APPLIED STATISTICAL MODELING AND DATA ANALYTICS

KMITL Digital Analytics and Intelligence Center

Linear Regression

Dr. Yuwadee Klomwises

Causation and Correlation

Causation

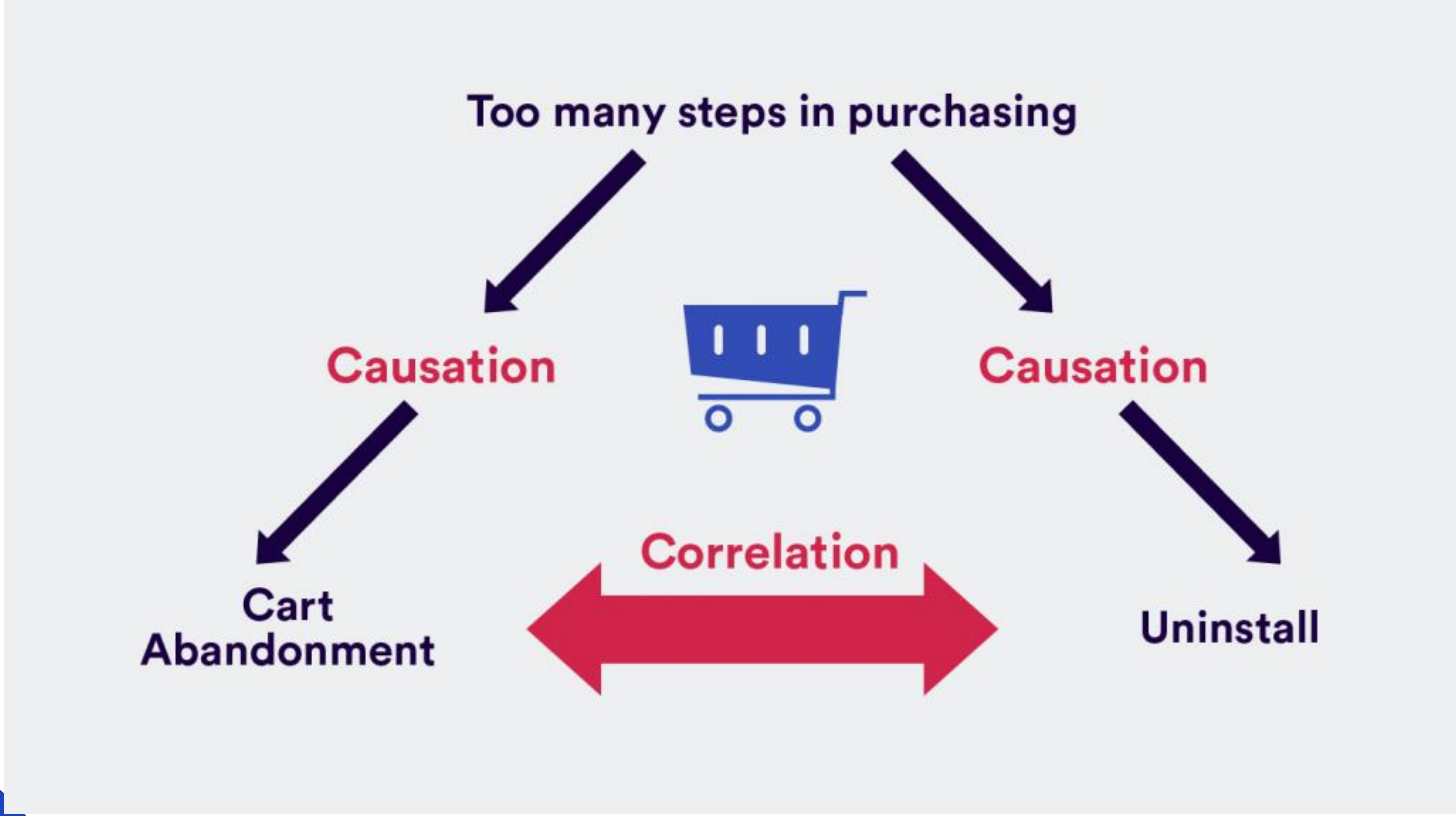
One variable producing an effect in another variable

Correlation

Relationship between two variables

Requirements for causation

- Evidence of association between X and Y
- X must occur before Y
- No third factor driving both


Too many steps in purchasing
Causation
Causation
Correlation
Cart
Abandonment
Uninstall

Correlation coefficients

How to explore the correlation between different types of variables

Dependent variable : Y

		Categorical	Quantitative
Exploratory Variable: X	Categorical	Chi-square test	Analysis of Variance (ANOVA)
	Quantitative	Chi-square test (transform quantitative to categorical)	Pearson correlation Spearman correlation Kendall correlation

Correlation between quantitative variables

the strength of relationship between variables

Pearson correlation

Restrictions: continuous variable

with normal distributed

Spearman correlation

Kendall correlation

Alternative for Pearson correlation

When assumptions are violated

Hypothesis testing for correlation

H_0 : The population correlation coefficient IS NOT significantly different from zero

H_1 : The population correlation coefficient IS significantly DIFFERENT FROM zero

Reject H_0 if p-value < significance level

Regression model depended on types of dependent variable

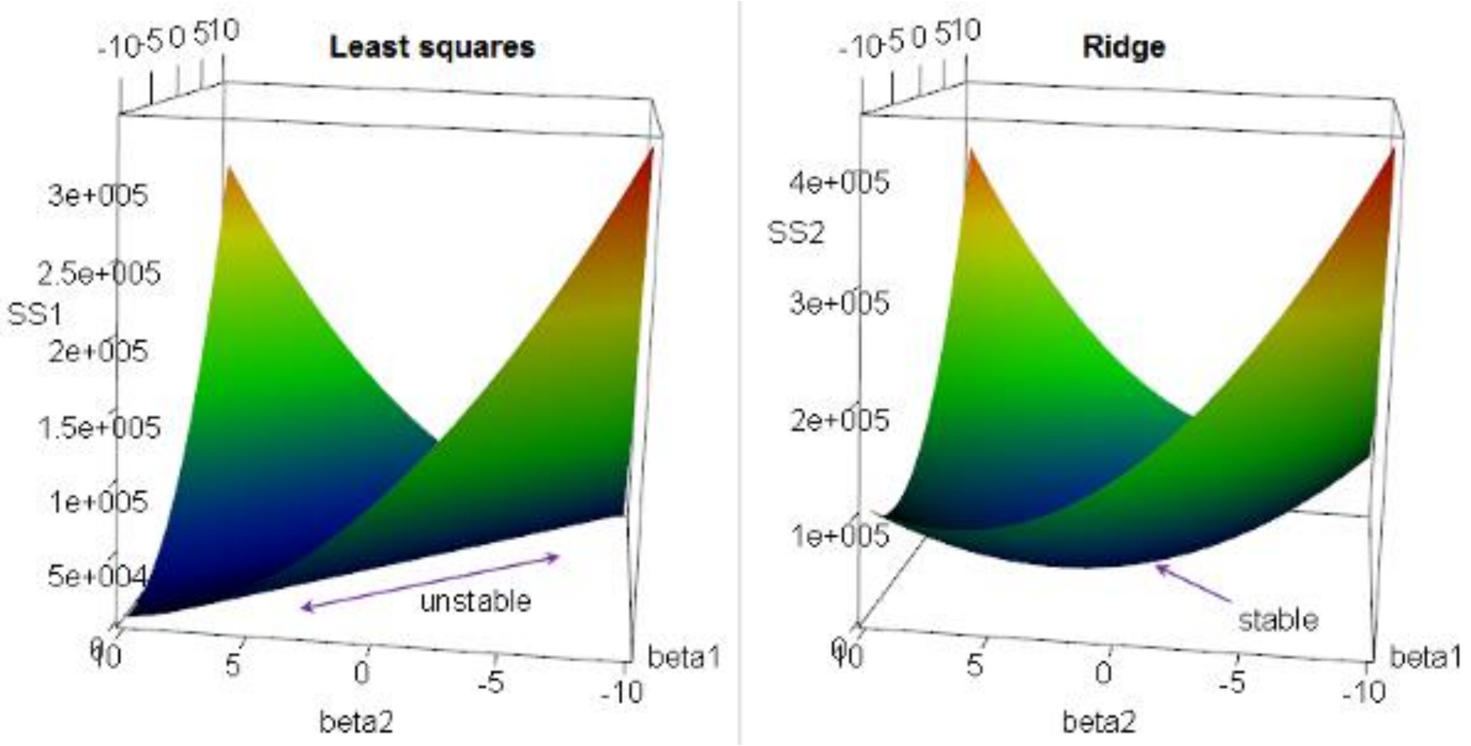
- Linear regression
- Logistic regression
- Beta regression
- Poisson regression
- Multinomial logistic regression

Regression model concerning complexity of optimization

- Ridge regression

majorly used
to prevent overfitting

works well even in presence of highly correlated features



Source: https://people.eecs.berkeley.edu/

Regression model concerning complexity of optimization

- Lasso regression

performs feature selection:: some of the coefficients become exactly zero

selects any one independence variable among the highly correlated ones

Simple linear Regression

What is regression analysis?

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions:

Which factors matter most?

Which can we ignore?

How do those factors interact with each other?

How certain are we about all of these factors?

Source https://hbr.org/

What is regression analysis?

In regression analysis, those factors are called **variables**.

You have your **dependent variable** — the main factor that you're trying to understand or predict such as monthly sales.

And then you have your **independent variables** — the factors you suspect have an impact on your dependent variable.

Source https://hbr.org/

Notation for data used in Regression analysis

Observation	Dependent variable	Independent variable			
	Y	X_1	X_2	…	X_p
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
	…	…	…		…
n	y_n	x_{n1}	x_{n2}	…	x_{np}

Model specification

In simple linear regression,

it assumes that there is approximately a linear relationship between X and Y.

Mathematically, we can write linear relationship between X and Y as

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

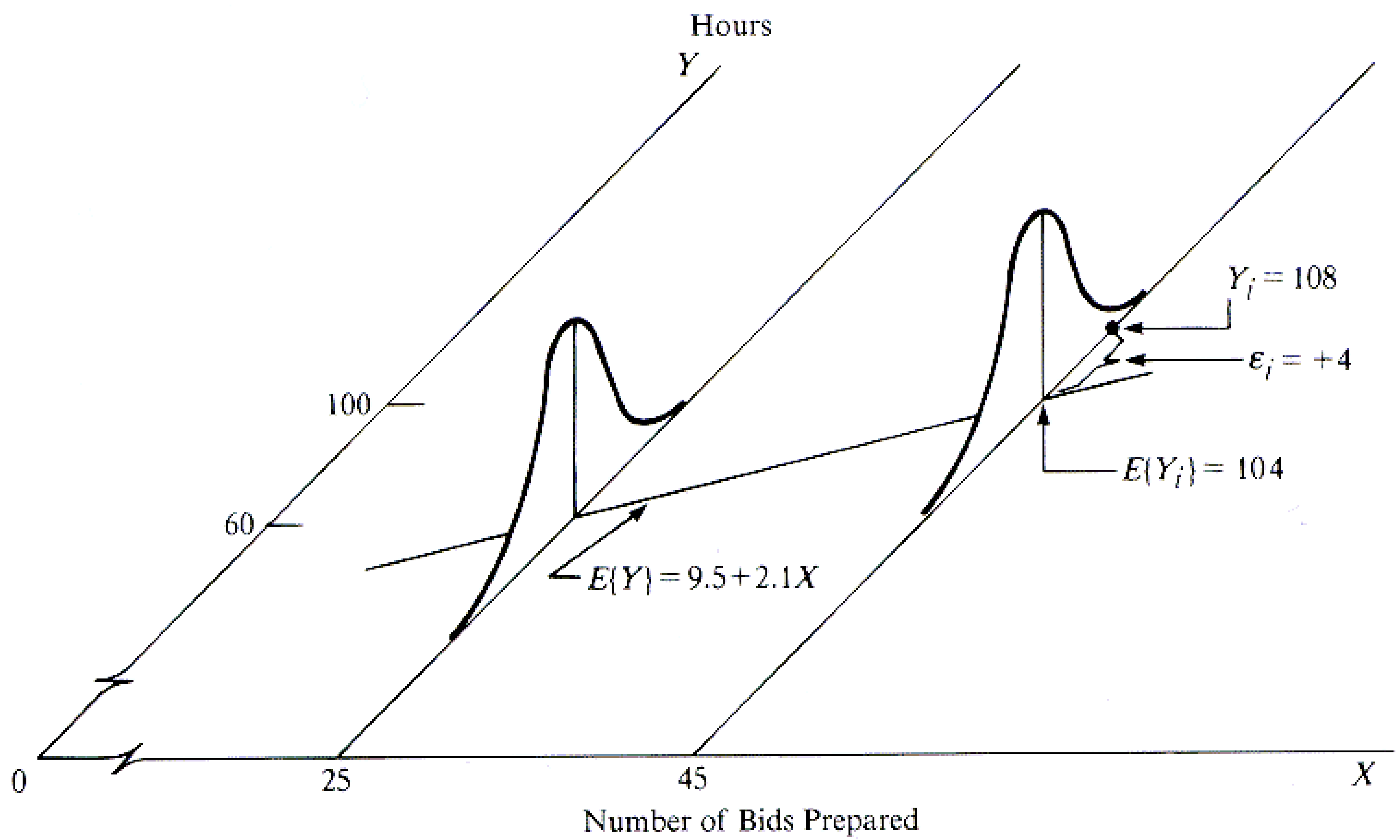
Model specification

$$E(Y) = \beta_0 + \beta_1 X_1 + \varepsilon$$

β_0, β_1 the so-called *coefficients* of the variables

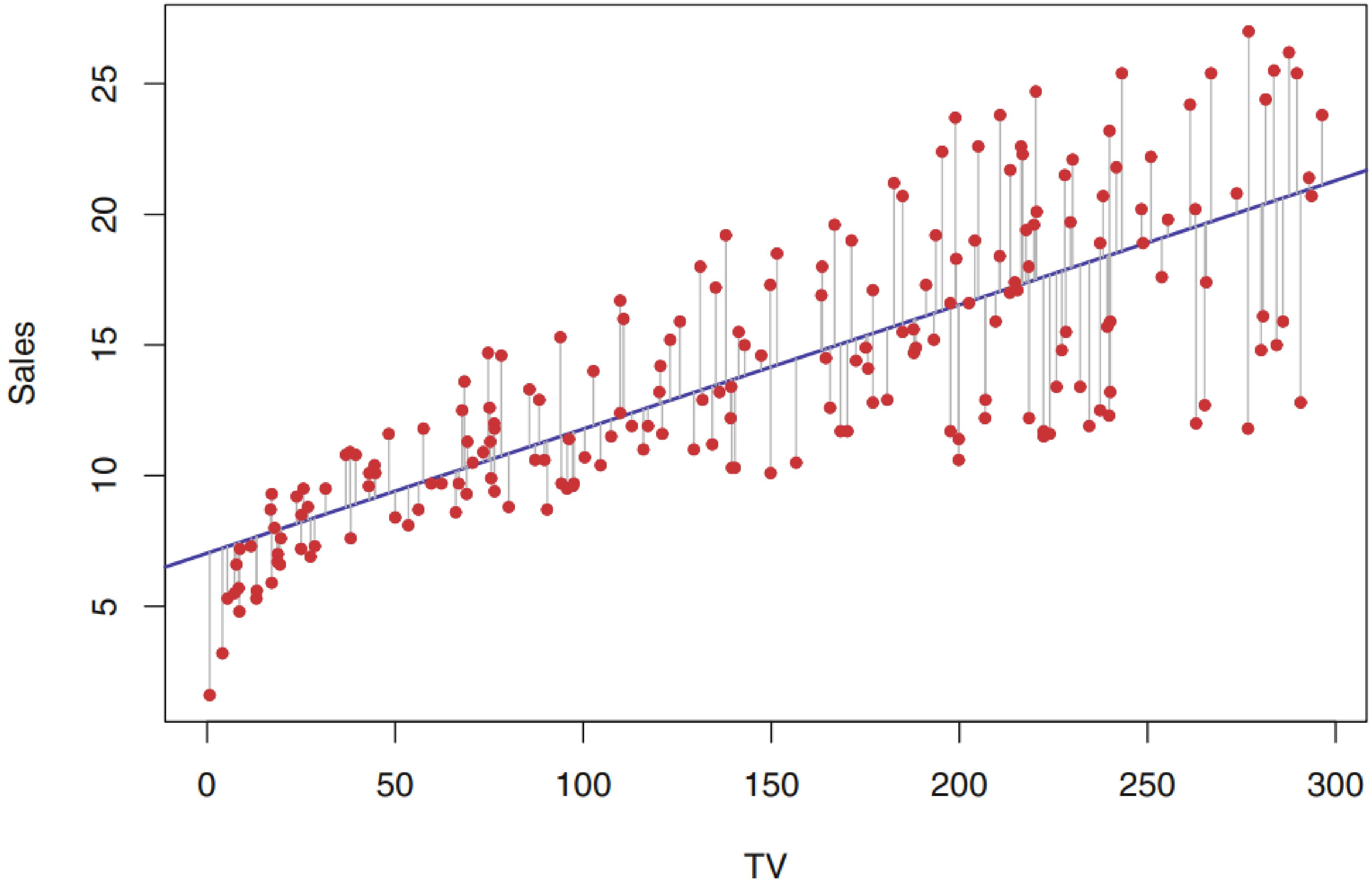
β_1 is the change in the predicted value of Y per unit of change in X_1

β_0 the so-called intercept, is the prediction that the model would make if X_1 was zero



Source: http://www.unc.edu/~nielsen/soci709/m1/m1005.gif

Parameter estimation



Parameter estimation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

We define the *residual sum of squares* (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

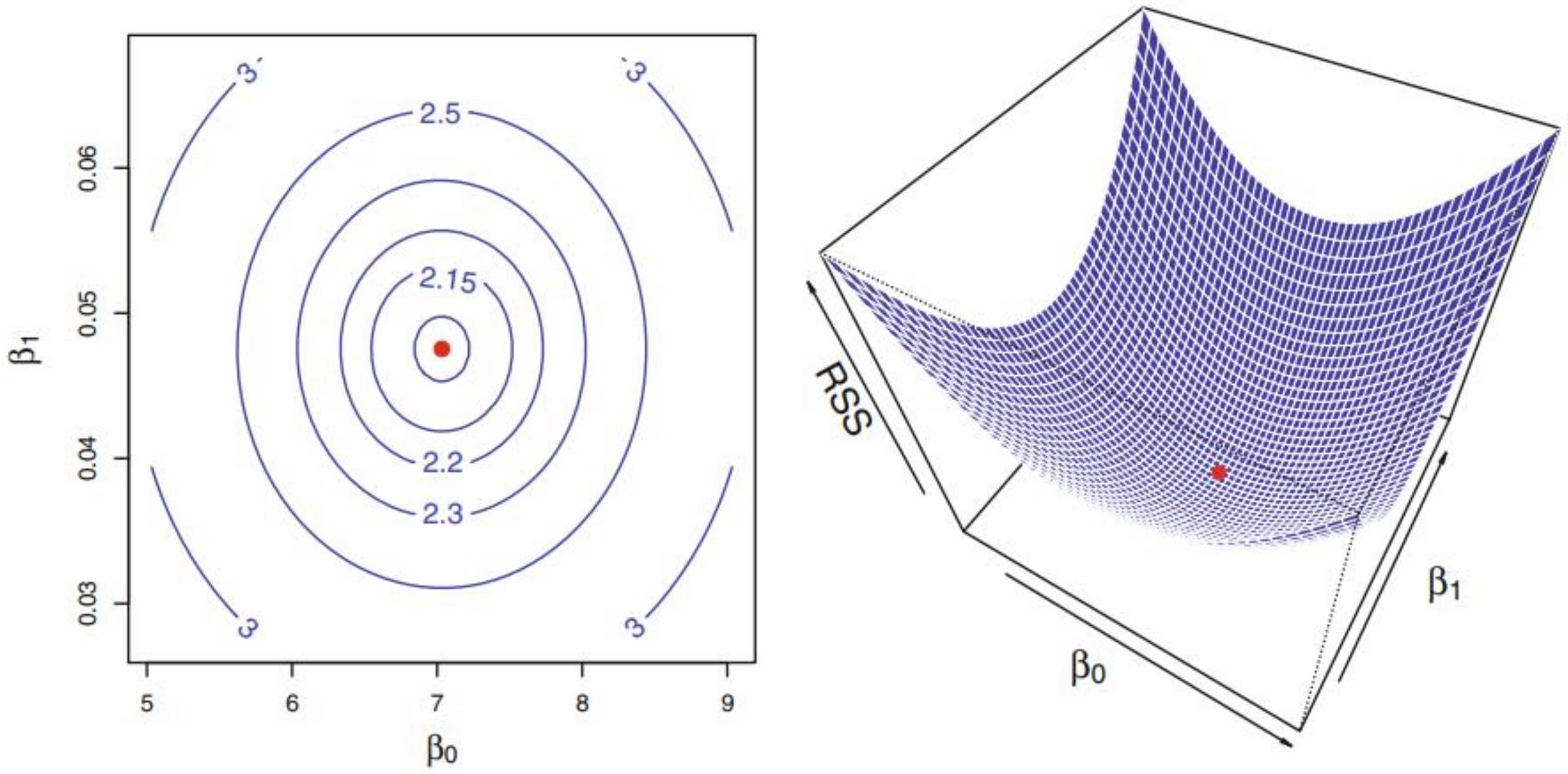
The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.

Parameter estimation

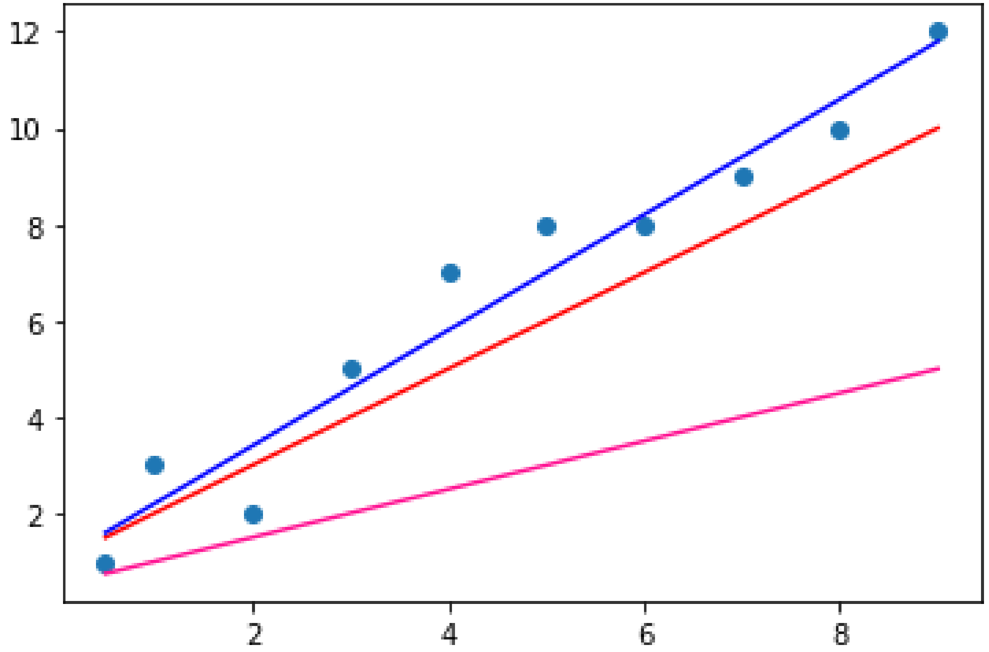
Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Parameter estimation

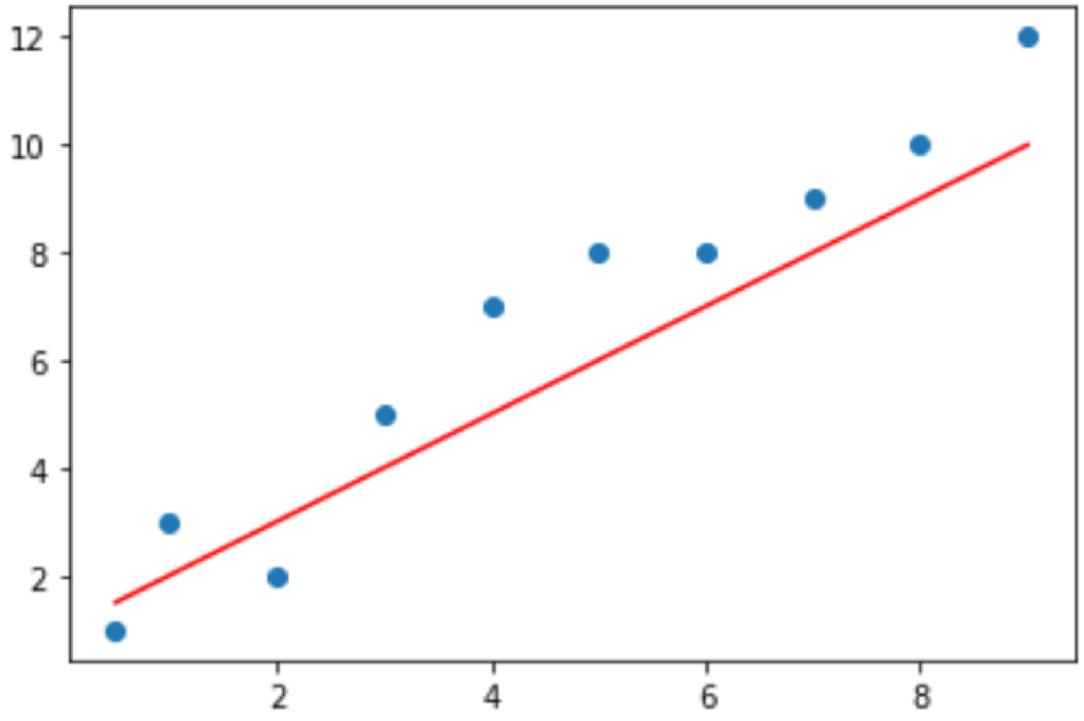


How to choose the best line



Cost of using specific line

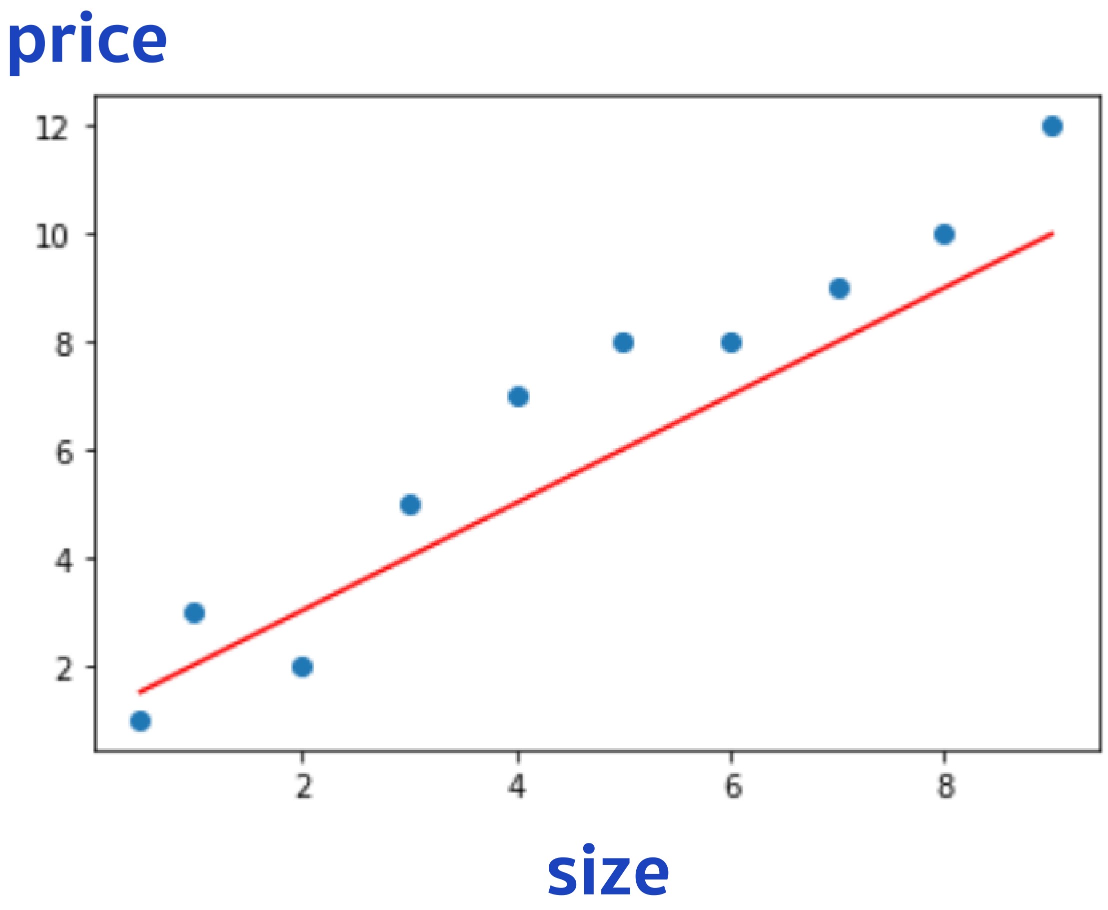
price



size

$$\mathrm{RSS}(\beta_0,\beta_1) = \left(\mathrm{price}_1 - \beta_0 - \beta_1 \mathrm{size}_1\right)^2$$
$$+\left(\mathrm{price}_2 - \beta_0 - \beta_1 \mathrm{size}_2\right)^2$$
$$+\ldots+\left(\mathrm{price}_n - \beta_0 - \beta_1 \mathrm{size}_n\right)^2$$

Cost of using specific line

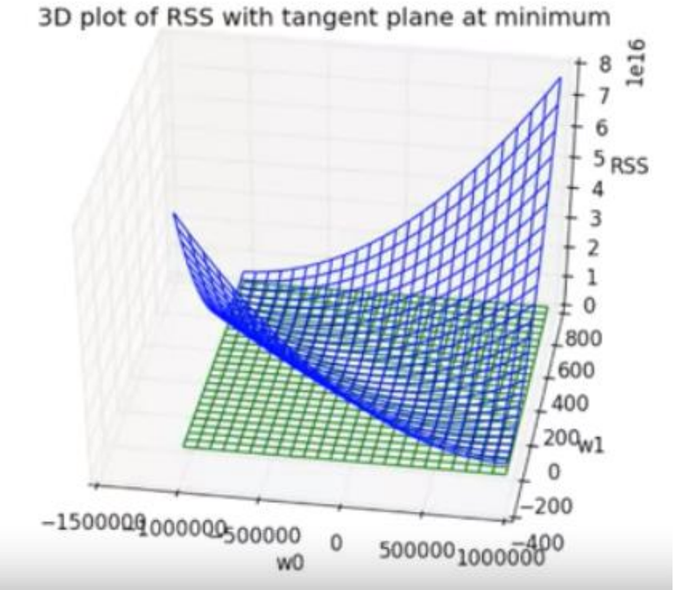


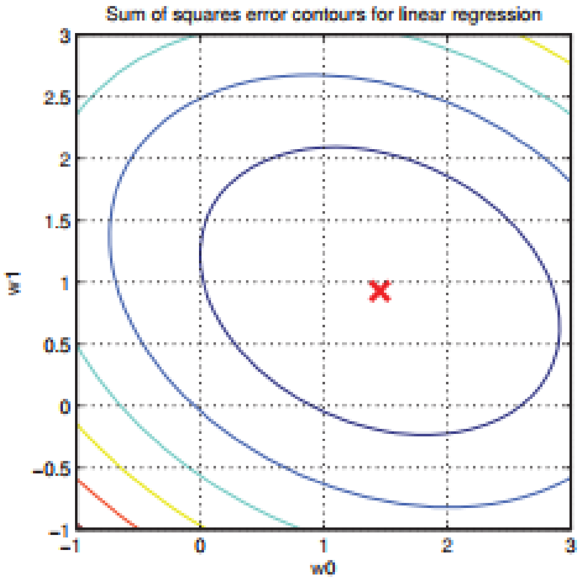
$$\text{RSS}(\beta_0, \beta_1) = \left(\text{price}_1 - \beta_0 - \beta_1 \text{size}_1\right)^2$$
$$+\left(\text{price}_2 - \beta_0 - \beta_1 \text{size}_2\right)^2$$
$$+\ldots+\left(\text{price}_n - \beta_0 - \beta_1 \text{size}_n\right)^2$$

$$\text{RSS} = \sum_{i=1}^{n} \left(\text{price}_i - \beta_0 - \beta_1 \text{size}_i\right)^2$$

Finding solution for RSS

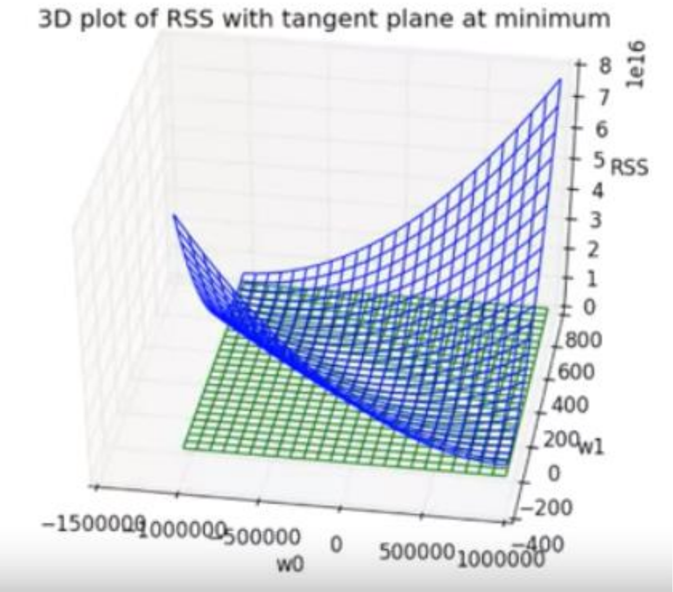
Defining Gradient of RSS





$$\nabla \text{RSS}(\beta_0, \beta_1) = \begin{bmatrix} \dfrac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_0} \\ \dfrac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_1} \end{bmatrix}$$

Defining Gradient of RSS



$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_0} =$$

$$\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_1} =$$

Defining Gradient of RSS

$$\nabla \mathrm{RSS}(\beta_0, \beta_1) = \begin{bmatrix} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \end{bmatrix}$$

- Closed form solution
- Gradient descent

Closed form solution

$$\nabla \mathrm{RSS}(\beta_0, \beta_1) = \begin{bmatrix} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \end{bmatrix}$$

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

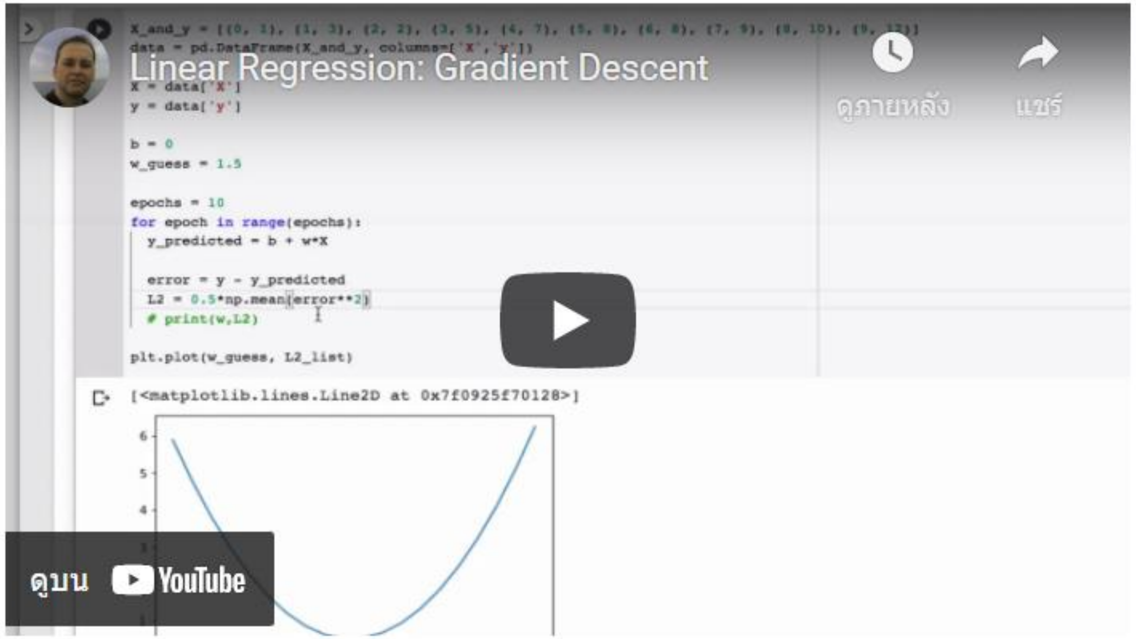
Closed form solution

$$\nabla \mathrm{RSS}(\beta_0, \beta_1) = \begin{bmatrix} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \end{bmatrix}$$

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Optional reading

https://sites.google.com/view/ml-basics/linear-regression-and-gradient-descent



- Gradient
- Gradient descent
- Epochs
- Learning rate
- Convergence

Assumptions in linear regression

Assumptions in linear regression

<u>mostly based on predicted values and residuals</u>

- Linearity
- Independence
- Homogeneity of variance (homoscedasticity)
- Normality

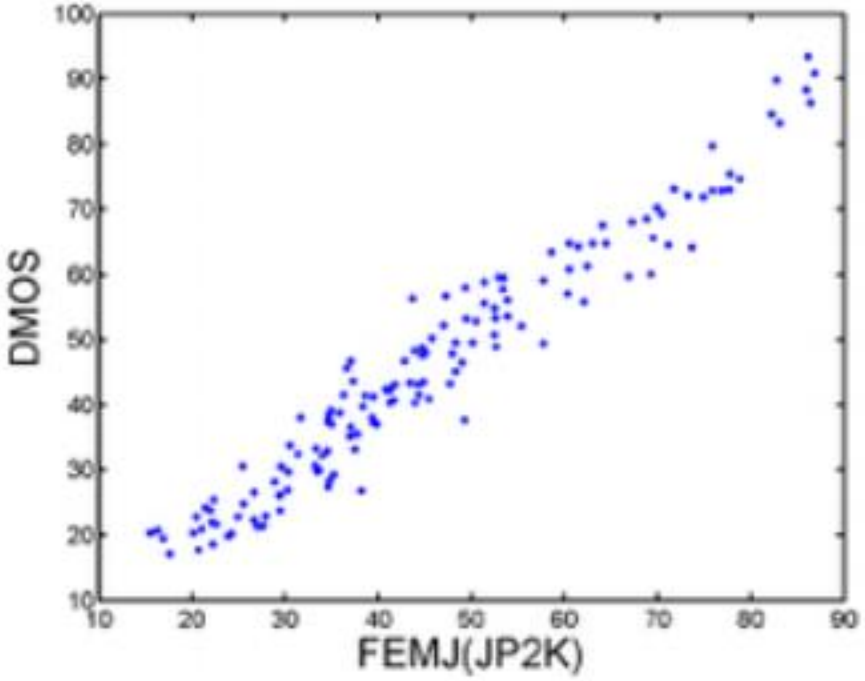
Assumptions in linear regression

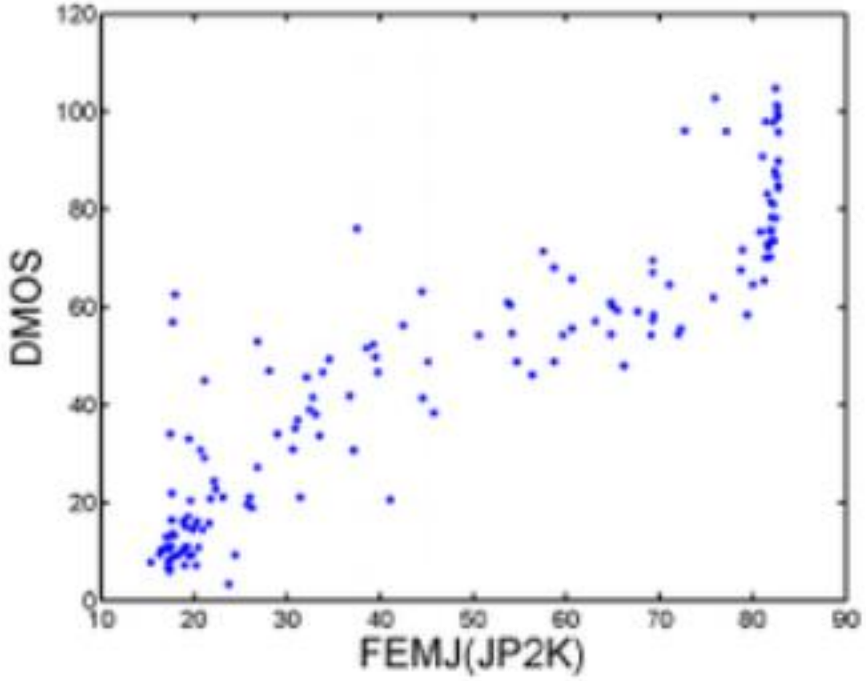
Linearity

the relationships between the predictors and the outcome variable should be linear.

Big deal if violated.

Linearity

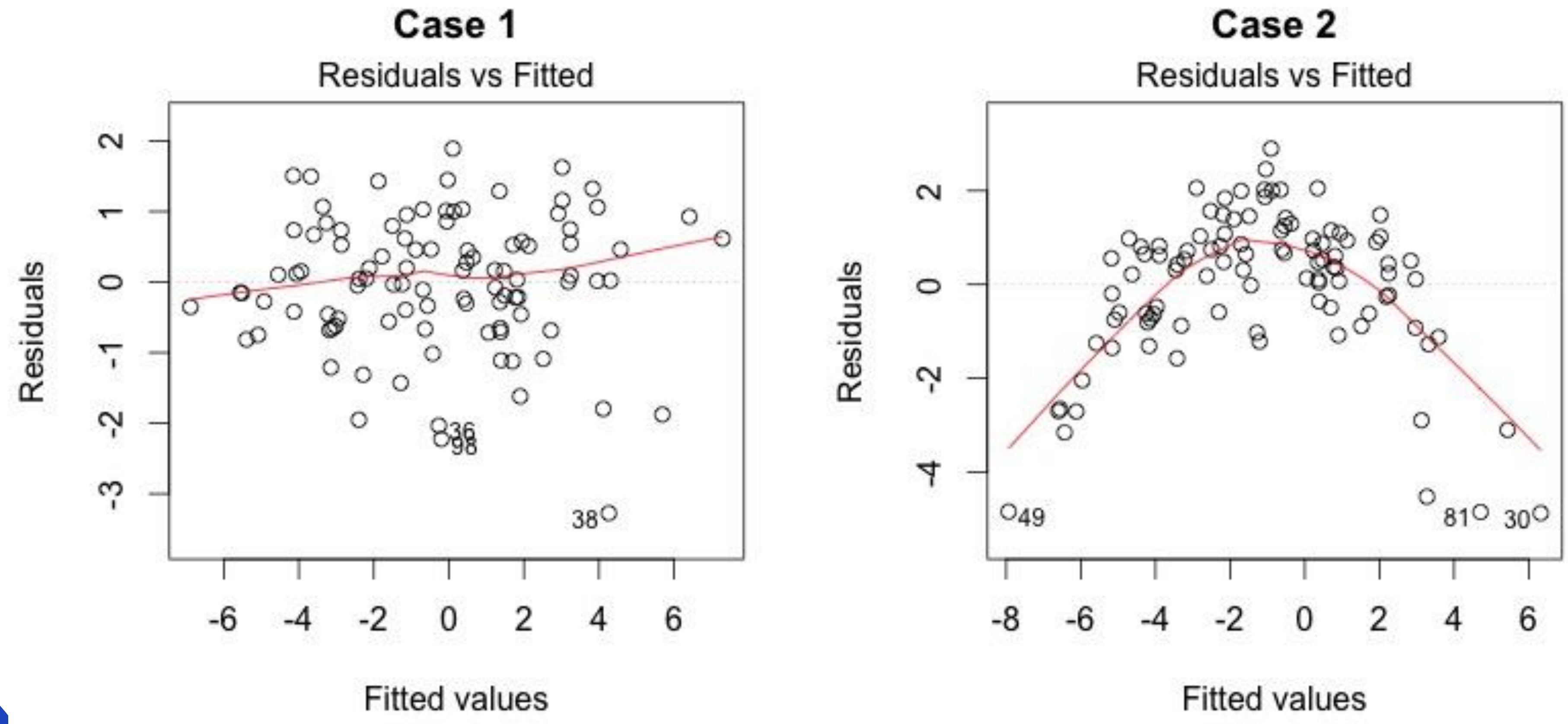




source https://www.researchgate.net/profile/Ke-Gu-2

Linearity: : How to create diagnostic plot

residuals versus predicted values



Linearity: : How to fix

nonlinear transformation : if it appropriate

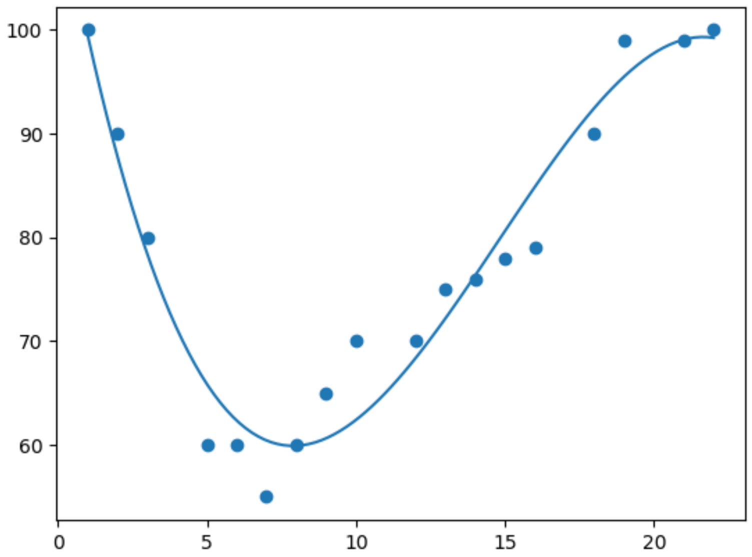
take natural log only to dependent variable → Y grows exponentially as a function of X

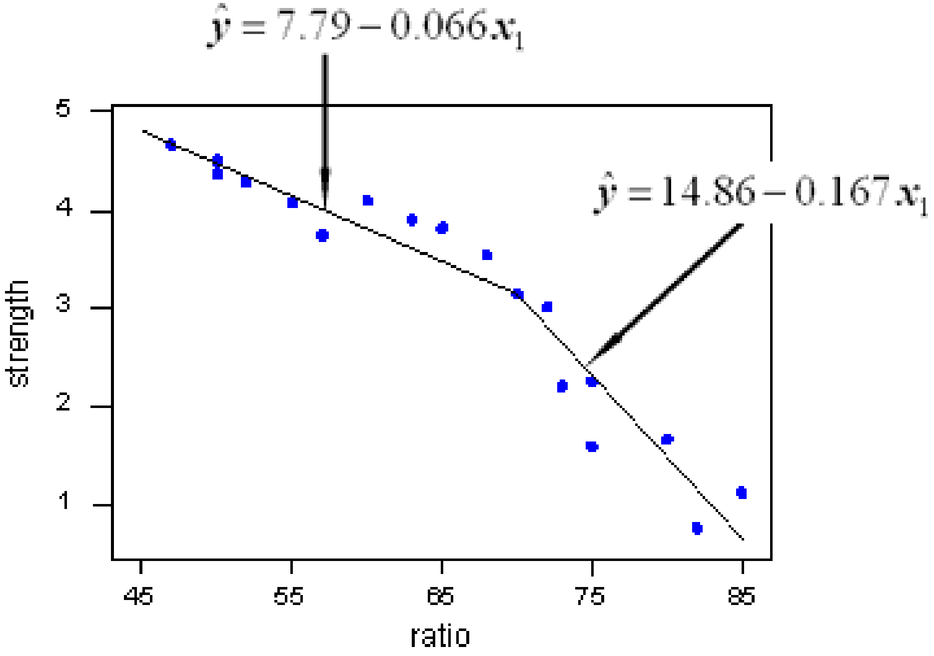
take natural log to dependent and independent variable → X has multiplicative effects on Y

Linearity: : How to fix

artificially constructed variables (may lead to overfitting)

- Polynomial regression
- Piecewise Linear Regression





ที่มา https://online.stat.psu.edu/stat501/lesson/8/8.8

Assumptions in linear regression

Independence

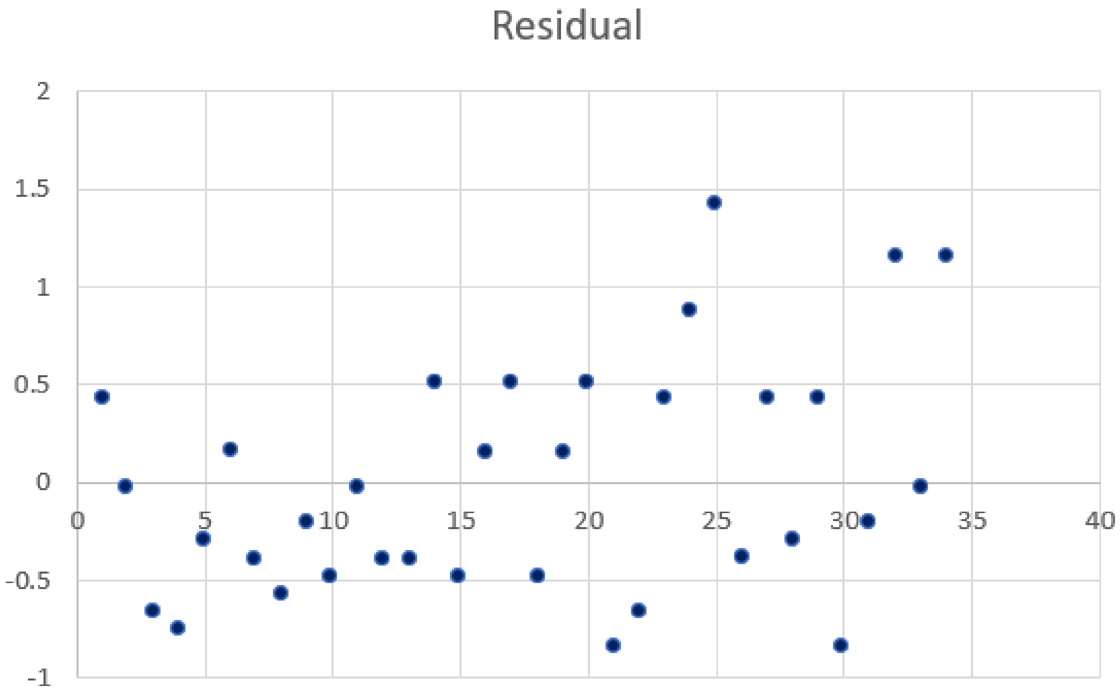
no relationship among the residuals

no relationship between the residuals and independent variable

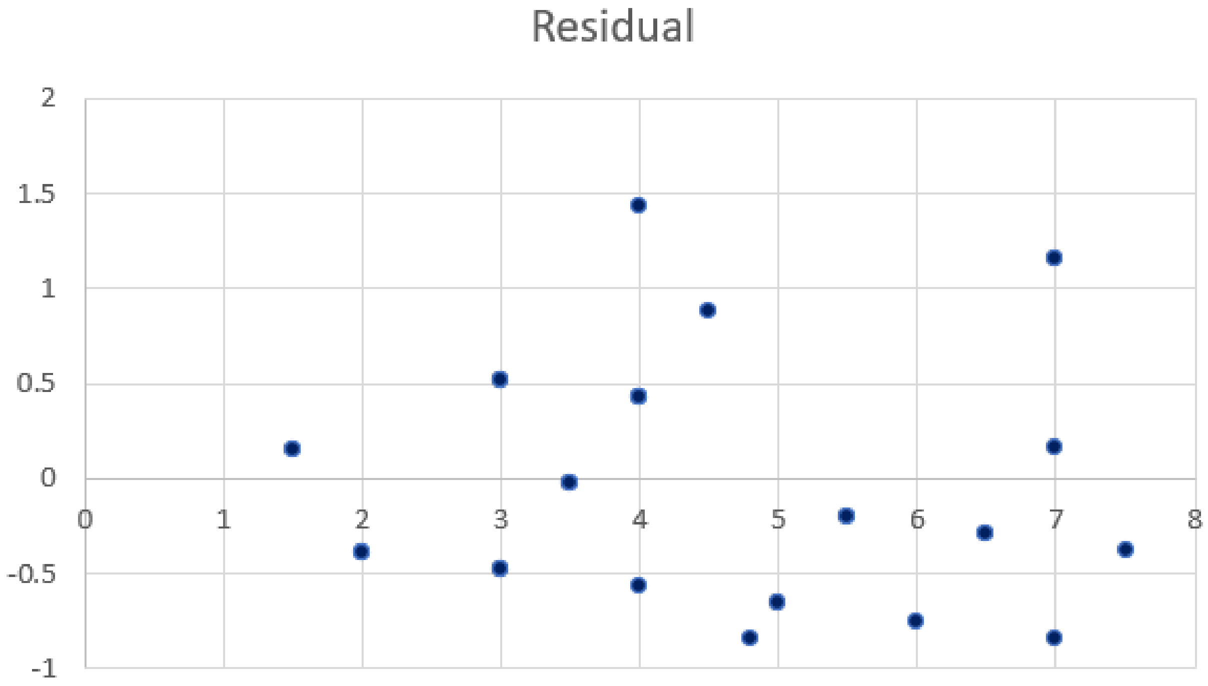
Huge deal if violated

Independence: : How to create diagnostic plot

- residuals versus row number



- residuals versus independent variables



Independence: : How to use statistics

Durbin Watson (DW) statistic should be close to 2.0

This statistic will always be between 0 and 4.

The closer to 0 the statistic, the more evidence for positive serial correlation.

The closer to 4, the more evidence for negative serial correlation.

Alternative test: Ljung-Box test

Ljung-Box test

```
from statsmodels.stats.diagnostic import acorr_ljungbox
lb = acorr_ljungbox(res.resid)
print(lb)
```

```
     lb_stat  lb_pvalue
1   0.003400   0.953500
2   0.774305   0.678988
3   1.412020   0.702720
4   1.890551   0.755881
5   2.176684   0.824197
6   2.397583   0.879749
7   3.186928   0.867188
8   3.639602   0.888089
9   3.793818   0.924451
10  5.639786   0.844565
```

Independence: : How to fix

If it due to

- Violation of the linearity assumption

- Omitted variable bias

Assumptions in linear regression

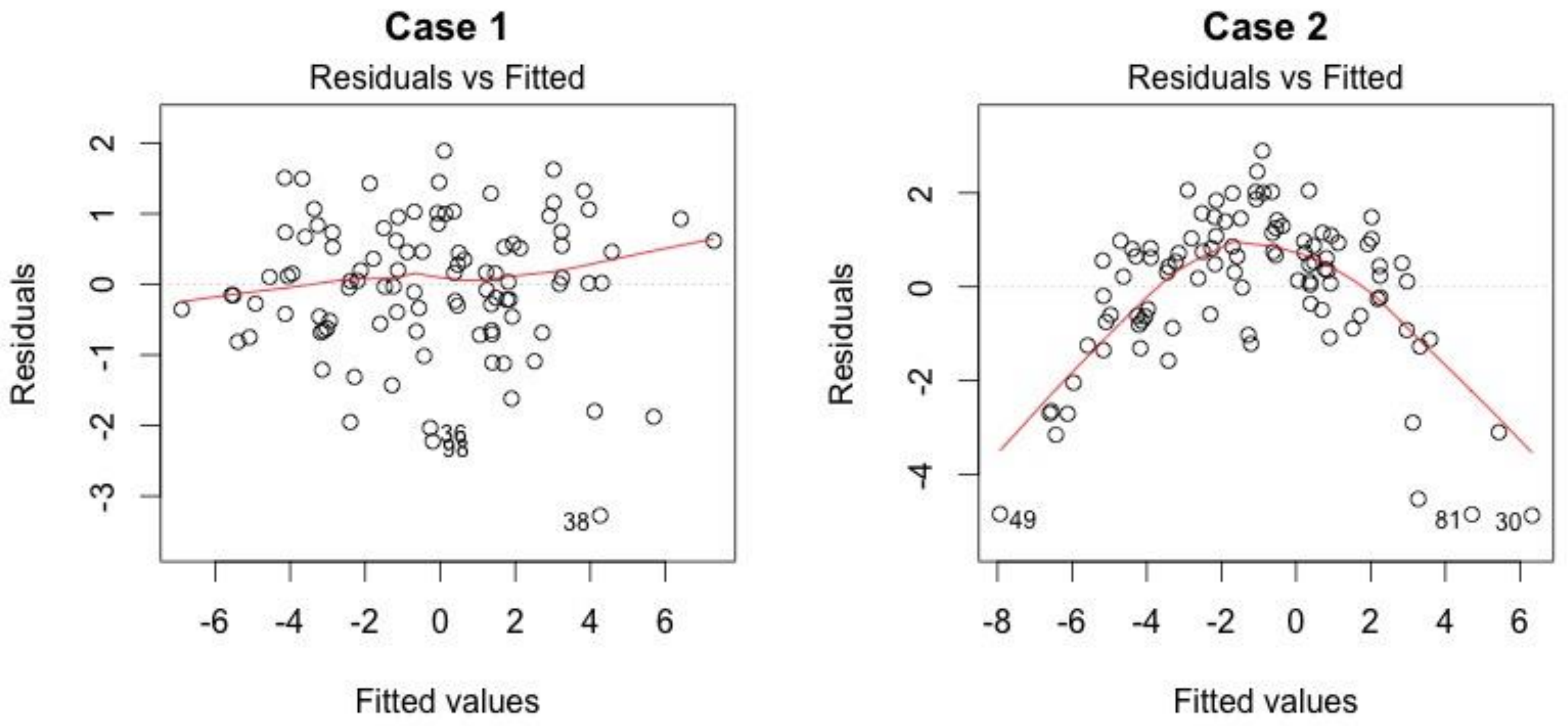
Homogeneity of variance (homoscedasticity)

the variance of error should be constant

Not as big deal if violated.

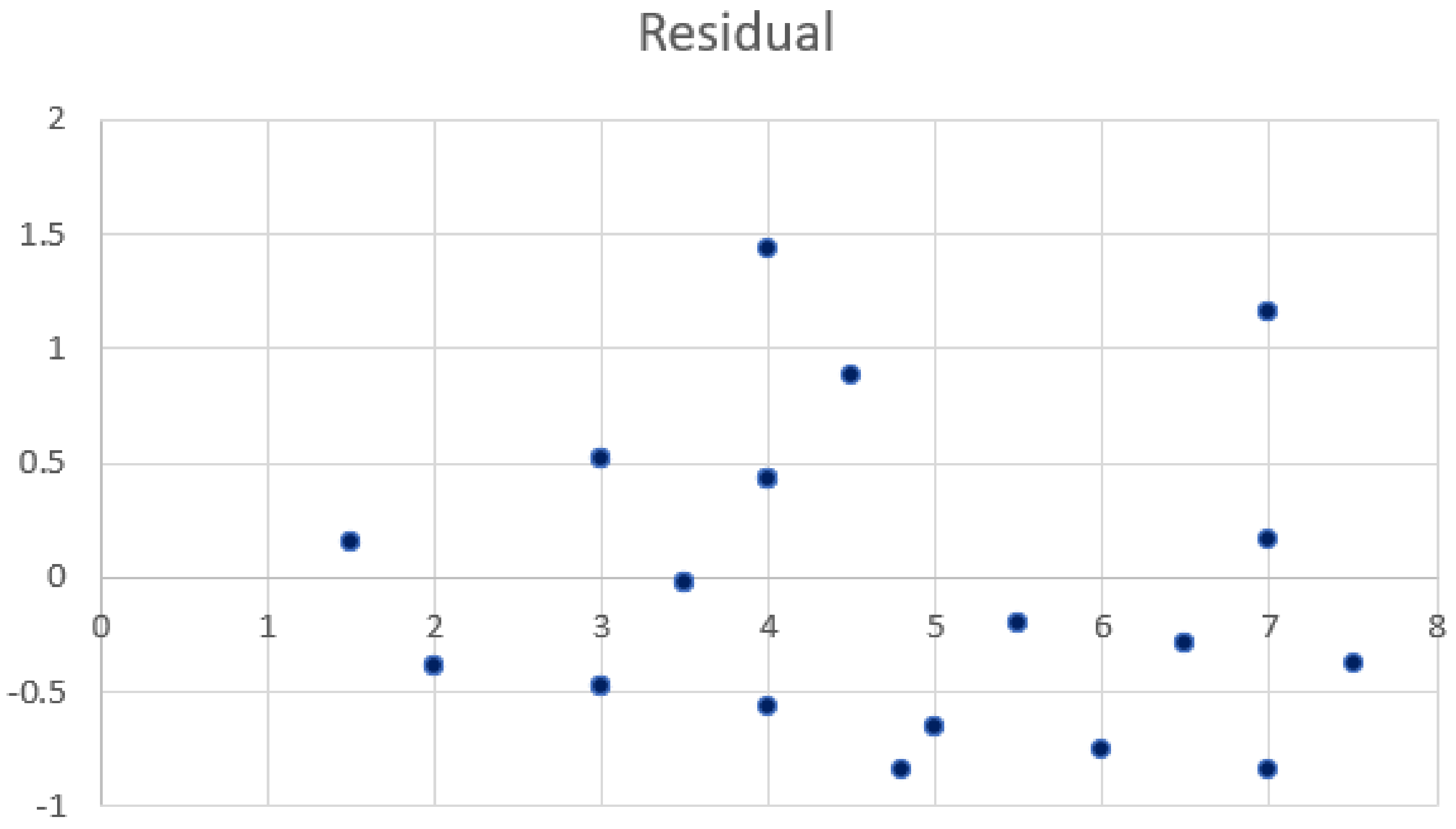
Homogeneity of variance:: How to create diagnostic plot

residuals (or standardized residual) versus predicted values



Homogeneity of variance:: How to create diagnostic plot

residuals versus independent variables



Homogeneity of variance : : How to use statistics

White Test

H_0: the errors are homoscedasticity (have constant variance)
H_1: the errors are heteroscedasticity (have non-constant variance)

Reject H_0 if p-value is less than significant level

Homogeneity of variance:: How to fix

Relevance to

- linearity assumption
- Independent assumption

Assumptions in linear regression

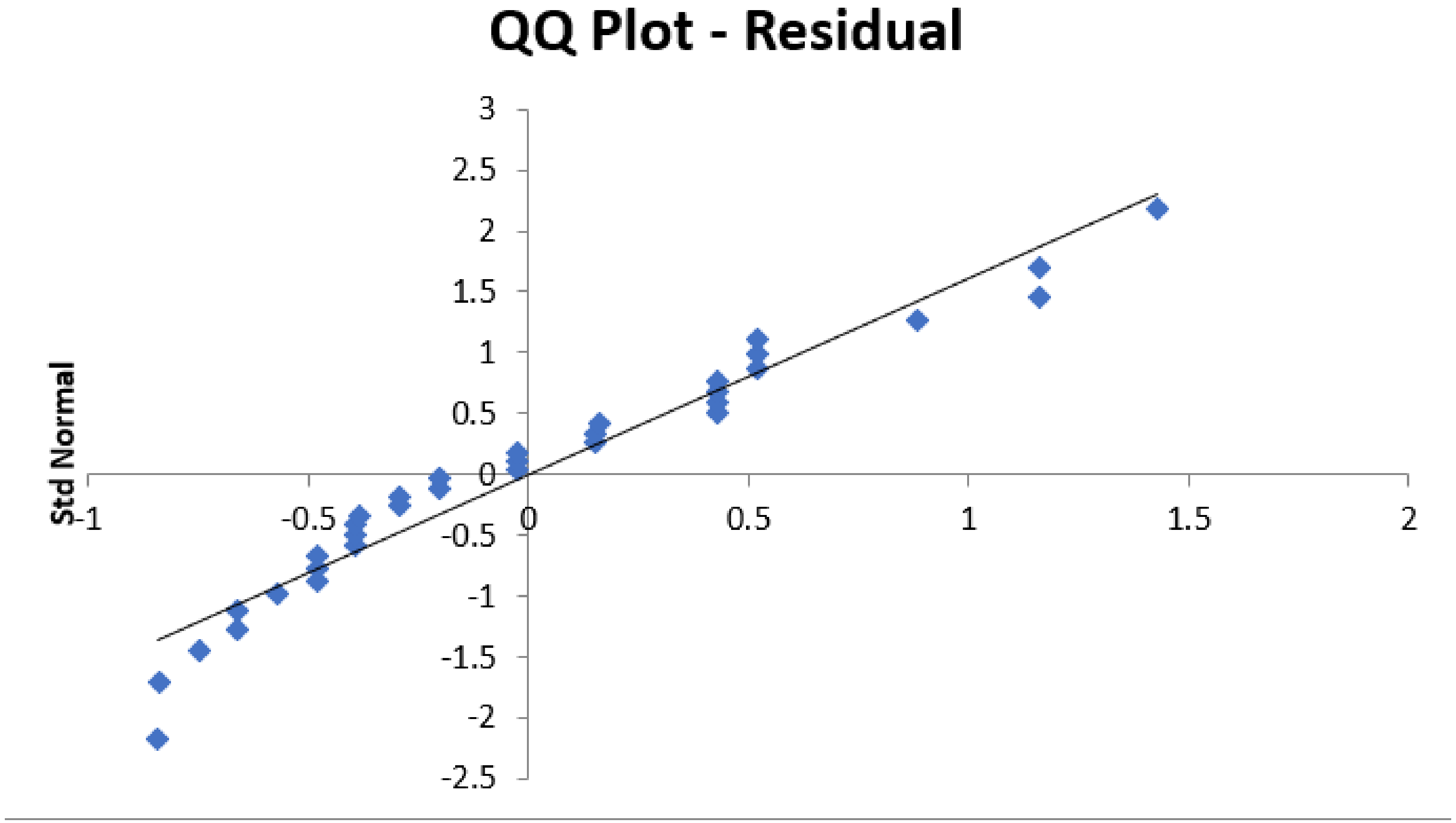
Normality

the errors should be normally distributed – normality is necessary for defining test statistics for regression coefficients

Not as big deal if violated.

Normality: How to create diagnostic plot

Q-Q plot of residuals



Normality: How to do hypothesis testing

- D'Agostino-Pearson test

  based on skewness and kurtosis

- Shapiro-Wilk test

  does not work well when several values are identical

H_0: a dependent variable is normally distributed

H_1: a dependent variable is not normally distributed

Reject H_0 if p-value is less than significant level

Normality:: How to fix

violations of normality often arise either because

- non-normal distributions of the dependent and/or independent variables
- violation of linearity assumption

In such cases, a nonlinear transformation of variables might cure both problems.

Normality:: How to fix

- case separate models should be built

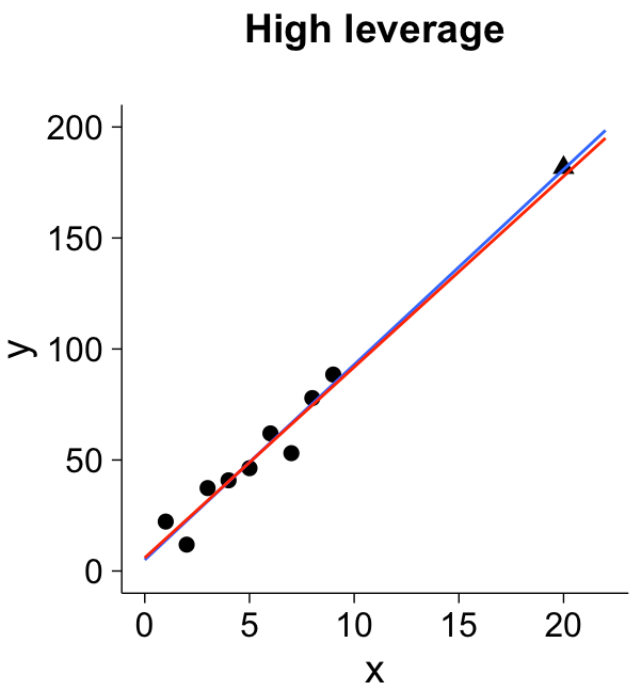
- some data point should be excluded
  if such events not likely to be repeated

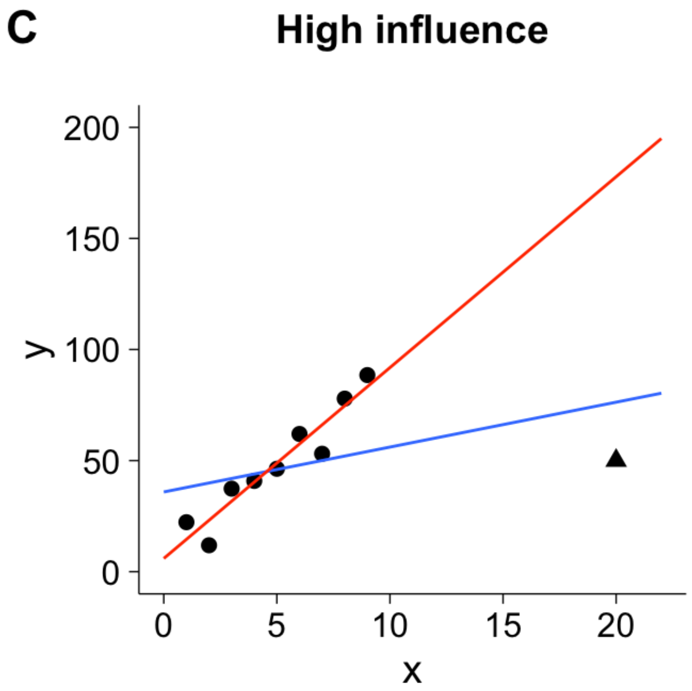
In addition to assumptions, user should concern unusual and influential data

- **Outliers**: observations with large residuals
- **Leverage**: measures the extent to which the predictor differs from the mean of the predictor; the red residual has lower leverage than the blue residual
- **Influence**: observations that have high leverage and are extreme outliers, changes coefficient estimates drastically if not included

Leverage:
focus on value of x

Influence:
focus on slop of the line



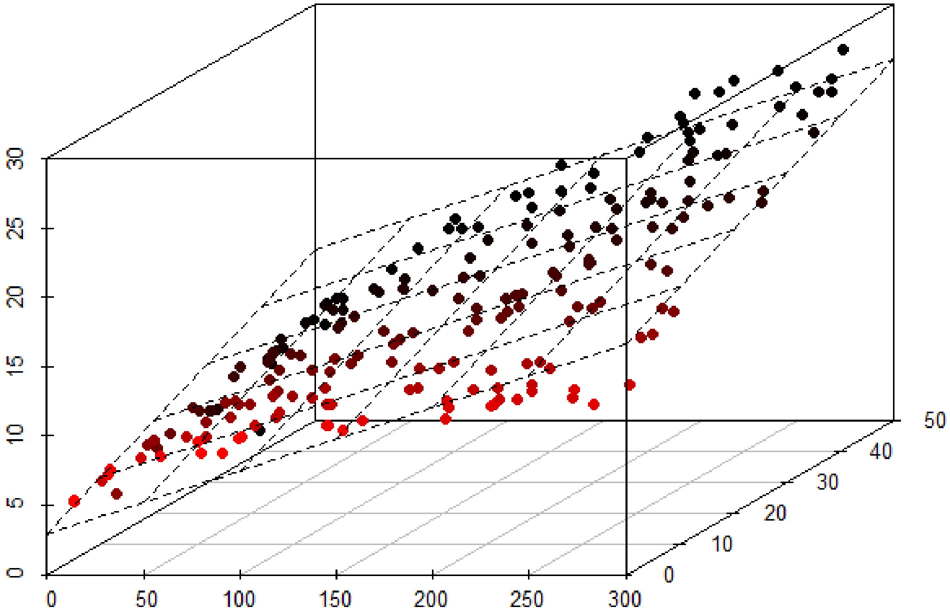


How to measure the influence of each observation

Cook's distance

- **usage** measure of <u>the influence of each observation</u> on the regression coefficients.
- **algorithm** extent of change in model estimates when that particular observation is omitted
- **interpretation** when Cook's distance is close to 1 → highly influential data point

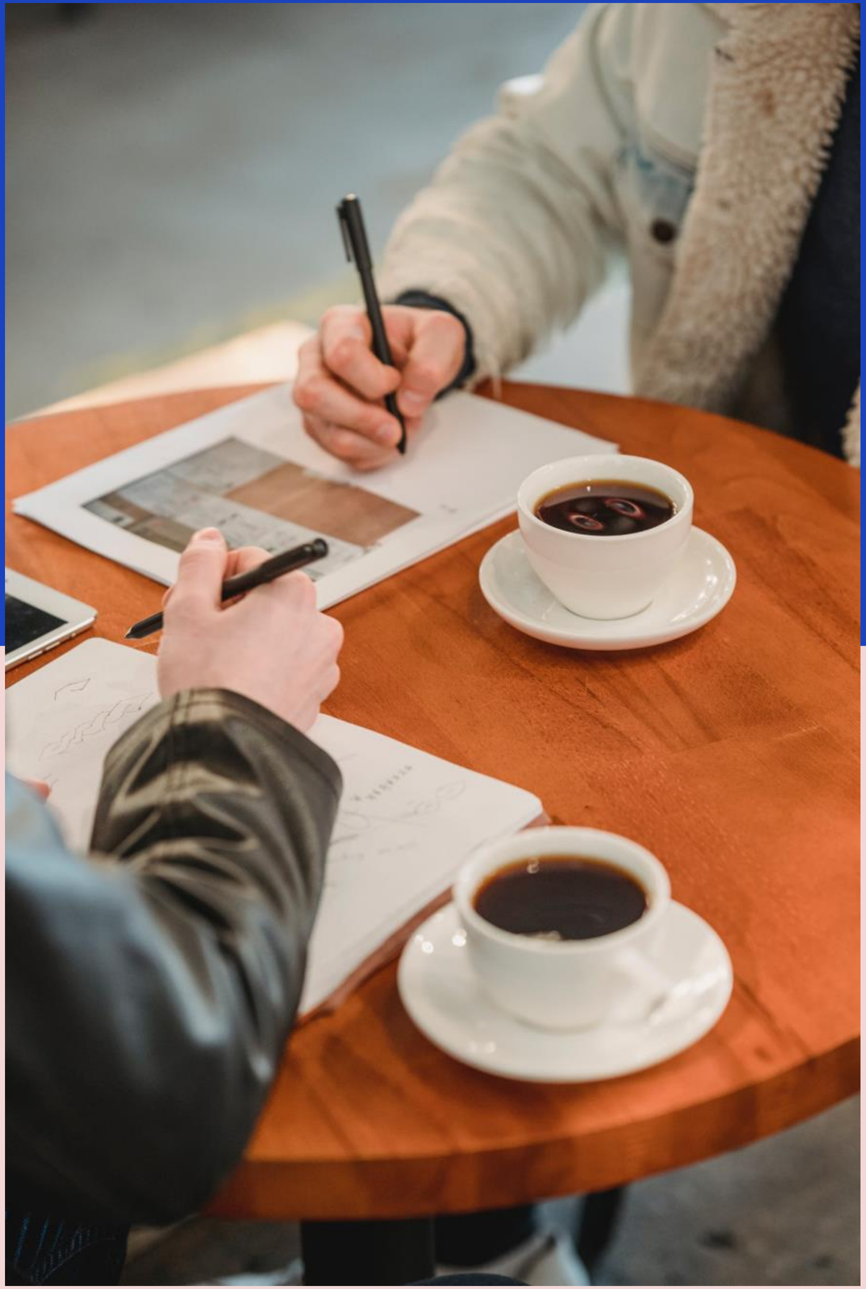
Multiple independent variables



$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

source: http://www.d4t4v1z.com/

Vector and Matrix notation



Vector notation

Obs.	Y	Independent variable			
		X_1	X_2	…	X_p
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
	…	…	…		…
n	y_n	x_{n1}	x_{n2}	…	x_{np}

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_3 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_i$$

Vector notation

Obs.	Y	Independent variable				
		X_0	X_1	X_2	…	X_p
1	y_1	1	x_{11}	x_{12}	…	x_{1p}
2	y_2	1	x_{21}	x_{22}	…	x_{2p}
	…		…	…		…
n	y_n	1	x_{n1}	x_{n2}	…	x_{np}

$$y_1 = \sum_{j=0}^{p} \beta_j x_{1j} + \varepsilon_1 \quad \Rightarrow \quad y_1 = \boldsymbol{\beta}^T \mathbf{x}_1 + \varepsilon_1$$

$$= \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_p \end{pmatrix} \begin{pmatrix} 1 \\ x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix} + \varepsilon_1$$

$$\Rightarrow \quad y_1 = \mathbf{x}_1^T \boldsymbol{\beta} + \varepsilon_1$$

$$= \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \varepsilon_1$$

Matrix notation

$$\mathbf{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

More general expression

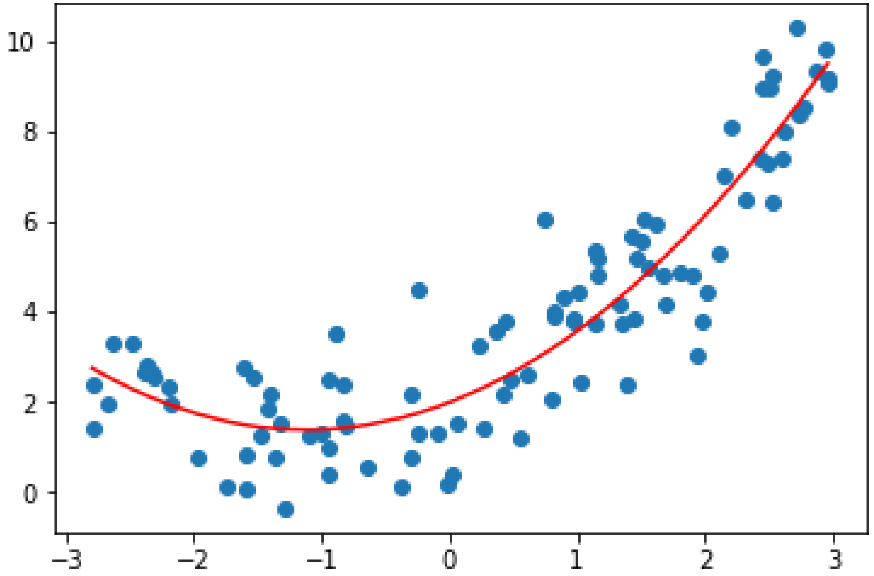
$$\mathbf{y} \quad = \quad \mathbf{H}\boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

$$\mathbf{H} = \begin{pmatrix} h_0(1) & h_1(x_{11}) & h_2(x_{12}) & \cdots & h_p(x_{1p}) \\ h_0(1) & h_1(x_{21}) & h_2(x_{22}) & \cdots & h_p(x_{2p}) \\ \vdots & \vdots & \vdots & & \vdots \\ h_0(1) & h_1(x_{n1}) & h_2(x_{n2}) & \cdots & h_p(x_{np}) \end{pmatrix}$$

where $h(x)$ is a non-linear transformation of the input x

This model is still called linear model as model parameter appear only linearly.

Example of non-linear transformation : Polynomial regression



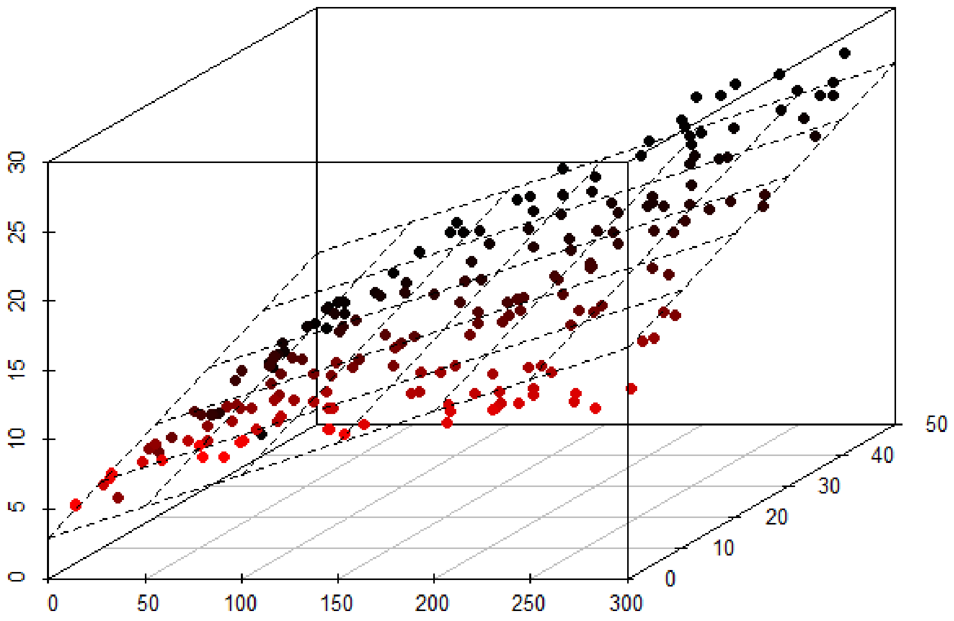
$$\mathbf{y} = \mathrm{h}^T(\mathbf{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $$\mathrm{h} = \begin{pmatrix} h_0(1) \\ h_1(x) \\ h_2(x) \\ \vdots \\ h_k(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x^2 \\ x^3 \\ \vdots \\ x^k \end{pmatrix}$$

This means that we “lift” the original one-dimensional input space into a *(K+1)*-dimensional feature space.

Cost function

Cost function



$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathrm{h}^T(\mathbf{x}_i)\boldsymbol{\beta} \right)^2$$

Cost function

Obs.	Y	Independent variable				
		X_0	X_1	X_2	…	X_p
1	y_1	1	x_{11}	x_{12}	…	x_{1p}
2	y_2	1	x_{21}	x_{22}	…	x_{2p}
	…		…	…		…
n	y_n	1	x_{n1}	x_{n2}	…	x_{np}

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \mathrm{h}^T(\mathbf{x}_i)\boldsymbol{\beta}\right)^2$$

Deriving the gradient

$$\nabla \mathrm{RSS}(\boldsymbol{\beta}) = \nabla\left[\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)^{T}\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)\right]$$

$$= -2\mathbf{H}^{T}\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)$$

$$\nabla \mathrm{RSS}(\boldsymbol{\beta}) = \nabla\left[\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)\right]$$

$$= -2\mathbf{H}^T\left(\mathbf{y}-\mathbf{H}\boldsymbol{\beta}\right)$$

$$\nabla \cdot (y - \square \beta)^{2.}$$

$$= 2(y - \square \beta) \square (-1)$$

Closed form solution

$$\nabla \mathrm{RSS}(\boldsymbol{\beta}) = -2\mathbf{H}^T\left(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\right)$$

$$-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$$

$$\mathbf{H}^T\mathbf{H}\boldsymbol{\beta} = \mathbf{H}^T\mathbf{y}$$

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

Framework for model evaluation

Creating model → Evaluating model

Performance evaluation

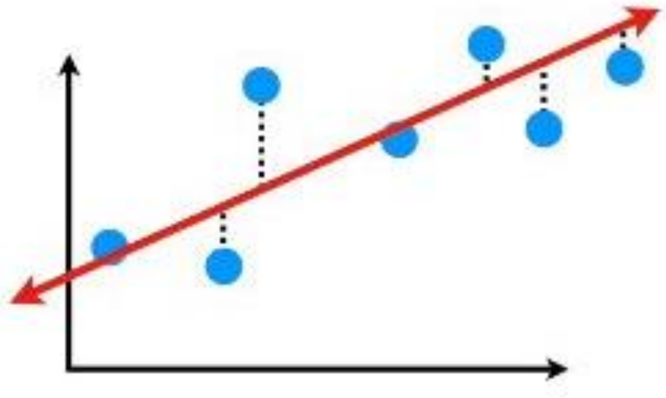
Obs.	y	$\hat{y}$

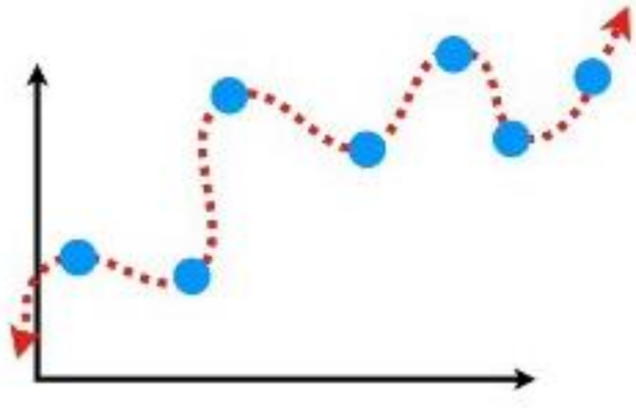
$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}\left(y_i - \mathrm{h}^T(\mathbf{x}_i)\boldsymbol{\beta}\right)^2}{n}$$

$$= \frac{\left(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\right)^T\left(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\right)}{n}$$

The bias-variance tradeoff

The bias-variance tradeoff





Low model complexity

High bias

Low variance

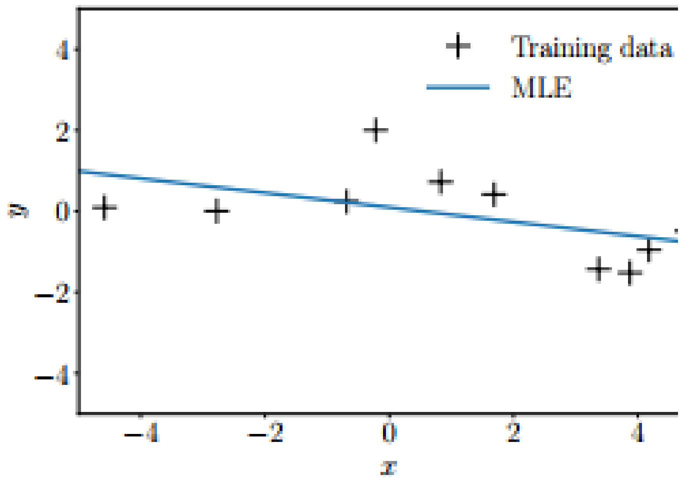
High model complexity

Low bias

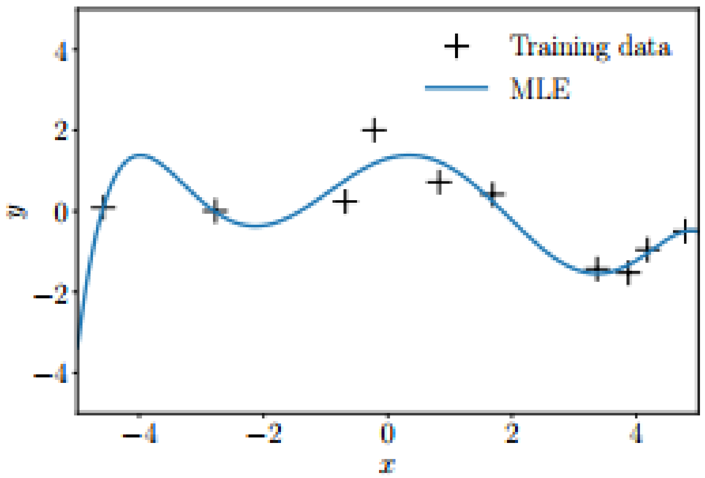
High variance

The bias-variance tradeoff

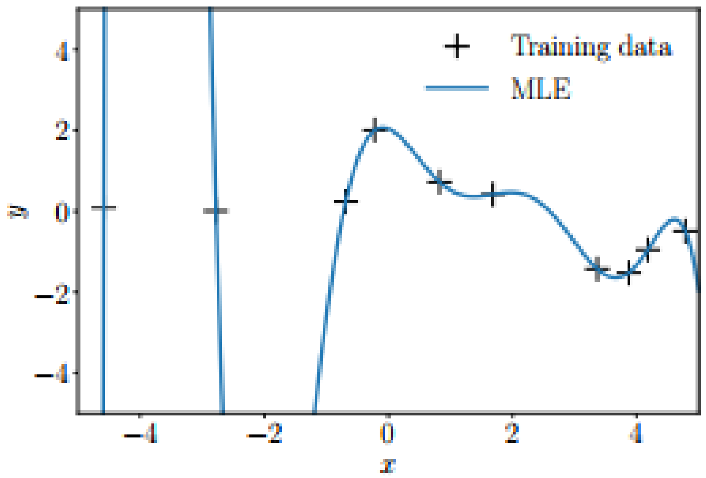
We notice that polynomials of low degree fit the training data poorly. When we go to higher-degree, it provide the better results of fitting.



Polynomial degree 1



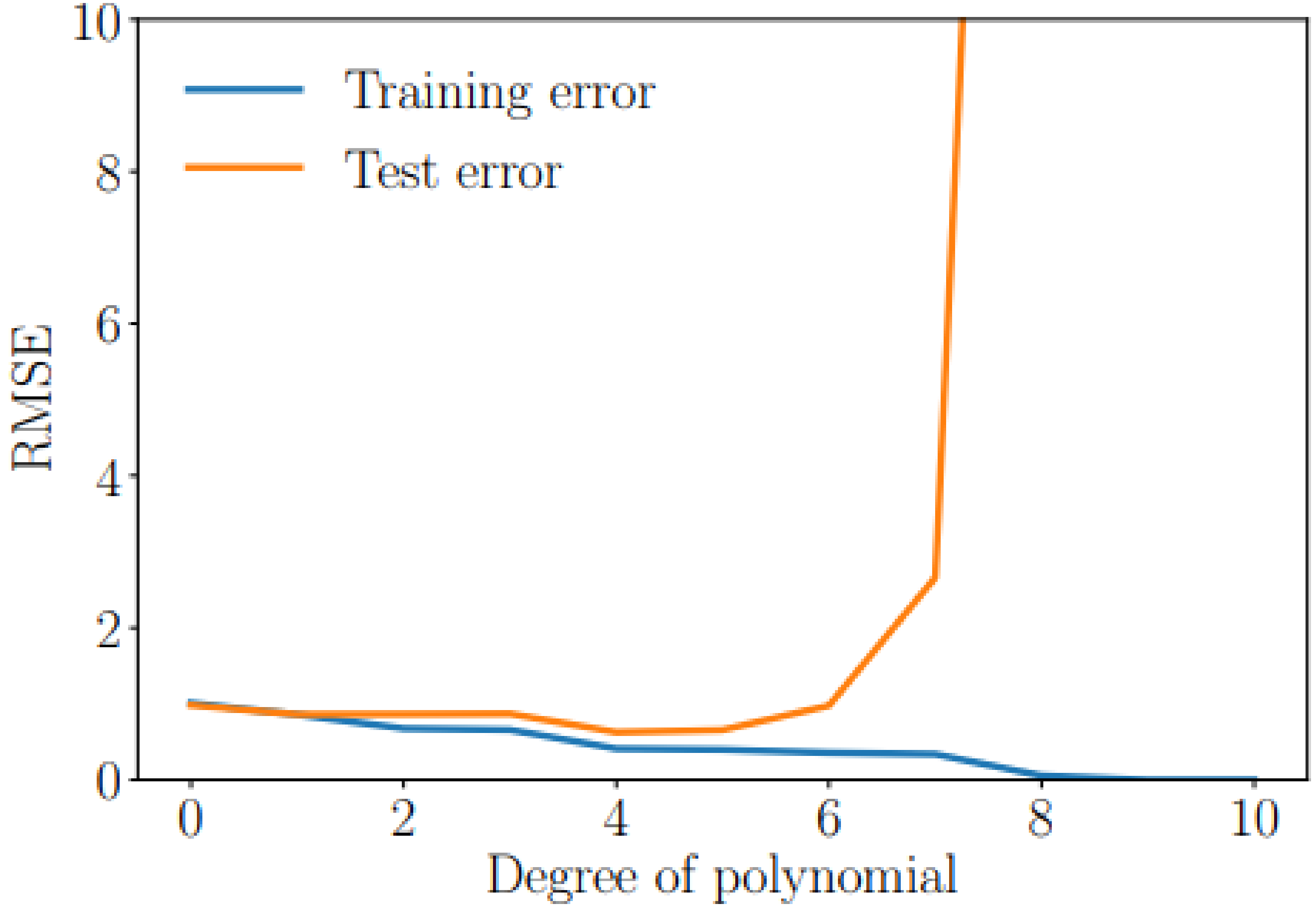
Polynomial degree 6



Polynomial degree 9

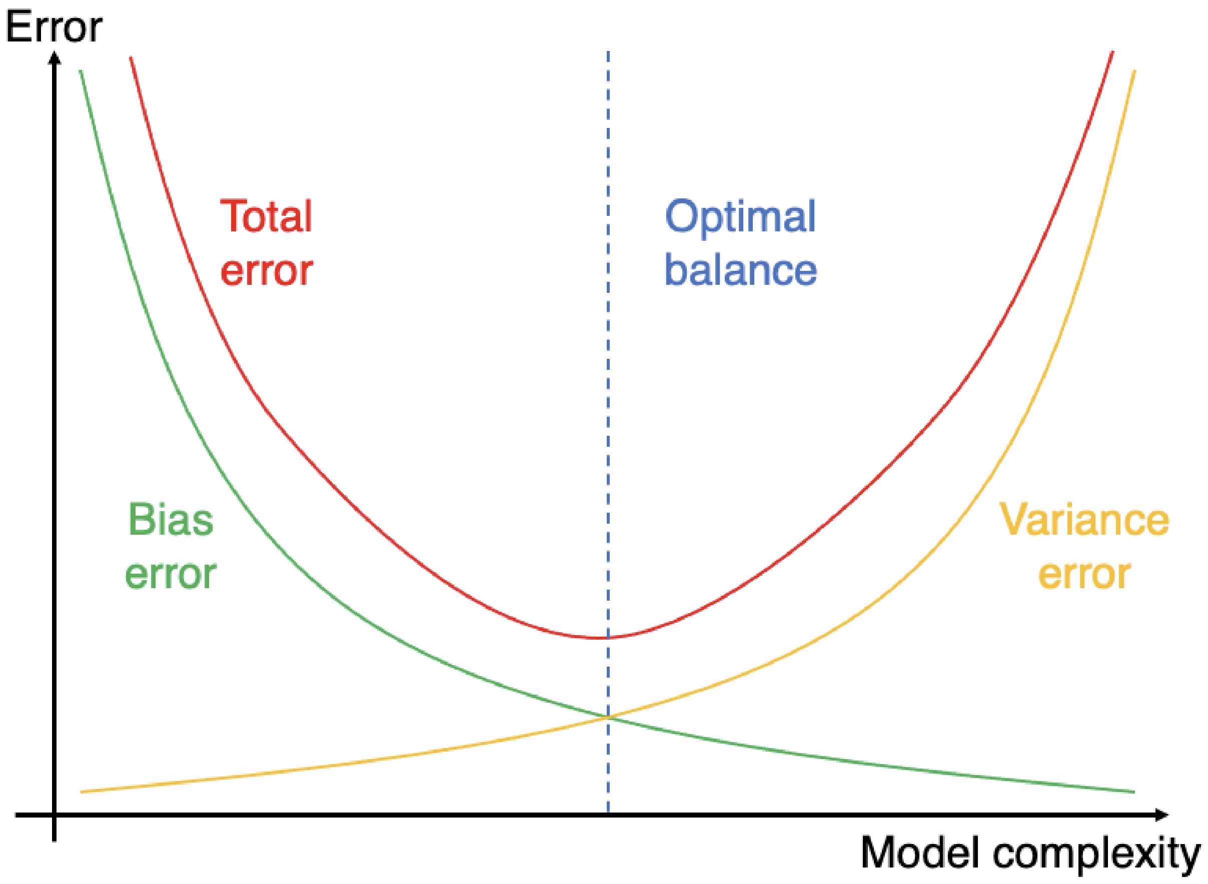
Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

The bias-variance tradeoff



Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

The bias-variance tradeoff



https://artsciencemillennial.substack.com/p/bias-variance-tradeoff-a-data-science

Ridge regression

When we increase the number of features, it can result in overfitting. To cope with this issue, the ridge regression introduces a penalty term by way of a tuning parameter called lambda. The idea is to make the fit small by making the residual sum or squares small plus adding a shrinkage penalty.

Cost function of ridge regression

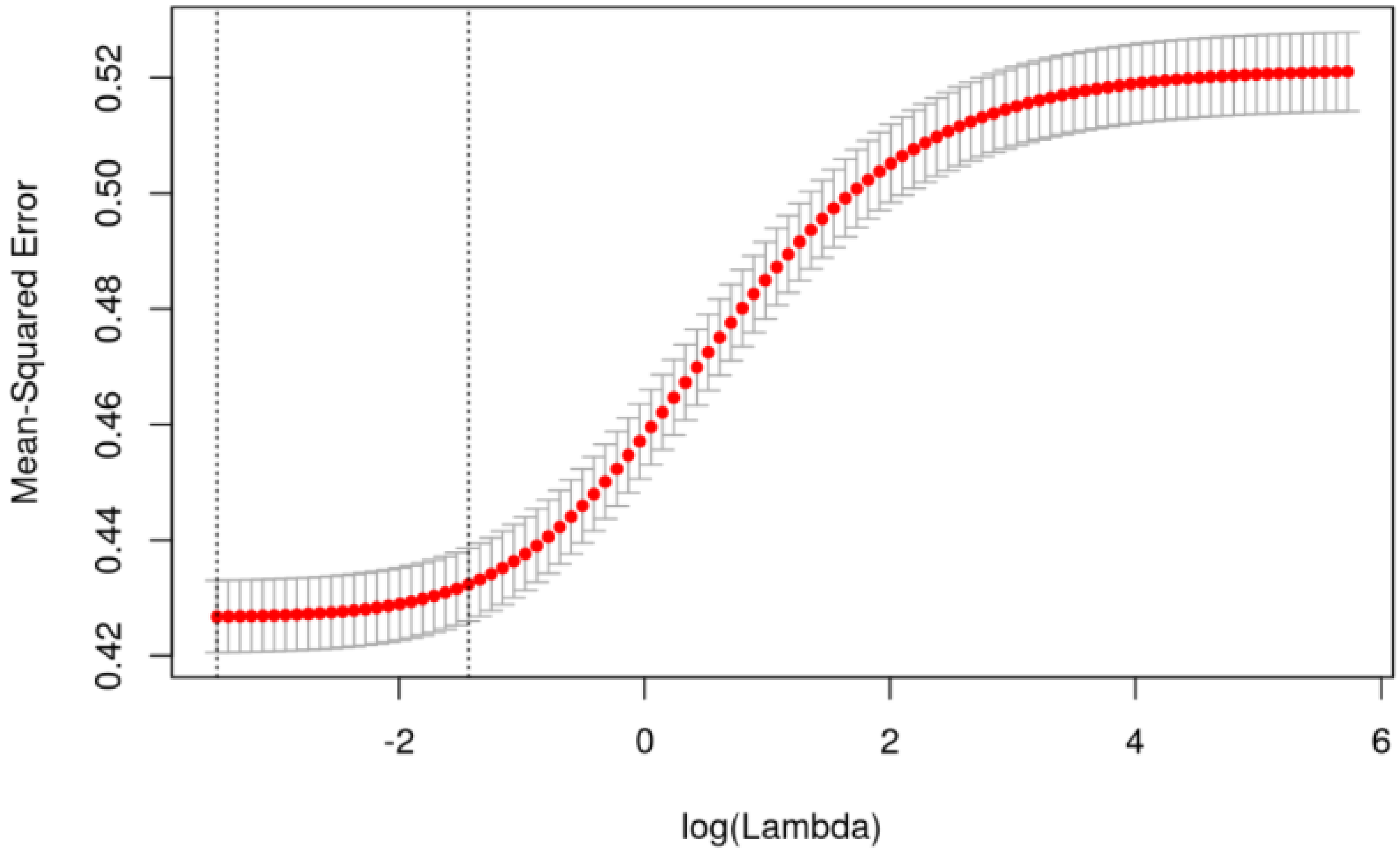
$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \mathrm{h}^T(\mathbf{x}_i)\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

where $\lambda \geq 0$

$$\|\beta\|_2^2 = \beta_0^2 + \beta_1^2 + \ldots + \beta_p^2$$

$$= \left(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\right)^T\left(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\right) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

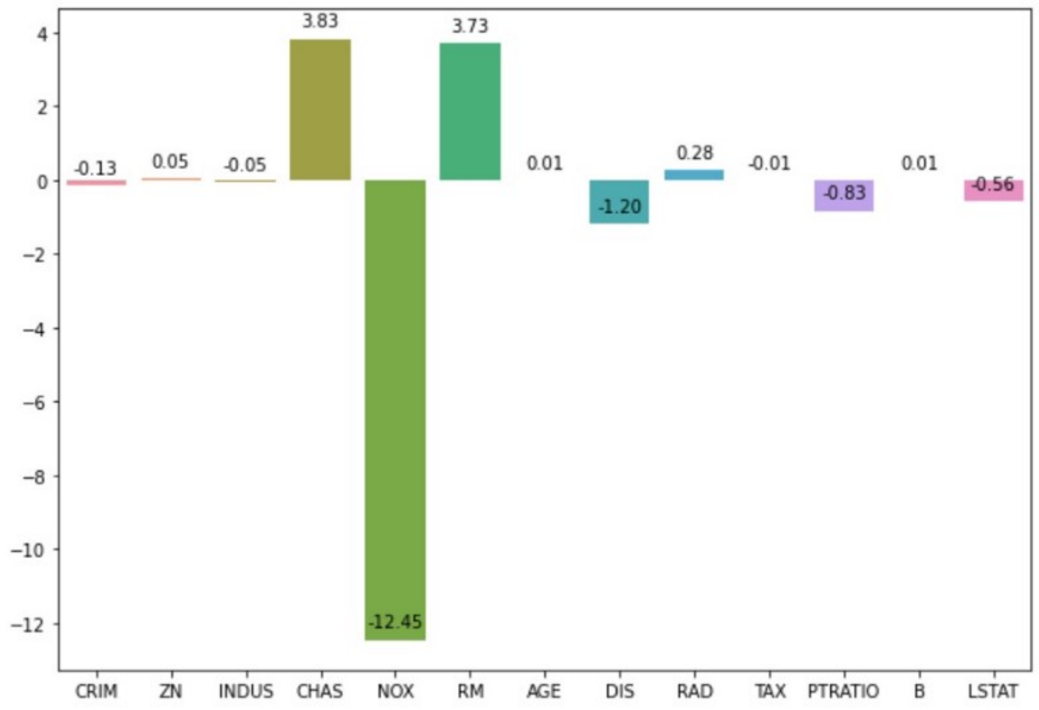
Tuning for penalty term



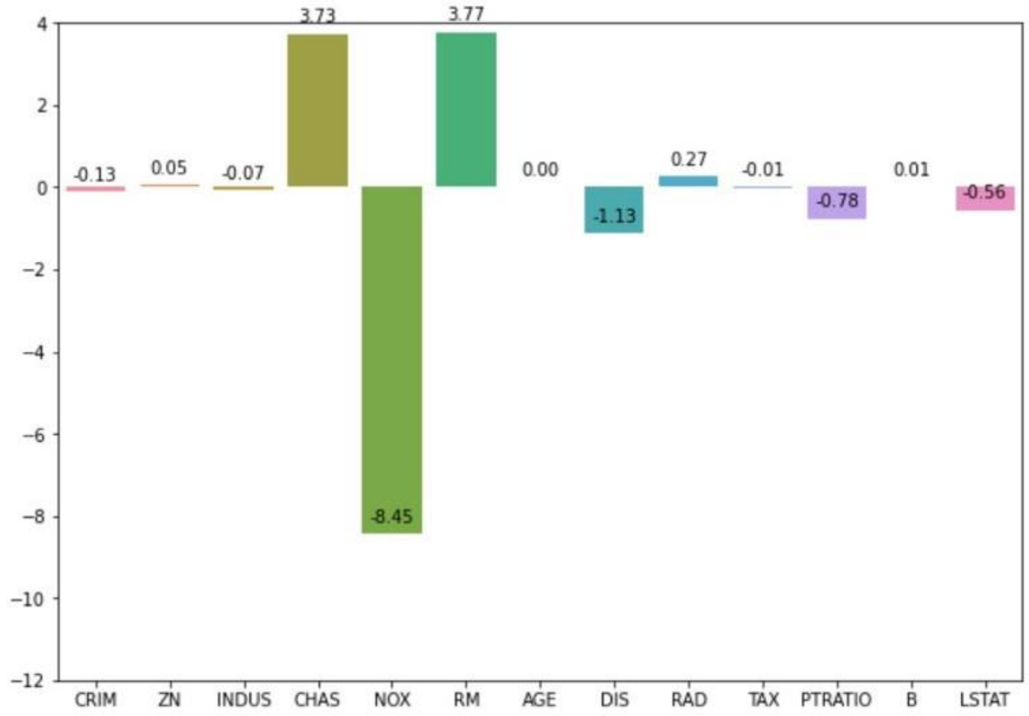
Source: http://wavedatalab.github.io/machinelearningwithr/post4.html

Magnitude of coefficients

Linear Regression

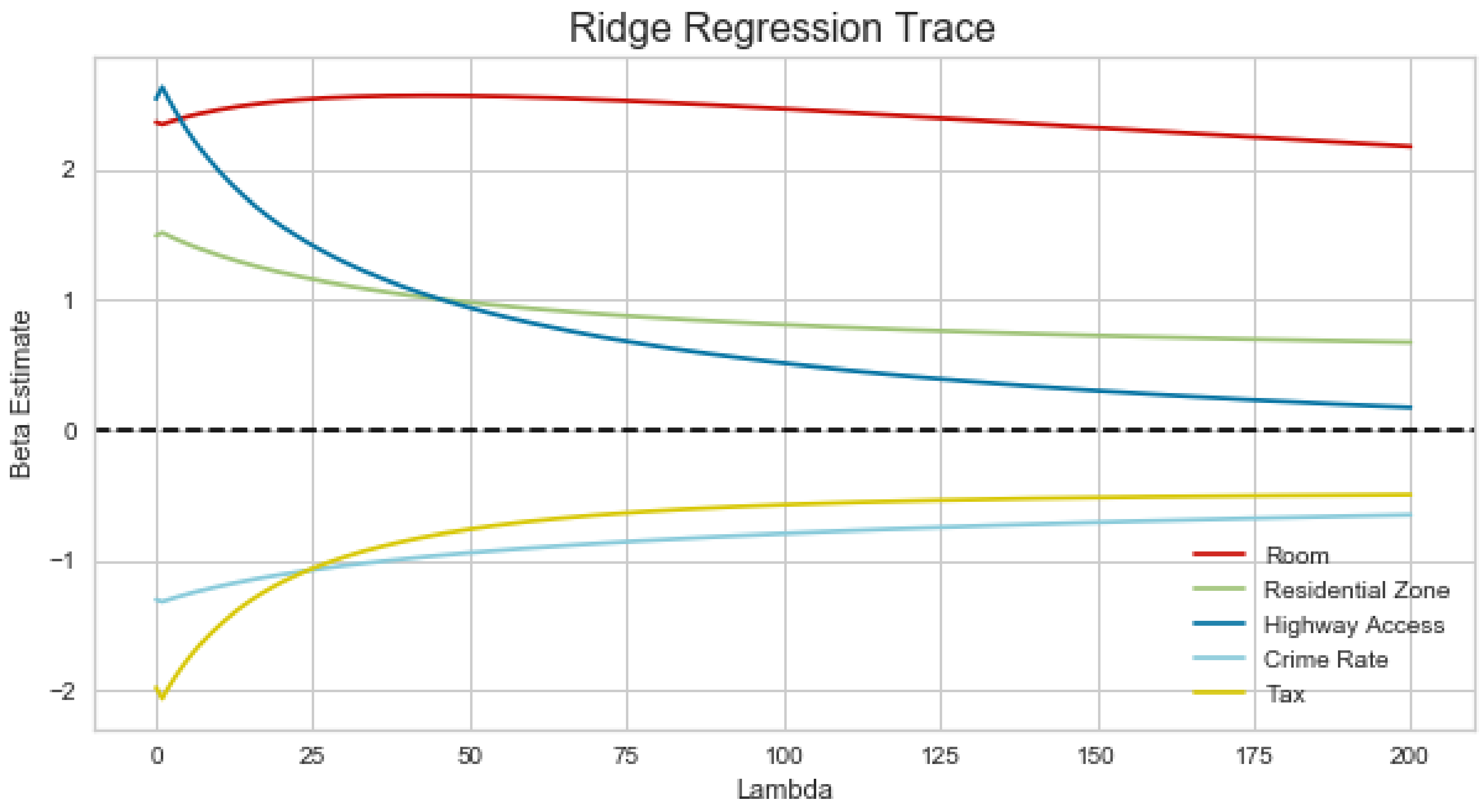


Ridge Regression



source: https://medium.com/geekculture/ridge-and-lasso-regression-51705b608fb9

Ridge coefficients and lambda value



https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Lasso regression

In lasso, the penalty is the sum of the absolute values of the coefficients. Lasso shrinks the coefficient estimates towards zero and it has the effect of setting variables exactly equal to zero when lambda is large enough while ridge does not. Hence, much like the best subset selection method, lasso performs variable selection.

Cost function of lasso regression

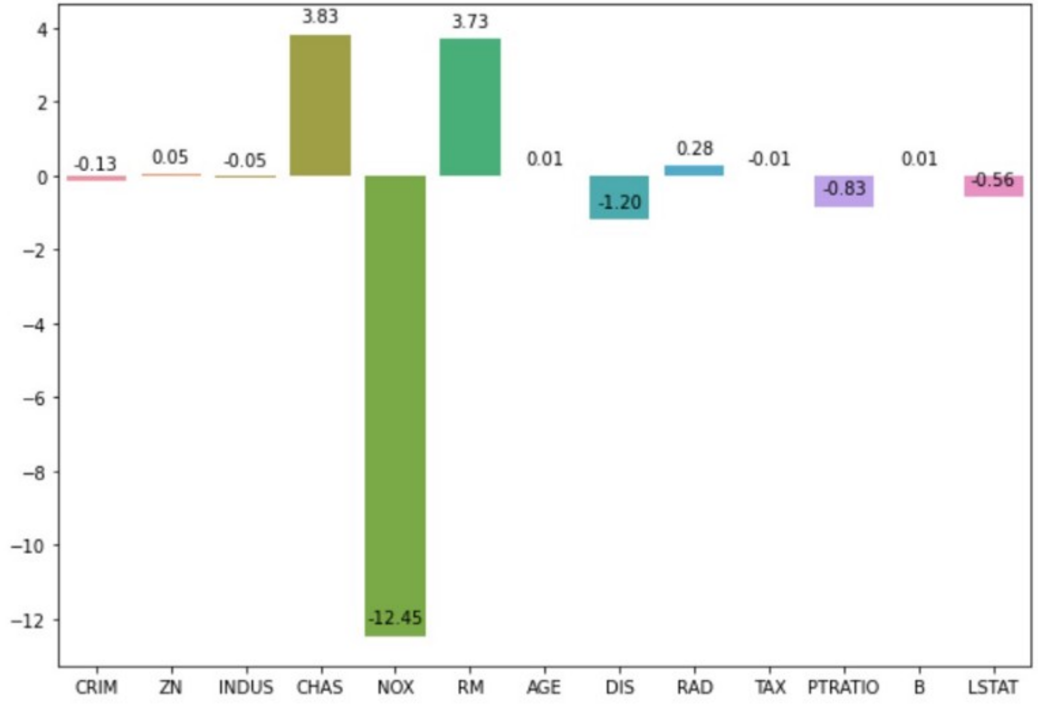
$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \mathrm{h}^T(\mathbf{x}_i)\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad \text{where } \lambda \geq 0$$

$$\|\beta\|_1 = |\beta_0| + |\beta_1| + \ldots + |\beta_p|$$

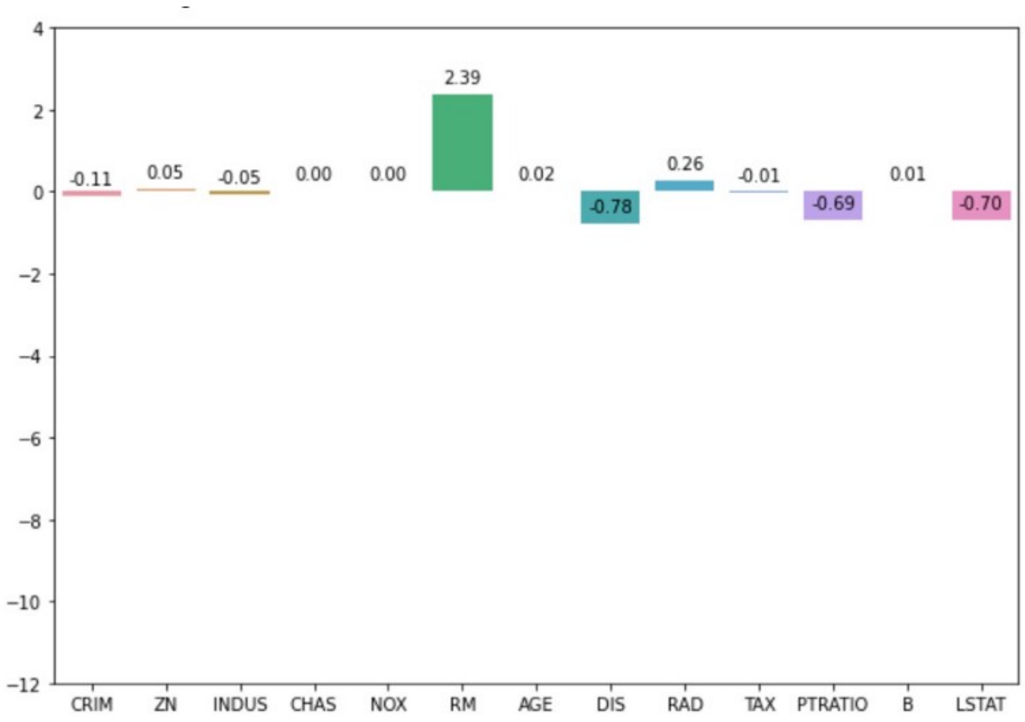
can be solved by coordinate descent

Magnitude of coefficients

Linear Regression

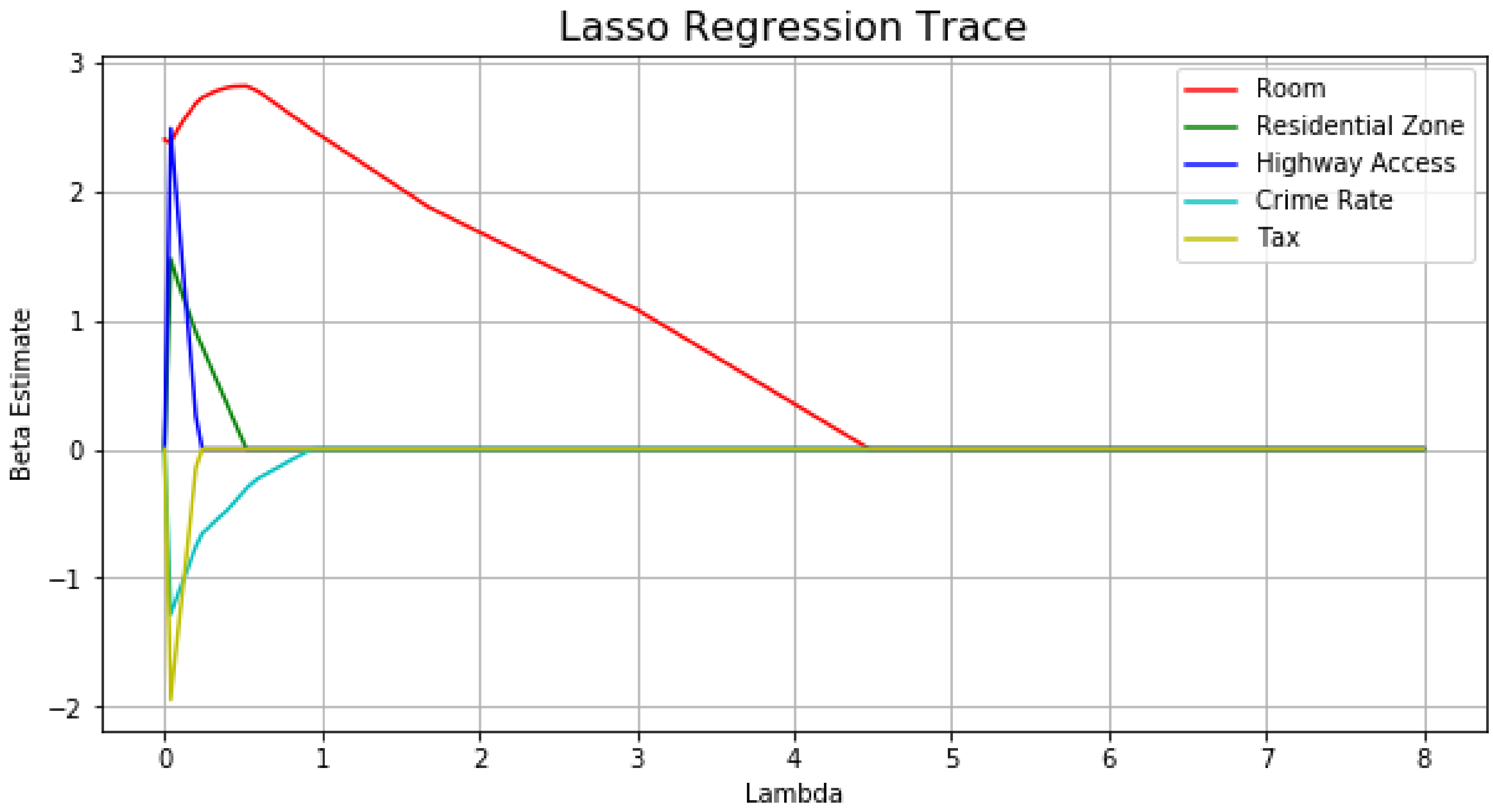


Lasso Regression



source: https://medium.com/geekculture/ridge-and-lasso-regression-51705b608fb9

Lasso coefficients and lambda value



https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Additional topic on linear regression

Log Transformations

- In standard mathematical notation and in Excel

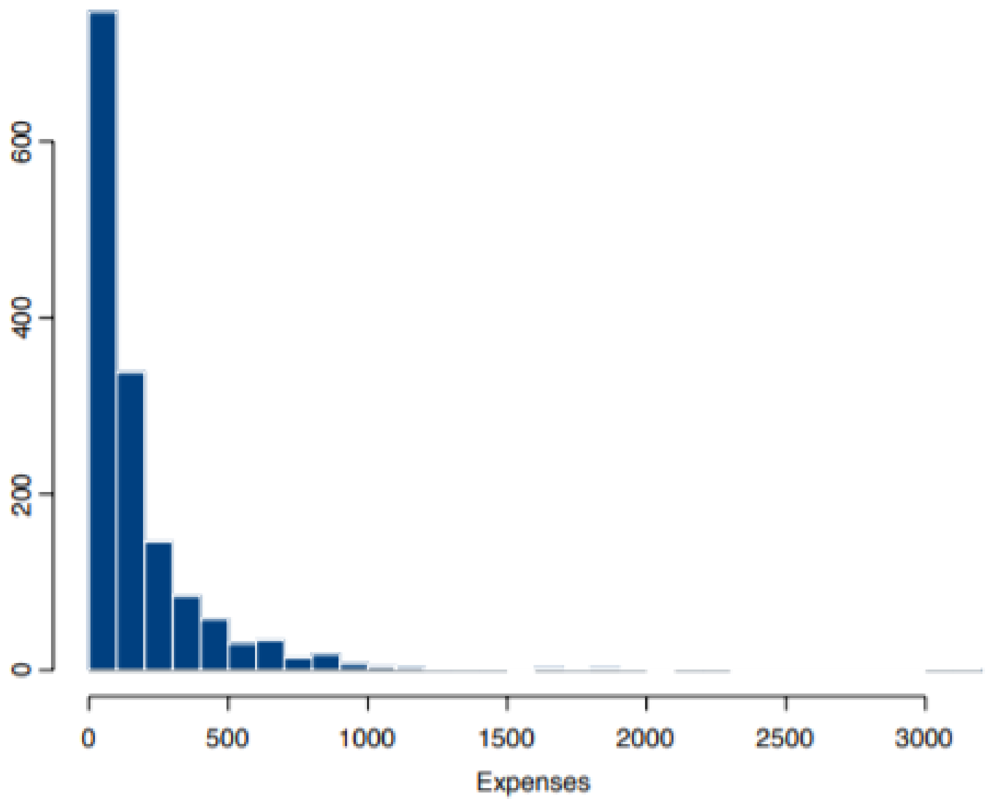
  LN(X) is the natural log of X

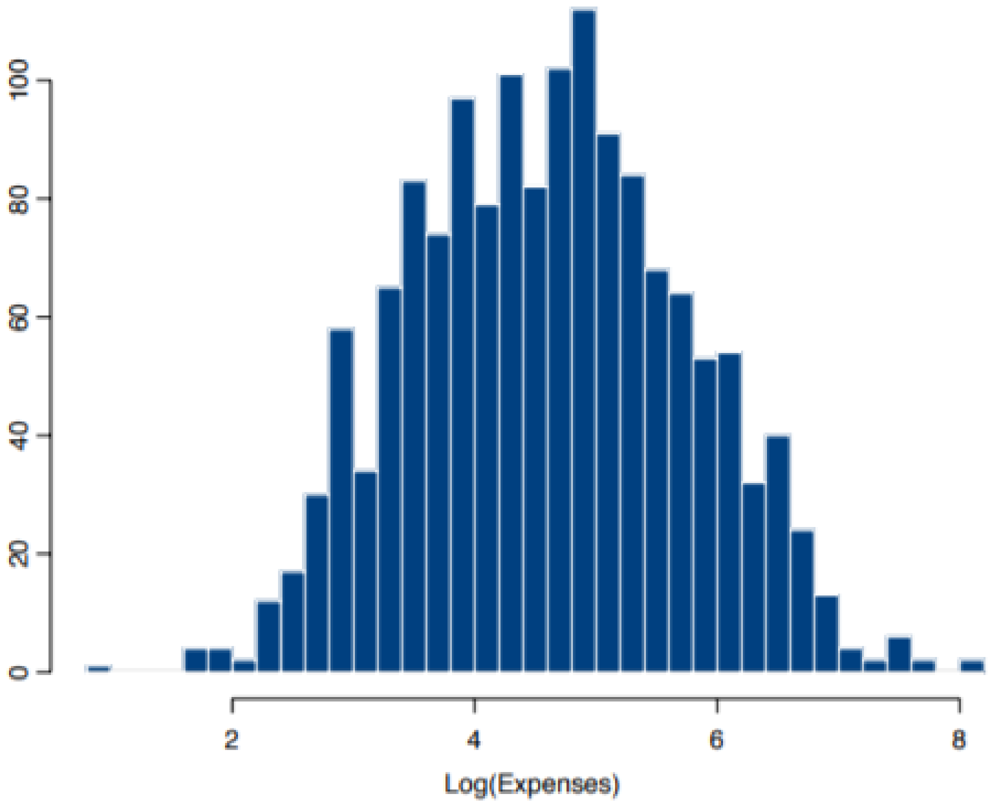
  LOG(X) is often used for the base-10 log

- In statistics method and most of statistical software such as R and SAS

  the function that is called LOG is the natural log

Why we employ Log transformations ?





ที่มา https://kenbenoit.net

- Only the dependent/response variable is log-transformed.
- Only independent/predictor variable(s) is log-transformed.
- Both dependent/response variable and independent/predictor variable(s) are log-transformed.

Interpreting Log transformations in a linear model

- **Only the dependent/response variable is log-transformed : log linear model**

$$\log Y = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$\hat{\beta}_1$:one-unit increase in x_1 will produce an expected increase in $\log Y$ of $\hat{\beta}_1$ units

$$Y = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1) = \exp(\hat{\beta}_0)\exp(\hat{\beta}_1 x_1)$$

Interpreting Log transformations in a linear model

Additive relationship	Multiplicative relationship
$Y = \hat{\beta}_0 + \hat{\beta}_1 x_1$	$Y = \exp(\hat{\beta}_0)\exp(\hat{\beta}_1 x_1)$

Ex.

$$Y = 1.7 + 1.22x_1$$

$$Y = \exp(0.53)\exp(0.198x_1)$$
$$= e^{0.53}\left(e^{0.198}\right)^{x_1}$$

One-unit increase in x_1

We add 1.22 to Y value.

We multiply exp(0.198) to Y value.

Interpreting Log transformations in a linear model

$$\log(\text{price}) = 12.15954 \; + \; 0.000417\text{sqft\_living}$$

$$\exp(0.000417) = 1.000417$$

For every one-unit increase in sqft_living, price is multiplied by about 1.000417.

Interpreting Log transformations in a linear model

- **Only independent/predictor variable(s) is log-transformed**

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1)$$

1% increase in the independent variable

increases (or decreases) the dependent variable by ($\hat{\beta}_1$ /100) units

$$(\hat{\beta}_0 + \hat{\beta}_1 \log(101)) - (\hat{\beta}_0 + \hat{\beta}_1 \log(100)) \quad = \hat{\beta}_1 \log(1.01)$$

$$= 0.00995(\hat{\beta}_1)$$

$$\approx 0.01(\hat{\beta}_1)$$

Next Week

Practical guide to linear regression with Python

วิทยากร คุณภาคิน เตชธีรโภคิน

ตำแหน่ง Data Scientist

บริษัท ทรู คอร์ปอเรชั่น จำกัด (มหาชน)

<u>ศิษย์เก่ารุ่นที่ 1 ของหลักสูตรวิทยาการข้อมูลและการวิเคราะห์</u>