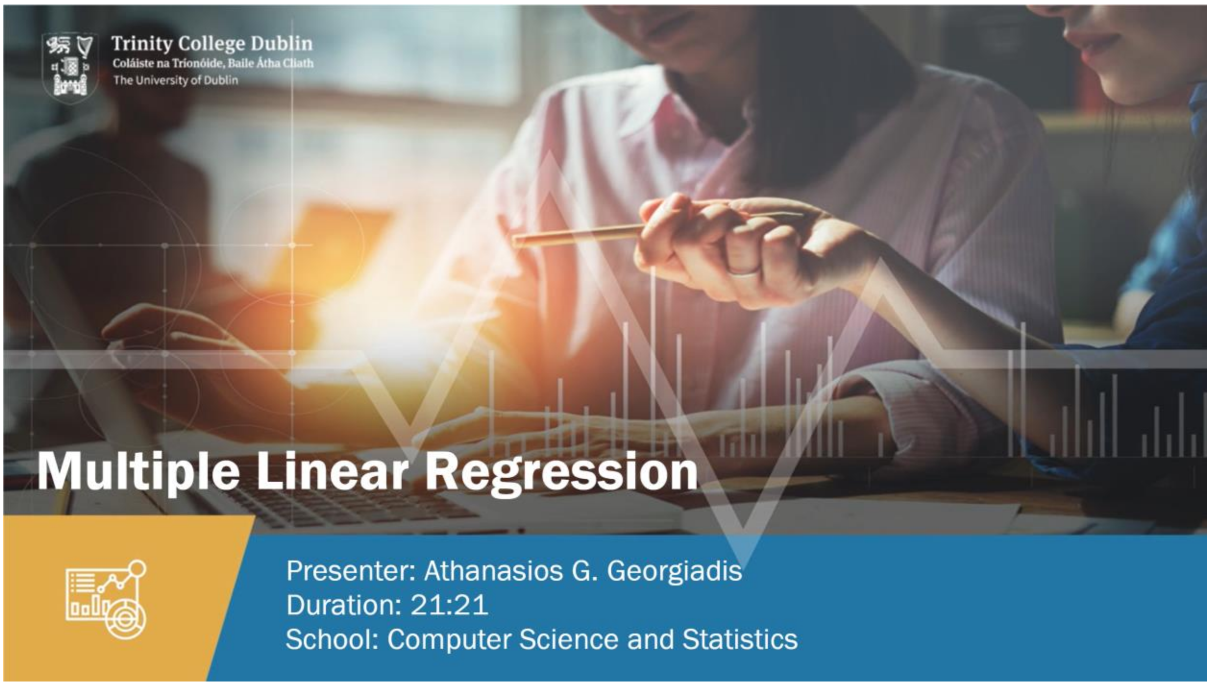
Multiple Linear Regression

Contents

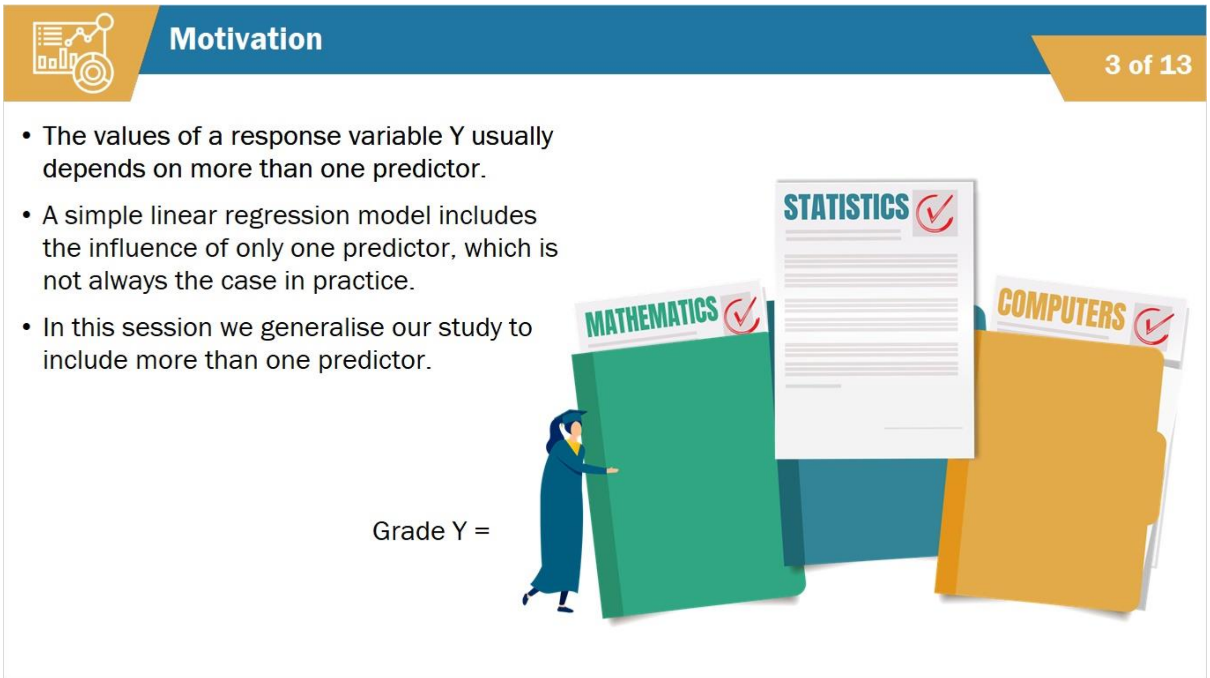
Slide 1: Introduction 2
Slide 2: Overview 2
Slide 3: Motivation 3
Slide 4: Multiple Linear Regression I 4
Slide 5: Multiple Linear Regression II 5
Slide 6: The least squares estimates. 6
Slide 7: Multiple linear regression equation. 7
Slide 8: Multiple linear regression in matrix notation I 7
Slide 9: Multiple linear regression in matrix notation II 8
Slide 10: Example: Italian restaurants in NYC 9
Tab 1: Undertake the Analysis 10
Tab 2: Read the Scatter Matrix 10
Tab 3: Obtaining the Regression Equation. 11
Tab 4: Predictions 12
Tab 5: The quality of the fit. 12
Tab 6: Discussion on the fitted equation. 13
Slide 11: Distributions of the estimators 14
Tab 1: Assumptions. 15
Tab 2: The estimator of σ^2. 15
Tab 3: The distribution of ˆβi's. 16
Tab 4: Simple linear regression model 17
Slide 12: Inference 17
Tab 1: Confidence Intervals (We have to keep it. Confidence intervals for the regression coefficients, the same in the next tab) 18
Tab 2: Hypothesis Testing 19
Tab 3: Hypothesis testing in R 19
Tab 4: Example: Italian Restaurants 20
Tab 5: Comments on the table. 21
Tab 6: Actions. 22
Slide 13: Summary 22

Slide 1: **Introduction**



Hello and welcome to this presentation. My name is Athanasios Georgiadis and I will be your instructor for this Session.

Slide 2: **Overview**



During this Session we will look at:

- Some Motivation
- Multiple Linear Regression

- A representative Example
- The Distribution of the estimators
- And Inference on the regression parameters

Slide 3: ## Motivation

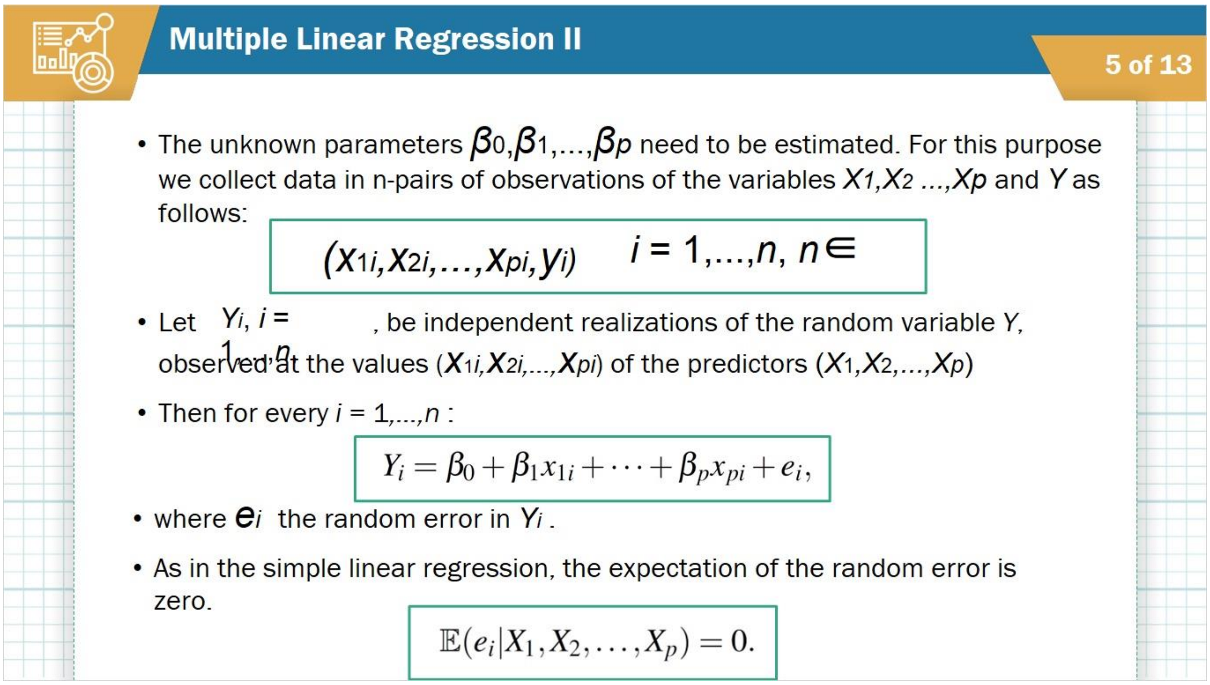
The values of a response variable Y usually depends on more than one predictor.

For example, the grade Y of a Student in Statistics, may depend on hers/his grades in a sequel of other modules; Mathematics, Computers etc.

A simple linear regression model, by definition, includes the influence of only one predictor, which is not always the case in practice.

In this Session we generalize our study so as to include more than one predictor.

Slide 4: Multiple Linear Regression I



The slides in this section will pause at the end to give you an opportunity to review the contents, when you’re ready click next to continue.

Let Y be a response variable and X1,X2,...,Xp, be p predictors.

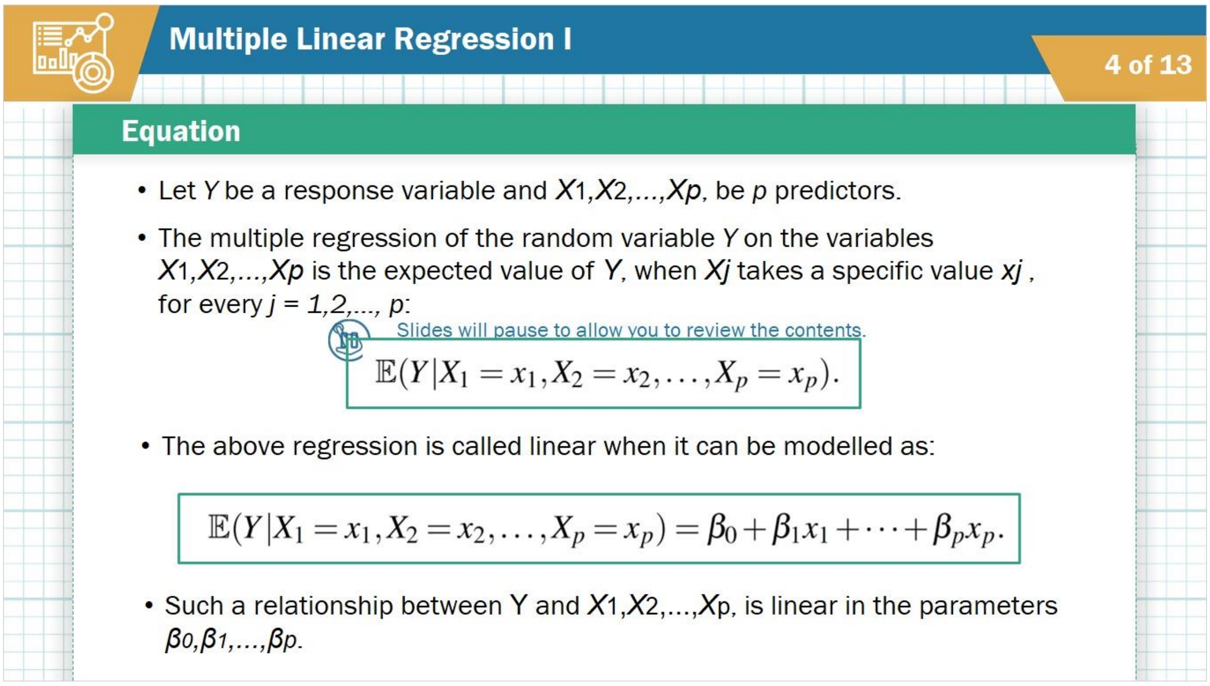
The multiple regression of the random variable Y on the variables X1,X2,...,Xp is the expected value of Y, when Xj takes a specific value xj , for every j = 1,2,..., p as in this equation.

The above regression is called linear when it can be modelled as:

E(Y|X1 = x1,X2 = x2,...,Xp = xp) = β0+β1x1+-+βpxp.

Such a relationship between Y and X1,X2,...,Xp, is linear in the parameters β0,β1,...,βp.

Slide 5: **Multiple Linear Regression II**



The unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$ need to be estimated. For this purpose we collect data in n-pairs of observations of the variables $X_1, X_2, \ldots, X_p$ and Y as follows:

$(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)$, for $i=1,2,\ldots,n$.

Let Y_i, $i = 1,\ldots,n$, be independent realizations of the random variable Y, observed at the values $(x_{1i}, x_{2i}, \ldots, x_{pi})$ of the predictors $(X_1, X_2, \ldots, X_p)$.

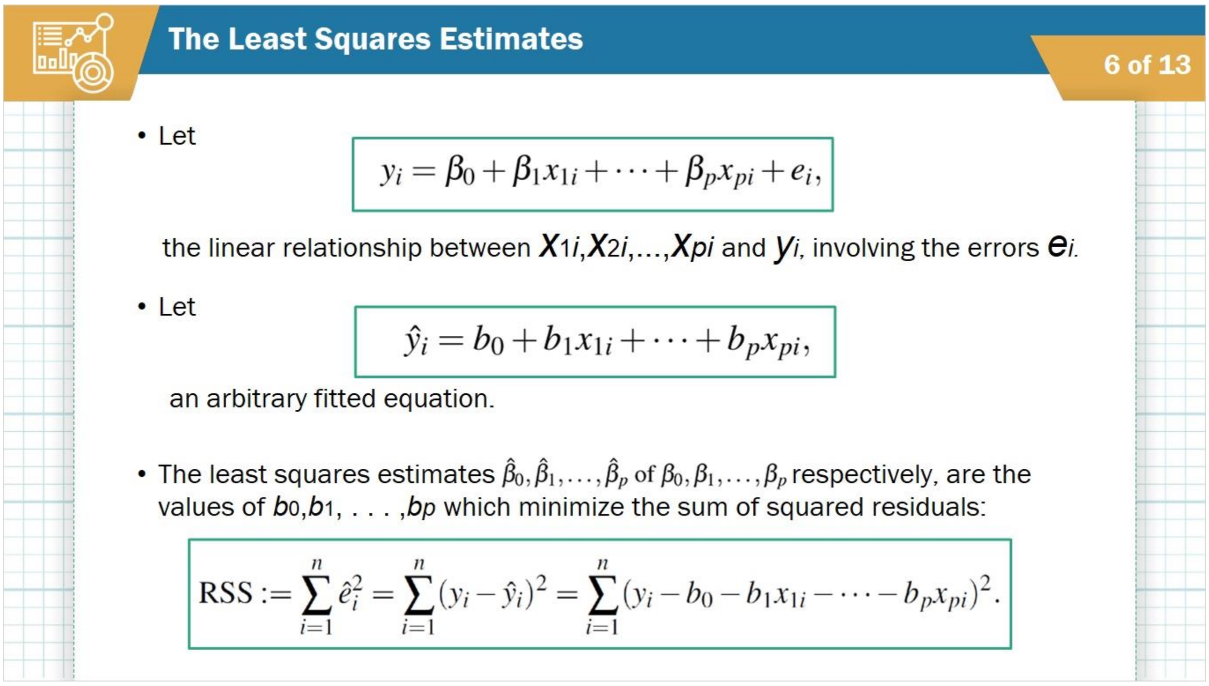
Then for every i:

$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i$,

where e_i the random error in Y_i .

As in the simple linear regression, the expectation of the random error is zero.

Slide 6: The least squares estimates.



Let

$y_i = \beta_0 + \beta_1 x_{1i} + - + \beta_p x_{pi} + e_i$,

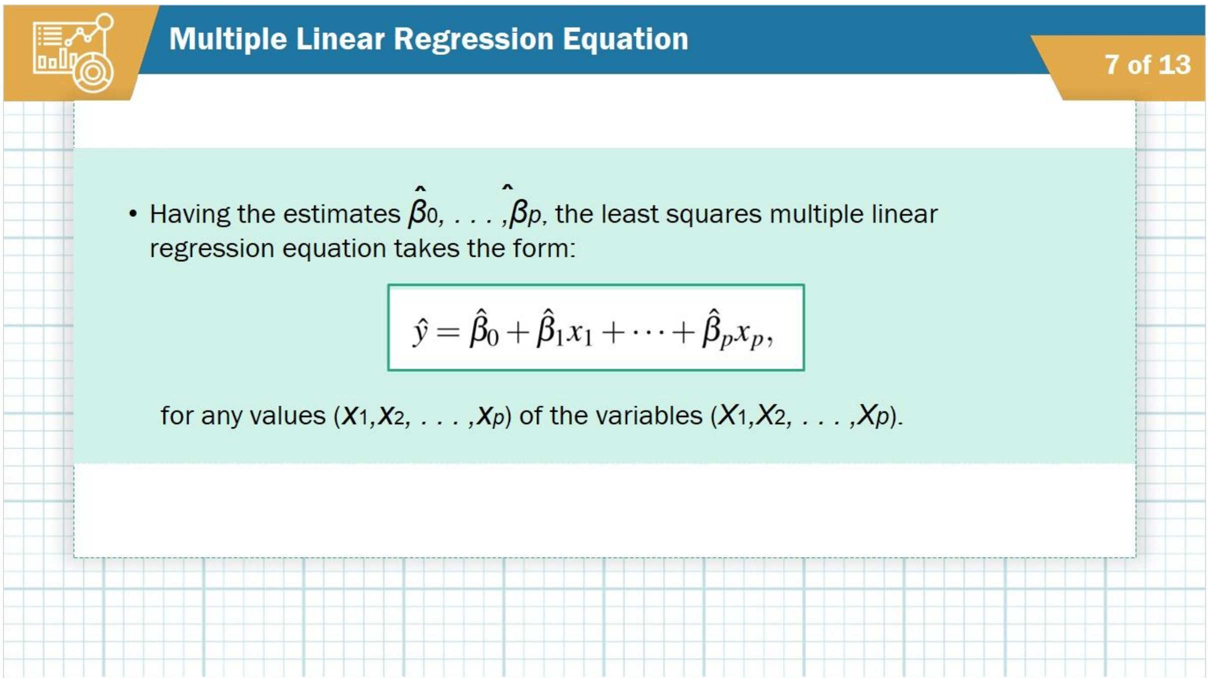
the linear relationship between x_{1i} , x_{2i} ,..., x_{pi} and y_i , involving the errors e_i .

Let $\hat{y}_i = b_0 + b_1 x_{1i} + - + b_p x_{pi}$,

an arbitrary fitted equation.

The least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$ of $\beta_0, \beta_1, \ldots, \beta_p$ respectively, are the values of $b_0, b_1, \ldots, b_p$ which minimize the sum of squared residuals, explicitly written here.

Slide 7: Multiple linear regression equation.

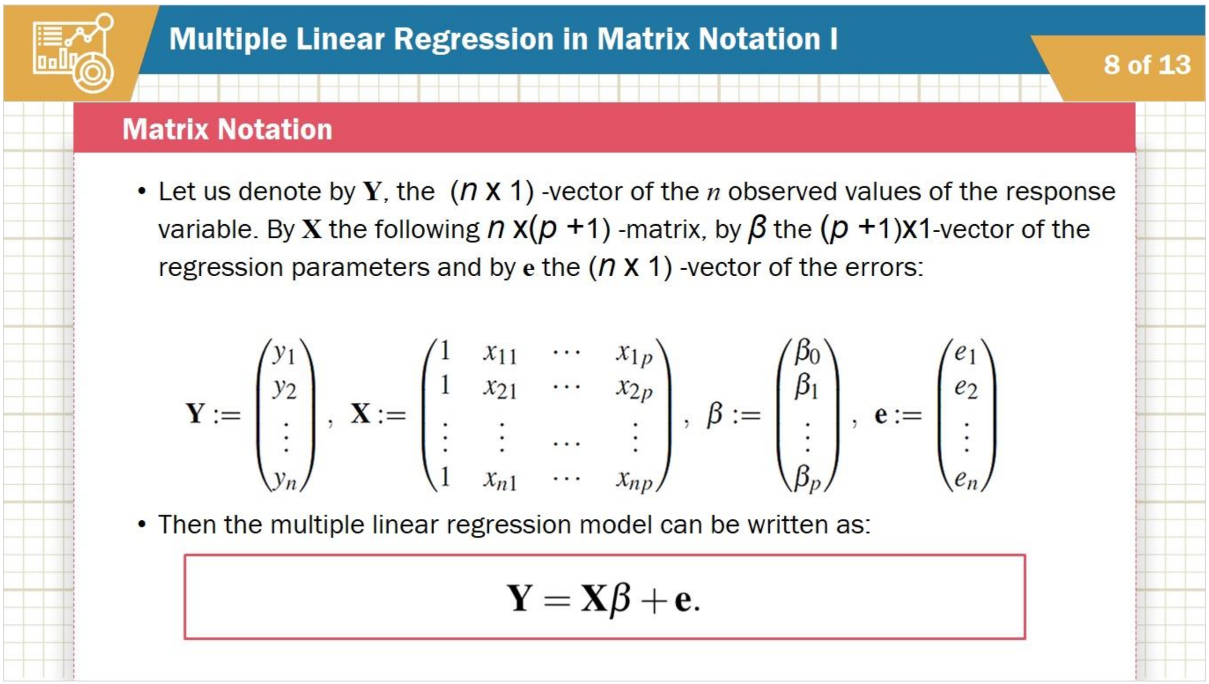


Having the estimates ˆβ0,..., ˆβp, the least squares multiple linear regression equation takes the form

yˆ = ˆβ0 + ˆβ1x1 +-+ ˆβpxp,

for any values (x1, x2,..., xp) of the variables (X1,X2,...,Xp).

Slide 8: Multiple linear regression in matrix notation I

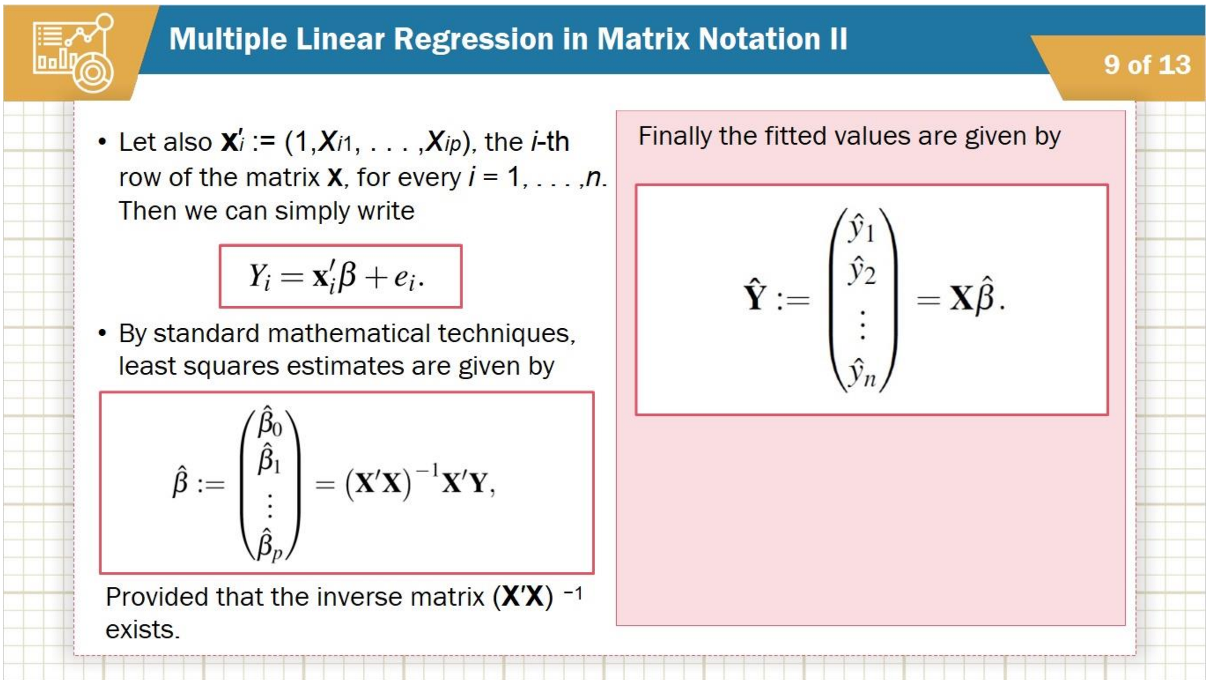


A convenient way to study multiple linear regression, is by using matrix-vector notation.

Let us denote by **Y** , the (n×1)-vector of the n observed values of the response variable, y1, y2,..., yn. By **X** the following n × (p + 1)-matrix, by β the (p + 1) × 1-vector of the regression parameters β0, β1,...,βp and by **e** the (n × 1)- vector of the errors e1, e2,..., en.

Then the multiple linear regression model can be written as: **Y** = **X**β +**e.**

Slide 9: Multiple linear regression in matrix notation II



Let also **x** ′ i := (1, xi1,..., xip), the i-th row of the matrix **X**, for every i = 1,...,n. Then we can simply write

Yi = **x** ′ iβ +ei.

By standard mathematical techniques, it turns out that the least squares estimates are given by the vector β ̂, provided that the inverse matrix (**X** ′**X**) ^{−1} exists.

Finally, the fitted values are given by this vector equation.

Slide 10: Example: Italian restaurants in NYC

This is a very representative example in multiple linear regression.

The Zagat guide contains restaurant ratings and reviews for many major world cities. Let us focus on Italian restaurants in NYC. Specifically, we want to know how customer ratings (measured on a scale of 0 to 30) of the "Food", "Decor", and "Service", as well as whether the restaurant is located to the east or west of 5th Avenue, affect the average "Price" of a meal. The dataset contains ratings and prices for 168 Italian restaurants in 2001 and is in your disposal (file rests.csv).

Click each tab to go through the steps of obtaining the equation, when you are ready, click next to continue.

Tab 1: **Undertake the Analysis**

Example: Italian Restaurants in NYC

Undertake the Analysis | Read the Scatter Matrix | Obtaining the Regression Equation | Predictions | The Quality of the Fit | Discussion on the Fitted Equation

Undertake the Analysis

- Let us undertake the multiple regression analysis! We would enter the dataset in R and use the plot-command as below.

```
rests=read.csv(file="C:\\Users\\nasge\\Desktop\\External\\1driveold\\R+Latex\\
      lectdata\\rests.csv")
plot(rests[, c("Price", "Food", "Decor", "Service", "East")])
```

- Then R returns the following plot we refer to as the scatter-matrix.

Let us undertake the multiple regression analysis! We would enter the dataset in R and use the plot-command as below.

Then R returns the following plot we refer to as the scatter-matrix.

Tab 2: **Read the Scatter Matrix**

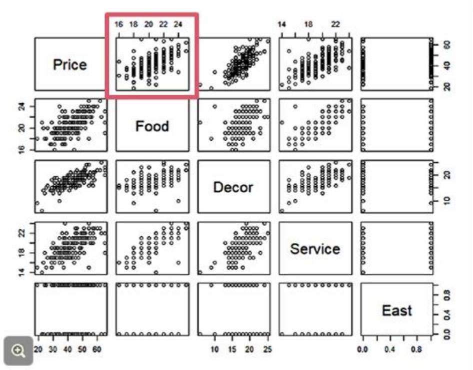
Example: Italian Restaurants in NYC

10 of 13

Undertake the Analysis | Read the Scatter Matrix | Obtaining the Regression Equation | Predictions | The Quality of the Fit | Discussion on the Fitted Equation

Read the Scatter Matrix

- The scatter-matrix contains the following:
- The response variable Y:=Price and the four predictors X1 :=Food, X2 :=Decor, X3 :=Service and X4 :=East.
- Each graph contains the scatterplot when we restrict to two of the variables of the problem.
- All the graphs in the first row can give us the influence of each predictor on the response variable.
- The predictors “Food”, “Decor" and “Service" range between 0 and 30. The predictor “East" is a categorical variable. It takes the values 1 when the restaurant is located at the East and 0 when it is not.



Click the image to enlarge.

Let us now “Read" the scatter-matrix.

The scatter-matrix contains the response variable Y:=Price and the four predictors X1 :=Food, X2 :=Decor, X3 :=Service and X4 :=East.

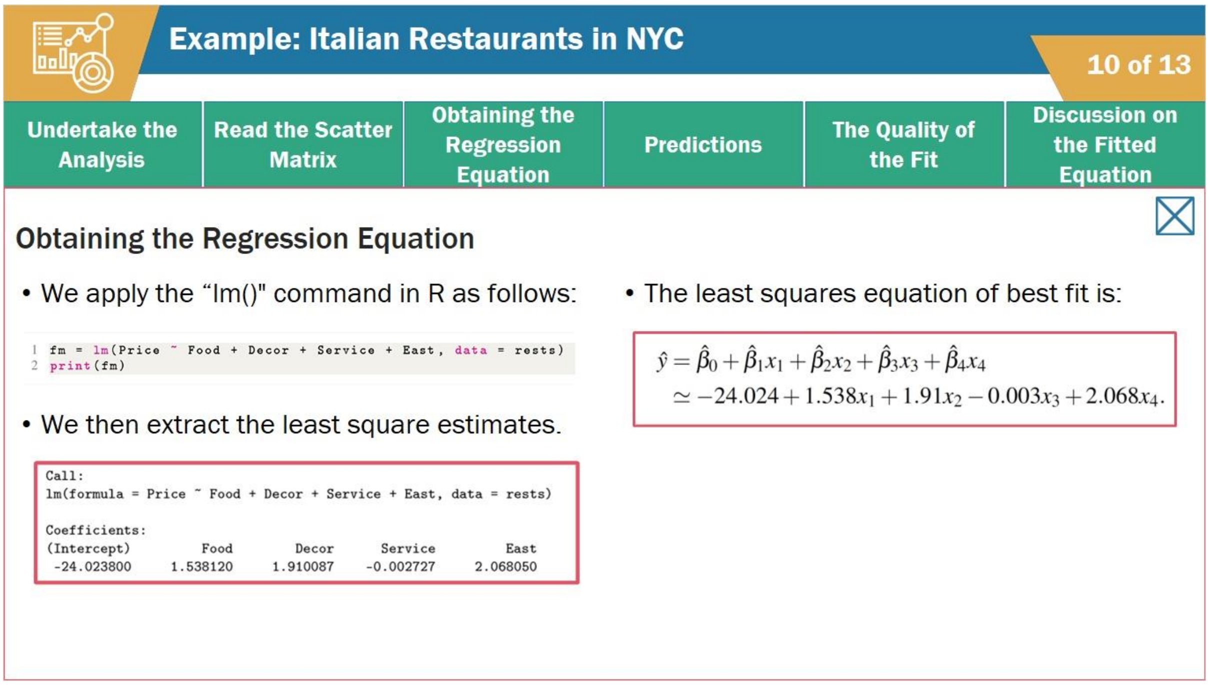
Each graph contains the scatterplot when we restrict to two of the variables of the problem.

For example, the very first graph, contains the points of the form (Food, Price). That is the x-axis contains the scores about the quality of Food (0-30) and the y-axis the Price in Dollars.

All the graphs in the first row can give us the influence of each predictor (one per time), on the response variable.

The predictors “Food", “Decor" and “Service" range between 0 and 30. The predictor “East" is a categorical variable. It takes the values 1 when the restaurant is located at the East and 0 when it is not.

Tab 3: Obtaining the Regression Equation.



We apply the “lm()" command in R as in the grey frame below.

We then extract the least square estimates.

The least squares equation of best fit is:

yˆ = ˆβ0 + ˆβ1x1 + ˆβ2x2 + ˆβ3x3 + ˆβ4x4 which equals with

≃ −24.024+1.538x1 +1.91x2 −0.003x3 +2.068x4.

Tab 4: **Predictions.**

Predictions

- R language allows to predict the average Price (Y) for given values of all the predictors.
- For example, for a restaurant scoring 20, 15 and 20, in "Food", "Decor" and "Service" respectively and is located at the "East", after using the standard commands in R:

```
a <- data.frame(Food=20, Decor=15, Service=20, East=1)
result <-  predict(fm,a)
print(result)
```

```
       1
37.40341
```

- We obtain that the expected price is around $\hat{y} \simeq 37.4$ Dollars.

R language allows to predict the average Price (Y) for given values of all the predictors. For example, for a restaurant scoring 20, 15 and 20, in "Food", "Decor" and "Service" respectively and is located at the "East", after using the standard commands in R, as below,

we obtain that the expected price is around 37.4Dollars.

Tab 5: **The quality of the fit.**

Example: Italian Restaurants in NYC

10 of 13

Undertake the Analysis | Read the Scatter Matrix | Obtaining the Regression Equation | Predictions | The Quality of the Fit | Discussion on the Fitted Equation

The Quality of the Fit

- Using R and the standard command:

```
summary(fm)$r.squared
```

we derive that $R^2 \simeq 63\%$, which is good enough for a fit.

Using R and the standard command presented here, we derive that R^2 is around 63%, which is good enough for a fit.

Tab 6: **Discussion on the fitted equation.**

We have obtained this equation).

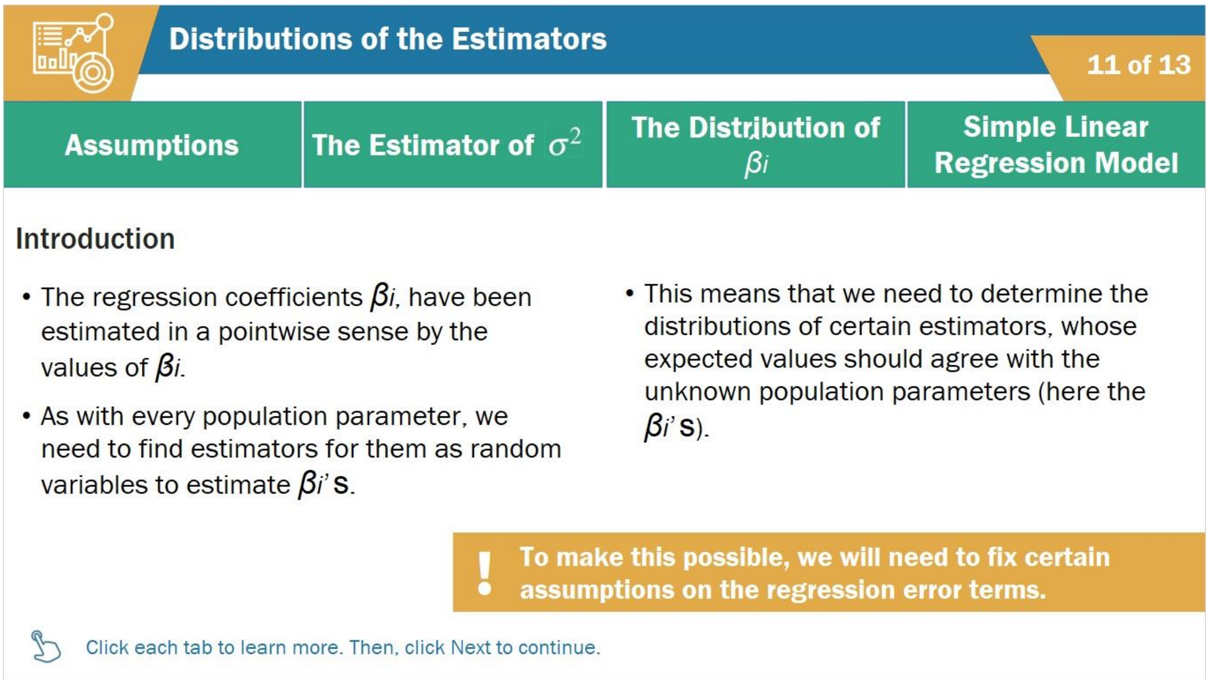
The coefficient of each predictor represents the expected influence of the predictor on the Price. Here we can see that “Decor" turns to have largest average influence than “Food".

The dummy variable “East", has a positive (and relatively large) coefficient. A restaurant on the East of the 5th avenue, is expected to be more expensive than one on the West, even if the restaurants are of the same standards in terms of “Food", “Decor" and “Service".

The coefficient ˆβ3, is really small. This means that the quality of “Service" has a low effect on the price. We will return on it in the next sections.

Click next to continue.

Slide 11: Distributions of the estimators



The regression coefficients βi , have been estimated in a pointwise sense by the values of ˆβi .

As with every population parameter, we need to find estimators for them as random variables.

This means that we need to determine the distributions of certain estimators, whose expected values should agree with the unknown population parameters (here the βi's).

To make this possible, we will need to fix certain assumptions on the regression error terms.

Click each tab to learn more, when you are ready, click next to continue.

Tab 1: **Assumptions.**

Distributions of the Estimators

Assumptions	The Estimator of σ^2	The Distribution of βi	Simple Linear Regression Model

Assumptions

- Consider the multiple linear regression model

$$\boxed{\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}.}$$

- We assume that the errors e_i :
  - Are independent of each other
  - Are normally distributed
  - Have zero mean
  - Have a common (but unknown) variance σ^2

Consider the multiple linear regression model **Y** = Xβ +**e.**

We assume that the errors ei :

are independent of each other,

are normally distributed,

have zero mean,

have a common (but unknown) variance σ^2.

Tab 2: **The estimator of σ^2.**

Distributions of the Estimators

Assumptions	The Estimator of σ^2	The Distribution of βi	Simple Linear Regression Model

The Estimator of σ^2

- The common variance σ^2 is of fundamental importance for the regression analysis.
- We need to define an estimator for this. As it can be proved

$$\boxed{S^2 := \frac{\mathrm{RSS}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n} \hat{e}_i^2,}$$

is an unbiased estimator of σ^2.

- Let us remark here that the denominator $n - p - 1 = n - (p+1)$, represents the estimation of the $(p+1)$–parameters $\beta_0, \beta_1, \ldots, \beta_p$, by the dataset of size n.
- R language computes for us $S = \sqrt{S}$, "lm()" command.

The common variance σ^2 is of fundamental importance for the regression analysis.

We need to define an estimator for this.

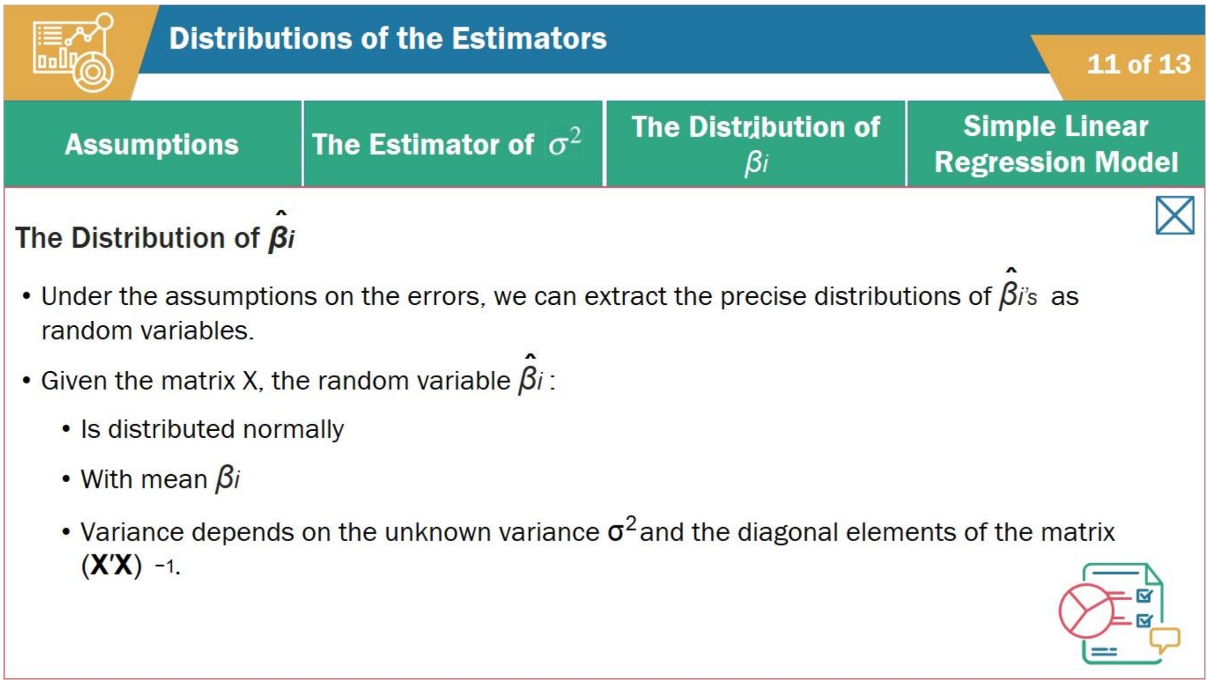
As it can be proved

$S^2 := RSS/(n-p-1)$ is an unbiased estimator of σ^2 .

Let us remark here that the denominator $n-p-1 = n-(p+1)$, represents the estimation of the $(p+1)$-parameters $\beta_0, \beta_1, \ldots, \beta_p$, by the dataset of size n.

R language computes for us S, the square root of S^2 in the "lm()" command. We will come back on it later.

Tab 3: **The distribution of $\hat{\beta}_i$'s.**



Under the assumptions on the errors, we can extract the precise distributions of $\hat{\beta}_i$'s, as random variables.

Given the matrix **X**, the random variable $\hat{\beta}_i$:

is distributed normally,

with mean β_i ,

and its variance depends on the unknown variance σ^2 and the diagonal elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$.

Tab 4: **Simple linear regression model**

Distributions of the Estimators

Assumptions	The Estimator of σ^2	The Distribution of βi	Simple Linear Regression Model

Simple Linear Regression Model

- Let us see the precise values on the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i,$$

where $p = 1$ predictors are included.

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{SXX} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

and

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SXX}, \quad \text{where} \quad SXX = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Let us see the precise values on the simple linear regression model

Yi = β0 +β1xi +ei ,

where p = 1 predictors are included. Then the variance of ˆβi are given here.

Slide 12: **Inference**

Inference

Confidence Intervals	Hypothesis Testing	Hypothesis Testing in R	Example: Italian Restaurants	Comments on the Table	Actions

Introduction

- The key for doing inference about a population parameter is the definition of a random variable which involves the parameter under study and its distribution is known.
- Here, for every $i = 0,1, \ldots, p$, the random variable

$$T_i := \frac{\hat{\beta}_i - \beta_i}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}.$$

- Note that above the standard error $\mathrm{se}(\hat{\beta}i)$ is the estimator of the standard deviation of $\hat{\beta}i$, when σ is estimated by S^2 .
- The degrees of freedom in the above Student's distribution are obtained by:
- Datasize−number of estimated parameters = n−(p+1)

Click each tab to learn more. Then, click Next to continue.

In this section, we shall develop methods for determining confidence intervals and for performing hypothesis tests for the regression parameters.

The key for doing inference about a population parameter is the definition of a random variable which involves the parameter under study and its distribution is known.

Here, for every i = 0,1,..., p, the random variable Ti := (ˆβi −βi)/se(ˆβi) is distributed according to the Student's distribution.

Note that above the standard error se(ˆβi) is the estimator of the standard deviation of ˆβi , when σ^2 is estimated by S^2 .

The degrees of freedom in the above Student's distribution are obtained by:

Datasize−number of estimated parameters = n−(p+1).

Click each tab to learn more, when you are ready, click next to continue.

Tab 1: Confidence Intervals (We have to keep it. Confidence intervals for the regression coefficients, the same in the next tab)

Inference

12 of 13

Confidence Intervals	Hypothesis Testing	Hypothesis Testing in R	Example: Italian Restaurants	Comments on the Table	Actions

Confidence Intervals

- We run the following command in R to find the 95% confidence intervals for the regression parameters:

```
confint(fm)
```

we then extract

```
> confint(fm)
                   2.5 %      97.5 %
(Intercept) -33.3210407 -14.7265586
Food          0.8095797   2.2666601
Decor         1.4815846   2.3385897
Service      -0.7851371   0.7796821
East          0.1985964   3.9375039
```

The 95% confidence interval for the parameter $\beta 1$ is (0.81,2.27) and for the parameter $\beta 2$ is (1.48,2.34).

Let us return to the Example of Italian restaurants.

We run the following command in R to find the 95% confidence intervals for the regression parameters.

We then extract this table.

For example, the 95% confidence interval for the parameter β1 is (0.81,2.27) and for the parameter β2 is (1.48,2.34).

Tab 2: **Hypothesis Testing**

Hypothesis Testing

- The next target is to perform hypothesis tests for the parameters βi.
- Let i = 0,1, . . . , p and βi^* a specific real number.
- The hypothesis tests with null hypothesis

$$\boxed{H_0 : \beta_i = \beta_i^*,}$$

can be performed just because

$$\boxed{T_i^* := \frac{\hat{\beta}_i - \beta_i^*}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}, \quad \text{under } H_0.}$$

The next target is to perform hypothesis tests for the parameters βi .

Let i = 0,1,..., p and β^{∗}_ i a specific real number.

The hypothesis tests with null hypothesis shown here

can be performed just because

T^{∗}_ i := (ˆβi −β^{∗}_i)/se(ˆβi) ∼ tn−p−1, under H0.

Tab 3: **Hypothesis testing in R**

Hypothesis Testing in R

- R performs for us directly in the “lm()" command, the following tests:

$$\boxed{H_0^i : \beta_i = 0 \text{ VS } H_1^i : \beta_i \neq 0, \quad \text{for every } i = 0, 1, \ldots, p.}$$

- Precisely, R gives columns with the observed values of the test statistics Ti^* and the p-values of the above tests.
- When we reject $H0^i$, we can accept that the corresponding regression coefficient $\beta i \neq$ 0. In this case we say that the intercept (when i=0) or the predictor Xi (when i = 1,2, . . . , p), is significant for the regression model and we have to include it in our predictions!
- For simplicity R evaluates the significances with a number of ∗s, according to the interval where the p-value lies in. For two or three ∗s, we have a clear significance with p-value smaller than 0.01. One ∗ corresponds to a p-value ∈(0.01,0.05), which is still significant.

R performs for us directly in the “lm()" command, the following tests:

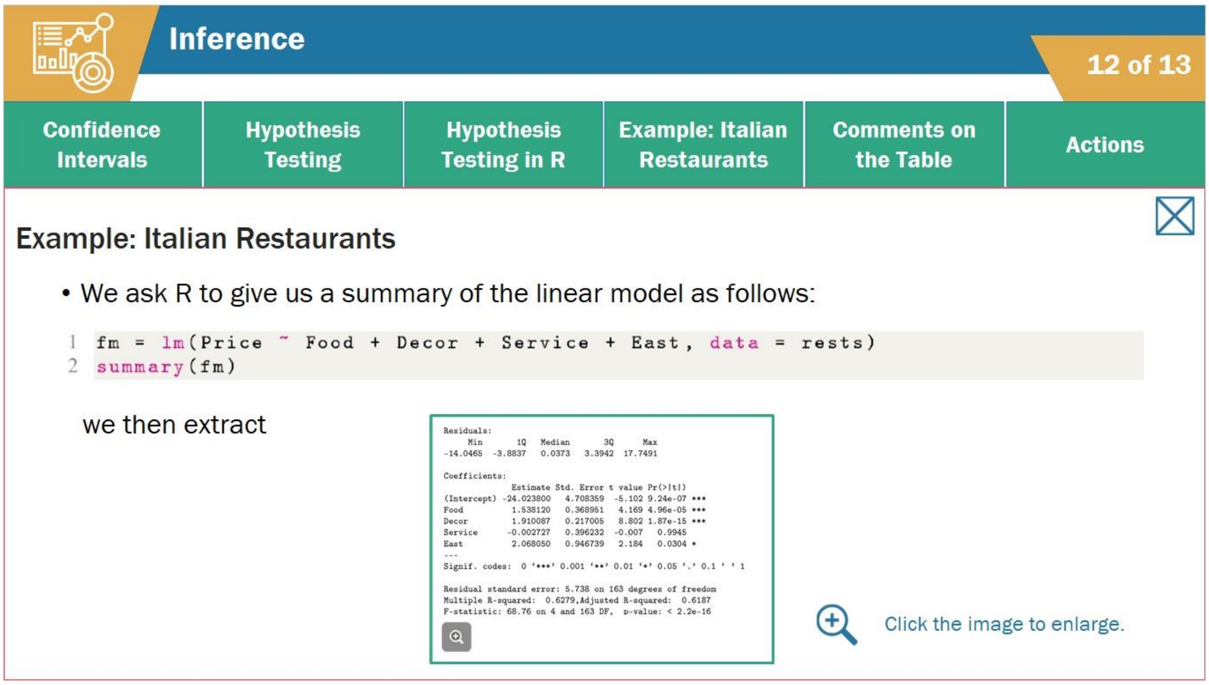
$H_{i0} : \beta_i = 0$ VS $H_{i1} : \beta_i \neq 0$, for every $i = 0,1,\dots, p$.

Precisely, R gives columns with the observed values of the test statistics T ^{∗}_ i and the p-values of the above tests.

When we reject H_{i0} , we can accept that the corresponding regression coefficient $\beta_i \neq 0$. In this case we say that the intercept (when $i = 0$), or the predictor X_i (when $i = 1,2,\dots, p$), is *significant* for the regression model and we have to include it in our predictions!

For simplicity R evaluates the significances with a number of stars, according to the interval where the p-value lies in. For two or three stars, we have a clear significance with p-value smaller than 0.01. One star corresponds to a p-value that belongs to (0.01,0.05), which is still significant, but less than when we have two or three stars.

Tab 4: Example: Italian Restaurants

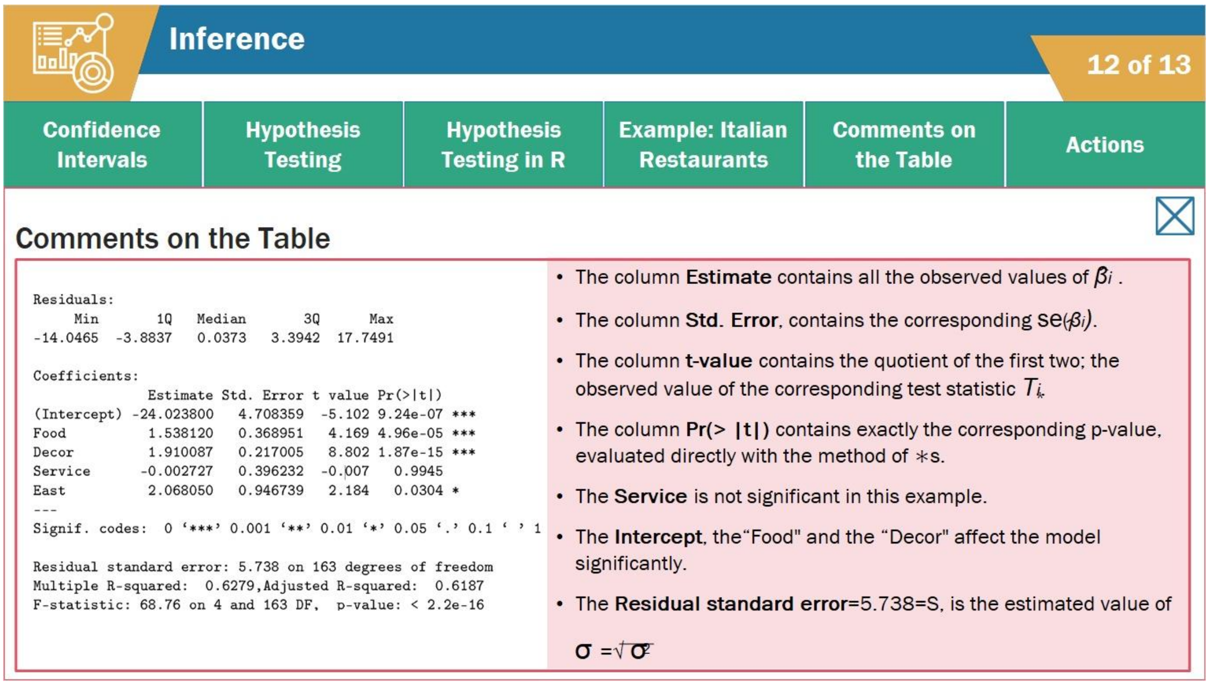


Let us now see everything in the Example of Italian restaurants.

We ask R to give us a summary of the linear model with the command in the grey frame.

We then extract this large table.

Tab 5: Comments on the table.



The column "Estimate" contains all the observed values of $\hat{\beta}_i$. Exactly these are going to be used in the model. The column "Std. Error", contains the corresponding standard errors.

The column "t-value" contains the quotient of the first two; the observed value of the corresponding test statistic T^{*}_{i} .

The column "Pr(> |t|)" contains exactly the corresponding p-value, evaluated directly with the method of stars.

In the present example: The "Service" is not significant at all. We will finally exclude it from the study. The intercept, the "Food" and the "Decor" affect the model significantly.

The "Residual standard error"=5.738=S, is the estimated value of σ the square root of σ^2.

Tab 6: **Actions.**

Inference

Confidence Intervals	Hypothesis Testing	Hypothesis Testing in R	Example: Italian Restaurants	Comments on the Table	Actions

Actions

- We may exclude the “Service" of the study and refit, obtaining the new model:

```
nm = lm(Price ~ Food + Decor + East, data = rests)
summary(nm)
```

we then extract

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
```

- Here there are not important changes. The fitted model is:

$$\boxed{\text{Price} \simeq -24 + 1.5\text{Food} + 1.9\text{Decor} + 2.1\text{East}.}$$

We may exclude the “Service" of the study and refit, obtaining the new model.

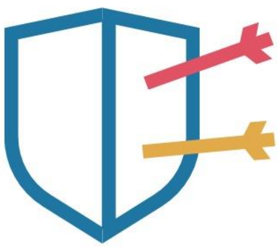
We then extract this table.

Here there are not important changes. The new fitted model is as shown here.

Slide 13: **Summary**

Summary

- Having completed this presentation, you should now be able to:
  - Define what a multiple linear regression model is
  - Fit a multiple linear regression model
  - Perform inferential controls on the regression parameters
  - Use R to obtain a model of multiple linear regression and predict values based on it.



By completing this Session, you should be able to:

- Define what a multiple linear regression model is.

- Fit a multiple linear regression model.
- Perform inferential controls on the regression parameters.
- Use R to obtain a model of multiple linear regression and predict values based on it.