Ch3 - Simple Linear Regression

Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka

Last Updated August 20, 2025

Outline

Outline

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

Scatter Plot, Covariance and Correlation

- We already know objects like probability distriution, 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 ....

- 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)				
	Difficult	Middle	Higher Middle	Rich	Total
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)				
	Difficult	Middle	Higher Middle	Rich	Total
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

- 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,

	Family Income Categories (X)				
Tried/Not Tried (Y)	1	2	3	4	Total
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

- 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

- **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).

- **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).

- **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.

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

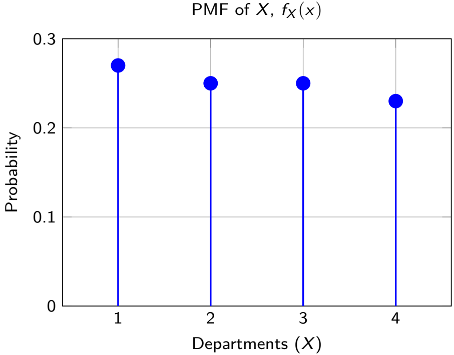
$$\begin{aligned}\mathbb{E}(Y \mid X = 1) &= \sum_y y \cdot f_{Y|X}(y \mid X = 1) \\ &= 1 \cdot 0.4 + 0 \cdot 0.6 = 0.4\end{aligned}$$

- 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 $X = 1$). Similat 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.
- Similatly we can also calculate two conditional distributions of X given $Y = 1$ and $Y = 0$, and then calculate conditional expectation and conditional variance.

- Here an example plot for marginal PMF of X



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

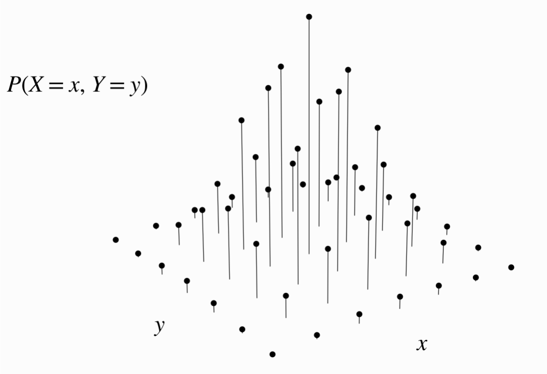


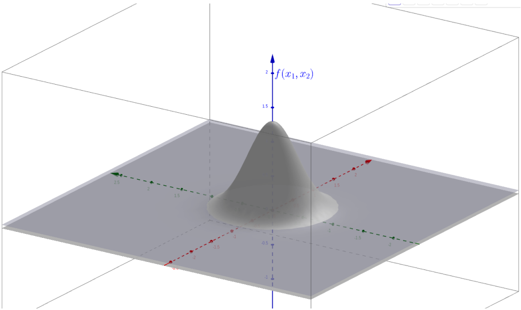
Figure 1: Figure above shows a sketch of what the joint PMF of two discrete random variables could look like. The height of a vertical bar at (x, y) represents the probability $\mathbb{P}(X = x, Y = y)$ or $f(x, y)$. For the joint PMF to be valid, the total height of the vertical bars must be 1 .

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 $f_{Y|X}(y \mid x)$ as follows,
- Here is an example of *bi-variate Normal or jointly Normal*,



- 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 $f_{Y|X}(y \mid x)$ with integration, I won't go to details here but impotant 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,

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

- 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,

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

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

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

- 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 $\mathbb{E}(Y \mid X) = g(X)$, where $g(X)$ is a function of X.

- 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

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

And the Correlation between two random variables X and Y is

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

- 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.

- 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.

- 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,

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y x \cdot y \cdot f(x, y) \\
&= 1 \cdot 1 \cdot 0.12 + 1 \cdot 0 \cdot 0.15 + 2 \cdot 1 \cdot 0.08 + 2 \cdot 0 \cdot 0.17 + 3 \cdot 1 \cdot 0.15 + 3 \cdot 0 \cdot 0.10 \\
&+ 4 \cdot 1 \cdot 0.16 + 4 \cdot 0 \cdot 0.07 \\
&= 0.12 + 0 + 0.16 + 0 + 0.45 + 0 + 0.64 + 0 \\
&= 2.95
\end{aligned}$$

- 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,

$$\begin{aligned}
\mathbb{E}(X) &= 1 \cdot 0.27 + 2 \cdot 0.25 + 3 \cdot 0.25 + 4 \cdot 0.23 = 2.45 \\
\mathbb{E}(Y) &= 1 \cdot 0.51 + 0 \cdot 0.49 = 0.51
\end{aligned}$$

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 coorelation. 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.

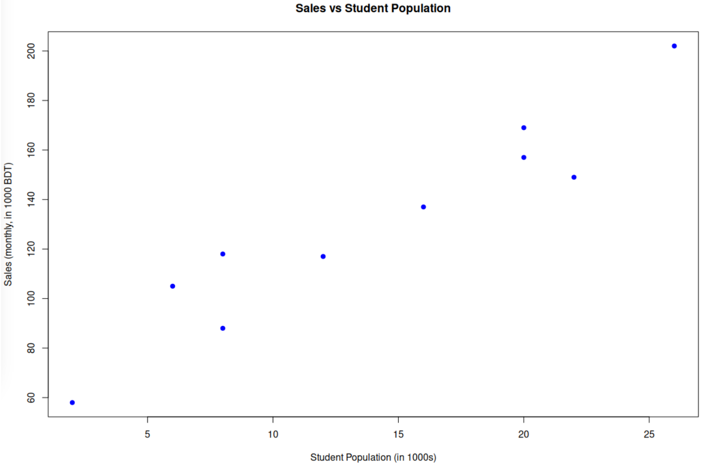
Scatter Plot, Covariance and Correlation

- 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. 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

- 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.

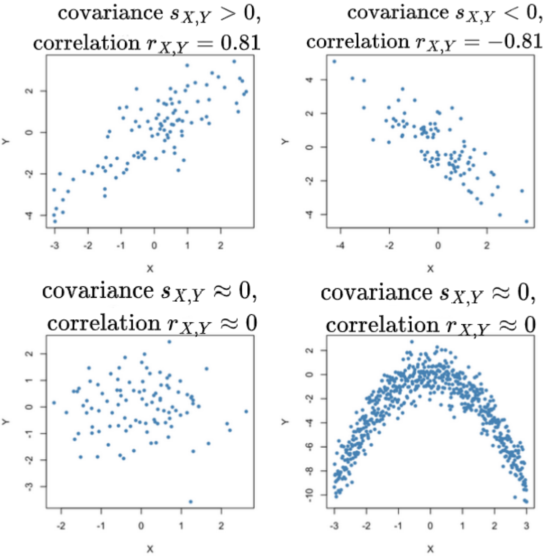
$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 315.5$$

- 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 pciture,

Scatter Plot, Covariance and Correlation



Problem of Regression and CEF

Problem of Regression and CEF

Best Function to Predict

- 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 Rstudio

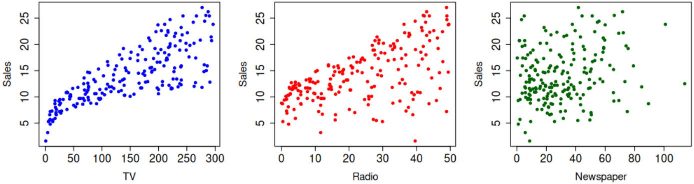
Filter

	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.

- 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.

- Back to predicition 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 $X_1, X_2, X_3, \ldots,$ (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*.

- **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 predicition 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,

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

- 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,

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

- 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

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

- So now we can rephrase the question -

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

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

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

- 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?

- 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, ...

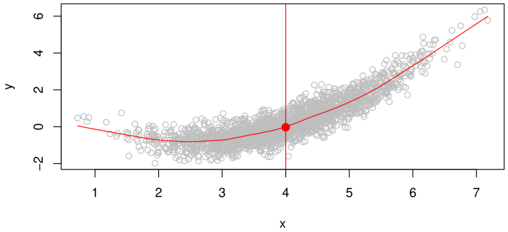


Figure 2: This is a scatter plot of population data of Y and X. The red line is the conditional expectation function, which is a function of X, at 4 the dot shows the conditional expectation of Y given $X = 4$, which is $\mathbb{E}(Y \mid X = 4)$

- 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 $f(x)$.

- 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.

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

$$\textbf{LIE:} \quad \mathbb{E}\left(\mathbb{E}(Y|X)\right) = \mathbb{E}(Y)$$

- 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

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

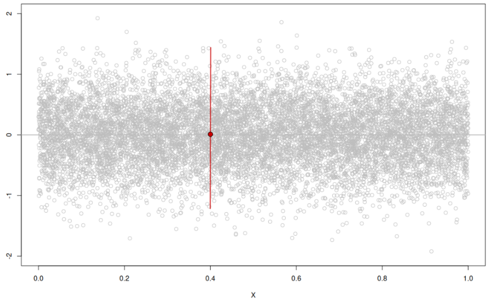
- We can easily see that

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

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

Understanding CEF and CEF ϵ

CEF and CEF Error - Mean and Variance

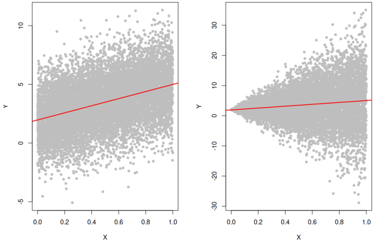


- Here we plotted x values on the x-axis and ϵ values on the y-axis. So for each x value, we have many ϵ values and the figure shows if we take average of these ϵ values at every x, then the average will be 0 at every x.
- If $\mathbb{E}(\epsilon|X = x) = 0$, then with LIE we know that $\mathbb{E}(\epsilon)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

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

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

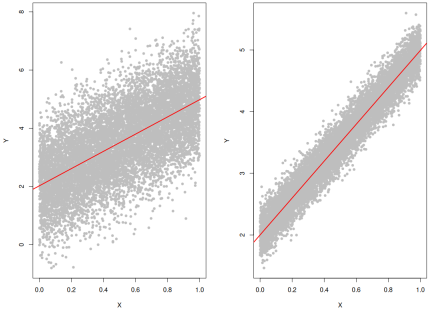
- So what is conditional variance? It means variance of Y, conditional on x values. In the following we plotted two *population data* where the red line is the CEF function.



- One 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 $\mathbb{V}(\epsilon) = \mathbb{V}(\epsilon|X = x)$

Simple Linear Regression Model (SLR)

Simple Linear Regression Model (SLR)

1. The Problem of Estimation

- 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 (slope and intercept) and also linear in* X and we have just one feature X.
- Let's exaplain this in detail.

- It's helpful to always keep a data example at the back of your mind, so we will use the following example. Assume we have *only one independent variable* which is **Student Population (SPop)** in 1000s and a depedent variable which is **Monthly Sales (Msales)** in 1000 BDT. We want to predict Msales based on SPop.

- We will write the data from the independent variable with x_i, so $x_1, x_2, \ldots, x_n$ and dependent variable or response variable with y_i, so $y_1, y_2, \ldots, y_n$, so a pair with (x_i, y_i) is a data point. So we can write the data as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where n is the sample size.

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 2: Two Variable Data for SLR, here Independent Variable is SPop and Dependent Variable is Msales

- This was one sample.... let's see what we assume in the population ....

- In the population we assume we have

$$\mathbb{E}(Y_i|X_i = x) = f(x_i) = \beta_0 + \beta_1 x$$

- This actually means, our true CEF looks like the red linear line 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)

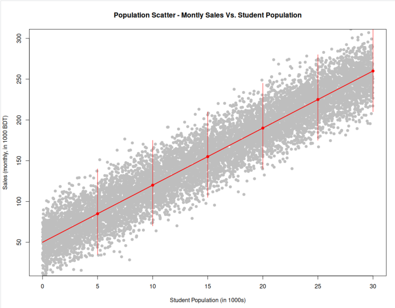


Figure 3: Scatter plot of the Population Data (gray points) the Conditional Expectation Function (red line)

- Now, if we want use the best prediction function to predict Y for any given values of x 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. It's obvious that just from the sample we can never get β_0 and β_1, since these are population quantities.... so what do we do? We try to guess the values from a sample data. You should immediately recognize this an *estimation problem*.

- Essentially our goal is to find the following red line - which can be called *the best fitted linear line*

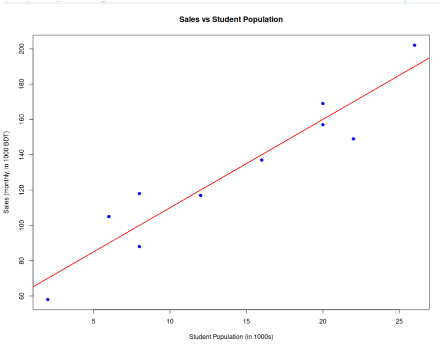


Figure 4: Scatterplot of Sales Vs. Student Population

- The equation of the line will be something like this

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

- Here the $\widehat{y_i}$ is used for predicted value and $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the unknown *intercept* and *slope* of the linear line ... note that if we know the intercept and slope we have our magical equation to predict ...
- Following R command will give us the result
- You can also get the similar output in Excel, we will see this in class.

R code: SLR results for the Armands data

```
# set the directory
setwd("...")

# turn off scientific printing
options(scipen = 100)

# get the data
Fast_Food_Data_SLR <- read_excel("Fast_Food_Data_SLR.xlsx")


# fit the model with the data
model <- lm(Msales ~ Spop, data = Fast_Food_Data_SLR)
summary(model)
```

- You should see following output,

```
Call:
lm(formula = Msales ~ Spop, data = Fast_Food_Data_SLR)

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 ***
Spop         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
```

- Here intercept $\widehat{\beta}_0 = 60$ and slope $\widehat{\beta}_1 = 5$

- So finally we can write the equation of the *best fitted line*,

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

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

Simple Linear Regression Model (SLR)

2. Interpretations

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

$$\widehat{y}_i = 60 + 5\,x_i$$

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

$$\widehat{\text{Monthly Sales}} = 60 + (5 \times \text{Student Population})$$

- The "hat" symbol is for predicted values (note it's not actual y_i)
- Let's see the interpretations,

Interpretation of $\widehat{\beta}_1 = 5$

- The slope co-efficient $\widehat{\beta}_1$ is the *predicted change in the dependent variable* (here monthly sales) for a unit change in the independent variable (here student population). So we can say - *if the student population is increased by 1000, then approximately monthly sales is predicted to increase by 5000 taka. Or we can also say an additional increase of 1000 student population is associated with approximately 5000 taka of additional sales.*

- Notice for the interpretation *the units are very important*. Here the student population is in 1000s, and the data of monthly sales is in 1000 taka, so we need to be careful when interpreting the coefficients. Also it must not be a causal interpretation, we cannot say - *change in student population causes change in sales*... so careful with the wordings...

- **Interpretation of intercept $\widehat{\beta}_0 = 60$**
- if the student population is 0, then the predicted sales is $60,000$ taka. This kind of interpretation for intercept often doesn't make any sense unless we come up with a story, so perhaps we can say - *if there is no student population, then the sales is still* $60,000$ *taka, this might be because of some other factors*.

Simple Linear Regression Model (SLR)

3. The Least Squares Problem

- Now a 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,

- Essentially here "best" means here - it's a line which has least error in some sense, in particular, here we are minimizing *the sum of squared errors* or in short *SSE* in the sample. So this line has the least SSE. What is SSE?

- First let's explain what is the error here, the idea of the error in this case is,

$$\text{error} \;= \text{actual} - \text{predicted}$$

- So if e_i is the error for the i_{th} data point, then using our notation this means

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

- and since our predicted value is $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, this means

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

- the squared error is

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

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

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

- So no we can write the problem clearly, *our problem is we need to find a line which minimizes SSE*, in particular we have the following minimization problem,

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

- In words this means, we need to *find the $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that the sum of squared errors is minimized*.

- I will skip the details here (some details are in the Appendix, if you have taken Mat 211, then you can understand it easily, otherwise you will see more in the Econometrics course),.... but if we solve the minimization problem we get,

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

- There is another way we can write $\widehat{\beta}_1$,which is using he sample covariance and variance formulas, recall

$$s_{x,y} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1} \quad \text{sample covariance} \tag{1}$$

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

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

- 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*).

Simple Linear Regression Model (SLR)

4. In-Sample and Out-of-Sample Predictions

- Using the estimated regression line we can also get *in-sample predicted* values, these are also sometimes called *fitted values*. These are essentially predicted values for the sample data points.... Manually we can calculate the fitted values using the estimated regression equation, $\hat{y}_i = 60 + (5 \times x_i)$.

	Spop in 1000s (x_i)	Msales (in 1000 taka) (y_i)	Fitted Values (in 1000 taka) ($\widehat{y}_i$)
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 R you can get the fitted values with the command `fitted(model)`. Note that these fitted values values are within the sample data points, so this is why we call this *in-sample prediction*.

- Note that in sample prediction may or may not be equal to the y_i from the data. In the next section we will learn about a quantity - which is called *R-squared* or in short R^2, which is a measure about how good is our in-sample prediction, or how good the line fits the data.

- With the same equation we can also do *out-of-sample prediction*, which was our initial goal.

- For example we can predict when the student population is 30 thousands (notice 30 is not in the sample, nor in the range). Recall this was initial goal .... If we do this we get $60 + (5 \times 30) = 210$ so, $210,000$ taka sales. So this is a *predicted value for which we don't know y_i*.

Be Careful With Perfect In-Sample Predictions

- We need to be careful regarding very good in-sample prediction. A *good in-sample prediction does not automatically mean we will get a very good out-of-sample prediction*. The reason is - *we already used the data to fit the line*, meaning, the *line is such that it fits the data points very well*, this is by construction. So of course we will get a very good in-sample prediction.
- There is a way we can evaluate out-of-sample prediction, using *training and test sample*. The idea is we randomly separate some data as a test data, which we don't use to get the line and then we get our best fitted line, do prediction and then we compare the predicted values with the actual values.

- You will do another example in your homework ....

Assessing the Fit - R^2 and RSE

Assessing the Fit - R^2 and RSE

1. Goodness of fit - R^2

- Now we will learn two summary measures that tells *how good the line fits the data*
    - *Coefficient of Determination* or in short R^2
    - *Residual Standard Error* or in short *RSE*
- Let's start with R^2. The basic formula is,

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

- where

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

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

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

- where $\bar{y}$ is the sample mean

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

Question is - what does this formula mean? To understand this let's decompose $y_i - \bar{y}$

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

We can visually understand this in the following picture, below the black horizontal line is for $\bar{y}$

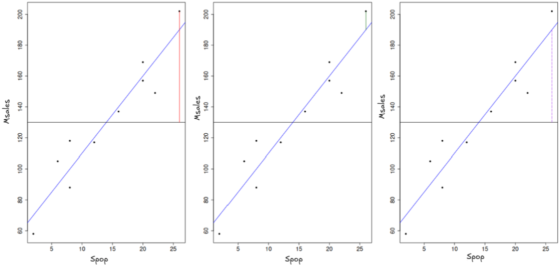


Figure 5: On the left we have $y_i - \bar{y}$, then on the middle we have $(y_i - \widehat{y}_i)$ and on the right we have $(\widehat{y}_i - \bar{y})$

- Now we can take squares and sum on both sides of the decomposition and we get (the product term becomes 0)

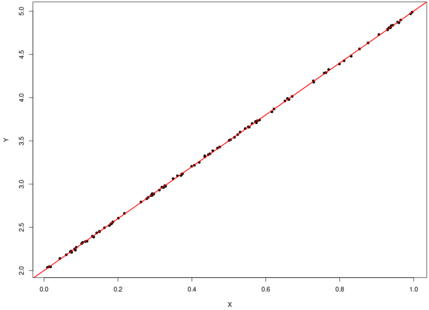
$$\underbrace{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}_{\text{SST}}=\underbrace{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}_{\text{SSE}}+\underbrace{\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{y}\right)^2}_{\text{SSR}}$$

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

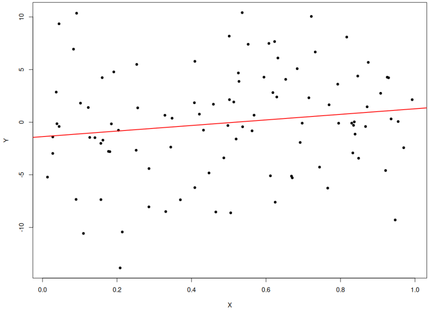
- So this means R^2 tells "*out of the total variation of y how much we can explain by regression*".
- Also note R^2 is a ratio of explained sum of squares and total sum of squares. So this means we will always have $0 \leq R^2 \leq 1$ (in other words the value of R^2 will always lie between 0 and 1).
- So high R^2 means the least-squares line fits very well with the data. Here are some examples of high R^2 with a different data sets .... please try to understand carefully,

Assessing the Fit

Goodness of Fit or R^2



- The black dots are the sample points, the red line is the fitted line. Here are the regression line perfectly fits the data. For this data set if we calculate R^2 we will get 0.99. It's a different data set, not our Msales-Spop data (so don't get confused)



- Here is another data set, here obviously the fit is not good, if we calculate the R^2, in this case we get $R^2 = 0.02$, which is almost close to 0.

- So the above discussion shows R^2 tells us how good is our *least-squares line* or the *regression 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.
- There are different names of R^2, one name is *Coefficient of Determination*, sometimes we also call it *Goodness of Fit*.
- In our Monthly Sales and Student Population, R^2 is 0.9027, which means 90% of the variability in sales can be explained by the student population. So this is a good fit.
- *Again be careful about out of sample prediction:* Probably you have already understood that *high R^2 does not automatically mean that we did a good job with our prediction problem for any data*, since this is an in-sample measure ....But still we can say high R^2 is something that is generally desirable.

Issues with SST, SSR, SSE short forms - BE CAREFUL if you read different books

- If you read Anderson, Sweeney, Williams, Camm, Cochran, Fry and Ohlmann (2020) or Newbold, Carlson and Thorne (2020) you will see the words SST (Total Sum of Squares), SSR (Regression Sum of Squares) and SSE (Sum of Squared Errors) or (Error Sum of Squares), we used this.
- If you read James, Witten, Hastie and Tibshirani (2023), you will see the words like TSS (Total Sum of Squares), RSS (Residual Sum of Squares), and ESS (Explained Sum of Squares)
- There
  - TSS is same as SST ,
  - ESS (Explained Sum of Squares) is same as SSR
  - RSS (Residual Sum of Squares) is same as SSE.
- So again, one option is to use TSS, RSS and ESS
- The other option is to use SST, SSR, SSE.
- We will use SST, SSR and SSE like Anderson, Sweeney, Williams, Camm, Cochran, Fry and Ohlmann (2020), because I think this is more common.
- Suppose we use TSS, RSS and ESS, then we can write R^2 as

Assessing the Fit - R^2 and RSE

2. Residual Standard Error or RSE

- Another useful measure to assess how good is the fit, is the *Mean Squared Error* or the square root of this quantity which is called *Residual Standard Error* or *Standard Error of the Estimate*. The *Mean Squared Error* is defined as

$$\mathrm{MSE} = \frac{\mathsf{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

- Here $n-2$ comes since we need to estimate two quantities to calculate e_i, which are $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Note that this can be also seen as as the variance of the residuals, or the variance of the errors since

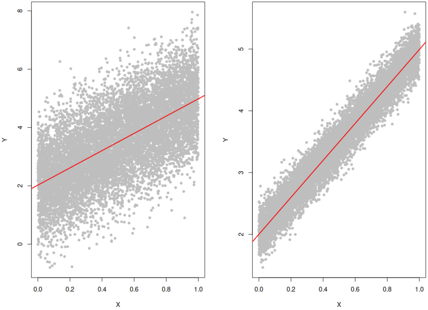
$$\frac{1}{n-2}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

- this equality comes since we can easily show that $\bar{e} = 0$ (you can check this with the data!).
- The square root of this is called *Residual Standard Error* or *Standard Error of the Estimate*.

$$\mathsf{RSE} = \sqrt{\mathrm{MSE}}$$

- In our regression result of Monthly Sales and Student Population, it is 13.83, how do we interpret this?
  - One way to interpret this is - on average sales deviate from the regression line by approximately 13,830 taka

- 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.

- 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,

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

- 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*.

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

- 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 $\sqrt{\mathsf{MSE}} = \mathsf{RSE}$ is an estimate of σ.

Model Assumptions, Interval Estimations and Testing

- An important question is

  Question: How do you know that the population regression function is linear like $\beta_0 + \beta_1 x$? why not some non-linear function?

  Answer: It's simply an assumption to make our life easier

- You will see that in Statistics / Econometrics often we will assume something about the unknown world, and this will make our life easier ... in fact help us to get some possible solutions...
- You might object by saying - *wait why did we assume*, the answer is the *real life scenarios are often so complex that it is almost impossible to learn from data without making any assumption at all*... so *there is no free lunch..*
- There is famous quote by George Box - "*All models are wrong, but some are useful*".

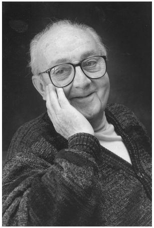


Figure 6: George Box (1919 - 2013), source - Wikipedia

- What Box meant here is, when we assume a model about the real life, it maybe wrong, but still the model may be useful to learn something about the world.
- Sometimes the assumptions are very strong and sometimes we can relax certain assumptions. In simple linear regression model, often we will often have following 4 assumptions,

Simple Linear Regression Model - Assumptions

- *Assumption 1* - We have an iid random sample, $\{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\}$. So all these pairs are independent and identically distributed random variables.
- *Assumption 2* - The CEF (also known as population regression function) is a linear function in X_i,

$$\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i \tag{3}$$

  Here β_0 is the intercept and β_1 is the slope but this is for the population.
- *Assumption 3* - Define

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

  We assume $\mathbb{V}(\epsilon_i|X_i = x) = \sigma^2$ for all x values, where σ^2 is a constant. This is known as *Homoskedasticity* which means the variance of the error term is constant for all x values.
- *Assumption 4** - Conditional on x, ϵ_i is Normally distributed with mean 0 and variance σ^2, so we can write $\epsilon_i|X_i = x \sim \mathcal{N}(0, \sigma^2)$

- The last assumption can be dropped if we have large sample size.

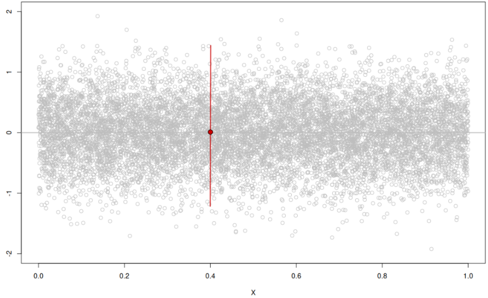
- We need to mention some important points regarding the CEF error ϵ_i, particularly the *conditional expectation or conditional mean* and the *conditional variance* of the CEF error. Recall CEF error is

$$\epsilon_i = Y_i - \mathbb{E}(Y_i|X_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

- First note that because of the model assumptions, it is possible to show that the conditional mean of CEF error is 0 (this is very to show, see Appendix)

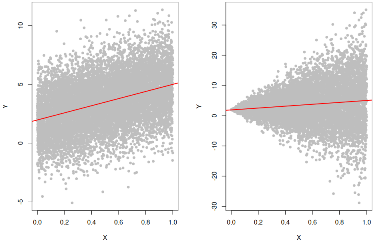
$$\mathbb{E}(\epsilon_i|X_i) = 0$$

- Also visually you can argue like this..... Let's plot 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.



- Interestingly because of this the overall expectation or unconditional expectation of ϵ_i is also 0, which means $\mathbb{E}(\epsilon_i) = 0$ (this is an application of *law of iterated expectation*, but we will not go into details here).

- Now let's talk about the conditional variance with $\mathbb{V}(\epsilon_i \mid X_i = x)$.
- We assume Homoskedasticity which means the conditional variance of ϵ_i is constant for all x values. Consider following picture where we plotted two *population data* and the red line is the CEF function.



- On the left the variance of ϵ_i seems to be constant with x values, so this means $\mathbb{V}(\epsilon_i|X_i = x)$ is constant. But on the right the variance of ϵ_i is changing with x values (in particular increasing), so this means $\mathbb{V}(\epsilon_i|X_i = x)$ is NOT constant, it is called *heteroskedasticity*! In the assumption we don't allow heteroskedasticity, so we assume $\mathbb{V}(\epsilon_i|X_i = x)$ is constant for all x values.

- Just using the definition of variances, we can show that conditional variance of ϵ_i is equal to conditional variance of Y_i, so this means (this is easy to understand from the picture)

$$\mathbb{V}(\epsilon_i \mid X_i = x) = \mathbb{V}(Y_i \mid X_i = x)$$

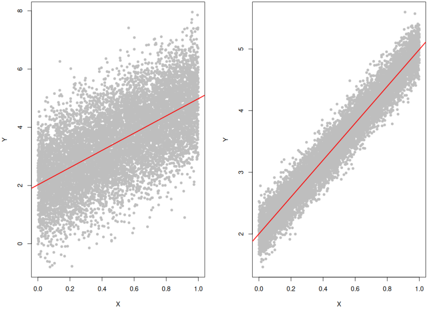
- Finally if we assume homoskedasticity, then we can show that the unconditional variance of ϵ_i is also σ^2, so

$$\mathbb{V}(\epsilon_i) = \mathbb{V}(\epsilon_i | X_i = x) = \sigma^2$$

- And also we have

$$\mathbb{V}(\epsilon_i) = \mathbb{V}(Y_i) = \sigma^2$$

- Now let's see what happens if σ^2 is high versus σ^2 is low, again consider two population data, for both $\mathbb{V}(\epsilon_i|X_i = x) = \sigma^2$ is constant. But on the left it is is high and on the right it is low



- Definitely, if the conditional variance is high then unconditional variance $\mathbb{V}(\epsilon)$ is also high.
- High variance of ϵ_i means the errors are large, so in a random sample we may have a data which could give us a line, that may not be close to the true line / population line....

Model Assumptions, Interval Estimations and Testing

4. Confidence Interval for β_0 and β_1

Recall from old discussions:

- When we have a sample mean $\overline{X}$, the formula for the *$(1-\alpha)$ percent confidence interval for population mean μ* would be,

$$\overline{X} + t_{1-\alpha/2,n-1}\,\widehat{\text{SE}}(\overline{X})$$

- For example if we want 95% confidence interval then $\alpha = 0.05$

- Here $t_{1-\alpha/2,n-1}$ is the $(1-\alpha) \times 100$ percent quantile of the t distribution with $n-1$ degrees of freedom. Following functions can be used in R and Excel
  - In R you can use `qt(1 - `α`/2, `$n-1$`)`,
  - and in **Excel**, you can use the function `T.INV(`$1-\alpha/2$`, `$n-1$`)`.

- And $\widehat{\text{SE}}(\overline{X})$ is the *estimate of the standard error of the sample mean*, which is calculated as

$$\widehat{\text{SE}}(\overline{X}) = \frac{s}{\sqrt{n}}$$

Recall the *standard error* is $\text{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, but we never know σ, so we use s and then it becomes *estimate of the standard error*, this is why we used "hat" symbol. The standard error is coming from the sampling distribution of $\overline{X}$, and it is the standard deviation of the sampling distribution of $\overline{X}$.

- One important point *if the sample size becomes large, the t distribution becomes Normal distribution*, on that case we can use $z_{1-\alpha/2}$, rather then $t_{1-\alpha/2,n-1}$. Usually when the sample size is more than 30 is considered as a large sample.

Confidence Interval for β_0 and β_1

- Now in the regression problem we have two unknown parameters,

$$\beta_0 \text{ and } \beta_1$$

- And for each of them it is possible to construct $(1-\alpha) \times 100\%$ percent confidence Interval, let's see them one by one,

- **The confidence interval formula for β_1 is**

$$\widehat{\beta}_1 \pm t_{1-\alpha/2, n-2}\, \widehat{\mathrm{SE}}(\widehat{\beta}_1)$$

- Excel automatically gives you 95% confidence interval and also in the setting you can change, in R you need to use the function `confint(model)`.

- Note and important point is, in this case the sampling distribution is t distribution with $n-2$ degrees of freedom, rather than $n-1$, the reason is we need to estimate two objects $\widehat{\beta}_1$ and $\widehat{\beta}_2$.
- And again if the sample size becomes large we can use $z_{1-\alpha/2}$, in this case the confidence interval would be

$$\widehat{\beta}_1 \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}(\widehat{\beta}_1)$$

Confidence Interval for β_0 and β_1

- For our problem, the 95% *confidence interval estimate for* β_1 is

$$(3.67, 6.34)$$

- What is the interpretation? *It's a fixed interval, the true value of* β_1 *is either in this interval or not. The 95% confidence interval means, if we construct this kind of intervals* 100 *times then* 95 *of them will contain the true value of* β_1..
- Similarly we can construct confidence interval for β_0 ... please construct and do the interpretation.

Model Assumptions, Interval Estimations and Testing

5. Significance Testing - t - test

Significance Testing - t test

- Standard errors can also be used to perform hypothesis tests on the *unknown coefficients*. The most common hypothesis test involves testing the null hypothesis of

Recall from old discussions:

- When we have a sample mean $\overline{X}$, the t-test, for example the two tail test for μ, can be done with following hypotheses,

$$H_0 : \mu = 30$$
$$H_a : \mu \neq 30$$

- In this case we used to calculate t_{calc}, which is

$$t_{calc} = \frac{\overline{x} - 30}{\widehat{SE}(\overline{X})}$$

- And then using critical value approach we reject the null if $t_{calc} > t_{1-\alpha/2,n-1}$ or $t_{calc} < t_{\alpha/2,n-1}$
- Or using p-value approach, we reject the Null if p-value $< \alpha$

Significance Testing - t test

- The testing problem in Regression is similar, we can different testing for β_0 and β_1, the most common test is called the *significance testing*, which is following,

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- Recall the population regression function,

$$\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

- So if we *accept the Null*, this means *there is no significant relationship between X variable and Y variable*, in our case this means there is no significant relationship between student population and monthly sales.
- Similarly if we *reject the Null*, then this means *there is a significant relationship between student population and monthly sales*.
- In our case, we have

$$t_{calc} = \frac{\widehat{\beta}_1 - 0}{\widehat{SE}\left(\widehat{\beta}_1\right)},$$

- Both in R and Excel output you already have the p value, so you don't need to manually do the testing, note that in page 31, we have `p-value:  0.00002549`, this means we can reject the Null and the conclusion is - *there is a significant relationship between student population and monthly sales*

- If you use the critical value approach, you need to compare the t_{calc} with $t_{\alpha/2,n-2}$ and $t_{1-\alpha/2,n-2}$, or in large samples just compare with $z_{\alpha/2}$ and $z_{1-\alpha/2}$

Model Assumptions, Interval Estimations and Testing

6. Some Algebraic Details*

Some Algebraic Details

- So far the story is following, we minimize sum of squared errors to get a line, this means we are minimizing the following function,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) \right)^2$$

- This problem is known as *Ordinary Least Squares* or OLS, and the solution is given by the following equations,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{S_{X,Y}}{S_X^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

- Where $S_{X,Y}$ is the sample covariance between X and Y, and S_X^2 is the sample variance of X.
- Now understanding the details regarding the solution is actually helpful, question is how to derive these equations?
- There are two approaches,

Some Algebraic Details

- **1. Plugin Approach or Method of Moments Approach:** Solve the population problem and the replace the Expectation with Sample Mean, this is called *plugin approach* or *method of moments approach*. So in this case our task is to derive first

$$\beta_1 = \frac{\mathbb{E}\Big[(X_i - \mu_{X_i})(Y_i - \mu_{Y_i})\Big]}{\mathbb{V}\left(X_i\right)} = \frac{\mathbb{C}\text{ov}(X_i, Y_i)}{\mathbb{V}(X_i)}$$
$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$

  and then we can replace the population quantities with sample quantities.

- **2. Directly Finding Least Squares Solution:** The other approach is to directly minimize the sample MSE, which is called *Least Squares Approach* or *Ordinary Least Squares* or OLS. In this case we will minimize the following function,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \Big(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\Big)^2$$

  Which is actually same as minimizing the sample MSE, in the minimization problem we will differentiate w.r.t. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and then we will get the solution.

- Let's see the first one, the plugin approach, which is actually easier to understand,...

Some Algebraic Details

Population Problem: Minimizing MSE for linear CEF function,

$$\min_{\beta_0,\beta_1} \mathbb{E}\left[\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2\right]$$

Differentiate w.r.t. β_0 gives :

$$\mathbb{E}\left[Y_i - \beta_0 - \beta_1 X_i\right] = 0$$

$$\Rightarrow \beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$

Differentiate w.r.t. β_1 gives:

$$\mathbb{E}\left[X_i\left(Y_i - \beta_0 - \beta_1 X_i\right)\right] = 0$$

Substitute β_0 into the second equation, then we get,

Some Algebraic Details

$$\underbrace{\mathbb{E}[X_i Y_i] - \mathbb{E}[X_i]\mathbb{E}[Y_i]}_{\mathbb{C}\text{ov}(X_i, Y_i)} - \beta_1 \underbrace{\left(\mathbb{E}\left[X_i^2\right] - \mathbb{E}[X_i]^2\right)}_{\mathbb{V}(X_i)} = 0$$

$$\Rightarrow \beta_1 = \frac{\mathbb{C}\text{ov}(X_i, Y_i)}{\mathbb{V}(X_i)} \quad (\text{ requires } \mathbb{V}(X_i) > 0)$$

and we already have

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$

Note in the derivaion we used the following definitions,

$$\mathbb{C}\text{ov}(X_i, Y_i) = \mathbb{E}\Big[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])\Big] = \mathbb{E}[XY_i] - \mathbb{E}[X_i]\mathbb{E}[Y_i]$$

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

Some Algebraic Details

- So what we proved is, in the population we have

$$\beta_1 = \frac{\mathbb{C}\text{ov}(X_i, Y_i)}{\mathbb{V}(X_i)}$$

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$

- So in the population the slope coefficient *β_1 is the ratio of population covariance of X_i and Y_i by sample variance of X_i* and the intercept coefficient *β_0 is the difference between the mean of Y_i and β_1 times mean of X_i.*
- Now note in the sample we just have,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

- Where the hat quantities are just the sample estimates (or estimators) of the population quantities,

Some Algebraic Details

- Now we can also derive the same estimates by directly by minmizing the sample MSE, which we will see in a minute, let's derive some results, for CEF error $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$,
- First note that assuming the CEF is linear, we can show that the conditional expectation of ϵ_i is 0, which means

$$\mathbb{E}(\epsilon_i | X_i) = Y_i - (\beta_0 + \beta_1 X_i) = 0$$

- Using LIE we can also show that the unconditional expectation of ϵ is also 0, which means

$$\mathbb{E}(\epsilon_i) = \mathbb{E}(\epsilon | X_i) = 0$$

- Finally we can also show that the error is uncorrelated with X_i, this is because

$$\mathbb{C}\text{ov}(X_i, \epsilon_i) = \mathbb{E}(X_i \epsilon_i) - \mathbb{E}(X_i)\mathbb{E}(\epsilon_i) = \mathbb{E}(X_i \epsilon_i) - 0 = 0$$

- Where using LIE we used

$$\mathbb{E}(\epsilon_i X_i) = 0$$

- Which can be showed using LIE.

Some Algebraic Details

- Now let's derive the sample estimates by minimizing the sample MSE, which is defined as

$$\text{Sample MSE} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

- Note that this is same as minimizing the following function, which is called *Sum of Squared Errors* or SSE (since we can ignore multiplid constants when we are minimizing or maximizing) with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$,

$$\text{SSE} = \sum_{i=1}^{n}\left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\right)^2$$

So we need to take the partial derivatives of this function with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$, set them to 0 and then solve the resulting linear system.

$$\frac{\partial\ \text{SSE}}{\partial \widehat{\beta}_0} = -2\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) = 0 \Rightarrow n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\frac{\partial\ \text{SSE}}{\partial \widehat{\beta}_1} = -2\sum_{i=1}^{n} X_i\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) = 0 \Rightarrow \widehat{\beta}_0 \sum_{i=1}^{n} X_i + \widehat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

Solve the 2×2 linear system

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}, \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}_i$$

where $\bar{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

$$S_x^2 = \sum \left(X_i - \bar{X}_i\right)^2, \quad S_{xy} = \sum \left(X_i - \bar{X}_i\right)\left(Y_i - \bar{Y}\right),$$

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_x^2}, \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}_i.$$

- So bottom line we derived the same estimates, which are called *Ordinary Least Squares* or OLS estimates, but now directly from the sample MSE, the previous approach is called moment approach or plugin method, where we have a population problem and then we derived the sample estimates. Often this is an easier approach and more intuitive.

Some Algebraic Details

- Note that using the OLS esimates we can easily show that

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} X_i e_i = 0, \quad \text{where } e_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i.$$

- These are useful properties, the first one means the sum of residuals is 0, and the second one means the sum of residuals weighted by X_i is also 0.

Some Algebraic Details

- Now we will derive the conditional mean and conditonal variance of $\widehat{\beta}_1$ ...
- We can show that the conditional expectation and conditional variance of $\widehat{\beta}_1$ are following,

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1,$$

$$\mathbb{V}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2}$$

  where σ^2 is the population variance of the CEF error ϵ_i, or as we wrote before $\mathbb{V}(\epsilon_i|X_i = x) = \sigma^2$ for all x values and $\mathbb{V}\left(\widehat{\beta}_1\right)$ is the variance of the sampling distribution of $\widehat{\beta}_1$ (think about repeated sampling).
- Now the standard error of $\widehat{\beta}_1$ is defined as

$$\text{SE}\left(\widehat{\beta}_1\right) = \sqrt{\mathbb{V}\left(\widehat{\beta}_1\right)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2}} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2}}$$

- In a practical scenario we never know σ^2, so we use the sample variance of the prediction error to estimate it, which is defined as

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right)^2 = \text{MSE}$$

Some Algebraic Details

- So the estimate of the standard error of $\widehat{\beta}_1$ is

$$\widehat{\mathrm{SE}}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2}} = \frac{\mathrm{RSE}}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2}}$$

- Where RSE is the **Residual Standard Error**, or Standard Error of the Estimate, or sometimes also called **Root Mean Squared Error** (in short RMSE) and defined as, $\mathrm{RSE} = \sqrt{\mathrm{MSE}}$
- This is the quantity which any software calculates and gives you in the output, so you can use it to construct confidence intervals and do hypothesis testing.

Some Algebraic Details

- Let's see how to derive this ....

References

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.

Newbold, P., Carlson, W. L. and Thorne, B. M. (2020), *Statistics for Business and Economics*, 9th, global edn, Pearson, Harlow, England.