Ch3 - Simple Linear Regression (SLR)

ECO 204
Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East **W**est **U**niversity, Dhaka
Last updated: April 13, 2025

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 - 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

The Regression Problem

The Regression Problem

1. Dependent and Independent Variables

- 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*.

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.

- 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...

- 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 *ECOStat* 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.

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

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

- Often we will use Y as a dependent variable, $X_1, X_2, X_3, X_4 \ldots,$ 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.... ???

- 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

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

- 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.

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,

Dependent Variable	*Independent Variable*
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

- 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 $x_1, x_2, \ldots, x_n$ and dependent variable or response variable with y_i, so $y_1, y_2, \ldots, y_n$.

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

- 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

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

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

- 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.

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

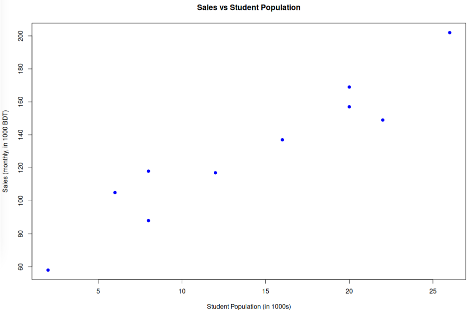


Figure 2: Scatter plot of SPop and Msales

- 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.

- 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..

Simple Linear Regression Model (SLR)

Simple Linear Regression Model (SLR)

1. The Problem of Estimation

- Our first task is to learn *Linear Regression Model*. In particular we will talk about *Simple Linear Regression Model* or in short SLR in this chapter.

- 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

- Recall the Scatter plot

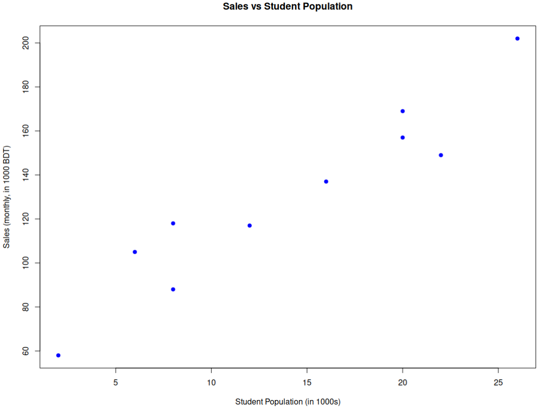


Figure 3: Scatterplot of Sales Vs. Student Population

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

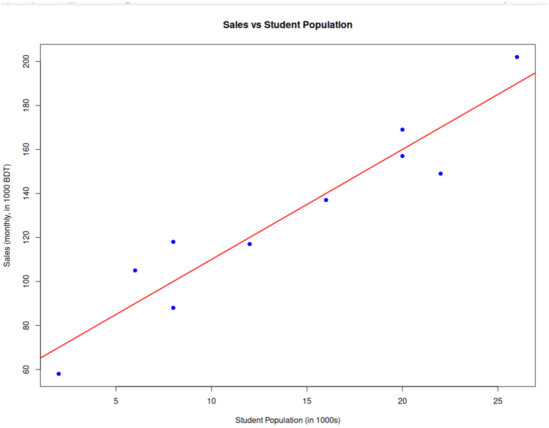


Figure 4: Scatterplot of Sales Vs. Student Population

- The equation of the line will be something like this

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

- Here $\hat{\beta}_0$ 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 ....
- 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 = 999) # turn off scientific printing

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

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

- You should see following output,

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

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 $\hat{\beta}_1 = 5$

- 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 R after plotting the scatter plot, you can plot this line using the `abline()` function.

- 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*.

- 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 $\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, 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 (you will see more details in the Econometrics course (or if you are impatient please look at the Appendix), but if we solve this minimization problem

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

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

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

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

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*).

- 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 ....

Simple Linear Regression Model (SLR)

2. Assessing the Fit - R^2 and RSE

- Now we will learn a measure which will help us to measure *how good does the linear line fit with the data*? It's a number or a summary measure called R^2. The basic formula is,

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

- where

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

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

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

- Recall the sample mean

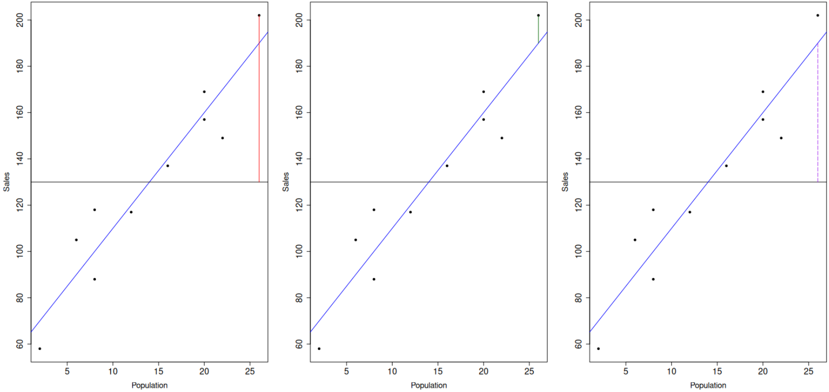
$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Question is - what does this formula mean?

To understand this first let's decompose $y_i - \bar{y}$

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

- This can be visually understood, on the left for an *ith* point, we have $y_i - \bar{y}$, then on the middle we have a residual or error $(y_i - \widehat{y}_i)$ and on the right we have $(\widehat{y}_i - \bar{y})$

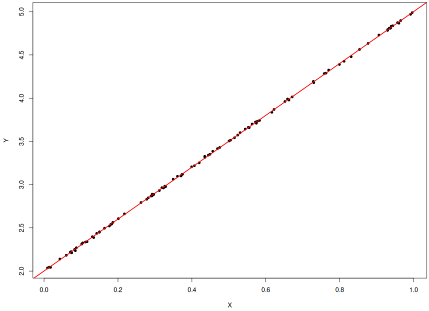


- Now we can take squares and sum on both sides of the decomposition and we get (the product term becomes 0)

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

- 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).

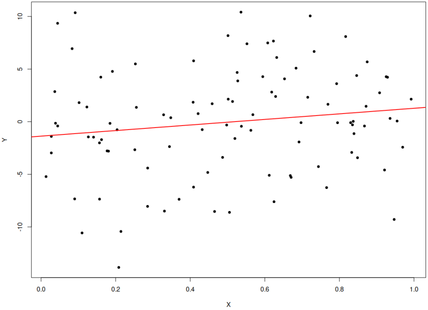
- We already know SSE, which is $\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2$. This is the sum of squared errors, or the *Error Sum of Squares* which shows how much variability of error remains after we fitted the line.
- And the term $\sum_{i=1}^{n}\left(\widehat{y}_i - \bar{y}\right)^2$ is called *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 is an example of high R^2 with a different data



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

Simple Linear Regression

Goodness of Fit or R^2



- Here is a different data, that does not show any linear pattern and we try to fit a linear line, obviously the fit won't be good, and R^2 will be low, for example consider following data If we fit a line (which is the red line), then R^2 in this case is 0.02, which is almost close to 0.

- So the above discussion shows R^2 tells us how well our least-squares line fits the data. High R^2 means the fit is quite good, on the other hand low R^2 means fit is not that good with the data.
- R^2 is also known as *Coefficient of Determination*, sometimes it is also called *Goodness of Fit*.
- In 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.
- As you already know there is an important caveat regarding R^2, since this is an in sample prediction - that is high R^2 does not automatically mean that we did a good job with our prediction problem for any data ...
- 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

Simple Linear Regression Model (SLR)

3. What is the Prediction Problem Here?

- 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.

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
11	**15**	135
12	**27**	195

Table 4: The question mark ? means we would like to predict what would be the sales if the student population is $15,000$ and $27,000$.

- *Question is - what are we predicting here?* Recall in Statistical Inference problem we always do prediction or estimation for a population quantity. *Ques.* 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 $10,0000$ restaurants or something!)

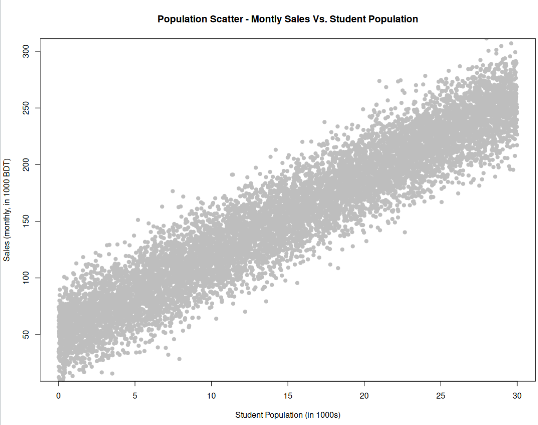


Figure 5: Scatter plot of the Population Data

- 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 6: Scatter plot of the Population Data (gray points) and Sample Data (blue points)

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

- *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

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

- Visually this means

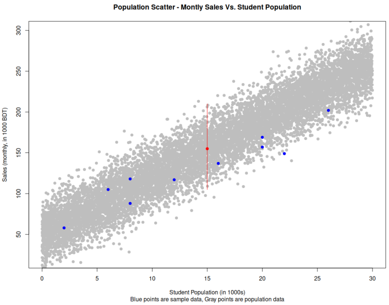


Figure 7: Scatter plot of the Population Data (gray points) and Sample Data (blue points) with the conditional expectation at $X = 15$, in this case the conditional average is 155

- It turns out that if we have the population data, this is the best prediction we can make (in some sense!)

Restaurant	SPop (x_i)	Msales (y_i)	best prediction using conditional average
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	
11	**15**	...	155

Table 5: Conditional Expectation is the ideal prediction we would like to have

- If we have the population we can calculate the conditional average at all any points, since in this case we have the population, here are the conditional average values....

Restaurant	SPop (x_i)	Msales (y_i)	best prediction: conditional average
1	2	58	64
2	6	105	92
3	8	88	106
4	8	118	106
5	12	117	134
6	16	137	162
7	20	157	190
8	20	169	190
9	22	149	204
10	26	202	232
11	**15**	...	155

Table 6: Conditional Expectation is the best prediction we would like to have

- Visually this means,

Figure 8: Scatter plot of the Population Data (gray points) and Sample Data (blue points) with the Conditional Expectation (red points) at all the sample points..

- Now if we calculate this conditional expectation at all x values in the population, and draw the line, this would give us the *conditional expectation function*, here is the line

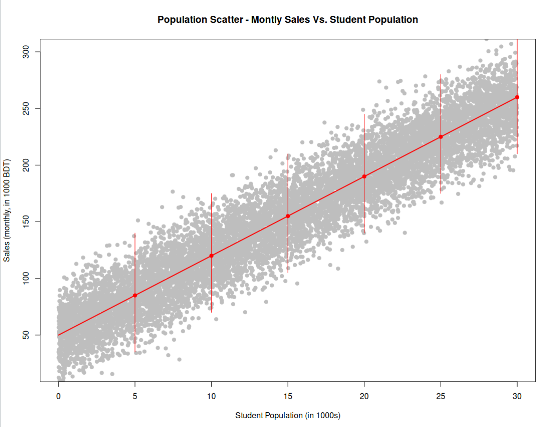


Figure 9: Scatter plot of the Population Data (gray points) the Conditional Expectation Function (red line)

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

$$\mathbb{E}(Y|X=x) = 50 + 7x$$

  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 $\mathbb{E}(Y|X=x) = 50 + 7x = 50 + 7 \times 15 = 155$, and we can predict the sales for $x = 15$ is 155.

- 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*, so in the next section we will estimate this conditional expectation function, in fact we will estimate the intercept and slope of this line. We will write the function with

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

  where β_0 is the intercept and β_1 is the slope.
- These are unknown quantities, since we don't know the population but we will see that it is possible to estimate them from the sample data.
- So the table we saw before is unrealistic, the realistic scenario is following

Restaurant	SPop (x_i)	Msales (y_i)	conditional average	estimate from data
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	??	...
11	**15**	...	???	...

Table 7: We can never calculate the conditional expectation, since we don't have the population and also we don't know the conditional expectation function, but we can estimate the conditional expectation function get the estimate from data

- We will get the prediction from the data using the *method of least squares*, which is a method to estimate the unknown β_0 and β_1 from the sample data, and then in the next section we can fill the last column...

Simple Linear Regression Model (SLR)

4. What are the Model Assumptions?

- In Statistics often there will be some assumptions about the unknown world, and the truth is nothing works if we don't have any assumption at all. This is because the real life scenarios are often so complex that it is almost impossible to learn from data without making any assumption at all. There is famous quote by George Box - "*All models are wrong, but some are useful*".

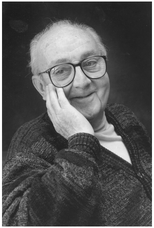


Figure 10: George Box (1919 - 2013), source - Wikipedia

- 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.
- *Assumption 2* - The population regression function or CEF is a linear function in X_i for all i (extensions possible, we will see later).

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

- *Assumption 3* - Define $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Homoskedasticity of ϵ, this means $\mathbb{V}(\epsilon_i|X_i = x) = \sigma^2$ for all x values, where σ^2 is a constant.
- *Assumption 4** - Conditional on x, ϵ_i is Normally distributed with mean 0 and variance σ^2, so we can write $\epsilon|x \sim \mathcal{N}(0, \sigma^2)$

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

Simple Linear Regression Model (SLR)

4. Assessing the Accuracy of the Estimated Coefficients

- How do we assess the *accuracy of the estimated coefficients*?
- We already know that in Statistics one way to measure the accuracy of the estimates is thinking about *random samples* or *repeated sampling*.
- *Repeated sampling* idea is very helpful, since we can think about how the values vary if we perform estimation more than once or multiple times. Here are 4 situations that may happen if we do repeated sampling and then do estimation multiple times.
- In the following suppose we are considering the parameter β_1 and $\hat{\beta}_1$ for different samples.
- The true value β_1 is at the center, and the black dots are estimates or values of $\hat{\beta}_1$ calculated from different samples.

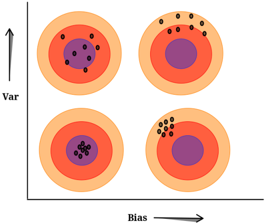


Figure 11: bias variance situations, true value β_1 is at the center, and the black dots are estimates or values of $\hat{\beta}_1$

- 1. *top-left:* Here sometimes the estimates are hitting the target, but their accuracy overall is really bad. You can say on average they are performing well, but there is a lot of variability. This is what we call *low-bias & high-variance* situation.
- 2. *bottom-left:* This is better than the last one (in fact this is the best one) here estimates are always very close to the truth and also the variability is very low. This is what is called *low-bias & low-variance* situation. This is ideally what we want.
- 3. *bottom-right:* In this case the variability is not high, but the estimates are more or less always very off from the target. This is called *high-bias & low-variance* situation. This is not good, even if we have low variance.

- 4. *top-right:* This is the worst case, here the estimates are always very off from the target and also the variability is very high. This is called *high-bias & high-variance* situation.

- Recall whenever we think about random sample or repeated sampling automatically the idea of *estimator* comes. And an estimator nothing but a random variable (or the formula) that we think whenever we are thinking about repeated sampling.
- Here we have *two estimators*,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \tag{4}$$

- For a fixed sample when we calculate the values applying this formula we get an estimate (in the last slides the black dots are different estimates).
- The formula is exactly same as (**??**), but we used uppercase letters to specify that now we are thinking these quantities as random variables or estimators or sample statistics, where the values may change if we have a different sample.
- So $\widehat{\beta}_1$ is now a random variable. This means for a fixed sample (for example the data set that we used) it will give us one possible value. But if we change the sample, and use a different sample, the value will change. Same interpretation can be given for $\widehat{\beta}_0$.

- Usually the bias variance picture that we saw is used to explain the quality of an estimator. For example think about $\hat{\beta}_1$ is an estimator.
- When we say an estimator has *low bias* this means, *on average* the values or the estimates will be *close to the true value*. And When we say an estimator has *low variance* this means, the values or the estimates *will not vary much*.
- We already saw the idea of bias and variance, now we can re-write the results using an estimator. In this case you can think about our parameter is β_1 and $\hat{\beta}_1$ is an estimator. But this is can be understood with any parameter and an estimator.
  - 1. *top-left:* The estimator has *low-bias & high-variance*.
  - 2. *bottom-left:* The estimator has *low-bias & low-variance*.
  - 3. *bottom-right:* The estimator has *high-bias & low-variance*.
  - 4. *top-right:* The estimator has *high-bias & high-variance*.

- Definitely we desire an estimator to be unbiased, for example, for example for the parameter β_1 and estimator $\hat{\beta}_1$ if the following holds we say $\hat{\beta}_1$ is an unbiased estimator for β_1,

$$\mathbb{E}(\hat{\beta}_1) = \beta_1$$

- Similarly if we have $\mathbb{E}(\hat{\beta}_0) = \beta_0$, we say $\hat{\beta}_0$ is an unbiased estimator for β_0.
- If we assume we have linear model assumption and the random sample is an iid random sample it is possible to show that the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Also we can show that the least square estimators have low variance. We will not go into the details of the proof of this claim, but you will see the details about these results in the Econometrics course.
- Under *homoskedasticity* assumption we can show that the variance of the least square estimators are

$$\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad \text{and} \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{5}$$

- The square root of these quantities are the *standard errors*.

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]} \quad \text{and} \quad \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- Actually in reality we don't know σ^2, but we can use MSE, which in an estimator for σ^2,

- Here an important point is, under homoskedasticity assumption MSE in repeated sampling is an unbiased estimator for σ^2, so $\mathbb{E}(\mathrm{MSE}) = \sigma^2$.
- Now we can replace σ^2 with MSE and we can get an estimate of the standard errors.

$$\widehat{\mathrm{SE}}(\hat{\beta}_0) = \sqrt{\mathrm{MSE}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \quad \text{and} \quad \widehat{\mathrm{SE}}(\hat{\beta}_1) = \sqrt{\frac{\mathrm{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Now using these estimated standard errors we can construct confidence intervals. For example a 95% confidence interval of β_1 would be

$$\left[\hat{\beta}_1 - 1.96 \cdot \widehat{\mathrm{SE}}\left(\hat{\beta}_1\right) \quad , \quad \hat{\beta}_1 + 1.96 \cdot \widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right]$$

- We will usually omit "hat" symbol and write

$$\left[\hat{\beta}_1 - 1.96 \cdot \mathrm{SE}\left(\hat{\beta}_1\right) \quad , \quad \hat{\beta}_1 + 1.96 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

- Here 1.96 is the 97.5 percentile of the normal distribution, we can also use t distribution under the normality assumption of ϵ.
- For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$
- SideNote: Strictly speaking, we need to think about the sampling distribution of $\hat{\beta}_1$ under the distributional assumption of ϵ, or large sample assumption. But I am avoiding the technical details here. You will see that in the Econometrics course.

Simple Linear Regression Model (SLR)

5. Significance Testing

- 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

  H_0 : There is no relationship between X and Y versus the alternative hypothesis

  H_a : There is some relationship between X and Y

- Mathematically, this corresponds to testing (note that it is a two tail test)

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

- This is because if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y. We can do the t-test here by calculating the value of the t-statistic, which is

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

- Under the Null, this will have a t-distribution with $n-2$ degrees of freedom, or with large sample we can also use Normal distribution.
- Then we can do the test using the critical value approach or p-value approach.
- For example, for regressing Sales on TV (regression result in page 42), the estimate of the standard error for $\hat{\beta}$ is 0.00269 and the value of the t-statistic is 17.67, and we can see that p value is almost close to 0.

- Using this we see that, for this testing we can reject the Null at $\alpha = 0.01$ or bigger.
- If we reject the Null then we say *statically there is a significant relationship between the variable X and Y*
- You should be able to do the test for yourself, only from the $\hat{\beta}_1$ and estimate of the standard error of $\hat{\beta}_1$, it is possible to calculate the value of the *t*-statistic.
- In the R output, 43 it is also possible to read this information using $*$, $**$ or $***$ (How?)

- There is another approach of doing significance testing. This approach is known as *analysis of variance* (in short ANOVA) approach. In this approach we will F-test.
- Following is the ANOVA table for the Sales Vs. TV problem

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TV	1	3314.62	3314.62	312.14	0.0000
Residuals	198	2102.53	10.62		

- You can just run the function `anova()` in R to get this table.
- Let's explain this table,

Simple Linear Regression

Significance Testing - F test

SSR

Df for SSR

$\text{MSR} = \dfrac{\text{SSR}}{\text{Df for SSR}}$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TV	1	3314.62	3314.62	312.14	0.0000
Residuals	198	2102.53	10.62		

p value for F statistic

SSE

Df for SSE

$\text{MSE} = \dfrac{\text{SSE}}{\text{Df for SSE}}$

$\text{F}_{\text{calc}} = \dfrac{\text{MSR}}{\text{MSE}}$

name of the regressor

- The first column is the *source of variation*. In this case we have two sources of variation, one is the *regression*, which is written with `TV` and the other is the *residuals* (or error).
- The second column is the *degrees of freedom* (Df). In this case we have 1 Df for the regression and 198 Df for the residuals (We will see why in a minute).
- The third column is the *sum of squares* (SS). Here we have two sum of squares $\text{SSR} = 3314.62$ and $\text{SSE} = 2102.53$ for the residuals. Note that, in this case we can automatically calculate SST (how?)

- The fourth column is the *mean sum of squares* (MS). The first one is the mean squared regression

$$\text{MSR} = \frac{\text{SSR}}{\text{Df of SSR}} = 3314.62$$

and the second one is

$$\text{MSE} = \frac{\text{SSE}}{\text{Df of SSE}} = 10.62$$

- The fifth column is the *F statistic*. We will use this statistic to do another test of significance. Here the value of the statistic is

$$F = \frac{\text{MSR}}{\text{MSE}}$$

- The sixth column is the *p-value* for the F statistic. In this case the p-value is almost close to 0.

- Now let's explain the F-test. First note that, in this case, we are still doing the same test

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

- And the testing procedure of the F test is same as t-test, when p-value $< \alpha$ we reject the Null, so in this case we can reject the Null.
- Now let's understand why F test works.
- Recall, we know that $\mathbb{E}(\text{MSE}) = \sigma^2$ (this was an unbiased estimator)!
- Now it is possible to also show that under the Null (I am skipping the detailed calculations)

$$\mathbb{E}(\text{MSR}) = \sigma^2$$

- This means if $\beta_1 = 0$, then

$$\mathbb{E}(\text{MSR}) = \mathbb{E}(\text{MSE}) = \sigma^2$$

- So under the Null, we may expect that the value of MSE will be close the value of MSR and the value of the F statistic is close to 1.
- This means larger values of F means higher chances of rejecting null $H_0 : \beta_1 = 0$.
- We can use F distribution to do the test. Note that F distribution is an asymmetric distribution, and F test is an upper tail test.

- So we will reject the Null if $F_{calc} > F_{1-\alpha}$, where $F_{1-\alpha}$ is the $1-\alpha$ percentile of the F distribution with 1 and $n-2$ degrees of freedom.
- But it is easy to do the test using p-value, because we already know the p-value.

- Now let's explain how did we calculate the Df in SS.
- In one line - *the degrees of freedom are the number of independent components that are needed to calculate the respective sum of squares*

- The total sum of squares, $\mathrm{SST} = \sum (y_i - \bar{y})^2$, is the sum of n squared components. However, since $\sum (y_i - \bar{y}) = 0$, only $n-1$ components can independently come in the calculation. The n^{th} component can always be calculated from $(y_n - \bar{y}) = -\sum_{i=1}^{n-1} (y_i - \bar{y})$. Hence, SST has $n-1$ degrees of freedom.

- $\mathrm{SSE} = \sum e_i^2$ is the sum of the n squared residuals. However, there are two restrictions among the residuals, coming from the two normal equations (you can think we are estimating two quantities $\widehat{\beta}_0$, $\widehat{\beta}_1$). So it has $n-2$ degrees of freedom.
- And we will always have,

$$\text{Df of SST} = \text{Df of SSE} + \text{Df of SSR}$$

- So for SSR the Df will be

$$\text{Df of SSR} = (n-1) - (n-2) = 1$$

Simple Linear Regression Model (SLR)

5. Prediction: Confidence Intervals and Prediction Intervals

- Recall using the estimated line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we can easily get a *point estimate of* $f(x^*)$ for any new point x^* by $\hat{f}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$
- For example, recall for the advertisement data where our *estimated regression function* was $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.03 + 0.048x_i$. Now suppose $x^* = 100$. This means we want to predict the sales for a TV advertising budget of \$100,000. We can see the predicted sales would be $11.786 \times 1000 = 11,786$ units. This is because $7.03 + (0.048 \times 100) = 11.786$
- Now how good is our prediction for the CEF at this new point.
- To answer the first question we can construct confidence intervals of mean at x^*, or confidence intervals around $f(x^*) = \mathbb{E}(Y|X = x^*) = \beta_0 + \beta_1 x^*$
- We will skip the derivation (see Abraham and Ledolter (2006) page 36 for a derivation if you are interested) but a $100(1-\alpha)$ percent confidence interval for $f(x^*)$ at a new point x^* is given by

$$\hat{f}(x^*) \pm t_{1-\alpha/2,\, n-2} \times \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]}$$

- where

$$\mathrm{SE}\left(\hat{f}(x^*)\right) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]}$$

- So if these confidence intervals are narrow, that means $\hat{f}(x^*)$ will be to close to $f(x^*)$ in repeated sampling, and we have more precision in the estimation of $f(x^*)$
- If these intervals are wider, this means there is a lot of uncertainty about the prediction.
- This confidence interval is what we call *confidence interval for the mean at a new point* x^*.
- You don't have to memorize the formula, it is very easy to construct this interval in R.

- There is another kind of uncertainty that we can consider, that is *how good can we predict the unknown response Y^* at a new point x^**?
- We can use *prediction intervals* to answer this question.
- Prediction intervals are intervals of the random Y^* at a new point x^* (recall at x^* there are many possible values of Y^*)
- $100(1-\alpha)$ percent prediction interval for Y^* at a new point x^* is given by

$$\hat{f}(x^*) \pm t_{1-\frac{\alpha}{2},n-2} \times \sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

- The uncertainty in this case will be definitely higher. This is because even if we knew $f(x)$ that is, even if we knew the true values for β_0 and β_1, the response value cannot be predicted perfectly because of the random error ϵ in the model.
- Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

Appendix

Appendix

1. Proof OLS Estimators for Simple Linear Regression

Proof OLS Estimators for Simple Linear Regression

Recall the problem is

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

There are different ways to solve it, here are the details for at least one,

- 1. Start with the sum of squared errors, or SSE
- 2. Take partial derivatives with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- 3. Set the derivatives equal to zero to get the normal equations (these are called normal equations)
- 4. Solve the first normal equation for $\widehat{\beta}_0$, which gives „,
- 5. Substitute $\widehat{\beta}_0$ into the second normal equation and manipulate the resulting equation to solve for $\widehat{\beta}_1$, leading to the formula involving the sums of products of deviations and squares of deviations.

Proof OLS Estimators for Simple Linear Regression

We start with partial derivatives

$$\frac{\partial \mathrm{SSE}}{\partial \widehat{\beta}_0} = -2\sum_{i=1}^{n}\left[y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right] = 0$$

$$\frac{\partial \mathrm{SSR}}{\partial \widehat{\beta}_1} = -2\sum_{i=1}^{n} x_i\left[y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right] = 0$$

Start with the first one, this simplifies to

$$\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

Dividing by n and solving for $\bar{\beta}_0$:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Substituting $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ into the equation:

$$\sum_{i=1}^{n} x_i\left[(y_i - \bar{y}) - \widehat{\beta}_1 (x_i - \bar{x})\right] = 0$$

This simplifies to:

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

Proof OLS Estimators for Simple Linear Regression

Solving for $\widehat{\beta}_1$:

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

Appendix

2. Other Technical Details

Some Technical Details

- Just using the definition of conditional variance, we can also show that

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

$$\begin{aligned}
\epsilon &= Y - f(X) \\
\mathbb{E}(\epsilon|X = x) &= \mathbb{E}(Y - f(X)|X = x) \text{ [take cond. expec. on both sides]} \\
&= \mathbb{E}(Y|X = x) - \mathbb{E}(f(X)|X = x) \text{ [expectation of sums = sum of expectations]} \\
&= f(x) - f(x) \text{ [conditioning means fixing so } f(X = x) = f(x)\text{]} \\
&= 0
\end{aligned}$$

References

Abraham, B. and Ledolter, J. (2006), *Introduction to Regression Modeling*, Duxbury applied series, Thomson Brooks/Cole, Belmont, CA.

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.