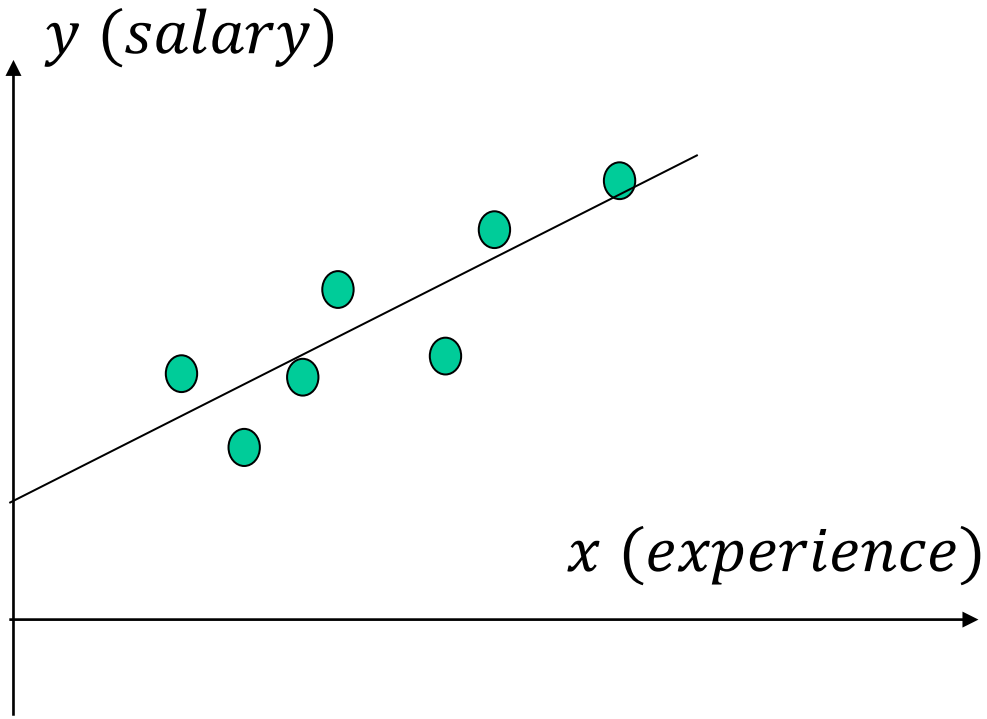
# 8. Regression

In previous parts, you have learned some popular models of classification used to solve a wide array of prediction problems where the target variable is a categorical variable such as “normal” vs “abnormal”, “positive” vs “negative” etc. In this part, we will build concepts on prediction of numerical variables – which is another key area of supervised learning known as regression.

## 8.1 Simple and multiple linear regression models

Simple linear regression roots from the statistical concept of fitting a straight line as shown below, where the experience is the predictor variable, also known as independent variable, while the salary is the target variable, also known as dependent variable.




Simple linear regression, also known as univariate regression, uses a single predictor variable and a linear function to predict the target variable as given below:

$$y = \theta_0 + \theta_1 x + \varepsilon$$

Where

$y$ --- target (dependent) variable
$x$ --- predictor (independent) variable
$\theta_0$ --- intercept
$\theta_1$ --- slope
$\varepsilon$---residual (error)

In **multiple linear regression**, more than one independent variables are used to predict the dependent variable as shown below:

$$y = \theta_0 + \theta_1 x_1 + \cdots \theta_m x_m + \varepsilon$$

Where $x_1$, $x_2$,…, $x_m$ are the independent variables, $\theta_0$, $\theta_1$, $\theta_2$,…, $\theta_m$ are the coefficients or parameters.

## 8.2 Ordinary least square estimation

Assume there are $N$ training data pairs (labelled data):

$$\{\mathbf{x}_1, y_1\}\ \{\mathbf{x}_2, y_2\}\ \cdots\ \{\mathbf{x}_N, y_N\}$$

and

$$\mathbf{x_i} = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{im}]^T$$

For the $i^{th}$ training data pair:

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots \theta_m x_{im} + \varepsilon_i$$

The $N$ equations, corresponding to the $N$ training data pairs, can be summarized into a matrix equation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nm} \end{bmatrix} \times \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

Define:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \qquad \mathbf{\Phi} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nm} \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

We have:

$$\mathbf{y} = \mathbf{\Phi} \times \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Define the following loss function:

$$J = \frac{1}{2}\sum_{i=1}^{N} \varepsilon_i^2 = \frac{1}{2}\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$$

The loss function $J$ is a function of $\boldsymbol{\theta}$. We want to find such $\boldsymbol{\theta}$ that the loss function is minimized:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

$J(\boldsymbol{\theta})$ can be expanded as:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N} \varepsilon_i^2 = \frac{1}{2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \frac{1}{2}(\mathbf{y} - \boldsymbol{\Phi\theta})^T(\mathbf{y} - \boldsymbol{\Phi\theta})$$

Then:

$$\frac{\boldsymbol{\partial} J(\boldsymbol{\theta})}{\boldsymbol{\partial\theta}} = \boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi\theta}) = \boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\boldsymbol{\Phi\theta}$$

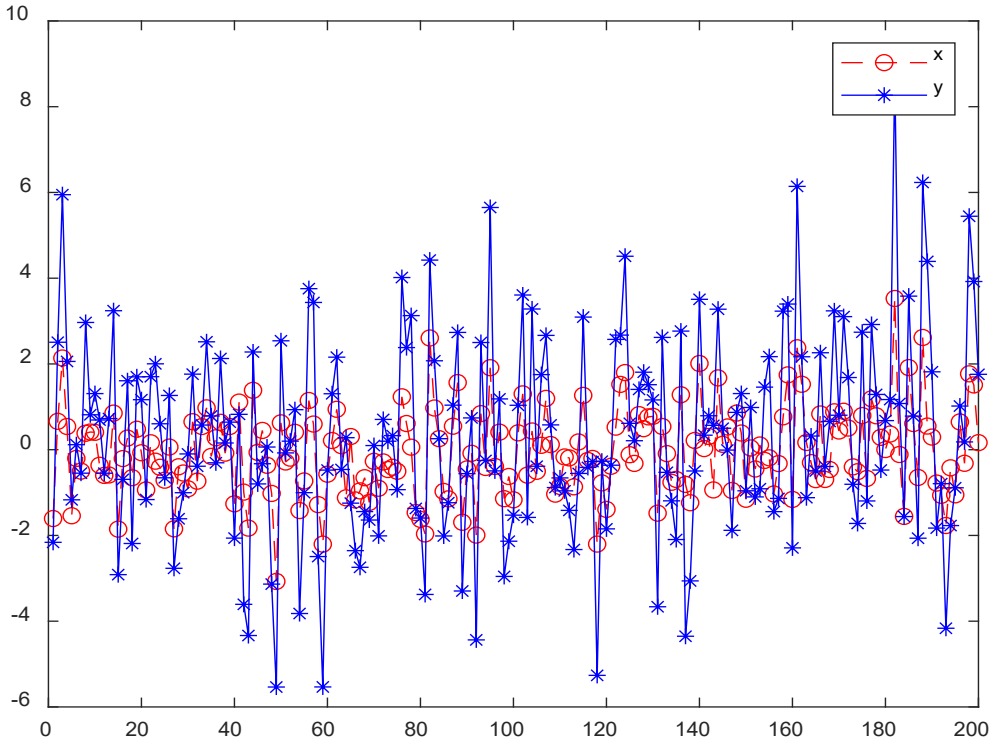
Solving $\frac{\boldsymbol{\partial} J(\boldsymbol{\theta})}{\boldsymbol{\partial\theta}} = 0$, obtains:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$$

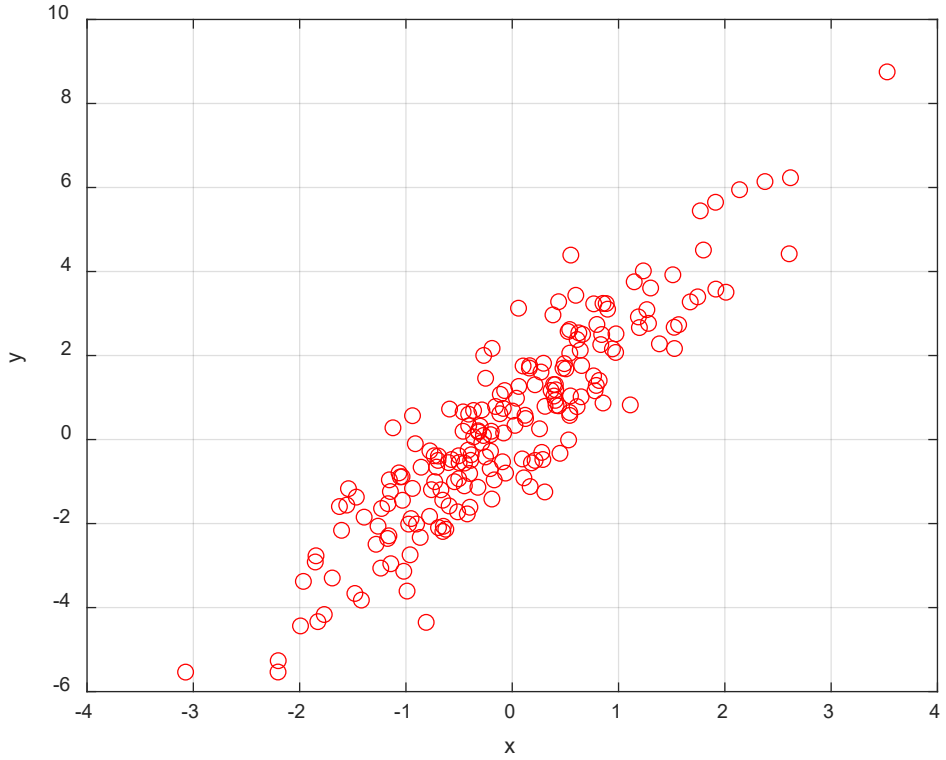
This estimate of $\boldsymbol{\theta}$ is called ordinary least square (OLS) estimation since it minimises the squared error.

## Example

The following figure shows 200 data pairs:



The scatter plot of $y$ against $x$ is as follow:




which shows that $y$ and $x$ may have the following linear relationship, where $\theta_0$ and $\theta_1$ are unknown parameters:

$$y = \theta_0 + \theta_1 x$$

To estimate the values of $\theta_0$ and $\theta_1$, we may use the OLS estimation method.

Define:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{200} \end{bmatrix} \qquad \mathbf{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{200} \end{bmatrix}$$

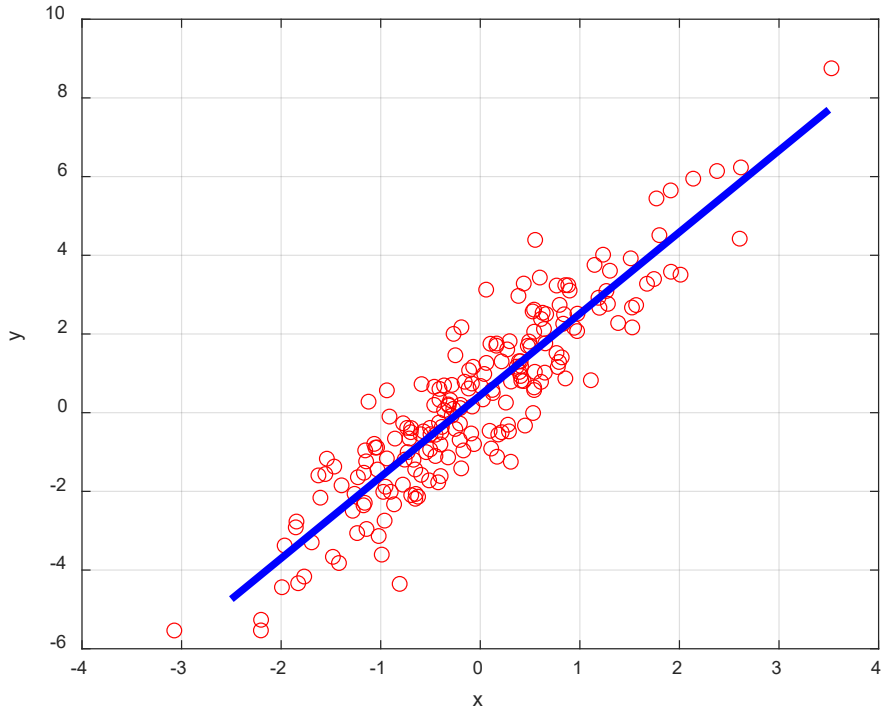
$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

Then the OLS estimate of $\boldsymbol{\theta}$ is obtained as follows:

$$\widehat{\boldsymbol{\theta}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{y} = \begin{bmatrix} 0.4433 \\ 2.0728 \end{bmatrix}$$

Thus, we obtain the following simple regression model:

$$y = \theta_0 + \theta_1 x = 0.4433 + 2.0728x$$




The above line shows that the OLS estimate well capture the relationship between $y$ and $x$.

Actually, the 200 data pairs are generated by using the following equation, where $\varepsilon$ is a Gaussian random noise with zero mean and standard deviation of one:

$$y = 0.5 + 2x + \varepsilon$$

The estimation error of $\theta_0$ and $\theta_1$, i.e. the difference between the estimated value and the true value, is due to the noise $\varepsilon$. The higher the noise level, the larger the estimation error.

If the standard deviation of $\varepsilon$ is reduced to 0.5, 0.1 and 0 (noise free), respectively, the OLS estimates obtained are:

$$\boldsymbol{\theta}_{\sigma_\varepsilon=0.5} = \begin{bmatrix} 0.4717 \\ 2.0364 \end{bmatrix} \qquad \boldsymbol{\theta}_{\sigma_\varepsilon=0.1} = \begin{bmatrix} 0.4943 \\ 2.0073 \end{bmatrix}$$

$$\boldsymbol{\theta}_{noisefree} = \begin{bmatrix} 0.5 \\ 2.0 \end{bmatrix}$$

# 8.3 Ridge regression

For ordinary least square (OLS), we have the following observations:

(1) It is not unusual to see the number of predictor (independent) variables greatly exceed the number of observations. With many predictors, fitting the full model without penalization, the OLS estimate may not uniquely exist.

(2) Ill-conditioned $\boldsymbol{\Phi}$. OLS estimates depend on $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$, we will have problems in computing the OLS estimate if $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ is singular or nearly singular.

(3) The OLS estimate may provide a good fit to the training data, but it may not generalize well to the test data.

One solution to the above problems is ridge regression, which introduces a penalty term on the parameters to the loss function:

$$J_{ridge}(\boldsymbol{\theta}) = \frac{1}{2}\left\{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=0}^{m}\theta_i^2\right\}$$

$$= \frac{1}{2}\{(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^T(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\}$$

Where weight $\lambda$ is a real-valued positive number.

Solving $\frac{\boldsymbol{\partial} J_{ridge}(\boldsymbol{\theta})}{\boldsymbol{\partial}\boldsymbol{\theta}} = 0$, obtains:

$$\widehat{\boldsymbol{\theta}}_{ridge} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I})^{-\mathbf{1}}\boldsymbol{\Phi}^T\mathbf{y}$$

Where $\mathbf{I}$ is a $(m+1) \times (m+1)$ identity matrix.

Ridge regression imposes constraints on the parameters in the linear model, taking the form of squared norm of the parameter vector. This means that if the parameters take on large values, the optimization function is penalized. Ridge regression shrinks the parameters towards zero.

Ridge regression is a kind of so-called regularization technique, which is an important concept in machine learning. Regularization aims to avoid overfitting of the data, especially when the trained and test data are much varying.

Regularization is implemented by adding a “penalty” term to the best fit derived from the trained data, to achieve a *lesser variance* with the tested data and also restricts the influence of predictor variables over the target variable by compressing their coefficients.

## 8.4 Lasso regression

Ridge regression is also named as L2-regulation since the penalty term is the L2-norm of the parameter vector.

In contrast, Least Absolute Shrinkage and Selection Operator (Lasso) is L1-regularization, where the penalty term is the L1-norm of the parameter vector as shown below:

$$J_{lasso}(\boldsymbol{\theta}) = \frac{1}{2}\left\{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{m}\left|\theta_j\right|\right\}$$

$$= \frac{1}{2}\{(\mathbf{y} - \boldsymbol{\Phi\theta})^T(\mathbf{y} - \boldsymbol{\Phi\theta}) + \lambda\|\boldsymbol{\theta}\|\}$$

Where $\left|\theta_j\right|$ and $\|\boldsymbol{\theta}\|$ denote the absolute value of scalar $\theta_j$ and L1-norm of vector $\boldsymbol{\theta}$, respectively.

By comparing the loss functions of the ridge regression and Lasso regression, the difference is subtle: L2-penalty and L1-penalty are used, respectively. However, this subtle difference brings important change. Lasso regression selects only the most important predictor variables for predicting the target variable by shrinking the regression coefficients associated with the least important predictor variables to zero. Thus, Lasso regression produces sparse models.

Unlike ridge regression, there is no analytic solution for the Lasso. Numerical algorithms to optimize the objective function in Lasso include iterative shrinkage threshold algorithm (ISTA), fast iterative shrinkage-thresholding algorithms (FISTA) etc. In practice, we use the Lasso function in toolbox such as Matlab.

The choice of the regularization parameter λ is crucial in Lasso regression. A larger λ value increases the amount of regularization, leading to more coefficients being pushed towards zero. Conversely, a smaller λ value reduces the regularization effect, allowing more variables to have non-zero coefficients.

# 8.5 Model Evaluation

The essential step in any machine learning model is to evaluate the accuracy of the model. The Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and R-Squared or Adjusted R-Squared metrics are commonly used to evaluate the performance of the model. These metrics are often used on both training and test datasets.

## (1) Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$y_i$: actual value of target variable of the $i^{th}$ sample
$\hat{y}_i$: predicted value of target variable of the $i^{th}$ sample
$n$: number of samples

## (2) Root Mean Squared Error (RMSE)

The root mean squared error is the square root of mean squared error. It measures the standard deviation of residuals:

$$\text{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

## (3) Mean Absolute Error (MAE)

The mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

## (4) R-squared ($R^2$)

R-squared is defined as the fraction by which the variance of the errors is less than the variance of the dependent variable.

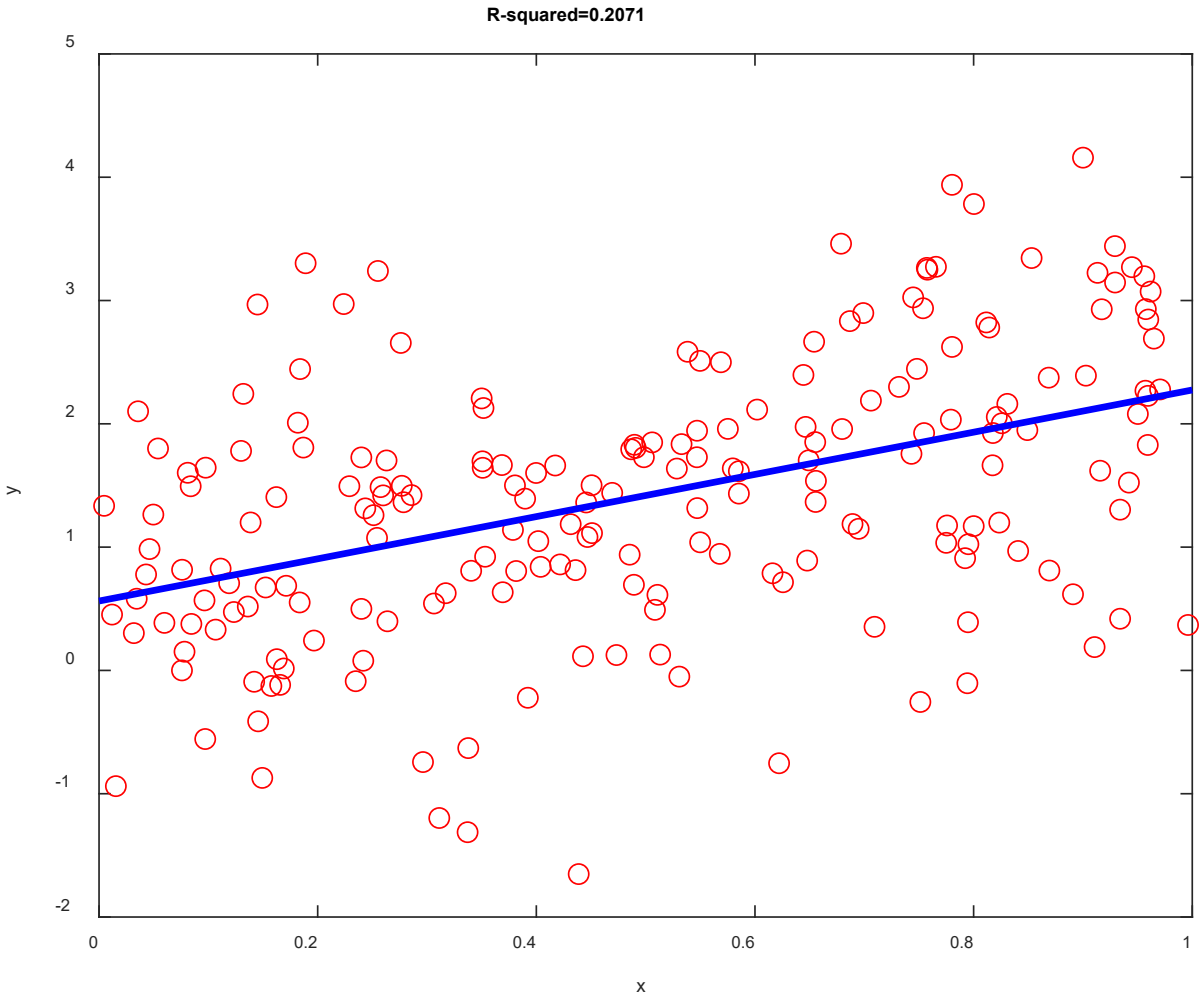
$$R^2 = 1 - \frac{RSS}{TSS}$$

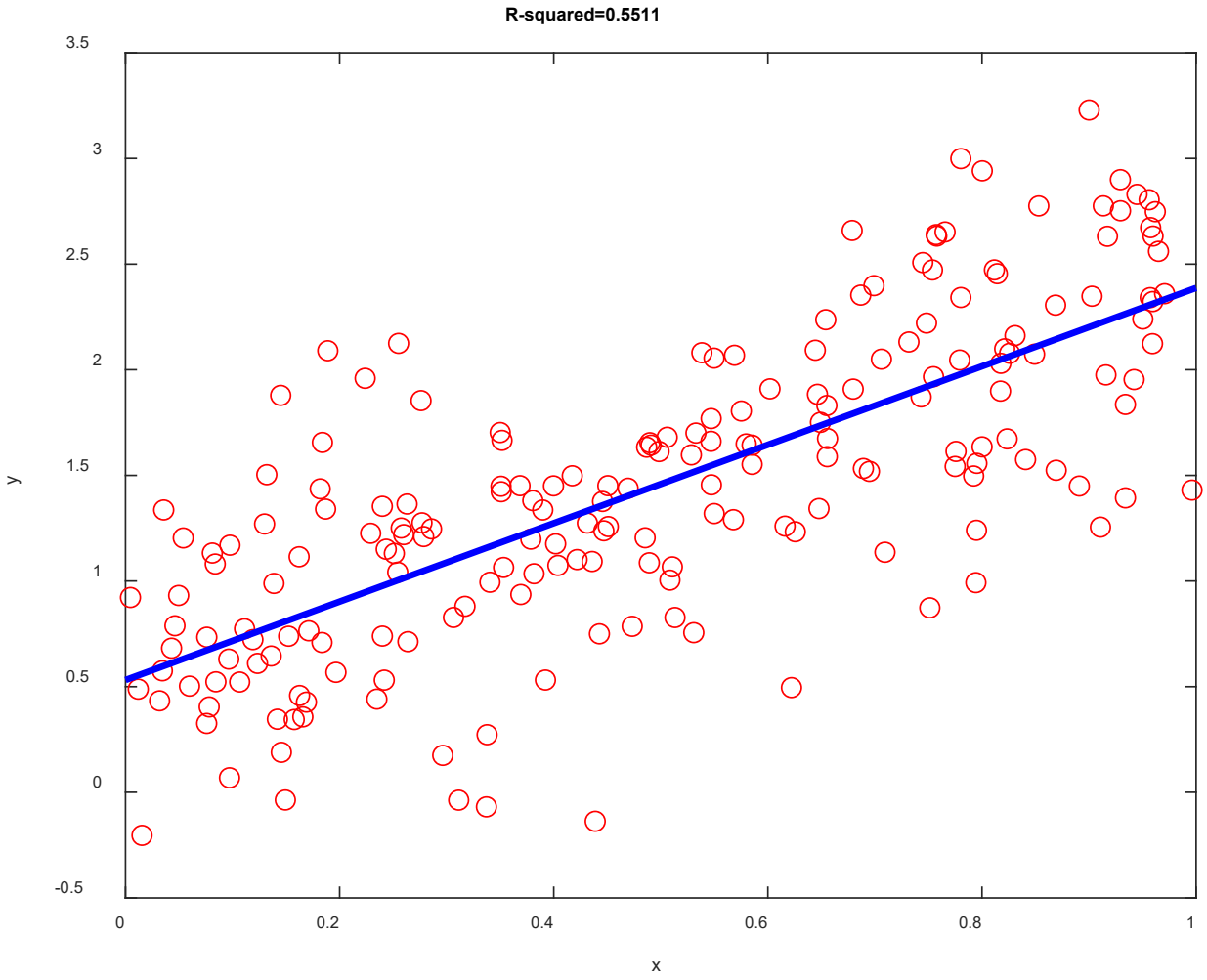
where $RSS$ and $TSS$ denote the sum of squares of residual and total of sum of squares, respectively,

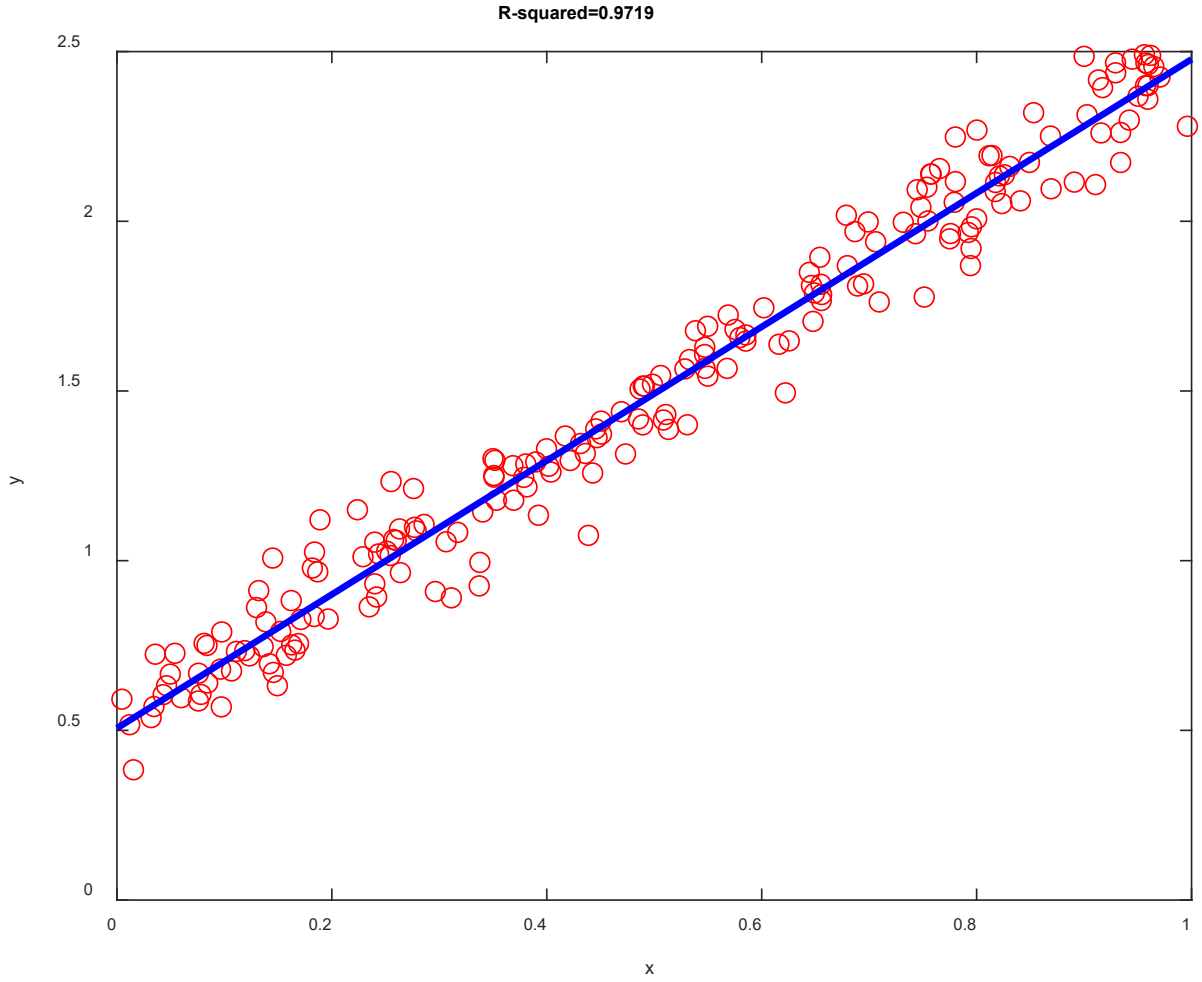
$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$TSS = \sum_{i=1}^{n} (\bar{y} - y_i)^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$


R-squared=0.2071
5
4
3
2
1
0
-1
-2
y
0
0.2
0.4
0.6
0.8
1
x


R-squared=0.5511
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
y
0
0.2
0.4
0.6
0.8
1
x


R-squared=0.9719
2.5
2
1.5
1
0.5
0
y
0
0.2
0.4
0.6
0.8
1
x

$R^2$ measures the fraction of variance explained by the model. It is between 0 and 1. 0 indicates that the model explains none of the variability of the target variable around its mean, while 1 indicates that model explains all the variability of the target variable around the mean.

## How high does $R^2$ need to be?

The $R^2$ value depends on the application area. Different questions have different amounts of variability that are inherently unexplainable. For example, humans are hard to predict. Any study that attempts to predict human behaviour will tend to have $R^2$ values less than 0.5. However, if you analyse a physical process and have very good measurements, you might expect $R^2$ values over 0.9. Consequently, the so-call "good" $R^2$ value depends on the amount of variability that is actually explainable.

The $R^2$ metric is not perfect. In fact, it suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model. That is, even if we are adding insignificant variables to the data, the value of $R^2$ does not decrease. It either remains the same or increases with the addition of new independent variables. This clearly does not make sense because some of the independent variables might not be useful in determining the target variable.

## (5) Adjusted R-squared (Adjusted $R^2$)

To address the flaw of the R-squared, the Adjusted R-squared takes into account the number of predictor variables used for predicting the target variable. The formula for calculating the adjusted $R^2$ is as follow:

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - m - 1)}$$

Where

$n$: number of samples

$m$: number of predictor variables used in the model

The advantage of Adjusted R-squared is that it penalizes the inclusion of unnecessary variables. This means that as you add more predictors to the model, the Adjusted $R^2$ value will only increase if the new variables significantly improve the model's performance.

In summary, a higher Adjusted $R^2$ value indicates that more of the variation in the dependent variable is explained by the model, while also considering the model's simplicity. It is a valuable tool for model selection, helping you strike a balance between explanatory power and complexity.

## Summary of the evaluation metrics

- Mean Squared Error (MSE) and Root Mean Square Error (RMSE) penalizes the large prediction errors in comparison with Mean Absolute Error (MAE). However, RMSE is more widely used than MSE to evaluate the performance of the regression model with other random models as it has the same units as the dependent variable.
- MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE. Therefore, in many models, MSE is used as a default loss function, despite being harder to interpret than MAE.

- The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of $R^2$ is considered desirable.
- $R^2$ and Adjusted $R^2$ are used for explaining how well the independent variables in the linear regression model explains the variability in the dependent variable. $R^2$ value always increases with the addition of the independent variables which might lead to the addition of the redundant variables in our model. However, the Adjusted $R^2$ solves this problem.
- For comparing the accuracy among different linear regression models, RMSE is a better choice than $R^2$.

## 8.6 Model Validation

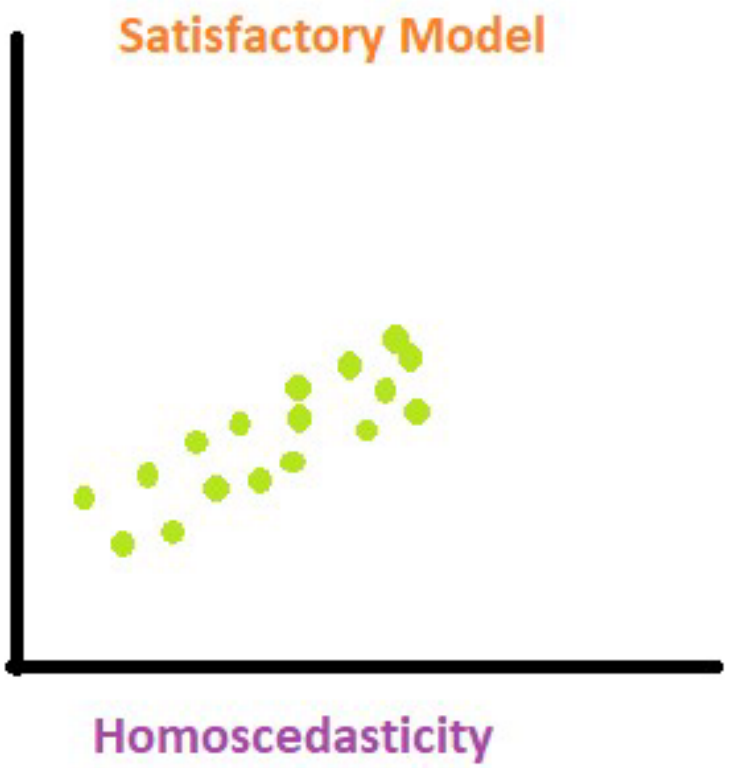
In order to validate a model, we need to check few assumption of linear regression model. The following are the common assumptions for linear regression models.

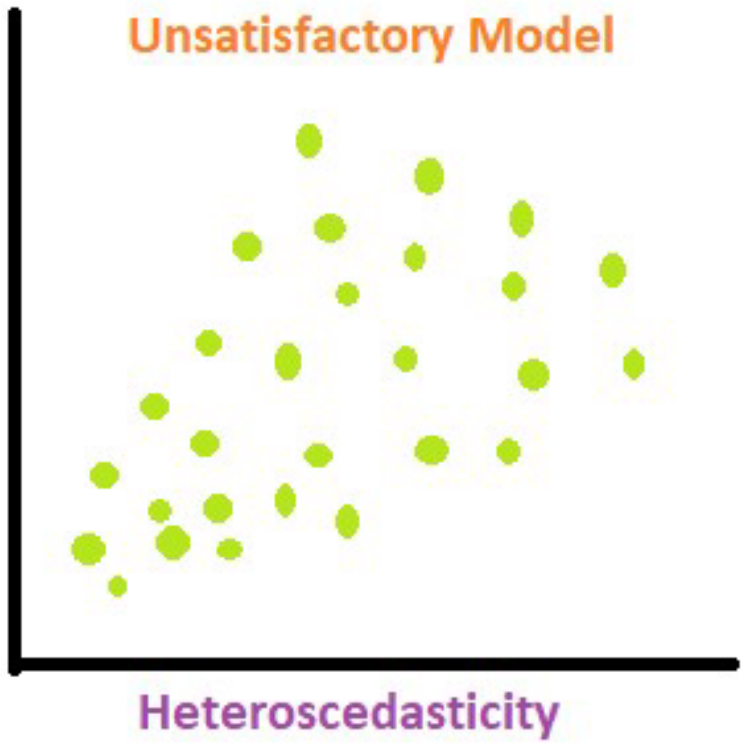
(1) Linear relationship. In linear regression, the relationship between the dependent and independent variable is assumed to be **linear**. This can be checked by scatter plot of actual value vs predicted value of the dependent variable.
(2) The residual (error) should be **normally distributed**.
(3) The **mean of residual (error) should be 0** or close to 0 as much as possible
(4) The linear regression require all independent variables to be **multivariate normal**. This assumption can be checked with Q-Q plot.

## (5) Homoscedasticity

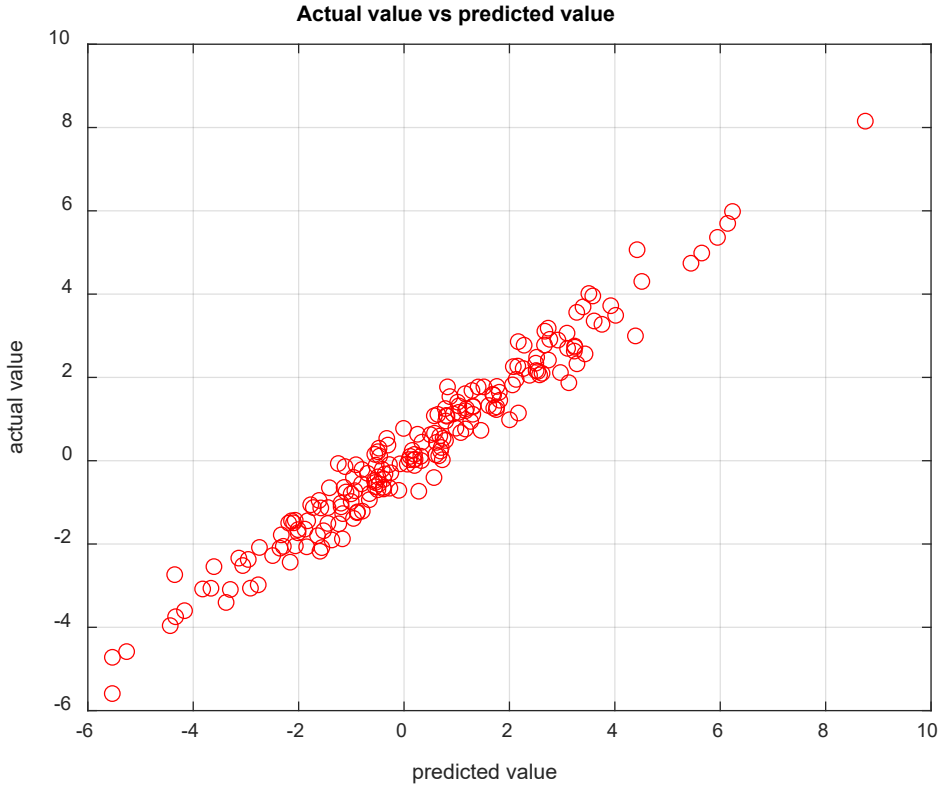
The assumption of homoscedasticity (meaning "same variance") is central to linear regression models. Homoscedasticity describes a situation in which the residual is the same across all values of the independent variables.

Heteroscedasticity (the violation of homoscedasticity) is present when the size of the residual differs across values of an independent variable. The easiest way to check homoscedasticity is to make a scatterplot with the residuals against the predicted dependent variable. A scatterplot in a busted homoscedasticity assumption would show a pattern to the data points. If you happen to see a funnel shape to your scatter plot, this indicates a busted assumption.

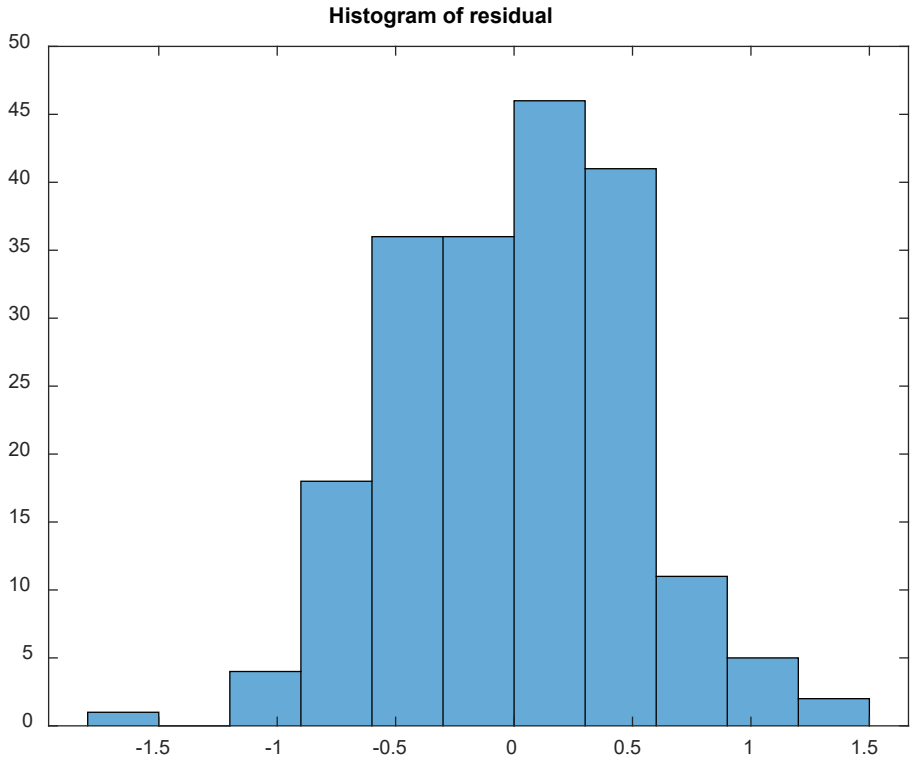

Satisfactory Model
Homoscedasticity


Unsatisfactory Model
Heteroscedasticity

# Let’s check the above assumptions for the above example



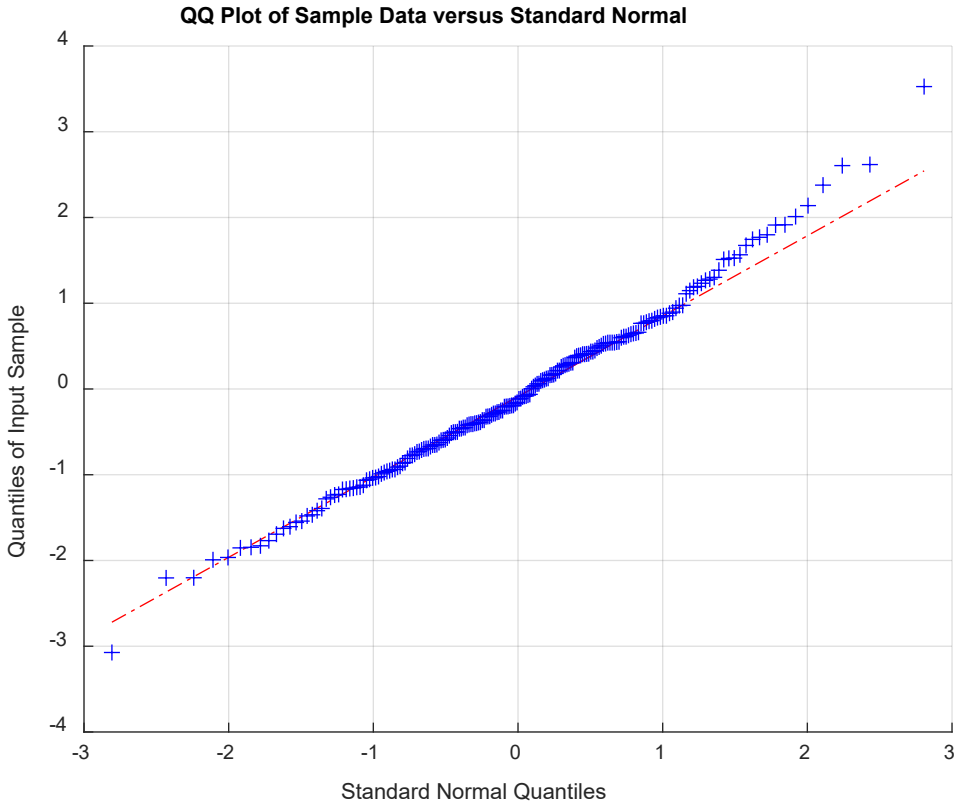

The above shows the assumption of linear relationship is true

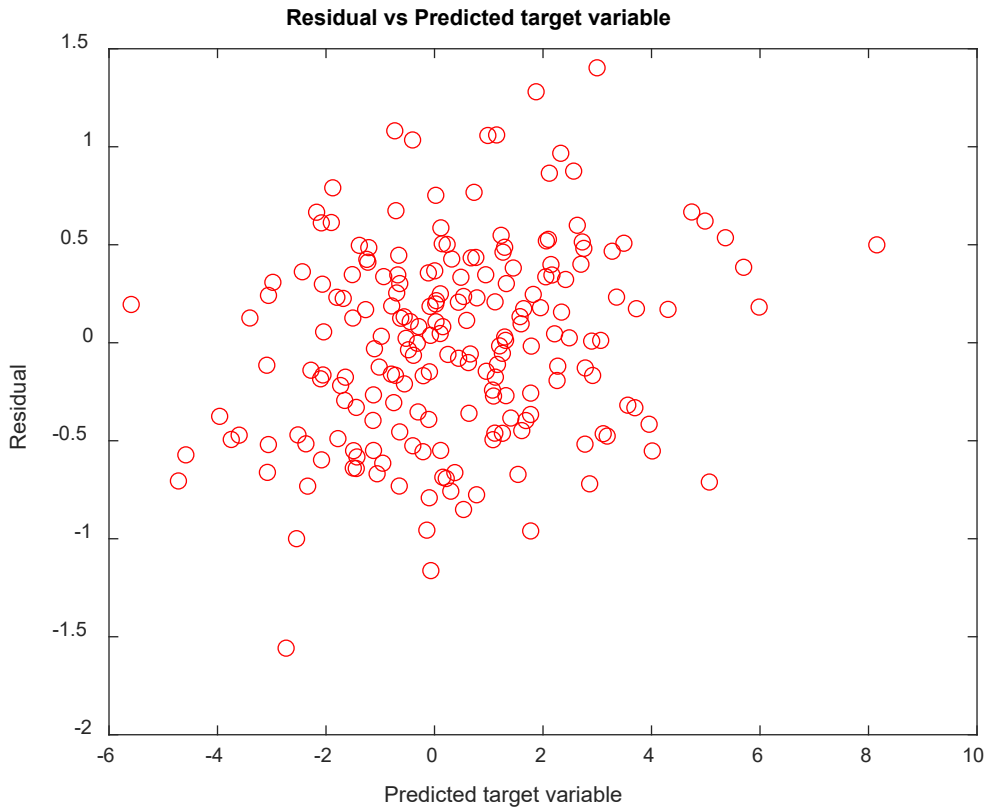



The residual roughly follows normal distribution

$$\text{mean}(\varepsilon) = 2.05 \times 10^{-17} \approx 0$$

Q-Q plot displays a quantile-quantile plot of the quantiles of the sample data $x$ versus the theoretical quantile values from a normal distribution

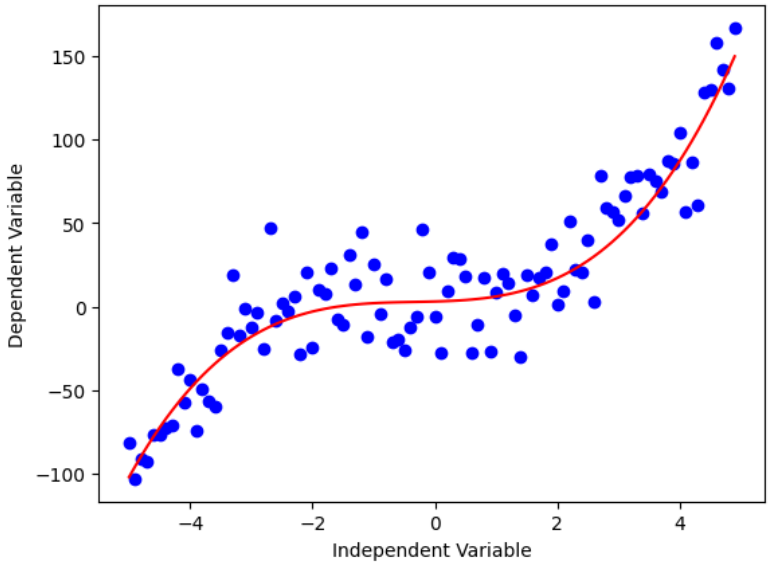


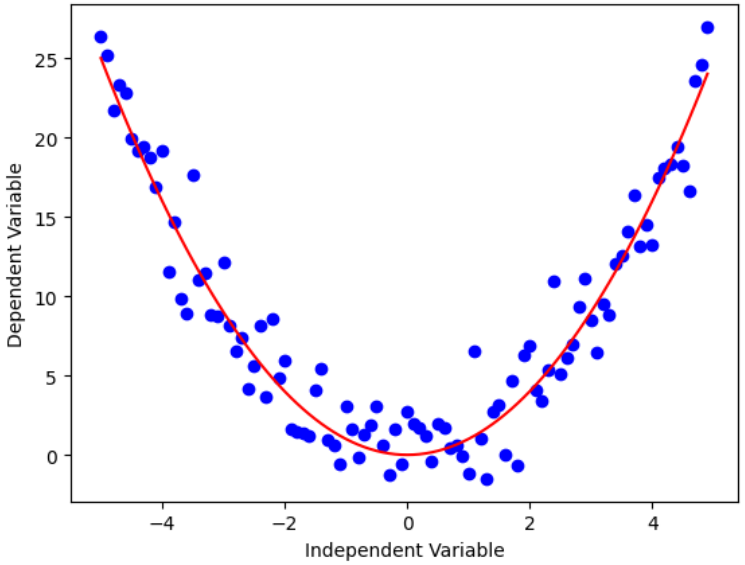



Obviously, the residual does not increase or decrease with the increased predicted target (dependent) variable. This verifies the assumption of homoscedasticity.

# 8.7 Nonlinear Regression

Although the linear regression model is often adequate, there are many cases, as shown below, in which a nonlinear function is more suitable.



Nonlinear regression is a method of finding a nonlinear model of the relationship between the dependent variable and a set of independent variables. Unlike traditional linear regression, which is restricted to estimating linear models, nonlinear regression can estimate models with arbitrary relationships between independent and dependent variables.

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon$$

Some examples of nonlinear models include:

(1) Polynomial model

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \varepsilon$$

(2) Rational model

$$y = \frac{\theta_0 + \theta_1 x + \theta_2 x^2}{\theta_3 + \theta_4 x + \theta_5 x^2} + \varepsilon$$

With increasing popularity of artificial neural networks, nonlinear regression receives less attention in recent years. Nonlinear regression will not be covered in this course. If you are interested, please refer to the vast materials on the Internet.