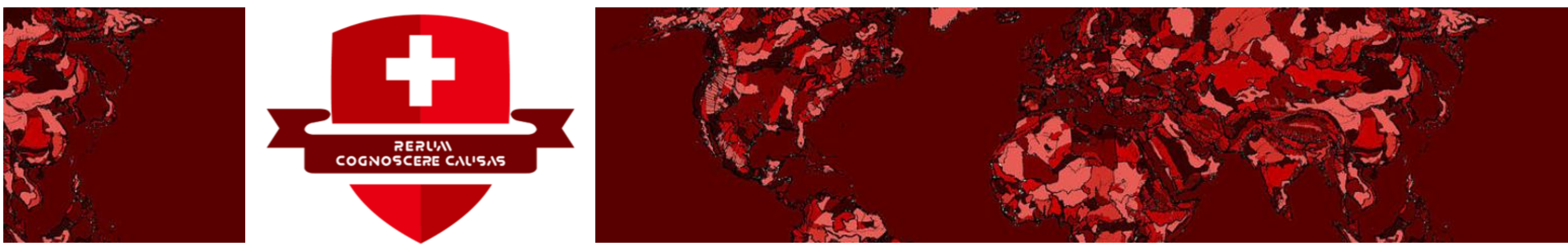
Insert the class / client / audience logo here

Math and Stats for MBA

Lecture Note 2

Swiss Institute of
Artificial Intelligence



Review on Regression Analysis

Swiss Institute of
Artificial Intelligence

Review

- Basic of regression analysis

Regression equations

■ **Simple regression**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

■ **Multiple regression**

– In scalar form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} + \epsilon_i$$

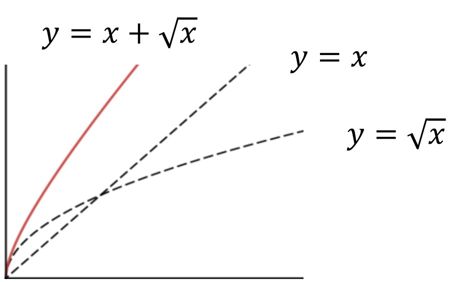
– In matrix form

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

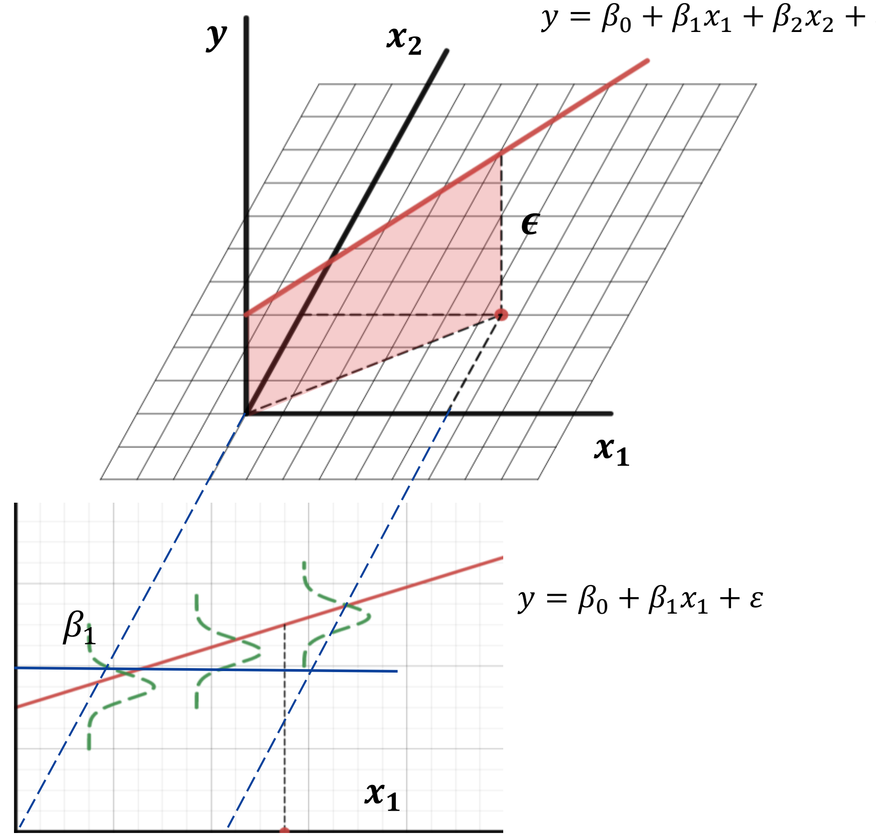
■ **Polynomial regression**

– For non-linear relationship between exploratory variables and dependent variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$



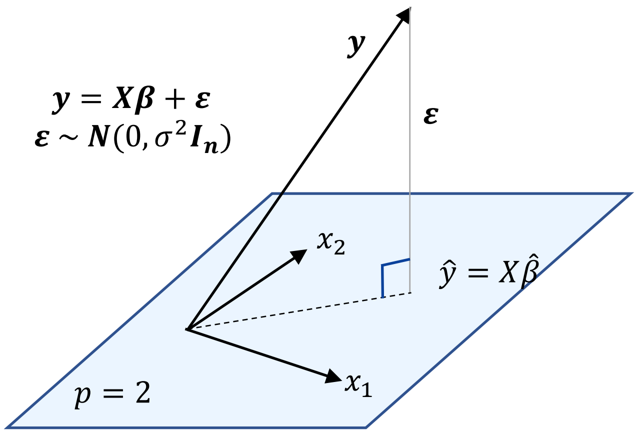
Graphical view of linear regression



Review

- Understanding vector space and projection metrics

Projection matrix



Ordinary Least square estimators

$$\underset{\boldsymbol{\beta}}{argmin}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2, \text{ where } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$$

$$\varepsilon = Y - X\beta$$

$$(Y - X\beta)'(Y - X\beta) = \varepsilon'\varepsilon$$

$$\begin{aligned}
\hat{\epsilon} &= y - \hat{y} \\
&= y - X\hat{\beta} \\
&= y - \overbrace{(X(X^TX)^{-1}X^Ty)}^{=H} \\
&= (I - H)y \\
&= (I - H)(X\beta + \epsilon) \\
&= (I - H)X\beta + (I - H)\epsilon \\
&= (I - X(X^TX)^{-1}X^T)X\beta + (I - H)\epsilon \\
&= (\underbrace{X - X(X^TX)^{-1}X^TX}_{=0})\beta + (I - H)\epsilon \\
&= (I - H)\epsilon
\end{aligned}$$

Note that $\boldsymbol{H}$ matrix is

- *symmetric*
- *idempotent*
  $(\boldsymbol{HH} = \boldsymbol{H})$
- *positive semi-definite*
  $(\boldsymbol{H} \geq \boldsymbol{0})$

Two variables in a same vector space

- There are interest rate data and credit data.

No	Interest Rate(%)
1	2.656
2	2.473
3	3.625
…	…
998	3.163
999	3.762

No	Credit(pts)
1	833
2	836
3	847
…	…
998	746
999	936

- Data A and B look very different from each other, but they represent same information (i.e. $A = X, B = X(X^TX)^{-1}X^T$)
- The ways to verify that data is in the same vector space
  - Is it Linear independence?
  - Is there no inverse matrix?
  - Is there multicollinearity?
- In order to explain according to the situation, it is necessary to understand the data.

collect data	→	abstraction from data	→	Logical thinking	→	Making a decision
A		$\lvert A \rvert = 0$		$same\ data!$		$\cancel{A}$

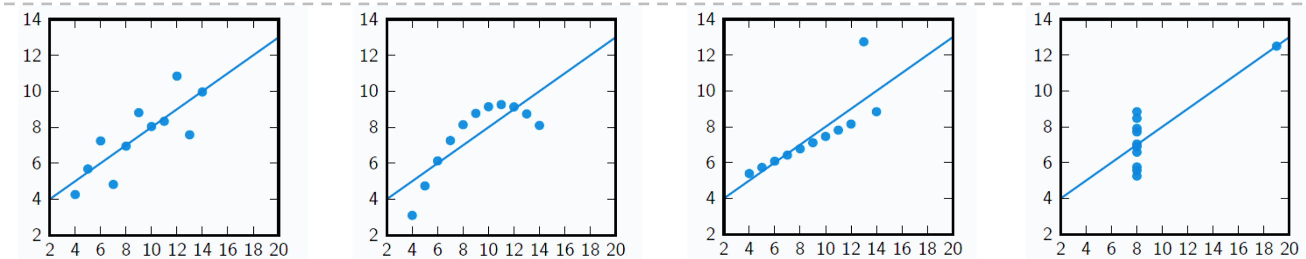
Regression Diagnostics and Advanced Regression Topics

Swiss Institute of Artificial Intelligence

Residual Analysis

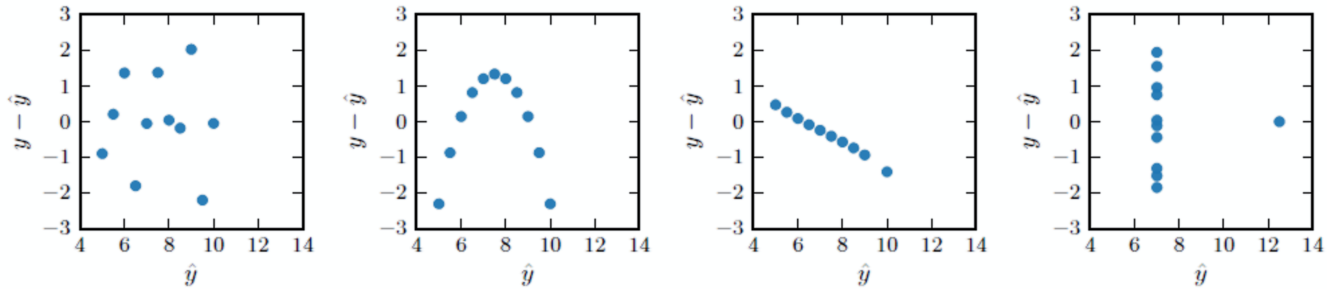
- Four datasets with very similar statistical properties

Linear regression lines fitted



Recall that four datasets have same moment information

Residuals vs. fitted value ($y_i - \hat{y}_i$ vs. $\hat{y}_i$)



Note that four plots show how much above and below the fitted line the data points are

ANSCOMBE'S QUATET (REVISITED)

- **Residuals vs. fitted value ($y_i - \hat{y}_i$ vs. $\hat{y}_i$)**
  - – The residual do not look anything like random noise.
  - – Hence, a linear fit is not appropriate for dataset 2, 3, and 4.
- **Pattern in residual plot**
  - – If there are pattern in error, the model lacks variables describing the dependent variable.
  - – *Mis-specification* model has *mis-specification error* of $\hat{\epsilon}$

What should we do for this case?

From the residual pattern, we can figure out the problem of the model.

- **Mis-specification case** : Append a new exploratory variable that can offset the pattern in residual plot. For instance, a squared exploratory variable into the model of second plot above.
- **Outlier case** : Filter out the outliers in the dataset. By removing one data point in the model of third and fourth plot above, the regression will be fitted almost perfectly.

Residual Analysis

- Gauss-Markov Assumption

Independence between X and ϵ

- **The assumption on X and ϵ**

$$Y = X\beta + \epsilon$$
$$X^T Y = X^T X\beta + X^T\epsilon$$
$$(X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T\epsilon$$
$$= \beta + (X^T X)^{-1} X^T\epsilon$$

(1) X is fixed so that we have:

$$E[\hat{\beta}] = E[\beta] + E\left[(X^T X)^{-1} X^T\epsilon\right]$$
$$= \beta + (X^T X)^{-1} X^T E[\epsilon]$$

(2) X is stochastic but independent of ϵ so that we have:

$$E[\hat{\beta}] = E[\beta] + E\left[(X^T X)^{-1} X^T\epsilon\right]$$
$$= \beta + (X^T X)^{-1} E[X^T\epsilon] \quad \text{where } E[X_T\epsilon] = 0$$

- X and ϵ should not be related to each other. If the ϵ value changes with X value, it is difficult to minimize ϵ.

Appendix – Gauss-Markov Assumption

The standard Gauss-Markov Assumptions are:

- **A1 : Linearity**
  - $\mathbf{y} = \mathbf{X}\beta + \epsilon$
  - This assumption states that there is a linear relationship between $\mathbf{y}$ and $\mathbf{X}$
- **A2 : Full rank**
  - X is a full rank matrix
  - This assumption states that there is no perfect multicollinearity.
  - This assumption is known as the identification condition.
- **A3 : Zero conditional mean**
  - $E[\epsilon|X] = 0$ (**A3F,** for fixed sample)
  - This assumption states that the disturbances average out to $\mathbf{0}$ for any value of X
  - $E[X'\epsilon] = 0$ (**A3Rsru**) , $E[\epsilon_i|X_1, X_2, \dots, X_n] = 0$ (**A3Rmi**)
- **A4: Homoskedasticity and no autocorrelation**
  - $Var(\epsilon_i) = \sigma^2 < \infty$, for $\forall_i$ and $Cov(\epsilon_i, \epsilon_j) = 0$, $\forall_i \neq j$
  - This assumption states assumption of homoskedasticity and no autocorrelation
- **A5: Normality condition**
  - $\epsilon_i \sim iid\ N(0, \sigma^2)$

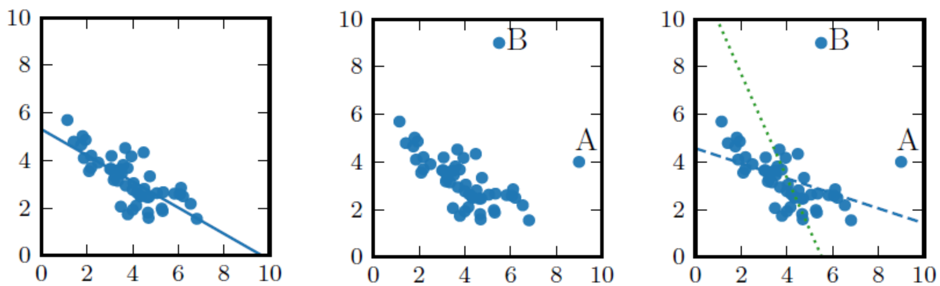
Outliers

- Analysis of outliers

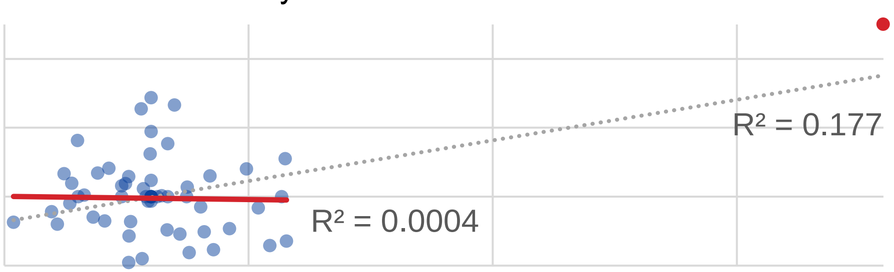
Outliers and Influential Points

Real life datasets often have some unexpected points that have significantly different aspects from the rest of the data.

- **Why outliers matter?**
  - Outliers distort the property of data



- **Influential points**
  - In the regression model, the influential points significantly affect the coefficients as well as R^2 and therefore mislead the researcher's analysis



Diagnosis of outliers

Diagnosis of influential outliers is based on the error term

- **Leverage index**

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- **Cook's distance**

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{p \cdot MSE} = \frac{1}{p \cdot MSE}\frac{H_{ii}}{(1 - Hii)^2}\hat{\epsilon}_i{}^2$$

There are several indices to detect outliers / anomalies

Index	Description	Cutoff value
Leverage (hat index)	Measure how far each observation point far from the mean of dependent variable	$2(k-1)/n$ $\sim 3(k-1)/n$
Standard error	Residual from fitted regression model	$2 \sim 3$
Cook's distance	Measure combining leverage and residual	$\frac{4}{n}$
DFFITS	Measure the change between restrict model and unrestricted model	1~2 or $2\sqrt{(k-1)/n}$
DFBETAS	Measure the change of each coefficient for restricted model and unrestricted model	$2/\sqrt{n}$

Source: Gorden, R.A. (2010), Regression Analysis for the Social Science, p. 367.

Advanced Regression: Robustness

- Optimization view of robustness

Various types of error

Suppose we tried to adjust the optimization problem:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X\beta)^2 = \sum_{i=1}^{n} \rho(r_i)$$

where $r_i = (y_i - X_i\beta)$ is the residual and $\rho(r) = r^2$ is squared error function. Recall that the squared error gives very large penalties on large error. (i.e., $\rho(2) = 4,\ \rho(10) = 100$)

Solution to the optimization problem

If the model is too sensitive to errors, we can consider a different function $\rho(\cdot)$ other than $\rho(r) = r^2$.

- **LAD (Least Absolute Deviations)**
  – Often used when the dataset follows Laplacian distribution

$$\boldsymbol{\rho(r) = |r|}$$

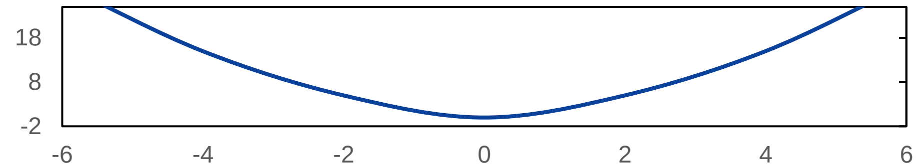
- **Huber**
  – Similar to LAD. Huber function can be differentiable at $r = 0$

$$\boldsymbol{\rho(r) = \begin{cases} r^2/2\,, & |r| < k \\ k(|x| - k/2), & |r| \geq k \end{cases}}$$

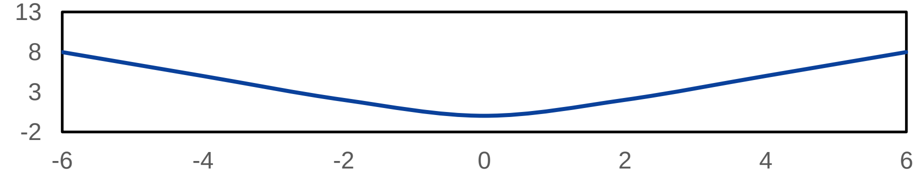
- **Bisquare**
  – Similar to squared loss. It can level off a certain point.

Loss functions

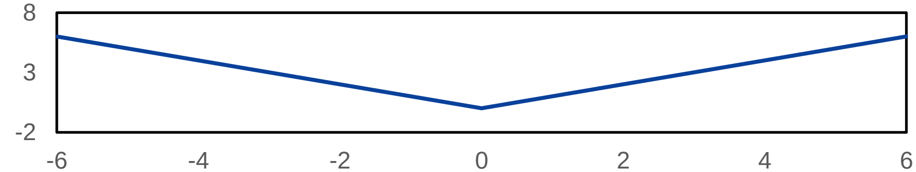
Squared error



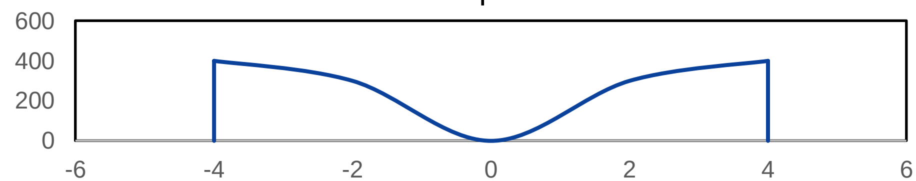
Huber



LAD



Bisquare



(Note that the difference in y-axis scales)

This kind of functions also called '*Kernel*' or '*Activation function*'

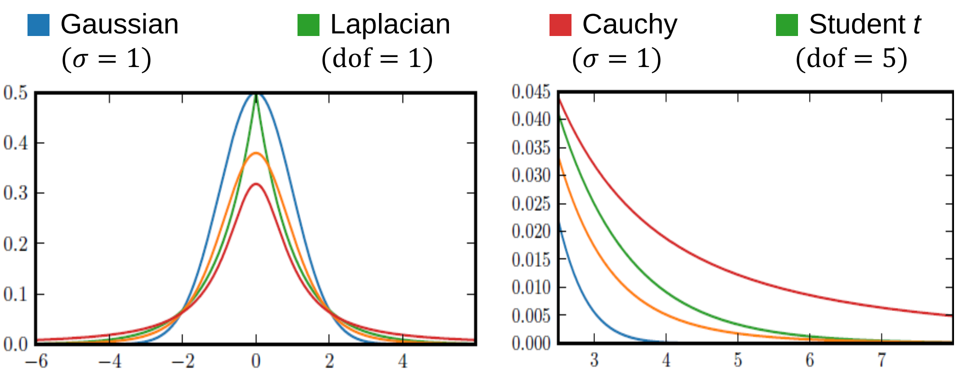
Advanced Regression: Robustness

- Distribution View of Robustness

Distributions insensitive to outliers or extreme points

How to make the model less sensitive to outliers

One way to be less sensitive to outliers is to assume distribution with heavier tails: assigning higher probability to improbable events.



Student *t* distribution, the Laplacian distribution, the Cauchy distribution, and any power-law distribution all have **heavier tails** than the Gaussian we usually assume effectively.

Any other heavy-tailed distribution?

- **One-tailed**
  - Pareto, Log-normal, Weibull, log-logistic, log-gamma, Half-Cauchy, log-Cauchy
- **Two-tailed**
  - Cauchy, Student *t*, Laplacian(heavier than Gaussian)

RANSAC

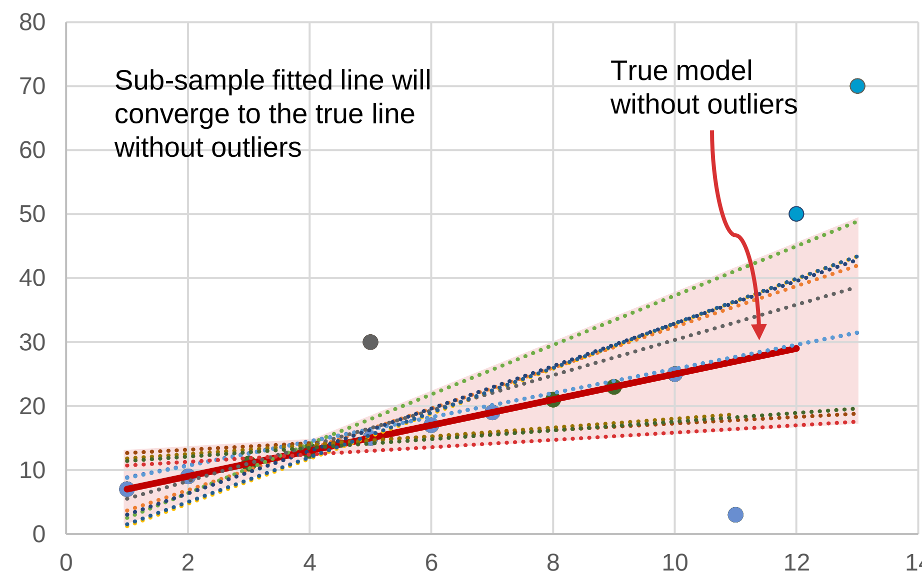
RANdom Sample Consensus (RANSAC)

The basic assumption of RANSAC is just that the data consists primarily of non-outliers.

(repeated)

Randomly select subsamples → Fit a model → Find points having less error than α → Fit a model again

dashed lines are fitted models from randomly selected samples



Advanced Regression: Sparsity

- Ridge regression and Lasso regression

Ridge regression

- **Regularization**
  - Adding in a sparsity constraint in these settings often helps prevent overfitting, and leads to simpler, more interpretable models.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_i^2 + \underbrace{\beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5}_{\text{Ignore high-order terms}} + \epsilon_i$$

- **Ridge regression (L2 regularization)**
  - The coefficient to be close to zero due to a regularization term

$$\min_{\boldsymbol{\beta}} \left[\underbrace{\sum_{i=1}^{n} (y_i - X_i\beta)^2}_{\text{data term}} + \underbrace{\lambda \sum_{k=1}^{p} \beta_k^2}_{\text{regularization term}} \right]$$

  - The coefficient will have a significant penalty, 2λ

$$\frac{\partial y}{\partial \beta_2} = 2\sum_{i=1}^{n} (y_i - X_2\beta_2)X_2 + 2\lambda\beta_2 = 0$$

$$\therefore \beta_2 = \frac{f(\cdot)}{X_2^2 + 2\lambda} \quad where\ \lambda > 0$$

Matrix form (Ridge)

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

  - The regularization term of ridge regression ($\lambda \sum_{k=1}^{p} \beta_k^2$) would not produce sparsity.

Lasso (Least Absolute Shrinkage and Selection Operator)

- **Lasso regression (L1 regularization)**
  - Different from ridge regression, penalize non-sparsity directly

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \underbrace{\sum_{k=1}^{p} \mathbb{I}(\beta_k^2 \neq 0)}_{\#\text{ of non-zeros in } \beta} \right]$$

Approximate

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right]$$

  - Lasso gives a solution as sparse as possible
- **A Bayesian view on ridge regression**
  - Bayesian updates hypothesis by adding new data

$$\boldsymbol{p}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) \propto \boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{X},\boldsymbol{y}|\ \boldsymbol{\beta})$$

  - Taking log on both sides above, then we get:

$$\underbrace{\boldsymbol{ln}[\boldsymbol{p}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y})]}_{\text{Ridge / Lasso}} \propto \underbrace{\boldsymbol{ln}[\boldsymbol{p}(\boldsymbol{\beta})]}_{\text{Prior (Penalty term)}} + \underbrace{\boldsymbol{ln}[\boldsymbol{p}(\boldsymbol{X},\boldsymbol{y}|\ \boldsymbol{\beta})]}_{\min(\hat{\epsilon})^2}$$

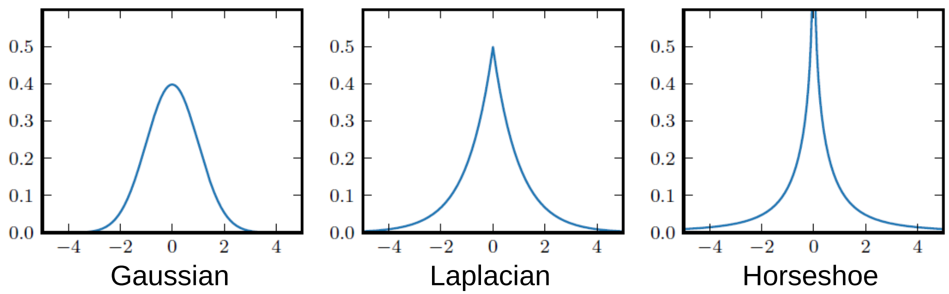
  - The left hand-side will become
    - *Lasso* if prior follows Laplacian
    - *Ridge* if prior follows Gaussian
- **Find a compromise between regularization and optimization**
  - Generalization vs. Overfitting

Advanced Regression: Sparsity

- Ridge regression and Lasso regression

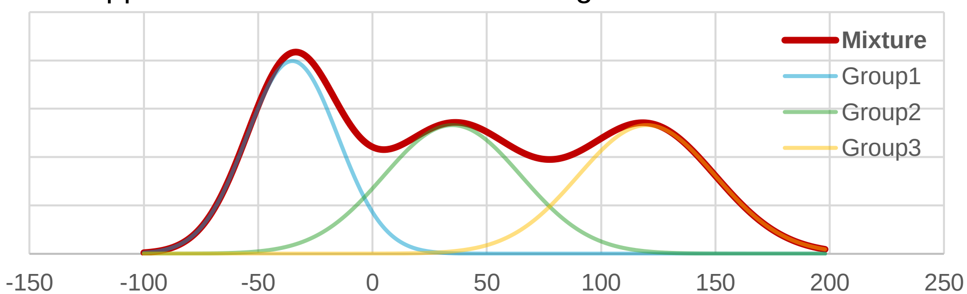
Sparsity and shrinkage: a graphical view

Several coefficient priors for sparse regression



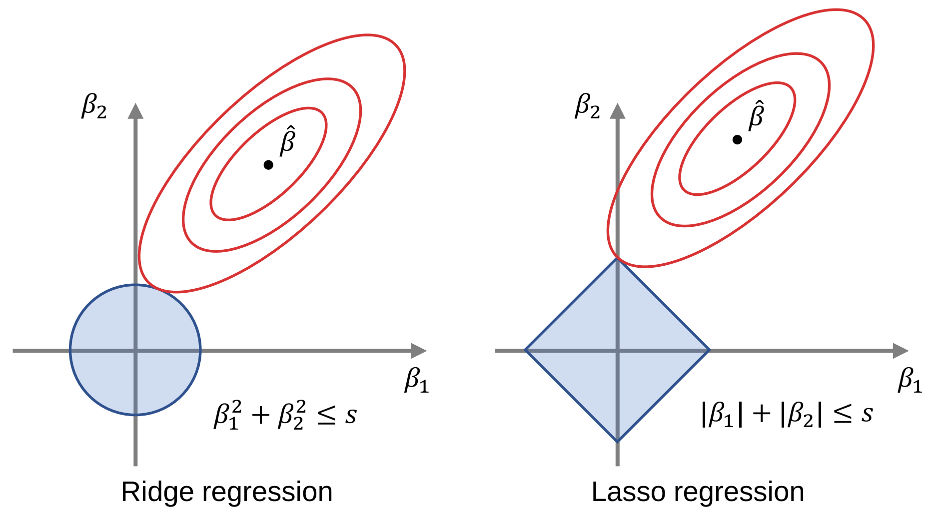
What if the data follows unknown pattern other than plots above?

- **Mixture of distribution**
  - Some dataset have mixture of distribution that can be grouped
  - Suppose we have a dataset following the distribution below:



  - We cannot apply any prior for this distribution (**RED LINE**)
  - However, in the view of mixture model, this can be divided into three Gaussian distribution

Graphical approach



Graphical interpretation on Ridge and Lasso regression

Red contours represent the error (RSS) and blue objects represent the constraints of each regression.

- **General difference between Ridge and Lasso**

Ridge	Lasso
L2-norm regularization	L1-norm regularization
Closed form solution (differentiation)	Numerical optimization
Good performance in presence of collinearity	Model selection property
Tend to shrink a large coefficient first	

Generalized Linear Models

- GLS

Generalized Linear Models

Generalized linear model(GLM) generalizes various forms of regression (i.e. non-linear model) into a general form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 \mathrm{x}_{2,i} + \beta_3 x_{1,i}^2 + \cdots + \epsilon_i$$
$$= f(\boldsymbol{x_i})$$

GLMs are a family of methods that assume the following:

$$\boldsymbol{y} = \boldsymbol{\mu_y} + \boldsymbol{\epsilon} \ where\ \boldsymbol{\mu_y} = \boldsymbol{X\beta}$$
$$\boldsymbol{\mu_y} = \boldsymbol{g^{-1}}(\boldsymbol{X\beta})$$

where $\mathrm{g}(\cdot)$ is called the link function and is usually nonlinear. The interaction between the input $\boldsymbol{X}$ and the parameters $\boldsymbol{\beta}$ remains linear, but the result of that linear interaction is passed through the inverse link function to obtain the output $\boldsymbol{y}$.

- **Link function (Kernel)**
  - There are infinite number of non-linear functions that can be used to explain the output.
  - We can choose a function similar to the dataset we have
- **EXAMPLE – logistic regression**
  - Sigmoid link function can be useful to map a real number to a number between 0 and 1.

$$g^{-1}(\boldsymbol{\eta}) = \frac{1}{1 + exp(-\boldsymbol{\eta})}$$

APPENDIX – Logit & Probit

Why are Logit & Probit model needed and what are those?

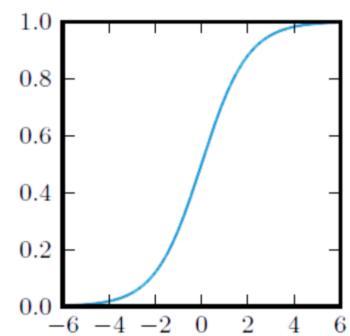
Linear model is hard to explain non-linear relationship between exploratory variables and dependent variable. Hence, we need a new probability model that has two properties:

- The dependent variable is confined between 0 and 1, $\boldsymbol{y} \in (\boldsymbol{0}, \boldsymbol{1})$
- The probability model become slower in change as $\boldsymbol{y}$ is approaching to $\boldsymbol{0}$ or $\boldsymbol{1}$ (sigmoid)
- OLS is not feasible due to the non-linear relationship between link function and $\boldsymbol{\beta}$

- **Logit model**
  - based on Logistic regression
  - $L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_i$

    where $P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}$
- **Probit**
  - based on Normal CDF
  - $P_i = F(I_i) = \frac{1}{\sqrt{2\pi}} \int_0^{I_i} e^{-\frac{z^2}{2}} dz$

    where $I_i = \beta_0 + \beta_1 X_i + \epsilon_i$



A Sigmoid function

Nonparametric statistics and model selection

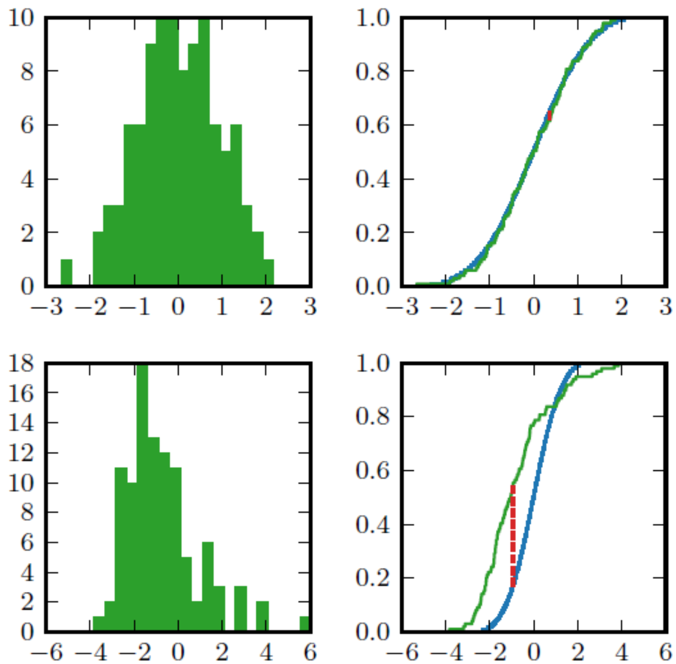
Swiss Institute of
Artificial Intelligence

Estimating distributions and distribution-free tests

- Normality test

Estimating distributions

Comparing two arbitrary distributions



- **There are several tests to check normality by comparing two distributions.**
  - – Kolmogorov-Smirnov test
  - – Sharpiro test
  - – Wilcoxon's signed-rank test
  - – Mann-Whitney *U* test

EXAMPLE – CHICAGO TEACHING SCANDAL

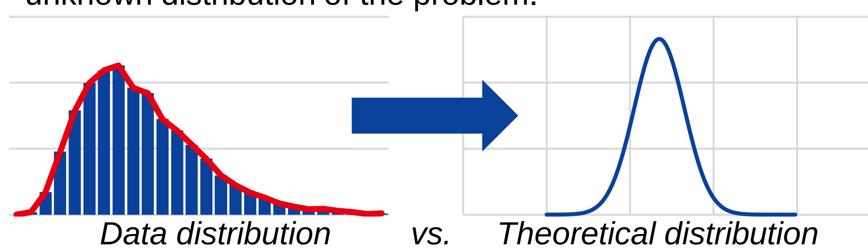
In 2002, economists Steven Levitt and Brian Jacob investigated cheating in Chicago public schools

They went through test scores from thousands of classrooms in Chicago schools, and for each classroom, computed two measures:

- – How unexpected is that classroom's performance?
- – How suspicious are the answer sheets?

- They tried to obtain a null distribution from the dataset and conduct a permutation test to evaluate the values they observed.
- If the distribution cannot be defined well, we can try to divide the data into groups and conduct the analysis

- **How to solve this kind of problem?**
  - – We could not be able to define the distribution of this problem.
  - – In general, we can still find out the distribution similar to the unknown distribution of the problem.



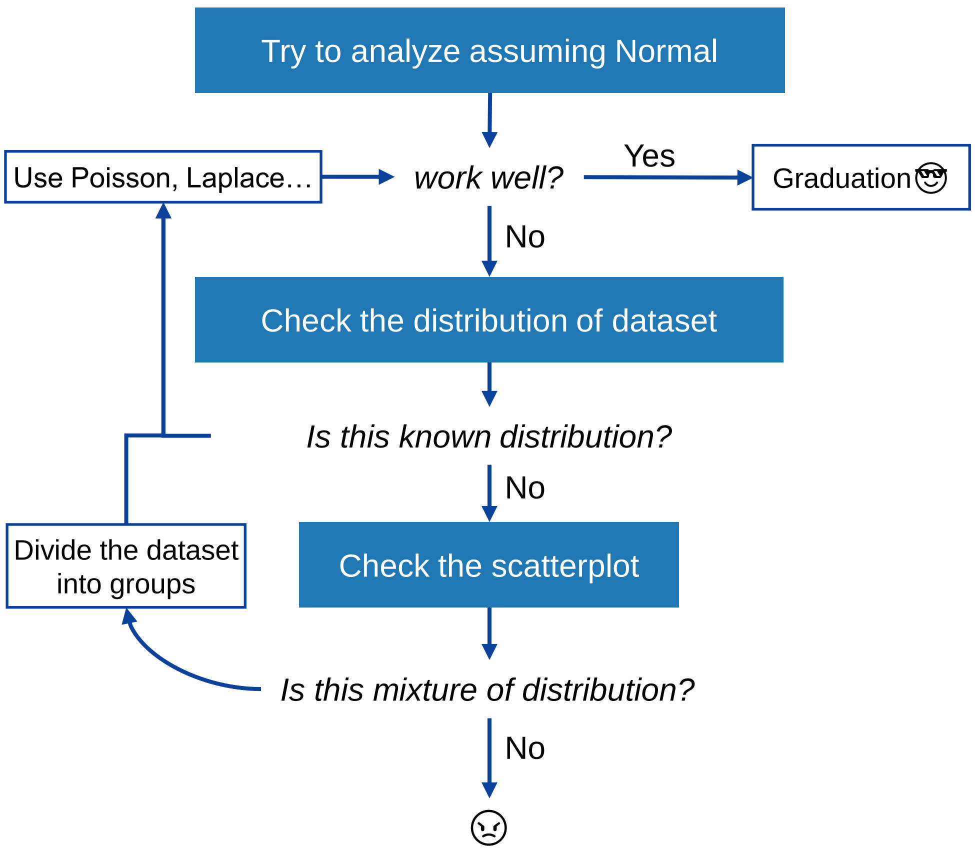
Resampling-based methods

- Model selection

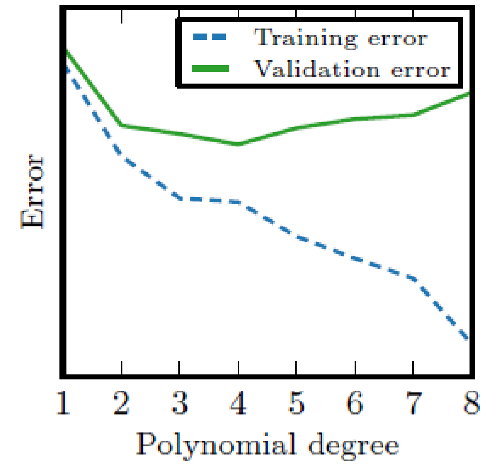
Approaches

General approach to check the distribution

"Eyeballing" can be a good approach to estimate distributions



Model selection



- We should find a compromise between training error and validation error. (regularization vs. optimization)
- In the case of LASSO we discussed in the previous slide, we need to select proper level of λ.
  - – High λ lead to simpler model (sparsity)
  - – Small λ lead to complex model
- There is no answer for the level of degree. (*No rule-of-thumbs*)
- In addition, error varies in every trial. Therefore, the result of the plot above is not guaranteed to every dataset.