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 6, 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 not always theory, it actually gives different perspectives to understand many real life situations or problems. Although we need to understand some theory but at the same time we need to understand how to apply them to real life situations....

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

▶ So 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...

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 *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.

The Regression Problem
A Motivating Example

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

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

- 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.... ???

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

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

▶ 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

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 $x_1, x_2, \ldots, x_n$ 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

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

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

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

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

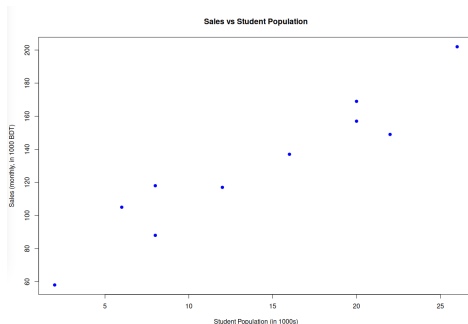▶ Next we can have a loot at the scatterplot, here is the scatter plot in ®

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.

▶ 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

Simple Linear Regression
The Problem of Estimation (method of least squares)

▶ 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.

▶ Here we will estimate a linear line that best fits the data.

▶ So we take the following data and in the SLR estimation we find a line.

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)

▶ Recall the Scatter plot

Figure 3: Scatterplot of Sales Vs. Student Population

Simple Linear Regression
The Problem of Estimation (method of least squares)

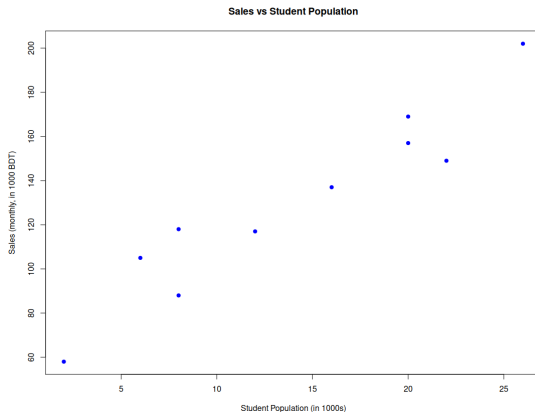▶ Our goal is to find the following red line - which is known as *best fitted line*
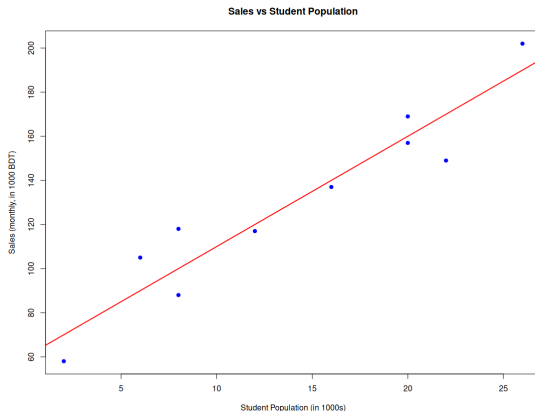

Figure 4: Scatterplot of Sales Vs. Student Population

Simple Linear Regression
The Problem of Estimation (method of least squares)

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

ℝ code: SLR results for the Armands data

```r
# 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
```

Simple Linear Regression
The Problem of Estimation (method of least squares)

- We will often write the estimates with hat symbol for these numbers.. So $\hat{\beta}_0$ and $\hat{\beta}_1$. Using the data we found $\hat{\beta}_0 = 60$ and $\hat{\beta}_1 = 5$.

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

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

- We can plot the fitted line with the data, this is the red line in the following figure.

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

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

Simple Linear Regression
The Problem of Estimation (method of least squares)

▶ Question is - Why the name *best fitted line*, what is the meaning of "best" or *how did we calculate* 5 *and* 60? Let's explain this now

▶ Essentially here best means minimizing *sum of squared errors* or SSE in the sample. What is SSE?

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

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

▶ the squared error is

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

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

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

▶ So no we can write the problem clearly, *our problem is we need to find a line which minimizes SSE*

Simple Linear Regression
The Problem of Estimation (method of least squares)

▶ Since we are fitting linear line, this means we need to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that SSE is minimized.

▶ We can write this as a following optimization (in particular minimization) problem

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

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

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

▶ We gave $*$ to represent that these are optimal points, usually we will omit $*$.

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

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

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

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

▶ Recall, there is a difference between sample variance and population variance.

▶ This method is famously known as *method of least-squares* and the fitted line is called the *least squares line* (often also called *estimated regression line* also *sample regression function*).

▶ Finally we write the formula for the estimated coefficients again

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

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

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

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

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

- What is the interpretation of the estimated coefficients?

- For the interpretation the units are very important. Recall, the data of the student population is in 1000s, and the data of sales is in 1000$

- What is the interpretation of the the *slope co-efficient* $\widehat{\beta}_1 = 5$? Here is how we can interpret,

  if the student population is increased by 1000, then approximately the sales is predicted to increase by 5000$ units.

- or

  An additional increase of 1000 student population is associated with approximately 5000$ units of additional sales.

Simple Linear Regression
The Problem of Estimation (method of least squares)

▶ What is the interpretation of the the *intercept co-efficient* $\widehat{\beta}_0$? if the student population is 0, then the predicted sales is 60, 000\$. This kind of interpretation for intercept often doesn't make any sense unless we come up with a story. So we often avoid interpetating the intercept co-efficient.

▶ Using the estimated regression line we can also get in-sample predicted values, these are also sometimes called *fitted values*.

▶ We can calculate the fitted values using the estimated regression equation,
$\hat{y}_i = 60 + (5 \times x_i)$.

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

- in ℝ you can get the fitted values with the command `fitted(model_result)`

- These fitted values values are within the sample data points, so this is why we call this *in-sample prediction*. But we can also do *out-of-sample prediction*.

Simple Linear Regression
The Problem of Estimation (method of least squares)

▶ For example we can also use the estimated regression line to predict the sales when the population is 30 thousands (notice 30 is not in the sample, nor in the range).

▶ If we do this we get $60 + (5 \times 30) = 210$ 1000\$ sales.

▶ So this is a predicted value for which we don't know y_i.

▶ We need to be careful on out of sample prediction, if the fit is good, our estimated function will always give very good in-sample prediction, but it does not automatically mean for an unseen data we we will also get good out-of-sample prediction.

▶ There is a way we can evaluate out-of-sample prediction, using *training and test sample*. We will see this in our lab session.

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

Simple Linear Regression Model (SLR)

2. Assessing the Fit - R^2 and RSE

Simple Linear Regression
Goodness of Fit or R^2

▶ So far we are fitting a line? Question is - *how good does the linear line fit with the data*?

▶ Actually there is a quantity / summary measure R^2, which answers this question for us. The formula for R^2 is

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

▶ where

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

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

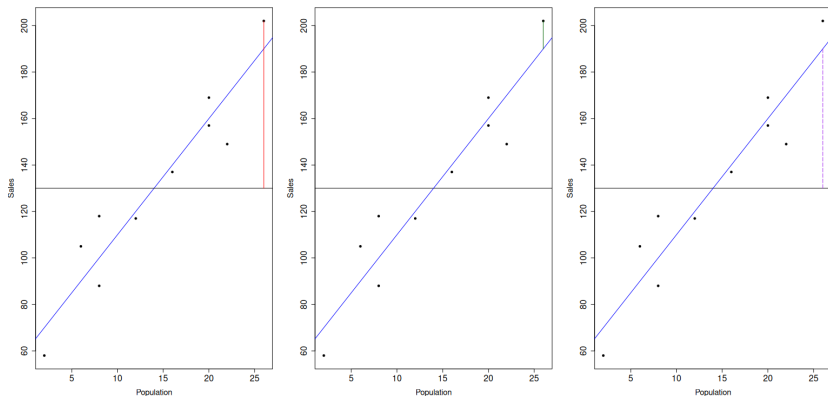$$\text{SSR} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2, \text{ Regression Sum of Squares}$$

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

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

▶ This can be visually understood

Simple Linear Regression

Goodness of Fit or R^2

▶ 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 $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$. 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 $\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$ is called *Regression Sum of Squares* or SSR in short, which shows how much variability of y_i is explained by the regression or can be explained by x_i.

▶ So this means R^2 tells "*out of the total variation of y how much we can explain by regression*".

Simple Linear Regression
Goodness of Fit or R^2

▶ Also note R^2 is a ratio of explained sum of squares and total sum of squares. So this means we will always have $0 \leq R^2 \leq 1$ (in other words the value of R^2 will always lie between 0 and 1).

▶ So high R^2 means the least-squares line fits very well with the data. Here is an example of high R^2 with a different data
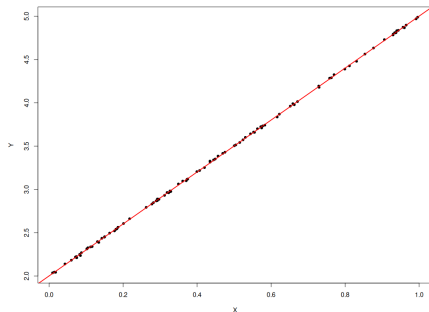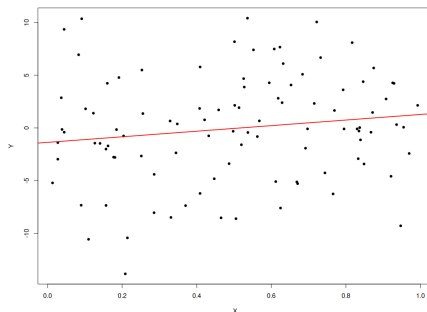

▶ The black dots are the sample points, the red line is the fitted line. Here R^2 is 0.99.

Simple Linear Regression
Goodness of Fit or R^2

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

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

Simple Linear Regression
Goodness of Fit or R^2

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

Simple Linear Regression
Residual Standard Error

- Let's think about the variance of ϵ again. For now assume homoskedasticity, which means $\mathbb{V}(\epsilon) = \mathbb{V}(\epsilon|X = x) = \sigma^2$, where σ^2 is some constant. So we can think about the unconditional variance $\mathbb{V}(\epsilon)$

- In the following we plotted same figure we plotted before.

- Recall on the left $\mathbb{V}(\epsilon)$ is high and on the right $\mathbb{V}(\epsilon)$ is low
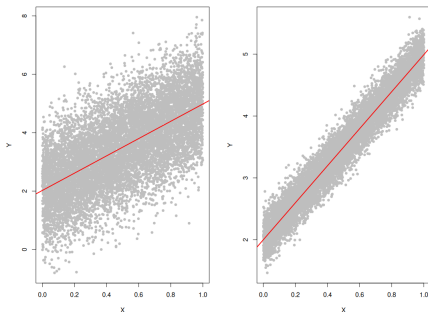
- It's important to understand that high variance of ϵ indicates our lack of certainty in prediction. Why? Because ϵ is the error that remains after we do prediction using CEF. So if there is a lot of noise, even if we use CEF, we won't be able to predict well.

Simple Linear Regression
Residual Standard Error

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

- Here is an estimate, it's called MSE,

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

- The last equality holds because we can show that $\bar{e} = 0$ (you can check this with the data!)

- Since this is an estimate of the variance of ϵ, we can take square root of this and get an estimate of the standard deviation of ϵ, which is called *Residual Standard Error* or *Standard Error of the Estimate*.

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

- So this gives an estimate of σ. If this is high we may conclude our uncertainty of prediction is high. If this is low, this is good for our prediction.

- So just to clearly mention again, for a fixed sample, MSE is an estimate of σ^2 and $\sqrt{\text{MSE}} = \text{RSE}$ is an estimate of σ.

- In the case of the advertising data, we see from the linear regression output in page 7 that the RSE is 3.26. How to interpret this?
  - One way to interpret this is - sales deviate from the regression line by approximately 3,260 units, on average.
  - Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.

- Is this an acceptable error? this depends on the problem context. In the advertising data set, the mean value of sales is approximately 14,000 units, and so the percentage error is $3,260/14,000 = 23$, so you might conclude the error is quite large.

- So high RSE is considered *a lack of fit*.

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.

Simple Linear Regression Model (SLR)

3. What are the Model Assumptions?

Simple Linear Regression
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 5: 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, $\{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\}$. 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).

$$\mathbb{E}(Y_i | X_i = x) = f(x) = \beta_0 + \beta_1 x \tag{4}$$

- ▶ *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

Simple Linear Regression
Assessing the Accuracy of the Coefficient Estimates

- 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.
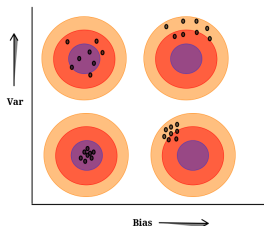
Figure 6: bias variance situations, true value β_1 is at the center, and the black dots are estimates or values of $\hat{\beta}_1$

▶ 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.

Simple Linear Regression
Assessing the Accuracy of the Coefficient Estimates

▶ 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*,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \tag{5}$$

▶ 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 (3), 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*.

Simple Linear Regression
Assessing the Accuracy of the Coefficient Estimates

▶ 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

$$\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \quad \text{and} \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (6)$$

▶ 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,

Simple Linear Regression
Assessing the Accuracy of the Coefficient Estimates

▶ Here an important point is, under homoskedasticity assumption MSE in repeated sampling is an unbiased estimator for σ^2, so $\mathbb{E}(\text{MSE}) = \sigma^2$.

▶ Now we can replace σ^2 with MSE and we can get an estimate of the standard errors.

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

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

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

▶ We will usually omit "hat" symbol and write

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

▶ 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

Simple Linear Regression
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

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

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

▶ 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.

Simple Linear Regression

- ▶ 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?)

Simple Linear Regression
Significance Testing - F test

▶ 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

- 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 SSR = 3314.62 and 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.

Simple Linear Regression
Significance Testing - F **test**

▶ 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 $\mathbb{E}(\mathrm{MSE}) = \sigma^2$ (this was an unbiased estimator)!

▶ Now it is possible to also show that under the Null (I am skipping the detailed calculations)

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

▶ This means if $\beta_1 = 0$, then

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

▶ 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, $\text{SST} = \sum (y_i - \bar{y})^2$, 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 nth 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.

- $\text{SSE} = \sum e_i^2$ 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

Simple Linear Regression
Confidence Intervals for $f(x^*)$ at a new point x^*

▶ 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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.03 + 0.048 x_i$. 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

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

Simple Linear Regression
Confidence Intervals for $f(x^*)$ at a new point x^*

- 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 .

Simple Linear Regression
Prediction intervals for Y at x^*

- 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

Some Technical Details

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

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

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

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.