Ch3 - Simple Linear Regression

Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka

Last Updated November 25, 2023

Outline

Outline

1. **Statistical Learning and the Problem of Regression**
   - Best Function to Predict
   - Estimation of f Parametric vs. Nonparametric
2. **Simple Linear Regression Model (SLR)**
   - 1. The Problem of Estimation
   - 2. Assessing the Fit - R^2 and RSE
   - 3. What are the Model Assumptions?
   - 4. Assessing the Accuracy of the Estimated Coefficients
   - 5. Significance Testing
   - 5. Prediction: Confidence Intervals and Prediction Intervals
3. **Appendix**

Comments and Acknowledgements

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

  For theoretical discussion I primarily followed James et al. (2023). Anderson et al. (2020) is a good book and very easy to read with lots of easy examples, but James et al. (2023) is truly amazing when it comes to explaining the concepts in an accessible way. We thank the authors of this book for making everything publicly available at the website `https://www.statlearning.com/`.
- I thank my students who took this course with me in Summer 2022, Fall 2022 and currently Fall 2023. Their engaging discussions and challenging questions always helped me to improve these notes. I think often I learned more from them than they learned from me, and I always feel truly indebted to them for their support.
- You are welcome to give me any comments / suggestions regarding these notes. If you find any mistakes, then please let me know at `tanvir.hossain@ewubd.edu`.
- I apologize for any unintentional mistakes and all mistakes are mine.

Thanks,
Tanvir

1. **Statistical Learning and the Problem of Regression**
    - Best Function to Predict
    - Estimation of f Parametric vs. Nonparametric

2. **Simple Linear Regression Model (SLR)**
    - 1. The Problem of Estimation
    - 2. Assessing the Fit - R^2 and RSE
    - 3. What are the Model Assumptions?
    - 4. Assessing the Accuracy of the Estimated Coefficients
    - 5. Significance Testing
    - 5. Prediction: Confidence Intervals and Prediction Intervals

3. **Appendix**

Statistical Learning and the Problem of Regression

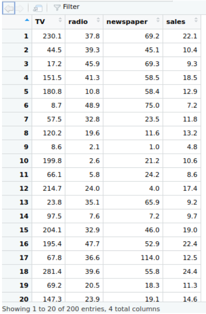
Statistical Learning and the Problem of Regression

Best Function to Predict

Best Function to Predict

Conditional Expectation Function

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



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

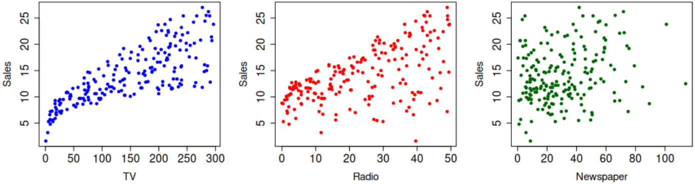
Showing 1 to 20 of 200 entries, 4 total columns

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

Best Function to Predict

Conditional Expectation Function

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



- We will see how to do scatter plots in our lab session.
- To make company's problem more concrete, we can think about a function f, which will predict sales based on TV, radio and newspaper expenditure.

$$\text{Sales } \approx f(TV, \text{Radio, Newspaper })$$

- We often call this function *a model*.

Best Function to Predict

Conditional Expectation Function

- Here Sales is a *response or target* that we wish to predict. Usually we denote the *response with Y*.

- TV, Radio and Newspaper are called *features, or inputs, or predictors or covariates*, we usually denote this with X. In this case, we have 3 features, we can refer to the inputs as X_1, X_2, and X_3 and often we refer them collectively with a vector,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

- Sometimes we also call Y as *dependent variable* and X as an *independent variable*.
- The problem is we want to predict Y with a function of X which we write as $f(X)$
- Of course our prediction will not be 100% accurate since we may have measurement errors or leave other variables in our model.
- This will lead to some errors in the prediction. Let's denote the error with ϵ, where $\epsilon = Y - f(X)$, this is called the *CEF error*.
- With this we can think about the complete model with

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

- Now there are many possible options, so quite naturally our question is *what is the "best" possible function of X* that we can use to predict Y?

Best Function to Predict

Conditional Expectation Function

- First let's define *what do we mean by "best"* here?. Here by "best" we mean minimizing the *mean squared error* (in short MSE). MSE is defined as $\mathbb{E}\left[\left(Y - f(X)\right)^2\right]$.
- So now we can rephrase the question -

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

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

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

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

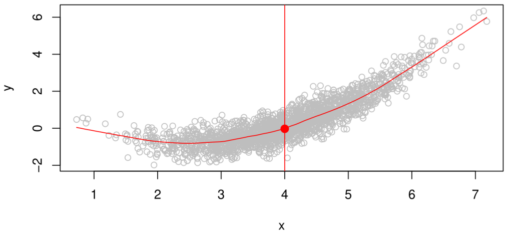
$$f(x) = \mathbb{E}(Y \mid X = x)$$

- This means conditioning on the value of X (conditioning means "fixing"), we are taking average of Y values. What does conditional expectation function mean visually?

Best Function to Predict

Conditional Expectation Function

- Let's explain this with a single variable. Suppose we have only one variable now, so TV expenditure and we wish to find the best possible function for X
- How does this looks like? First, recall *Expectation is population concept* so we need to bring population to explain the concept Conditional Expectation.
- Suppose following is the scatter plot of the population data of all sales and TV expenditures of the company



- Then at $X = 4$ there can be many Y values but if we take expectation, we get the red dot in the picture, mathematically we can write, $\mathbb{E}(Y \mid X = 4)$
- This is conditional expectation (or you can think average) of Y at $X = 4$ (or when we think the value of X as 4)

Best Function to Predict

Conditional Expectation Function

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

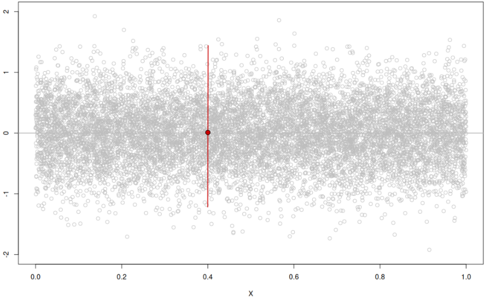
$$f(x) = \mathbb{E}(Y \mid X = x)$$

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

Understanding ϵ

CEF Error - Mean and Variance

- We need to mention some important points regarding the CEF error, particularly the conditional expectation and conditional variance of the CEF error.
- We write conditional expectation of error with $\mathbb{E}(\epsilon|X=x)$ and conditional variance with $\mathbb{V}(\epsilon \mid X=x)$.
- First of all it is easy to show that $\mathbb{E}(\epsilon|X=x)=0$ for all x (see Appendix). Here x can be any fixed value, for example 0.4 in the figure. In this setting when we are modeling CEF, this will always hold.
- What does this mean visually? Consider following population data of ϵ



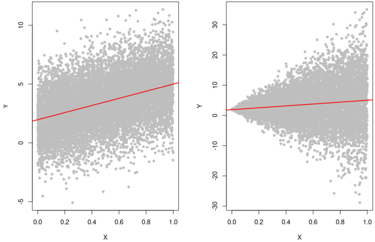
Understanding ϵ

CEF Error - Mean and Variance

- Here we plotted x values on the x-axis and ϵ values on the y-axis. So for each x value, we have many ϵ values and the figure shows if we take average of these ϵ values at every x, then the average will be 0 at every x.
- If $\mathbb{E}(\epsilon|X = x) = 0$, it is possible to show that unconditional expectation $\mathbb{E}(\epsilon) = 0$ (this is an application of law of iterated expectation, but we will not go into details here)
- We can also talk about conditional variance $\mathbb{V}(\epsilon \mid X = x)$ of the CEF error.
- Just using the definition of variances, we can show that conditional variance of ϵ is equal to conditional variance of Y.
- So this means $\mathbb{V}(\epsilon \mid X = x) = \mathbb{V}(Y \mid X = x)$ and also unconditional variance are also equal $\mathbb{V}(\epsilon) = \mathbb{V}(Y)$
- So what is conditional variance? It means variance of Y, conditional on x values. In the following we plotted two *population data* where the red line is the CEF function.

Understanding ϵ

CEF Error - Mean and Variance

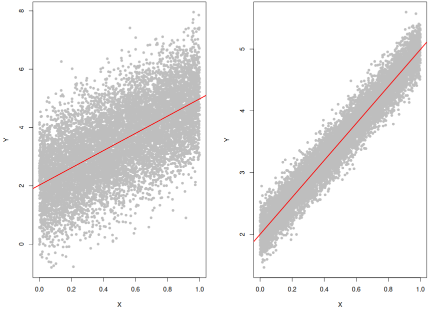


- One the left the variance of Y seems to be constant with x values, so this means $\mathbb{V}(Y|X)$ or $\mathbb{V}(\epsilon|X)$ is constant. This is called *homoskedasticity*!
- On the right the variance of Y is changing with x values (in particular increasing), so this means $\mathbb{V}(\epsilon|X)$ is NOT constant, it is called *heteroskedasticity*!

Understanding ϵ

CEF Error - Mean and Variance

- Now again consider two population data, for both $\mathbb{V}(\epsilon|X = x)$ is constant. But on the left $\mathbb{V}(\epsilon|X = x)$ is high and on the right $\mathbb{V}(\epsilon|X = x)$ is low



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

Statistical Learning and the Problem of Regression

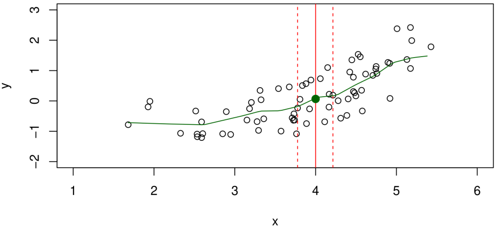
Estimation of f Parametric vs. Nonparametric

Estimation of f Parametric vs. Nonparametric

- Is it possible to know the population CEF $f(x)$? The answer is NO, why? - because usually we never have access to the population data (unless in some very special cases).
- Can we *estimate* it? The answer is YES we can.
- There are two kinds of methods for estimating this function,
  - *Parametric* method - imposing structural assumption on the function
  - *Non-Parametric* method - data driven method, but requires tuning parameters and often need lots of data for good performance.

Estimation of f Parametric vs. Nonparametric

- In this course we will only consider parametric methods of regression. But now we will just give a short overview of a simple **Nonparametric Method** known as **Nearest Neighbor Method**.
- In general in any nonparametric method, we will *not impose any functional form on f*, and the entire method will be a data driven method, but there will be *external or tuning parameters*, which often influences the quality of the estimation.
- Here is a nonparametric way to estimate f? The idea is *rather than taking average at x, we will take average on the neighbors of x*. This is called the *Nearest Neighbor Method*
- Why nearest neighbors? First of all note that typically *in a sample* we have very few data points at x, and sometimes no data points at x (think about $X = 4$), look at the following picture (note this is a scatter plot of a sample taken from the population data in page 10)



Estimation of f Parametric vs. Nonparametric

- Here at $X = 4$ we don't have any point, so we cannot compute $\mathbb{E}(Y \mid X = x)$, but what we can do is we can take *average of the nearest neighbors*, which we write as,

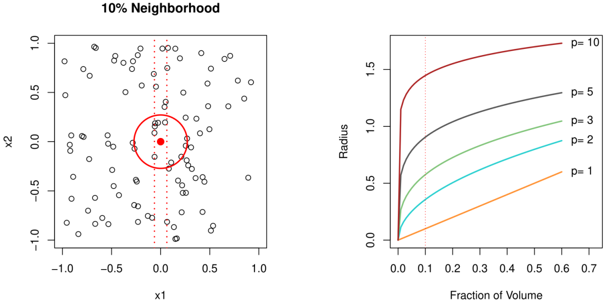
$$\widehat{f}(x) = \text{Ave}(Y \mid X \in \mathcal{N}(x))$$

- here $\widehat{f}(x)$ is the estimate of $f(x)$, and $\mathcal{N}(x)$ means the neighbors around x
- Here the tuning parameter is how many neighbors we will consider, or how we will define the neighbors.
- Notice if we take the neighbors as all data points, we will get a horizontal straight line, but if we get the neighbor very small, then the curve will be very wiggly.
- So tuning parameters influence the performance, and this is something we give.
- Nearest neighbor averaging can be good for small p - i.e. $p \leq 4$ and large sample size n, but it can give poor performance when p is large.
- Reason: *the curse of dimensionality*, which means the neighbors tend to be far away in high dimensions, and we need lots and lots of data for good performance.
- *Curse of dimensionality* is typically THE problem of nonparametric methods in general.
- Following picture might give you an idea

Estimation of f Parametric vs. Nonparametric

Figure 1: In the left plot we have only two features, the red point is the point for which we will find nearest neighbor, the red circle shows 10% of the data points as nearest neighbors. Here is the radius of the circle
is roughly 0.4.

In the right, the plot shows how we need to increase the radius to get 10% of the data points as we increase the dimension or the number of features. For example if we have 10 covariates, to get 10% of the data points as neighbors, the radius has be almost 1.5



Estimation of f Parametric vs. Nonparametric

- Let's talk about **Parametric Methods** now.
- **Parametric Methods** generally we will impose assumptions on the structure of f. For example suppose we assume our model is a linear function of X (extensions possible, we will come back to the extension later!)

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

- Here we are assuming the unknown CEF is linear and *β_0 and β_1 are population parameters*.
- Why linear model? Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.
- Once we assume the unknown CEF is linear, the remaining work is to get the parameters β_0 and β_1, if we want to do prediction.
- We can also assume the model has p variables, where X is a p-dimensional vector so, $X = (X_1, X_2, \ldots, X_p)$

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- In the next section we will start talking about *Simple Linear Regression* problem, where we will assume we have only one covariate to predict the response and the CEF has a linear form.
- In the next chapter (Chapter - 4) we will start talking about *Multiple Linear Regression*, where we have many covariates to predict Y.

1. **Statistical Learning and the Problem of Regression**
   - Best Function to Predict
   - Estimation of f Parametric vs. Nonparametric

2. **Simple Linear Regression Model (SLR)**
   - 1. The Problem of Estimation
   - 2. Assessing the Fit - R^2 and RSE
   - 3. What are the Model Assumptions?
   - 4. Assessing the Accuracy of the Estimated Coefficients
   - 5. Significance Testing
   - 5. Prediction: Confidence Intervals and Prediction Intervals

3. **Appendix**

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 parametric method is known as *Linear Regression Model*. In particular we will talk about *Simple Linear Regression Model* or in short SLR in this chapter.

- According to Simple Linear Regression Model we will assume the unknown CEF is linear in X and we have just one feature X.

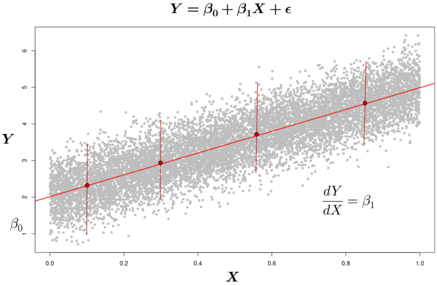
- This means we will assume the true and unknown CEF $f(X)$ has following form,

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

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

Simple Linear Regression

The Problem of Estimation (method of least squares)



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

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

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

- It's obvious that just from the sample we can never get β_0 and β_1, so what do we do? We try to guess the values from a sample data. You should immediately recognize this an *estimation problem*.

- We explain the estimation method here with a concrete example from Anderson et al. (2020). Keep in mind, our goal is to estimate the unknown β_0 and β_0 using a sample.

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

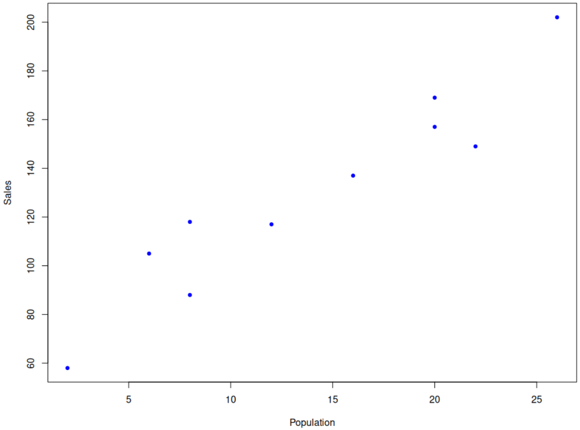
Restaurant	Population (x_i), in 1000s	Sales (y_i), in 1000\$
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

Figure 2: Scatterplot of Armand's Pizza Parlor data from Anderson et al. (2020)



Simple Linear Regression

The Problem of Estimation (method of least squares)

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

R code: SLR results for the Armands data

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

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

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

- You should see following output,

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.0000    9.2260   6.503 0.000187 ***
Population   5.0000    0.5803   8.617 0.0000255 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 13.83 on 8 degrees of freedom
Multiple R-squared: 0.9027,^^IAdjusted R-squared: 0.8906
F-statistic: 74.25 on 1 and 8 DF, p-value: 0.00002549
```

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

Table 1: Regression results of Quarterly Sales on Population

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

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

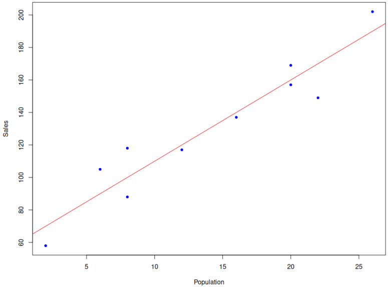
Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

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



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

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

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

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

- 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*

Simple Linear Regression

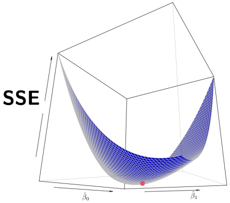
The Problem of Estimation (method of least squares)

- 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

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

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

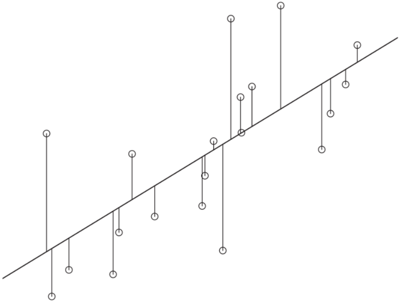


- Why? Because you can think *SSE* is a function of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and we are looking for the optimal $\widehat{\beta}_0$ and $\widehat{\beta}_1$ which will minimize this function.

Simple Linear Regression

The Problem of Estimation (method of least squares)

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



Simple Linear Regression

The Problem of Estimation (method of least squares)

- So following is the minimization problem

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

- 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

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

- 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

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

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

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

- 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,

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

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

- What is the interpretation of the the *slope co-efficient* $\widehat{\beta}_1 = 5$? 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.*

- 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, $\hat{y}_i = 60 + (5 \times x_i)$.

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

- in 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)

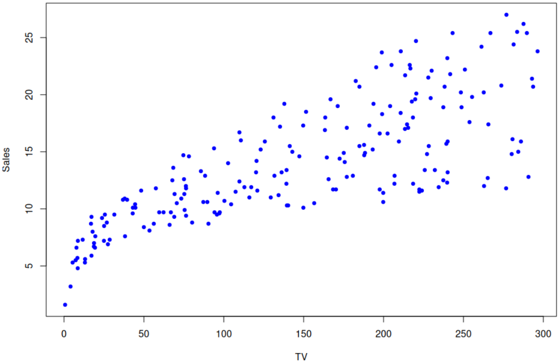
- For example we can also use the estimated regression line to predict the sales when the population is 30 thousands (notice 30 is not in the sample, nor in the range).
- If we do this we get $60 + (5 \times 30) = 210$ 1000$ sales.
- So this is a predicted value for which we don't know y_i.
- We need 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.

Simple Linear Regression

The Problem of Estimation (method of least squares)

- Let's do another example, recall the advertisement example from page 7.

- Suppose we want to use the TV expenditure (in 1000$) variable to predict sales (in 1000 unit). We already have seen the scatter plot, but here is it again



Simple Linear Regression

The Problem of Estimation (method of least squares)

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

R code - Regression Results for Advertisement Data - SLM

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

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

# see the results
summary(slr_result)
```

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

Coefficients:
 Estimate  Std. Error t value   Pr(>|t|)
(Intercept) 7.032594  0.457843   15.36     <0.0000000000000002 ***
TV          0.047537  0.002691   17.67     <0.0000000000000002 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099
F-statistic: 312.1 on 1 and 198 DF, p-value: < 0.00000000000000022
```

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

code: Regression Results for Advertisement Data - SLM

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

Table 2: Regression Results for Sales and TV Expenditure

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

Simple Linear Regression

The Problem of Estimation (method of least squares)

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

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

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

- or

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

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

Simple Linear Regression Model (SLR)

2. Assessing the Fit - R^2 and RSE

- 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

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

- where

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

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

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

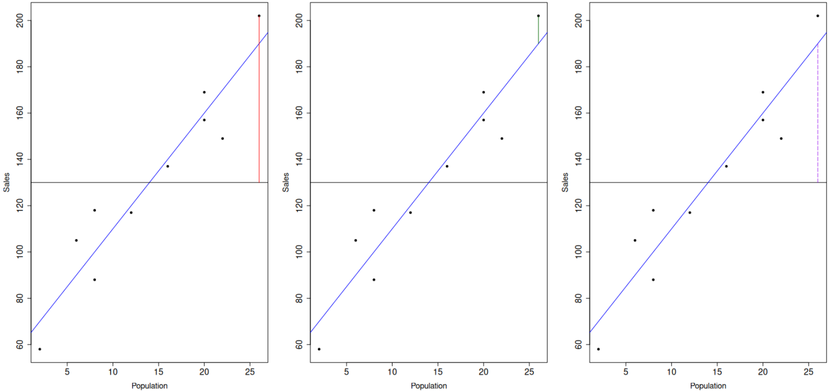
- Also recall $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Question is - what does this formula mean? To understand this first let's decompose $y_i - \bar{y}$

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

- 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 $(\widehat{y}_i - \bar{y})$

Simple Linear Regression

Goodness of Fit or R^2

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

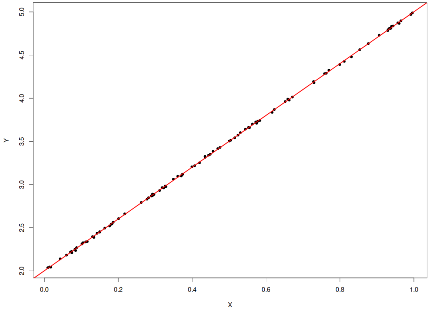
$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}_{\text{SSR}}$$

- 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

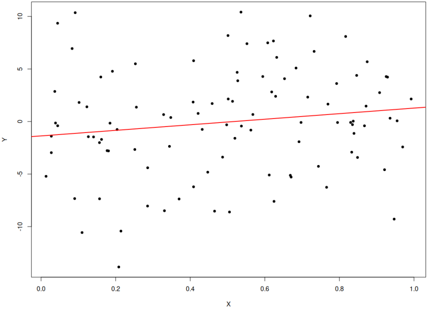


- The black dots are the sample points, the red line is the fitted line. Here R^2 is 0.99.

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

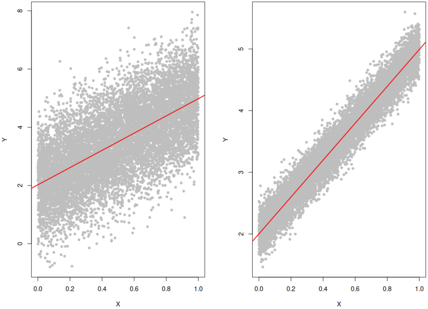
Goodness of Fit or R^2

- 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,

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

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

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

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

Simple Linear Regression

Residual Standard Error

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

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*".

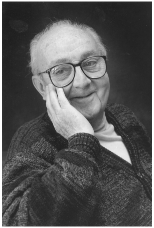


Figure 3: George Box (1919 - 2013), source - Wikipedia

- 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, $\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)\}$. 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.

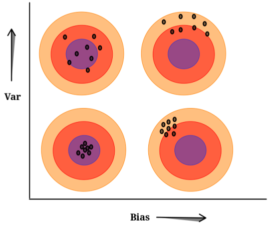


Figure 4: bias variance situations, true value β_1 is at the center, and the black dots are estimates or values of $\hat{\beta}_1$

- 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*,

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

- 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*.

- 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

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

- 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,

- 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.

Simple Linear Regression

Significance Testing - t test

- 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

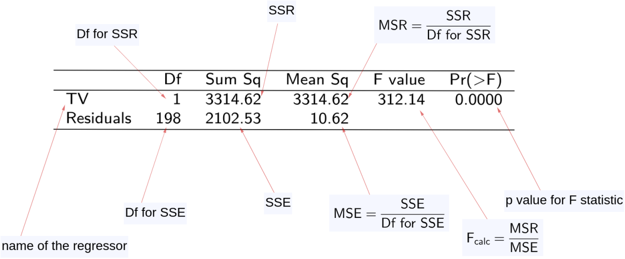
- There is another approach of doing significance testing. This approach is known as *analysis of variance* (in short ANOVA) approach. In this approach we will F-test.
- Following is the ANOVA table for the Sales Vs. TV problem

	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



	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		

SSR

Df for SSR

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

name of the regressor

Df for SSE

SSE

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

p value for F statistic

$F_{\text{calc}} = \dfrac{\text{MSR}}{\text{MSE}}$

- 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?)

Simple Linear Regression

Significance Testing - F test

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

Simple Linear Regression

Significance Testing - F 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

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 R.

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).

1. **Statistical Learning and the Problem of Regression**
    - Best Function to Predict
    - Estimation of f Parametric vs. Nonparametric

2. **Simple Linear Regression Model (SLR)**
    - 1. The Problem of Estimation
    - 2. Assessing the Fit - R^2 and RSE
    - 3. What are the Model Assumptions?
    - 4. Assessing the Accuracy of the Estimated Coefficients
    - 5. Significance Testing
    - 5. Prediction: Confidence Intervals and Prediction Intervals

3. **Appendix**

Appendix

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.

Hansen, B. (2022), *Econometrics*, Princeton University Press, Princeton.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2023), *An introduction to statistical learning*, Vol. 112, Springer.

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