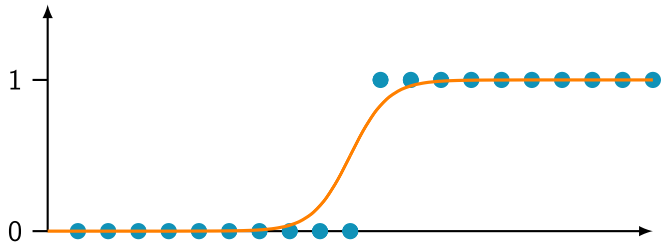
STA302H1 Methods of Data Analysis I

Haoran Yu

August 18, 2024



Contents

1 Basic concepts and introduction

1.1 Different data formats

In any data analysis course, understanding the different types of data is crucial because it informs how we can process, analyze, and model the data. This section introduces three main types of data: scalar data, vector data, and matrix data, each of which can be used in different contexts in data analysis and linear regression.

1.1.1 Scalar data

Definition 1.1: Scalar data

A single value or element that represents a specific measurement or attribute. It is the most basic form of data, consisting of just one number or categorical item without any additional structure. Scalars are typically used to represent individual quantities or characteristics.

Definition 1.2: Numerical scalar data

A scalar data with integer value $\mathbb{Z}$ or a continuous value $\mathbb{R}\backslash\mathbb{Z}$, which can take any value within a given range and is not restricted to whole numbers.

Example.

> The number of students in a classroom (e.g., 30 students) is a numerical scalar data with integer value.

Example.

> The amount spent on a cup of coffee (e.g., 2.99 CAD) is a numerical scalar data with continuous value.

Definition 1.3: Categorical scalar data

A data that can take on one of a limited, fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category based on some qualitative property.

Definition 1.4: Dummy variables

They are variables used to represent categorical data numerically. For a categorical variable with two categories, a dummy variable can take on values of 0 or 1.

Example.

> Suppose we have information that describes students' affiliation with the Department of Statistical Sciences (DoSS). We have two categories of students:
>
> - **Category 1:** Students from DoSS.
> - **Category 2:** Students not from DoSS.
>
> And the categorical data in this case is the students' affiliation with DoSS.
> For the categorical data above, we can create a dummy variable x_i to represent the two categories numerically:
>
> - Let $x_i = 1$ if student i is from DoSS.
> - Let $x_i = 0$ if student i is not from DoSS.
>
> Or alternatively:
>
> - Let $x_i = 1$ if student i is not from DoSS.
> - Let $x_i = 0$ if student i is from DoSS.

1.1.2 Vector data

Definition 1.5: Vector data

It is ordered collection of numerical data points, often used for representing series of data. Each element in the vector corresponds to a specific value or measurement.

Remark.

> When vector data has independent coordinates, each element of the vector is an individual data point that does not depend on the other elements. When vector data has dependent coordinates, each element in the vector is related to the others, often in a sequential manner. This is typical in time series data where each data point is dependent on previous data points.

Remark.

> Vectors in this course are represented as column vectors by default.

Example.

> Consider a vector representing the prices of coffees bought by students:
>
> $$\mathbf{x} = (2.99, 1.99, 3.45, \ldots, 3.99)^T \in \mathbb{R}^{20}$$
>
> Here, each element in the vector represents the price of a different coffee, and there are 20 such prices in total.

Example.

> Consider a vector representing the final grades of students in the STA302 course:
>
> $$\mathbf{x} = (99, 97, 98, 85, 100, \ldots)^T \in \mathbb{R}^{135}$$
>
> Each element in this vector represents the final grade of a different student, and there are 135 grades in total.

1.1.3 Matrix data

Definition 1.6: Matrix data

An $n \times p$ matrix $X \in \mathbb{R}^{n \times p}$ stores np scalars in n rows and p columns.

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix}$$

When $p = 1, X$ reduces to vector data

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

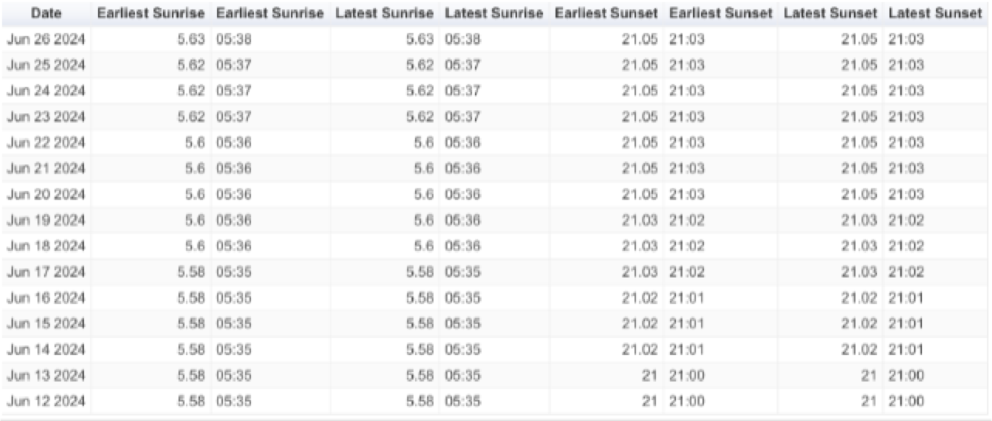
Definition 1.7: Matrix data transpose

The transpose of X, denoted X^T, is a $p \times n$ matrix where the rows and columns are switched.

$$X^T = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{n,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1,p} & X_{2,p} & \cdots & X_{n,p} \end{pmatrix}$$

Example.

Consider the following data set representing Toronto sunrise-sunset data:



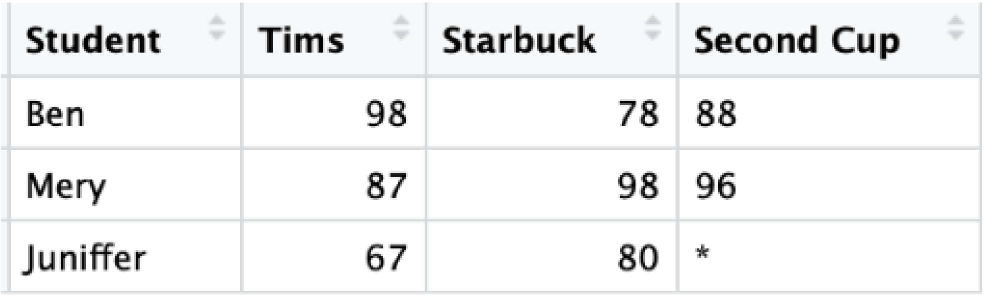
Date	Earliest Sunrise	Earliest Sunrise	Latest Sunrise	Latest Sunrise	Earliest Sunset	Earliest Sunset	Latest Sunset	Latest Sunset
Jun 26 2024	5.63	05:38	5.63	05:38	21.05	21:03	21.05	21:03
Jun 25 2024	5.62	05:37	5.62	05:37	21.05	21:03	21.05	21:03
Jun 24 2024	5.62	05:37	5.62	05:37	21.05	21:03	21.05	21:03
Jun 23 2024	5.62	05:37	5.62	05:37	21.05	21:03	21.05	21:03
Jun 22 2024	5.6	05:36	5.6	05:36	21.05	21:03	21.05	21:03
Jun 21 2024	5.6	05:36	5.6	05:36	21.05	21:03	21.05	21:03
Jun 20 2024	5.6	05:36	5.6	05:36	21.05	21:03	21.05	21:03
Jun 19 2024	5.6	05:36	5.6	05:36	21.03	21:02	21.03	21:02
Jun 18 2024	5.6	05:36	5.6	05:36	21.03	21:02	21.03	21:02
Jun 17 2024	5.58	05:35	5.58	05:35	21.03	21:02	21.03	21:02
Jun 16 2024	5.58	05:35	5.58	05:35	21.02	21:01	21.02	21:01
Jun 15 2024	5.58	05:35	5.58	05:35	21.02	21:01	21.02	21:01
Jun 14 2024	5.58	05:35	5.58	05:35	21.02	21:01	21.02	21:01
Jun 13 2024	5.58	05:35	5.58	05:35	21	21:00	21	21:00
Jun 12 2024	5.58	05:35	5.58	05:35	21	21:00	21	21:00

Figure 1: Toronto Sunrise-Sunset Dataset

This is a matrix data set representing the times of sunrise and sunset in Toronto over a period. Rows might represent different days. And columns might represent the times for sunrise and sunset.

Example.

Consider the following data on users' ratings or preferences for various coffee shops in Toronto:



Student	Tims	Starbuck	Second Cup
Ben	98	78	88
Mery	87	98	96
Juniffer	67	80	*

Figure 2: User rating data

Rows represent different users. And columns represent different items. Elements represent ratings given by users to items. Typically, such matrices have missing entries due to some users not rating all items.

Definition 1.8: Recommender system

Recommender systems are algorithms designed to suggest relevant items to users based on various factors.

The system uses this matrix to predict and recommend items that the user might like based on the available data.

Example.

Consider the following graph (e.g., friendship network):

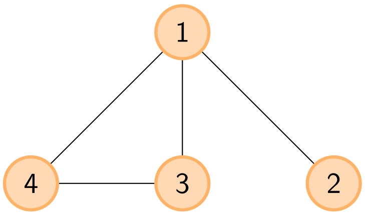


Figure 3: friendship network

In a matrix data, we take rows and columns represent nodes in the network. The network consists of four nodes and four edges:

1. Node 1 is connected to Nodes 2,3, and 4 .
2. Node 3 is connected to Node 4.

And take elements represent the presence or absence of edges between nodes. Specifically $A_{ij} = 1$ if there is an edge between node i and node j, and $A_{ij} = 0$ otherwise.
We first create a square matrix with dimensions equal to the number of nodes. Initially, fill it with zeros. For a graph with 4 nodes, the matrix will be 4×4:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- Edge between Node 1 and Node 2: Set $A_{12} = 1$, and set $A_{21} = 1$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- Edge between Node 1 and Node 3: Set $A_{13} = 1$, and set $A_{31} = 1$

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- Edge between Node 1 and Node 4: Set $A_{14} = 1$, and set $A_{41} = 1$

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- Edge between Node 3 and Node 4:Set $A_{34} = 1$, and set $A_{43} = 1$

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

After processing all the edges, the final adjacency matrix A for this graph is:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

According to linear algebra, we know a matrix is symmetric if $A_{ij} = A_{ji}$ for all i, j.

- In a directed graph, edges have a direction. This means that an edge from node i to node j does not imply an edge from node j to node i. The adjacency matrix for a directed network is not necessarily symmetric.
- In a weighted graph, edges have weights associated with them, representing the strength or capacity of the connection. The adjacency matrix for a weighted graph contains the weights of the edges instead of just 0 s and $1 \mathrm{~s}$. For example, if the weight of the edge between nodes i and j is w_{ij}, then $A_{ij} = w_{ij}$.

Thus, the adjacency matrix is symmetric if the graph it represents is undirected. So, for our A, it is binary and symmetry.

From the network matrix example, we have the following facts

Fact 1.9

A network is considered sparse if it has relatively few edges compared to the number of possible edges. In other words, a sparse network has a low density of connections. Thus, Simpler the network, simpler the matrix data.

Fact 1.10

The adjacency matrix is symmetric if the network it represents is undirected.

Example.

Matrix data can also be used to represent a gray scale image. A gray scale image is represented by a matrix where each element corresponds to a pixel in the image. The value of each element represents the intensity of the pixel, usually ranging from 0 (black) to 255 (white).

$$\begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1512} \\ g_{21} & g_{22} & \cdots & g_{2512} \\ \vdots & \vdots & \ddots & \vdots \\ g_{3841} & g_{3842} & \cdots & g_{384512} \end{pmatrix}$$

Here, g_{ij} represents the intensity of the pixel at row i and column j.

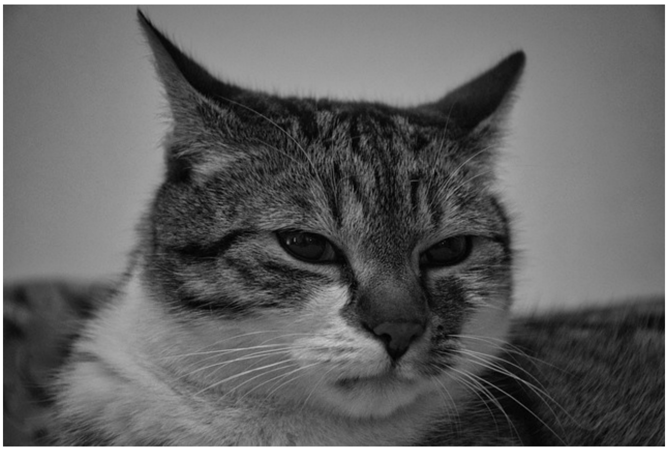


Figure 4: Example gray scale image

Example.

An RGB image can be represented as a $384 \times 512 \times 3$ array. This means the image has a height of 384 pixels, a width of 512 pixels, and 3 color channels / layers (Red, Green, Blue). Each pixel in the image is described by three values, corresponding to the intensities of the red, green, and blue channels. Each pixel in the image is accessed by three indices: one for the row, one for the column, and one for the color channel. For example, the pixel at position (i, j) has three values:

$$\text{Pixel}(i, j) = (\text{Red}[i, j], \text{Green}[i, j], \text{Blue}[i, j])$$

Alternatively, the RGB image can be thought of as three separate 384×512 matrices:

- Red Channel Matrix: A matrix where each element represents the red intensity of the corresponding pixel.
- Green Channel Matrix: A matrix where each element represents the green intensity of the corresponding pixel.
- Blue Channel Matrix: A matrix where each element represents the blue intensity of the corresponding pixel.

Each of these matrices has the same dimensions (384 rows and 512 columns), but they store different information corresponding to the intensity of their respective color channels. By combining these three matrices, we can reconstruct the original RGB image.

Example.

Consider the following table as a specific example:

Coffee shop	min price	max price	daily customers	working hours	working days
A	1.99	7.99	300	10	5
B	2.99	9.99	345	8	6
C	2.45	6.54	423	9	7
D	3.99	8.99	199	7	6
E	4.45	10.00	250	12	5

Table 1: Coffee shop data

Definition 1.11: Covariate

They are independent variables used to predict or explain the dependent variable in a model. They can be classified as univariate or multivariate based on the number of variables being considered.

- Univariate refers to a single variable or covariate. When we analyze or model data with one predictor variable, we are dealing with a univariate covariate.
- Multivariate refers to multiple variables or covariates. When we analyze or model data with more than one predictor variable, we are dealing with multivariate covariates.

- min price: The minimum price of coffee at the shop. This covariate indicates the lowest price at which coffee is sold at each shop.
- max price: The maximum price of coffee at the shop. This covariate indicates the highest price at which coffee is sold at each shop.
- daily customers: The average number of customers visiting the shop daily. This covariate provides an estimate of the shop's daily customer traffic.

> - working hours: The number of working hours per day. This covariate indicates how many hours each coffee shop is open each day.
>
> - working days: The number of working days per week. This covariate indicates how many days each coffee shop is open each week.
>
> When working with multiple covariates, the data can be organized into a matrix where each row represents an observation, and each column represents a different covariate:
>
> $$\begin{pmatrix} 1.99 & 7.99 & 300 & 10 & 5 \\ 2.99 & 9.99 & 345 & 8 & 6 \\ 2.45 & 6.54 & 423 & 9 & 7 \\ 3.99 & 8.99 & 199 & 7 & 6 \\ 4.45 & 10.00 & 250 & 12 & 5 \end{pmatrix}$$
>
> Regression analysis is particularly suitable for this kind of data because it allows us to understand the relationships between the dependent variable (response variable) and one or more independent variables (covariates or predictor variables). In this example, we can use regression to study how various features of coffee shops (like prices, working hours, etc.) influence the number of daily customers.

1.2 Introduction to methods to analyze the data

The purpose of analyze data is to understand the relationship between a dependent variable (also called the response variable) and one or more independent variables (also called predictor variables or covariates). There are different methods to achieve this purpose, such as:

- Regressions
- Classifications
- Reinforcement learning

The primary focus of this course is on regression analysis. Regression is a powerful statistical method that allows us to examine the relationship between two or more variables of interest.

1.2.1 Functional and statistical relationships

Definition 1.12: Function relationship

A precise, mathematical relationship between two variables where one variable is a deterministic function of the other. This means that for every value of the independent variable, there is exactly one corresponding value of the dependent variable.

Example.

> Suppose the price of salad is $\$2.99$ per 100 grams. If x is the weight of the salad a customer buys (in grams), then the price y they pay can be calculated using the functional relationship:
>
> $$y = \frac{2.99}{100} \times x$$
>
> In this case, every pair (x_i, y_i) lies exactly on the line described by this function.

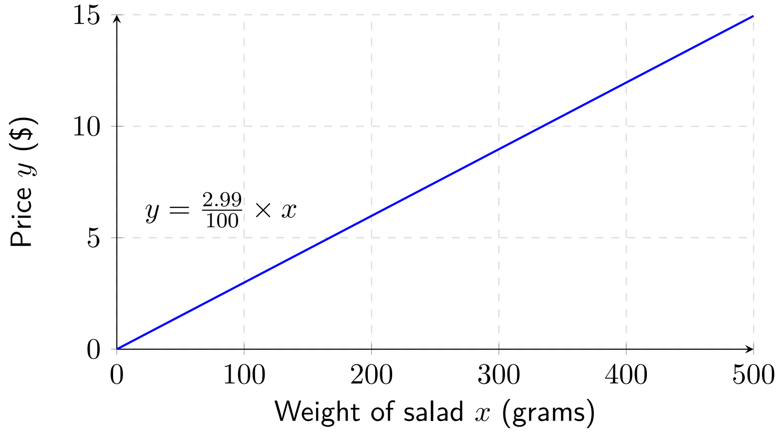


Figure 5: Caption

Definition 1.13: Statistical relationship

A statistical relationship, on the other hand, describes a relationship between two variables where the dependent variable is influenced by the independent variable but with some degree of randomness or error. This means that the data shows a trend, but the relationship is not exact.

Example.

> Suppose the relationship between the number of hours studied and the exam scores of students. The relationship shows a general trend but includes variability that cannot be captured by a simple deterministic function. This probabilistic nature is characteristic of statistical relationships, distinguishing them from functional relationships where the dependent variable is an exact function of the independent variable.

1.2.2 Basic definition of regressions

Definition 1.14: Regression

It is a statistical method used to examine the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables or covariates). The goal of regression is to model the dependent variable as a function of the independent variables and to understand the nature of their relationship. Mathematically, it means regression tries to estimate a function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$y_i \approx f(\boldsymbol{x}_i)$$

for all $i = 1, \ldots, n$ as close as possible.

- $\boldsymbol{x_i}$ is called the explanatory variable and y_i is called the response variable
- $\boldsymbol{x_i}$ is also called predictor, and y_i is also called the dependent variable

Remark.

> - When $n = 1, x_i$ is just a scalar, Simple regression.
> - When $n > 1, \boldsymbol{x}_i$ is a vector, Multiple regression.

Definition 1.15: Parametric regression

A type of regression where we specify a functional form for the relationship between the dependent and independent variables, and then estimate the parameters of that function.

Remark.

> Parametric regression simplifies the modeling process by assuming a specific functional form for the relationship between variables. This reduces the complexity of the model and makes it easier to estimate and understand.

Example.

> Linear regression assumes a linear relationship between the dependent and independent variable:
>
> $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
>
> Polynomial regression extends linear regression by including polynomial terms, allowing for a curved relationship:
>
> $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \epsilon_i$$

2 Simple linear regression and polynomial regression

2.1 Simple linear regression

Definition 2.1: Simple Linear Regression

Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we propose simple linear regression model

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

where $\mathbb{E}(\epsilon_i) = 0$, for $i = 1, 2, \ldots, n$

- y_i is the dependent variable.
- x_i is the independent variable.
- β_0 is the intercept, which is the expected value of y when x is zero.
- β_1 is the slope, which is the expected change in y for a one-unit change in x.
- ϵ_i is the error term.

Remark.

> Formally, given $X = x_i$, we have
>
> $$\begin{aligned} EY_i =& E(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= E(\beta_0 + \beta_1 x_i) + E\epsilon_i \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

Consider observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. These are the actual data points collected from a sample. Each pair (x_i, y_i) represents an observation where x_i is the value of the independent variable and y_i is the value of the dependent variable. These observations are considered realizations of the random variables X and Y.

At the population level, we consider the theoretical distribution of the variables. The random variables X and Y follow a joint distribution, which describes how X and Y are related in the population. The "true" underlying relationship between X and Y is modeled as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 is the intercept.
- β_1 is the slope.
- ϵ is the error term, representing the variability in Y that cannot be explained by X. The randomness of Y_i comes from ϵ_i.

Remark.

> For simplicity, we treat X as a given (fixed) value and Y as the random variable. This simplification helps in focusing on the variability and distribution of Y. Under this treatment, we simplify the conditional expectation $E(Y \mid X)$ to $E(Y)$, meaning the expected value of Y. Similarly, we simplify the conditional variance $\operatorname{var}(Y \mid X)$ to var Y, meaning the $\operatorname{Var}(Y)$.

2.1.1 Least square estimation

We want to find the "best" line that fits all the data by estimating the parameters β_0 and β_1. We approach this objective by quantify the closeness of the data to the model using this equation $y_i - \beta_0 - \beta_1 x_i$, which could be positive, negative, or zero. As a result, the best β_0 and β_1 are found by solving the following optimization problem:

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

To formalize this optimization question, we define the least squares function:

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

Compute the derivatives and set them to zero:

$$\begin{aligned} \frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}$$

We reorganize the term as:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Since $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we have:

$$\begin{aligned} \beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\ \beta_0 &= \overline{y} - \beta_1 \overline{x} \end{aligned}$$

Definition 2.2: The least square estimator

From our optimization, we have the least square estimator as:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}. \end{aligned}$$

With the estimating regression line is

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

Remark.

> The estimator $\widehat{\beta}_0$ is the estimated value of the intercept of the regression line. It represents the expected value of the response variable y when the predictor variable x is zero. The estimator $\widehat{\beta}_1$ is the estimated value of the slope of the regression line. It represents the average change in the response variable y for a one-unit increase in the predictor variable x.

Remark.

> These least sqaure estimators are also called ordinary least squares estimator becuase we are using ordinary least squares method, which is to minimize the sum of the squared differences (residuals) between the observed values and the values predicted by the model.

Example.

> Suppose we have data:
>
> $$x = (1, 2, 3, 4, 5)^\top, y = (0.5, 2.5, 2.5, 4.5, 5)^\top$$

> And sample averages:
>
> $$\overline{x}=3, \overline{y}=3$$
>
> $$\sum_{i=1}^{5}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)=11$$
>
> $$\sum_{i=1}^{5}\left(x_i-\overline{x}\right)^2=10$$
>
> $\widehat{\beta}_1=\frac{11}{10}=1.1, \quad \widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x}=3-1.1\times 3=-0.3$

Definition 2.3: Fitted value, residue, residue sum of squares

We have the fitted value from the regression line as:

$$\widehat{y}_i=\widehat{\beta}_0+\widehat{\beta}_1 x_i, \text{ for } i=1,2,\ldots,n$$

The residue as:

$$\widehat{e}_i=y_i-\widehat{y}_i, \text{ for for } i=1,2,\ldots,n$$

The residue sum of squares RSS:

$$RSS=\sum_{i=1}^{n}\widehat{e}_i^2=\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2$$

Remark.

> We square residuals to prevent cancellation of positive and negative residuals. And ensures the RSS function is differentiable.

Proposition 2.4

With some new definition above, we have the following propositions:

1. $\sum_{i=1}^{n}\hat{e}_i=0$
2. $\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i=0$
3. $\sum_{i=1}^{n}x_i\hat{e}_i=0$

Proof. We will prove the identities in the proposition above one by one:

1. The residuals are given by:

   $$\hat{e}_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i$$

   Summing the residuals:

   $$\sum_{i=1}^{n}\hat{e}_i=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)$$

   Substitute $\hat{\beta}_0$:

   $$\sum_{i=1}^{n}\hat{e}_i=\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\left(\bar{y}-\hat{\beta}_1\bar{x}\right)-\hat{\beta}_1\sum_{i=1}^{n}x_i$$

   $$\sum_{i=1}^{n}\hat{e}_i=\sum_{i=1}^{n}y_i-n\bar{y}+\hat{\beta}_1\sum_{i=1}^{n}x_i-\hat{\beta}_1\sum_{i=1}^{n}x_i$$

   Note that:

   $$n\bar{y}=\sum_{i=1}^{n}y_i \quad \sum_{i=1}^{n}\hat{e}_i=\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}y_i=0$$

2. Since the sum of the residuals $\sum_{i=1}^{n} \hat{e}_i$ is zero, the mean of the residuals is also zero.

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = \frac{1}{n} \cdot 0 = 0$$

3. The sum of the product of the predictors x_i and the residuals $\hat{e}_i$ is also zero. This property arises from the fact that the residuals are orthogonal to the predictors in the least squares regression.

$$\sum_{i=1}^{n} x_i \hat{e}_i = \sum_{i=1}^{n} x_i \left(y_i - \hat{y}_i\right) = 0$$

■

2.2 Perform simple linear regression in R

See corresponding R file.

3 Multiple linear regression

3.1 Polynomial regression and its matrix representation

Consider the following data set that documents the working experience in years (x), and annual salary in thousand dollars (y):

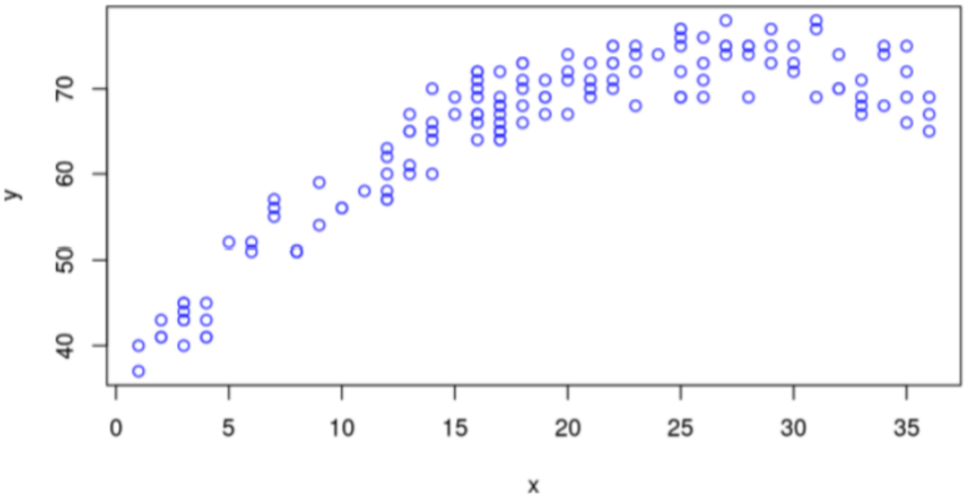


Figure 6: Professor data set

We can construct a simple linear regression $y = \beta_0 + \beta_1 x + \epsilon$:

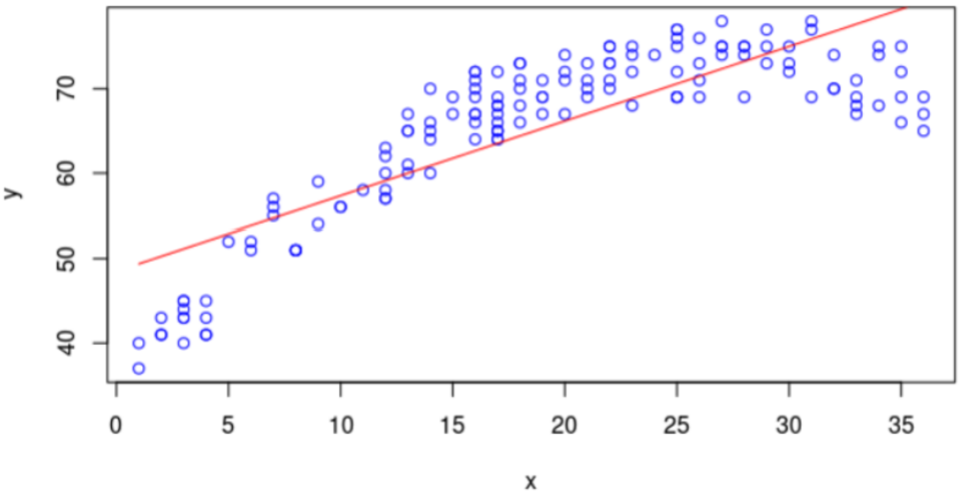


Figure 7: Simple linear regression

Now we fit a quadratic regression line $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$:

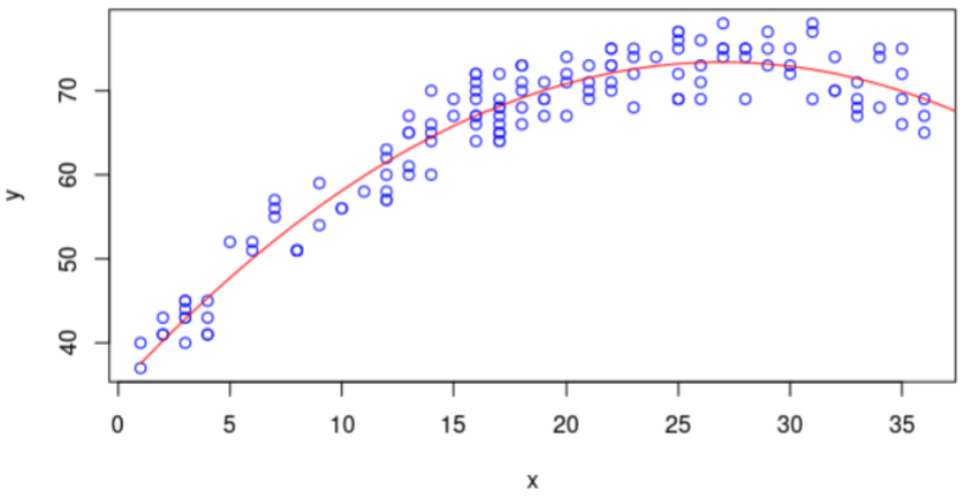


Figure 8: Quadratic regression line

Now we fit a cubic regression line $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$:

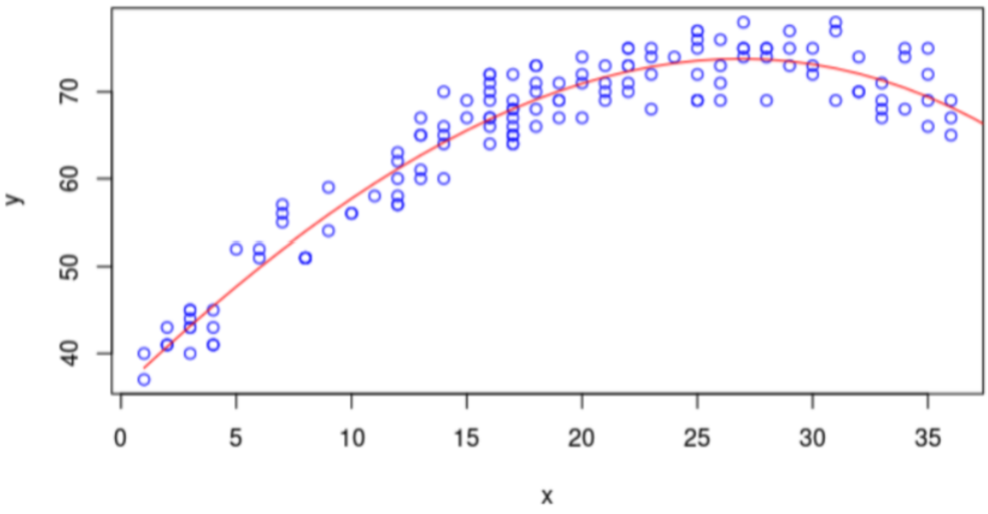


Figure 9: Cubic regression line

This illustration is a motivational example for polynomial regression:

- A quadratic model $(y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon)$ includes the linear model $(y = \beta_0 + \ \beta_1 x + \epsilon)$ as a special case. This means that the quadratic model can represent any relationship that a linear model can, plus additional flexibility to capture curvature in the data.

- Similarly, a cubic model $(y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon)$ includes both the linear and quadratic models as special cases. This provides even more flexibility to fit the data by accounting for more complex relationships.

- RSS measures the discrepancy between the data and the estimation model. A smaller RSS indicates a model that fits the data better.

- More general models (those with more terms, like cubic compared to quadratic) tend to have a smaller RSS because they can more closely fit the data by accounting for additional patterns or trends.

However, there is a trade off between complexity and regression accuracy. While more complex models can fit the data better (i.e., have a smaller RSS), they are not always better, especially when it comes to prediction. In regression analysis, the goal is often to create a model that predicts new data accurately, not just to fit the existing data well. Thus, a balance must be found between model complexity and prediction accuracy.

The size of the coefficients for higher-order terms can indicate whether the additional complexity is necessary.

Example.

> In the cubic model example, if the coefficient for x^3 is relatively small, the contribution of the x^3 term to the model is minimal. This suggests that a simpler quadratic model might be sufficient for capturing the data's relationship.

Remark.

> Polynomial regression is a simple example for multiple regression, because it has more than one covariate i.e. $x^2, x^3, \cdots$

3.1.1 Matrix for simple linear regression

Given the data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ and } \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \beta_0 \cdot 1 + \beta_1 x_i + \epsilon_i, \text{ for } i = 1, 2, \ldots, n$$

We can rewrite it in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1, & x_1 \\ 1, & x_2 \\ \vdots & \vdots \\ 1, & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

For shorthand notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

3.1.2 Matrix for quadratic regression

Given the data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$:
Denote:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ and } \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

The design matrix is:

$$\boldsymbol{X} = \begin{bmatrix} 1, & x_1, & x_1^2 \\ 1, & x_2, & x_2^2 \\ \vdots, & \vdots, & \vdots \\ 1, & x_n, & x_n^2 \end{bmatrix} \in \mathbb{R}^{n \times 3}$$

The quadratic regression model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i = \beta_0 \cdot 1 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \text{ for } i = 1, 2, \ldots, n$$

We can rewrite it in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1, & x_1, & x_1^2 \\ 1, & x_2, & x_2^2 \\ \vdots, & \vdots, & \vdots \\ 1, & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

For shorthand notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

3.1.3 Matrix for cubic regression

Given the data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$:
Denote:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ and } \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

The design matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1, & x_1, & x_1^2, & x_1^3 \\ 1, & x_2, & x_2^2, & x_2^3 \\ \vdots, & \vdots, & \vdots & \vdots \\ 1, & x_n, & x_n^2 & x_n^3 \end{bmatrix} \in \mathbb{R}^{n \times 4}$$

The cubic regression model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i = \beta_0 \cdot 1 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \text{ for } i = 1, 2, \ldots, n$$

We can rewrite it in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1, & x_1, & x_1^2 & x_1^3 \\ 1, & x_2, & x_2^2 & x_2^3 \\ \vdots, & \vdots, & \vdots & \vdots \\ 1, & x_n, & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

For shorthand notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

3.1.4 Matrix for polynomial regression of degree p

Polynomial regression model of degree p is, for $i = 1, 2, \ldots, n$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_p x_i^p + \epsilon_i,$$

We can rewrite it in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1, & x_1, & x_1^2 & x_1^3, & \ldots, & x_1^p \\ 1, & x_2, & x_2^2 & x_2^3 & \ldots, & x_2^p \\ \vdots, & \vdots & \vdots & \vdots, & \vdots, & \vdots \\ 1, & x_n, & x_n^2 & x_n^3 & \ldots, & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}$

3.2 Perform polynomial regreesion in R

See corresponding R file.

3.3 Multiple linear regression

Definition 3.1: Multiple linear regression

We define the following data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \boldsymbol{x}_i \in \mathbb{R}^p, p > 1$
The multiple linear regression model is given by:

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

- y_i : Response variable.
- β_0 : Intercept.
- $\beta_1, \beta_2, \ldots, \beta_p$: Regression coefficients.
- $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$: Predictor variables for the i-th observation.
- ϵ_i : Error term.

Example.

Consider the following table again as a specific example:

Coffee shop	min price	max price	daily customers	working hours	working days
A	1.99	7.99	300	10	5
B	2.99	9.99	345	8	6
C	2.45	6.54	423	9	7
D	3.99	8.99	199	7	6
E	4.45	10.00	250	12	5

Table 2: Coffee shop data

The model can be written as: daily customers for each coffee shop i $= \beta_0 + \beta_1($ min price of i $) + \beta_2($ max price of i $) + \beta_3($ working hours of i $)+$ $\beta_4($ working days of i $) + \epsilon$

In multiple linear regression, we have multiple predictors (independent variables). Each predictor adds a dimension to the space in which we are working. For instance:

- With one predictor (simple linear regression), the data points lie on a line in two dimensional space.
- With two predictors, the data points lie on a plane in three-dimensional space.
- With p predictors, the data points lie on a hyperplane in $(p+1)$-dimensional space.

Each observation in the dataset can be represented as a point in $(p+1)$-dimensional space, where p is the number of predictors, and the additional dimension is the response variable y. The regression model seeks to find the best-fitting hyperplane that represents the relationship between the predictors and the response variable. The equation of this hyperplane is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The goal of the regression analysis is to minimize the differences (residuals) between the observed values and the values predicted by the hyperplane. The best-fitting hyperplane is the one that minimizes the sum of the squared residuals (RSS).

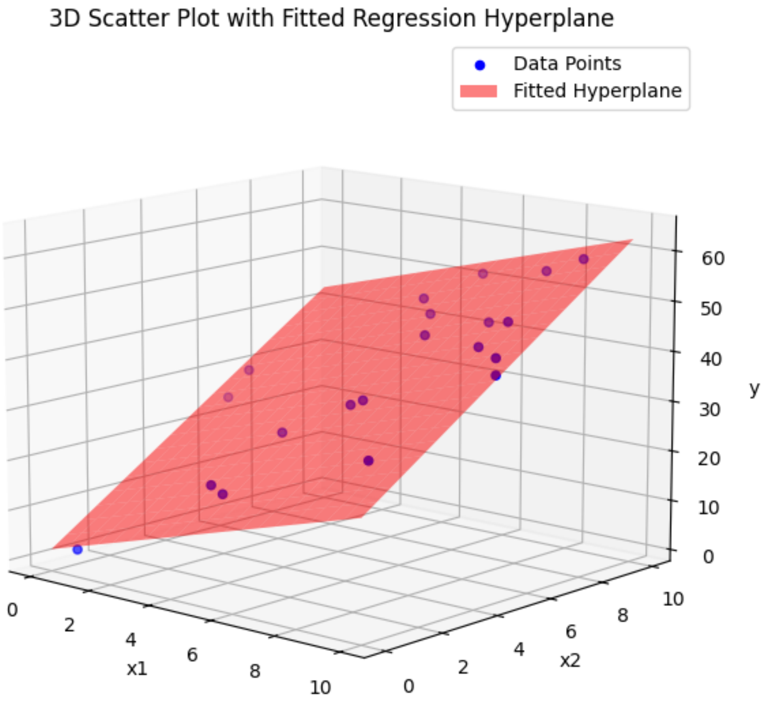


Figure 10: 3D example of the best fitted hyperplane

Remark.

> In population level, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \ldots, + \beta_p X_p$ determines a hyperplane in $\mathbb{R}^{p+1}$, meaning that the relationship is "true".

Just as the polynomial regression, multiple linear regression also has its own matrix representation. Collectively,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1, & \boldsymbol{x}_1^1 \\ 1, & \boldsymbol{x}_2^1 \\ \vdots & \vdots \\ 1, & \boldsymbol{x}_n^1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

And:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

3.3.1 Least square estimator for multiple linear regression

The objective is to solve for the coefficients β in the multiple linear regression model by minimizing the sum of squared residuals. The goal is to find the values of β that minimize:

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}))^2$$

More concisely:

$$\min_{\beta \in \mathbb{R}^{p+1}} \|y - X\beta\|^2$$

Consider the following linear function:

$$\boldsymbol{\beta} = (\beta_1 \beta_2 \cdots \beta_p)^\top$$

in the matrix form

$$F(\boldsymbol{\beta}) = \boldsymbol{a}^\top \boldsymbol{\beta} = \boldsymbol{\beta}^\top \boldsymbol{a} = \sum_{j=1}^{p} a_j \beta_j, \quad \boldsymbol{a} = \begin{pmatrix} a_1 & a_2 & \cdots & a_p \end{pmatrix}^\top$$

It's obvious that

$$\frac{\partial F(\beta)}{\partial \beta_j} = a_j, \quad j = 1, \ldots, p$$

So the derivative of $F(\beta)$ w.r.t. β is

$$\frac{\partial F(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial F(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial F(\boldsymbol{\beta})}{\partial \beta_2} \cdots \frac{\partial F(\boldsymbol{\beta})}{\partial \beta_p} \right)^\top = \boldsymbol{a}$$

For a quadratic function $Q(\beta) = \beta^T \boldsymbol{Q} \beta$, where $\boldsymbol{Q}$ is a square matrix, we have:

$$Q(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{Q} \boldsymbol{\beta} = \sum_{s,t} \beta_{\boldsymbol{s}} \beta_t q_{s,t}, \quad \boldsymbol{Q} = (q_{s,t})_{s=1,\cdots,p;t=1,\cdots,p}$$

The term including β_i is:

$$q_{ii}\beta_i^2 + \sum_{j \neq i} (q_{ij} + q_{ji}) \beta_j \beta_i.$$

The derivative of $Q(\beta)$ with respect to β_i is:

$$\frac{\partial Q(\beta)}{\partial \beta_i} = 2q_{ii}\beta_i + \sum_{j \neq i} (q_{ij} + q_{ji}) \beta_j = \sum_{j=1}^{p} (q_{ij} + q_{ji}) \beta_j.$$

The vector of derivatives is:

$$\frac{\partial Q(\beta)}{\partial \beta} = \left(\frac{\partial Q(\beta)}{\partial \beta_1}, \frac{\partial Q(\beta)}{\partial \beta_2}, \ldots, \frac{\partial Q(\beta)}{\partial \beta_p} \right)^T = \left(\boldsymbol{Q} + \boldsymbol{Q}^T \right) \beta.$$

Remark.

> If $\boldsymbol{Q}$ is symmetric, $\frac{\partial Q(\beta)}{\partial \beta} = 2\boldsymbol{Q}\beta$.

Our goal is to minimize the quadratic difference:

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^\top \boldsymbol{y} \\ &= \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{y}. \end{aligned}$$

- The quadratic term can be written as a quadratic form: $\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}$, where $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{X}$, which is a symmetric matrix.

  The gradient of $\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ is $2\boldsymbol{A}\boldsymbol{\beta}$.

  Thus, the gradient of $\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ is:

  $$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \right) = 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$$

- The linear term can be written as $\boldsymbol{\beta}^T \boldsymbol{b}$, where $\boldsymbol{b} = -2\boldsymbol{X}^T \boldsymbol{y}$.
  The gradient of $\boldsymbol{\beta}^T \boldsymbol{b}$ with respect to $\boldsymbol{\beta}$ is $\boldsymbol{b}$.
  Thus, the gradient of $-2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y}$ with respect to $\boldsymbol{\beta}$ is:
$$\frac{\partial}{\partial \boldsymbol{\beta}}\left(-2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y}\right) = -2\boldsymbol{X}^T \boldsymbol{y}$$
- The term $\boldsymbol{y}^T \boldsymbol{y}$ is constant with respect to $\boldsymbol{\beta}$.
  The gradient of a constant term with respect to $\boldsymbol{\beta}$ is zero:
$$\frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{y}^T \boldsymbol{y}\right) = 0$$

Combining the gradients of the individual terms, we get:

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

This can be factored as:

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

Setting the derivative equal to zero to find the minimum:

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

Solving for β :

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

This leads to the least square estimator:

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

Remark.

> The estimator is also OLS because we find the coefficient vector β that minimizes the sum of the squared residuals.

Example.

> Suppose $x = (1,2,3,4,5)^T$ and $y = (0.5, 2.5, 2.5, 4.5, 5)^T$. Compute the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
> Design matrix X :
> $$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$$
> Transpose and product:
> $$X^T X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}, \quad X^T y = \begin{bmatrix} 15 \\ 56 \end{bmatrix}$$
> Inverse of $X^T X$:
> $$\left(X^T X\right)^{-1} = \frac{1}{5 \cdot 55 - 15 \cdot 15} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix}$$
> Coefficients:
> $$\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix} \begin{bmatrix} 15 \\ 56 \end{bmatrix} = \begin{bmatrix} -0.3 \\ 1.1 \end{bmatrix}$$

Definition 3.2: Fitted values, Hat matrix and residues

We have the OLS estimator $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$:

- The fitted value is $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$
- And hat matrix is $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top, \widehat{y} = \boldsymbol{H}\boldsymbol{y}$. $\boldsymbol{H}^\top = \boldsymbol{H}$ is symmetric and $\boldsymbol{H}^2 = \boldsymbol{H}$ is idempotent.
- The residual is $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$

Proposition 3.3

$\boldsymbol{X}^\top \widehat{\boldsymbol{e}} = 0$

Proof. The OLS estimate also satisfies the following equation

$$\boldsymbol{X}^\top(-\widehat{\boldsymbol{e}}) = \boldsymbol{X}^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) = \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{0},$$

Therefore,

$$\boldsymbol{X}^\top \widehat{\boldsymbol{e}} = \boldsymbol{0}$$

■

Remark.

> $\widehat{\beta}_0$ is the estimated averaged response when the predictor is 0. $\widehat{\beta}_j$ is the estimated averaged change in response for a one-unit increase in the value of the j-th covariate in the coefficient vector, when all the other covariates are given.

3.4 Perform multiple linear regression in R

See corresponding R file.

4 Regression assumptions and diagnostic

4.1 Introduction to random vector and matrix

Definition 4.1: Random vector

A vector whose elements are random variables.

Definition 4.2: Expectation and covariance of a random vector

Defined element-wise. For a random vector y :

$$E[y] = \begin{pmatrix} E\left[y_1\right] \\ E\left[y_2\right] \\ \vdots \\ E\left[y_n\right] \end{pmatrix}$$

For a random vector y :

$$\operatorname{Cov}(y) = E\left[(y - Ey)(y - Ey)^T\right]$$

Definition 4.3: Covariance matrix

The covariance matrix $\operatorname{Cov}(Y)$ is a symmetric matrix:

$$\operatorname{Cov}(Y) = \begin{pmatrix} \operatorname{Var}\left(y_1\right) & \operatorname{Cov}\left(y_1, y_2\right) & \ldots & \operatorname{Cov}\left(y_1, y_n\right) \\ \operatorname{Cov}\left(y_2, y_1\right) & \operatorname{Var}\left(y_2\right) & \ldots & \operatorname{Cov}\left(y_2, y_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}\left(y_n, y_1\right) & \operatorname{Cov}\left(y_n, y_2\right) & \ldots & \operatorname{Var}\left(y_n\right) \end{pmatrix}$$

Remark.

> $\operatorname{Cov}(Y)$ is symmetric as $\operatorname{Cov}\left(y_i, y_j\right) = \operatorname{Cov}\left(y_j, y_i\right)$.

Proposition 4.4

Suppose A is a constant matrix and $z = Ay$, where y is a random vector. Then:

- **Expectation of A:** If A is a constant matrix, $E(A) = A$.
- **Expectation of z:** Using the linearity property of expectation, $E(z) = E(Ay) = AE(y)$.
- **Covariance of z:** The covariance of the transformed vector $z = Ay$ is given by:

$$\operatorname{Cov}(z) = \operatorname{Cov}(Ay) = A \operatorname{Cov}(y) A^\top$$

Definition 4.5: Multivariate normal distribution

Let $X \sim N_m(\mu, \Sigma)$. The density function $f_X(x)$ for the multivariate normal distribution is given by:

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}$$

Where:

- m is the number of dimensions.
- μ is the mean vector.

- Σ is the covariance matrix.
- $|\Sigma|$ denotes the determinant of the covariance matrix.
- Σ^{-1} denotes the inverse of the covariance matrix.
- $(x-\mu)^T$ is the transpose of the vector $(x-\mu)$.

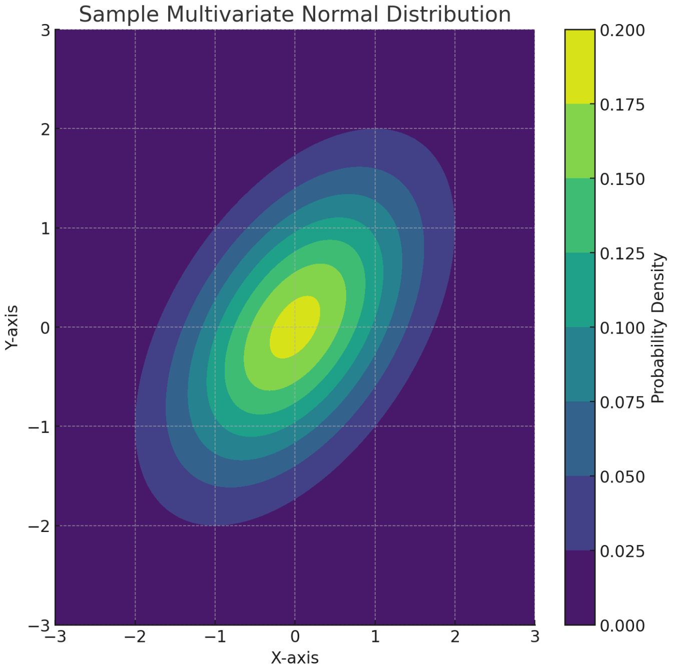


Figure 11: A example of multivariate normal distribution

4.2 Assumptions for linear regressions

The assumption for linear regressions are the following:

- **Linearity:** The relationship between the dependent variable y and the independent variable x is linear. Mathematically, this can be expressed as:

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

  The expected value of the error term ϵ_i is zero:

  $$E[\epsilon_i] = 0$$

  This implies that the expected value of y_i given x_i is:

  $$E[y_i] = \beta_0 + \beta_1 x_i$$

- **Uncorrelated errors:**

  $$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \text{for} \quad i \neq j$$

  The covariance between any two error terms should be zero, indicating no correlation between them.

- **Constant Variance (Homoskedasticity):**

  $$\mathrm{Var}(\epsilon_i) = \sigma^2 \quad \text{for any} \quad i$$

  The variance of the errors should be constant across all observations.

- **Normality:**

  $$\epsilon_i \sim N\left(0, \sigma^2\right)$$

  The errors are normally distributed with mean 0 and variance σ^2.

Remark.

> Under normality assumptions, ϵ_i and ϵ_j are uncorrelated iff they are independent

Proof. We will prove both directions:

- Assume ϵ_i and ϵ_j are uncorrelated i.e. $\operatorname{Cov}(\epsilon_i, \epsilon_j) = 0$.
  We know $\epsilon_i \sim N(0, \sigma^2)$ and $\epsilon_j \sim N(0, \sigma^2)$. This means the mean vector μ for (ϵ_i, ϵ_j) is:

  $$\mu = \begin{pmatrix} E[\epsilon_i] \\ E[\epsilon_j] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  And covariance matrix Σ for (ϵ_i, ϵ_j) is:

  $$\Sigma = \left(\begin{array}{cc} \operatorname{Var}(\epsilon_i) & \operatorname{Cov}(\epsilon_i, \epsilon_j) \\ \operatorname{Cov}(\epsilon_i, \epsilon_j) & \operatorname{Var}(\epsilon_j) \end{array} \right) = \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & \sigma^2 \end{array} \right)$$

  With the mean vector and covariance matrix defined, the joint distribution of (ϵ_i, ϵ_j) can be written as:

  $$(\epsilon_i, \epsilon_j) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & \sigma^2 \end{array} \right) \right)$$

  This indicates that (ϵ_i, ϵ_j) follows a bivariate normal distribution with the specified mean and covariance structure.
  The joint probability density function (pdf) of (ϵ_i, ϵ_j) for the bivariate normal distribution is given by:

  $$f_{\epsilon_i, \epsilon_j}(x_i, x_j) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \begin{pmatrix} x_i \\ x_j \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x_i \\ x_j \end{pmatrix} \right)$$

  Given:

  $$\Sigma = \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & \sigma^2 \end{array} \right)$$

  The inverse of $\mathbf{\Sigma}$ is:

  $$\Sigma^{-1} = \left(\begin{array}{cc} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{array} \right)$$

  The determinant of $\mathbf{\Sigma}$ is:

  $$|\Sigma| = \sigma^4$$

  So, the joint pdf simplifies to:

  $$f_{\epsilon_i, \epsilon_j}(x_i, x_j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2} \left(\frac{x_i^2}{\sigma^2} + \frac{x_j^2}{\sigma^2} \right) \right)$$
  $$f_{\epsilon_i, \epsilon_j}(x_i, x_j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x_i^2 + x_j^2}{2\sigma^2} \right)$$

  The marginal distributions of ϵ_i and ϵ_j are both normal:

  $$\epsilon_i \sim N\left(0, \sigma^2\right)$$
  $$\epsilon_j \sim N\left(0, \sigma^2\right)$$

The marginal pdf of ϵ_i is:

$$f_{\epsilon_i}\left(x_i\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_i^2}{2\sigma^2}\right)$$

The marginal pdf of ϵ_j is:

$$f_{\epsilon_j}\left(x_j\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_j^2}{2\sigma^2}\right)$$

To show independence, we need to prove that the joint pdf factorizes into the product of the marginal pdfs:

$$f_{\epsilon_i,\epsilon_j}\left(x_i,x_j\right)=f_{\epsilon_i}\left(x_i\right)f_{\epsilon_j}\left(x_j\right)$$

From above, we have:

$$f_{\epsilon_i,\epsilon_j}\left(x_i,x_j\right)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x_i^2+x_j^2}{2\sigma^2}\right)$$

And:

$$f_{\epsilon_i}\left(x_i\right)f_{\epsilon_j}\left(x_j\right)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_i^2}{2\sigma^2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_j^2}{2\sigma^2}\right)\right)$$

$$f_{\epsilon_i}\left(x_i\right)f_{\epsilon_j}\left(x_j\right)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x_i^2}{2\sigma^2}\right)\exp\left(-\frac{x_j^2}{2\sigma^2}\right)$$

$$f_{\epsilon_i}\left(x_i\right)f_{\epsilon_j}\left(x_j\right)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x_i^2+x_j^2}{2\sigma^2}\right)$$

Since:

$$f_{\epsilon_i,\epsilon_j}\left(x_i,x_j\right)=f_{\epsilon_i}\left(x_i\right)f_{\epsilon_j}\left(x_j\right)$$

We have shown that the joint pdf factorizes into the product of the marginal pdfs, proving that ϵ_i and ϵ_j are independent.

- Assume ϵ_i and ϵ_j are independent. Independence implies that the joint distribution of ϵ_i and ϵ_j can be written as the product of their marginal distributions:

  $$f\left(\epsilon_i,\epsilon_j\right)=f\left(\epsilon_i\right)f\left(\epsilon_j\right)$$

  For normally distributed variables, independence also implies that the covariance is zero:

  $$\operatorname{Cov}\left(\epsilon_i,\epsilon_j\right)=E\left[\left(\epsilon_i-E\left[\epsilon_i\right]\right)\left(\epsilon_j-E\left[\epsilon_j\right]\right)\right]=E\left[\epsilon_i\epsilon_j\right]-E\left[\epsilon_i\right]E\left[\epsilon_j\right]$$

  Given that $\epsilon_i\sim N\left(0,\sigma^2\right)$ and $\epsilon_j\sim N\left(0,\sigma^2\right)$:

  $$E\left[\epsilon_i\right]=0\quad\text{and}\quad E\left[\epsilon_j\right]=0$$

  Therefore, the covariance simplifies to:

  $$\operatorname{Cov}\left(\epsilon_i,\epsilon_j\right)=E\left[\epsilon_i\epsilon_j\right]-0\cdot0=E\left[\epsilon_i\epsilon_j\right]$$

  Since ϵ_i and ϵ_j are independent:

  $$E\left[\epsilon_i\epsilon_j\right]=E\left[\epsilon_i\right]E\left[\epsilon_j\right]=0$$

  Thus:

  $$\operatorname{Cov}\left(\epsilon_i,\epsilon_j\right)=0$$

  This proves that if ϵ_i and ϵ_j are independent, they are uncorrelated.

■

Remark.

> If these assumptions are true, then the observed residuals should behave in the similar fashion.

4.3 General regression assumptions

We have the following assumptions for regression in general:

1. The regression model can be written in matrix form as:

$$y = X\beta + \epsilon$$

   where:

   - y is the $n \times 1$ vector of observations of the dependent variable.
   - X is the $n \times p$ matrix of observations of the independent variables (including the intercept).
   - β is the $p \times 1$ vector of regression coefficients.
   - ϵ is the $n \times 1$ vector of error terms.

2. The error terms ϵ are normally distributed with mean 0 and covariance matrix $\sigma^2 I$:

$$\epsilon \sim N_n \left(0, \sigma^2 I\right)$$

Remark.

> If these assumptions are true, then the observed residuals should behave in the similar fashion.

4.4 Introduction to residue plot and Normal QQ plot

Definition 4.6: Residue plot

A residual plot is a graphical tool used in regression analysis to assess the goodness-of-fit of a regression model and check the underlying assumptions. It plots the residuals (the differences between observed and predicted values) on the vertical axis against a predictor variable or the predicted values on the horizontal axis.

Definition 4.7: Normal QQ plot

A Normal Q-Q (Quantile-Quantile) plot is a tool used to assess whether a set of data approximates a normal distribution. It compares the quantiles of the residuals from running the model on sample data to the quantiles of a standard normal distribution. The x axis is the theoretical quantiles from a standard normal distribution (with mean 0 and standard deviation 1). And y axis is the sample quantiles from the data.

Remark.

> The reference line in a normal Q-Q plot is a 45 degrees straight line that represents the expected values if the residue were perfectly normally distributed. If the residuals are normally distributed, the points will lie on or close to this reference line. Deviations from the line indicate departures from normality assumption.

Remark.

> Usually in a residual plot, the reference line is a horizontal line at zero on the y-axis. This line represents the ideal case where the residuals are perfectly centered around zero.

Example.

Consider the following data set:

case	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Table 3: Example Dataset

We have the following example residue plot with fitted value:

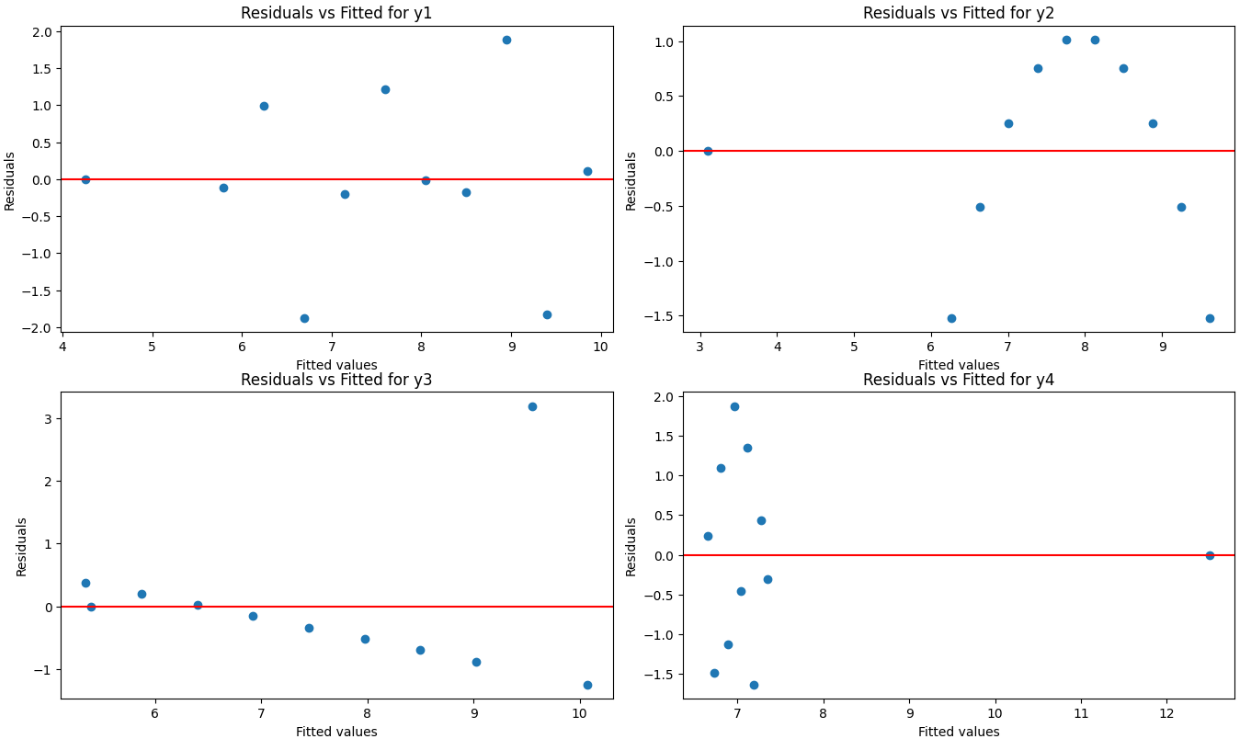


Figure 12: Example residue plot

We also have the following example normal QQ plot:

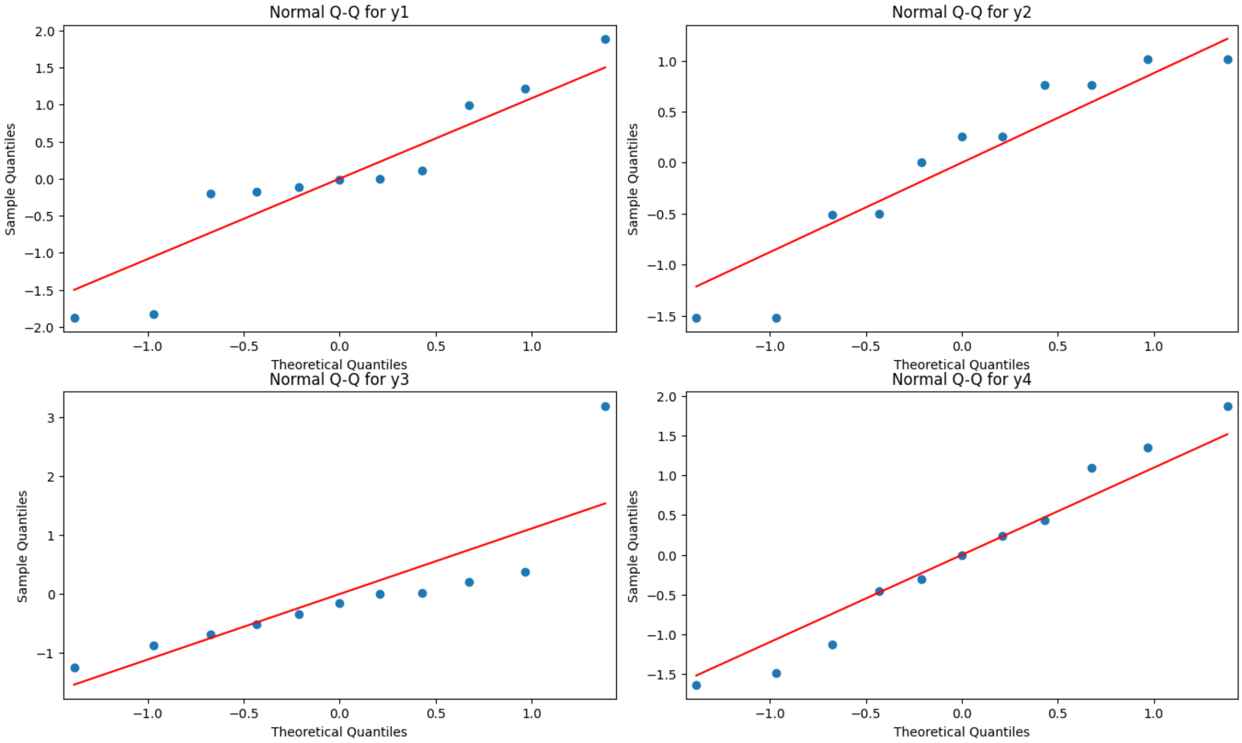


Figure 13: Example Normal QQ plot

We have the following example residue plot with predictor:

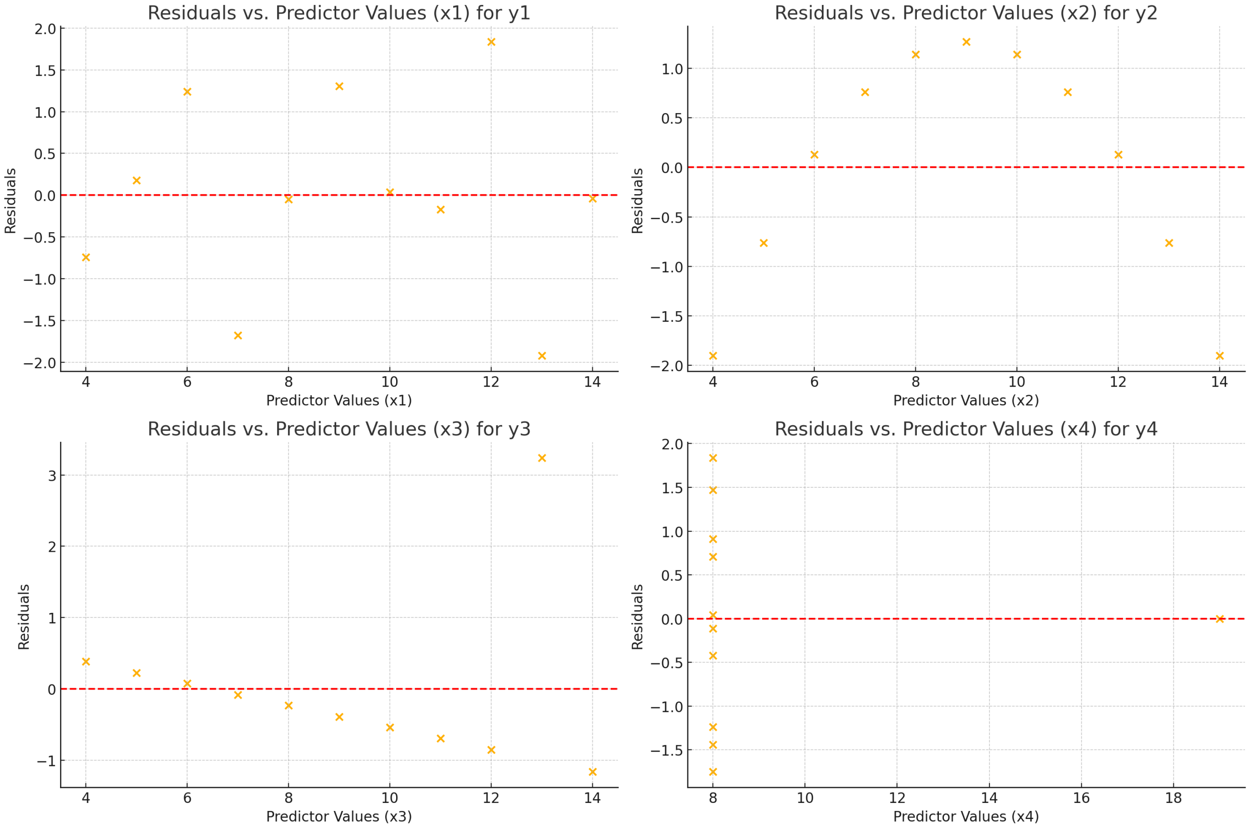


Figure 14: Example residue plot

4.5 Linear Regression diagnostic

Linear regression diagnostics are used to check the validity of the assumptions of a linear regression model when dealing with a specific data set. These assumptions include:

- **Linearity:** The relationship between the independent variables and the dependent variable is linear.
- **Homoscedasticity:** The residuals have constant variance at every level of the independent variables.
- **Independence:** The residuals (errors) are independent. There should be no correlation between consecutive residuals.
- **Normality:** The residuals of the model are normally distributed.

Linear regression diagnostics are crucial for ensuring the validity, reliability, and accuracy of a linear regression model when dealing with a specific data set. They help in understanding the relationships in the data, refining the model, communicating results effectively, and preventing misleading conclusions.

4.5.1 Diagnostic methods for detecting non-linearity

This is a straightforward method where the response variable (y_i) is plotted against the predictor variable (x_i). It helps to visually inspect the relationship between y_i and x_i.

- If the plot shows a clear linear trend, it suggests that the linear regression model might be appropriate.
- However, if the plot shows a curved or non-linear pattern, it indicates that a simple linear regression model might not be suitable.

Residuals ($e_i = y_i - \hat{y}_i$) are plotted against the predictor variables (x_i). This plot helps to identify any non-linear patterns in the residuals.

- In a well-fitted linear regression model, the residuals should be randomly scattered around zero, without any systematic pattern.
- If a pattern (such as a curve) is observed in the residual plot, it suggests non-linearity.

If non-linearity is detected, fit a different model such as quadratic regression model and plot the residuals e_i against x_i again. Check if the quadratic model removes the pattern observed in the simple linear regression residuals. A random scatter of residuals around zero without any pattern suggests a good fit. The model does not necessarily need to be quadratic.

Example.

Consider the data set that provide information on Salary versus Experience for professors:

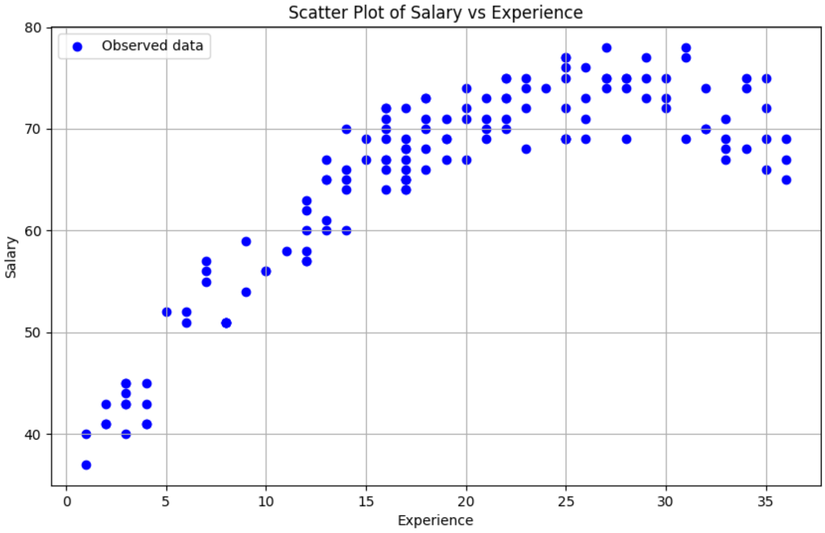


Figure 15: Example data set

The scatter plot shows a general increasing trend, but there are indications that the relationship might not be strictly linear. Specifically, the spread of Salary for different levels of Experience suggests potential non-linearity.

Next, we fit a simple linear regression model to the data and plot the residuals to check for patterns:

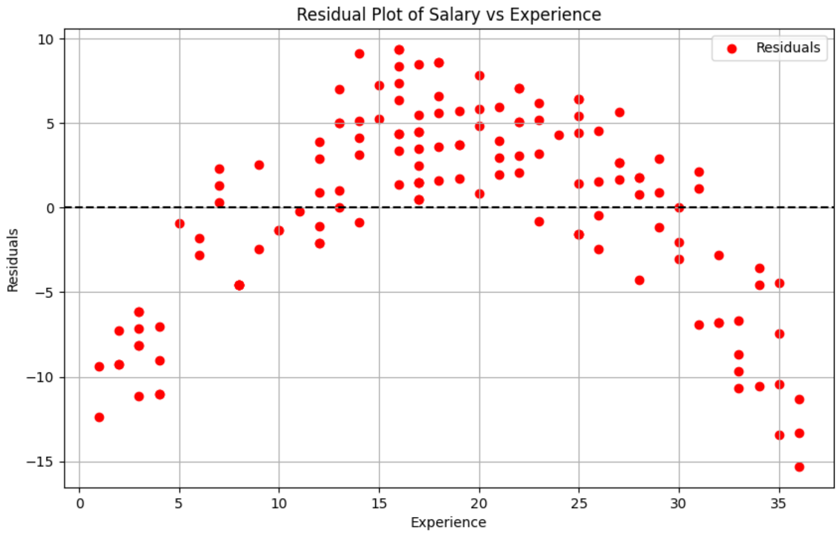


Figure 16: Residue plot for simple linear regression

The residual plot shows a clear pattern, indicating non-linearity. The funnel shape (wider spread of residuals at higher levels of Experience) suggests that the simple linear regression model does not adequately capture the relationship.

To address the non-linearity, we fit a quadratic regression model, including an additional term for the squared Experience. We choose quadratic regression model because the distribution of the initial data indicates a quadratic relationship.

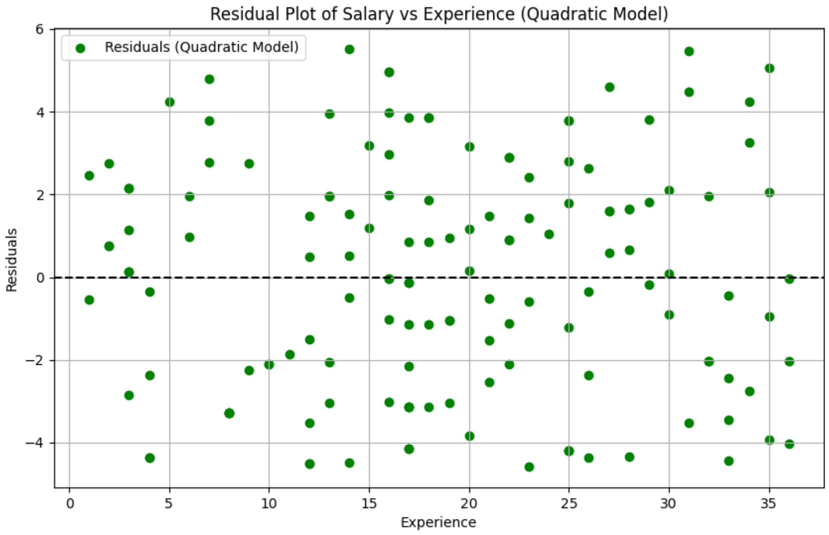


Figure 17: Residue plot for quadrative regression

The residual plot of the quadratic model shows a more random scatter around zero compared to the simple linear model. This indicates that the quadratic model better captures the non-linear relationship between Salary and Experience.

Example.

Consider the following result:

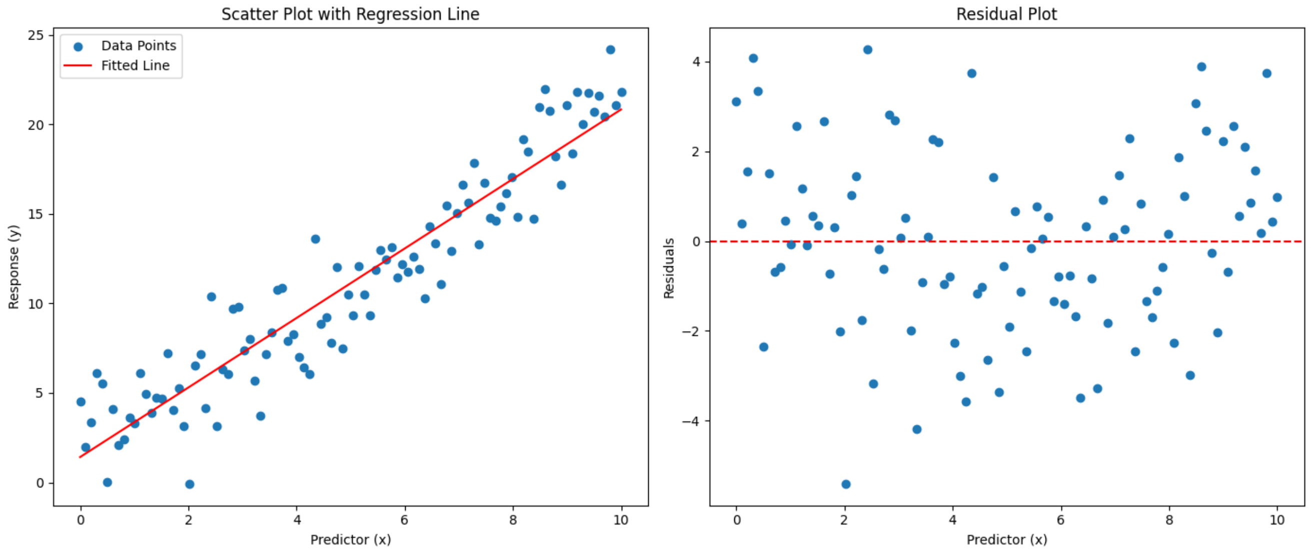


Figure 18: A regression that fits the linearity assumption

In this example, the scatter plot shows the data points along with the fitted regression line, while the residual plot shows the residuals against the predictor variable.

4.5.2 Diagnostic methods for detecting heteroskedsticity

Non-constant variance, or heteroskedasticity, violates one of the key assumptions of linear regression, which assumes that the variance of the error terms (ϵ_i) is constant (homoskedasticity).

Plotting the observed data points to inspect the spread of the response variable $(y_i\ )$ across different values of the predictor variable (x_i). Plotting the residuals ($e_i = y_i - \hat{y}_i$) against the predictor variable (x_i).

Inspect y_i versus x_i to check for changing spread in y_i values. Inspect e_i versus x_i to check for patterns in the spread of residuals. When ϵ_i has the same variance, the y_i and e_i values are expected to have a consistent spread for any value of the predictor.

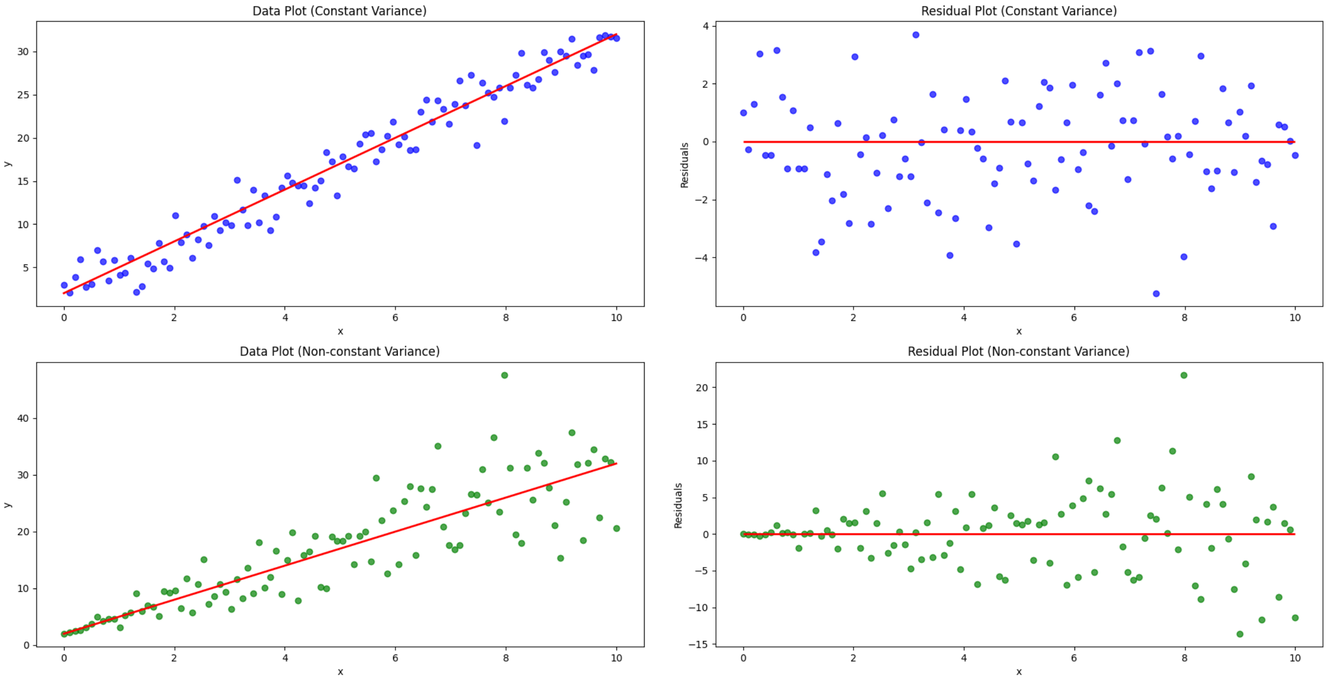


Figure 19: Non-constant and constant variance

We can see from the graph that:

1. For constant variance, the blue points represent the observed data with constant variance. The red line represents the fitted linear regression line. The spread of the data points around the regression line is consistent across all values of x.

   The residuals (differences between observed and fitted values) are plotted against x. The spread of the residuals is consistent and shows no pattern, indicating constant variance.

2. However, for non-constant variance, the green points represent the observed data with non-constant variance (variance increasing with x). The red line represents the fitted linear regression line. The spread of the data points around the regression line increases with x.

   The residuals are plotted against x. The residuals show a funnel shape, indicating that the variance increases with x, suggesting heteroskedasticity.

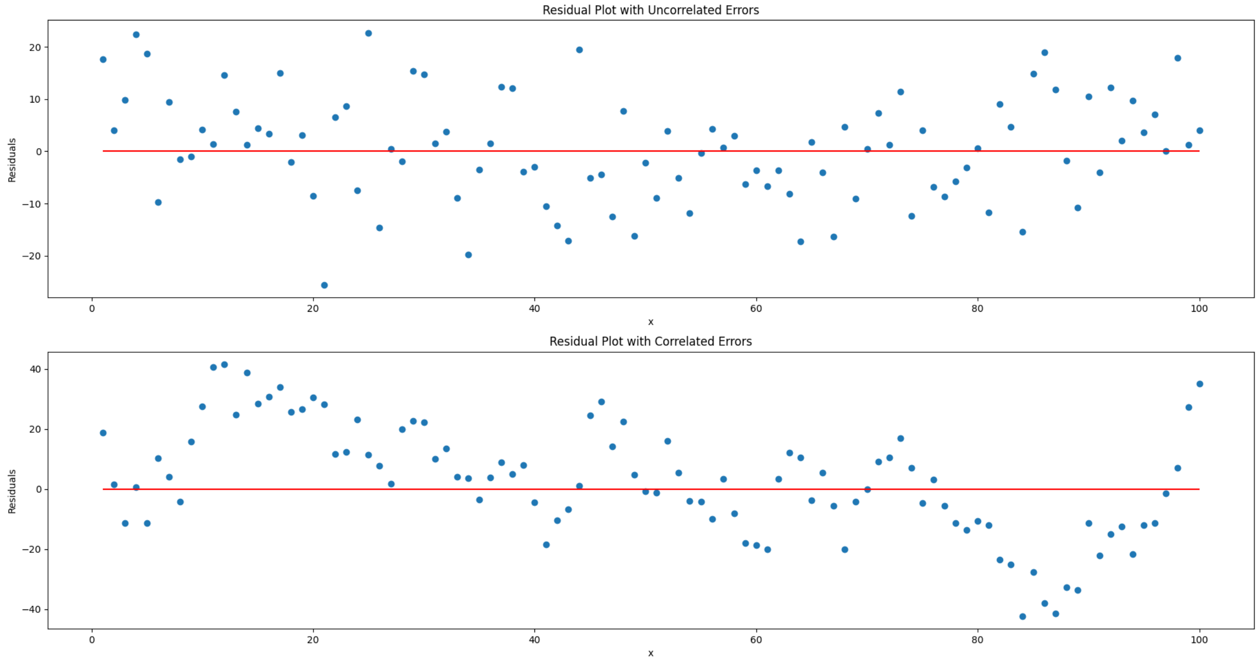
4.5.3 Diagnostic methods for detecting correlated errors

Plotting the data can sometimes reveal patterns that indicate correlated errors. For instance, if the errors ϵ_i show a trend over time or across spatial locations, this might suggest correlation. If residuals show a pattern (such as a wave or trend), it indicates correlated errors.

Example.

> We have a numeric example. Consider $\epsilon_1 \sim N(0,1)$ and $\epsilon_2 \sim N\left(\epsilon_1, 1\right)$. We have $E\epsilon_2 = E_1\left(E_2\left(\epsilon_2 \mid \epsilon_1\right)\right) = E_1\epsilon_1 = 0$. This means
>
> $$\begin{aligned} \operatorname{Cov}\left(\epsilon_1, \epsilon_2\right) &= E\left(\epsilon_1\epsilon_2\right) - E\left(\epsilon_1\right) E\left(\epsilon_2\right) \\ &= E\left(\epsilon_1\epsilon_2\right) \\ &= E_1\left(E_2\left(\epsilon_1\epsilon_2 \mid \epsilon_1\right)\right) \\ &= E_1\left(\epsilon_1 E_2\left(\epsilon_2 \mid \epsilon_1\right)\right) \\ &= E_1\epsilon_1^2 \\ &= 1 \end{aligned}$$

Example.



Residual Plot with Uncorrelated Errors

Residual Plot with Correlated Errors

Figure 20: Example of uncorrelated error and correlated error

For uncorrelated errors, points are randomly scattered without any visible pattern. This randomness suggests

> that the residuals are independent. For correlated errors, points show a clear pattern, such as clustering together, forming a wave, or displaying a trend over time. This pattern indicates that the residuals are not independent and are correlated.

4.5.4 Diagnostic methods for detecting normality

The normal QQ plot helps to visually assess if the residuals follow a normal distribution. If the residuals are normally distributed, the points on the QQ plot should fall approximately along a straight line. If the points deviate significantly from the line, it suggests that the residuals are not normally distributed.

Consider the case when the residuals are normally distributed with unit variance after applying linear regression, all the points on the QQ plot will fall approximately along a straight line:

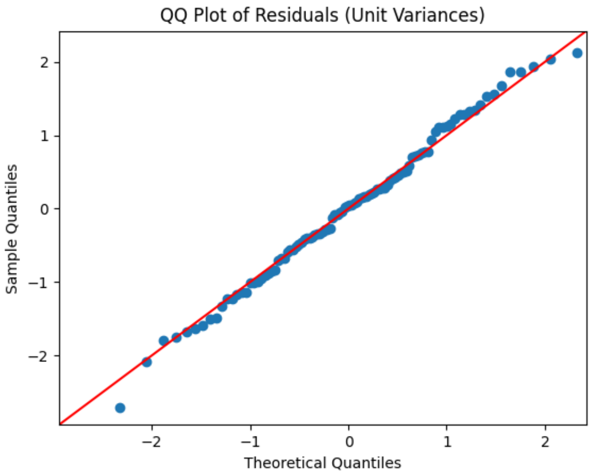


Figure 21: Residue with unit variance

Consider the case when the residuals are normally distributed with smaller variance than the unit variance after the linear regression, we can see significant deviations from the line, especially in the tails:

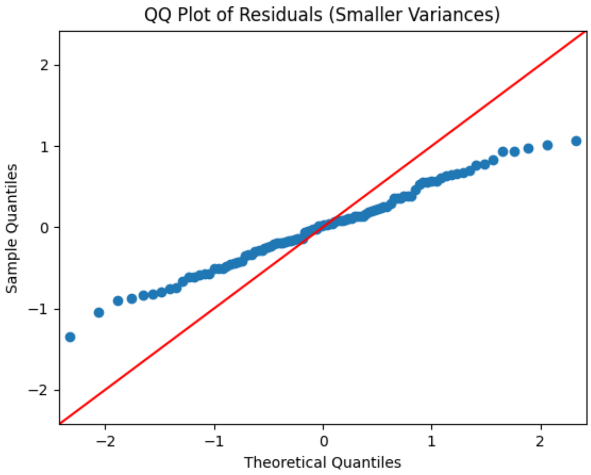


Figure 22: Residue with smaller variance

Consider the case when the residuals are normally distributed with bigger variance than the unit variance after applying linear regression, we can also see significant deviations from the line, especially in the tails:

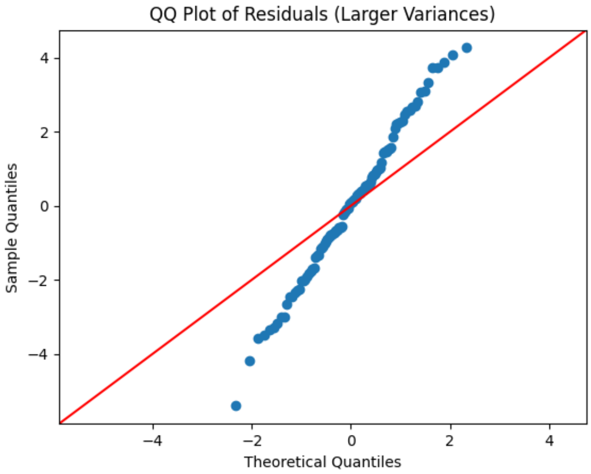


Figure 23: Residue with bigger variance

For uniform noise (uniformly distributed residue), the residuals will not follow a normal distribution. In the QQ plot, the points will deviate significantly from the reference line, particularly in the tails and central part of the plot:

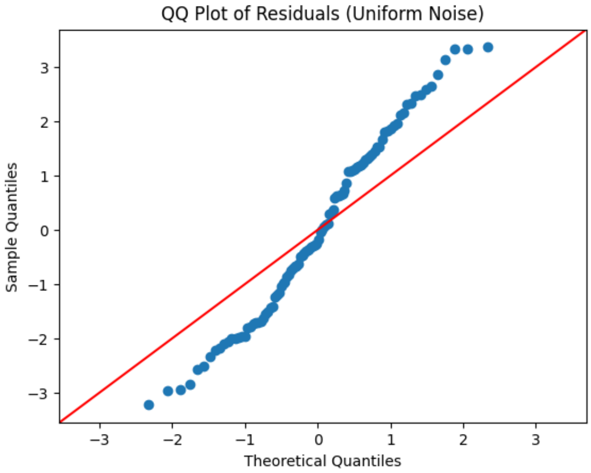


Figure 24: Residue with uniform noise

The scale of residuals (smaller or larger variances) affects the spread of points in the QQ plot but does not necessarily indicate a deviation from normality if the points still align with the reference line. Standardizing residuals ensures that residuals are on the same scale, making it easier to assess normality.

To achieve standardizing residuals, we can use the R:

```
# Set seed for reproducibility
set.seed(1234)

# Generate some sample data
a <- rnorm(100)
b <- rnorm(100)
y <- 2 * a - 3 * b + rnorm(100, mean = a^2 + b^2, sd = 2)

# Fit a linear model
model <- lm(y ~ a + b)

# Summary of the model
summary(model)

# Generate diagnostic plots, including the Normal Q-Q plot
plot(model)
```

We have the following normal QQ plot:

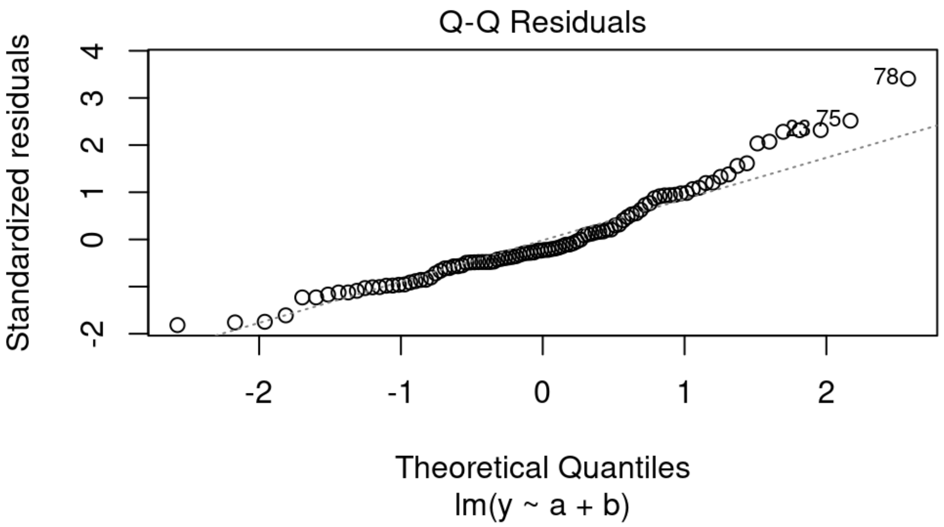


Figure 25: Example standardized normal QQ plot

When R generates the Normal Q-Q plot using standardized residuals, it follows these steps:

1. It calculates the standardized residuals from your regression model.
2. It sorts these standardized residuals and uses them as empirical quantiles.
3. It calculates the theoretical quantiles from a standard normal distribution.

We can see that the points in the Q-Q plot are generally on the reference line from the lower left to the upper right, so the data is basically normally distributed.

In addition to visually observing the plot, we can use Shapiro-Wilk test to statistically assess normality in R:

```
shapiro.test(resid(model))
```

This test exam the null hypothesis is that the population is normally distributed(low p value means not normally distributed). A low p-value (< 0.05) indicates that the residuals are not normally distributed.

4.5.5 Special section from the tutorial: Handling influential point

Definition 4.8: Outlier

An outlier is a data point that has an extreme response value (y-value) that differs significantly from the other observations.

Definition 4.9: Leverage point

A leverage point is a data point that has an extreme value for one or more predictor variables (x-values). It is far from the mean of the predictor variables.

Remark.

> Outliers can affect the fit of the regression model by increasing the residual sum of squares, leading to biased estimates of the coefficients.

Remark.

> Leverage points have the potential to influence the regression line because of their position in the predictor space. They can affect the slope and intercept even if their residuals are not large.

Remark.

> Leverage points do not necessarily have large residuals (differences between observed and predicted values). They are identified based on their position in the predictor space.

Remark.

> A data point can be both a leverage point and an outlier if it has an extreme predictor value and an extreme response value. Such points have a high influence on the regression model and can significantly distort the results.

Consider the following hypothetical scenario:

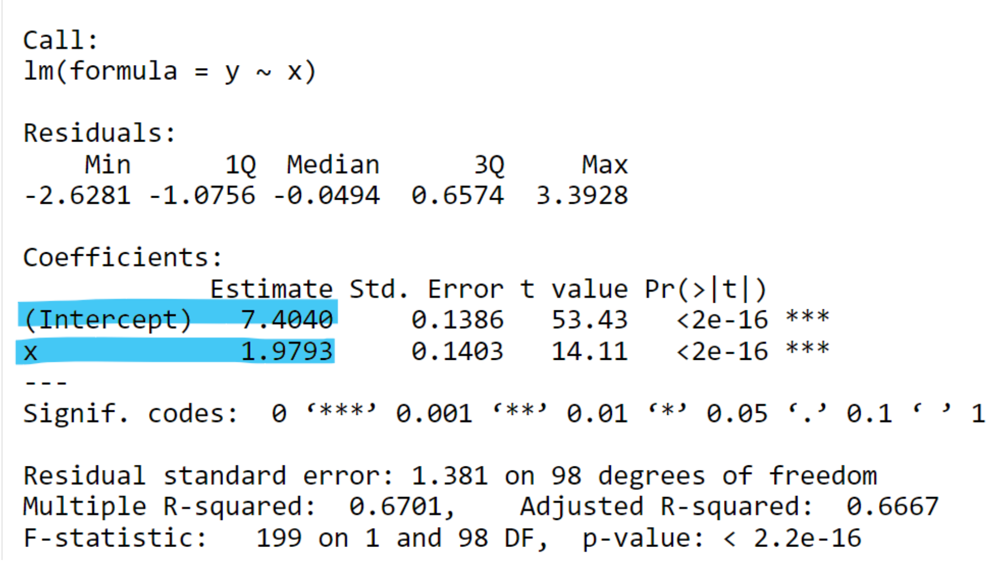
```r
set.seed(2212)

x = rnorm(100, mean = 0, sd = 1)
y = 7+2*x+rnorm(100, mean = 0, sd = 1.5)

model = lm(y~x)

summary(model)
```

We have the following result:



```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6281 -1.0756 -0.0494  0.6574  3.3928 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.4040     0.1386   53.43   <2e-16 ***
x             1.9793     0.1403   14.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.381 on 98 degrees of freedom
Multiple R-squared:  0.6701,	Adjusted R-squared:  0.6667 
F-statistic:   199 on 1 and 98 DF,  p-value: < 2.2e-16
```

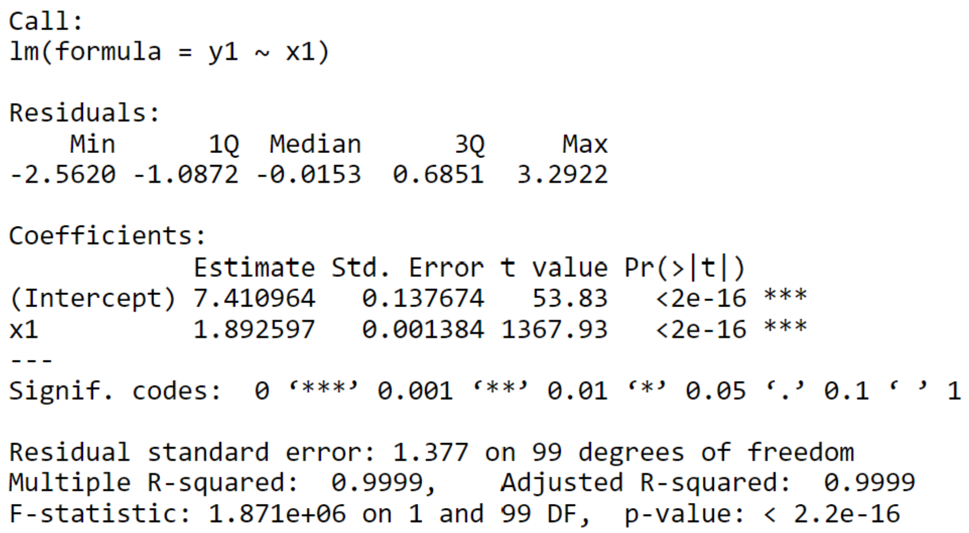
Figure 26: Summary of the model generated by R

```r
# Now adds a high leverage point to the original data.
x1 = c(x, 1000)
y1 = c(y, 1900)

# Fits a linear model with the outlier included.
model1 = lm(y1~x1)

summary(model1)
```

Now we have this result after adding the leverage point:



```
Call:
lm(formula = y1 ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5620 -1.0872 -0.0153  0.6851  3.2922 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.410964   0.137674   53.83   <2e-16 ***
x1          1.892597   0.001384 1367.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.377 on 99 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 1.871e+06 on 1 and 99 DF,  p-value: < 2.2e-16
```

Figure 27: Summary of the new model after adding the leverage point

The point at $x = 1000$ is far from the mean of x, making it a high leverage point. High leverage points have a significant influence on the regression line because they can pull the line towards themselves. This means if a high leverage point is included in the dataset, the algorithm adjusts the line to fit that point as closely as possible, often at the expense of fitting the other points less accurately.

We have a way to handle outliers and leverage point in the data set, it is called cook's distance.

Definition 4.10: Cook's distance

Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. It is good to use this method because it gives us influential points that are worth checking for validity. In the Cook's distance plot, points that lie above the red line 0.5 are flagged as potential influential points.

```r
# Fit a linear model
model1 <- lm(y ~ x, data = sample_data)

# Generate Cook's distance plot
plot(model1, which=4)
```

This generates the following graph with the numbered potential influential observations:

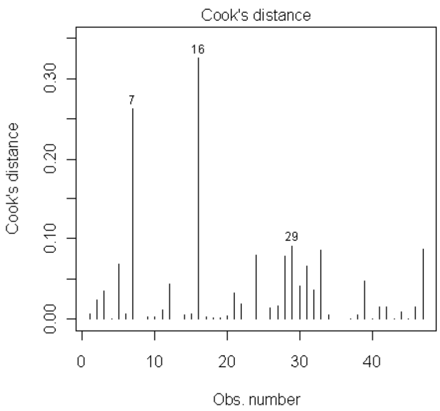


Figure 28: Example cook distance graph

4.5.6 Special section from the tutorial: Added variable plot in R

Definition 4.11: Added variable plot

An Added Variable Plot (also known as a Partial Regression Plot) is a diagnostic tool used in multiple regression analysis. It shows the relationship between a response variable and a predictor variable, while holding constant and adjusting for the effects of other predictor variables in the model.

Definition 4.12: Omitted-variable bias

Omitted-variable bias (OVB) occurs when a statistical model leaves out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to those that were included. Using added variable plot is a good way to avoid omitted-variable bias.

To create added variable plots in R, we will use the avPlots() function from the car library. Here is an example code block:

```r
# Load the car library
library(car)
# Fit a linear regression model
model <- lm(mpg ~ disp + hp + drat + am, data = mtcars)
# Create added variable plots
avPlots(model)
```

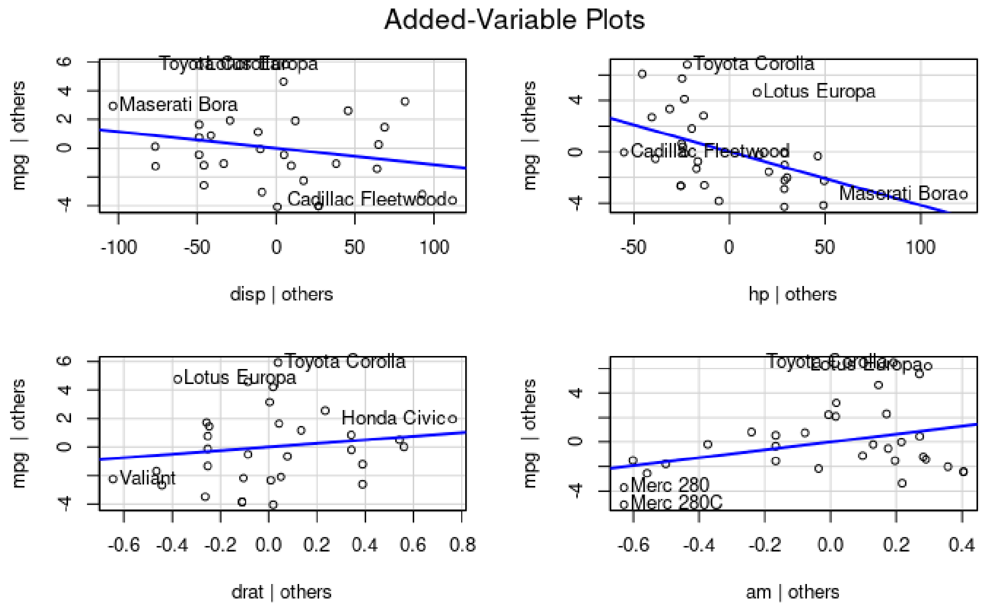


Figure 29: Added variable plot generated

When we regress Y on all the predictor variables except X, we get:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon$$

The residuals from this regression, $\hat{e}_Y$, represent the part of Y not explained by all the variables other than X. Similarly, when we regress X on all the other predictor variables, we get:

$$X = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_{p-1} X_{p-1} + \eta$$

The residuals from this regression, $\hat{e}_X$, represent the part of X not explained by the other variables. The added variable plot is a scatter plot of $\hat{e}_Y$ versus $\hat{e}_X$. The simple regression of $\hat{e}_Y$ on $\hat{e}_X$:

$$\hat{e}_Y = \gamma \hat{e}_X + \epsilon$$

The slope γ of this regression line is equivalent to the coefficient β_X from the full multiple regression model. The intercept will be 0. A strong linear relationship in the added variable plot indicates the increased importance of the contribution of X to the model already containing the other predictors.

4.6 Relevant coding in R

See corresponding R code.

5 Transformation and sampling distribution

5.1 Transformation

In the previous few sections, we have introduced assumptions we have for linear regression model, and some potential issues with using the linear model. These issues includes but not limited to:

1. important predictor variables are omitted.
2. error terms are not independent.
3. outliers exist.
4. unequal error variances.
5. response is not a linear function for the predictor.

Transforming response and/or predictor variables therefore has the potential to remedy a number of model problems.

Definition 5.1: counting data

A data type describing countable quantities, data which can take only the counting numbers, nonnegative integer values $\{0, 1, 2, 3, \ldots\}$, and where these integers arise from counting rather than ranking.

Definition 5.2: Binary data

A data type consisting of categorical data that can take exactly two possible values.

5.1.1 Square root transformation

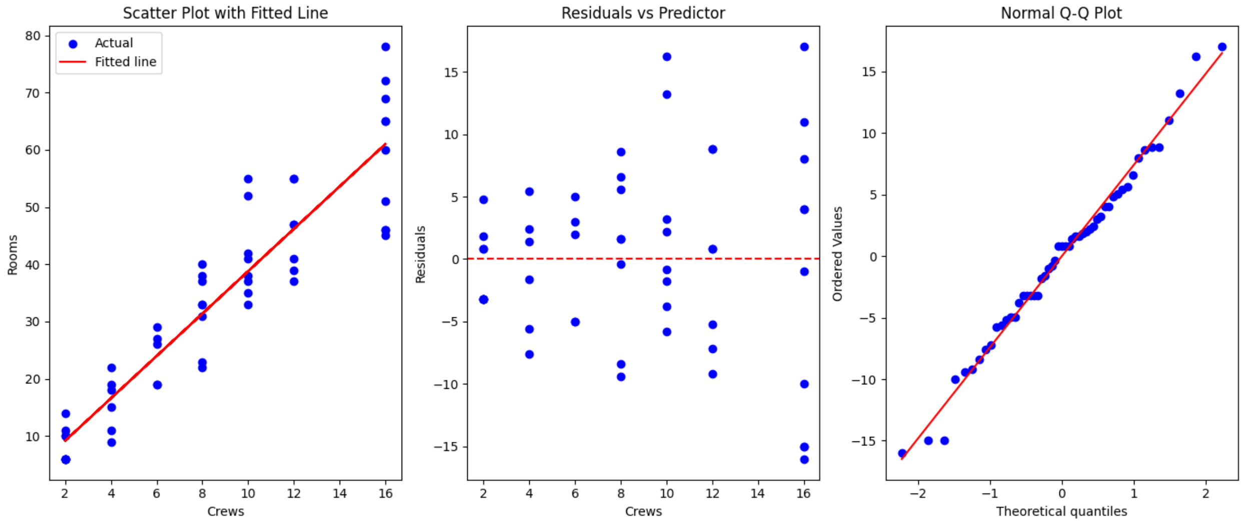
To introduce basic ideas behind data transformations we first consider a simple linear regression model example.

Example.

> A building maintenance company wants to bid on a contract to clean corporate offices. And we have data contains number of rooms cleaned and number of cleaning crews over 53 days:
>
> Table 4: Number of crews and rooms cleaned for each case
>
> | Case | Number of crews | Rooms cleaned | Case | Number of crews | Rooms cleaned |
> |---|---|---|---|---|---|
> | 1 | 16 | 51 | 28 | 4 | 18 |
> | 2 | 10 | 37 | 29 | 16 | 72 |
> | 3 | 12 | 37 | 30 | 8 | 22 |
> | 4 | 16 | 46 | 31 | 10 | 55 |
> | 5 | 16 | 45 | 32 | 16 | 65 |
> | 6 | 4 | 11 | 33 | 6 | 26 |
> | 7 | 2 | 6 | 34 | 10 | 52 |
> | 8 | 4 | 19 | 35 | 12 | 55 |
> | 9 | 6 | 29 | 36 | 8 | 33 |
> | 10 | 2 | 14 | 37 | 10 | 38 |
> | 11 | 12 | 47 | 38 | 8 | 23 |
> | 12 | 8 | 37 | 39 | 8 | 38 |
> | 13 | 16 | 60 | 40 | 2 | 10 |
> | 14 | 2 | 6 | 41 | 16 | 65 |
> | 15 | 2 | 11 | 42 | 8 | 31 |
> | 16 | 2 | 10 | 43 | 8 | 33 |
> | 17 | 6 | 19 | 44 | 12 | 47 |

18	10	33	45	10	42
19	16	46	46	16	78
20	16	69	47	2	6
21	10	41	48	2	6
22	6	19	49	8	40
23	2	6	50	12	39
24	6	27	51	4	9
25	10	35	52	4	22
26	12	55	53	12	41
27	4	15			

Both number of crews and number of rooms are counting data. And after the run the linear model, we have:


Scatter Plot with Fitted Line

Residuals vs Predictor

Normal Q-Q Plot

Figure 30: Data scatter plot, residue plot, and normal QQ

We see from the scatter plot of the data shows increasing variability in the number of rooms cleaned as the number of crews increases. The residue plot also shows uneven dispersion. Thus, the assumption that the variance of the errors is constant appears to be violated in this case. This indicates a need for transformation.

Definition 5.3: Square root transformation

We have y_i (number of rooms cleaned for example), x_i (number of crews for example). We transformed Variables $\tilde{y}_i = \sqrt{y_i}$ and $\tilde{x}_i = \sqrt{x_i}$. The transformed variables are used in a simple linear regression model $\tilde{y}_i = \beta_0 + \beta_1 \tilde{x}_i + \epsilon_i$. we can use the square root transformation to stabilize the variance in the data.

Example.

Given the same scenario, after we apply the transformation, we have the following result:

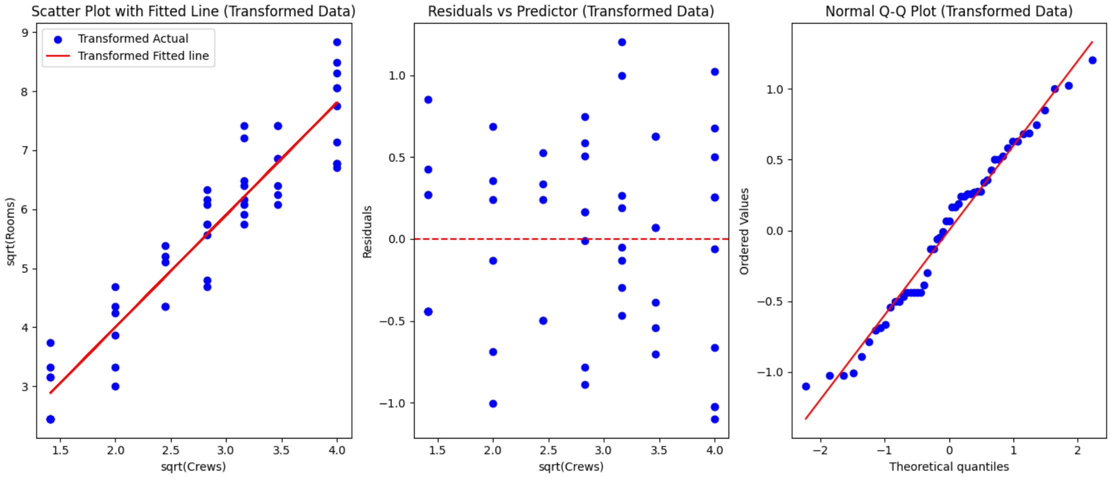


Figure 31: Results of the transformation

The residual plot after the square root transformation shows improved variance stability, indicating the transformation was effective.

5.1.2 Log transformation

Definition 5.4: Log transformation

The log transformation is another technique used to address violations of regression assumptions, specifically non-linearity. The purposes are to linearize relationships between variables. We have the original model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, and Log-Transformed model $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$. Here, both the response variable y_i and the predictor variable x_i are log-transformed.

Example.

Consider the following example with data set describing the relationship between price and sales for canned food:

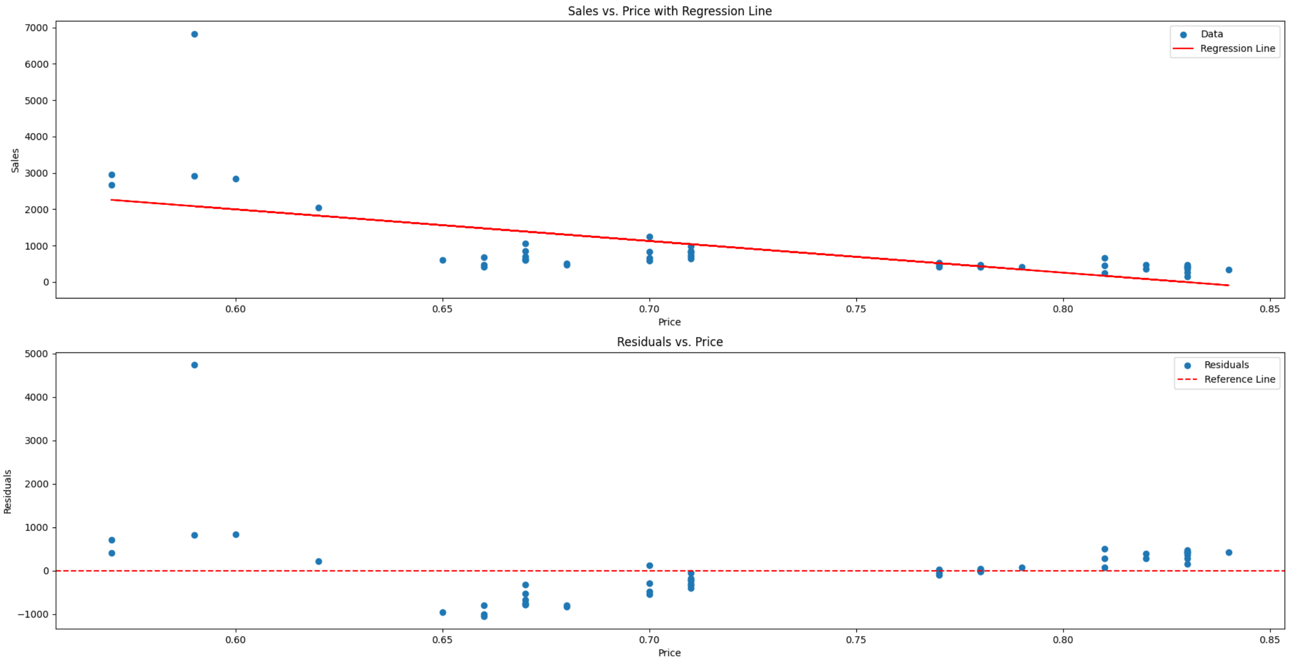
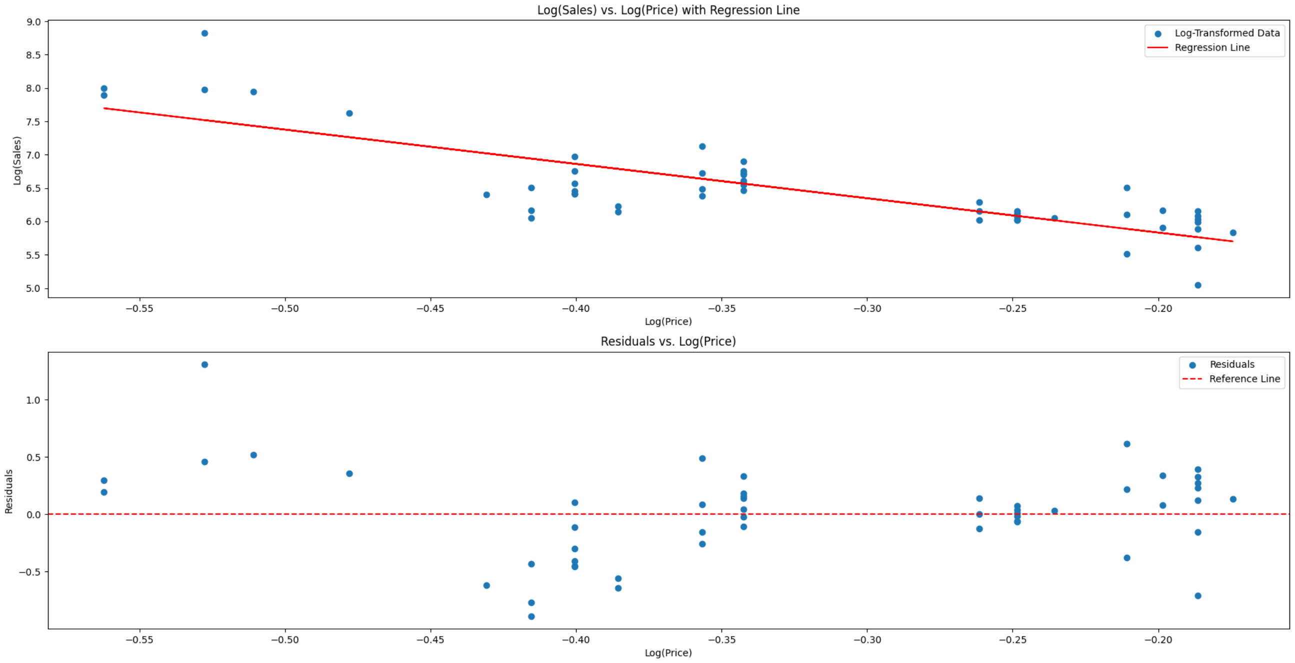


Figure 32: non-linearity violation

It is clear that linear model does not fit with the data and the residue plot has shown a trend. Thus, we have

> a linearity assumption violation.

Example.

> To solve the issue above, we have use logarithm transformation:
>
> 
>
>
> Figure 33: Improvements with log transformation
>
> Thus, we have seen some improvements in linearity.

5.1.3 Box-Cox Transformation

The Box-Cox transformation is a powerful tool used to stabilize variance and make data more closely follow a normal distribution. It is especially useful when dealing with non-normal response variables that violate the assumptions of linear regression.

Definition 5.5: Box-Cox transformation

The Box-Cox transformation is defined by a parameter λ and transforms the positive response variable y_i as follows: $\widetilde{y}_i(\lambda) = \Psi\left(y_i, \lambda\right) \operatorname{gm}(\boldsymbol{y})^{1-\lambda}$, where

$$\Psi\left(y_i, \lambda\right) = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y_i & \text{if } \lambda = 0 \end{cases}$$

is the power transformation, and

$$\operatorname{gm}(\boldsymbol{y}) = \Pi_{i=1}^{n} y_i^{1/n} = \sqrt[n]{\prod_{i=1}^{n} y_i}$$

is the geometric mean.

After applying the Box-Cox transformation, the transformed response variable $y_i(\lambda)$ is used in a regression model:

$$y_i(\lambda) = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where ϵ_i are the residuals assumed to be normally distributed.

5.2 MLE approaches to coefficients for the transformations

Assume we have the regression assumption $\epsilon_i \overset{i,i,d}{\sim} N\left(0, \sigma^2\right)$. And we know $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$ independently. The likelihood function represents the probability of observing the given data as a function of the parameters $\left(\beta_0, \beta_1, \sigma^2\right)$. For a normal distribution, the probability density function is:

$$f\left(y_i; \beta_0, \beta_1, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right)$$

The likelihood function for all observations $y_1, y_2, \ldots, y_n$ is the product of individual densities:

$$\begin{aligned} L\left(\beta_0, \beta_1, \sigma^2; y\right) &= \prod_{i=1}^{n} f\left(y_i; \beta_0, \beta_1, \sigma^2\right) \\ L\left(\beta_0, \beta_1, \sigma^2; y\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right) \end{aligned}$$

The log-likelihood function is often used because it simplifies the product of probabilities into a sum, making it easier to differentiate.

$$\begin{aligned} \log L\left(\beta_0, \beta_1, \sigma^2; y\right) &= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right)\right] \\ \log L\left(\beta_0, \beta_1, \sigma^2; y\right) &= \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right] \\ \log L\left(\beta_0, \beta_1, \sigma^2; y\right) &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right] \\ \log L\left(\beta_0, \beta_1, \sigma^2; y\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 \end{aligned}$$

To find the estimates of β_0, β_1, and σ^2, we need to maximize the log-likelihood function with respect to these parameters. First, consider the part of the log-likelihood that involves β_0 and β_1 :

$$\log L\left(\beta_0, \beta_1; y\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

This is equivalent to minimizing the residual sum of squares (RSS):

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

Taking partial derivatives with respect to β_0 and β_1 and setting them to zero gives:

$$\begin{aligned} \frac{\partial \mathrm{RSS}}{\partial \beta_0} &= -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \\ \frac{\partial \mathrm{RSS}}{\partial \beta_1} &= -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) x_i = 0 \end{aligned}$$

Solving these equations yields the ordinary least squares (OLS) estimators:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Next, to estimate σ^2, we take the derivative of the log-likelihood with respect to σ^2 :

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Setting this to zero and solving for σ^2 gives:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Remark.

> MLE above assumes normal distribution for the residuals based on the regression model's assumptions. However, MLE can be adapted to other distributions, such as Poisson or Bernoulli, depending on the nature of the response variable.

Remark.

> MLE provides a general framework for parameter estimation that is not limited to specific types of data or models. This is particularly useful in the context of models with residue following a different distribution and transformation method.

Given the model:

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

Apply the log transformation:

$$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$$
$$\text{where } \epsilon_i \sim N\left(0, \sigma^2\right).$$

The likelihood function for the normal distribution is:

$$L\left(\beta_0, \beta_1, \sigma^2; y\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y_i) - \beta_0 - \beta_1 \log(x_i))^2}{2\sigma^2} \right)$$

Taking the natural logarithm of the likelihood function gives the log-likelihood function:

$$\log L\left(\beta_0, \beta_1, \sigma^2; y\right) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log(y_i) - \beta_0 - \beta_1 \log(x_i))^2$$

To maximize the log-likelihood, we need to minimize the residual sum of squares (RSS):

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

The estimators for β_0 and β_1 can be found by solving the normal equations derived from the partial derivatives of the log-likelihood function with respect to β_0 and β_1.

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

The variance σ^2 is estimated by:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 \log(x_i) \right)^2$$

We have the following Box-Cox transformation:

$$\tilde{y}_i(\lambda) = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. The likelihood function for normally distributed residuals is:

$$L\left(\beta_0, \beta_1, \sigma^2; \tilde{y}(\lambda)\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\tilde{y}_i(\lambda) - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right)$$

The log-likelihood function is:

$$\log L\left(\beta_0, \beta_1, \sigma^2 ; \tilde{y}(\lambda)\right)=-\frac{n}{2} \log (2 \pi)-\frac{n}{2} \log \left(\sigma^2\right)-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(\tilde{y}_i(\lambda)-\beta_0-\beta_1 x_i\right)'$$

To find the parameter estimates, we maximize the log-likelihood function. This involves minimizing the residual sum of squares (RSS):

$$\operatorname{RSS}(\lambda)=\sum_{i=1}^n\left(\tilde{y}_i(\lambda)-\beta_0-\beta_1 x_i\right)^2$$

The estimators for β_0 and β_1 are derived by setting the partial derivatives of the loglikelihood function with respect to β_0 and β_1 to zero:

$$\hat{\beta}_0=\overline{\tilde{y}(\lambda)}-\hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right)\left(\tilde{y}_i(\lambda)-\overline{\tilde{y}(\lambda)}\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}$$

These are the ordinary least squares (OLS) estimators for the transformed model. The variance σ^2 is estimated by:

$$\hat{\sigma}^2=\frac{1}{n} \sum_{i=1}^n\left(\tilde{y}_i(\lambda)-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$$

To find the optimal λ, compute the RSS for different values of λ and choose the value that minimizes the RSS. A common approach is to use a grid search over a range of λ values $\left\{-1,-\frac{1}{2},-\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1\right\}$ and select the λ that gives the smallest RSS.

5.3 Sampling distributions for the coefficients of a multiple linear regression

Recall that multiple linear regression extends simple linear regression to include multiple predictor variables.

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

Where:

- y_i is the response variable.
- $x_{i, j}$ are the predictor variables for sample i.
- β_j are the regression coefficients.
- ϵ_i are the error terms assumed to be normally distributed with mean 0 and variance σ^2.

The model can be written in matrix form as:

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

Where:

- $\mathbf{y}$ is an $n \times 1$ vector of responses.
- $\mathbf{X}$ is an $n \times(p+1)$ design matrix, including a column of ones for the intercept.
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors.

Assume The errors ϵ_i are independently and identically distributed as $N\left(0, \sigma^2\right)$. And the predictors $x_{i, j}$ are fixed (non-random).

The least squares estimator for the regression coefficients $\hat{\boldsymbol{\beta}}$ is given by:

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

To find the expectation of $\hat{\boldsymbol{\beta}}$, we start by substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

Simplify this expression:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}$$

Using the fact that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}$ (the identity matrix), we get:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}$$

Taking the expectation:

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

Since $\boldsymbol{\beta}$ is a constant and the expectation of the error term $\boldsymbol{\epsilon}$ is zero (i.e., $E[\boldsymbol{\epsilon}] = \mathbf{0}$):

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E[\boldsymbol{\epsilon}]$$
$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{0}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

The variance of $\hat{\boldsymbol{\beta}}$ is given by:

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

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

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \operatorname{Var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right)$$

Since $\boldsymbol{\beta}$ is a constant, it does not contribute to the variance. We focus on the error term:

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \operatorname{Var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right)$$

Using the property that $\operatorname{Var}(\mathbf{A}\mathbf{b}) = \mathbf{A}\operatorname{Var}(\mathbf{b})\mathbf{A}^T$ for any matrix $\mathbf{A}$ and vector $\mathbf{b}$, we get:

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\operatorname{Var}(\boldsymbol{\epsilon})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

Since the errors $\boldsymbol{\epsilon}$ are assumed to be independently and identically distributed with variance σ^2, we have $\operatorname{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$:

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

Simplify the expression:

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$
$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

Given the assumptions, $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution:

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

5.3.1 Sampling distribution for simple linear regression

In simple linear regression, we model the relationship between a dependent variable y and an independent variable x using the equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 is the intercept, β_1 is the slope, and ϵ represents the error term.

The design matrix X for n observations is given by:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

To find the sampling distribution of the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, we need $X^T X$:

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

For a 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The inverse is:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Applying this to $X^T X$:

$$\left(X^T X\right)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

Using the sum of squares of x_i (denoted as S_{XX}):

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, we can rewrite the inverse matrix as:

$$\left(X^T X\right)^{-1} = \frac{1}{S_{XX}} \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} & \frac{-\bar{x}}{S_{XX}} \\ \frac{-\bar{x}}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix}$$

Given that:

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

Where $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$, we can derive the sampling distribution.

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\hat{x}^2}{S_{xx}} & \frac{-\bar{x}}{S_{XX}} \\ \frac{-\bar{x}}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix} \right)$$

This gives:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\hat{x}^2}{S_{XX}}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

6 Inference on linear regression

6.1 Inference on Simple Linear Regression

6.1.1 Estimation of variance σ^2

From our previous section, we know that for simple linear regression, the MLE estimator for σ^2 is:

$$\hat{\sigma}^2_{ML} = \frac{1}{n}RSS = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\hat{e}^T\hat{e}$$

We have:

$$E\left[\hat{e}^T\hat{e}\right] = E\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]$$

In multiple linear regression, we have:

$$\hat{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

where $\boldsymbol{H}$ is the hat matrix given by $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$. To find the expected value of the residuals, we take the expectation of both sides:

$$E[\hat{\boldsymbol{e}}] = E[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}]$$

Since $\boldsymbol{y}$ is a vector of observed values and can be expressed as:

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

where ϵ is the vector of error terms. Thus,

$$E[\hat{\boldsymbol{e}}] = E[(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})]$$

Expanding this and using the linearity of expectation, we get:

$$E[\hat{\boldsymbol{e}}] = (\boldsymbol{I} - \boldsymbol{H})E[\boldsymbol{X}\boldsymbol{\beta}] + (\boldsymbol{I} - \boldsymbol{H})E[\boldsymbol{\epsilon}]$$

Since $\boldsymbol{\beta}$ is a constant vector and $E[\boldsymbol{\epsilon}] = 0$ (by the assumption that the error terms have a mean of zero), we get:

$$E[\hat{\boldsymbol{e}}] = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \boldsymbol{H})0 = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}$$

Now, recall that the hat matrix $\boldsymbol{H}$ projects $\boldsymbol{X}\boldsymbol{\beta}$ onto the column space of $\boldsymbol{X}$, meaning:

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}$$

Therefore,

$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = 0$$

Thus, the mean of the residuals is:

$$E[\hat{\boldsymbol{e}}] = 0$$

To find the variance of the residuals, we need to compute the covariance matrix of $\hat{e}$. Given that $\boldsymbol{y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$:

$$\operatorname{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$$

Now, considering $\hat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$:

$$\operatorname{Cov}(\hat{\boldsymbol{e}}) = \operatorname{Cov}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}]$$

Using the property that $\operatorname{Cov}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\operatorname{Cov}(\boldsymbol{y})\boldsymbol{A}^\top$ for any matrix $\boldsymbol{A}$:

$$\operatorname{Cov}(\hat{\boldsymbol{e}}) = (\boldsymbol{I} - \boldsymbol{H})\operatorname{Cov}(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{H})^\top$$

$$\operatorname{Cov}(\hat{\boldsymbol{e}}) = (\boldsymbol{I} - \boldsymbol{H})\left(\sigma^2\boldsymbol{I}\right)(\boldsymbol{I} - \boldsymbol{H})^\top$$

$$\operatorname{Cov}(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top$$

Next, we simplify $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top$:

$$(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})$$

Expanding the multiplication:

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}) &= \boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H})-\boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H}) \\ &= \boldsymbol{I}-\boldsymbol{H}-\boldsymbol{H}+\boldsymbol{H}^2\end{aligned}$$

Simplifying further:

$$=\boldsymbol{I}-2\boldsymbol{H}+\boldsymbol{H}^2$$

Since $\boldsymbol{H}$ is an idempotent matrix (i.e., $\boldsymbol{H}^2=\boldsymbol{H}$):

$$=\boldsymbol{I}-2\boldsymbol{H}+\boldsymbol{H}=\boldsymbol{I}-\boldsymbol{H}$$

Putting it all together:

$$\operatorname{Cov}(\hat{\boldsymbol{e}})=\sigma^2(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})^{\top}=\sigma^2(\boldsymbol{I}-\boldsymbol{H})$$

Thus, the covariance of the residuals $\hat{\boldsymbol{e}}$ is $\sigma^2(\boldsymbol{I}-\boldsymbol{H})$.

Since the residuals $\hat{e}$ are normally distributed with mean zero and covariance matrix $\sigma^2(I-H)$

$$\hat{e}\sim N\left(0,\sigma^2(I-H)\right)$$

To find the expected value of $\hat{\boldsymbol{e}}^{\top}\hat{\boldsymbol{e}}$, we start by using the properties of the trace operator:

$$E\left(\hat{\boldsymbol{e}}^{\top}\hat{\boldsymbol{e}}\right)=E\operatorname{tr}\left(\hat{\boldsymbol{e}}^{\top}\hat{\boldsymbol{e}}\right)$$

Using the linearity of the expectation and the trace operator:

$$=\operatorname{tr}E\left(\hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^{\top}\right)$$

From the covariance of the residuals, $\operatorname{Cov}(\hat{e})=\sigma^2(\boldsymbol{I}-\boldsymbol{H})$:

$$=\sigma^2\operatorname{tr}(\boldsymbol{I}-\boldsymbol{H})$$

The trace of the matrix $\boldsymbol{I}-\boldsymbol{H}$ is calculated as:

$$\operatorname{tr}(\boldsymbol{I}-\boldsymbol{H})=\operatorname{tr}(\boldsymbol{I})-\operatorname{tr}(\boldsymbol{H})$$

Since the trace of the identity matrix $\boldsymbol{I}$ of size $n\times n$ is n :

$$=n-\operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\right)$$

Using the property of the trace and the hat matrix:

$$\begin{aligned}&=n-\operatorname{tr}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right) \\ &=n-\operatorname{tr}\left(\boldsymbol{I}_{p+1}\right)\end{aligned}$$

where $\boldsymbol{I}_{p+1}$ is the identity matrix of size $(p+1)\times(p+1)$. Therefore:

$$\begin{aligned}&=n-(p+1) \\ &=n-p-1\end{aligned}$$

Using the previous results:

$$E\hat{\sigma}_{ML}^2=\frac{1}{n}E\hat{\boldsymbol{e}}^{\top}\hat{\boldsymbol{e}}=\frac{\sigma^2}{n}\operatorname{tr}(\boldsymbol{I}-\boldsymbol{H})=\frac{n-p-1}{n}\sigma^2$$

Thus, $\hat{\sigma}_{ML}^2$ is biased because:

$$E\hat{\sigma}_{ML}^2=\frac{n-p-1}{n}\sigma^2<\sigma^2$$

To obtain an unbiased estimator, we need to adjust $\hat{\sigma}_{ML}^2$ by dividing by the correct degrees of freedom:

$$S^2 = \frac{1}{n-p-1} \hat{\boldsymbol{e}}^\top \hat{\boldsymbol{e}}$$

The expectation of this estimator is:

$$E\left[S^2\right] = E\left[\frac{1}{n-p-1} \hat{\boldsymbol{e}}^\top \hat{\boldsymbol{e}}\right] = \frac{1}{n-p-1} E\left[\hat{\boldsymbol{e}}^\top \hat{\boldsymbol{e}}\right] = \frac{1}{n-p-1} \sigma^2(n-p-1) = \sigma^2$$

Thus, S^2 is an unbiased estimator for σ^2. For simple linear regression where $p=1$:

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Note as n approaches infinity:

$$\lim_{n \to \infty} \frac{n-p-1}{n} \sigma^2 = \sigma^2$$

This means that the bias diminishes as the sample size increases, making $\hat{\sigma}_{ML}^2$ asymptotically unbiased.

Because $\hat{\sigma}_{ML}^2$ converges to the true value σ^2 as n increases, it is a consistent estimator:

$$\hat{\sigma}_{ML}^2 \xrightarrow{p} \sigma^2$$

This means that for any $\epsilon > 0$:

$$\lim_{n \to \infty} P\left(\left|\hat{\sigma}_{ML}^2 - \sigma^2\right| \geq \epsilon\right) = 0$$

6.1.2 Making the inference and confidence interval for simple linear regression

To make inferences about β_0 and β_1, we standardize the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{XX}}} \sim N(0,1)$$

Since σ is usually unknown, we replace it with its estimate S :

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim t_{n-2}$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{S_{XX}}} \sim t_{n-2}$$

To construct a confidence interval for β_0 :

$$\hat{\beta}_0 \pm t_{n-2, \alpha/2} \cdot \operatorname{se}\left(\hat{\beta}_0\right)$$

where $\operatorname{se}\left(\hat{\beta}_0\right)$ is:

$$\operatorname{se}\left(\hat{\beta}_0\right) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

Thus, the $100(1-\alpha)\%$ confidence interval for β_0 is:

$$\left(\hat{\beta}_0 - t_{n-2,\alpha/2} \cdot \operatorname{se}\left(\hat{\beta}_0\right), \hat{\beta}_0 + t_{n-2,\alpha/2} \cdot \operatorname{se}\left(\hat{\beta}_0\right)\right)$$

The value $t_{n-2,\alpha/2}$ is the critical value of the t-distribution such that the area under the t distribution curve to the right of $t_{n-2,\alpha/2}$ is $\alpha/2$.

To construct a confidence interval for β_1 :

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot \operatorname{se}\left(\hat{\beta}_1\right)$$

where $\operatorname{se}\left(\hat{\beta}_1\right)$ is:

$$\operatorname{se}\left(\hat{\beta}_1\right)=\frac{S}{\sqrt{S_{XX}}}$$

Thus, the $100(1-\alpha)\%$ confidence interval for β_1 is:

$$\left(\hat{\beta}_1-t_{n-2,\alpha/2}\cdot\operatorname{se}\left(\hat{\beta}_1\right),\hat{\beta}_1+t_{n-2,\alpha/2}\cdot\operatorname{se}\left(\hat{\beta}_1\right)\right)$$

The value $t_{n-2,\alpha/2}$ is the critical value of the t-distribution such that the area under the t distribution curve to the right of $t_{n-2,\alpha/2}$ is $\alpha/2$.

6.1.3 Confidence interval for multiple linear regression

For multiple linear regression, the formula for the $100(1-\alpha)\%$ confidence interval for a coefficient β_j is:

$$\left(\hat{\beta}_j-t_{n-p-1,\alpha/2}\cdot S\sqrt{(X^TX)^{-1}_{j+1,j+1}},\hat{\beta}_j+t_{n-p-1,\alpha/2}\cdot S\sqrt{(X^TX)^{-1}_{j+1,j+1}}\right)$$

6.1.4 Perform hypothesis testing on coefficients for simple linear regression

To test the hypothesis $H_0:\beta_0=\beta_0^*$:

$$T=\frac{\hat{\beta}_0-\beta_0^*}{\operatorname{se}\left(\hat{\beta}_0\right)}\sim t_{n-2}$$

We compute the p-value:

$$\text{p-value }=2\left(1-F_{n-2}(|T|)\right)$$

where F_{n-2} is the cumulative distribution function (CDF) of the t-distribution with $n-2$ degrees of freedom. We reject H_0 if the p-value is less than the chosen significance level (typically 0.05).

To test the hypothesis $H_0:\beta_1=\beta_1^*$:

$$T=\frac{\hat{\beta}_1-\beta_1^*}{\operatorname{se}\left(\hat{\beta}_1\right)}\sim t_{n-2}$$

We compute the p-value:

$$\text{p-value }=2\left(1-F_{n-2}(|T|)\right)$$

where F_{n-2} is the cumulative distribution function (CDF) of the t-distribution with $n-2$ degrees of freedom. We reject H_0 if the p-value is less than the chosen significance level (typically 0.05).

6.2 Prediction

An important goal of regression analysis is to make predictions about new observations based on the fitted model. After fitting a regression model, we can predict the response for a new sample.

Given a new observation x^*, without its response y^*, we can predict the response using the fitted model:

$$\widehat{y}^*=\left[1,(x^*)^\top\right]\widehat{\beta}$$

If the new sample follows the sample linear regression model, its response shall be

$$y^*=\left[1,(\boldsymbol{x}^*)^\top\right]\beta+\epsilon^*,\quad \epsilon^*\sim N\left(0,\sigma^2\right)$$

From the sampling distribution, we know:

$$\hat{\beta}\sim N_{p+1}\left(\beta,\sigma^2\left(X^TX\right)^{-1}\right)$$

Since $\hat{\beta}$ follows a normal distribution, $\hat{y}^*$ is also normally distributed:

- $E\widehat{y}^*=E\left(\left[1,(x^*)^\top\right]\widehat{\beta}\right)=\left[1,(x^*)^\top\right]E\widehat{\beta}=\left[1,(x^*)^\top\right]\beta$

- $\operatorname{Var}(\widehat{\boldsymbol{y}}^*) = \operatorname{Var}\left(\left[1, (\boldsymbol{x}^*)^\top\right]\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left[1, (\boldsymbol{x}^*)^\top\right]\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\begin{bmatrix} 1 \\ \boldsymbol{x}^* \end{bmatrix}$

Thus, $\widehat{y}^* \sim N\left(\left[1, (\boldsymbol{X}^*)^\top\right]\beta, \sigma^2\left[1, (\boldsymbol{x}^*)^\top\right]\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\begin{bmatrix} 1 \\ \boldsymbol{x}^* \end{bmatrix}\right)$.

6.2.1 Simple linear regression case

In simple linear regression, we have the following:

$$\hat{y}^* \sim N\left(\begin{bmatrix} 1 & x^* \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \sigma^2\begin{bmatrix} 1 & x^* \end{bmatrix}\left(X^TX\right)^{-1}\begin{bmatrix} 1 \\ x^* \end{bmatrix}\right)$$

And we know:

$$\begin{aligned} &[1, x^*]\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\begin{bmatrix} 1 \\ x^* \end{bmatrix} \\ =&[1, x^*]\frac{1}{SXX}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} 1 \\ x^* \end{bmatrix} \\ =&\frac{\frac{1}{n}\sum_{i=1}^n x_i^2 - 2\bar{x}x^* + (x^*)^2}{SXX} \\ =&\frac{\frac{1}{n}SXX + \bar{x}^2 - 2\bar{x}x^* + (x^*)^2}{SXX} \\ =&\frac{1}{n} + \frac{(\bar{x} - x^*)^2}{SXX} \end{aligned}$$

Thus:

$$\hat{y}^* \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}\right)\right)$$

6.2.2 Confidence interval for $\hat{y}^*$

To make inferences about the mean response for the new observation x^*, we construct a confidence interval for $\hat{y}^*$. Standardizing $\hat{y}^*$:

$$\frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}} \sim N(0, 1)$$

Since σ is unknown, we replace it with its estimator S :

$$\frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

Thus, the $100(1 - \alpha)\%$ confidence interval for the mean response $\beta_0 + \beta_1 x^*$ is:

$$\left(\hat{y}^* - t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}, \hat{y}^* + t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}\right)$$

6.2.3 Prediction Interval for Actual Value y^*

A prediction interval accounts for both the variability in the estimation of the mean response and the inherent variability in the actual response y^*.

The actual response y^* for the new observation x^* is:

$$y^* = \beta_0 + \beta_1 x^* + \epsilon^*$$

where $\epsilon^* \sim N\left(0, \sigma^2\right)$. The prediction error $y^* - \hat{y}^*$ has mean 0 and variance:

$$\operatorname{Var}\left(y^* - \hat{y}^*\right) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}\right)$$

And

$$E\left(y^* - \hat{y}^*\right) = \left(\beta_0 + \beta_1 x_i + E\epsilon^*\right) - \left(\beta_0 + \beta_1 x^*\right) = 0$$

Standardizing:

$$\frac{y^* - \hat{y}^*}{\sigma\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}}} \sim N(0,1)$$

Replacing σ with S :

$$\frac{y^* - \hat{y}^*}{S\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}}} \sim t_{n-2}$$

Thus, the $100(1-\alpha)\%$ prediction interval for the actual response y^* is:

$$\left(\hat{y}^* - t_{n-2,\alpha/2} S\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}}, \hat{y}^* + t_{n-2,\alpha/2} S\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{XX}}}\right)$$

7 Multi-collinearity and penalized regression

7.1 Multi-collinearity

Definition 7.1: Multip-collineairty

Multi-collinearity is a situation in multiple linear regression models where two or more predictor variables (independent variables) are highly correlated. This means that one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.

Example.

We have the following extreme example of perfect collinearity. Consider an extreme case where two covariates $x_{i,1}$ and $x_{i,2}$ are perfectly linearly related. This means $x_{i,1} = \lambda x_{i,2}$ for some constant $\lambda \neq 0$. In this scenario, the regression equation:

$$\beta_1 x_{i,1} + \beta_2 x_{i,2}$$

can be rewritten in various forms due to the perfect linear relationship. For example:

$$\begin{aligned}\beta_1 x_{i,1} + \beta_2 x_{i,2} &= 0 \cdot x_{i,1} + (\beta_1 \lambda + \beta_2)\, x_{i,2} \\ &= \left(\beta_1 + \frac{\beta_2}{\lambda}\right) x_{i,1} + 0 \cdot x_{i,2} \\ &= (\beta_1 + 100)\, x_{i,1} + (\beta_2 - 100\lambda)\, x_{i,2}\end{aligned}$$

and so on.
Consider the linear combination of the column vectors $\boldsymbol{X}_{\cdot,2}$ and $\boldsymbol{X}_{\cdot,3}$. Let:

$$\boldsymbol{v} = \beta_1 \boldsymbol{X}_{\cdot,2} + \beta_2 \boldsymbol{X}_{\cdot,3}$$

Since $\boldsymbol{X}_{\cdot,2}$ and $\boldsymbol{X}_{\cdot,3}$ are not linearly independent, we can express $\boldsymbol{X}_{\cdot,2}$ in terms of $\boldsymbol{X}_{\cdot,3}$:

$$\boldsymbol{X}_{\cdot,2} = \lambda \boldsymbol{X}_{\cdot,3}$$

Therefore, the linear combination becomes:

$$\begin{aligned}\boldsymbol{v} &= \beta_1 (\lambda \boldsymbol{X}_{\cdot,3}) + \beta_2 \boldsymbol{X}_{\cdot,3} \\ \boldsymbol{v} &= (\beta_1 \lambda + \beta_2)\, \boldsymbol{X}_{\cdot,3}\end{aligned}$$

This shows that the linear combination can be written in different forms depending on the values of β_1 and β_2.
The linear combination $\boldsymbol{v}$ is not unique because there are infinitely many pairs (β_1, β_2) that can produce the same vector $\boldsymbol{v}$. For instance, if we choose different values for β_1 and β_2, we can still achieve the same result by appropriately adjusting the coefficients. To illustrate this further, consider the following example:

$$\boldsymbol{v} = \beta_1 (\lambda \boldsymbol{X}_{\cdot,3}) + \beta_2 \boldsymbol{X}_{\cdot,3} = (\beta_1 \lambda + \beta_2)\, \boldsymbol{X}_{\cdot,3}$$

If $\beta_1 = 1$ and $\beta_2 = -\lambda$, we have:

$$\boldsymbol{v} = (1 \cdot \lambda - \lambda)\boldsymbol{X}_{\cdot,3} = 0 \cdot \boldsymbol{X}_{\cdot,3}$$

Similarly, if $\beta_1 = 0$ and $\beta_2 = 1$, we have:

$$\boldsymbol{v} = (0 \cdot \lambda + 1)\boldsymbol{X}_{\cdot,3} = 1 \cdot \boldsymbol{X}_{\cdot,3}$$

This shows that the linear combination of the column vectors $\boldsymbol{X}_{\cdot,2}$ and $\boldsymbol{X}_{:,3}$ is not unique. And as a result, its design matrix does not has full column rank because its columns are not linearly independent.
Mathematically, when the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$ does not have full column rank (i.e., there is perfect multi-collinearity), the matrix $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)$ is not invertible. This means:

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

does not exist. Consequently, the least squares estimator:

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

is not defined.

Example.

> Suppose we have a simple linear regression model with perfect collinearity:
>
> $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$
> $$\text{where } x_{i,1} = \lambda x_{i,2}.$$
>
> We can rewrite the regression equation as:
>
> $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 \left(\frac{x_{i,1}}{\lambda}\right) + \epsilon_i$$
> $$y_i = \beta_0 + \left(\beta_1 + \frac{\beta_2}{\lambda}\right) x_{i,1} + \epsilon_i$$
>
> Let:
>
> $$\beta_1' = \beta_1 + \frac{\beta_2}{\lambda}$$
>
> The regression equation becomes:
>
> $$y_i = \beta_0 + \beta_1' x_{i,1} + \epsilon_i$$
>
> Here, we see that any combination of β_1 and β_2 that satisfies $\beta_1 + \frac{\beta_2}{\lambda} = \beta_1'$ will result in the same fit. Therefore, there are infinitely many combinations of β_1 and β_2 that satisfy the regression equation.

Fact 7.2

In perfect collinearity, least square estimators still exit, but there will be infinitely many.

Fact 7.3

In multi-collinearity:

1. The design matrix $\boldsymbol{X}$ consists of the predictor variables in a regression model. When there is collinearity, one or more columns of $\boldsymbol{X}$ are nearly linear combinations of other columns. This nearlinear dependence reduces the effective rank of $\boldsymbol{X}$, making it nearly rank-deficient. When $\boldsymbol{X}$ has collinear columns, $\boldsymbol{X}^\top\boldsymbol{X}$ becomes nearly singular. This means that the determinant $\det\left(\boldsymbol{X}^\top\boldsymbol{X}\right)$ is very close to zero.

2. The inversion process $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$ is highly sensitive to small changes in the elements of $\boldsymbol{X}$ when $\det\left(\boldsymbol{X}^\top\boldsymbol{X}\right)$ is near zero.

3. This sensitivity can lead to large variations in the estimated coefficients $\widehat{\boldsymbol{\beta}}$, including changes in sign and magnitude.

7.2 Detecting Multip-collineairty

Pair-wise collinearity refers to the situation where two variables are highly correlated with each other. We can identify pair-wise collinearity using pair-wise covariate plots (scatter plots) or by calculating the correlation coefficient between each pair of predictor variables.

However, pair-wise collinearity does not capture more complex relationships involving three or more variables. For more general detection of multi-collinearity, we need to look at how one covariate can be predicted by the others. Consider the following regression equation:

$$x_{i,p} = \gamma_0 + \gamma_1 x_{i,1} + \ldots + \gamma_{p-1} x_{i,p-1} + \delta_i$$

Here:

- $x_{i,p}$ is the i-th observation of the p-th predictor.
- γ_0 is the intercept.
- $\gamma_1, \gamma_2, \ldots, \gamma_{p-1}$ are the coefficients for the other predictors.
- δ_i is the error term.

By fitting this regression model, we can determine how much of the variance in $x_{i,p}$ is explained by the other predictors.

7.3 Penalized regression

Definition 7.4: Penalized regression

A penalized method for regression involves adding a penalty term to the regression objective function to constrain the magnitude of the regression coefficients. This technique helps address issues involving multi-collinearity.

Definition 7.5: Ridge regression

Ridge regression is a penalized method that addresses multi-collinearity by adding a penalty term to the objective function. This penalty term is proportional to the sum of the squares of the coefficients. In multiple linear regression, the objective function for ridge regression is:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 + \lambda \sum_{j=0}^{p} \beta_j^2$$

Or more compactly:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \|\beta\|^2$$

Where:

- $\boldsymbol{y}$ is the vector of observed values.
- $\boldsymbol{X}$ is the design matrix of predictor variables.
- β is the vector of coefficients.
- λ is a tuning parameter that controls the strength of the penalty.

Rescaling can lead to easier analysis and more stable computation:

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \|\beta\|^2$$

The ridge regression objective function is:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2$$

To find the ridge regression estimator, we need to minimize this objective function. We do this by setting the derivative of the objective function with respect to β to zero. From lecture 2, we know that:

$$\frac{\partial \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2}{\partial \beta} = -2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = 2\boldsymbol{X}^\top\boldsymbol{X\beta} - 2\boldsymbol{X}^\top\boldsymbol{y}$$

For the penalty term:

$$\frac{\partial \lambda \|\beta\|^2}{\partial \beta} = \frac{\partial \lambda \beta^\top \beta}{\partial \beta} = 2\lambda\beta$$

Setting the derivative of the objective function to zero:

$$2\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{ridge}} - 2\boldsymbol{X}^\top\boldsymbol{y} + 2\lambda\hat{\boldsymbol{\beta}}^{\text{ridge}} = 0$$

Solving for $\hat{\beta}^{\text{ridge}}$:

$$\left(\boldsymbol{X}^{\top}\boldsymbol{X}+\lambda\boldsymbol{I}\right)\hat{\beta}^{\text{ridge}} = \boldsymbol{X}^{\top}\boldsymbol{y}$$
$$\hat{\beta}^{\text{ridge}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

- Even if $\boldsymbol{X}^{\top}\boldsymbol{X}$ is not invertible, $\boldsymbol{X}^{\top}\boldsymbol{X}+\lambda\boldsymbol{I}$ will be invertible for any $\lambda > 0$.
- λ can be chosen by cross-validation to optimize the model performance.

Definition 7.6: LASSO regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression is another penalized regression method. The objective function for LASSO is:

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\beta\|^2+\lambda\|\boldsymbol{\beta}\|_1$$

Where $\|\boldsymbol{\beta}\|_1$ is the sum of the absolute values of the coefficients. LASSO can perform variable selection by shrinking some coefficients exactly to zero.

Definition 7.7: Elastic net regression

Elastic Net regression combines the penalties of both Ridge and LASSO regression. The objective function is:

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|^2$$

Where:

- λ_1 controls the LASSO penalty.
- λ_2 controls the Ridge penalty.

8 Decomposition of variance

8.1 Sums of squares

In regression analysis, the sum of squares is a measure of the variability of the data. It helps us understand how much of the variability in the dependent variable y can be explained by the independent variable $\boldsymbol{x}$.

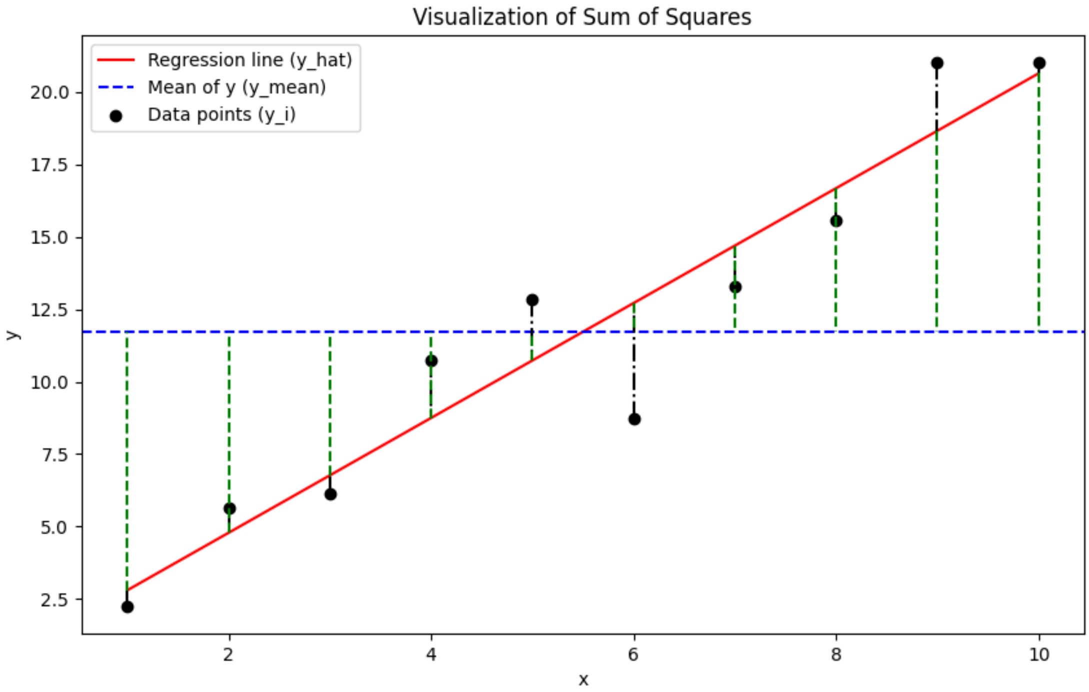


Figure 34: Sum of squares visualization

From this example, we have:

- **45-degree line** $(y = x)$ **:** This line represents a perfect correlation where each value of y is exactly equal to its corresponding x. In the context of simple linear regression, this line can be considered as the ideal regression line if the slope were 1 and the intercept were 0 .
- **Horizontal line** $(y = \bar{y})$ **:** This line represents the mean of y. It is a reference line showing the average value of the dependent variable.
- **Black + green dash line:** This line represents the deviation of each observed value y_i from the mean of $y, \bar{y}$. This deviation is calculated without considering the predictor variables $x_1, \ldots, x_n$.
- **Black dash line:** This is the residual $\hat{e}_i = y_i - \hat{y}_i$, which is the difference between the observed value y_i and the predicted value $\hat{y}_i$ obtained from the regression model.
- **Green dash line:** This line represents the difference between the predicted value $\hat{y}_i$ and the mean $\bar{y}$. This is the portion of the variability in y that is explained by the predictors $x_1, \ldots, x_n$

Definition 8.1: TSS, RSS, SSreg

Total Sum of Squares (TSS): $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$. This measures the total variability in the dependent variable y before fitting the model.
Residual Sum of Squares (RSS): $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. This measures the variability in y that is not explained by the model (i.e., the leftover or unexplained variation).
Regression Sum of Squares (SSreg): $\text{SSreg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$. This measures the variability in y that is explained by the model.

For TSS, we can express it in the matrix notation. Let's define $\boldsymbol{y}$ as the vector of observed values:

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\bar{y}$ as the mean of the observed values, which can be written as:

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

A vector of means, $\overline{\boldsymbol{y}}$, where each element is $\bar{y}$:

$$\overline{\boldsymbol{y}}=\begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}$$

Define the matrix $\boldsymbol{J}$ as:

$$J=11^{\top} \in \mathbb{R}^{n\times n}$$

where $\mathbf{1}$ is a column vector of ones:

$$1=\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The vector of means $\overline{\boldsymbol{y}}$ can be written as:

$$\bar{y}=\frac{1}{n}\boldsymbol{J}y$$

Now, TSS can be expressed as:

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

Substituting $\overline{\boldsymbol{y}}$ with $\frac{1}{n}\boldsymbol{J}\boldsymbol{y}$:

$$TSS=\left\|y-\frac{1}{n}Jy\right\|^2$$

This can be rewritten using the matrix subtraction:

$$TSS=\left\|\left(I-\frac{1}{n}J\right)y\right\|^2$$

where $\boldsymbol{I}$ is the identity matrix of size $n\times n$.
The norm squared of a vector $\boldsymbol{a}$ is equal to the inner product $\boldsymbol{a}^{\top}\boldsymbol{a}$:

$$\left\|\left(I-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}\right\|^2=\left(\left(I-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}\right)^{\top}\left(\left(I-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}\right)$$

Simplifying further:

$$TSS=y^{\top}\left(I-\frac{1}{n}J\right)^{\top}\left(I-\frac{1}{n}J\right)y$$

Since $\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}$ is symmetric and idempotent:

$$\left(I-\frac{1}{n}J\right)^{\top}\left(I-\frac{1}{n}J\right)=I-\frac{1}{n}J-\frac{1}{n}J+\frac{1}{n^2}J^2=I-\frac{1}{n}J$$

Therefore:

$$TSS = y^\top \left(I - \frac{1}{n}J\right) y$$

For RSS, we can represent it using matrix representation as well. Let's define $\boldsymbol{y}$ as the vector of observed values:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\hat{\boldsymbol{y}}$ as the vector of predicted values:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

The predicted values $\hat{\boldsymbol{y}}$ can be expressed using the hat matrix $\boldsymbol{H}$:

$$\hat{y} = Hy$$

where $\boldsymbol{H}$ is defined as:

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$$

The residuals are the differences between the observed and predicted values:

$$e = y - \hat{y} = y - Hy$$

RSS can be expressed as:

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

Substituting $\hat{\boldsymbol{y}}$ with $\boldsymbol{Hy}$:

$$RSS = (\boldsymbol{y} - \boldsymbol{Hy})^\top (\boldsymbol{y} - \boldsymbol{Hy})$$

Simplifying the expression:

$$RSS = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$
$$RSS = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

The matrix $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent:

$$(\boldsymbol{I} - \boldsymbol{H})^\top (\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})$$

Therefore:

$$RSS = \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

For SSreg, we can deduce its matrix representation as well. SSreg can be expressed as:

$$SSreg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \|\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2$$

Substituting $\hat{\boldsymbol{y}}$ with $\boldsymbol{Hy}$ and $\overline{\boldsymbol{y}}$ with $\frac{1}{n}\boldsymbol{Jy}$:

$$\text{SSreg } = \left\| Hy - \frac{1}{n}Jy \right\|^2$$

This can be rewritten as:

$$\text{SSreg } = \left\| \left(H - \frac{1}{n}J\right) y \right\|^2$$

The norm squared of a vector $\boldsymbol{a}$ is equal to the inner product $\boldsymbol{a}^{\top}\boldsymbol{a}$:

$$\text{SSreg } = \left(\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}\right)^{\top}\left(\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}\right)$$

Simplifying further:

$$\text{SSreg } = \boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$$

Since $\boldsymbol{H}$ and $\boldsymbol{J}$ are symmetric, we have:

$$\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)^{\top}=\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}$$

Therefore:

$$\text{SSreg } = \boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$$

Since $\boldsymbol{H}$ and $\frac{1}{n}\boldsymbol{J}$ are both idempotent, we simplify to:

$$\text{SSreg } = \boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}+\frac{1}{n^2}\boldsymbol{J}\right)\boldsymbol{y}$$

Given that $\boldsymbol{H}$ and $\boldsymbol{J}$ are symmetric and $\boldsymbol{J}^2=n\boldsymbol{J}$, we simplify further to:

$$\text{SSreg } = \boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$$

Therefore, we have the following matrix notations:

- $TSS=\boldsymbol{y}^{\top}\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$
- $RSS=\boldsymbol{y}^{\top}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$
- $SSreg=\boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$

Fact 8.2

$TSS=SSreg+RSS$

Proof. Add SSreg and RSS:

$$\text{SSreg } + \text{ RSS } = \boldsymbol{y}^{\top}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}+\boldsymbol{y}^{\top}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$$

Combine the terms inside the quadratic forms:

$$\text{SSreg } + \text{ RSS } = \boldsymbol{y}^{\top}\left[\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)+(\boldsymbol{I}-\boldsymbol{H})\right]\boldsymbol{y}$$

Simplify the expression inside the brackets:

$$\text{SSreg } + \text{ RSS } = \boldsymbol{y}^{\top}\left[\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}+\boldsymbol{I}-\boldsymbol{H}\right]\boldsymbol{y}$$

Notice that $\boldsymbol{H}-\boldsymbol{H}=0$, so this reduces to:

$$\text{SSreg } + \text{ RSS } = \boldsymbol{y}^{\top}\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$$

This is exactly the expression for TSS :

$$TSS=y^{\top}\left(I-\frac{1}{n}J\right)y$$

Thus, we have shown that:

$$TSS= \text{ SSreg } + \text{ RSS}$$

■

Proof. Without using matrix notation, we can also construct a similar proof with sigma notation. Given a linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n} ((\hat{y}_i - \bar{y}) + \underbrace{(y_i - \hat{y}_i)}_{\hat{\varepsilon}_i})^2 \\
&= \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y})^2 + 2\hat{\varepsilon}_i (\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i^2 \right) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^{n} \hat{\varepsilon}_i (\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^{n} \hat{\varepsilon}_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} - \bar{y} \right) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \left(\hat{\beta}_0 - \bar{y} \right) \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i}_{0} + 2\hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i x_{i1}}_{0} + \cdots + 2\hat{\beta}_p \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i x_{ip}}_{0} \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \text{ESS} + \text{RSS}
\end{aligned}$$

■

Degrees of freedom (df) refer to value of independent variable minus the number of parameters or relationships. Thus:

- Total Sum of Squares (TSS) has $n - 1$ degrees of freedom. This is becuase we have n number of indepdent variable and only 1 parameter $\overline{y}$.
- Residual Sum of Squares (RSS) has $n-p-1$ degrees of freedom, where p is the number of predictors. This is becuase in this case we have n number of indepdent variable, and there are $p+1$ number of predictors or parameters to calculate $\hat{y}$.
- Regression Sum of Squares (SSreg) has p degrees of freedom. This is because we have n number of independent variable, and p

8.2 Analysis of variance (ANOVA)

Given a set of data points $(x_i, y_i)_{i=1}^{n}$, we want to determine whether there is a linear relationship between the variables X and Y. This can be approached using two methods:

- **Solution A: Hypothesis Testing**

  We test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_A : \beta_1 \neq 0$.

  - *Null Hypothesis H_0 :* There is no linear relationship between X and Y $(\beta_1 = 0)$.
  - *Alternative Hypothesis H_A :* There is a linear relationship between X and Y $(\beta_1 \neq 0)$.

  If we reject the null hypothesis, it suggests that there is a significant linear relationship between X and Y.

- **Solution B: Compare Two Models**

  We compare two models to assess whether including the predictor X improves the fit of the model:

  - Model 1: $y_i = \beta_0 + \epsilon_i$
    This is a simple model where the response y_i is modeled as the mean β_0 plus random error ϵ_i.
  - Model 2: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
    This model includes the predictor x_i, indicating a potential linear relationship between X and Y.

ANOVA, or Analysis of Variance, is a statistical method used to compare the means of different groups to see if there are any statistically significant differences between them.

Here, ANOVA is used to compare the fit of different regression models. Specifically, it helps us compare a simpler model (e.g., a model without predictors) to a more complex model (e.g., a model with one or more predictors) to determine if the added complexity (additional predictors) significantly improves the model's explanatory power.

To implement ANOVA for comparing these models, we follow these steps:

1. Estimate the parameters of both Model 1 and Model 2 using the data:
   - For Model 1: $RSS_1 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(1)}\right)^2$
   - For Model 2: $RSS_2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(2)}\right)^2$
2. We can show that always $RSS_1 \geq RSS_2$

   Proof. The null model does not include any predictors and simply models the response variable y as the mean of all observed values.

   $$y_i = \beta_0 + \epsilon_i$$

   Here, β_0 is estimated as the mean of $y, \bar{y}$. So the fitted values for this model are:

   $$\hat{y}_i^{(1)} = \bar{y}$$

   TSS measures the total variability in the response variable y around its mean.

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

   For the null model, the residuals are the differences between the observed values y_i and the mean $\bar{y}$:

   $$RSS_1 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(1)}\right)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

   Since $\hat{y}_i^{(1)} = \bar{y}$, we see that:

   $$RSS_1 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

   Thus, for the null model:

   $$TSS = RSS_1$$

   The full model includes the predictor x_i and models the response variable y as a linear function of x_i :

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

   Here, β_0 and β_1 are estimated using the least squares method, resulting in the fitted values:

   $$\hat{y}_i^{(2)} = \beta_0 + \beta_1 x_i$$

   As before, TSS measures the total variability in y :

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

   For the full model, the residuals are the differences between the observed values y_i and the predicted values $\hat{y}_i^{(2)}$:

   $$RSS_2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(2)}\right)^2$$

SSreg measures the variability explained by the predictor x_i :

$$\text{SSreg } =\sum_{i=1}^{n}\left(\hat{y}_i^{(2)}-\bar{y}\right)^2$$

TSS can be decomposed into the SSreg and the residual variability RSS_2:

$$TSS = SSreg + RSS_2$$

Rewriting this, we get:

$$RS_1 = SSreg + RSS_2$$

Since $SSreg \geq 0$, it follows that:

$$RSS_1 \geq RSS_2$$

■

If $RSS_1 \gg RSS_2$, model 2 is effective in capturing variation unexplained by Model 1 , indicating that the predictor x_i is important.

If RSS_1 and RSS_2 are Very Close, model 2 introduces an unimportant parameter β_1 that does not significantly improve the model fit.

To quantify the improvement:

$$SSreg = RSS_1 - RSS_2 = TSS - RSS$$

8.3 F-test

The F test is used to determine if there is a significant linear relationship between the dependent variable and the independent variables in a regression model. The intuition is based on the following points:

- **Significant Linear Relationship:** There is a significant linear relationship if a significant amount of variation in the response variable y is explained by the model.
- **SSreg Relative to TSS:** If the regression sum of squares (SSreg) is large relative to the total sum of squares (TSS), it indicates that the model explains a substantial portion of the variability in y.
- **RSS Relative to TSS:** Conversely, if the residual sum of squares (RSS) is small relative to TSS, it suggests that the unexplained variability is low, and the model fits well.
- **Overall Comparison:** To check the model's effectiveness, we compare SSreg to RSS.

Definition 8.3: MSreg and MSR

Mean Square for Regression (MSreg):

$$\text{MSreg } = \frac{\text{SSreg}}{p} = \frac{\text{SSreg}}{1} = \text{ SSreg}$$

Mean Square for Residuals (MSR):

$$\text{MSR} = \frac{\text{RSS}}{n-p-1} = \frac{\text{RSS}}{n-2}$$

8.3.1 Simple linear regression

In a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where ϵ_i are independent and identically distributed normal random variables with mean 0 and variance σ^2.

We want to test:

$$H_0 : \beta_1 = 0$$

against the alternative:

$$H_A : \beta_1 \neq 0$$

Definition 8.4: F statistics

The F-statistic is the ratio of the mean square for regression to the mean square for residuals:

$$F=\frac{\mathrm{MSreg}}{\mathrm{MSR}}=\frac{\mathrm{SSreg}/1}{\mathrm{RSS}/(n-2)}$$

In simple linear regression, the estimator $\hat{\beta}_1$ is given by:

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

Under the null hypothesis $H_0: \beta_1=0$:

$$y_i=\beta_0+\epsilon_i$$

The fitted values $\hat{y}_i$ can be written as:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators.
When $\beta_1=0$, the variability in $\hat{y}_i$ is solely due to $\hat{\beta}_1 x_i$. Since $\hat{\beta}_1$ is a linear combination of normal random variables ϵ_i, it follows a normal distribution:

$$\hat{\beta}_1 \sim N\left(0, \sigma^2 / S_{xx}\right)$$

where $S_{xx}=\sum_{i=1}^n\left(x_i-\bar{x}\right)^2$.
Now, consider the definition of SSreg:

$$\mathrm{SSreg}=\sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2=\sum_{i=1}^n\left(\hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y}\right)^2$$

Because $\bar{y}$ is the mean of y_i, and under $H_0, \hat{\beta}_0=\bar{y}$, we get:

$$\text{SSreg }=\sum_{i=1}^n\left(\hat{\beta}_1 x_i\right)^2$$

To simplify, rewrite SSreg:

$$\text{SSreg }=\hat{\beta}_1^2 \sum_{i=1}^n x_i^2$$

Given that $\hat{\beta}_1 \sim N\left(0, \sigma^2 / S_{xx}\right)$, we have:

$$\frac{\hat{\beta}_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$

Thus:

$$\left(\frac{\hat{\beta}_1}{\sqrt{\sigma^2 / S_{xx}}}\right)^2 \sim \chi^2(1)$$

Since SSreg is a multiple of $\hat{\beta}_1^2$, we can write:

$$\text{SSreg }=\hat{\beta}_1^2 S_{xx}$$

And thus:

$$\frac{\text{SSreg}}{\sigma^2}=\frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2}=\left(\frac{\hat{\beta}_1}{\sqrt{\sigma^2 / S_{xx}}}\right)^2$$

Since $\left(\frac{\hat{\beta}_1}{\sqrt{\sigma^2 / S_{xx}}}\right)^2 \sim \chi^2(1)$, it follows that:

$$\frac{\text{SSreg}}{\sigma^2} \sim \chi^2(1)$$

Therefore, SSreg follows a chi-square distribution with 1 degree of freedom when $\beta_1 = 0$.

Under the null hypothesis $H_0 : \beta_1 = 0$, the model simplifies to:

$$y_i = \beta_0 + \epsilon_i$$

These residuals e_i are still normally distributed with mean 0 , because ϵ_i are normally distributed with mean 0 :

$$e_i \sim N\left(0, \sigma^2\right)$$

The degrees of freedom for the RSS is the number of observations n minus the number of estimated parameters. In simple linear regression, we estimate two parameters: β_0 and β_1. Therefore, the degrees of freedom for the residuals is:

$$\mathrm{df} = n - 2$$

Given that the residuals e_i are normally distributed, the sum of the squared residuals, normalized by the variance σ^2, follows a chi-square distribution:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} e_i^2 \sim \chi^2(n-2)$$

Therefore:

$$\frac{\mathrm{RSS}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \sim \chi^2(n-2)$$

This shows that RSS is a scaled chi-square distribution:

$$\mathrm{RSS} \sim \sigma^2 \chi^2(n-2)$$

Because the F-statistic is then given by:

$$F = \frac{\mathrm{SSreg}/1}{\mathrm{RSS}/(n-2)}$$

Under the null hypothesis and the definition of the F distribution, this follows an F-distribution with degrees of freedom $(1, n-2)$:

$$F \sim F(1, n-2)$$

To determine if we reject the null hypothesis:

- p-value Approach: If the p-value $< \alpha$, reject H_0.
- F-statistic Approach: If $F > F_{1-\alpha}(1, n-2)$, reject H_0.

Rejecting H_0 provides evidence that a statistically significant linear relationship exists.

The ANOVA table for simple linear regression summarizes the sources of variation in the response variable y. It breaks down the total variability into components attributable to the regression model and residual (error) variability.

Source	df	SS	MS	F	p-value
Regression	1	SSreg	MSreg	F_0	$P\left(F_{1,n-2} > F_0\right)$
Residual	$\mathrm{n}-2$	RSS	MSR		
Total	$\mathrm{n}-1$	TSS			

The t -test evaluates the null hypothesis $H_0 : \beta_1 = 0$ using the test statistic:

$$t_1 = \frac{\hat{\beta}_1}{\mathrm{SE}\left(\hat{\beta}_1\right)}$$

where $\hat{\beta}_1$ is the estimated slope and $\mathrm{SE}\left(\hat{\beta}_1\right)$ is the standard error of $\hat{\beta}_1$. The t-test follows a t-distribution with $n-2$ degrees of freedom under H_0. For simple linear regression, the relationship between the F-statistic and the t-statistic can be derived as follows:

- $t_1 = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$

  The standard error $\mathrm{SE}\left(\hat{\beta}_1\right)$ is:

$$\mathrm{SE}\left(\hat{\beta}_1\right) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

  where $\sigma^2 = \frac{\mathrm{RSS}}{n-2}$.

$$t_1 = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{RSS}}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}}} = \frac{\hat{\beta}_1\sqrt{(n-2)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}{\sqrt{\mathrm{RSS}}}$$

- $F = \frac{\mathrm{MSreg}}{\mathrm{MSR}} = \frac{\mathrm{SSreg}/1}{\mathrm{RSS}/(n-2)} = \frac{\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\mathrm{RSS}\,/(n-2)} = \frac{\left(\hat{\beta}_1\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2}{\frac{\mathrm{RSS}}{n-2}} = t_1^2$

Therefore, the F-test statistic in the ANOVA table is equivalent to the square of the t-test statistic for the slope parameter β_1 in the coefficient table, and their p-value is the same:

$$F = t_1^2$$

8.3.2 Multiple linear regression

For multiple linear regression, the F test is used to test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_A : \text{not all } \beta_k = 0 (k \geq 1)$$

Definition 8.5: F-statistics

The F test statistic is defined as:

$$F = \frac{\mathrm{MSreg}}{\mathrm{MSR}} = \frac{\mathrm{SSreg}/p}{\mathrm{RSS}\,/(n-p-1)} \sim F_{p,n-p-1}$$

The proof of the F statistics following this F distribution is omitted because it is not the scope of this course.

Under H_0, the F statistic follows an F distribution with degrees of freedom $(p, n-p-1)$. To determine if we reject the null hypothesis:

- p-value Approach: If the p -value $< \alpha$, reject H_0.
- F-statistic Approach: If $F > F_{1-\alpha}(p, n-p-1)$, reject H_0.

Rejecting H_0 indicates that there is evidence of a statistically significant linear relationship.

The ANOVA table for multiple linear regression:

Source	df	SS	MS	F	p-value
Regression	p	SSreg	MSreg	F_0	$P\left(F_{p,n-p-1} > F_0\right)$
Residual	$n-p-1$	RSS	MSR		
Total	$n-1$	TSS			

T Test vs F Test:

- T-test: Used to test if a single coefficient $\beta_j = 0$.
- F-test: Used to test if all coefficients $\beta_1 = \beta_2 = \ldots = \beta_p = 0$.

When predictors are highly correlated (multicollinearity), the individual T-tests may not correctly identify significant predictors. Multicollinearity inflates the standard errors of the coefficients, leading to less reliable T-test results. F-tests accounts for the joint effect of all predictors, making it more robust in the presence of multicollinearity. It evaluates the overall significance of the regression model, even if individual coefficients are not significant due to multicollinearity.

8.4 Partial F test

Consider the following model with increasing complexity:

1. Model 1: $y_i = \beta_0 + \epsilon_i$

   This is a simple model with only an intercept β_0 and error term ϵ_i.

2. Model 2: $y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$

   This model includes one predictor variable $x_{i,1}$.

3. Model 3: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$

   This model includes two predictor variables $x_{i,1}$ and $x_{i,2}$.

4. Model 4: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$

   This model includes three predictor variables $x_{i,1}, x_{i,2}$, and $x_{i,3}$.

5. Model 5: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i$

   This model includes four predictor variables $x_{i,1}, x_{i,2}, x_{i,3}$, and $x_{i,4}$.

When comparing models, particularly to see if adding more predictors significantly improves the model, we perform hypothesis testing:

- Null Hypothesis (H_0) : The additional predictors have no effect $(\beta_2 = \beta_3 = \beta_4 = 0$).
- Alternative Hypothesis (H_a) : At least one of the additional predictors has a significant effect ($\beta_2, \beta_3, \beta_4 \neq 0)$.

Different statistical tests are used for different purposes:

- **F Test (ANOVA Test):** Tests if at least one of the predictors is linearly related to the response variable Y.
- **T Test:** Tests if a single predictor X_j is linearly related to Y.
- **Partial F Test:** Tests if a subset of predictors (e.g., X_2, X_3, X_4) is linearly related to Y.

For a full model: $E(Y \mid X) = \boldsymbol{X}\boldsymbol{\beta}$, It uses all p predictors. For a reduced model: $E(Y \mid X) = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$, it uses $p-k$ predictors (removing k predictors). Because both models use the same dataset, leading to the same Total Sum of Squares (TSS).

Generally, $\mathrm{RSS}_{\text{reduced}} \geq \mathrm{RSS}_{\text{full}}$. This is becuase adding more predictors (as in the full model) generally leads to a better fit to the data, reducing the RSS. Removing predictors (as in the reduced model) typically worsens the fit, increasing the RSS.

Generally, $\mathrm{SSreg}_{\text{reduced}} \leq \mathrm{SSreg}_{\text{full}}$. This follows directly from $\mathrm{RSS}_{\text{reduced}} \geq \mathrm{RSS}_{\text{full}}$ with the same TSS.

We have:

- For the full model RSS: $n-p-1$
- For the reduced model RSS: $n-(p-k)-1$
- For the difference: $k = (n-(p-k)-1) - (n-p-1)$

Without loss of generality, we want to test $H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$, vs H_a : at lest one of $\beta_1, \ldots \beta_k$ is non-zero. We can replace $\beta_1, \beta_2, \ldots \beta_k$ with any subset of $\{\beta_1 \ldots \beta_p\}$ with cardinality k.

Definition 8.6: Partial F-test

The test statistic is defined as:

$$F^* = \frac{\left(\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}}\right)/k}{\text{RSS}_{\text{full}}/(n-p-1)}$$

Under H_0 :

$$F^* \sim F(k, n-p-1)$$

Example.

> Consider the following two models:
>
> - **Model 2:** $y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$
> - **Model 5:** $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i$
>
> We have:
>
> - **Null Hypothesis** (H_0) **:** Model 2 (simpler) is correct.
> - **Alternative Hypothesis** (H_a) **:** Model 5 (more complex) is correct.
>
> By definition, we have the test statistics $F^* = \frac{(\text{RSS}_2 - \text{RSS}_5)/3}{\text{RSS}_5/(n-5)}$, and p-value $P\left(F_{3,n-5} \geq F^*\right)$

We have the following rejection criteria:

- If $F^* \geq F_{1-\alpha}(k, n-p-1)$ or p -value $< \alpha$, reject H_0. This indicates a significant linear relationship between Y and at least one of the k predictors, favoring the full model.
- If $F^* < F_{1-\alpha}(k, n-p-1)$ or p-value $> \alpha$, fail to reject H_0. This indicates no significant linear relationship, favoring the reduced model.

8.5 Coefficient of determination (R^2)

In regression analysis, the total sum of squares (TSS) can be decomposed into the regression sum of squares (SSreg) and the residual sum of squares (RSS):

$$\text{TSS} = \text{SSreg} + \text{RSS}$$

Definition 8.7: Coefficient of determination

The coefficient of determination R^2 quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

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

This value lies in the interval $[0, 1]$. An R^2 close to 1 indicates that a large proportion of the variance in the dependent variable is accounted for by the model.

8.5.1 R^2 in simple linear regression

- $R^2 = 1$: All observations fall perfectly on the fitted regression line, meaning RSS $= 0$ and X accounts for all variation in Y.
- $\widehat{\beta}_1 = 0$ then $R^2 = 0$: there is no linear association between X and Y in the sample data, or the covariates X is of no help in reducing the variation in Y.
- Closer R^2 to 1: Indicates a stronger linear association between X and Y.

8.5.2 R^2 in multiple linear regression

- $R^2 = 1$: All observations fall perfectly on the fitted regression hyperplane, meaning RSS $= 0$ and $X_1, \ldots, X_p$ account for all variation in Y.
- When $\widehat{\beta}_1 = \widehat{\beta}_2 = \ldots = \widehat{\beta}_p = 0$, and then $R^2 = 0$: The predictors $X_1, \ldots, X_p$ do not reduce the variation in Y; there is no linear relationship.
- Closer R^2 to 1: Indicates a stronger linear association between the predictors $X_1, \ldots, X_p$ and Y.

Definition 8.8: Coefficient of correlation

In simple linear regression, the coefficient of correlation r is defined as:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

It represents the sample correlation between X and Y when both are treated as random variables. The relationship between r and R^2 is given by:

$$r = \operatorname{sign}\left(\widehat{\beta}_1\right)\sqrt{R^2}$$

Remark.

> In simple linear regression, $r^2 = R^2$.

Proof. Consider the simple linear regression model:

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

The estimated regression line is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The least squares estimate for β_1 is:

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

$$\text{SSreg } = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Substitute $\hat{y}_i$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\text{SSreg } = \sum_{i=1}^{n} \left(\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) - \bar{y}\right)^2$$

Since $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$:

$$\begin{aligned}
\text{SSreg} &= \sum_{i=1}^{n} \left(\left(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \right) - \bar{y} \right)^2 \\
\text{SSreg} &= \sum_{i=1}^{n} \left(\hat{\beta}_1 \left(x_i - \bar{x} \right) \right)^2 \\
\text{SSreg} &= \hat{\beta}_1^2 \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \\
\text{SSreg} &= \left(\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} \right)^2 \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \\
\text{SSreg} &= \frac{\left(\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}
\end{aligned}$$

$$\begin{aligned}
R^2 &= \frac{\text{SSreg}}{\text{TSS}} = \frac{\frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \\
R^2 &= \frac{\left(\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right)^2}{\left(\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right) \left(\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 \right)} \\
r^2 &= \left(\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} \sqrt{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}} \right)^2 \\
r^2 &= \frac{\left(\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right)^2}{\left(\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right) \left(\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 \right)}
\end{aligned}$$

Thus, $R^2 = r^2$. ■

While R^2 and r are useful measures, they have limitations:

- They provide no information about the absolute precision of the model for estimating a mean response or predicting new observations.
- They quantify only the linear association between predictors and the response variable. A low R^2 or r does not necessarily imply no relationship, but rather no linear relationship.

Definition 8.9: Adjusted R^2

Adjusted R^2 accounts for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of predictors. It is defined as:

$$R_a^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}\,/(n-1)} = 1 - \frac{(n-1)}{(n-p-1)} \frac{\text{RSS}}{\text{TSS}}$$

The adjusted R^2 can be negative and is always less than or equal to 1:

$$-\frac{p}{n-p-1} \leq R_a^2 \leq 1$$

8.6 Multi-collinearity: R^2 perspective

Consider the multiple linear regression model:

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

The least squares estimator for the regression coefficients is:

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

From statistical theory, we know that:

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

The variance of $\widehat{\beta}_j$ is influenced by the presence of multi-collinearity, which occurs when the predictor variables are highly linearly correlated. Specifically:

$$\operatorname{Var}\left(\widehat{\beta}_j\right)$$

This variance becomes large when other variables are highly linearly correlated with x_j.
To quantify the linear association between x_j and the other predictors, consider the following linear regression model:

$$x_{i,j} = \gamma_0 + \gamma_1 x_{i,1} + \ldots + \gamma_{j-1} x_{i,j-1} + \gamma_{j+1} x_{i,j+1} + \ldots + \gamma_p x_{i,p} + \epsilon_i$$

Let R_j^2 be the coefficient of determination for this regression model. It can be shown that:

$$\operatorname{Var}\left(\widehat{\beta}_j\right) = \sigma^2\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{j+1,j+1} = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{(n-1)S_j^2}$$

where S_j^2 is the sample variance for the j-th covariate.

The term $\frac{1}{1-R_j^2}$ is called the Variance Inflation Factor (VIF):

$$\mathrm{VIF} = \frac{1}{1-R_j^2}$$

The VIF measures how much the variance of $\widehat{\beta}_j$ is inflated due to multi-collinearity. Technically, VIF > 1 indicates some multi-collinearity, but generally, a VIF > 5 is considered to indicate severe multicollinearity.

For the special case where $p = 2$, consider the model:

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

Here, $R_1^2 = R_2^2 = r_{1,2}^2$, where $r_{1,2}$ is the sample correlation between X_1 and X_2 :

$$r_{1,2} = \frac{\sum_{i=1}^n \left(x_{i,1} - \bar{x}_1\right)\left(x_{i,2} - \bar{x}_2\right)}{\sqrt{\sum_{i=1}^n \left(x_{i,1} - \bar{x}_1\right)^2}\sqrt{\sum_{i=1}^n \left(x_{i,2} - \bar{x}_2\right)^2}}$$

The sample means are:

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^n x_{i,1}$$
$$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^n x_{i,2}$$

In this case, the VIF is:

$$\mathrm{VIF} = \frac{1}{1-R_1^2} = \frac{1}{1-R_2^2}$$

9 Problematic observations

9.1 General intuition

Definition 9.1: Leverage point

A leverage point is a data point where the independent variable value (denoted as x) is far from the average of all the independent variables, $\overline{x}$. These points have the potential to influence the fitting of the regression model significantly.

Definition 9.2: Outlier

An outlier is a data point that does not follow the general trend or pattern established by the majority of the data. This could be due to errors, variability in data, or other factors.

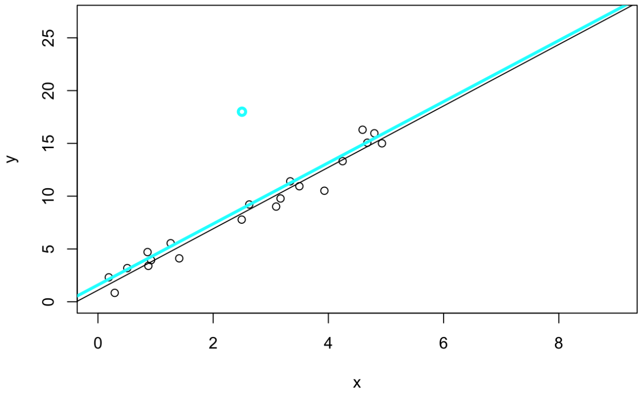


Figure 35: Non-leverage outlier

Definition 9.3: Good leverage point

A good leverage point is a leverage point that is not an outlier. It's far from the mean $\overline{x}$ but still follows the overall trend of the data, so it does not distort the regression line.

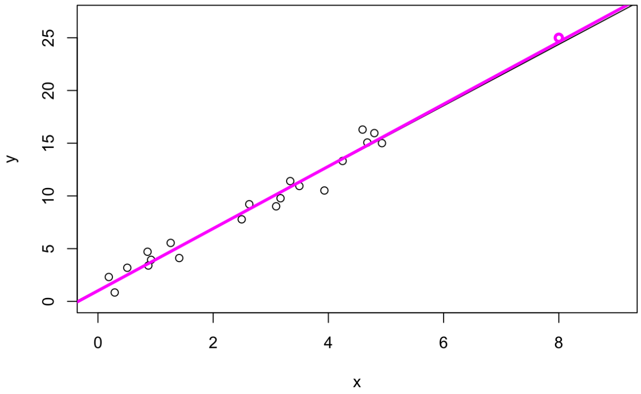


Figure 36: Leverage point (non-influential)

Definition 9.4: Bad leverage point

A bad leverage point is a leverage point that is also an outlier. It is far from $\overline{\boldsymbol{x}}$ and does not follow the trend of the rest of the data, potentially causing a significant shift in the regression line.

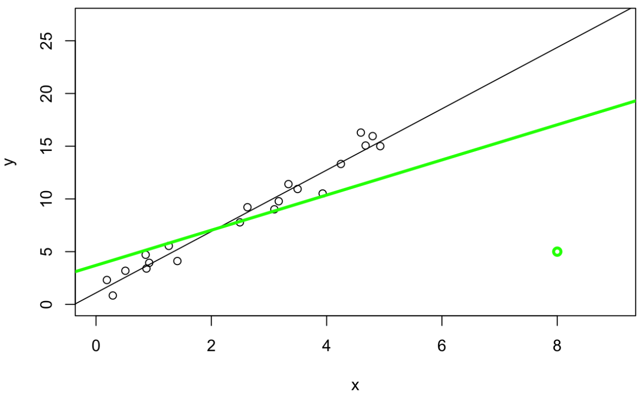


Figure 37: Influential point

9.2 Leverage point

Leverage points can potentially shift the regression line, but they do not always do so. Their impact depends on whether they are also outliers (bad leverage points). For a leverage point (x_i, y_i), the fitted value $\widehat{y}_i$ is significantly influenced by the actual value y_i. This relationship is expressed as:

$$\widehat{y}_i = \sum_{j=1}^{n} h_{i,j} y_j = h_{i,i} y_i + \sum_{j \neq i} h_{i,j} y_j$$

where $h_{i,i}$ represents the leverage of the i^{th} data point. Leverage quantifies how much the fitted value $\widehat{yi}$ is determined by the actual value y_i. A larger hi, i indicates that the regression line is more influenced by that particular data point because it is further from the center of the data.

9.2.1 Properties of hat matrix

The hat matrix $\boldsymbol{H}$ in simple linear regression has the following properties:

- **Symmetric:** $\boldsymbol{H}^\top = \boldsymbol{H}$
- **Idempotent:** $\boldsymbol{H}^2 = \boldsymbol{H}$, implying $\boldsymbol{H}^k = \boldsymbol{H}$ for any power k.
- **Trace:** The trace of $\boldsymbol{H}$, which is the sum of its diagonal elements, equals $p + 1$ (where p is the number of predictors).

9.2.2 Properties of the entries of the hat matrix

In simple linear regression, each entry of $\boldsymbol{H}$ is given by:

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

and the diagonal elements (leverage values) are:

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

These values indicate that the leverage is minimized when $x_i = \bar{x}$. Define the averaged leverage $\bar{h} = \frac{1}{n}\sum_{i=1}^n h_{i,i} = \frac{p+1}{n}$.

The hat matrix $\boldsymbol{H}$ is defined as:

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$$

When we multiply the hat matrix $\boldsymbol{H}$ by the design matrix $\boldsymbol{X}$, we get:

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X} \cdot \boldsymbol{I} = \boldsymbol{X}$$

Given this identity, we know that

$$\sum_{j=1}^n h_{i,j} = \sum_{i=1}^n h_{i,j} = 1, \text{ for any } i \text{ and } j$$

Also given $\boldsymbol{H}^2 = \boldsymbol{H}$, the (i,i)-th entry yields that:

$$h_{i,i} = \sum_{j=1}^n h_{i,j}^2$$

Proof. First, we wish to prove $\sum_{j=1}^n h_{i,j} = \sum_{i=1}^n h_{i,j} = 1$, for any i and j. For each row i and each column k of this product, we have:

$$(\boldsymbol{H}\boldsymbol{X})_{ik} = \sum_{j=1}^n h_{i,j} x_{jk}$$

where x_{jk} is the element in the j-th row and k-th column of the matrix $\boldsymbol{X}$. The equation $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ implies:

$$\sum_{j=1}^n h_{i,j} x_{jk} = x_{ik} \quad \text{for each } k \text{ and } i$$

Let's now consider the sum over all columns k for a fixed row i :

$$\sum_{k=1}^p \sum_{j=1}^n h_{i,j} x_{jk} = \sum_{k=1}^p x_{ik}$$

Notice that $\sum_{k=1}^p x_{ik}$ is just a summation over all the columns for a fixed row i. The summation on the left side can be rewritten as:

$$\sum_{j=1}^n h_{i,j} \sum_{k=1}^p x_{jk} = \sum_{k=1}^p x_{ik}$$

Since $\sum_{k=1}^p x_{jk}$ is not generally zero for any j, we must have:

$$\sum_{j=1}^n h_{i,j} = 1 \quad \text{for each } i$$

Given that $\boldsymbol{H}$ is symmetric $(h_{i,j} = h_{j,i})$, the sum over columns must be equal to the sum over rows:

$$\sum_{i=1}^n h_{i,j} = 1 \quad \text{for each } j$$

■

Proof. Consider the matrix product $\boldsymbol{H} \cdot \boldsymbol{H}$. The (i,i)-th entry of the matrix $\boldsymbol{H}^2$ is given by:

$$\left[\boldsymbol{H}^2\right]_{ii} = \sum_{j=1}^n h_{i,j} h_{j,i}$$

Since $\boldsymbol{H}$ is a symmetric matrix, $h_{j,i} = h_{i,j}$, so this simplifies to:

$$\left[\boldsymbol{H}^2\right]_{ii} = \sum_{j=1}^n h_{i,j} h_{i,j} = \sum_{j=1}^n h_{i,j}^2$$

Using the idempotence property $\boldsymbol{H}^2 = \boldsymbol{H}$, we know that:

$$\left[\boldsymbol{H}^2\right]_{ii} = h_{i,i}$$

Thus, we have:

$$h_{i,i} = \sum_{j=1}^{n} h_{i,j}^2$$

■

9.2.3 Ranges of diagonal terms

We know $h_{i,i} = \sum_{j=1}^{n} h_{i,j}^2$ implies $h_{i,i} > 0$. $h_{i,i} = h_{i,i}^2 + \sum_{j \neq i} h_{i,j}^2$ implies $h_{i,i} \leq 1$. Thus, we have:

$$1 = \left(h_{i,1} + \ldots + h_{i,n}\right)^2 \leq \left(1^2 + \ldots + 1^2\right)\left(h_{i,1}^2 + \ldots, +h_{i,n}^2\right) = nh_{i,i}$$

The inequality implies $h_{i,i} \geq \frac{1}{n}$, and thus $\frac{1}{n} \leq h_{i,i} \leq 1$.

9.2.4 Criteria for leverge point

leverage values and those that are unusually large. A common rule of thumb is to consider a point as having high leverage if its leverage is at least twice the average leverage. Thus, the criterion for a high leverage point becomes:

$$h_{i,i} \geq 2\bar{h}$$

Substituting the expression for $\bar{h}$, we get:

$$h_{i,i} \geq 2 \cdot \frac{p+1}{n} = \frac{2(p+1)}{n}$$

9.3 Outliers

Residuals are the differences between observed values and the values predicted by the model. Mathematically, the residual vector is given by:

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

Here, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{H}$ is the hat matrix, $\boldsymbol{y}$ is the vector of observed values, and $\widehat{\boldsymbol{y}}$ is the vector of predicted values. For the i-th data point, the residual can be expressed as:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \sum_{j=1}^{n} h_{i,j} y_j = y_i - \left(h_{i,i} y_i + \sum_{j \neq i} h_{i,j} y_j\right) = y_i - h_{i,i} y_i - \sum_{j \neq i} h_{i,j} y_j = \left(1 - h_{i,i}\right) y_i - \sum_{j \neq i} h_{i,j} y_j$$

High leverage points tend to have smaller residuals because these points have a strong influence on the fitted value $\widehat{y}_i$. As a result, the model fits these points closely, leading to smaller residuals. Residuals alone are not reliable for outlier detection, especially in the presence of high leverage points. This is because high leverage points can have small residuals despite being outliers.

9.3.1 Standarized residual

From earlier lectures, it's established that the residuals $\widehat{\boldsymbol{e}}$ follow a multivariate normal distribution:

$$\widehat{\boldsymbol{e}} \sim N_n\left(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right)$$

This indicates that the residuals are centered around zero with a variance that depends on the hat matrix $\boldsymbol{H}$. The variance of the i-th residual $\widehat{e}_i$ is not constant and is given by:

$$\operatorname{Var}\left(\hat{e}_i\right) = \sigma^2\left(1 - h_{i,i}\right)$$

This variance depends on the leverage $h_{i,i}$ of the i-th data point, meaning that points with higher leverage have smaller residual variance. To account for the varying variance of residuals, we use the standardized residual, defined as:

$$r_i = \frac{\hat{e}_i}{S\sqrt{1 - h_{i,i}}}$$

Here, S is an estimate of the standard deviation of the errors. The standardized residual r_i provides a more consistent measure across all data points by adjusting for the leverage.

9.3.2 Criteria for outlier detection

- Small Datasets $(n < 50)$: For smaller datasets, a data point $(\boldsymbol{x}_i, y_i)$ is considered an outlier if the absolute value of the standardized residual exceeds 2:
$$|r_i| > 2$$
- Large Datasets $(n \geq 50)$: For larger datasets, a stricter criterion is applied. A data point is flagged as an outlier if the absolute value of the standardized residual exceeds 4:
$$|r_i| > 4$$

9.4 Influential observations

Both leverage points and outliers can be problematic. These points are considered influential observations because they have the potential to alter the estimated regression relationship significantly. A problematic observation can influence the model in three main ways:

- **Affect its own fitted value:** The observation may skew its predicted value.
- **Affect all fitted values:** The observation may impact the overall fit of the model, altering predictions for other data points.
- **Affect the estimated coefficients:** The observation can change the estimated regression coefficients, leading to a different model.

To address these issues, it is necessary to quantify the influence of each observation on the fitted values and the estimated coefficients.

9.4.1 Influence on own fitted value - DFFITS

Definition 9.5: Difference in Fitted value

DFFITS (Difference in FITted Values) is a measure that quantifies how much an observation influences its own fitted value. It is computed as:

$$\text{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{S_{(i)}\sqrt{h_{i,i}}}$$

Here, $\widehat{y_i(i)}$ is the fitted value for observation i calculated from the model after excluding the i-th observation. $S^2_{(i)}$ is the estimated error variance from the model excluding i, and $h_{i,i}$ is the leverage of the i-th point.

DFFITS measures the change in the fitted value when the i-th observation is removed from the model. A high DFFITS value indicates that the observation has a large influence on its own prediction.

An observation is considered influential on its own fitted value if:

$$\text{DFFITS}_i > 2\sqrt{\frac{p+1}{n}}$$

where p is the number of predictors, and n is the sample size. This threshold is based on the sample size and the number of predictors in the model.

9.4.2 Cook distance

Definition 9.6: Cook distance

Cook's Distance is used to measure the influence of an observation on all fitted values in the model. It is defined as:

$$D_i = \frac{\sum_{j=1}^{n}\left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{(p+1)S^2}$$

where $\widehat{y}_{j(i)}$ is the fitted value for the j-th observation after removing the i-th observation from the model. Cook's Distance can also be calculated using:

$$D_i = \frac{r_i^2}{p+1} \cdot \frac{h_{i,i}}{1-h_{i,i}}$$

This formulation avoids the need to refit the model $n+1$ times, making it computationally efficient.

An observation is considered influential on the overall fitted values if:

$$D_i > F_{0.5}(p+1, n-p-1)$$

where $F_{0.5}$ is the median of the F-distribution with degrees of freedom $(p+1, n-p-1)$. This criterion provides a statistical benchmark for detecting influential points.

9.4.3 Influence on regression coefficients - DFBETAS

Definition 9.7: DFBETAS

DFBETAS measures the influence of an observation on the estimation of the regression coefficients. For the k-th coefficient, it is defined as:

$$\mathrm{DFBETAS}_{k(i)} = \frac{\widehat{\beta}_k - \widehat{\beta}_{k(i)}}{S_{(i)}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{k+1,k+1}}}$$

Here, $\widehat{\beta}k$ is the estimated coefficient with all data points included, and $\widehat{\beta}k(i)$ is the coefficient estimated after excluding the i-th observation. $S_{(i)}$ is the standard error from the model without the i-th observation.

DFBETAS quantifies the impact of a single observation on the estimation of a specific coefficient. High values indicate that the observation significantly alters the coefficient estimate.

An observation is considered influential on the estimation of the k-th coefficient if:

$$\mathrm{DFBETAS}_{k(i)} > \frac{2}{\sqrt{n}}$$

This threshold is designed for larger datasets.

9.5 More remarks

To determine whether an observation is influential, it is sometimes useful to directly compare the regression model's inference with and without the observation in question. This involves assessing changes in the coefficients, fitted values, and overall model fit.

Identifying influential observations is somewhat subjective. While statistical measures like DFFITS, Cook's Distance, and DFBETAS provide guidelines, the final decision often depends on the context and the specific goals of the analysis.

All the measures discussed are based on the concept of Leave-One-Out Cross-Validation (LOOCV). LOOCV involves fitting the model multiple times, each time excluding a different observation, to assess the impact of each observation on the model's performance.

10 Model selection and interaction

10.1 Variable selection

Variable selection is crucial in many fields, especially when dealing with large datasets where not all variables (or features) contribute equally to the outcome. This section gives examples of where variable selection is applied:

1. **Genetic Data:** Imagine you're studying a disease like cancer, and you have data on thousands of genes. Not all of these genes will be relevant to the disease. Variable selection helps identify which specific genes are most strongly associated with the disease.

2. **Brain Imaging (fMRI):** In studies of brain function, researchers often want to link specific brain regions to behaviors or conditions, like memory loss. Here, variable selection would help pinpoint the regions most relevant to memory.

3. **Network Data:** In social networks, some individuals (nodes) are more influential than others. Variable selection can identify these key players.

4. **Handwritten Numerical Data:** Consider digit recognition (like reading handwritten numbers). Variable selection helps determine which pixels in the image are most important for correctly identifying the number.

In linear regression, when you have too many predictors, some of them might not be relevant, or they might make the model overly complex. In some cases, the number of predictors (p) can be larger than the number of observations (n). For example, if you have data on 50 patients but 1000 potential genes, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ (used to calculate the regression coefficients) becomes non-invertible, meaning you can't solve the linear regression problem directly. This course will focus on situations where the number of predictors (p) is less than the number of observations (n), making the problem more tractable.

10.2 Selection criteria

Suppose we have m potential predictors $X_1, \ldots, X_m$, which are the ones we shall include to the linear regression model? Suppose we choose p predictors to fit a linear regression model (Model 1), and choose p predictors (can be overlapped with the previous choice) to fit a linear regression model (Model 2). Which model shall we pick? Given the m predictors, which $p = 1, 2, \ldots, m$ shall we prefer?

Given those motivational questions, we need a criterion to compare the models and an algorithm to select the variables.

10.2.1 Adjusted R^2

R^2 is a measure of how well the independent variables explain the variation in the dependent variable. It ranges from 0 to 1, with 1 indicating a perfect fit. However, a key issue with R^2 is that it always increases as more variables are added to the model, even if those variables are not useful. To address the problem of R^2 increasing regardless of the usefulness of the predictors, Adjusted R^2 was introduced. It adjusts R^2 based on the number of predictors in the model.

$$R_a^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

Adjusted R^2 only increases if the new variable improves the model more than would be expected by chance. If a variable doesn't add meaningful information, adjusted R^2 can decrease, which helps in avoiding overfitting. Even though Adjusted R^2 is more robust than R^2, it can still lead to overfitting in some cases, especially when the model complexity increases without improving the generalizability of the model.

10.2.2 MLE

The MLE is a method used to estimate the parameters (like coefficients in regression) that maximize the likelihood of the observed data given the model. The likelihood function measures how probable the observed data is, given a set of parameters. The log-likelihood is simply the logarithm of this likelihood, which simplifies calculations.

$$\log L\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{RSS}{2\sigma^2}$$

The MLE of the coefficients is:

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

This equation gives us the values of the coefficients that maximize the likelihood of observing the given data. The MLE for the variance of the errors is:

$$\widehat{\sigma^2} = \frac{RSS(\widehat{\boldsymbol{\beta}})}{n}$$

Therefore,

$$\log L\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}; \boldsymbol{y}\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{RSS(\widehat{\boldsymbol{\beta}})}{n}\right) - \frac{n}{2}$$

10.2.3 Akaike's Information Criterion (AIC)

AIC is a criterion for model selection based on the likelihood function. It is defined as:

$$AIC = 2\left[-\log L\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}; \boldsymbol{y}\right) + p + 2\right]$$

Simplifying, this becomes:

$$AIC = n\log(2\pi) + n\log\left(\frac{RSS(\widehat{\boldsymbol{\beta}})}{n}\right) + n + 2(p+2)$$

The smaller the AIC, the better the model. AIC penalizes models with more parameters to avoid overfitting. The goal is to find a model that minimizes:

$$n\log(\mathrm{RSS}(\widehat{\boldsymbol{\beta}})) + 2p$$

10.2.4 Corrected AIC

AIC_C is a corrected version of AIC for small sample sizes, suggested by Hurvich and Tsai (1989) and Anderson (2004):

$$AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-1}$$

This correction is important when the sample size is small, typically when $n \leq 40(p+2)$.

10.2.5 Bayesian Information Criterion

BIC is another model selection criterion, similar to AIC but with a stronger penalty for models with more parameters. It is defined as:

$$BIC = -2\log L\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}; \boldsymbol{y}\right) + (p+2)\log n$$

Simplifying, this becomes:

$$BIC = n\log(2\pi) + n\log\left(\frac{RSS(\widehat{\boldsymbol{\beta}})}{n}\right) + n + (p+2)\log n$$

The smaller the BIC, the better the model. BIC tends to select simpler models compared to AIC because of the logarithmic penalty on the number of parameters. The goal is to find a model that minimizes:

$$n\log(RSS(\widehat{\boldsymbol{\beta}})) + p\log n$$

10.2.6 Other criteria

There are numerous other information criteria, each with its own characteristics and applications. Some of these include CIC, DIC, ZIC, and AAIC. These criteria are designed to address specific issues in model selection, such as dealing with complex models, different data structures, or varying sample sizes.

10.2.7 An example

Suppose we have a dataset with 50 observations $(n=50)$ and 5 potential predictors (X_1, X_2, X_3, X_4, X_5) that could be used to predict a response variable Y. We will consider two models:

- Model 1: Includes predictors X_1, X_2, and X_3.
- Model 2: Includes predictors X_1, X_4 and X_5.

For both models, let's assume the following:

- RSS_1 (Residual Sum of Squares for Model 1) $=100$
- RSS_2 (Residual Sum of Squares for Model 2) $=120$

The number of predictors (p) in both models is 3. Assume the Total Sum of Squares TSS for the dataset is 200 .

The formula for Adjusted R^2 is:

$$R_a^2=1-\frac{RSS/(n-p-1)}{\text{TSS }/(n-1)}$$

- $R_{a1}^2=1-\frac{100/(50-3-1)}{200/(50-1)}=1-\frac{100/46}{200/49}=1-\frac{2.17}{4.08}\approx 0.4686$
- $R_{a2}^2=1-\frac{120/(50-3-1)}{200/(50-1)}=1-\frac{120/46}{200/49}=1-\frac{2.61}{4.08}\approx 0.3596$

Model 1 has a higher Adjusted $R^2(0.4686)$ compared to Model $2(0.3596)$, so Model 1 is preferred.

The AIC formula is:

$$AIC=n\log(2\pi)+n\log\left(\frac{RSS(\widehat{\boldsymbol{\beta}})}{n}\right)+n+2(p+2)$$

To simplify the comparison, we focus on minimizing:

$$AIC=n\log\left(\frac{RSS}{n}\right)+2p$$

For both models, $n=50$ and $p=3$.

- $AIC_1=50\log\left(\frac{100}{50}\right)+2(3+2)=50\log(2)+10\approx 34.66+10=44.66$
- $AIC_2=50\log\left(\frac{120}{50}\right)+2(3+2)=50\log(2.4)+10\approx 43.83+10=53.83$

Model 1 has a lower AIC (44.66) compared to Model 2 (53.83), so Model 1 is preferred.

The formula for AIC_C is:

$$AIC_C=AIC+\frac{2(p+2)(p+3)}{n-p-1}$$

Let's calculate this for both models:

- $AIC_{C1}=44.66+\frac{2(3+2)(3+3)}{50-3-1}=44.66+\frac{2(5)(6)}{46}\approx 44.66+1.30=45.96$
- $AIC_{C2}=53.83+\frac{2(3+2)(3+3)}{50-3-1}=53.83+\frac{2(5)(6)}{46}\approx 53.83+1.30=55.13$

Model 1 has a lower $\text{AIC}_C(45.96)$ compared to Model 2 (55.13), so Model 1 is preferred.

The BIC formula is:

$$BIC=n\log(2\pi)+n\log\left(\frac{RSS(\widehat{\boldsymbol{\beta}})}{n}\right)+n+(p+2)\log n$$

Focusing on the simplified comparison:

$$BIC=n\log\left(\frac{RSS}{n}\right)+p\log n$$

For $n=50, p=3$:

- $BIC_1=50\log\left(\frac{100}{50}\right)+3\log 50\approx 34.66+10.95=45.61$
- $BIC_2=50\log\left(\frac{120}{50}\right)+3\log 50\approx 43.83+10.95=54.78$

Model 1 has a lower BIC (45.61) compared to Model 2 (54.78), so Model 1 is preferred.

10.3 Selection methods for variables

10.3.1 General procedures

The general procedure for variable selection involves choosing the best set of predictors from a given dataset to build an optimal model. The steps involved are:

1. **Pick a Selection Criterion:** Start by deciding on a criterion to evaluate the models, such as the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), or others.

2. **Fit Different Models:** Fit multiple models on the same dataset, each with a different combination of predictors.

3. **Compare Models:** Evaluate these models using the chosen selection criterion. For example, if using BIC, find the model with the smallest BIC value.

4. **Break Ties:** If two models have the same BIC, prefer the model with fewer predictors. If the number of predictors is also the same, break the tie arbitrarily by picking one of the model arbitrarily.

10.3.2 Best subset selection

Best Subset Selection involves trying all possible combinations of the predictors to find the best model. For m predictors, there are 2^m possible models, ranging from no predictors to all m predictors. However, the method is exhaustive and ensures that the global optimum model is found, but it can become computationally infeasible when m is large due to the exponential number of possible models. For instance, with 10 predictors, you would need to evaluate $2^{10} = 1024$ models, but with 30 predictors, it jumps to over a billion models.

10.3.3 Backwards elimination

Backward Elimination is a stepwise selection method that starts with the full model (all predictors included) and then sequentially removes predictors:

1. **Start with Full Model:** Begin with a model that includes all m predictors.

2. **Delete Predictors:** Remove one predictor at a time, selecting the one whose removal leads to the largest improvement (or smallest worsening) in the selection criterion, such as BIC or AIC.

3. **Stopping Rule:** Continue removing predictors until no further improvement in the criterion can be achieved, or all predictors have been removed.

Greedy Algorithm: This method is greedy because it makes the best local choice at each step (i.e., removing the predictor that most improves the criterion) but does not necessarily result in the globally optimal model.

10.3.4 Forward selection

Forward Selection is another stepwise selection method, but it starts with an empty model (no predictors included) and then sequentially adds predictors:

1. **Start with Empty Model:** Begin with no predictors in the model.

2. **Add Predictors:** Add one predictor at a time, selecting the one that leads to the largest improvement in the selection criterion.

3. **Stopping Rule:** Continue adding predictors until no further improvement in the criterion can be achieved, or all predictors have been added.

Greedy Algorithm: Like backward elimination, forward selection is also a greedy algorithm and may not lead to the globally optimal model.

10.3.5 Penalized regression for variable selection

Penalized regression techniques, such as LASSO, Ridge Regression, and Elastic Net, incorporate penalties on the coefficients to perform variable selection. These methods are particularly useful when dealing with high-dimensional data with a large p and aim to reduce model complexity by shrinking some coefficients to zero.

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a popular penalized regression technique:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Remark.

> If we do not want to penalize β_0, replace $\|\beta\|_1$ to $\|\beta\|_1 - |\beta_0|$

10.4 Dummy variables

Before diving into dummy variables, it's important to understand the different types of variables used in statistical modeling:

- **Quantitative Variables:** These are numeric values that can be measured. Examples include the price of a cup of coffee or the salary of a professor.
- **Qualitative Variables:** These variables represent categories or groups and are typically encoded as integers. They can be further classified into:
  - *Ordinal Variables:* These have a natural order. For example, size (high, medium, low) or evaluation grades $(A+, A, A - B+)$.
  - *Nominal Variables:* These do not have a natural order. Examples include blood type, eye color, gender, and academic major.

A dummy variable is used to represent qualitative variables in statistical models. It is an indicator variable that takes on a binary value (0 or 1) to encode different categories of a qualitative variable.

For example, consider a binary qualitative variable representing gender:

$$D = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if Male} \end{cases}$$

Alternatively, you could define the dummy variable in reverse:

$$D' = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

Note that both D and D' contain the same information, so it's unnecessary to define both in the same model.

Dummy variables can be used in regression models to account for different intercepts across categories. Let's consider a model where y_i is the dependent variable, x_i is an independent variable, and d_i is a dummy variable representing a category:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i$$

This model can be rewritten based on the value of d_i :

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } d_i = 0 \\ (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i & \text{if } d_i = 1 \end{cases}$$

In this model:

- If $d_i = 0$, the intercept is β_0.
- If $d_i = 1$, the intercept is $\beta_0 + \beta_2$.

Example.

> Consider a dataset where x_i, y_i, and d_i represent the weight, height, and gender of person i, respectively. This model would create two parallel regression lines: one for males and one for females, each with its own intercept.

Example.

> Suppose we want to study the GPA of students based on their major. Here, the response variable y_i is the GPA of student i, and the categories (majors) include Statistics, Math, Engineering, and Computer Science. We can define dummy variables for each major:
>
> $$D_i^{(M)} = \begin{cases} 1 & \text{if Math;} \\ 0 & \text{otherwise;} \end{cases}$$
>
> $$D_i^{(E)} = \begin{cases} 1 & \text{if Engineering;} \\ 0 & \text{otherwise;} \end{cases}$$
>
> $$D_i^{(C)} = \begin{cases} 1 & \text{if Computer Science;} \\ 0 & \text{otherwise} \end{cases}$$
>
> The multiple linear regression model with multiple intercepts would be:
>
> $$y_i = \beta_0 + \beta_1 D_i^{(M)} + \beta_2 D_i^{(E)} + \beta_3 D_i^{(C)} + \epsilon_i$$
>
> In this model:
>
> - $E(y_i) = \beta_0$ if student i majors in Statistics (as all dummy variables would be 0).
> - $E(y_i) = \beta_0 + \beta_1$ if student i majors in Math.
> - $E(y_i) = \beta_0 + \beta_2$ if student i majors in Engineering.
> - $E(y_i) = \beta_0 + \beta_3$ if student i majors in Computer Science.

Remark.

> **Dummy Variable Trap:** You can only use $p-1$ dummy variables to model p categories. This is to avoid multicollinearity (also known as the dummy variable trap), where the predictors become linearly dependent. For example, if you had four majors, you would use only three dummy variables.

Remark.

> **Alternative Coding:** Sometimes, people use $+1, -1$ instead of 0,1 to code dummy variables, depending on the context or modeling technique.

10.5 Interaction terms in regression

Variable selection typically involves choosing which variables (e.g., $X_1, \ldots, X_p$) to include in a regression model. However, when variables are selected, they are often assumed to affect the dependent variable y independently. In many real-world situations, the effect of one variable on the dependent variable might depend on the level of another variable. This is where interaction terms come into play.

Interaction terms are products of different variables, representing the combined effect of two or more variables on the dependent variable. For instance:

- A simple interaction between two variables X_j and X_k can be represented as $X_j X_k$.
- More complex interactions can involve three or more variables, such as $X_1 X_2 X_3$ or even all variables in the model, $X_1 X_2 \ldots X_p$.

These interaction terms allow the model to capture the situation where the effect of one variable on the outcome changes depending on the value of another variable.

A linear regression model can include both main effects (the independent effects of each variable) and interaction effects (the combined effect of two or more variables). The general form of such a model is:

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

In this model:

- β_0 is the intercept.
- $\beta_1, \ldots, \beta_p$ are the coefficients for the main effects of the predictors.
- $\beta_{1,2}$ is the coefficient for the interaction between $x_{i,1}$ and $x_{i,2}$.
- ϵ_i is the error term.

Interaction terms can also be used to model scenarios where different categories have different slopes. Consider a model where d_i is a dummy variable:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i x_i + \epsilon_i$$

This model can be interpreted as:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } d_i = 0 \\ \beta_0 + (\beta_1 + \beta_2)\, x_i + \epsilon_i & \text{if } d_i = 1 \end{cases}$$

- When $d_i = 0:$ The slope is β_1.
- When $d_i = 1:$ The slope is $\beta_1 + \beta_2$.

Example.

> Consider two investment products, A and B , where $d_i = 1$ for product A and $d_i = 0$ for product B . Let x_i represent the investment period, and y_i represent the final fortune. This model allows the effect of the investment period on the final fortune to differ between the two products. The model describes two regression lines with different slopes, intersecting at $(0, \beta_0)$.

In some cases, both the intercept and the slope may vary across categories. The model for this scenario is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 d_i x_i + \epsilon_i$$

This can be interpreted as:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } d_i = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_i + \epsilon_i & \text{if } d_i = 1 \end{cases}$$

- When $d_i = 0:$ The intercept is β_0 and the slope is β_1.
- When $d_i = 1:$ The intercept is $\beta_0 + \beta_2$ and the slope is $\beta_1 + \beta_3$.

Example.

> This model could describe the income of two taxi companies with different charging schemes, where $d_i = 1$ for one company and $d_i = 0$ for the other. This setup results in two regression lines that differ in both intercept and slope.

When you include interaction terms in a model with p covariates, the number of possible interaction terms can be calculated as:

$$2^p - p - 1$$

Combining both the original covariates and their interaction terms, the total number of possible predictors becomes:

$$2^p - 1$$

Given this, the total number of possible linear models (with or without interaction terms) you can build is:

$$2^{2^p-1}$$

This exponential growth in the number of possible models demonstrates how quickly the complexity can increase when including interaction terms, particularly in models with a large number of predictors.

11 Model validation

11.1 Under fitting and over fitting

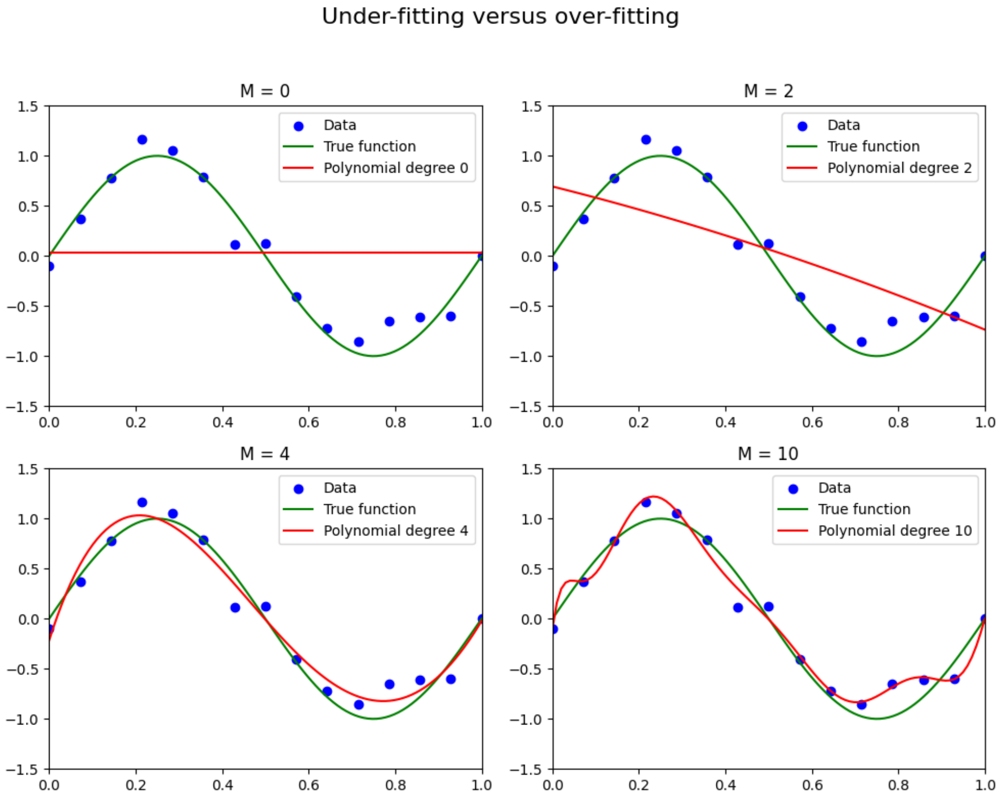


Figure 38: Under-fitting versus over-fitting

- **Under-fitting:** Occurs when the model is too simple to capture the underlying patterns in the data. The model might have high bias, leading to poor performance both on training and test data.
- **Over-fitting:** The opposite of under-fitting. Here, the model is too complex and fits the training data too closely, including its noise. Although it performs well on training data, it lacks the ability to generalize to new, unseen data, resulting in high variance.

11.1.1 Avoid over-fitting

One way to avoid over-fitting is to carefully select which variables (features) are included in the model. Removing irrelevant or redundant features can help in reducing the model's complexity.

Another approach to avoid over-fitting is to add a penalty to the loss function. The general form of the penalization method is given by:

$$\text{Objective function } = \text{ Loss function } + \lambda \times \text{ Penalty}$$

The parameter $\lambda > 0$ controls the model complexity. A small λ results in a model with less penalty, potentially leading to over-fitting, while a large λ increases the penalty, leading to a simpler model that might under-fit. By adjusting λ, one can control the trade-off between bias and variance, achieving an optimal model that generalizes well to new data.

- The objective function for ridge regression is:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

Here, $\|\boldsymbol{\beta}\|^2$ represents the $L2$ norm (sum of the squares of the coefficients), which penalizes large coefficients and thus prevents the model from becoming too complex.

- The objective function for LASSO regression is:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

  In this case, $\|\boldsymbol{\beta}\|_1$ represents the L1 norm (sum of the absolute values of the coefficients). LASSO not only penalizes large coefficients but can also shrink some coefficients to exactly zero, effectively performing variable selection.

Here, λ is a tuning parameter that controls the strength of the penalty, which in turn controls the complexity of the model.

11.2 Validation

After fitting a regression model (or any machine learning model), it's important to assess how well the model performs on new, unseen data. This is called **generalization**:

- When a new data point or dataset is introduced, the ability of the model to make accurate predictions is a key measure of its generalization capability.
- This term refers to the error rate (or accuracy) of the model when applied to unseen data. It's a crucial indicator of how well the model has learned the underlying patterns in the data, rather than just memorizing the training data.

When fitting a model, we don't have access to unseen data that would allow us to directly evaluate how well the model will generalize. Therefore, we must simulate this scenario using the data we do have. To assess generalization, we split the original dataset into two main parts:

1. **Training Set:** Used to fit (train) the model.
2. **Validation/Test Set (or Hold-Out Set):** Used to evaluate the model's performance on unseen data.

The dataset is typically split randomly to ensure that the training and validation sets are representative of the overall data distribution.

Example.

> Consider the data structure:
>
> $$(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, y_i)\,, \text{ for } i = 1, \dots, 100$$
>
> We want to validate the model:
>
> $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$$
>
> Randomly sample 80 integers from $\{1, \dots, 100\}$ and denote this set as T. This set T will serve as our training set. $\{(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, y_i) : i \in T\}$ set is used to fit the model, estimating the coefficients $\boldsymbol{\beta}$.
> The remaining 20 integers will be in the set $V = \{1, \dots, 100\}\backslash T$, which will serve as our validation set. $\{(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, y_i) : i \in V\}$ set is used to evaluate the model's performance, providing an estimate of how well the model will generalize to unseen data.
> We find the optimal coefficients $\widehat{\boldsymbol{\beta}}$ by minimizing the sum of squared errors (SSE) on the training set T :
>
> $$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i \in T} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \beta_3 x_{i,3})^2$$
>
> After fitting the model, we assess its performance on the validation set V using Mean Squared Error (MSE):
>
> $$MSE = \frac{1}{|V|} \sum_{i \in V} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \beta_3 x_{i,3})^2$$
>
> Here, $|V|$ is the number of observations in the validation set.

Similar to selecting the best model, validation can help us choose the model that exhibits the best generalization performance by comparing different models on the validation set. This method is particularly useful for choosing the best tuning parameter λ in penalized regression (like Ridge or LASSO). The λ that results in the lowest validation error is considered the optimal choice.

One potential issue with validation is that the results can be sensitive to how the data is split. For instance, if the validation set contains outliers, it could lead to biased results. This is why techniques like **cross-validation** are often used to mitigate the impact of any particular data split.

11.3 Cross-validation

To better estimate the generalization performance of a model by splitting the data into multiple subsets and using different subsets for training and validation.

1. **Multiple Data Splits:** The dataset is split in various ways.
2. **Validation for Each Split:** For each split, the model is trained on one subset and validated on another.
3. **Averaging Performance:** The generalization performance across all splits is averaged to provide a more reliable estimate.

Here, we will talk about:

1. **Leave-One-Out Cross-Validation (LOOCV):** Each data point is used once as a validation set, while the rest are used as the training set.
2. **K-Fold Cross-Validation:** The data is divided into K equally sized subsets, or "folds". The model is trained on $K-1$ folds and validated on the remaining fold. This process is repeated K times, each time with a different fold as the validation set.

11.3.1 Leave-One-Out Cross-Validation (LOOCV)

Let the dataset be z_i for $i = 1, \ldots, n$. In linear regression, $z_i = (\boldsymbol{x}_i, y_i)$. For each data point i :

- **Training Set:** All data points except z_i, denoted $z_{-i} = \{z_1, \ldots, z_n\} \setminus \{z_i\}$. The model is trained on this set to get $\widehat{m}^{(-i)}$, which represents the model that has been trained excluding the i-th data point.
- **Validation Set:** The left-out data point z_i. The performance is evaluated using $\operatorname{loss}\left(\widehat{m}^{(-i)}\left(z_i\right)\right)$, where $\widehat{m}^{(-i)}\left(z_i\right)$ is the prediction made by the model $\widehat{m}^{(-i)}$ for the input $\boldsymbol{x}_i$ in the data point $z_i = (\boldsymbol{x}_i, y_i)$. And loss function measures the difference between the actual value y_i and the predicted value $\widehat{m}^{(-i)}\left(z_i\right)$.

The average generalization performance:

$$\frac{1}{n}\sum_{i=1}^{n} \operatorname{loss}\left(\widehat{m}^{(-i)}\left(z_i\right)\right)$$

Example.

> Here we apply the LOOCV on Ridge Regression. Let $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ be the candidate set of regularization parameters, and $(\boldsymbol{x}_i, y_i)$ the data for $i = 1, \ldots, n$. For each $m = 1, \ldots, M$, conduct LOOCV:
>
> $$\widehat{\boldsymbol{\beta}}^{(-i,m)} = \arg\min_{\beta} \sum_{j \neq i} \left(y_j - \beta_0 - \beta_1 x_{j,1} - \ldots - \beta_p x_{j,p}\right)^2 + \lambda_m \|\beta\|^2$$
>
> $$MSE^{(i,m)} = \left(y_i - \widehat{\beta}_0^{(-i,m)} - \widehat{\beta}_1^{(-i,m)} x_{i,1} - \ldots - \widehat{\beta}_p^{(-i,m)} x_{i,p}\right)^2$$
>
> Average generalization performance:
>
> $$\frac{1}{n}\sum_{i=1}^{n} MSE^{(i,m)}$$
>
> Choose λ_{m^*} such that:
>
> $$m^* = \arg\min_{m=1,2,\ldots,M} \frac{1}{n}\sum_{i=1}^{n} MSE^{(i,m)}$$

11.3.2 K-Fold Cross-Validation (K-Fold CV)

Let z_i for $i = 1, \ldots, n$ be the data, and assume n is divisible by K. Randomly divide the data into K equal-sized folds: $Z_1, \ldots, Z_K$. For each fold k:

- **Training Set:** All folds except Z_k. The model is trained on $Z_1, \ldots, Z_{k-1}, Z_{k+1}, \ldots, Z_K$ to get $\widehat{m}^{(-k)}$.
- **Validation Set:** The k-th fold Z_k. The performance is evaluated using $\operatorname{loss}\left(\widehat{m}^{(-k)}\left(Z_k\right)\right)$

Average Generalization Performance:

$$\frac{1}{K} \sum_{k=1}^{K} \operatorname{loss}\left(\widehat{m}^{(-k)}\left(Z_k\right)\right)$$

Example.

> Here we perform K-Fold CV for Ridge Regression. Let $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ be the candidate set of regularization parameters, and $(\boldsymbol{x}_i, y_i)$ the data for $i = 1, \ldots, n$. Randomly divide the indices $\{1, 2, \ldots, n\}$ into K folds of sizes $n/K : N_1, \ldots, N_K$. For each $m = 1, \ldots, M$, conduct K -fold CV :
>
> $$\widehat{\boldsymbol{\beta}}^{(-k,m)} = \arg\min_{\beta} \sum_{j \notin N_k} \left(y_j - \beta_0 - \beta_1 x_{j,1} - \ldots - \beta_p x_{j,p}\right)^2 + \lambda_m \|\beta\|^2$$
>
> $$MSE^{(k,m)} = \frac{K}{n} \sum_{i \in N_k} \left(y_i - \widehat{\beta}_0^{(-k,m)} - \widehat{\beta}_1^{(-k,m)} x_{i,1} - \ldots - \widehat{\beta}_p^{(-k,m)} x_{i,p}\right)^2$$
>
> Average generalization performance:
>
> $$\frac{1}{K} \sum_{k=1}^{K} MSE^{(k,m)}$$
>
> Choose λ_{m^*} such that:
>
> $$m^* = \arg \min_{m=1,2,\ldots,M} \frac{1}{K} \sum_{k=1}^{K} MSE^{(k,m)}$$

Remark.

> Often, instead of MSE, the Root Mean Square Error (RMSE) is used to quantify generalization performance:
>
> $$\text{RMSE} = \sqrt{\text{Mean Square Error}}$$

11.4 Gaussian-Markov Theorem

Theorem 11.1: Gaussian-Markov Theorem

Under classical linear regression assumptions, the least squares estimator is the Best Linear Unbiased Estimator (BLUE). Formally, the least squares estimator $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is the BLUE, meaning it has the smallest variance among all unbiased linear estimators.

Definition 11.2: Semi-definite matrix

Definition: A symmetric matrix $\boldsymbol{A}$ is positive semi-definite if for any non-zero vector $\boldsymbol{x}$,

$$\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \geq 0$$

Comparison: For two positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite.

Proof. Here, we will provide the proof outline for this theorem. The expectation of the least squares estimator $\widehat{\boldsymbol{\beta}}$ is equal to the true parameter $\boldsymbol{\beta}$, indicating that it is unbiased:

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

For any linear estimator $\boldsymbol{C}\boldsymbol{y}$, the covariance matrix is:

$$\mathrm{Cov}(\boldsymbol{C}\boldsymbol{y}) = \sigma^2 \boldsymbol{C}\boldsymbol{C}^\top$$

The least squares estimator's covariance matrix is:

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

By introducing an arbitrary matrix $\boldsymbol{D}$, you can express any unbiased estimator as $\boldsymbol{C} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top + \boldsymbol{D}$. The proof shows that the difference between the covariance of any unbiased estimator $\boldsymbol{C}\boldsymbol{y}$ and the least squares estimator $\widehat{\boldsymbol{\beta}}$ is a positive semi-definite matrix:

$$\mathrm{Cov}(\boldsymbol{C}\boldsymbol{y}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{D}\boldsymbol{D}^\top \geq 0$$

This implies that the least squares estimator has the smallest variance among all linear unbiased estimators. ■

While fitting a regression or machine learning model, it is crucial not only to achieve a good fit on the training data but also to ensure that the model generalizes well to unseen data. The Gaussian-Markov theorem guarantees that, under the given assumptions, the least squares estimator is the best linear unbiased estimator, providing a strong foundation for inference in regression models.

To avoid overfitting, penalized regression techniques like Ridge Regression add a penalty term to the loss function. This can be tuned via cross-validation to strike a balance between fitting the data well and preventing the model from becoming too complex.

The fact that the least squares estimator is BLUE gives it a special status in linear regression. This is why, despite the development of more complex models, ordinary least squares (OLS) remains a cornerstone of statistical practice. It is optimal under the conditions specified by the Gaussian-Markov theorem.

12 Beyond this course

As the end of the course, we progress beyond simple and multiple linear regression, it introduces more advanced concepts that extend the regression framework to handle more complex types of data and relationships.

12.1 Scalar-on-Vector Regression

This is the most common form of regression that you've studied, where the response variable y_i is a scalar (a single number) and the predictors $x_{i,j}$ form a vector (a set of variables for each observation).

When working as a data analyst, for example, at a bank or insurance company, you might use this model to predict an outcome like customer risk or credit score based on various covariates (e.g., age, income, account history).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i$$

In matrix notation:

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

12.2 Multivariate Regression (Vector-on-Vector Regression)

This model extends scalar-on-vector regression to multiple response variables. Instead of predicting a single outcome, you predict several outcomes simultaneously. Each outcome has its own set of coefficients.

This could be used if you are interested in predicting multiple financial indicators or health metrics at once, each modeled by a different response vector.

$$\left[\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(q)}\right] = \boldsymbol{X}\left[\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(q)}\right] + \left[\epsilon^{(1)}, \ldots, \epsilon^{(q)}\right]$$

In a compact matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathcal{E}$$

Here, $\boldsymbol{\beta}$ is a matrix with dimensions $(p+1) \times q$, where q is the number of response variables.

12.3 Scalar-on-Matrix Regression

When the predictor is not a vector but a matrix (like an image), and the response is still a scalar, we use scalar-on-matrix regression.

In healthcare, for instance, an analyst might use this model to predict a health outcome (scalar) based on image data like MRI scans (matrix).

$$y_i = \sum_{j,k} X_{j,k}^{(i)} \beta_{j,k} + \epsilon_i = \operatorname{tr}\left(\boldsymbol{X}^\top \boldsymbol{\beta}\right) + \epsilon_i$$

Here, tr denotes the trace operation, which is the sum of the elements on the main diagonal of a matrix.

12.4 Scalar-on-Tensor Regression

This model extends scalar-on-matrix regression to even higher dimensions where the predictor is a tensor (a multi-dimensional array). Tensors generalize matrices to higher dimensions.

For instance, in video data analysis, where each frame of a video could be a matrix, the entire video sequence could be represented as a tensor.

$$y_i = \sum_{j,k,l} X_{j,k,l}^{(i)} \beta_{j,k,l} + \epsilon_i = \langle \mathcal{X}, \beta \rangle + \epsilon_i$$

The symbol $\langle \mathcal{X}, \beta \rangle$ represents the inner product between the tensor $\mathcal{X}$ and the coefficient tensor β.

12.5 Autoregressive Model (AR Model)

This is a special type of regression model used for time series data, where the response variable at time t is modeled as a function of its past values.

This model is particularly useful in finance for modeling stock prices, where the price at a given time depends on its previous prices.

- AR(1) Model (first-order autoregressive model):

$$x_t = \varphi x_{t-1} + \epsilon_t, \text{ for } t = 2, \ldots, T$$

- $\mathrm{AR(p)}$ Model (p-th order autoregressive model):

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + \epsilon_t, \text{ for } t = p+1, \ldots, T$$

Here, ϵ_t is the error term at time t, and φ_i are the coefficients that determine how much past values influence the current value.

12.6 Final words

The basic concepts of regression can be extended in various ways to handle complex types of data and relationships. Whether dealing with simple relationships between scalars or complex interactions between tensors, the principles of regression remain applicable.

Depending on the industry or field you work in, you may encounter and apply these advanced forms of regression. The ability to model relationships involving vectors, matrices, and tensors expands the utility of regression analysis in practical scenarios.

All the examples provided can be generalized to a tensor-on-tensor framework, reflecting the broad applicability and power of regression techniques in modern data analysis. Tensor regression, in particular, is a growing research area with numerous applications in machine learning and data science.

13 Appendix

13.1 Matrix algebra recap

13.1.1 Basic definitions and properties

Definition 13.1: Matrix

A matrix is a rectangular array of elements arranged in rows and columns. The dimension of the matrix is $n \times p$, where n is the number of rows and p is the number of columns.
A matrix with n rows and p columns is usually represented using boldface letters, say $\boldsymbol{A}$, which can be represented as

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,j} & \cdots & A_{1,p} \\ \vdots & \vdots & \cdots & \vdots & & \vdots \\ A_{i,1} & A_{i,2} & \cdots & A_{i,j} & \cdots & A_{i,p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,j} & \cdots & A_{n,p} \end{bmatrix}$$

Remark.

> A can also be written in a compact form,
>
> $$\boldsymbol{A} = (A_{i,j})_{i=1,\cdots,n;j=1,\cdots,p}$$
>
> where $A_{i,j}$ is the element in the i-th row and j-th column.

Definition 13.2: Matrix equality

Two matrices $\boldsymbol{A} = \boldsymbol{B}$, iff all of their corresponding elements are equal; i.e., $A_{i,j} = B_{i,j}$ for all i and j.

Definition 13.3: Square matrix

When $n = p, \boldsymbol{A}$ is a square matrix. And J is a square matrix with all elements equal to one.

Definition 13.4: Column vector and row vector

When $p = 1, \boldsymbol{A}$ is a column vector or simply a vector; when $n = 1, \boldsymbol{A}$ is a row vector. $\mathbf{1}$ is column vector with all elements equal to one. $\mathbf{0}$ is a column vector with all elements equal to zero.

Definition 13.5: Design matrix

A design matrix is a matrix containing data about multiple characteristics of several individuals or objects. Each row corresponds to an individual and each column to a characteristic.

Example.

> If we measure the height and weight of five individuals, we can collect the measurements in a design matrix having five rows and two columns. Each row corresponds to one of the ten individuals, the first column

> contains the height measurements and the second one reports the weights:
>
> $$X = \begin{bmatrix} h_1 & w_1 \\ h_2 & w_2 \\ h_3 & w_3 \\ h_4 & w_4 \\ h_5 & w_5 \end{bmatrix}$$
>
> where h_i denotes the height of the i-th individual and w_i her weight.

Example.

> If we collect the data about the gross domestic product (GDP) of four countries in three consecutive years, then the design matrix is the 4×3 matrix
>
> $$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & X_{43} \end{bmatrix}$$
>
> where, for example, X_{32} is the GDP of the third country in the second year.

Definition 13.6: Matrix transpose and symmetric matrix

For $\boldsymbol{A} = (A_{i,j})$, its transpose is $\boldsymbol{A}^\top = (A_{j,i})_{i=1,\cdots,n;j=1,\cdots,p}$
$\boldsymbol{A}$ is called a symmetric matrix if $\boldsymbol{A} = \boldsymbol{A}^\top$.

Remark.

> $(AB)^T = B^T A^T$

13.1.2 Matrix operations and multiplications

Definition 13.7: Matrix operations

We can perform the following operations on matrix:

1. Element-wise summation and subtraction

$$\boldsymbol{C} = \boldsymbol{A} \pm \boldsymbol{B}; \quad C_{i,j} = A_{i,j} \pm B_{i,j}$$

2. Inner product of $\boldsymbol{x} = (x_1, \cdots, x_n)^\top$ and $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ is

$$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k$$

3. The product of a scalar (a real number) and matrix $\boldsymbol{A}$ is

$$\lambda \boldsymbol{A} = (\lambda A_{i,j})$$

Definition 13.8: Matrix multiplication

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is an $n \times p$ matrix,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

the matrix product $\mathbf{C} = \mathbf{AB}$ (denoted without multiplication signs or dots) is defined to be the $m \times p$ matrix

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

such that

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj},$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, p$

Remark.

> Generally, $\boldsymbol{AB} \neq \boldsymbol{BA}$.

Proposition 13.9

The (i, j)-th element of $\boldsymbol{A}^\top \boldsymbol{A}$ is the inner product of i-th and j-th columns of $\boldsymbol{A}$.

Proof. Let $\boldsymbol{A}$ be an $m \times n$ matrix. The transpose of $\boldsymbol{A}$, denoted $\boldsymbol{A}^\top$, is an $n \times m$ matrix. The product $\boldsymbol{A}^\top \boldsymbol{A}$ is an $n \times n$ matrix. The (i, j)-th element of $\boldsymbol{A}^\top \boldsymbol{A}$ is denoted as $\left(\boldsymbol{A}^\top \boldsymbol{A}\right)_{ij}$. The i-th column of $\boldsymbol{A}$ is denoted as $\boldsymbol{a}_i$.

The (i, j)-th element of the product $\boldsymbol{A}^\top \boldsymbol{A}$ is given by:

$$\left(\boldsymbol{A}^\top \boldsymbol{A}\right)_{ij} = \sum_{k=1}^{m} \left(\boldsymbol{A}^\top\right)_{ik} \cdot \boldsymbol{A}_{kj}$$

The (i, k)-th element of $\boldsymbol{A}^\top$ is the (k, i)-th element of $\boldsymbol{A}$:

$$\left(\boldsymbol{A}^\top\right)_{ik} = \boldsymbol{A}_{ki}$$

Substituting this into the matrix multiplication definition, we get:

$$\left(\boldsymbol{A}^\top \boldsymbol{A}\right)_{ij} = \sum_{k=1}^{m} \boldsymbol{A}_{ki} \cdot \boldsymbol{A}_{kj}$$

The i-th column of $\boldsymbol{A}$, denoted as $\boldsymbol{a}_i$, is:

$$a_i = \begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{mi} \end{pmatrix}$$

Similarly, the $\boldsymbol{j}$-th column of $\boldsymbol{A}$, denoted as $\boldsymbol{a}_j$, is:

$$\boldsymbol{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

The inner product of the i-th and j-th columns of $\boldsymbol{A}$ is:

$$\boldsymbol{a}_i \cdot \boldsymbol{a}_j = \sum_{k=1}^{m} a_{ki} a_{kj}$$

By comparing the sum in the inner product definition with the sum in the matrix multiplication definition, we see that:

$$\left(\boldsymbol{A}^\top \boldsymbol{A}\right)_{ij} = \sum_{k=1}^{m} \boldsymbol{A}_{ki} \boldsymbol{A}_{kj} = \boldsymbol{a}_i \cdot \boldsymbol{a}_j$$

Therefore, the (i, j)-th element of $\boldsymbol{A}^\top \boldsymbol{A}$ is indeed the inner product of the i-th and j-th columns of $\boldsymbol{A}$. This completes the proof. ■

13.1.3 The rank of the matrix

Definition 13.10: Rank of a matrix

The rank of a matrix is the maximum number of linearly independent columns.

Fact 13.11

Here are some useful facts about rank:

1. $\text{Rank}(AB) \leq \min(\text{Rank}(A), \text{Rank}(B))$
2. $\text{Rank}(A) = \text{Rank}\left(A^T\right) \leq \min(n, p)$
3. $\text{Rank}\left(A^T A\right) = \text{Rank}\left(AA^T\right) = \text{Rank}(A)$

Proof. We will prove each fact:

1. Let A be an $m \times n$ matrix and B be an $n \times p$ matrix.

   The rank of a matrix is the dimension of its column space, which is the same as the maximum number of linearly independent columns.

   Consider AB. The column space of AB is a subspace of the column space of A. Therefore, the rank of AB cannot exceed the rank of A. Hence:

   $$\text{Rank}(AB) \leq \text{Rank}(A)$$

   Similarly, the row space of AB is a subspace of the row space of B. Therefore, the rank of AB cannot exceed the rank of B. Hence:

   $$\text{Rank}(AB) \leq \text{Rank}(B)$$

   Combining these results, we get:

   $$\text{Rank}(AB) \leq \min(\text{Rank}(A), \text{Rank}(B))$$

2. Let A be an $m \times n$ matrix.

   The rank of A is the dimension of the column space of A.

   The column space of A and the row space of A^T are the same. Hence:

$$\text{Rank}(A) = \text{Rank}\left(A^T\right)$$

   The maximum number of linearly independent columns of A (or rows of A) cannot exceed the smaller of the number of rows or columns of A.

   Hence, the rank of A is at most $\min(m, n)$. Therefore:

$$\text{Rank}(A) = \text{Rank}\left(A^T\right) \leq \min(m, n)$$

3. Let A be an $m \times n$ matrix.

   $\text{Rank}\left(A^T A\right)$ and $\text{Rank}\left(AA^T\right)$ are determined by the column spaces of these matrices. Note that $A^T A$ is an $n \times n$ matrix, and AA^T is an $m \times m$ matrix. The non-zero eigenvalues of $A^T A$ and AA^T are the same and equal to the singular values of A.

   The rank of $A^T A$ is equal to the number of non-zero singular values of A, which is the same as the rank of A. Hence:

$$\text{Rank}\left(A^T A\right) = \text{Rank}(A)$$

   Similarly, the rank of AA^T is equal to the number of non-zero singular values of A, which is the same as the rank of A. Hence:

$$\text{Rank}\left(AA^T\right) = \text{Rank}(A)$$

   Therefore:

$$\text{Rank}\left(A^T A\right) = \text{Rank}\left(AA^T\right) = \text{Rank}(A)$$

■

13.1.4 Matrix inverse

Definition 13.12: Matrix inverse

The inverse of $\boldsymbol{A}$ is another matrix, denoted by $\boldsymbol{A}^{-1}$, such that

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

$\boldsymbol{A}^{-1}$ exists if $\text{Rank}(\boldsymbol{A})$ equals to the number of rows or columns, when $\boldsymbol{A}$ is said to be invertible, or nonsingular, or of full rank.

Example.

> A simple example:
>
> $$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$
>
> where $D = ad - bc$ is the determinant of $\boldsymbol{A}$.

13.1.5 Traces of a matrix

Definition 13.13: Trace of a matrix

The trace of a square matrix $\mathbf{A}$ is the sum of its diagonal entries, denoted as $\mathrm{tr}(\mathbf{A})$.

Proposition 13.14

The trace of a product of two matrices $\mathbf{A} \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{q\times p}$ is invariant under cyclic permutations:

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$

The expectation of the trace of a random matrix $\mathbf{A}$ is the trace of the expectation:

$$E[\mathrm{tr}(\mathbf{A})] = \mathrm{tr}(E[\mathbf{A}])$$