

Ch3 - Linear Regression

Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka
Last Updated July 23, 2025

Outline

1. Recap of Joint Distribution, Covariance-Correlation and Scatterplots
2. Problem of Regression and CEF
 - Best Function to Predict
3. Simple Linear Regression Model (SLR)
 - 1. The Problem of Estimation
 - 2. Assessing the Fit - R^2 and RSE

Comments and Acknowledgements

- ▶ These lecture notes have been prepared while I was teaching the course ECO-204: Statistics for Business and Economics II, at East West University, Dhaka (Current Semester - Fall 2023)
- ▶ Most of the contents of these slides are based on
 - ▶ James et al. (2023) and
 - ▶ Anderson et al. (2020)

For theoretical discussion I primarily followed James et al. (2023). Anderson et al. (2020) is a good book and very easy to read with lots of easy examples, but James et al. (2023) is truly amazing when it comes to explaining the concepts in an accessible way. We thank the authors of this book for making everything publicly available at the website <https://www.statlearning.com/>.

- ▶ I thank my students who took this course with me in Summer 2022, Fall 2022 and currently Fall 2023. Their engaging discussions and challenging questions always helped me to improve these notes. I think often I learned more from them than they learned from me, and I always feel truly indebted to them for their support.
- ▶ You are welcome to give me any comments / suggestions regarding these notes. If you find any mistakes, then please let me know at tanvir.hossain@ewubd.edu.
- ▶ I apologize for any unintentional mistakes and all mistakes are mine.

Thanks,
Tanvir

1. Recap of Joint Distribution, Covariance-Correlation and Scatterplots

2. Problem of Regression and CEF

- Best Function to Predict

3. Simple Linear Regression Model (SLR)

- 1. The Problem of Estimation
- 2. Assessing the Fit - R^2 and RSE

Scatter Plot, Covariance and Correlation

- ▶ We already know objects like probability distribution, expectation and variance. So far we have seen only for a single variable cases, both for discrete and continuous random variables. We will now see how to extend these concepts for multiple variables, and how to use them to understand the relationship between two random variables. Important concepts are
 - ▶ *Joint Distribution*,
 - ▶ *Covariance* and *Correlation*.
 - ▶ *Marginal distribution* (related to Marginal Expectation and Marginal Variance)
 - ▶ *Conditional distribution* (related to Conditional Expectation and Conditional Variance)

Scatter Plot, Covariance and Correlation

► Recap of Expectation and Variance Formulas

- Recall for discrete random variable X with probability mass function $f_X(x)$, we have

$$\mathbb{E}(X) = \sum_x x \cdot f_X(x)$$

Suppose if we have X with following probability distribution,

Value of X	Probability $f_X(x)$
1	0.2
2	0.2
3	0.6

Then we can calculate expectation as follows,

$$\mathbb{E}(X) = 1 \cdot 0.2 + 2 \cdot 0.2 + 3 \cdot 0.6 = 2.4$$

- *Ques:* What is the intuition behind the Expectation formula? *Ans:* It gives you population mean without using the population.

Scatter Plot, Covariance and Correlation

- And for variance we have two formulas, the definition is

$$\mathbb{V}(X) = \mathbb{E}[X - \mathbb{E}(X)]^2$$

- We can directly apply this definition, and get

$$\mathbb{V}(X) = (1 - 2.4)^2 \cdot 0.2 + (2 - 2.4)^2 \cdot 0.2 + (3 - 2.4)^2 \cdot 0.6 = 0.64$$

- However there is a shortcut formula for variance (can you derive this?), which is

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

where we can calculate $\mathbb{E}(X^2)$ as follows,

$$\mathbb{E}(X^2) = 1^2 \cdot 0.2 + 2^2 \cdot 0.2 + 3^2 \cdot 0.6 = 6.4$$

Then we can calculate variance as follows, both will give you same result,

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 6.4 - (2.4)^2 = 0.64$$

- What is the intuition behind the Variance formula? *Ans:* It gives you population variance without using the population.

Scatter Plot, Covariance and Correlation

- Now we can start

Scatter Plot, Covariance and Correlation

- Suppose we have following data of 150 students at East West University (EWU) regarding their family income categories and whether they tried to go to abroad for higher studies or not. *For now assume this is the population data, so we have all 150 students in the population*

	Family Income Categories (X)				Total
	Difficult	Middle	Higher Middle	Rich	
Tried	18	13	22	24	77
Not Tried	22	25	16	10	73
Total	40	38	38	34	150

- From here we can easily calculate the joint probability table,

	Family Income Categories (X)				Total
	Difficult	Middle	Higher Middle	Rich	
Tried	0.12	0.08	0.15	0.16	0.51
Not Tried	0.15	0.17	0.10	0.07	0.49
Total	0.27	0.25	0.25	0.23	1

Scatter Plot, Covariance and Correlation

- Here we can X represents Family Income Categories, 1 for Difficult, 2 for Middle, 3 for Higher Middle and 4 for Rich and Y represents tried or not, 1 means the student tried 0 means the student didn't try
- Now we can write following table which is actually called the joint probability distribution of random variables X and Y ,

Tried/Not Tried (Y)	Family Income Categories (X)				Total
	1	2	3	4	
1	0.12	0.08	0.15	0.16	0.51
0	0.15	0.17	0.10	0.07	0.49
Total	0.27	0.25	0.25	0.23	1

Scatter Plot, Covariance and Correlation

- From joint probability distribution we can derive different type of probabilities and probability distributions, also Expectation and Variance.

- **Joint Probability** $\mathbb{P}(X = x, Y = y)$:

For example $\mathbb{P}(X = 1, Y = 0) = 0.15$ means if we randomly select a student from the *population of 150*, then there is a 15% chance that he/she is from Difficult income category and she didn't try to go abroad for higher studies. And all the joint probabilities will sum to 1, i.e.

$\sum_x \sum_y \mathbb{P}(X = x, Y = y) = 1$ and the 8 joint probabilities together is called *joint probability distribution* of X and Y . We will often use $f(x, y)$ to denote the joint probability distribution.

	$f(x, y)$			
	$x = 1$	$x = 2$	$x = 3$	$x = 4$
$y = 1$	0.12	0.08	0.15	0.16
$y = 0$	0.15	0.17	0.10	0.07

Scatter Plot, Covariance and Correlation

► **Marginal Probability** $\mathbb{P}(X = x)$:

This is the probability of X taking a specific value, regardless of the value of Y . For example, $\mathbb{P}(X = 1) = 0.27$ means if we randomly select a student from the *population of 150*, then there is a 27% chance that he/she is from Difficult income category. Similarly, we can find $\mathbb{P}(X = 2) = 0.25$, $\mathbb{P}(X = 3) = 0.25$, and $\mathbb{P}(X = 4) = 0.23$. From here we can calculate the *marginal probability distribution of X* as follows.

$$\mathbb{P}(X = 1) = 0.27, \quad \mathbb{P}(X = 2) = 0.25, \quad \mathbb{P}(X = 3) = 0.25, \quad \mathbb{P}(X = 4) = 0.23$$

We will use $f_X(x)$ to denote the *marginal probability distribution of X* ,

Departments (x)	Probability $f_X(x)$
1	0.27
2	0.25
3	0.25
4	0.23

- And using the marginal probability distribution, we can calculate Marginal Expectation $\mathbb{E}(X)$ and Marginal Variance $\mathbb{V}(X)$ (please do it as an exercise).

Scatter Plot, Covariance and Correlation

► **Marginal Probability** $\mathbb{P}(Y = y)$:

This is the probability of Y taking a specific value, regardless of the value of X . For example, $\mathbb{P}(Y = 1) = 0.51$ means if we randomly select a student from the *population of 150*, then there is a 51% chance that he/she tried to go abroad for higher studies. Similarly, we can find $\mathbb{P}(Y = 0) = 0.49$. From here we can calculate the *marginal probability distribution of Y* as follows,

$$\mathbb{P}(Y = 1) = 0.51, \quad \mathbb{P}(Y = 0) = 0.49$$

We will use $f_Y(y)$ to denote the *marginal probability distribution of Y* ,

Tried/Not Tried (y)	Probability $f_Y(y)$
1	0.51
0	0.49

- And using the marginal probability distribution, we can calculate Marginal Expectation $\mathbb{E}(Y)$ and Marginal Variance $\mathbb{V}(Y)$ (please do it as an exercise).

Scatter Plot, Covariance and Correlation

► Conditional Probability $\mathbb{P}(Y = y \mid X = x)$:

This is something new, this is the probability of Y taking a specific value given that X takes a specific value. For example, $\mathbb{P}(Y = 1 \mid X = 1)$ means if we randomly select a student from the *population of 150* and we know she is from Difficult income category (so we are fixing only for Difficult income category), then what is the probability that he/she tried to go abroad for higher studies. The calculation of conditional probability is straightforward, we can use the joint probability and marginal probability as follows,

$$\mathbb{P}(Y = 1 \mid X = 1) = \frac{\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(X = 1)} = \frac{0.12}{0.27} \approx 0.4444$$

Or using the $f(x, y)$ and $f_X(x)$ we can write it as (the symbol becomes complicated but the calculation is easy)

$$f_{Y|X}(y \mid X = 1) = f_{Y|X}(1 \mid X = 1) = \frac{f(x, y)}{f_X(x)} = \frac{f(1, 1)}{f_X(1)} = \frac{0.12}{0.27} \approx 0.4$$

- In fact conditioning on $X = 1$, we can calculate both $Y = 1$ (which we did) and $Y = 0$ as follows, and then write the conditional distribution of Y given $X = 1$, in a table we can write as follows,

Tried/Not Tried (y)	Probability $f_{Y X}(y \mid X = 1)$
1	0.4
0	0.6

- Note this is conditional distribution of Y given $X = 1$, this is different from marginal distribution of Y which is $f_Y(y)$, which we calculated earlier. And conditional distribution is a distribution so this will sum to 1.

Scatter Plot, Covariance and Correlation

- Now we can also Conditional Expectation $\mathbb{E}(Y \mid X = 1)$ as follows,

$$\begin{aligned}\mathbb{E}(Y \mid X = 1) &= \sum_y y \cdot f_{Y|X}(y \mid X = 1) \\ &= 1 \cdot 0.4 + 0 \cdot 0.6 = 0.4\end{aligned}$$

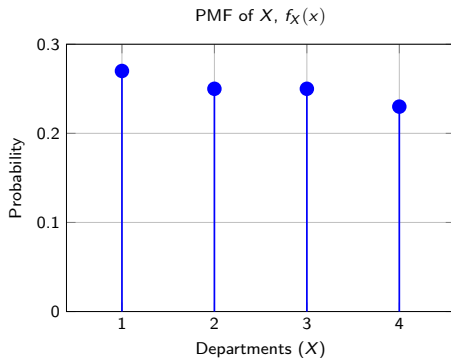
- And Conditional Variance $\mathbb{V}(Y \mid X = 1)$ as follows,

$$\begin{aligned}\mathbb{V}(Y \mid X = 1) &= \mathbb{E}[Y - \mathbb{E}(Y \mid X = 1)]^2 \\ &= (1 - 0.4)^2 \cdot 0.4 + (0 - 0.4)^2 \cdot 0.6 = 0.24\end{aligned}$$

- In this case you can think about conditional expectation as a population average of all Y values given $X = x$ (for example $X = 1$). Similar interpretation can be given for conditional variance.
- From this joint distribution we can calculate 4 conditional distribution of Y , given four possible values of X , i.e. $X = 1, 2, 3, 4$. This will give us 4 conditional mean and 4 conditional variance.
- Similarly we can also calculate two conditional distributions of X given $Y = 1$ and $Y = 0$, and then calculate conditional expectation and conditional variance.

Scatter Plot, Covariance and Correlation

- Here an example plot for marginal PMF of X



- Here is an example plot for joint PMF of X and Y , where X is Departments and Y is Tried or Not Tried.

Scatter Plot, Covariance and Correlation

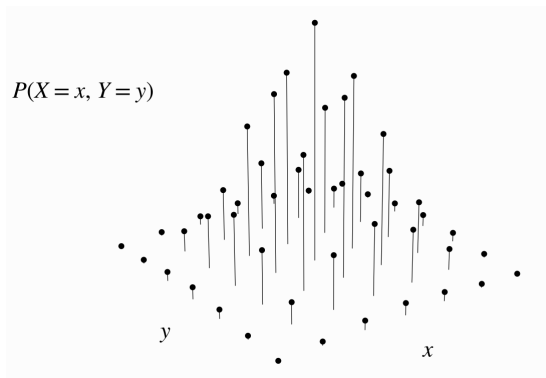


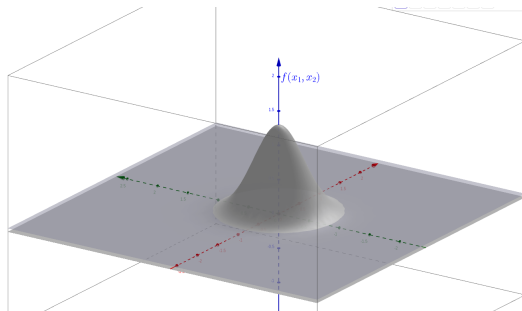
Figure 1: Figure above shows a sketch of what the joint PMF of two discrete random variables could look like. The height of a vertical bar at (x, y) represents the probability $\mathbb{P}(X = x, Y = y)$ or $f(x, y)$. For the joint PMF to be valid, the total height of the vertical bars must be 1 .

Scatter Plot, Covariance and Correlation

- We only looked at discrete random variables, but we can also extend this to continuous random variables. For example, if X and Y are two continuous random variables, then we can define joint probability density function (PDF) $f(x, y)$ such that

$$\mathbb{P}(X \in A, Y \in B) = \iint_{A \times B} f(x, y) dx dy$$

- The things become more complicated when we have continuous random variables, but the idea is similar. We can define marginal PDF $f_X(x)$ and $f_Y(y)$, and then we can define conditional PDF $f_{Y|X}(y | x)$ as follows,
- Here is an example of *bi-variate Normal or jointly Normal*,



Scatter Plot, Covariance and Correlation

- The functions looks a bit more scary, sorry,

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times e^{\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right] \right\}}$$

- Here we have two random variables, X and Y which are jointly normal. Now we have 5 parameters, μ_X , μ_Y , σ_X , σ_Y and ρ . here μ_X and μ_Y are the means of X and Y , σ_X and σ_Y are the standard deviations of X and Y , and ρ is the correlation between X and Y .

Scatter Plot, Covariance and Correlation

- ▶ For continuous random variables we can also define marginal PDF $f_X(x)$ and $f_Y(y)$, and then we can define conditional PDF $f_{Y|X}(y | x)$ with integration, I won't go to details here but important is here everything will be a function of x and y . I give one example below
- ▶ **Joint PDF of X and Y** is given by

$$f(x, y) = x + \frac{3}{2}y^2, \quad 0 < x < 1, \quad 0 < y < 1$$

- ▶ In this case from this joint just by integrating we can find **marginal PDF of X and Y as follows**,

$$f_X(x) = x + \frac{1}{2}$$

$$f_Y(y) = \frac{3}{2}y^2$$

- ▶ We can also calculate **conditional PDF of Y given X** as follows,

$$\begin{aligned} f_{Y|X}(y | X = x) &= \frac{f(x, y)}{f_X(x)} = \frac{x + \frac{3}{2}y^2}{x + \frac{1}{2}} \\ &= \frac{2x + 3y^2}{2x + 1} \end{aligned}$$

Scatter Plot, Covariance and Correlation

- Notice for each fixed x , this is a density function of y , so this is a conditional PDF of Y given $X = x$. For example, if $x = \frac{1}{2}$, then we can write the conditional PDF of Y given $X = \frac{1}{2}$ as follows,

$$f_{Y|X}(y | X = \frac{1}{2}) = \frac{2 \cdot \frac{1}{2} + 3y^2}{2 \cdot \frac{1}{2} + 1} = \frac{1 + 3y^2}{2}$$

- If we use $f_{Y|X}(y | X = x) = \frac{2x+3y^2}{2x+1}$ and calculate expectation of Y given $X = x$, then we can write as follows,

$$\mathbb{E}(Y | X = x) = \frac{1}{2(2x+1)} \left(x + \frac{3}{4} \right)$$

- Note that conditional expectation becomes a function of X . This is called conditional expectation function. How do we visualize this, there is a nice way to visualize this in scatter plot. We will come back to this later, however important is conditional expectation is a function of X , so we can write $\mathbb{E}(Y | X) = g(X)$, where $g(X)$ is a function of X .

Scatter Plot, Covariance and Correlation

- ▶ Before we end this section we will learn about two other quantities, which are very important in statistics, these are *Covariance* and *Correlation*. Probably you already know the sample covariance and sample correlation, but here we will learn about population covariance and population correlation. These two quantities will help us to understand the relationship between two random variables.
- ▶ Here is the formula or definition

Scatter Plot, Covariance and Correlation

Definition 3.1: Covariance and Correlation)

The population covariance between two random variables X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

And the Correlation between two random variables X and Y is

$$\rho_{X,Y} = \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{(\sqrt{\text{Var}(X)}) (\sqrt{\text{Var}(Y)})} = \frac{\text{Cov}(X, Y)}{\sigma_X \times \sigma_Y}$$

- ▶ where μ_X and μ_Y are the marginal Expected values of X and Y , and σ_X and σ_Y are the standard deviations of X and Y .
- ▶ **What does covariance mean?** If covariance is positive, then X and Y are *positively associated or related*, which roughly means if X increases, then Y also increases. If covariance is negative, then X and Y are *negatively associated / related*, which roughly means if X increases, then Y decreases. If covariance is close to 0, then there is almost no relationship between X and Y .
- ▶ Now **What does correlation mean?** Correlation is a normalized version of covariance, which means it gives a value between -1 and 1 (we will always have $-1 \leq \rho_{X,Y} \leq 1$). So it's a better measure of association than covariance, since we can understand the strength of association between X and Y from correlation.

Scatter Plot, Covariance and Correlation

- In particular if $\rho_{X,Y}$ is close to $+1$, then X and Y are *positively correlated*, which means if X increases, then Y also increases. If $\rho_{X,Y}$ is close to -1 , then X and Y are perfectly negatively correlated, which means if X increases, then Y decreases. If $\rho_{X,Y} = 0$, then there is no linear relationship between X and Y .

Scatter Plot, Covariance and Correlation

- ▶ Let's calculate covariance and correlation for our example of EWU students. We can use the joint distribution of X and Y that we calculated earlier, and then we can calculate the covariance and correlation as follows,
- ▶ But before that there is also a shortcut formula for covariance, which is (this is easy to derive, please do it as an exercise)

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

- ▶ Here for $\mathbb{E}(XY)$, we need the joint distribution of X and Y , which we can calculate as follows,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_x \sum_y x \cdot y \cdot f(x, y) \\ &= 1 \cdot 1 \cdot 0.12 + 1 \cdot 2 \cdot 0.08 + 1 \cdot 3 \cdot 0.15 + 1 \cdot 4 \cdot 0.16 + 2 \cdot 1 \cdot 0.15 + 2 \cdot 2 \cdot 0.17 + 2 \cdot 3 \cdot 0.10 \\ &= 0.12 + 0.16 + 0.45 + 0.64 + 0.30 + 0.34 + 0.60 + 0.28 \\ &= 2.95\end{aligned}$$

- ▶ We need to calculate $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, which we can calculate from the marginal distributions of X and Y that we calculated earlier,

$$\mathbb{E}(X) = 1 \cdot 0.27 + 2 \cdot 0.25 + 3 \cdot 0.25 + 4 \cdot 0.23 = 2.45$$

$$\mathbb{E}(Y) = 1 \cdot 0.51 + 0 \cdot 0.49 = 0.51$$

Scatter Plot, Covariance and Correlation

- Now we can calculate covariance as follows,

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y) \\ &= 2.95 - 2.45 \cdot 0.51 \\ &= 2.95 - 1.25 = 1.70\end{aligned}$$

- Now calculate the standard deviations of X and Y , which we can calculate from the marginal distributions of X and Y that we calculated earlier and then calculate correlation (do it as an exercise),
- Since Covariance is positive, we can say X and Y are positively associated, which means if a student is from higher income categories, then he/she is more likely to try to go abroad for higher studies.
- How strong is the relationship between Y and X ? We can use the correlation between Y and X to measure this (please do it as an exercise).
- What we learned is population covariance and correlation. There is also sample covariance and sample correlation from a sample data, the formulas are,

$$\begin{aligned}s_{X,Y} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ r_{X,Y} &= \frac{s_{X,Y}}{s_X s_Y}\end{aligned}$$

- where $s_{X,Y}$ is the sample covariance, $r_{X,Y}$ is the sample correlation, s_X and s_Y are the sample standard deviations of X and Y , and \bar{x} and \bar{y} are the sample means of X and Y .

Scatter Plot, Covariance and Correlation

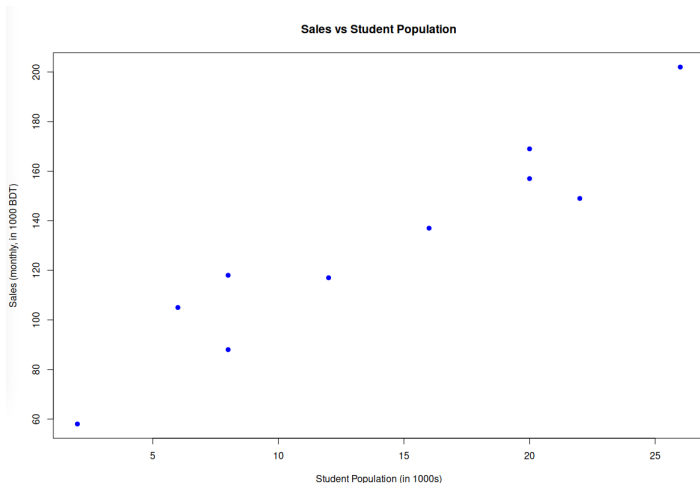
- There is a connection of covariance and correlation with scatter plot, what is a scatter plot? A scatter plot is a graphical representation of the relationship between two variables, where each point represents an observation in the dataset. The horizontal axis represents one variable (say X) and the vertical axis represents another variable (say Y).
- For example here suppose we collected a dataset from 10 restaurants asking about their *student population size* (what is approximate number of students live close to them) and *monthly sales*. We can think about the population size as x_i and monthly sales as y_i . Here is the data,

Restaurant	SPOP (in 1000s) - x_i	Msales (in 1000 BDT) - y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Table 1: Two Variable Data for SLR, here Independent Variable is SPOP and Dependent Variable is Msales

Scatter Plot, Covariance and Correlation

- With this sample we can plot a scatter plot, where we can see the relationship between x_i and y_i as follows,



Scatter Plot, Covariance and Correlation

- ▶ Roughly this shows that there is a positive relationship between x_i and y_i , which means if the student population size increases, then the monthly sales seems to increase.
- ▶ We can now calculate the covariance and correlation between X and Y as follows, which should be also positive, since we can see the positive relationship in the scatter plot.

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 315.5$$

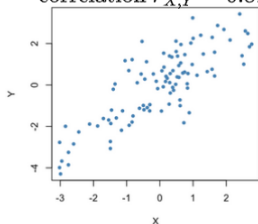
- ▶ Which doesn't give us how strong is the relationship. We can also calculate the correlation, which is

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} = 0.95$$

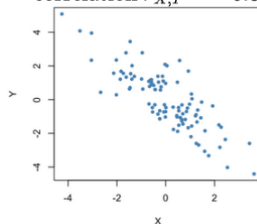
- ▶ Which shows the strong relationship between X and Y , which is also visible in the scatter plot.
- ▶ So scatterplot is a graphical representation of the relationship between two variables, and covariance and correlation are numerical measures of the strength and direction of that relationship.
- ▶ You should always remember the following picture,

Scatter Plot, Covariance and Correlation

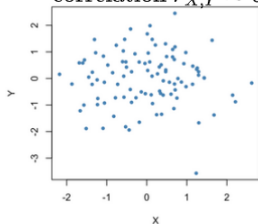
covariance $s_{X,Y} > 0$,
correlation $r_{X,Y} = 0.81$



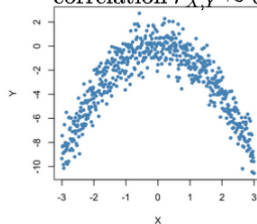
covariance $s_{X,Y} < 0$,
correlation $r_{X,Y} = -0.81$



covariance $s_{X,Y} \approx 0$,
correlation $r_{X,Y} \approx 0$



covariance $s_{X,Y} \approx 0$,
correlation $r_{X,Y} \approx 0$



1. Recap of Joint Distribution, Covariance-Correlation and Scatterplots

2. Problem of Regression and CEF

- Best Function to Predict

3. Simple Linear Regression Model (SLR)

- 1. The Problem of Estimation
- 2. Assessing the Fit - R^2 and RSE

Problem of Regression and CEF

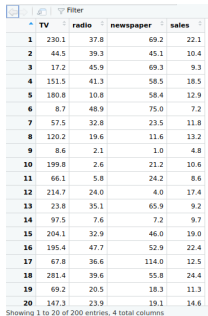
Problem of Regression and CEF

Best Function to Predict

Best Function to Predict

Conditional Expectation Function

- Suppose we have a data set of a company's sales and money spent on TV, radio and newspaper advertisement. Here is how the data looks like in R studio



	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1.0	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24.0	4.0	17.4
13	23.8	35.1	65.9	9.2
14	97.5	7.6	7.2	9.7
15	204.1	32.9	46.0	19.0
16	195.4	47.7	52.9	22.4
17	67.8	36.6	114.0	12.5
18	281.4	39.6	55.8	24.4
19	69.2	20.5	18.3	11.3
20	147.3	23.9	19.1	14.6

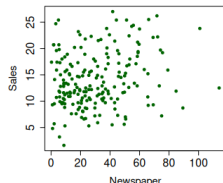
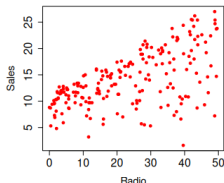
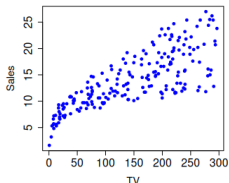
Showing 1 to 20 of 200 entries, 4 total columns

- It shows we have 200 observations (so sample size is 200), 20 of them is shown and we have 4 variables.
- The units are an important part of the data “Sales” variable is in 1000 unit and other variables are in 1000\$.
- Now suppose the company wants to *predict the sales* based on the other three variables.
- Doing some descriptive statistics is often a good idea before we go for inferential statistics.

Best Function to Predict

Conditional Expectation Function

- In this case we can see following *scatter plots* which shows some *association* between sales and each of the variables (what about causality?). Recall scatter plot is a graphical method to see association between two variables (what are some numerical methods to check association? Ans: Covariance and Correlation)



- We will see how to do scatter plots in our lab session.

Best Function to Predict

Conditional Expectation Function

- ▶ Back to prediction problem.
- ▶ Here Sales is called the *response or target* that we wish to predict with the help of *TV, Radio and Newspaper*.
- ▶ The target variable is often represented by Y and other variables that we will use to predict are often represented by X (if we have single variable) or X_1, X_2, X_3, \dots , (if we have multiple variables).
- ▶ Sometimes we also call Y as *dependent variable* and X or X_1, X_2, X_3 as *independent variables* or *explanatory variables or regressors or features or predictors or covariates*.

Best Function to Predict

Conditional Expectation Function

- **Question:** How do we solve the prediction problem? Answer is we need a function $f(X_1, X_2, X_3)$, which is following,

$$f(X_1, X_2, X_3) = 3 + 4X_1 + 5X_2 + 6X_3$$

- Assuming the function does a good job for our prediction problem. Then we use this function to predict Y
- For example if we know $X_1 = 10$, $X_2 = 20$ and $X_3 = 30$, then we can predict the sales as follows,

$$\begin{aligned}\text{predicted } Y &= f(X_1, X_2, X_3) = 3 + 4(10) + 5(20) + 6(30) \\ &= 3 + 40 + 100 + 180 \\ &= 323\end{aligned}$$

- Of course our prediction will not be 100% accurate since we may have measurement errors or leave other variables in our model, and there will be a *True Sales or True Y* at this combination of X_1, X_2, X_3 , which we will not be able to predict exactly. So we will have some *error* in our prediction.
- We will denote the *error* or *residual* or *prediction error* with ϵ , and we can write it as,

$$\epsilon = Y - f(X_1, X_2, X_3)$$

Best Function to Predict

Conditional Expectation Function

- ▶ In this chapter our goal is to find such function f that will help us to predict Y as accurately as possible ... this is called the regression problem.
- ▶ **From now first we will focus on a single variable case** which is called *simple linear regression problem* so we will assume X is a single variable, say TV expenditure, and then we will extend it to multiple variables later. So now we can write the model as,
- ▶ Note that if we have only one variable, then we can write the function as,

$$\text{predicted } Y = f(X) = 3 + 4X$$

Best Function to Predict

Conditional Expectation Function

- ▶ Now the question is **what is the best function f that we can use to predict Y ?**
- ▶ Here we need to be clear about *what do we mean by “best”?*.
- ▶ Here we will assume “best” means we mean minimizing the *mean squared error* (in short MSE).
- ▶ MSE is defined as

$$\mathbb{E} [(Y - f(X))^2]$$

- ▶ So now we can rephrase the question -

“is there a function f that will minimize MSE or $\mathbb{E} [(Y - f(X))^2]$, if YES, then what is the function?”

- ▶ The question can be also stated mathematically as an optimization problem,

$$\underset{f}{\text{minimize}} \quad \mathbb{E} [(Y - f(X))^2]$$

Best Function to Predict

Conditional Expectation Function

- ▶ I won't show the calculation here mathematically (but you can look into Hansen (2022) if you want to see the proof), but the answer is YES, there is a function and the function is the *conditional expectation function*, which we write as,

$$f(X) = \mathbb{E}(Y \mid X)$$

- ▶ Or when we write as a function of X , we can write as

$$f(x) = \mathbb{E}(Y \mid X = x)$$

- ▶ You already know conditional expectation (which is the average of Y values given a fixed X), the question is what is conditional expectation function?

Best Function to Predict

Conditional Expectation Function

- The idea is this is a function of X , where when we plug the value of X , we get the conditional expectation of Y given that value of X . For example it could be when both X (single variable) and Y are continuous random variables, then the conditional expectation function is

$$f(x) = 2 + 3x^2$$

- Here is how we can visualize this function in a scatterplot, suppose we have population of Y and X values, maybe lots of values,

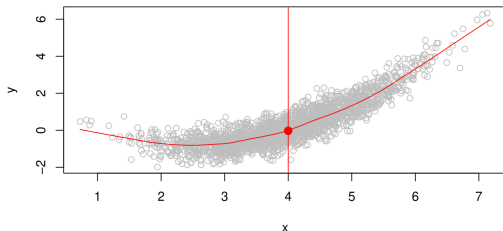


Figure 2: This is a scatter plot of population data of Y and X . The red line is the conditional expectation function, which is a function of X , at 4 the dot shows the conditional expectation of Y given $X = 4$, which is $E(Y | X = 4)$

Best Function to Predict

Conditional Expectation Function

- We can calculate the conditional expectation function for all X values, and then we can connect the points which gives us the conditional expectation function which is the red line in the picture and which is going to be a function of x , which we can write with $f(x)$.

Best Function to Predict

Conditional Expectation Function

- ▶ Why CEF could be useful?
- ▶ Two key reasons
 - ▶ *Prediction* - With a good f we can make predictions of Y at new points $X = x$. In this case we are not interested to know the true f per se, but if we can do good predictions we are happy.
 - ▶ *Inference regarding the function and related objects* - Prediction is one kind of inference, but there is another kind, where we want to infer about the true CEF. Maybe we are interested to understand the true nature of the relationships between the response and predictors, or which predictors are important in explaining the response. Sometimes this is more difficult and often we have no hope without imposing strong assumptions.

Understanding CEF and CEF ϵ

CEF and CEF Error - Mean and Variance

- ▶ We need to mention some important points regarding conditional expectation function and the CEF error ϵ .
- ▶ Conditional expectation always follow following properties,

$$\text{LIE: } \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

- ▶ This is called *law of iterated expectation* (LIE), which says the average of conditional expectation is equal to unconditional expectation. There are other properties of conditional expectation, but we will not go into details here.
- ▶ Now we come to Error, recall Error is defined as

$$\epsilon = Y - f(X) = Y - \mathbb{E}(Y|X)$$

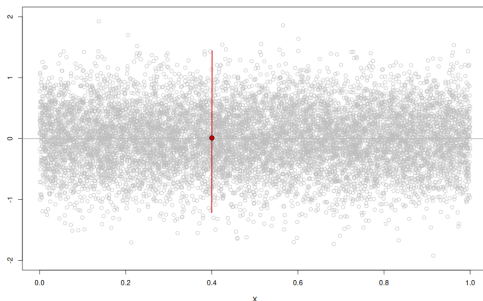
- ▶ We can easily see that

$$\mathbb{E}(\epsilon|X) = \mathbb{E}(Y|X) - \mathbb{E}(Y|X) = 0$$

- ▶ What does this mean visually? Consider following population data of ϵ

Understanding CEF and CEF ϵ

CEF and CEF Error - Mean and Variance



- ▶ Here we plotted x values on the x -axis and ϵ values on the y -axis. So for each x value, we have many ϵ values and the figure shows if we take average of these ϵ values at every x , then the average will be 0 at every x .
- ▶ If $\mathbb{E}(\epsilon | X = x) = 0$, then with LIE we know that $\mathbb{E}(\epsilon) = 0$ (this is an application of law of iterated expectation)
- ▶ We can also think about conditional variance of Y which is $\mathbb{V}(Y | X = x)$ and conditional variance of ϵ which is $\mathbb{V}(\epsilon | X = x)$. Using the definition of variance we can show that

$$\mathbb{V}(\epsilon | X = x) = \mathbb{V}(Y | X = x) = \mathbb{E}((Y - \mathbb{E}(Y | X = x))^2 | X = x)$$

Understanding CEF and CEF ϵ

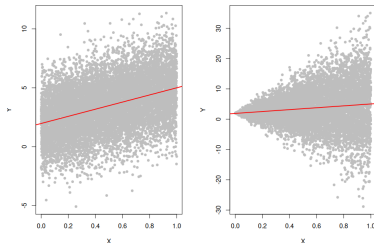
CEF and CEF Error - Mean and Variance

- Which means conditional variance of ϵ is equal to conditional variance of Y given $X = x$.

Understanding CEF and CEF ϵ

CEF and CEF Error - Mean and Variance

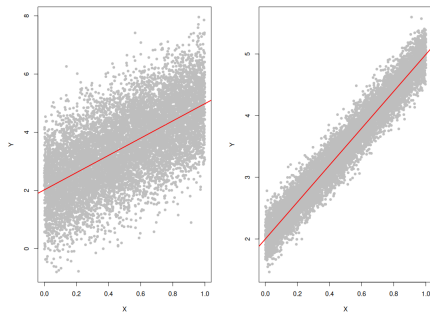
- So what is conditional variance? It means variance of Y , conditional on x values. In the following we plotted two *population data* where the red line is the CEF function.



- On the left the variance of Y seems to be constant with x values, so this means $\mathbb{V}(Y|X)$ or $\mathbb{V}(\epsilon|X)$ is constant. This is called *homoskedasticity*!
- On the right the variance of Y is changing with x values (in particular increasing), so this means $\mathbb{V}(\epsilon|X)$ is NOT constant, it is called *heteroskedasticity*!
- We can also show that unconditional variance are also equal $\mathbb{V}(\epsilon) = \mathbb{V}(Y)$
- Now again consider two population data, for both $\mathbb{V}(\epsilon|X = x)$ is constant. But on the left $\mathbb{V}(\epsilon|X = x)$ is high and on the right $\mathbb{V}(\epsilon|X = x)$ is low

Understanding CEF and CEF ϵ

CEF and CEF Error - Mean and Variance



- It's important to note that, if the conditional variance is high then unconditional variance $\mathbb{V}(\epsilon)$ is also high.
- If we have homoskedasticity for ϵ , which means constant conditional variance of ϵ , then it is possible to show that $\mathbb{V}(\epsilon) = \mathbb{V}(\epsilon|X = x)$

1. Recap of Joint Distribution, Covariance-Correlation and Scatterplots

2. Problem of Regression and CEF

- Best Function to Predict

3. Simple Linear Regression Model (SLR)

- 1. The Problem of Estimation
- 2. Assessing the Fit - R^2 and RSE

Simple Linear Regression Model (SLR)

Simple Linear Regression Model (SLR)

1. The Problem of Estimation

Simple Linear Regression

The Problem of Estimation (method of least squares)

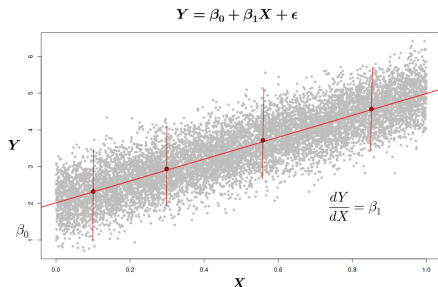
- ▶ The method we will learn first is known as *Linear Regression Model*. In particular we will talk about *Simple Linear Regression Model* or in short SLR in this chapter.
- ▶ According to Simple Linear Regression Model we will assume the unknown CEF is linear in parameters (constants) and also in X and we have just one feature X .
- ▶ This means we will assume the true and unknown CEF $f(X)$ has following form,

$$f(X) = \beta_0 + \beta_1 X$$

- ▶ This actually means, our true CEF looks like the red line in the following figure where β_0 is the intercept and β_1 is the slope (Notice here we are assuming following is the scatter plot of some population data)

Simple Linear Regression

The Problem of Estimation (method of least squares)



- In this case we can write the model with the error as

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

- Now, if we want to know the best prediction function CEF here, our job is to *only get the values of unknown β_0 and β_1* , then we can use CEF to predict Y for any values of X.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ It's obvious that just from the sample we can never get β_0 and β_1 , so what do we do? We try to guess the values from a sample data. You should immediately recognize this an *estimation problem*.
- ▶ We explain the estimation method here with a concrete example from Anderson et al. (2020). Keep in mind, our goal is to estimate the unknown β_0 and β_0 using a sample.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- Suppose Armand's Pizza Parlors is a chain of Italian-food restaurants located in a five-state area. Armand's most successful locations are near college campuses. The managers believe that quarterly sales for these restaurants (denoted by y) are related positively to the size of the student population (denoted by x); that is, restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population. Using regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x . Here is how the data looks like

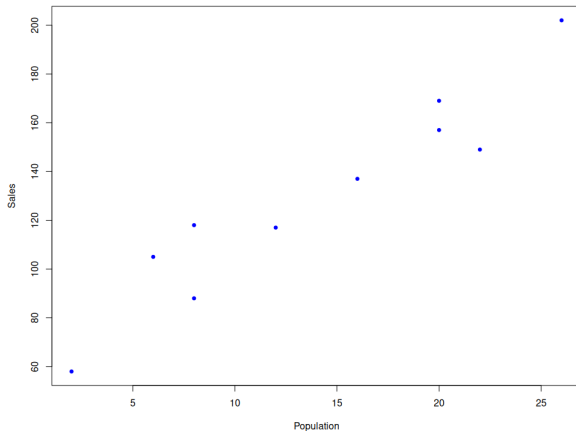
Restaurant	Population (x_i), in 1000s	Sales (y_i), in 1000\$
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

- Notice, there is a big difference between the sample data and population data. Sample data only have very few samples, in this case only 10. Here is the scatterplot of the sample data,

Simple Linear Regression


The Problem of Estimation (method of least squares)

Figure 3: Scatterplot of Armand's Pizza Parlor data from Anderson et al. (2020)



Simple Linear Regression

The Problem of Estimation (method of least squares)

- Using this data we can estimate of β_0 and β_1 . Following  command will give us the result

code: SLR results for the Armands data

```
# load the library to read the excel file
library(readxl)
armands <- read_excel("Armand's.xlsx") # load the data

# fit the model with the data
slr_fit <- lm(Sales ~ Population, data = armands)

## see the output
options(scipen = 999) # turn off scientific printing
summary(slr_fit)
```

- You should see following output,

Simple Linear Regression

The Problem of Estimation (method of least squares)

Call:

```
lm(formula = Sales ~ Population, data = armands)
```

Residuals:

Min	1Q	Median	3Q	Max
-21.00	-9.75	-3.00	11.25	18.00

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	60.0000	9.2260	6.503	0.000187 ***
Population	5.0000	0.5803	8.617	0.0000255 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.83 on 8 degrees of freedom

Multiple R-squared: 0.9027, Adjusted R-squared: 0.8906

F-statistic: 74.25 on 1 and 8 DF, p-value: 0.00002549

Simple Linear Regression

The Problem of Estimation (method of least squares)

- We can see a bit formatted output using the `stargazer` package.

Table 2: Regression results of Quarterly Sales on Population

	<i>Dependent variable:</i>
	Quarterly Sales (in 1000s)
Student Population (in 1000s)	5*** (0.580)
Constant	60*** (9.226)
Observations	10
R ²	0.903
Residual Std. Error	13.829 (df = 8)
F Statistic	74.248*** (df = 1; 8)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

- We will often write the estimates with hat symbol. Here the estimate of β_0 is $\hat{\beta}_0$ and the estimate of β_1 is $\hat{\beta}_1$. Using the data we found $\hat{\beta}_0 = 60$ and $\hat{\beta}_1 = 5$.

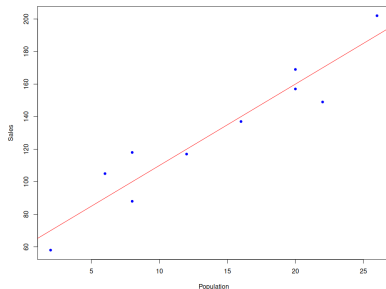
Simple Linear Regression

The Problem of Estimation (method of least squares)

- Using this we can write equation of *estimated line or the fitted line*, which is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 60 + 5x_i$$

- We can plot the fitted line with the data, this is the red line in the following figure.
- After plotting the scatter plot, you can plot this line using the `abline()` function.



- Just to make it clear, note the intercept of this line is $\hat{\beta}_0 = 60$ and the slope is $\hat{\beta}_1 = 5$, so the equation of this line is $\hat{y}_i = 60 + 5x_i$, we call this *the best fitted line with this data*.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ Question is - Why the name *best fitted line*, what is the meaning of “best” or *how did we calculate 5 and 60*? Let's explain this now
- ▶ Essentially here best means minimizing *sum of squared errors* or SSE in the sample. What is SSE?

- ▶ If we think $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the estimated line, then the error (also called residual) is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- ▶ the squared error is

$$e_i^2 = (y_i - \hat{y}_i)^2 = \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

- ▶ And *sum of squared errors*, in short SSE is

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

- ▶ So now we can write the problem clearly, *our problem is we need to find a line which minimizes SSE*

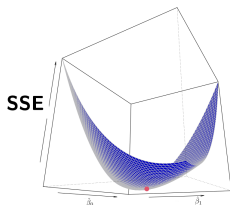
Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ Since we are fitting linear line, this means we need to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that SSE is minimized.
- ▶ We can write this as a following optimization (in particular minimization) problem

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{minimize}} \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

- ▶ Actually this is a multivariate minimization problem where we need to minimize the SSE function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Following picture might be useful to think what's happening here

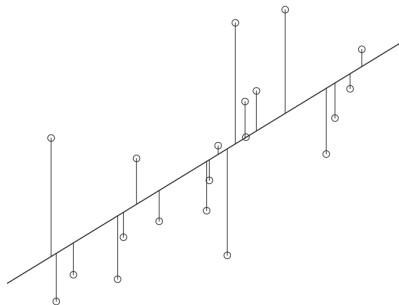


- ▶ Why? Because you can think SSE is a function of $\hat{\beta}_0$ and $\hat{\beta}_1$ and we are looking for the optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ which will minimize this function.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- So far we explain the algebraic way of understanding the problem, which is related to the optimization problem, there is another way to think about what's happening,
- Here the vertical lines are the errors, e_1, e_2, \dots, e_n and we are essentially minimizing the sum of the squared of these errors.



Simple Linear Regression

The Problem of Estimation (method of least squares)

- So following is the minimization problem

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{minimize}} \sum_{i=1}^n \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2$$

- I will skip the details here (you will see more details in the Econometrics course and I will try to give you some additional notes), but if we solve this minimization problem (this means taking derivatives, setting the equations to 0 and then solving), the optimal coefficients are

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

- We gave * to represent that these are optimal points, usually we will omit *.
- There is another way we can write $\hat{\beta}_1$, which is using the sample covariance and variance formulas

$$\widehat{\text{cov}}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \text{sample covariance} \quad (1)$$

$$s_X^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{sample variance} \quad (2)$$

where s_X^2 is the sample variance of X , so we can write $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(x, y)}{s_X^2}$

Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ Recall, there is a difference between sample variance and population variance.
- ▶ This method is famously known as *method of least-squares* and the fitted line is called the *least squares line* (often also called *estimated regression line* also *sample regression function*).
- ▶ Finally we write the formula for the estimated coefficients again

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3)$$

- ▶ Now let's interpret the coefficients. Recall the estimated equation is

$$\hat{y}_i = 60 + 5x_i$$

- ▶ We can also write the estimated equation with the original variable names, rather than x and y ,

$$\widehat{\text{sales}} = 60 + (5 \times \text{population})$$

- ▶ What is the interpretation of the estimated coefficients?
- ▶ For the interpretation the units are very important. Recall, the data of the student population is in 1000s, and the data of sales is in 1000\$

Simple Linear Regression


The Problem of Estimation (method of least squares)

- ▶ What is the interpretation of the the *slope co-efficient* $\hat{\beta}_1 = 5$? Here is how we can interpret,
if the student population is increased by 1000, then approximately the sales is predicted to increase by 5000\$ units.
- ▶ or
An additional increase of 1000 student population is associated with approximately 5000\$ units of additional sales.
- ▶ What is the interpretation of the the *intercept co-efficient* $\hat{\beta}_0$? if the student population is 0, then the predicted sales is 60,000\$. This kind of interpretation for intercept often doesn't make any sense unless we come up with a story. So we often avoid interpreting the intercept co-efficient.
- ▶ Using the estimated regression line we can also get in-sample predicted values, these are also sometimes called *fitted values*.
- ▶ We can calculate the fitted values using the estimated regression equation,
 $\hat{y}_i = 60 + (5 \times x_i)$.

Simple Linear Regression

The Problem of Estimation (method of least squares)

	Population in 1000s	Sales (in 1000\$)	Fitted Values (in 1000\$)
1	2	58	$60 + (5 \times 2) = 70$
2	6	105	$60 + (5 \times 6) = 90$
3	8	88	$60 + (5 \times 8) = 100$
4	8	118	$60 + (5 \times 8) = 100$
5	12	117	$60 + (5 \times 12) = 120$
6	16	137	$60 + (5 \times 16) = 140$
7	20	157	$60 + (5 \times 20) = 160$
8	20	169	$60 + (5 \times 20) = 160$
9	22	149	$60 + (5 \times 22) = 170$
10	26	202	$60 + (5 \times 26) = 190$

- ▶ in  you can get the fitted values with the command `fitted(model_result)`
- ▶ These fitted values are within the sample data points, so this is why we call this *in-sample prediction*. But we can also do *out-of-sample prediction*.

Simple Linear Regression

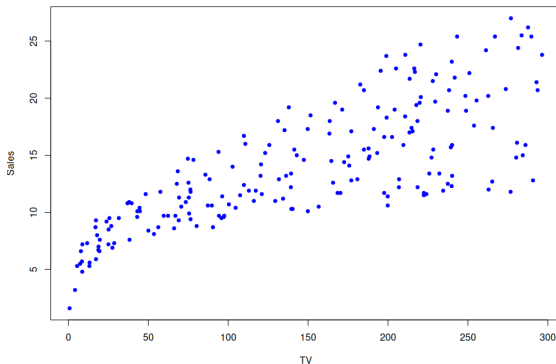
The Problem of Estimation (method of least squares)

- ▶ For example we can also use the estimated regression line to predict the sales when the population is 30 thousands (notice 30 is not in the sample, nor in the range).
- ▶ If we do this we get $60 + (5 \times 30) = 210$ 1000\$ sales.
- ▶ So this is a predicted value for which we don't know y_i .
- ▶ We need to be careful on out of sample prediction, if the fit is good, our estimated function will always give very good in-sample prediction, but it does not automatically mean for an unseen data we will also get good out-of-sample prediction.
- ▶ There is a way we can evaluate out-of-sample prediction, using *training and test sample*. We will see this in our lab session.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ Let's do another example, recall the advertisement example from page 7.
- ▶ Suppose we want to use the TV expenditure (in 1000\$) variable to predict sales (in 1000 unit). We already have seen the scatter plot, but here is it again



Simple Linear Regression

The Problem of Estimation (method of least squares)


- If we fit the best fitted line with the data, we get following results

code - Regression Results for Advertisement Data - SLM

```
# load the library, load the data
library(readxl)
advdata <- read_excel("Advertising.xlsx")

# fit the model
slr_result <- lm(sales ~ TV, data = advdata)

# see the results
summary(slr_result)
```

- Following is the output, you should see this in your console. Note that in this case we can also use directly the `read.csv()` function. Also use `options(scipen = 999)` to turn off scientific printing in .

Simple Linear Regression

The Problem of Estimation (method of least squares)

Call:

```
lm(formula = sales ~ TV, data = advdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.3860	-1.9545	-0.1913	2.0671	7.2124

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.032594	0.457843	15.36	<0.0000000000000002 ***
TV	0.047537	0.002691	17.67	<0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom

Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099

F-statistic: 312.1 on 1 and 198 DF, p-value: < 0.00000000000000022

Simple Linear Regression

The Problem of Estimation (method of least squares)

We can also use the stargazer library to get a little bit organized output (you need to install the library first)

code: Regression Results for Advertisement Data - SLM

```
library(stargazer)
stargazer(slm_result, type = "text")
```

Table 3: Regression Results for Sales and TV Expenditure

	<i>Dependent variable:</i>
	Sales (in 1000s)
TV Expenditure (in 1000\$)	0.048*** (0.003)
Constant	7.033*** (0.458)
Observations	200
R ²	0.612
Residual Std. Error	3.259 (df = 198)
F Statistic	312.145*** (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

Simple Linear Regression

The Problem of Estimation (method of least squares)

► So $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.048$.

► Let's interpret $\hat{\beta}_1 = 0.048$,

if we increase the spending on TV advertisement by 1000\$, then approximately the sales is predicted to increase by 48 units.

► or

1000\$ additional spending on TV advertisement, is associated with approximately 48 units of additional sales.

► Note that, this is a prediction type statement (not a causal statement), so we cannot say *the sales will increase by*, we can only say *the sales is predicted to increase by*, or *the increased sales is associated with*.

Simple Linear Regression Model (SLR)

2. Assessing the Fit - R^2 and RSE

Simple Linear Regression

Goodness of Fit or R^2

- So far we are fitting a line? Question is - *how good does the linear line fit with the data?*
- Actually there is a quantity / summary measure R^2 , which answers this question for us. The formula for R^2 is

$$R^2 = \frac{SSR}{SST}$$

- where

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2, \text{ Total Sum of Squares}$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2, \text{ Error Sum of Squares}$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \text{ Regression Sum of Squares}$$

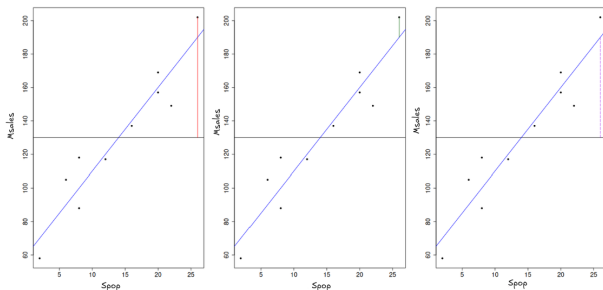
- Also recall $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Question is - what does this formula mean? To understand this first let's decompose $y_i - \bar{y}$

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

- This can be visually understood

Simple Linear Regression

Goodness of Fit or R^2



- ▶ On the left for an i th point, we have $y_i - \bar{y}$, then on the middle we have a residual or error $(y_i - \hat{y}_i)$ and on the right we have $(\hat{y}_i - \bar{y})$
- ▶ Now we can take squares and sum on both sides of the decomposition and we get (the product term becomes 0)

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SSR}}$$

Simple Linear Regression

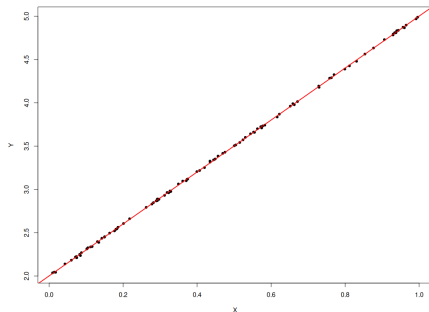
Goodness of Fit or R^2

- ▶ We mentioned SST stands for *Total Sum of Squares*. This is easy to explain. Recall, the total variability of y_i can be explained by the sample variance $\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$. And for SST we have the numerator of the sample variance of y_i . So SST measures the total variability of y_i (but it's not exactly variance).
- ▶ We already know SSE, which is $\sum_{i=1}^n (y_i - \hat{y}_i)^2$. This is the sum of squared errors, or the *Error Sum of Squares* which shows how much variability of error remains after we fitted the line.
- ▶ And the term $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is called *Regression Sum of Squares* or SSR in short, which shows how much variability of y_i is explained by the regression or can be explained by x_i .
- ▶ So this means R^2 tells “*out of the total variation of y how much we can explain by regression*”.

Simple Linear Regression

Goodness of Fit or R^2

- ▶ Also note R^2 is a ratio of explained sum of squares and total sum of squares. So this means we will always have $0 \leq R^2 \leq 1$ (in other words the value of R^2 will always lie between 0 and 1).
- ▶ So high R^2 means the least-squares line fits very well with the data. Here is an example of high R^2 with a different data

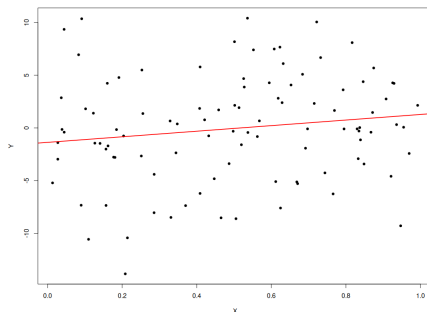


- ▶ The black dots are the sample points, the red line is the fitted line. Here R^2 is 0.99.

Simple Linear Regression

Goodness of Fit or R^2

- Now suppose we have a data which does not show any linear pattern and we try to fit a linear line, obviously the fit won't be good, and R^2 will be low, for example consider following data



- If we fit a line (which is the red line), then R^2 in this case is 0.02, which is almost close to 0.

Simple Linear Regression

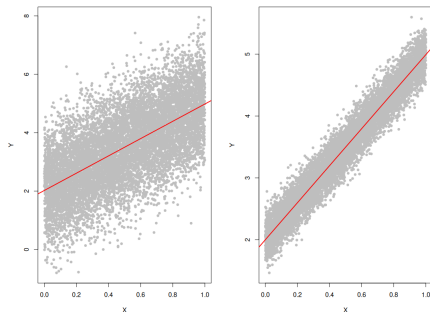
Goodness of Fit or R^2

- ▶ So the above discussion shows R^2 tells us how well our least-squares line fits the data. High R^2 means the fit is quite good, on the other hand low R^2 means fit is not that good with the data.
- ▶ R^2 is also known as *Coefficient of Determination*, sometimes it is also called *Goodness of Fit*.
- ▶ In the Sales Vs. TV example, we found $R^2 = .612$ (page 43), this means *almost 61.2% variability of y can be explained by the estimated regression line*.
- ▶ There is an important caveat regarding R^2 , that is high R^2 does not automatically mean that we did a good job with our prediction problem.
- ▶ There are always issues with out of sample prediction [on board discussion, watch the recorded class]
- ▶ But still we can say high R^2 is something that is generally desirable.

Simple Linear Regression

Residual Standard Error

- ▶ Let's think about the variance of ϵ again. For now assume homoskedasticity, which means $\mathbb{V}(\epsilon) = \mathbb{V}(\epsilon|X = x) = \sigma^2$, where σ^2 is some constant. So we can think about the unconditional variance $\mathbb{V}(\epsilon)$
- ▶ In the following we plotted same figure we plotted before.
- ▶ Recall on the left $\mathbb{V}(\epsilon)$ is high and on the right $\mathbb{V}(\epsilon)$ is low



- ▶ It's important to understand that high variance of ϵ indicates our lack of certainty in prediction. Why? Because ϵ is the error that remains after we do prediction using CEF. So if there is a lot of noise, even if we use CEF, we won't be able to predict well.

Simple Linear Regression

Residual Standard Error

- ▶ Now note that ϵ is not observable, so we cannot calculate its variance σ^2 or standard deviation σ , but using the estimated residuals we can get an *estimate of the standard deviation* of ϵ .
- ▶ Here is an estimate, it's called MSE,

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

- ▶ The last equality holds because we can show that $\bar{e} = 0$ (you can check this with the data!)
- ▶ Since this is an estimate of the variance of ϵ , we can take square root of this and get an estimate of the standard deviation of ϵ , which is called *Residual Standard Error* or *Standard Error of the Estimate*.

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

- ▶ So this gives an estimate of σ . If this is high we may conclude our uncertainty of prediction is high. If this is low, this is good for our prediction.
- ▶ So just to clearly mention again, for a fixed sample, MSE is an estimate of σ^2 and $\sqrt{\text{MSE}} = \text{RSE}$ is an estimate of σ .

Simple Linear Regression

Residual Standard Error

- ▶ In the case of the advertising data, we see from the linear regression output in page 7 that the RSE is 3.26. How to interpret this?
 - ▶ One way to interpret this is - sales deviate from the regression line by approximately 3,260 units, on average.
 - ▶ Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.
- ▶ Is this an acceptable error? this depends on the problem context. In the advertising data set, the mean value of sales is approximately 14,000 units, and so the percentage error is $3,260/14,000 = 23$, so you might conclude the error is quite large.
- ▶ So high RSE is considered *a lack of fit*.

- Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J. and Ohlmann, J. W. (2020), *Statistics for Business & Economics*, 14th edn, Cengage, Boston, MA.
- Hansen, B. (2022), *Econometrics*, Princeton University Press, Princeton.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2023), *An introduction to statistical learning*, Vol. 112, Springer.