MAT555E: Statistical Data Analysis for Computational Sciences

Fall22-Lecture 02: Simple Linear Regression

Gül İnan

İstanbul Technical University

Learning Objectives

- Introduction simple linear regression
- Least-squares estimation
- Building a simple linear regression model
- Implementing a simple linear regression model with Python statsmodels library

Simple Linear Regression

Regression

- The **idea of Regression** first appeared around 1886 (Fra Francis Galton: Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 15: 246-263; 1886).

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

Simple linear regression

- Simple linear regression (SLR) has been **around for a long time** and is one of the **widely** used statistical learning methods for predicting a **quantitative response**.
- **Structurally**, SLR may seem to be a bit **easy method** compared to some of the more modern statistical learning approaches.
- However, it serves as a **good jumping-off point** for newer approaches: many fancy statistical learning approaches can be seen as **generalizations** of linear regression.
- For that reason, having a **good understanding** of linear regression before studying more complex learning methods is very **important**.

Simple linear regression

- SLR is a very straightforward approach for predicting a **quantitative response** Y on the basis of a **single predictor variable** X.
  - Does the GPA of a student change with the family's income?
  - Does the number of Covid-19 deaths in a country change with gross domestic product (GDP)?
  - Does number of lung cancer deaths change with number of cigarettes smoked?

Simple linear regression

- SLR assumes that there is **approximately a linear relationship** between the variables Y and X.
- **Mathematically**, we can write this linear relationship as:

$$Y \approx \beta_0 + \beta_1 * X.$$

- We might read "$\approx$" as **is approximately modeled as**.
- The terms β_0 and β_1 refer to the **intercept** and **slope**, respectively, in the **equation of a straight line** (remember the slope-intercept form).

SLR model set-up

- **Statistically**, we assume a **simple linear regression model** where the response variable Y is related to the independent variable X as follows:

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

- The terms β_0 and β_1 refer to the unknown population **intercept** and **slope parameters**, respectively.
- The term ϵ is the **random error** and corresponds to the part of the response variable that **cannot** be explained or predicted by the independent variable (predictor).

SLR model set-up

- Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a **random sample** of **size n** from the SLR model given above.
- Then, at individual data point level, we can write the SLR as:

$$Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n.$$

SLR model assumptions

- In a SLR model, we further assume that the random errors $\epsilon_1, \epsilon_n, \ldots, \epsilon_n$:
  - have **zero mean**: $E(\epsilon_i) = 0$,
  - have a **common variance**: $Var(\epsilon_i) = \sigma^2$,
  - are **independent** of one another: $Cov(\epsilon_i, \epsilon_j) = 0 \quad \text{for} \quad i \neq j$, and
  - are **normally distributed**.

SLR model assumption results

- Then, we can see that:

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \quad \forall i = 1, \dots, n.$$

- This leads to the fact that:

$$Y_i \overset{\text{i.d.}}{\sim} N(\beta_0 + \beta_1 * X_i, \sigma^2) \quad \forall i = 1, \dots, n.$$

- Note that: i.i.d.: **independent and identically distributed** and i.d.: **independent distributed**.

Aim of SLR

- In SLR, our aim is to **estimate** the **unknown population parameters**, namely, β_0 and β_1, for a given sample of data.
- There are **different ways** to estimate the parameters from the sample such as:
  - Least-squares estimation method,
  - Maximum likelihood estimation method, or
  - Gradient descent (an iterative approach). (will see in a few weeks!!!)
- In this class, we will focus on the **least-squares estimation method** for SLR first.

Ordinary least-squares estimation method

Least-squares estimation method

- The method of least-squares estimation is a **standard** and **popular** technique for **parameter estimation** in statistics.
- The method of least-squares was **discovered** in the early 1800's by Gauss and Legendre.
- The **connection** between least-squares estimation and regression appeared in 1897 (George Udny Yule).

Least-squares estimation method

May, 1981

Gauss and the Invention of Least Squares

Stephen M. Stigler

Ann. Statist. 9(3): 465-474 (May, 1981). DOI: 10.1214/aos/1176345451

Abstract

The most famous priority dispute in the history of statistics is that between Gauss and Legendre, over the discovery of the method of least squares. New evidence, both documentary and statistical, is discussed, and an attempt is made to evaluate Gauss's claim. It is argued (though not conclusively) that Gauss probably possessed the method well before Legendre, but that he was unsuccessful in communicating it to his contemporaries. Data on the French meridian arc are presented that could, conceivably, permit a definitive verification of the claim.

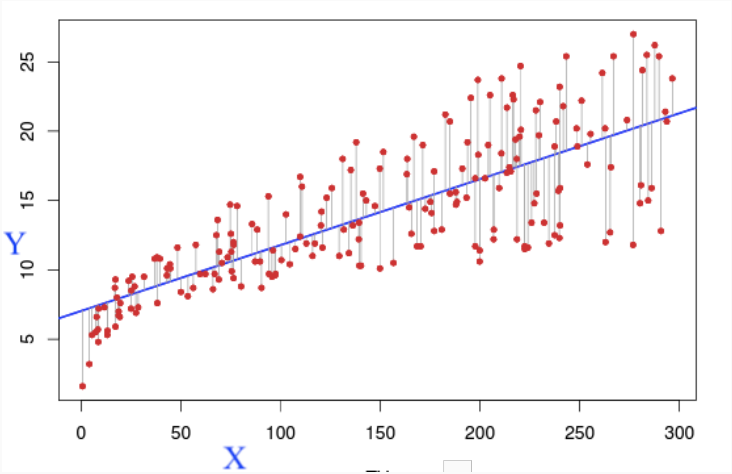
Least-squares estimation method

- Suppose we have given n (observed) samples $\{(x_i, y_i)\}_{i=1}^{n}$ now.
- Our goal is to obtain coefficient estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \in \Omega_\beta$ such that the linear model above fits the available data well-that is $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 * x_i \ (\forall i = 1, \ldots, n)$.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$ be the **predicted value** of y_i through the model.
- Let the **observed difference**:

$$e_i = y_i - \hat{y}_i$$

- be named as the **residual** $(\forall i = 1, \ldots, n)$.

Least-squares estimation method



- The **vertical distance** between a red dot and the blue line gives the residual.

Residual sum-of-squares

- The method of least-squares estimation chooses the coefficient pair $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \in \Omega_\beta$ which **minimizes** the **residual sum-of-squares (RSS)**:

$$RSS(\hat{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 * x_i)\right)^2.$$

- RSS, as a quadratic function, determines an equation for an ellipse in β_0 and β_1 variables. For given values of RSSs, we can draw a counter plot for β_0 and β_1 variables.

Geometric interpretation

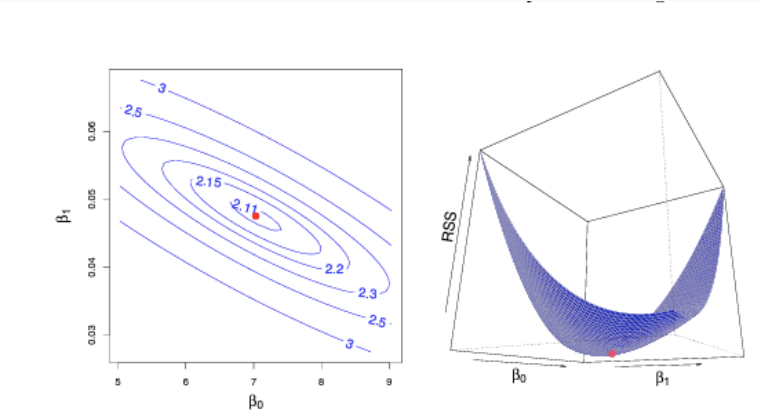


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on*

- The red dots, **minimizing** RSS, correspond to the least squares estimates.

Least-squares estimation method as an optimization

- Then, the **least-squares estimates** of $\widehat{\beta}$ is the solution to the following **optimization** problem:

$$\widehat{\beta}_{OLS} = \underset{\beta_0,\beta_1}{argmin} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\beta_0,\beta_1}{argmin} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2.$$

- Equivalently, in matrix notation:

$$\widehat{\beta}_{OLS} = \underset{\beta_0,\beta_1}{argmin} \left\|\epsilon\right\|_2^2 = \underset{\beta_0,\beta_1}{argmin} \left\|\mathbf{y} - \mathbf{x} * \beta\right\|_2^2.$$

- L_2 norm: $\left\|\epsilon\right\|_2 = \sqrt{\sum_{i=1}^{n} \epsilon_i^2}$.

Finding least-squares estimators

- Let $\mathcal{Q} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2$ be the **objective function** which is **differentiable everywhere** since it is the **sum of quadratic** functions.
- To minimize $\mathcal{Q}$, we take partial derivatives, set them to zero, and solve the resulting system of equations simultaneously, as follows:

$$\left.\frac{\partial \mathcal{Q}}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and}$$

$$\left.\frac{\partial \mathcal{Q}}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 y_i) x_i = 0.$$

Finding least-squares estimators

- Simplifying these two equations yield:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \text{and}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i.$$

- These equations are called **the least squares normal equations**.
- The solution to the normal equations results in **the least squares estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$.

The least-squares estimators

- The **least squares estimators** of the intercept and slope parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, in the simple linear regression model are:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and}$$

- where $\bar{x} = \frac{\sum_i^n x_i}{n}$ and $\bar{y} = \frac{\sum_i^n y_i}{n}$ are the sample means.
- Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ defined above are the **minimizers** of the objective function $\mathcal{Q}$.

On multiplying a differentiable function with a constant

- Note that this function: $\mathcal{Q} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2$ has the **same set of minimizers** as the following problems:

$$\mathcal{Q}_1 = \frac{\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2}{2},$$

$$\mathcal{Q}_2 = \frac{\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2}{n},$$

$$\mathcal{Q}_3 = \frac{\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2}{2n}, \quad \text{and}$$

$$\mathcal{Q}_4 = \frac{\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 * x_i)\right)^2}{2n} + C.$$

- Derivative will still be **zero** at the **same locations**. For mathematical and computational convenience, many Python libraries including scikit-learn uses this trick a lot!!!

Assesing the accuracy of coefficient estimates

The population regression line

- In SLR, we assumed that the **relationship** between Y and X takes the following form:

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

- This equation is also called as **population regression line**.
- Here, the term β_0 is the **intercept term**-that's, the expected value of Y when $X = 0$.
- The term, β_0 is the slope-the **average increase** in Y associated with a one-unit increase in X.
- The error term is a catch-all for **what we miss** in this model: there may be other variables associated with Y, there may be measurement errors etc.

The least-squares line

- The equation below is called **least-squares line**:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i \quad \forall i = 1, \dots, n.$$

- The **least-squares line** gives us the **values predicted by the linear model**.
- Note that here: $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimates.

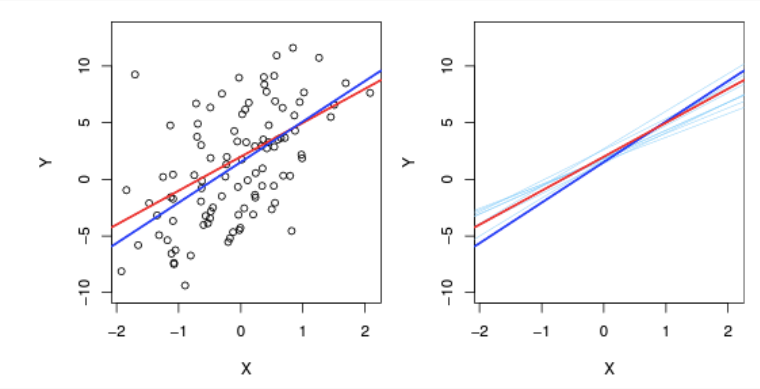
A synthetic data experiment

- When you generate 100 data points from the following model:

$$Y_i = 2 + (3 * X_i) + \epsilon_i \quad \forall i = 1, ..., 100,$$

- with $X_i \sim N(0, 4^2)$ and $\epsilon_i \sim N(0, 6^2)$.
- Note that in real applications, we **do not know the true values** of β_0 and β_1.
- The red line in the left-figure displays the line for the equation for $E(Y_i) = 2 + (3 * X_i)$, while the blue line is the least-squares line based on the observed data.

A synthetic data experiment

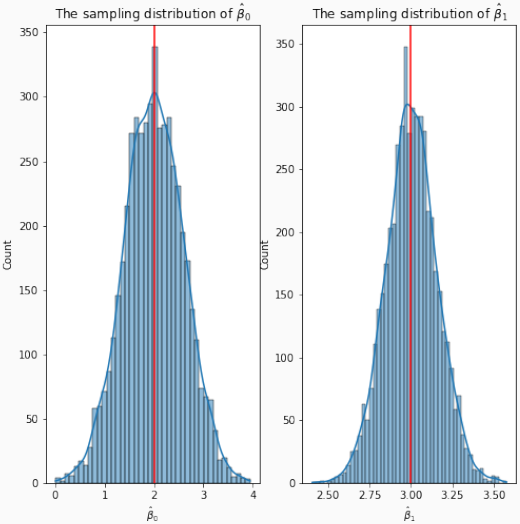


- When we generate 10 different data sets from the equation above, and calculate and plot the least-squares lines, we will see that each least-squares line will be **different** than each other (see right-panel).

A synthetic data experiment

- As in all branches of statistics, in SLR, we use a sample to estimate the characteristics of a **large population**, specifically, we are trying to estimate β_0 and β_1 on the basis of a particular data set.
- Of course, our estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ won't be **exactly equal to** β_0 and β_1.
- But, if we could **average the estimates obtained a huge number of data sets**, then the average of these estimates would be spot on!! (See next figure!!!)
- In fact, we can see from the right-hand figure that the average of many least-squares lines, each estimated from a separate data set, is pretty close to the red line.

A synthetic data experiment



A synthetic data experiment

- A natural follow-up question would be as follows: **How accurate** are the least-square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as an estimate of β_0 and β_1, respectively?
- The averages of $\hat{\beta}_0$ and $\hat{\beta}_1$'s over many data sets will be very **close** to β_0 and β_1, respectively (see previous figure), but how far off will that single estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$ be?
- In general, we can give answer this question by **computing the standard error** of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Roughly speaking, the standard error tells us the average amount that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ differ from β_0 and β_1, respectively.

The sampling distribution of $\hat{\beta}_1$

- The characteristics of $\hat{\beta}_1$ are:
  - Centered at β_1, i.e. $E(\hat{\beta}_1) = \beta_1$.
  - $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
  - $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$.

The sampling distribution of $\widehat{\beta}_0$

- The characteristics of $\widehat{\beta}_0$ are:
  - Centered at β_0, i.e. $E(\widehat{\beta}_0) = \beta_0$.
  - $Var(\widehat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$.
  - $\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})\right)$.

Estimation of σ^2

- The term $Var(\epsilon_i) = \sigma^2$ in the formulas of $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$ is **generally unknown**.
- However, we can **estimate it from the data** through the following formula:

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{RSS}{n-2} = MSR,$$

- RSS: Residual sum of squares.
- MSR: Mean squared residuals which is also **interchangeably** used as MSE: Mean squared error.

Estimated Variance and Standard Error

- When σ^2 is estimated from the data, we should write $\widehat{Var}(\hat{\beta}_0)$, $\widehat{Var}(\hat{\beta}_1)$, $\widehat{se}(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)}$ or $\widehat{se}(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}$ to indicate that an **estimate has been made**.
- But, for simplicity of notation, many textbooks **drop** this **extra** "hat" in standard errors $\widehat{se}(\hat{\beta}_0)$ and $\widehat{se}(\hat{\beta}_1)$ and just write $se(\hat{\beta}_0)$ and $se(\hat{\beta}_1)$.

Hypothesis testing for β_1

- Suppose we wish to test the hypothesis that the **slope parameter equals a constant**, say, $\beta_1 = \beta_{1,0}$.
- The appropriate **hypothesis** is:

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \\ H_1 &: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{1}$$

- Under H_0, the appropriate **test statistic** is:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-2)}.$$

- where $\widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Hypothesis testing for β_1

- The test statistic t_0 measures the number of standard deviations that $\widehat{\beta}_1$ is away from β_1.
- **Decision rule:** We would reject $H_0 : \beta_1 = \beta_{1,0}$ at α significance level, if $t_0 < -t_{1-\alpha/2,n-2}$ or $t_0 > t_{1-\alpha/2,n-2}$.
- The term α refers to the tolerance level for making a Type I error. (e.g., 10%, 5%, 1%).
- Alternatively, a P-value approach could also be used for decision making. We would reject $H_0 : \beta_1 = \beta_{1,0}$ at α significance level, if P-value $< \alpha$.

Hypothesis testing for β_1

- A very important **special case** of the hypothesis above is:

$$H_0 : \beta_1 = 0 \qquad (2)$$
$$H_1 : \beta_1 \neq 0$$

- Failure to reject $H_0 : \beta_1 = 0$ is equivalent to concluding that there is **no linear relationship** between the variables X and Y.

Confidence interval for β_1

- In addition to **point estimate** of the slope parameter, it is possible to obtain **confidence interval estimates**.
- A 100(1-α)% confidence interval on the **slope parameter** β_1 in the simple linear regression is:

$$\left[\hat{\beta}_1 - t_{(1-\alpha/2,n-2)}se(\hat{\beta}_1), \hat{\beta}_1 + t_{(1-\alpha/2,n-2)}se(\hat{\beta}_1)\right].$$

Confidence interval for β_1

- As $\widehat{\sigma^2}$ **decreases** in the formula of $se(\widehat{\beta}_1)$, the **width** of the interval **decreases**.
- As we **decrease** the **confidence level** (1-α), the t-multiplier decreases, and hence the width of the interval decreases.
- As we **increase the sample size n**, the t-multiplier decreases, and hence the width of the interval decreases.

Assesing the accuracy of model

Assesing the accuracy of model

- Once we have **rejected the null hypothesis** in favor of the alternative hypothesis, it is natural to want to quantify **the extent to which the model fits the data**.
- The **quality** of a linear regression fit is typically assessed using two related **quantities**:
  - Residual standard error (RSE) and
  - R^2 statistics.

Residual Standard Error

- Residual standard error (RSE) is an **estimate** of the **standard deviation** of ϵ.
- Roughly, it is the average amount that the model predicted value deviates from the observed value.
- It is computed using the formula:

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

Residual Standard Error

- RSE is considered a measure of the **lack of fit** of the model to the data.
- If the predictions obtained using the model are very close to the true outcome values-that's, if $\hat{y}_i \approx y_i$ for $i = 1, 2, \ldots, n$-then **RSE will be small**, and we can conclude that the model **fits the data very well**.
- On the other hand, if $\hat{y}_i$ is very **far from** y_i for one or more observations, then the RSE may be quite larger, indicating that the model **does not fit the data well**.

Coefficient of determination: R^2

- Let $TTS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ is the **total sum of squares** and measures the **total variance** in the response Y.
- It can also be thought of as the amount of variability inherent in the response before the regression is performed.
- In contrast, RSS measures the **amount of variability that is left unexplained** after performing the regression.
- Hence, TSS-RSS measures the amount of variability in the response that is explained by performing the regression.

Coefficient of determination: R^2

- To calculate R^2, we use the following formula:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{TSS - RSS}{SSTotal} = 1 - \frac{RSS}{TTS}.$$

- R^2 measures the **proportion of variability in Y that can be explained using X**.
- Since it defines a proportion, it always takes on a value **between 0 and 1**.
- An R^2 statistic that is **close to 1** indicates that a large proportion of the variability in the response is **explained by the regression**.
- An R^2 statistic that is **close to 0** indicates that the regression **does not explain** much of the variability in the response.

Relationship between R^2 and r^2

- The R^2 statistic is a measure of the linear relationship between X and Y.
- Recall that $r = Cov(X, Y)$, correlation coefficient, is also a measure of the linear relationship between X and Y.
- In fact, it can be algebraically shown that $R^2 = r^2$ in SLR (only).

References

- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R. New York: Springer.
- https://www2.stat.duke.edu/courses/Fall21/sta521.001/post/week01-1/ (with permission of Prof. Yuansi Chen)
- https://www.cs.ubc.ca/~fwood/CS340/lectures/L12.pdf
- Fan, J., Li, R., Zhang, C.H., and Zou, H. (2020). Statistical Foundations of Data Science. Chapman and Hall/CRC.