

PS5841

Data Science in Finance & Insurance

Linear Regression

Yubo Wang

Autumn 2021

Linear Regression Model

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

- where

$$\mathbf{y} = [Y_1, \dots, Y_n]^T,$$

Y_1, \dots, Y_n are independent RVs with

$$Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

$$\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_n^T]^T, \boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^T, \boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_n]^T$$

$$E(\mathbf{y}) = \boldsymbol{\mu}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$$

$$\text{Var}(\mathbf{y}) = E[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T] = \boldsymbol{\Sigma}$$

can accommodate heteroscedasticity

Least Squares Estimation

- Minimize

$$S_w = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Set

$$\frac{\partial S_w}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

de-emphasize high variance RV's

Get

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

Least Squares Estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

- If Y_i 's are homoscedastic $\boldsymbol{\Sigma} = \mathbf{I}\sigma^2$, LSE is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

with

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{y}) = \boldsymbol{\beta}$$
$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Var(\mathbf{y}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n - p} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Coefficient of Determination

- RSS of the model

$$\hat{S} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}$$

- \hat{S}_0 , the worst value of RSS, is associated with the naïve model $E(Y_i) = \mu, \forall i$

$$\hat{S}_0 = \mathbf{y}^T \mathbf{y} - N\bar{y}^2 = \sum (y_i - \bar{y})^2 \propto \text{Var}(y)$$

- Relative improvement

$$R^2 = \frac{\hat{S}_0 - \hat{S}}{\hat{S}_0} = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - N\bar{y}^2}{\mathbf{y}^T \mathbf{y} - N\bar{y}^2}$$

Proportion of the total variation in the data which is explained by the model

The Naïve Model

$$E(Y_i) = \mu, \forall i$$

$$\hat{\beta} = \hat{\mu} = \bar{y}$$

$$X = \mathbf{j}$$

$$\hat{S}_0 = \mathbf{y}^T \mathbf{y} - \hat{\beta} X^T \mathbf{y}$$

$$= \mathbf{y}^T \mathbf{y} - \bar{y} \Sigma y_i$$

$$= \mathbf{y}^T \mathbf{y} - N \bar{y}^2$$

$$= \Sigma (y_i - \bar{y})^2 \propto \text{Var}(y)$$

Gauss-Markov Theorem

- When $\Sigma = I\sigma^2$, the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ is BLUE

Categorical Features

- Categorical features are often modeled by binary (dummy) variables in a regression
- For a feature (factor) with J levels
 - Need J binary variables if there is no intercept
 - Need (J-1) binary variables if intercept
 - Baseline is the level with no dummy variable
- Example: 1 factor with 3 levels (A,B,C)
 - with baseline A (A as the reference level)
$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$
 - How are $\hat{\beta}_j$'s interpreted?

Example (1)

- 1 factor with 3 levels (A,B,C)
 - with baseline A (A as the reference level)

$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

- How are $\hat{\beta}_j$'s interpreted?

Example (2)

- 1 factor with 3 levels (A,B,C)
 - with baseline A (A as the reference level)

$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

- If level A is present, all else being equal

$$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

- If level B is present, all else being equal

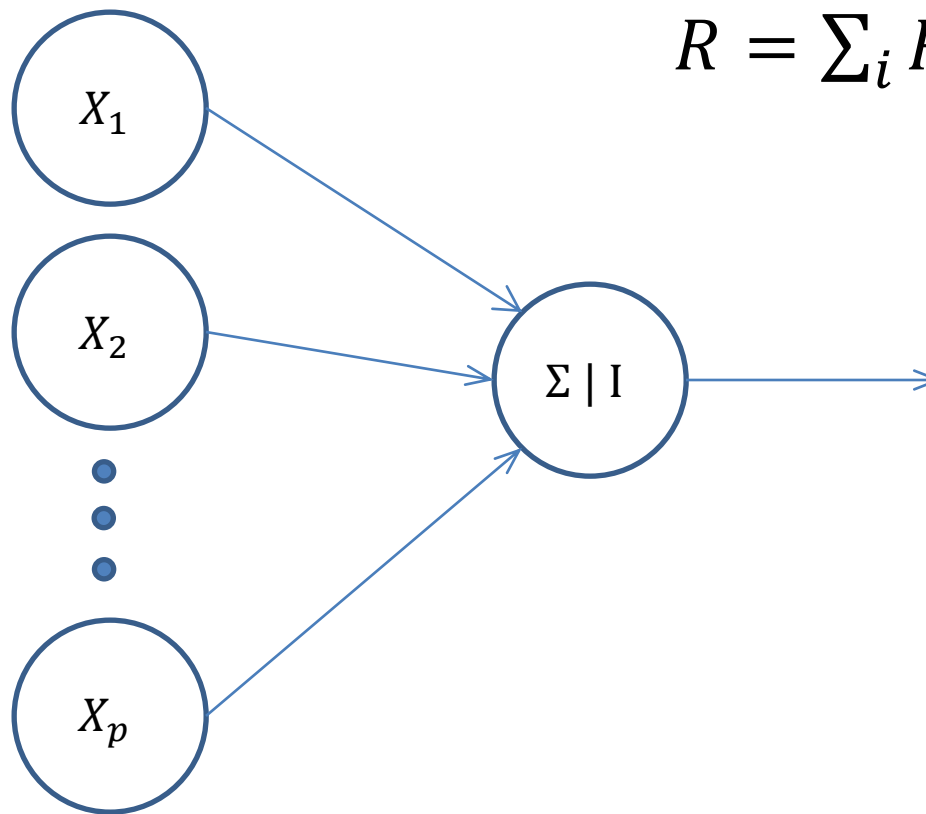
$$\hat{y}_B = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

- Interpretation of $\hat{\beta}_1$

$$\hat{y}_B - \hat{y}_A = \hat{\beta}_1$$

Linear Regression with a Neuron

$$R_i = (y_i - \hat{y}_i)^2$$
$$R = \sum_i R_i$$



Gradient Descent

- Gradient operator (del operator)

$$\nabla = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \right)^T$$

- when operating on a real valued function
 $f: \mathbb{R}^n \rightarrow \mathbb{R}$

$$\nabla(f) = \nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^T$$

Gradient

- For $f: \mathbb{R}^n \rightarrow \mathbb{R}$, differentiable at \mathbf{x} , the gradient of f at \mathbf{x} is

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)^T$$

- When $\nabla f(\mathbf{x}) \neq \mathbf{0}$
 - the maximum rate of increase of f is $|\nabla f(\mathbf{x})|$ and is in the direction of $\nabla f(\mathbf{x})$
 - the maximum rate of decrease of f is $|\nabla f(\mathbf{x})|$ and is in the direction of $-\nabla f(\mathbf{x})$

Example (1)

- $f(\mathbf{x}) = x_1 + x_2^2, \nabla f(\mathbf{x}) = \begin{pmatrix} 1 \\ 2x_2 \end{pmatrix}$
- Moving d units from $\mathbf{x}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ in the direction of $\nabla f(1,1) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, |\nabla f(1,1)| = \sqrt{5}$

lands at

$$\mathbf{x}_1 = \begin{pmatrix} 1 + \frac{d}{\sqrt{5}} \\ 1 + \frac{2d}{\sqrt{5}} \end{pmatrix}$$

Example (2)

Same as
predicted
rate

gradient ascent						
	$f[1+(1/\sqrt{5})d,$ $d \ 1+(2/\sqrt{5})d]$	$f(1,1)$	chg (f)	$\sqrt{5}*d$	error	
1	5.036067977	2	3.036068	2.23606798	0.8	
0.1	2.231606798	2	0.231607	0.2236068	0.008	
0.01	2.02244068	2	0.022441	0.02236068	8E-05	
non-optimal direction						
	$f[1+(1/\sqrt{2})d,$ $d \ 1+(1/\sqrt{2})d]$	$f(1,1)$	chg (f)	See! It's true!		
1	4.621320344	2	2.62132	<	3.036068	
0.1	2.217132034	2	0.217132	<	0.231607	
0.01	2.021263203	2	0.021263	<	0.022441	

Gradient Descent

- To minimize the objective function

$$R(\boldsymbol{\beta}) = \sum_{i=1}^n R_i(\boldsymbol{\beta})$$

with learning rate $\eta > 0$, updating involves all observations

$$\begin{aligned}\boldsymbol{\beta}^{(r+1)} &= \boldsymbol{\beta}^{(r)} - \eta \nabla R(\boldsymbol{\beta}^{(r)}) \\ &= \boldsymbol{\beta}^{(r)} - \eta \sum_{i=1}^n \nabla R_i(\boldsymbol{\beta}^{(r)})\end{aligned}$$

Stochastic Gradient Descent

- SGD is a stochastic approximation of GD. SGD uses randomly selected samples to evaluate the gradients
- At extreme, updating would involve only a single observation

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \eta \nabla R_i(\boldsymbol{\beta}^{(r)})$$

Normal Linear Model

- $Y_i \sim N(\mu_i, \sigma^2)$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$$

$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \sim \frac{\sigma^2}{n-p} \chi_{n-p}^2$$

NLM: Inference (3)

- Variance of fit at \mathbf{x}

$$\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} \sigma^2$$

- Variance of prediction at \mathbf{x}

$$[1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}] \sigma^2$$

General Linear Model

- $Y_i \sim N(\mu_i, \sigma^2)$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$
$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$$

- Features (Explanatory Variables, Independent Variables) can come from different sources
 - Quantitative or their transformations
 - Basis expansions (e.g. polynomial regression)
 - Numeric or dummy coding of categorical levels
 - Interactions between variables (e.g. $X_1 X_2$)
- Linear regression, ANOVA, ANCOVA

Power Transforms

- To make the assumption of normality (if desired) more plausible

Strictly Positive Data

- Box-Cox family of transforms ($y > 0$)

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

- Estimate λ via maximum likelihood
- In practice,
 - Typically, $\lambda = 1, 0.5, 0, -1$
 - No need to -1 and $\div \lambda$ for operations unaffected by location and scale shifts (e.g. some regressions)

General Data

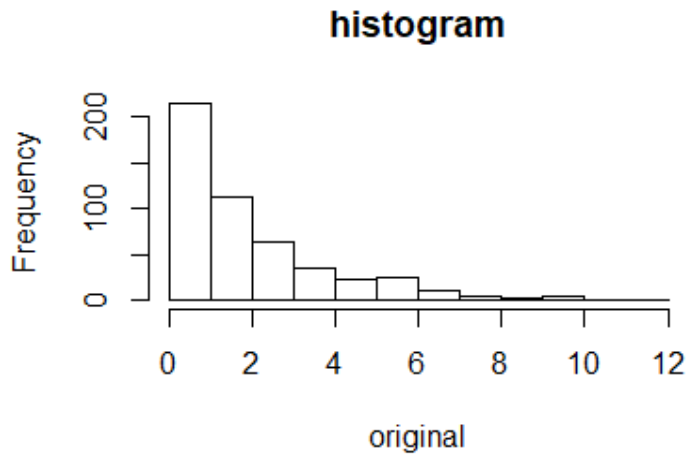
- Box-Cox family of transforms ($y > -\alpha$)

$$y^{(\lambda)} = \begin{cases} \frac{(y + \alpha)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y + \alpha), & \lambda = 0 \end{cases}$$

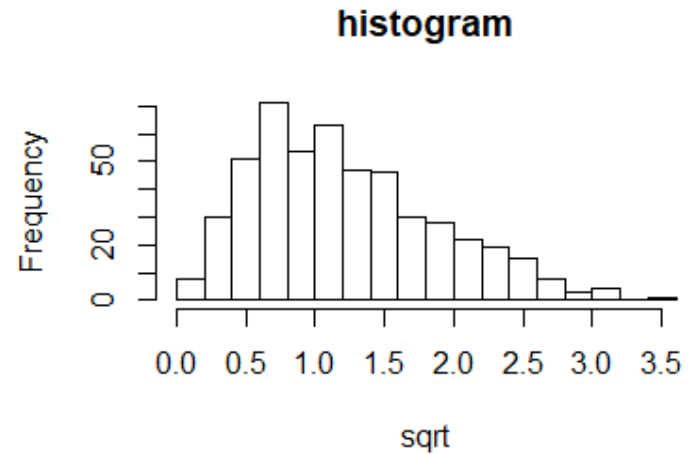
- Yeo-Johnson family of transforms

$$y^{(\lambda)} = \begin{cases} \frac{(y + 1)^\lambda - 1}{\lambda}, & \lambda \neq 0, y \geq 0 \\ \ln(y + 1), & \lambda = 0, y \geq 0 \\ -\frac{(-y + 1)^{2-\lambda} - 1}{2 - \lambda}, & \lambda \neq 2, y < 0 \\ -\ln(-y + 1), & \lambda = 0, y < 0 \end{cases}$$

Box-Cox Transform Example

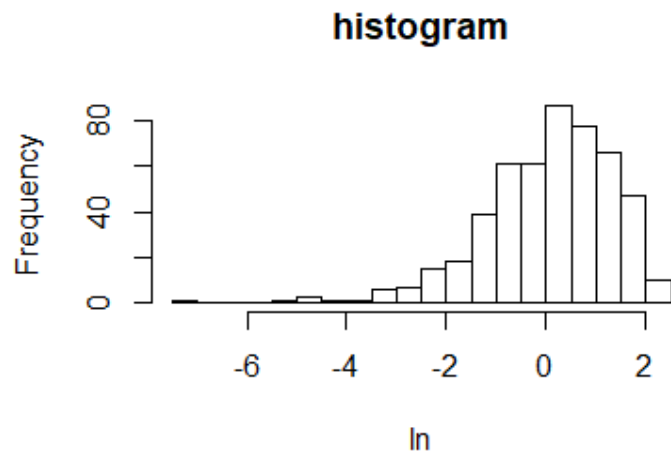


Heavily Skewed

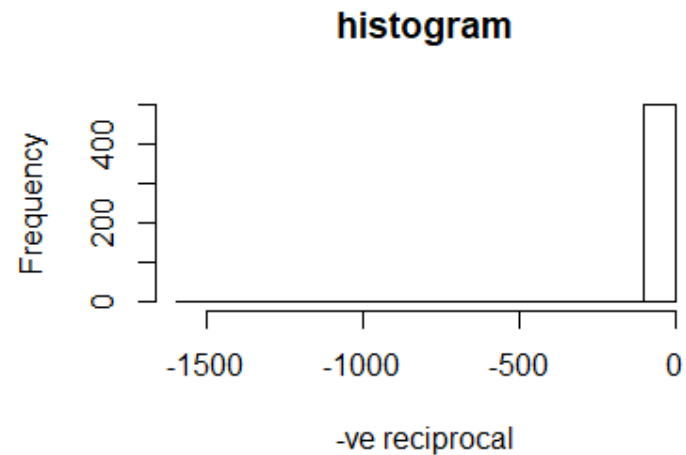


Approximately
Symmetric

Box-Cox Transform Example

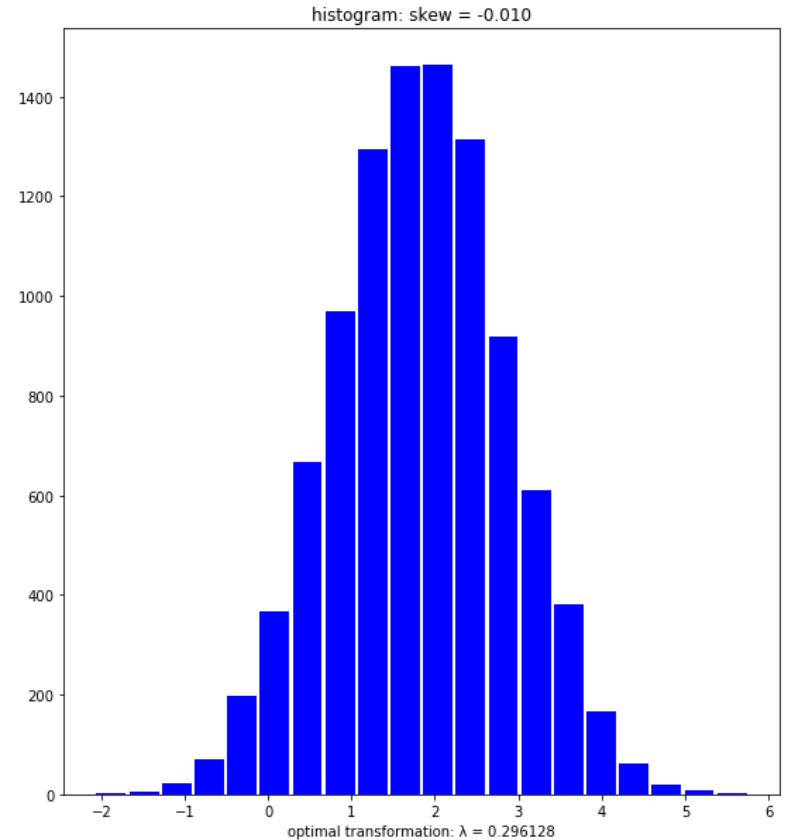
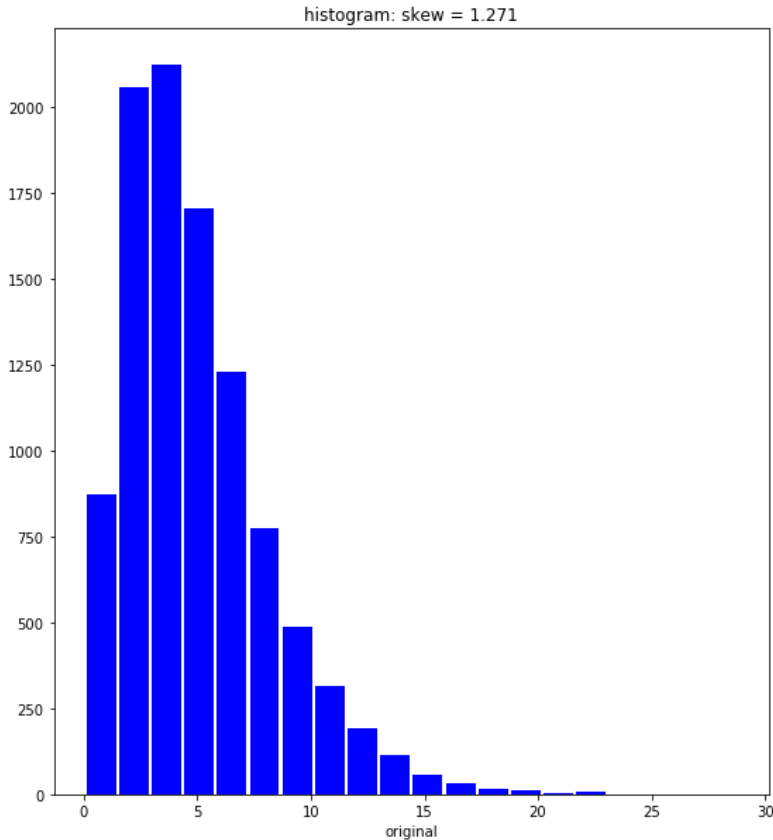


Approximately
Symmetric



Heavily Skewed

Box-Cox Transform Example



Generalized Linear Models

$$Y_i \sim D(\quad)$$

$$g[E(Y_i)] = \mathbf{x}_i^T \boldsymbol{\beta}$$

Example: Normal Linear Model

$$Y_i \sim N(\mu_i, \sigma^2), \quad Y \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

$$g[E(Y_i)] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

MLE

$$l \propto \sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2, \quad l \propto (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

That was

