PS5841

Data Science in Finance & Insurance

Linear Regression

Yubo Wang

Autumn 2021

Columbia University

Linear Regression Model

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

- where

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

$Y_1, \dots Y_n$ are independent RVs with

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

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

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

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

can accommodate heteroscedasticity

COLUMBIA UNIVERSITY

Least Squares Estimation

- Minimize

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

de-emphasize high variance RV's

Set

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

Get

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

Columbia University

Least Squares Estimator

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

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

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

with

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

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Var(y) \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2$$

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

Columbia University

Gauss-Markov Theorem

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

COLUMBIA UNIVERSITY

Normal Linear Model

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

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

$$\hat{\sigma}^2 = \frac{1}{n-p}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

COLUMBIA UNIVERSITY

NLM: Inference (3)

- Variance of fit at $\boldsymbol{x}$

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

- Variance of prediction at $\boldsymbol{x}$

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

Columbia University

Coefficient of Determination

- RSS of the model

  $\hat{S} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{y}^T\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{y}$
- $\hat{S}_0$, the worst value of RSS, is associated with the naïve model $E(Y_i) = \mu, \forall i$

  $\hat{S}_0 = \boldsymbol{y}^T\boldsymbol{y} - N\bar{y}^2 = \Sigma(y_i - \bar{y})^2 \propto Var(y)$
- Relative improvement

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

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

Columbia University

The Naïve Model

$$E(Y_i) = \mu, \forall i$$
$$\widehat{\boldsymbol{\beta}} = \hat{\mu} = \bar{y}$$
$$\boldsymbol{X} = \boldsymbol{j}$$

$$\begin{aligned}\hat{S}_0 &= \boldsymbol{y}^T\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \bar{y}\Sigma y_i \\ &= \boldsymbol{y}^T\boldsymbol{y} - N\bar{y}^2 \\ &= \Sigma(y_i - \bar{y})^2 \propto Var(y)\end{aligned}$$

COLUMBIA UNIVERSITY

General Linear Model

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

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

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$$

- 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_1X_2)

COLUMBIA UNIVERSITY

Power Transforms

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

Strictly Positive Data

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

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

- 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)

COLUMBIA UNIVERSITY

General Data

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

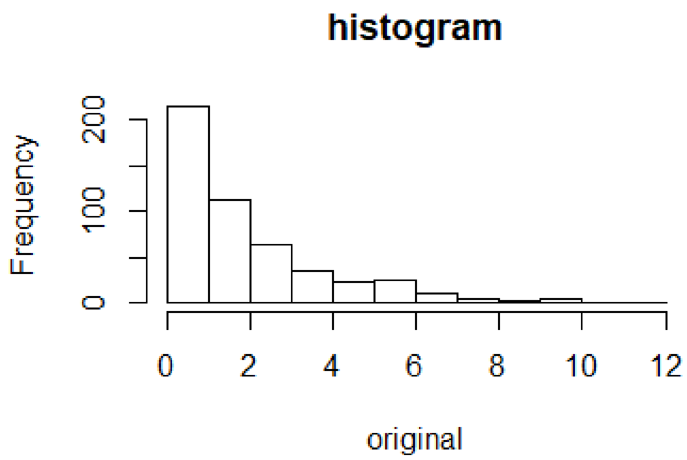
$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\alpha)^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y+\alpha), & \lambda = 0 \end{cases}$$

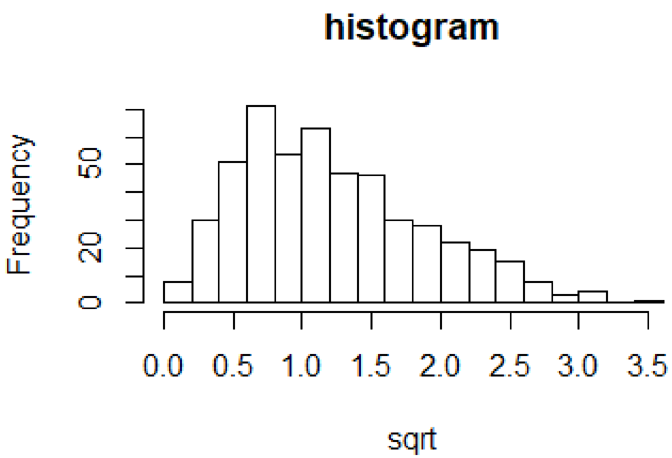
- Yeo-Johnson family of transforms

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

COLUMBIA UNIVERSITY

Box-Cox Transform Example

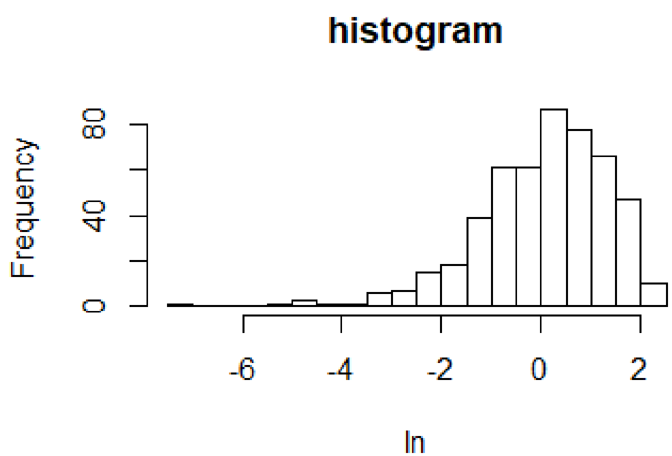


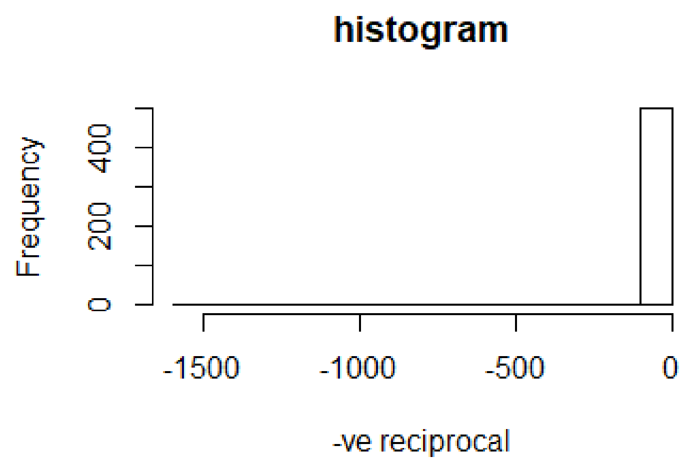


Heavily Skewed

Approximately Symmetric

Box-Cox Transform Example

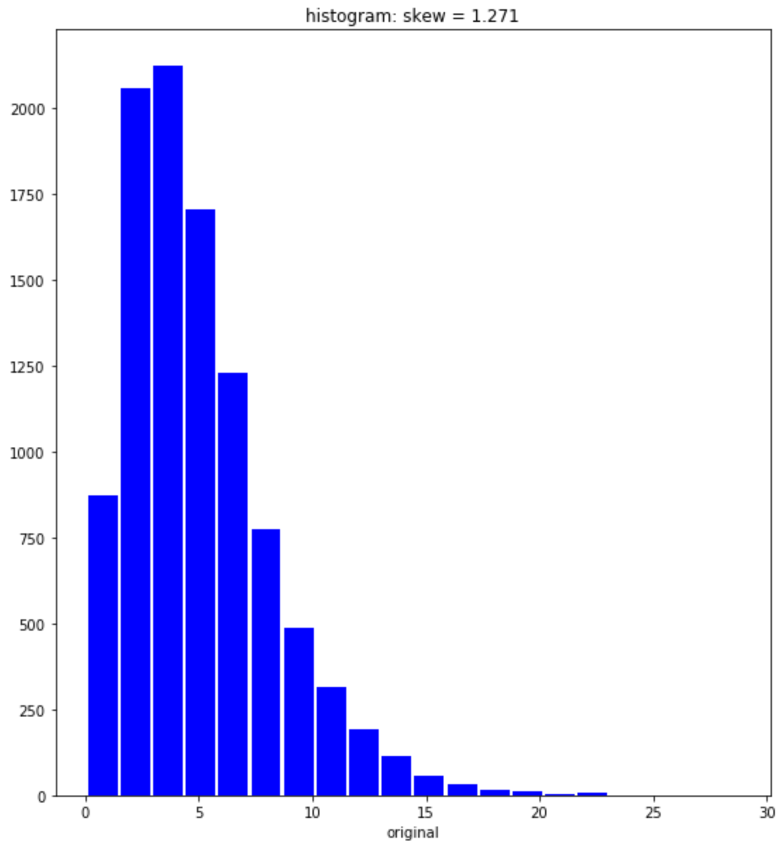


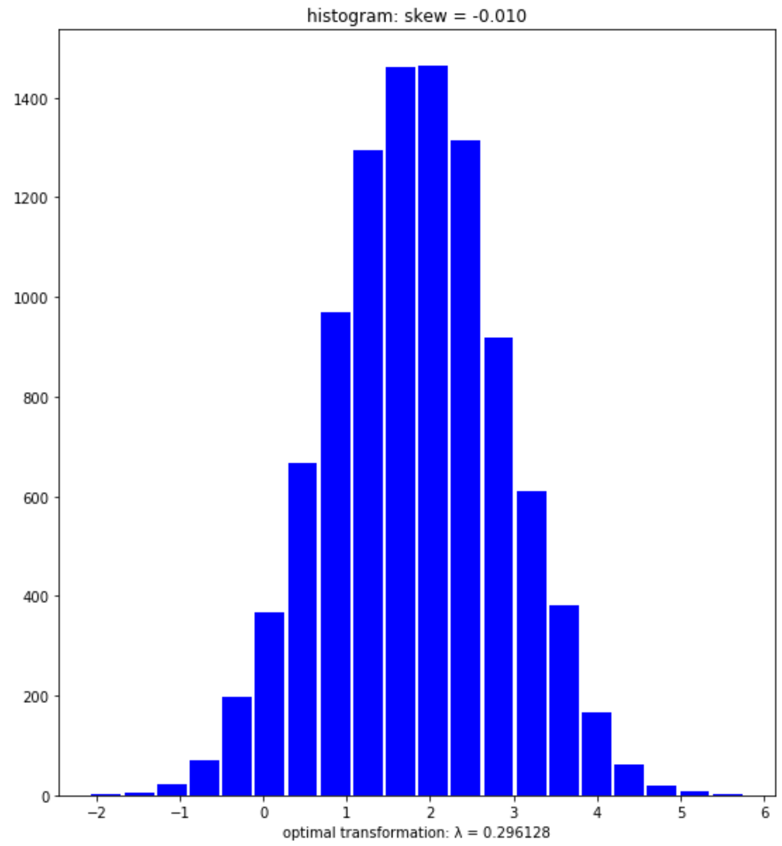


Approximately Symmetric

Heavily Skewed

Box-Cox Transform Example





Categorical Variables

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?

COLUMBIA UNIVERSITY

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?

COLUMBIA UNIVERSITY

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 $\hat{\beta}_1$

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

COLUMBIA UNIVERSITY

Generalized Linear Models

Columbia University

GLM

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

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

COLUMBIA UNIVERSITY

Example: Normal Linear Model

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

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

MLE

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

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

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

COLUMBIA UNIVERSITY

Gradient Descent

Gradient Operator

- Gradient operator (del operator)

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

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

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

Gradient

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

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

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

COLUMBIA UNIVERSITY

Example (1)

- $f(\boldsymbol{x}) = x_1 + x_2^2, \boldsymbol{\nabla} f(\boldsymbol{x}) = \binom{1}{2x_2}$
- Moving d units from $\boldsymbol{x}_0 = \binom{1}{1}$ in the direction of

$$\boldsymbol{\nabla} f(1,1) = \binom{1}{2}, |\boldsymbol{\nabla} f(1,1)| = \sqrt{5}$$

lands at

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

Columbia University

Example (2)

Same as predicted rate

gradient ascent					
d	f[1+(1/sqrt(5))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					
d	f[1+(1/sqrt(2))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

Columbia University

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

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

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

COLUMBIA UNIVERSITY

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

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

COLUMBIA UNIVERSITY

That was

COLUMBIA UNIVERSITY