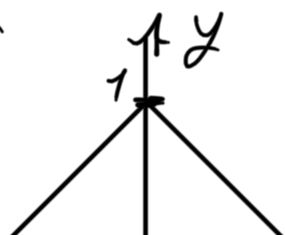
- This has a unique minimum, but no closed-form formula

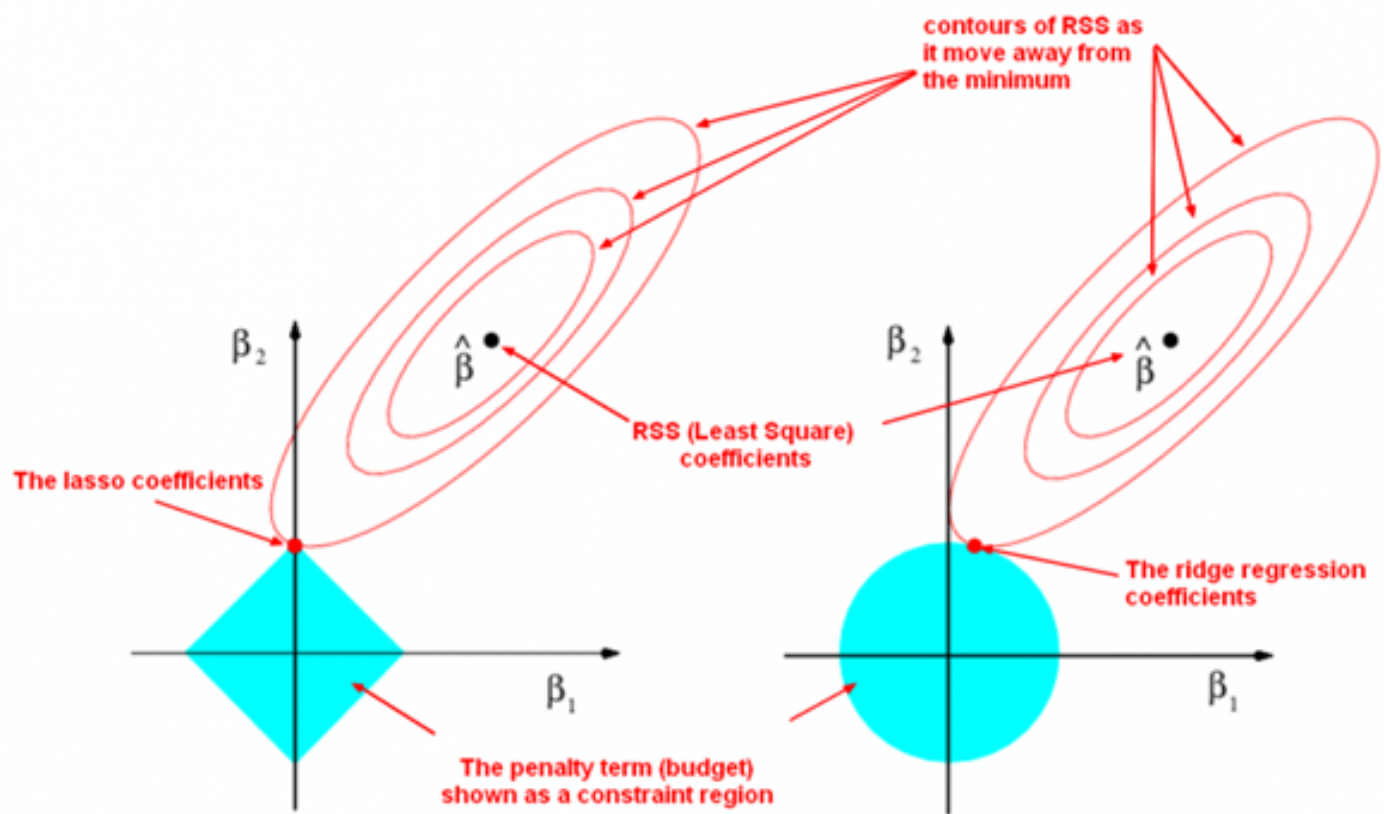
- Usually, use steepest descent method

- Find λ using CV

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

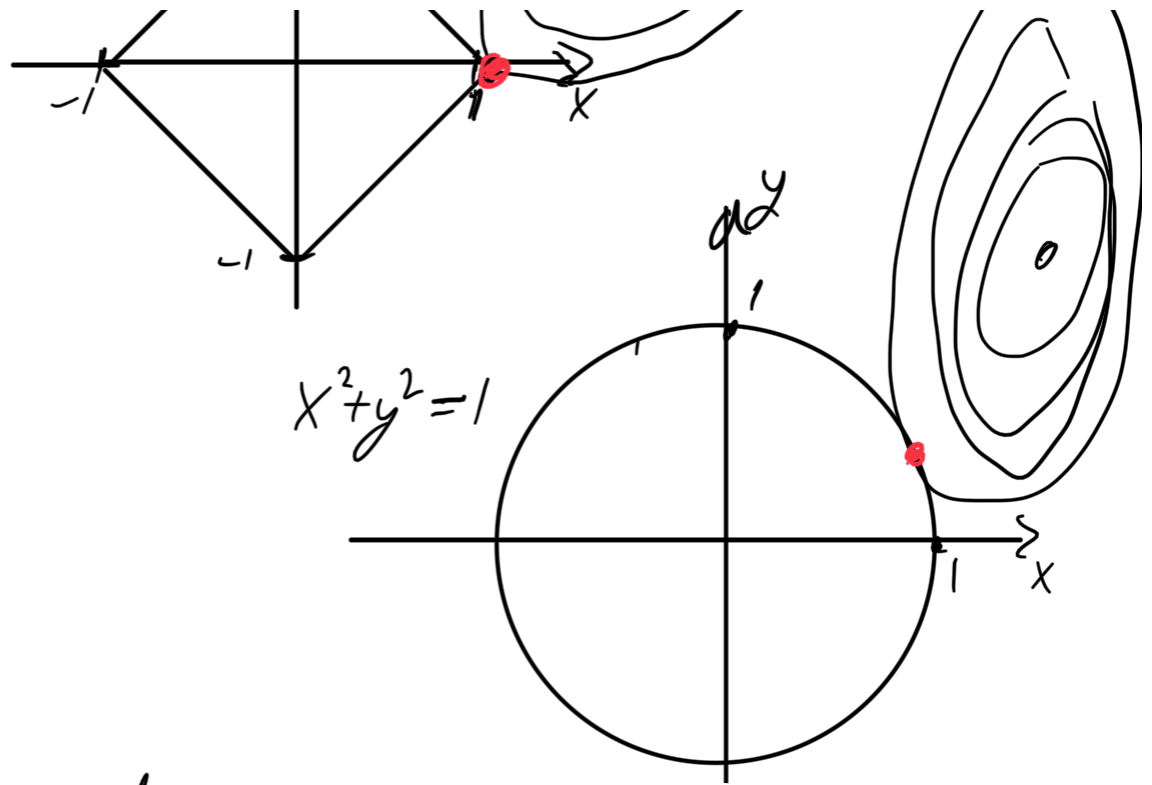
$$|x| + |y| = 1$$





LASSO

RIDGE REGRESSION



- sacrifices unbiasedness for less variance.

Violation of Assumptions

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

What can go wrong:

- bias $E(Y) \neq z\beta$

also ...

= non-normality

= Heteroscedasticity: Variances are not common across observations:
$$\text{Var}(Y) = \sigma^2 V \quad V \neq I_n$$

Bias

Missing Information/Variable

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

= If we regress y on Z :

$$\begin{aligned} \beta_{OLS} &= (Z^T Z)^{-1} Z^T (Z\beta + \tilde{z} \cdot \tilde{\beta}) \\ &= \hat{\beta} + (Z^T Z)^{-1} Z^T \tilde{z} \cdot \tilde{\beta} \end{aligned}$$

= if $Z^T \tilde{z} = 0$, then we get the usual $\hat{\beta}$, 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 $\hat{\epsilon}_i$ vs \tilde{z} and look for linear relationship
 - Better: plot $\hat{\epsilon}_i$ vs the residuals at \tilde{z} on Z (added variable plot)
(no need to regress $\hat{\epsilon}_i$ on \tilde{z})
(bc. they are orthogonal)
- In practice, you should check all administrative things that data may depend on: $\hat{\epsilon}_i$ vs:
 i , calendar time, file order/number, trends, blocking.



Non - Normality

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

Conditions for CLT:


- 1) $\lambda_{\min}(Z^T Z) \rightarrow \infty$
minimal eigenvalues $Z^T Z$ is a matrix version of n
- 2) No Z_{ij} is "too large"
- 3) ε_i is not heavy tailed

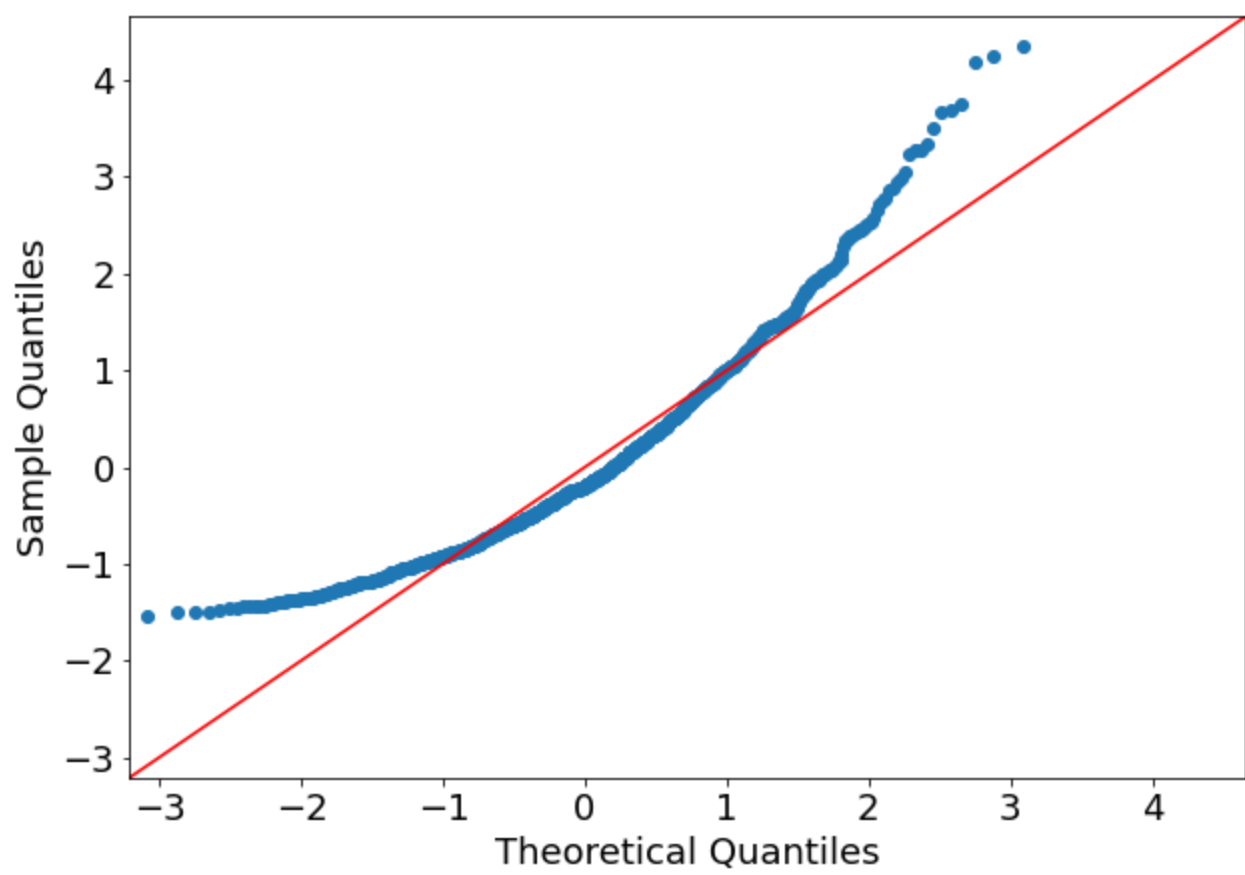
Detection

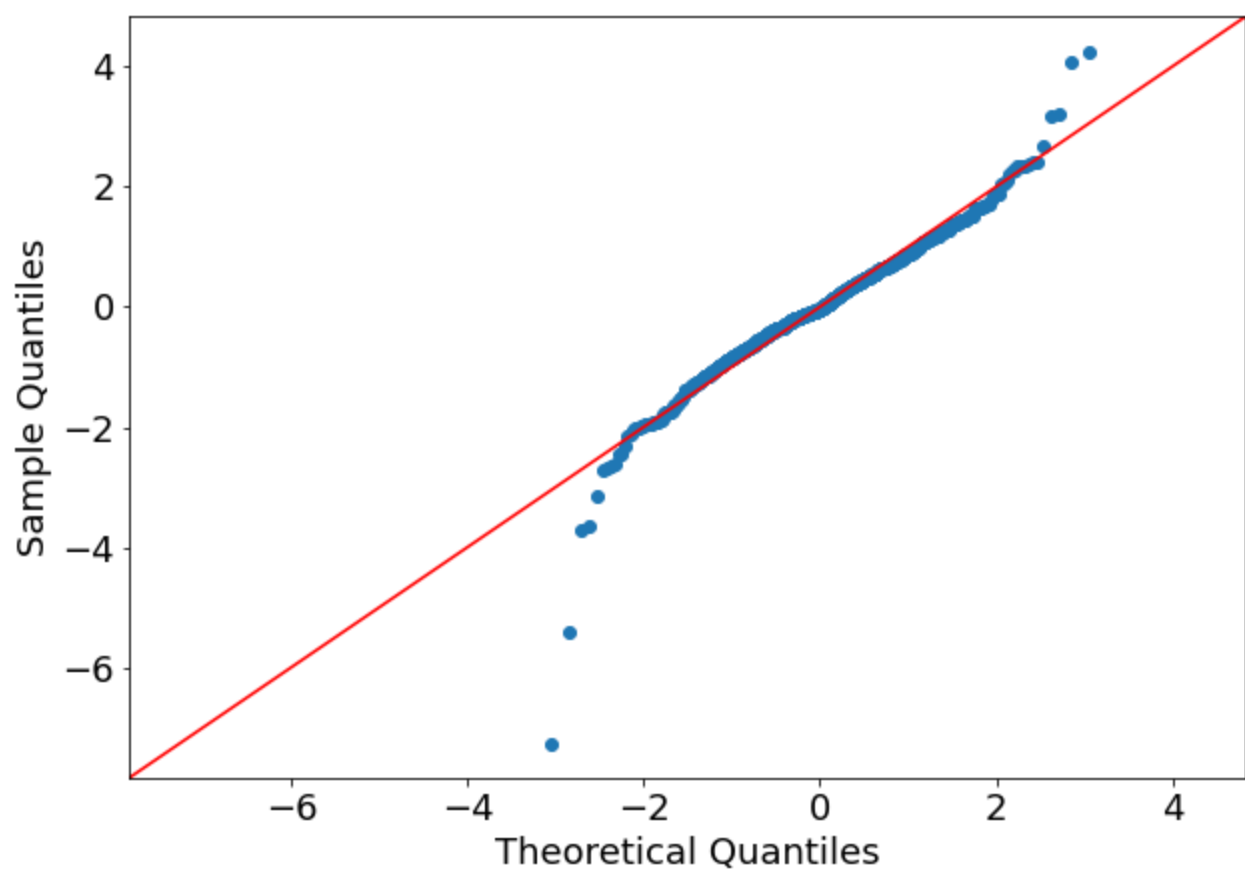
eyeball approach: QQ - Plot:

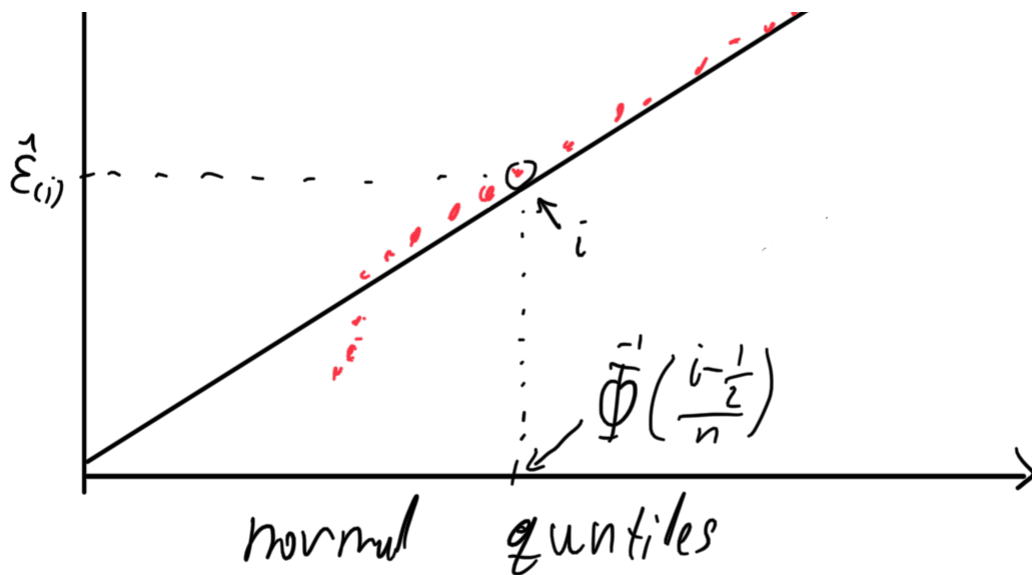
- 1) Sort $\hat{\varepsilon}_i$ so that $\hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \dots \leq \hat{\varepsilon}_{(n)}$
- 2) Plot $\hat{\varepsilon}_{(i)}$ vs. $\Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$

sample
quantiles
↑

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




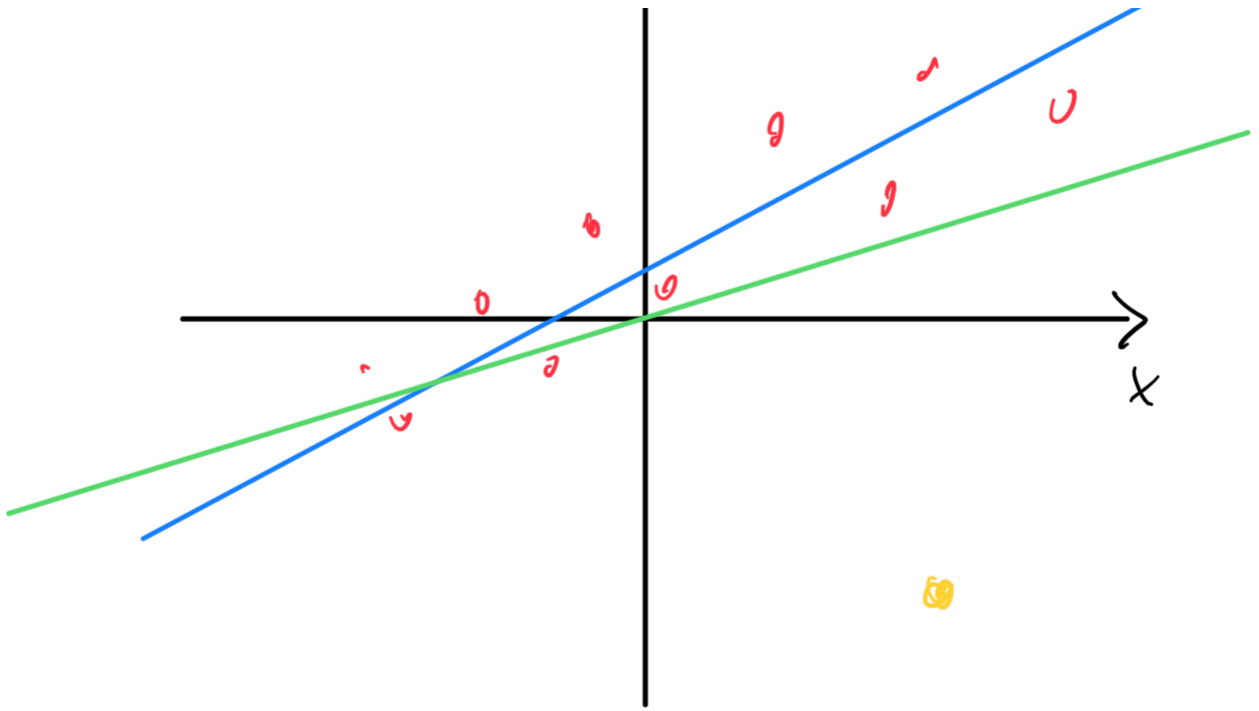




- Ideally, the plot follows a straight line where the slope is σ and the intercept is μ (which is zero with residuals)
- = There are many tests for normality of ε : Jarque-Bera, Anderson-Darling, Kolmogorov-Smirnov
- Non-normality is usually an issue only when it comes to outliers

Outliers

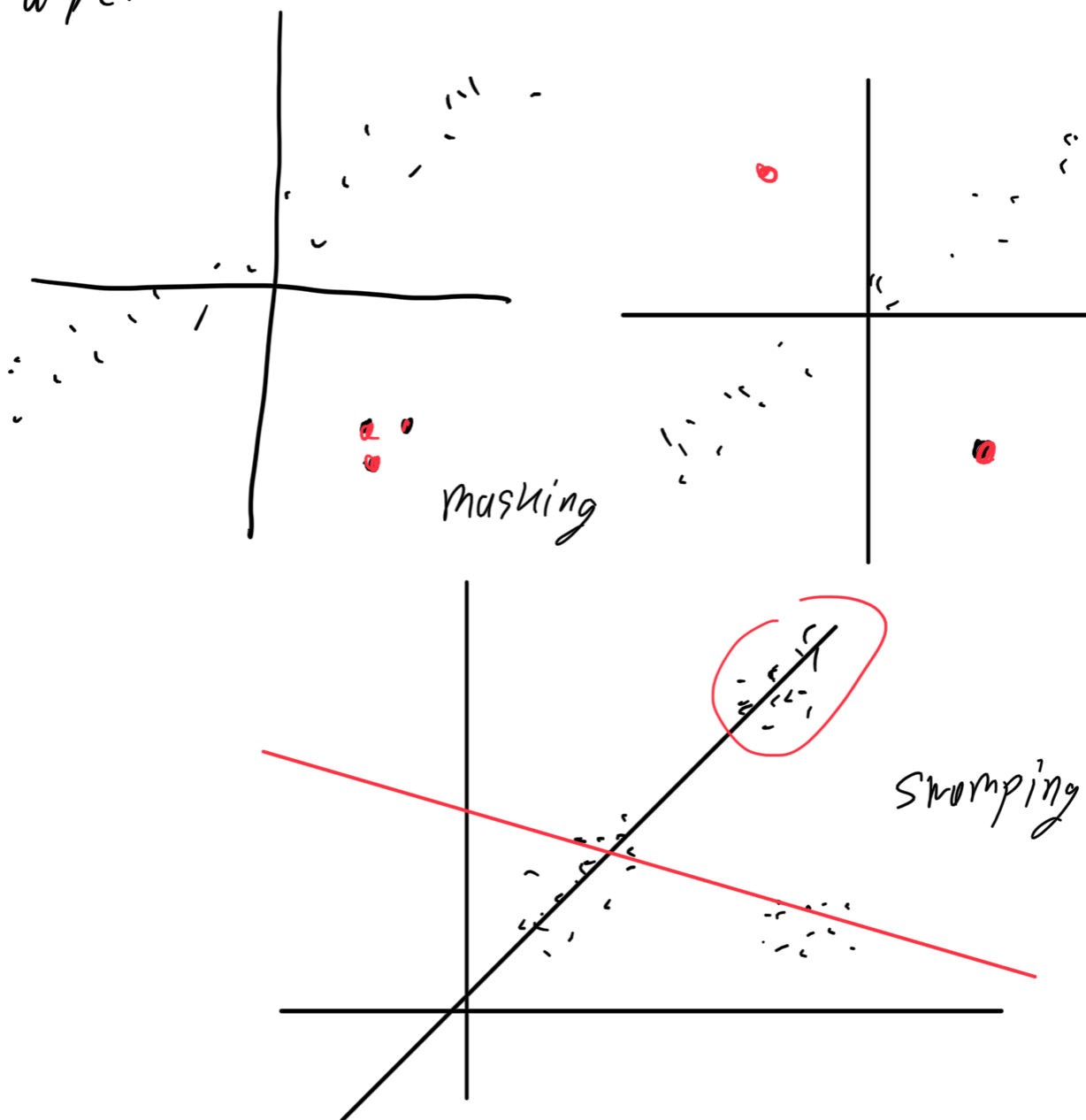
\uparrow



Detection

- A much larger $|\hat{\epsilon}_i|$ than the rest.
- Better: look at $\frac{\hat{\epsilon}_i^{(i)}}{s^{(i)}}$ (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:

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

r

Example 2:
Least trimmed regression:

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

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

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

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