

Ch3 - Simple Linear Regression (SLR)

ECO 204
Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka

Last updated: May 4, 2025



Outline

1. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

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 - Spring 2025)
- ▶ Most of the contents of these slides are based on
 - ▶ James et al. (2023),
 - ▶ Anderson et al. (2020), and
 - ▶ My own ideas and imaginations (which you are always welcome to criticize)...
- ▶ For theoretical discussion I primarily followed James et al. (2023). Anderson et al. (2020) is an excellent book with lots of nice and intuitive examples, but it lacks proper theoretical foundations. Here James et al. (2023) is truly amazing and I believe it explained the concepts in a very very easy and accessible way. We thank the authors of this book for making everything publicly available at the website <https://www.statlearning.com/>.
- ▶ Also I thank my students who took this course with me in Summer 2022, Fall 2022, Fall 2023 and Currently Spring 2025. 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. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

The Regression Problem

The Regression Problem

1. Dependent and Independent Variables

The Regression Problem

A Motivating Example

- Statistics is a blend of theory and practice. While the theories and formulas can seem abstract, their true power lies in how they help us make sense of the complex, messy data around us. ...

Experience without theory is blind, but theory without experience is mere intellectual play. - Immanuel Kant

- Let's start with a real life problem, suppose we would like to understand the *sales of a fast food restaurants located in different university areas in the Dhaka city.*

The Regression Problem

A Motivating Example



Figure 1: A snapshot of fast food restaurants close to the East West University campus.

- Note: This image, taken from Google Maps (October 2023), shows that several fast food places, in particular *Khan Tasty Food*, *Ka te Kacchi*, *Tasty Treat*, *Yummy Bite*, *CP Five Star*, and *Turkish Kabab House* are in the close proximity of the East West University campus. This also highlights the high concentration of fast food options available within a short distance of the campus.

The Regression Problem

A Motivating Example

- ▶ In particular our aim is to understand following questions:
 - *Q1. Which variables are associated to the sales of the fast food restaurants? and these variables are associated to sales?*
 - *Q2. Which variables can be used to predict sales if we have some data? and how to do “best” prediction?*
- ▶ Let's answer *Q1. and Q2.* ... following variables maybe positively / negatively associated to sales... (you can think more... but for now these are OK)
 - ▶ *Student Population:* More students means more customers and more sales.
 - ▶ *Average Pricing:* Cheaper prices could lead to larger sales.
 - ▶ *Advertising:* More advertising perhaps could increase sales.
 - ▶ *Local Economic Status:* Higher income area might lead to increased sales...

The Regression Problem

A Motivating Example

- ▶ Above we have already given a qualitative answer of *Q1. and Q2.* ... but now we will give a quantitative answer to the question and also answer *Q3.*
- ▶ To do this we will use a technique known as *Regression*
- ▶ In the regression problem *there is a dependent variable*, which we want to predict, and *there are some independent variables (or features or predictors)* which we will use to predict (or explain) the dependent variable.
- ▶ In our example,
 - the *dependent variable* is Monthly *Sales*
and the *independent variables* are
 - *Student Population*,
 - *Average Pricing*,
 - *Advertising*,
 - *Local Economic Status*.
- ▶ Here the units are important and also we can give some short names for convenience
 - ▶ **Monthly Sales** will be *Msales* in the data and measured in 1000 BDT.
 - ▶ **Student Population** will be *Spop* in the data and will be measured in 1000s.
 - ▶ **Average Pricing** will be *Aprice* and will be measured in BDT.
 - ▶ **Annual Advertising** will be *Adv* and will be measured in 1000 BDT.
 - ▶ **Local Economic Status** will be *ECOSat* and can be *High, Medium and Low* (Categorical)
- ▶ We would like to understand these variables influence on the **Sales** and whether we can use these variables to predict sales in the “best” possible way.

The Regression Problem

A Motivating Example

- ▶ We can express the relationship between dependent variable and independent variables as a function

$$\text{MSales} \approx f(\text{Spop}, \text{Aprice}, \text{Adv}, \text{LES})$$

- ▶ Often we will use Y as a dependent variable, $X_1, X_2, X_3, X_4 \dots$, as independent variables and $f(\cdot)$ as a function. So we can write this as

$$Y \approx f(X_1, X_2, X_3, X_4)$$

- ▶ You know what is a function right.... ???

The Regression Problem

A Motivating Example

- Suppose to do this we collected following data set, then in the regression our goal is to often estimate a function $f(\cdot)$, such that if we only have independent variables (X_1, X_2, X_3, X_4) but not the value for Y , we can predict Y , look at the data for restaurant 11

Restaurant	Msales (Y)	Spop (X_1)	Aprice (X_2)	Adv (X_3)	ECOSTat (X_4)
1	58	2	280	50	Low
2	105	6	260	120	Middle
3	88	8	270	100	Middle
4	118	8	250	150	High
5	117	12	240	200	High
6	137	16	230	180	Low
7	157	20	220	220	Middle
8	169	20	210	250	High
9	149	22	200	230	Middle
10	202	26	180	300	High
11	???	15	200	250	High

- For example if somehow magically you know the for ECOSTat = High, the function is

$$f(X_1, X_2, X_3, X_4) = 50 + 7X_1 + 0.5X_2 + 0.5X_3$$

The Regression Problem

A Motivating Example

- ▶ Then we can predict the sales for restaurant 11 as

$$Y = f(15, 200, 250) = 50 + 7 \times 15 + 0.5 \times 200 + 0.5 \times 250 = 155$$

- ▶ So we can predict the sales for restaurant 11 is 155.
- ▶ In the regression problem our goal is to learn this function in the best possible way....
- ▶ Note that Restaurant column is not a variable, it just represents which restaurant (you may call this an identifier), so even if you remove this column, it won't change anything.

The Regression Problem

A Motivating Example

Simple Linear Regression Problem:

*When we will try to understand how **one independent variable** is associated to **a dependent variable** we call this **Simple Linear Regression** (SLR) problem. For example, we might be interested to know how **Student Population** is associated to **Sales**.*

and

Multiple Linear Regression Problem:

*When we will try to understand how **more than one independent variables** is associated to **a dependent variable** we call this **Multiple Linear Regression** (MLR) problem. For example in this case we will see how **Student Population (Spop)**, **Average Pricing (Aprice)** and **Advertising (Adv)** together or jointly associated to **Sales**.*

- ▶ In this chapter first we will start with **Simple Linear Regression** problem and then in the next chapter we will move to **Multiple Linear Regression** problem, which is of course more realistic.
- ▶ The dependent and independent variables have different names, you should know them,

The Regression Problem

A Motivating Example

<i>Dependent Variable</i>	<i>Independent Variable</i>
Response Variable	Predictor Variable
Target Variable	Feature
Outcome Variable	Covariate
Label	Explanatory Variable
Output Variable	Input Variable

Table 1: Different names for dependent and independent variables

- In some moments you will understand why the independent variable is called input variable and dependent variable is called output variable.... hold on...
- We will learn that Regression is a new technique that will help us to understand the relationships between the response and predictor variables and also how to predict the response using the predictors.

The Regression Problem

2. Numerical and Graphical Measures of Association from ECO 104

The Regression Problem

Numerical and Graphical Measures from ECO 104

- ▶ Since from now on we will only focus on Simple Linear Regression, Let's consider *only one independent variable* which is **Student Population (SPop)** in 1000s (we will ignore the other variables in this chapter...).
- ▶ We will write the independent variable or predictor variable with x_i , so x_1, x_2, \dots, x_n and dependent variable or response variable with y_i , so y_1, y_2, \dots, y_n .

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

The Regression Problem

Numerical and Graphical Measures from ECO 104

- Before going to the regression problem, with some numerical and graphical measures we can also see whether there is an association between these two variables (this is from ECO 104), there are two measures you can see

1. **Sample Covariance and Correlation:**, where the formula for the *Sample Covariance* is

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

and the formula for the *Sample Correlation* is

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

where s_x and s_y are the sample *Standard Deviations* of x and y respectively which is

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

2. **Scatter Plot:** where we plot (x_i, y_i) on the x - y coordinate.

The Regression Problem

Numerical and Graphical Measures from ECO 104

- ▶ We can do this very easily in Excel and in R, let's see this in class....
- ▶ First we calculate the covariance, it will be

$$s_{x,y} = 315.5556$$


- ▶ Which means there is a positive association between the two variables, but from this number it doesn't tell us how strong the association is.
- ▶ Here is what correlation comes, if you calculate the correlation, it will be

$$r_{x,y} = 0.950123$$

- ▶ Which means there is a very strong positive correlation between the two variables. Since we know that always we will have $-1 \leq r_{x,y} \leq 1$ and $r_{x,y}$ close to 1 means strong positive association and $r_{x,y}$ close to -1 means strong negative association and $r_{x,y} = 0$ means no association.

The Regression Problem

Numerical and Graphical Measures from ECO 104

- Next we can have a look at the scatterplot, here is the scatter plot in 

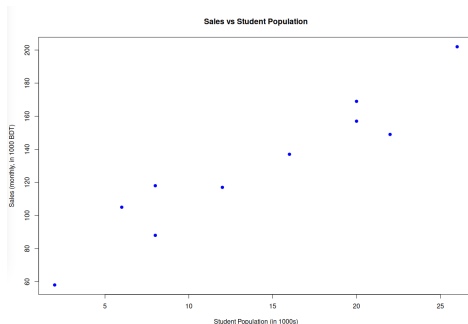


Figure 2: Scatter plot of SPop and Msales

- Note as Student Population (SPop) increases the Monthly Sales (Msales) also Increases for the restaurants. This already shows the positive relationship or association between the two variables.

The Regression Problem

Numerical and Graphical Measures from ECO 104

- But what about prediction? Suppose we would like to predict the sales of a restaurant with 15,000 students.... (note we don't have data for 15000 student population). Although the correlation and scatter-plot are good measure to talk about association, but we cannot directly use them for prediction, and here is where regression comes.... see next section..

The Regression Problem

Numerical and Graphical Measures from ECO 104

- ▶ Before we conclude this section, I need to emphasize that correlation, covariance and also regression slope co-efficient (which you will learn in the next section) are simply association measures. Just because two variables have high correlation or covariance, *this does not imply that one variable causes the other*.
- ▶ Causality is a completely different thing, and just from the data it's very hard to prove causality.
- ▶ We will not talk about causality in this course, but you should always remember that *correlation does not imply causation*.... Example - There is high correlation between chocolate consumption and Nobel Prize winners across countries... but we cannot say that eating chocolate causes people to win Nobel Prizes.

The Regression Problem

Numerical and Graphical Measures from ECO 104

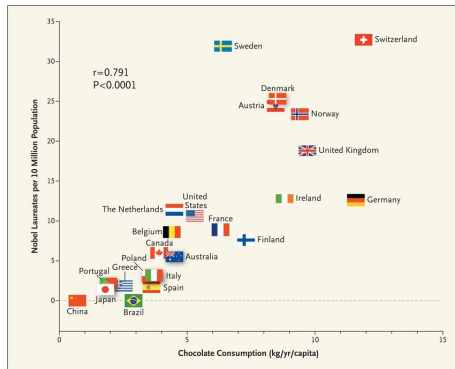


Figure 3: Scatterplot between chocolate consumption and Nobel Prize winners

1. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

Simple Linear Regression Model (SLR) - The Problem of Estimation

Simple Linear Regression Model (SLR) - The Problem of Estimation

1. Fitting a Linear Line

Simple Linear Regression

The Problem of Estimation (method of least squares)

- Our first task is to learn about *Simple Linear Regression Model* or in short SLR. Recall the following data, and the scatter plot

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

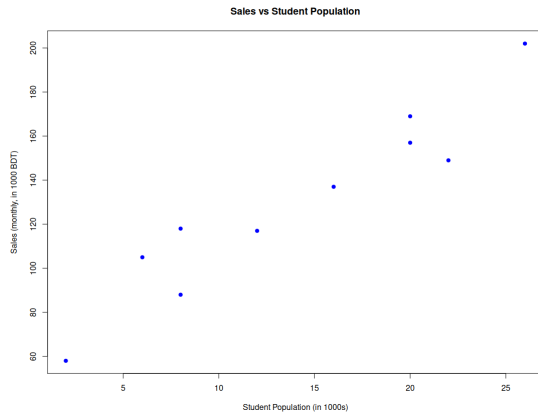


Figure 4: Scatterplot of Sales Vs. Student Population

Simple Linear Regression

The Problem of Estimation (method of least squares)

- We will start with the estimation problem (it will be clear later what are we estimating...), essentially our goal is to find the following red line - which can be called *the best fitted linear line*

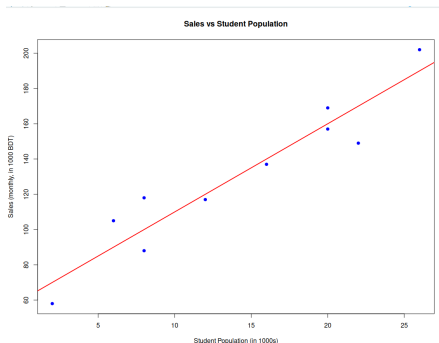



Figure 5: Scatterplot of Sales Vs. Student Population

Simple Linear Regression

The Problem of Estimation (method of least squares)

- ▶ The equation of the line will be something like this

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

- ▶ Here the \hat{y}_i is used for predicted value and $\hat{\beta}_0$ and $\hat{\beta}_1$ 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  command will give us the result
- ▶ You can also get the similar output in Excel, we will see this in class.

Simple Linear Regression

The Problem of Estimation (method of least squares)

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,

Simple Linear Regression

The Problem of Estimation (method of least squares)

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 $\hat{\beta}_0 = 60$ and slope $\hat{\beta}_1 = 5$

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

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

Simple Linear Regression Model (SLR) - The Problem of Estimation

2. Interpretations

Simple Linear Regression

Interpreting The Coefficients

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

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

- 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,

Simple Linear Regression

Interpreting The Coefficients

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

- The slope co-efficient $\hat{\beta}_1$ 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...

Simple Linear Regression

Interpreting The Coefficients

- **Interpretation of intercept** $\hat{\beta}_0 = 60$
- 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) - The Problem of Estimation

3. The Least Squares Problem

Simple Linear Regression

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

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

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

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

- ▶ the squared error is

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

Simple Linear Regression

The Least Squares Problem

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

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

- So now 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,

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

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

Simple Linear Regression

The Least Squares Problem

- 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,

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

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

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

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

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

Simple Linear Regression

The Least Squares Problem

- 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) - The Problem of Estimation

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

Simple Linear Regression

In-sample and Out-of-sample prediction

- 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) (\hat{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 are within the sample data points, so this is why we call this *in-sample prediction*.

Simple Linear Regression

In-sample and Out-of-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* .

Simple Linear Regression

In-sample and Out-of-sample prediction

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.

Simple Linear Regression

In-sample and Out-of-sample prediction

- ▶ You will do another example in your homework

1. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

Assessing the Fit - R^2 and RSE

Assessing the Fit - R^2 and RSE

1. Goodness of fit - R^2

Assessing the Fit

Goodness of Fit or 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,

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

- where

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

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

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

- where \bar{y} is the sample mean

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

Assessing the Fit

Goodness of Fit or R^2

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

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

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

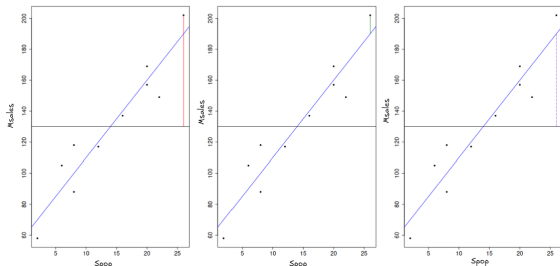


Figure 6: On the left we have $y_i - \bar{y}$, then on the middle we have $(y_i - \hat{y}_i)$ and on the right we have $(\hat{y}_i - \bar{y})$

Assessing the Fit

Goodness of Fit or R^2

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

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

- 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 .

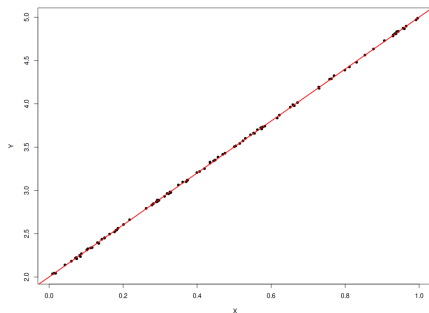
Assessing the Fit

Goodness of Fit or R^2

- ▶ 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

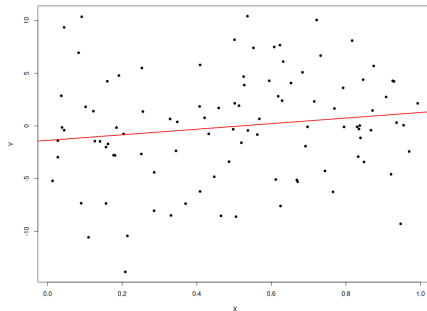
Goodness of Fit or R^2



- The black dots are the sample points, the red line is the fitted line. Here 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)

Assessing the Fit

Goodness of Fit or R^2



- 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.

Assessing the Fit

Goodness of Fit or R^2

- ▶ 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 measureBut still we can say high R^2 is something that is generally desirable.

Simple Linear Regression

Issues with Different Terminologies

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

Assessing the Fit

Residual Standard Error or RSE or Standard Error of the Estimate

- ▶ 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

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

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

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

- ▶ 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*.

$$\text{RSE} = \sqrt{\text{MSE}}$$

- ▶ 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

1. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

What is the Population Solution?

What is the Population Solution?

The CEF function - The Best Population Solution

The Regression Problem

The Population Solution to the Problem

- We already know how to do prediction here? We have the best fitted line $\hat{y}_i = 60 + 5x_i$, so we can use this to predict the sales for any given population.

	Spop in 1000s (x_i)	Msales (in 1000 taka) (y_i)	(\hat{y}_i)
1	2	58	70
2	6	105	90
3	8	88	100
4	8	118	100
5	12	117	120
6	16	137	140
7	20	157	160
8	20	169	160
9	22	149	170
10	26	202	190
11	15	?	135
12	27	?	195

- *Question is - what are we predicting here?* You might answer y_i , but it's not actually correct .. let's see why ...

The Regression Problem

The Population Solution to the Problem

- ▶ Recall in Statistical Inference problem we always do prediction or estimation for a population quantity. First of all we should ask *what is the population data here?* *Ans.* The data from *all fast food restaurants operating in different university areas in the Dhaka city...*
- ▶ Suppose we have the population ... then here is how the scatter plot would look like... (for example, think about population is 10,000 restaurants or something!)

The Regression Problem

The Population Solution to the Problem

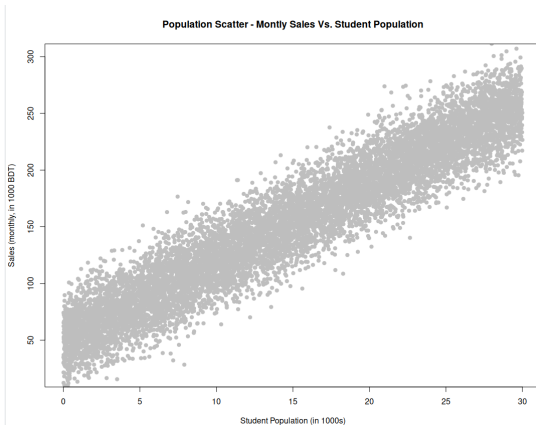


Figure 7: Scatter plot of the Population Data

The Regression Problem

The Population Solution to the Problem

- Our sample of 10 pairs of data points is a *random sample from the population*, we can also plot the sample in the same scatter plot,



Figure 8: Scatter plot of the Population Data (gray points) and Sample Data (blue points)

The Regression Problem

The Population Solution to the Problem

- Interestingly note that even if we have the population, *at $x = 15$ we have multiple y values*, so which value to take as a target for prediction? Any idea.... ?

The Regression Problem

The Population Solution to the Problem

- *Ans.* One solution is we can take the average of all y values which are paired with $x = 15$ in the population. In particular we can calculate what is known as *Conditional Average* or *Conditional Expectation* at $x = 15$, which can be written as

$$\mathbb{E}(Y_i \mid X_i = 15) \quad \text{Average of all } Y \text{ values when } X = 15$$

- Visually this means

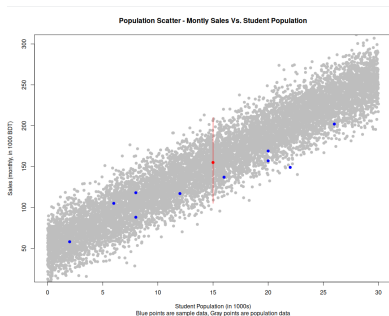


Figure 9: Scatter plot of the Population Data (gray points) and Sample Data (blue points) with the conditional expectation at $X = 15$, in this case the conditional average is 155

The Regression Problem

The Population Solution to the Problem

- ▶ It turns out that if we have the population data, this is the best prediction we can make (in some sense!)
- ▶ If we have the population we can calculate the conditional average at all points, and this will give us the following conditional expectation line, or conditional expectation function, or we also call this population regression function.

The Regression Problem

The Population Solution to the Problem

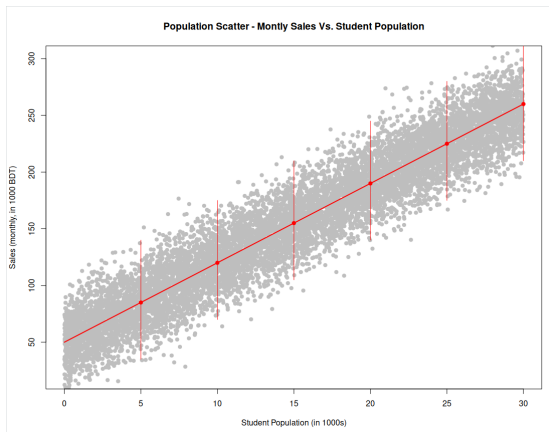


Figure 10: Scatter plot of the Population Data (gray points) the Conditional Expectation Function (red line)

The Regression Problem

The Population Solution to the Problem

- Note that knowing this conditional expectation function means we can calculate the conditional expectation at any point, for example if the function is

$$\mathbb{E}(Y_i|X_i = x) = 50 + 7x$$

we can plug the x value and get the conditional expectation at that point and this would give us the best prediction for any x .

- This is possibly the most important part of this section, that knowing conditional expectation function means we know the *best prediction line*, and then we can predict for any x value.
- For example $x = 15$, then we have $\mathbb{E}(Y_i|X_i = x) = 50 + 7x = 50 + 7 \times 15 = 155$, and we can predict the sales for $x = 15$ is 155.

The Regression Problem

The Population Solution to the Problem

- ▶ So now the question is - *Do we have the population data?* *Ans.* No, we don't have the population data, we only have the sample data.
- ▶ *Do we know the conditional expectation function?* *Ans.* No, we don't know the conditional expectation function either.
- ▶ But recall we are in the *Statistical Inference Class*, and we need to learn how to estimate this line and this is what we have learned beforefinding the best fitted line from the sample ...
- ▶ So in this case, we assume, the *population regression function* or *CEF* has following form,

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

where β_0 is the population intercept and β_1 is the population slope, and we we found before $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1 respectively.

- ▶ So again the population regression (from the population is),

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

- ▶ The sample regression function or the best fitted line is

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

1. The Regression Problem

- 1. Dependent and Independent Variables
- 2. Numerical and Graphical Measures of Association from ECO 104

2. Simple Linear Regression Model (SLR) - The Problem of Estimation

- 1. Fitting a Linear Line
- 2. Interpretations
- 3. The Least Squares Problem
- 4. In-Sample and Out-of-Sample Predictions

3. Assessing the Fit - R^2 and RSE

- 1. Goodness of fit - R^2
- 2. Residual Standard Error or RSE

4. What is the Population Solution?

- The CEF function - The Best Population Solution

5. Model Assumptions, Interval Estimations and Testing

- 4. Confidence Interval for β_0 and β_1
- 5. Significance Testing - t - test

Model Assumptions, Interval Estimations and Testing

Simple Linear Regression

Model Assumptions

- ▶ 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"*.

Simple Linear Regression

Model Assumptions



Figure 11: 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

Simple Linear Regression Model - Assumptions

- ▶ *Assumption 1* - We have an iid random sample, $\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)\}$. 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 ,

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

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.

Simple Linear Regression

Model Assumptions

- 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

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

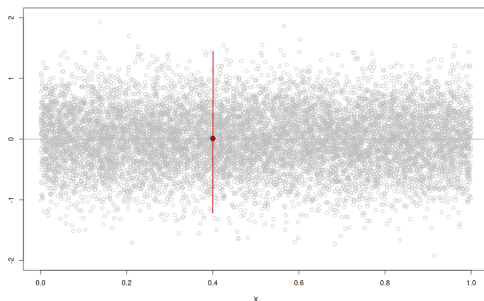
- First note that because of the model assumptions, it is possible to show that the conditional mean of CEF error is 0 (this is very to show, see Appendix)

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

- 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 .

Simple Linear Regression

Model Assumptions

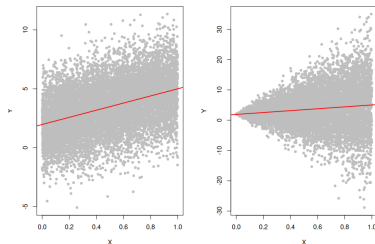


- Interestingly because of this the overall expectation or unconditional expectation of ϵ_i is also 0, which means $\mathbb{E}(\epsilon_i) = 0$ (this is an application of *law of iterated expectation*, but we will not go into details here).

Simple Linear Regression

Model Assumptions

- Now let's talk about the conditional variance with $\mathbb{V}(\epsilon_i | X_i = x)$.
- 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.

Simple Linear Regression

Model Assumptions

- ▶ 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

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

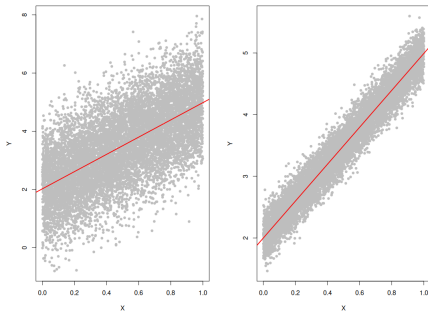
- ▶ And also we have

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

Simple Linear Regression

Model Assumptions

- 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 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

Confidence Interval for β_0 and β_1

Recall from old discussions:

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

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

- ▶ 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 - $\alpha/2$, $n - 1$)`,
 - ▶ and in **Excel**, you can use the function `T.INV(1 - $\alpha/2$, $n - 1$)`.
- ▶ And $\widehat{SE}(\bar{X})$ is the *estimate of the standard error of the sample mean*, which is calculated as

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

Recall the *standard error* is $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, 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 \bar{X} , and it is the standard deviation of the sampling distribution of \bar{X} .

Confidence Interval for β_0 and β_1

- 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 than $t_{1-\alpha/2, n-1}$. 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**

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

- 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 $\hat{\beta}_1$ and $\hat{\beta}_2$.
- And again if the sample size becomes large we can use $z_{1-\alpha/2}$, in this case the confidence interval would be

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

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 \bar{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

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

- ▶ 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 $p\text{-value} < \alpha$
- ▶ 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,

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Significance Testing - t test

- Recall the population regression function,

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

- 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

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

- 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 $t_{1-\alpha/2, n-1}$, or in large samples just compare with $z_{\alpha/2}$ and $z_{1-\alpha/2}$

- 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.
- 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.