Lecture 01: Performing Data Analysis using R-Language for Linear Regression

Data Analysis 2 [125J5041]

Instructor: Dr. Abar

Spring Semester 2025

Recap DA1 | Lecture 01: Linear Regression

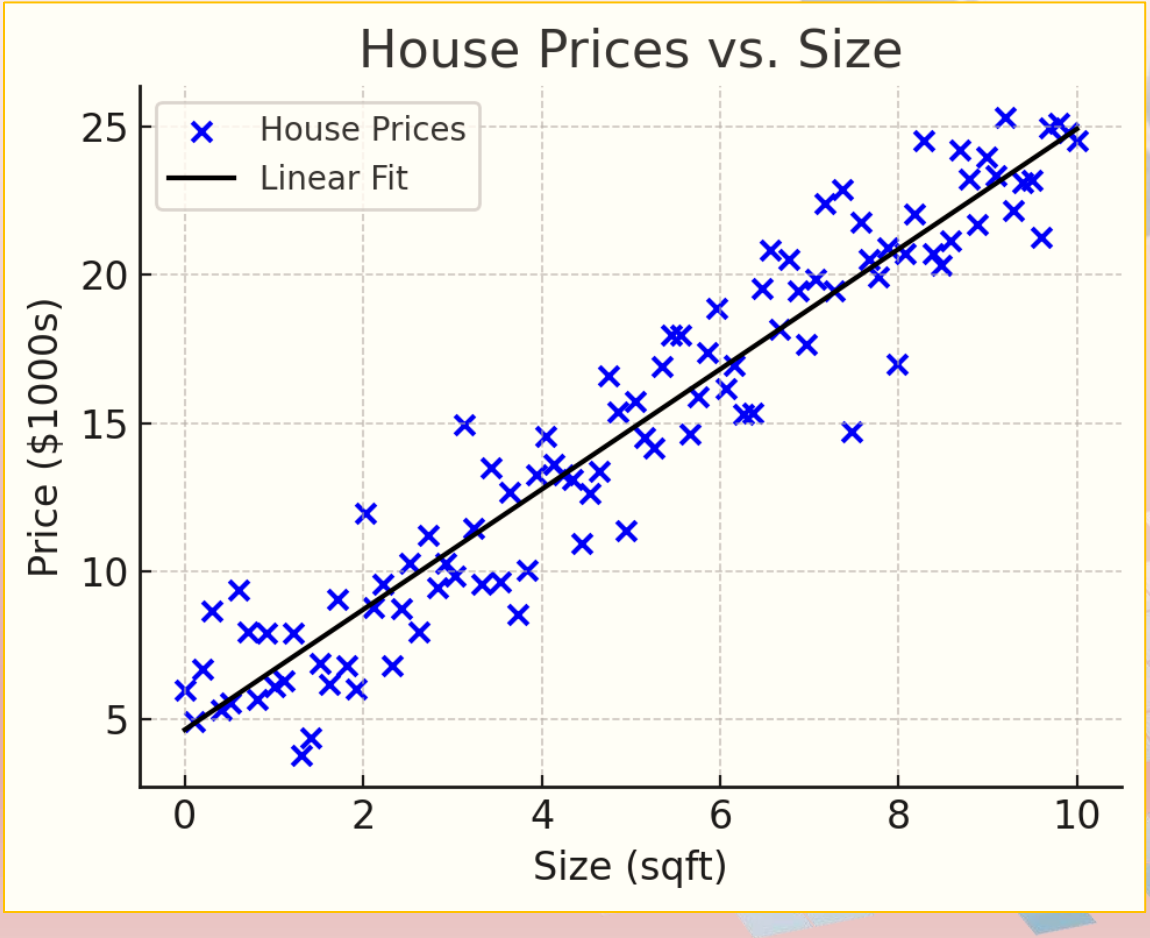
- **Regression analysis** is a method for modeling relationships between variables.
- A variable we want to infer or predict is called the dependent variable or outcome; and variables we use for prediction are called independent or explanatory variables or predictors.
- **Linear regression** models the relationship between the independent variables and a dependent variable as a straight/linear line; e.g. since a person's height increases as age increases, age (independent variable) and height (dependent variable) have a linear relationship.
- It is used when the dependent variable is continuous and can take any value within a range. For instance, estimating house prices, measuring temperature, or predicting sales revenue, billing electricity consumption, etc.
- To understand the performance of Linear Regression model, carrying out model evaluation is necessary. Error (i.e. deviation in predicted findings) is typically measured using evaluation metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

Recap DA1 | Lecture 01: Linear Regression

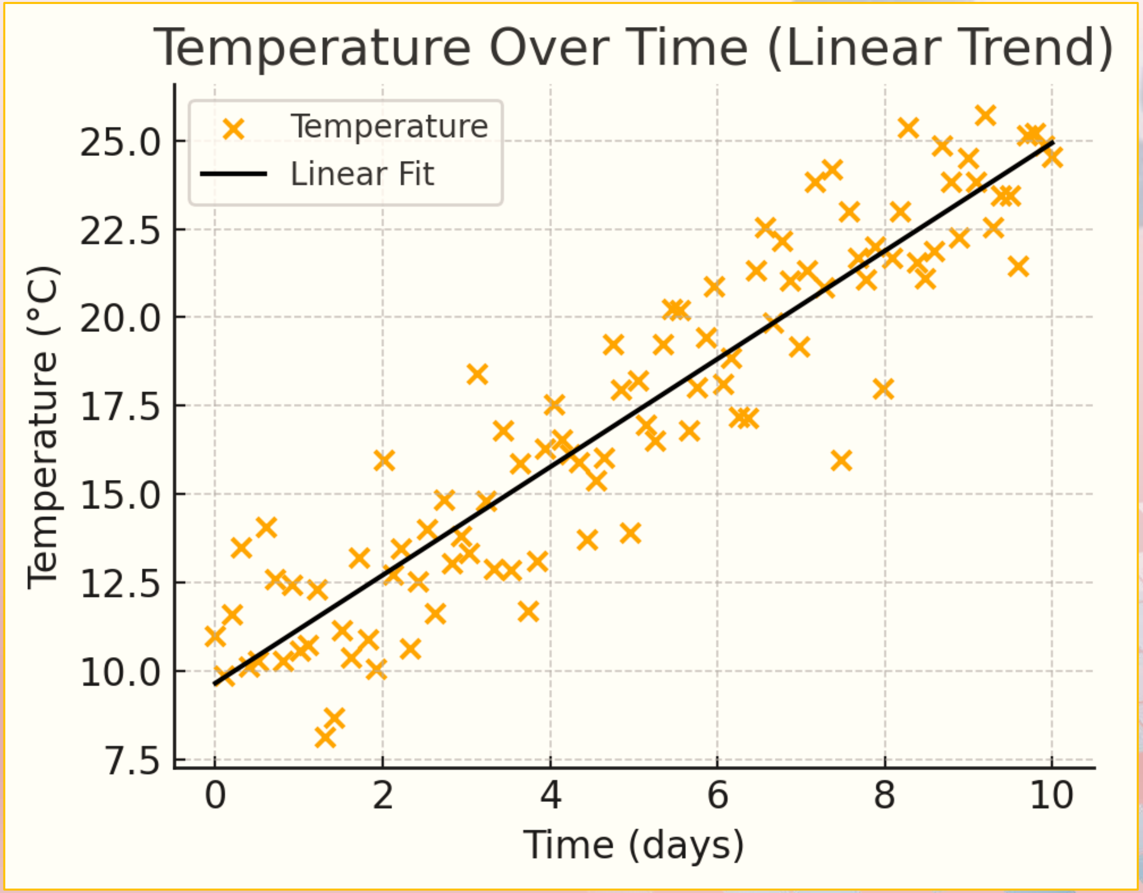
Why Linear regression for Machine Learning? | 4 Minutes Video

Recap DA1 | Lecture 01: Linear Regression

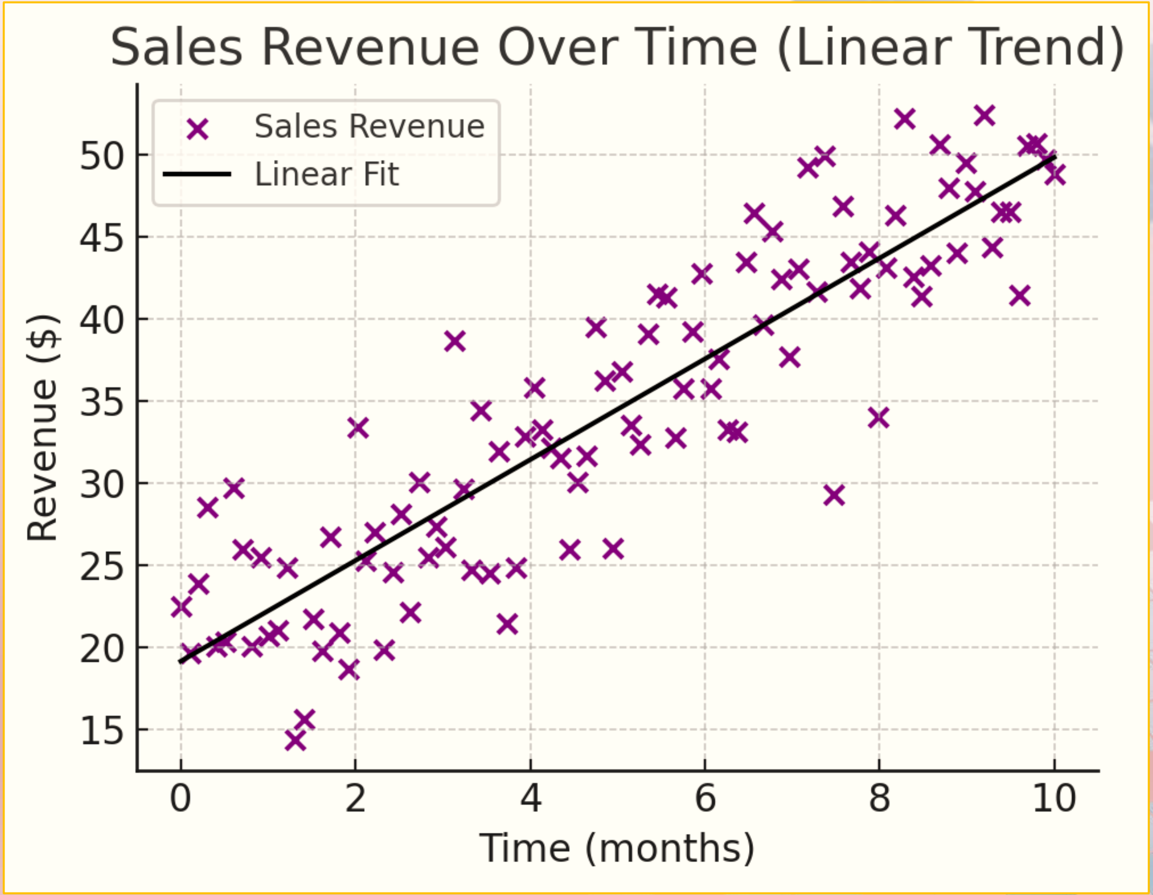
kcg.edu



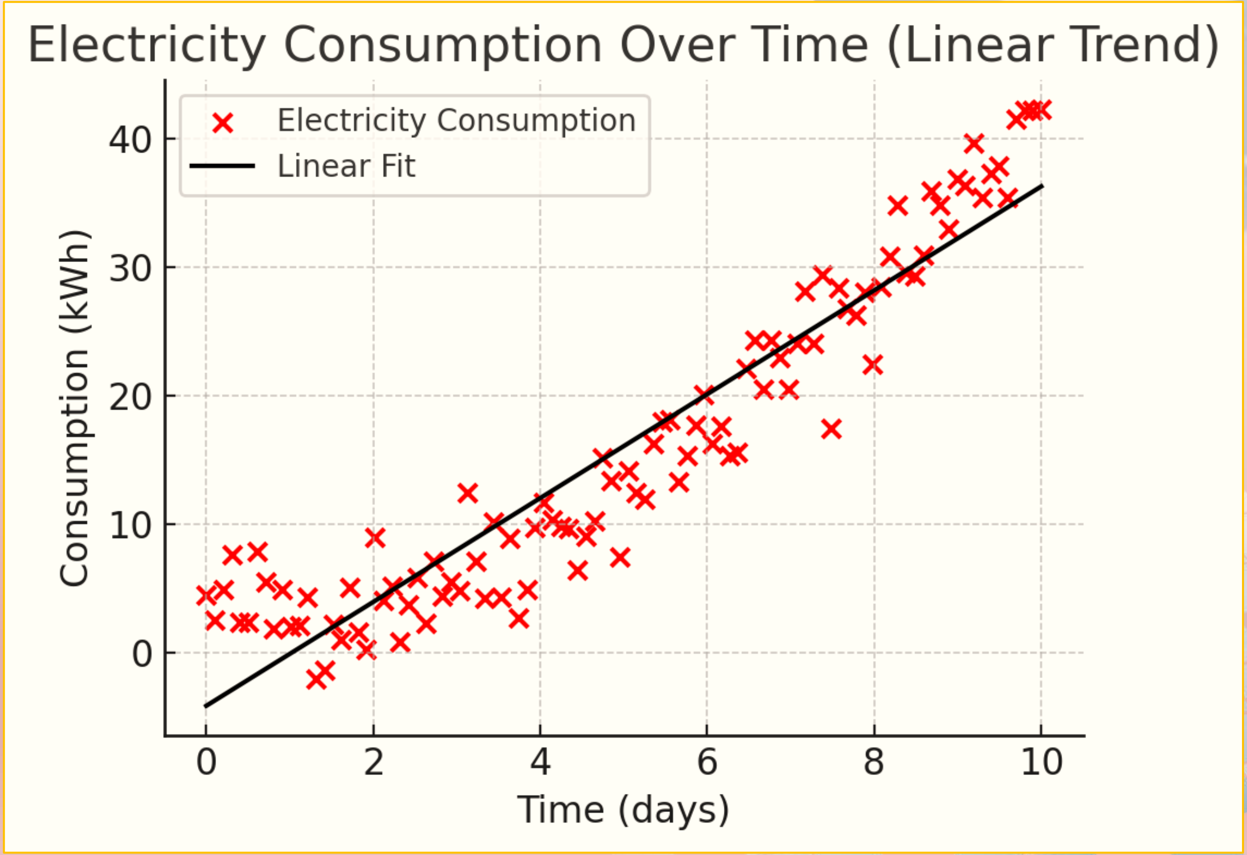
Recap DA1 | Lecture 01: Linear Regression



Recap DA1 | Lecture 01: Linear Regression



Recap DA1 | Lecture 01: Linear Regression



Exercise on Linear Regression [1/8]

kcg.edu

Perform linear regression analysis in RStudio using a sample dataset

- **Objective:** Perform a linear regression analysis to predict the sales of a product based on its advertising expenses using a sample dataset.

- **Dataset:** You will use the "Advertising" dataset, which contains three columns: "TV," "Radio," and "Newspaper" advertising expenses in thousands of dollars, and "Sales" in thousands of units.

- **Instructions:**

1. Open RStudio

2. Click File → New Project → New Directory → New Project → Directory name: → Create Project

3. Create a new R Notebook (from File menu + select R Notebook) and form a new chunk.

4. Copy file Advertising.csv in your Directory or Folder created in Step: 1.

5. Load dataset into your R session using following command:
advertising_data <- read.csv("Advertising.csv") (Slide: 11)

6. Explore dataset using functions like head(advertising_data) for viewing first few rows of your dataset, summary(advertising_data) for viewing statistical summaries of your dataset, and str(advertising_data) to understand its structure and contents.

Exercise on Linear Regression [2/8]

kcg.edu

Perform linear regression analysis in RStudio using a sample dataset

7. Click on Preview, R Notebook in HTML format will appear inside RStudio. (Slide: 12)

8. Perform a simple linear regression analysis to predict "Sales" based on "TV" advertising expenses. For this, use lm() function to create linear regression model:

```
model <- lm(Sales ~ TV, data = advertising_data)
```

9. Print the summary of regression model using summary(model). (Click on Preview again and refer to Slide: 14).

10. Interpret results from the regression summary, paying attention to coefficients, p-values, and R-squared. (see Slide: 17 Appendix)

11. Create a scatter plot to visualize the relationship between "TV" advertising expenses and "Sales":

```
plot(advertising_data$TV, advertising_data$Sales, xlab = "TV Advertising Expenses", ylab = "Sales", main = "Scatter Plot of TV vs. Sales")
abline(model, col = "red")
```

[R function abline() is used to draw regression lines to a graph]

(Click on Preview again and see Slide: 15)

12. Calculate predicted values using your regression model:

```
predicted_sales <- predict(model, newdata = data.frame(TV = advertising_data$TV))
```

[This line of code calculates predicted sales for each value of "TV" advertising expenses in your original dataset and stores the results in "predicted_sales" object]

```
predicted_sales
```

13. Calculate residual values (differences between actual and predicted values):

```
residuals <- advertising_data$Sales - predicted_sales
residuals
```

14. Calculate mean squared error (MSE) to measure model's accuracy:

```
mse <- mean(residuals^2)
mse
```

15. Discuss results, including interpretation of coefficients and model's accuracy. (Next Page)

Exercise on Linear Regression [3/8]

kcg.edu

Perform linear regression analysis in RStudio using a sample dataset

15. Discuss the results, including the interpretation of coefficients and the model's accuracy.

Let's discuss the results of linear regression analysis performed in this exercise, focusing on the interpretation of coefficients and model's accuracy.

1. Coefficients:

Linear regression model you created aims to predict "Sales" based on "TV" advertising expenses. Coefficient for "TV" predictor variable can be found in the regression summary output.

Coefficient for "TV" represents the estimated change in "Sales" for a one-unit increase in "TV" advertising expenses, holding other variables constant. If the coefficient is positive, it implies that as "TV" advertising expenses increase, "Sales" is expected to increase. Conversely, a negative coefficient would imply a negative relationship.

In your regression summary, if coefficient for "TV" is, for instance, 0.0475, this means that for each additional unit of money spent on TV advertising, the expected increase in sales is 0.0475 units, assuming all other factors remain constant.

2. P-values:

In regression summary, you will also observe p-value associated with the coefficient for "TV." This p-value is used to test the null hypothesis that coefficient is equal to zero (no effect). If the p-value is less than your chosen significance level (commonly 0.05), it indicates that the coefficient is statistically significant. In this case, a low p-value $2x10^{-16}$ would suggest that "TV" advertising expenses have a statistically significant impact on "Sales".

3. R-squared (R^2):

R-squared is a measure of the goodness of fit of your regression model. In your regression summary, you should see the R-squared value.

An R-squared value closer to 1 indicates that the model explains a larger proportion of the variance in "Sales." For example, if R-squared is 0.75, it means that 75% of the variability (unpredictability/inconsistency/unevenness) in "Sales" is explained by the "TV" advertising expenses in your model.

Interpretation:

- If the coefficient for "TV" is positive and statistically significant (i.e. $p\text{-value} < 0.05$), you can conclude that there is a positive linear relationship between "TV" advertising expenses and "Sales“. Therefore as you invest more in TV advertising, you can expect an increase in sales.
- R-squared value tells how well the model fits data. A higher R-squared suggests that "TV" advertising expenses explain a significant portion of the variance in "Sales."

Remember that linear regression assumes a linear relationship between the predictor and the response variable. Interpretation of the coefficients and model's accuracy is based on this assumption. If relationship is not truly linear, model may not be accurate in making predictions.

It's also essential to consider other factors, such as the residual plots and overall context of your analysis for assessing model's validity and whether it provides meaningful insights into the relationship between advertising expenses and sales.

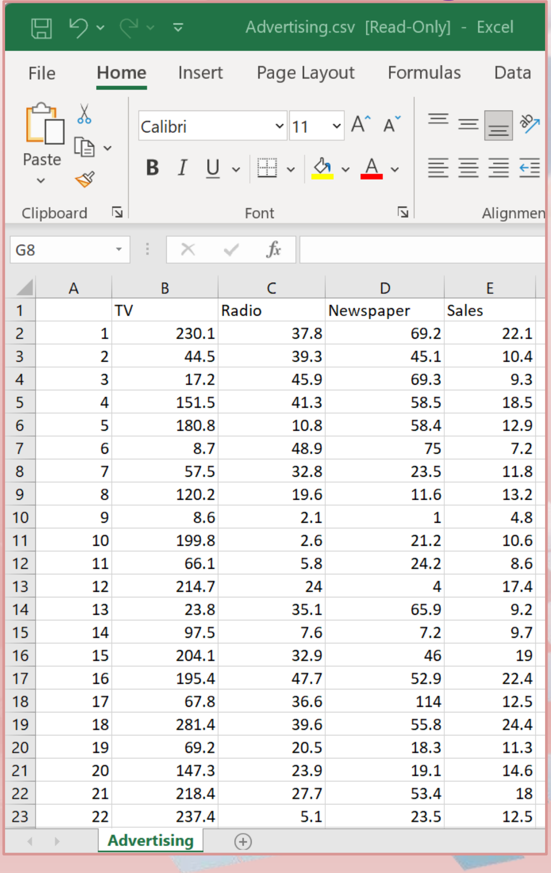
Exercise on Linear Regression [4/8]

kcg.edu

DATASET: Advertising.csv

Download this dataset from:

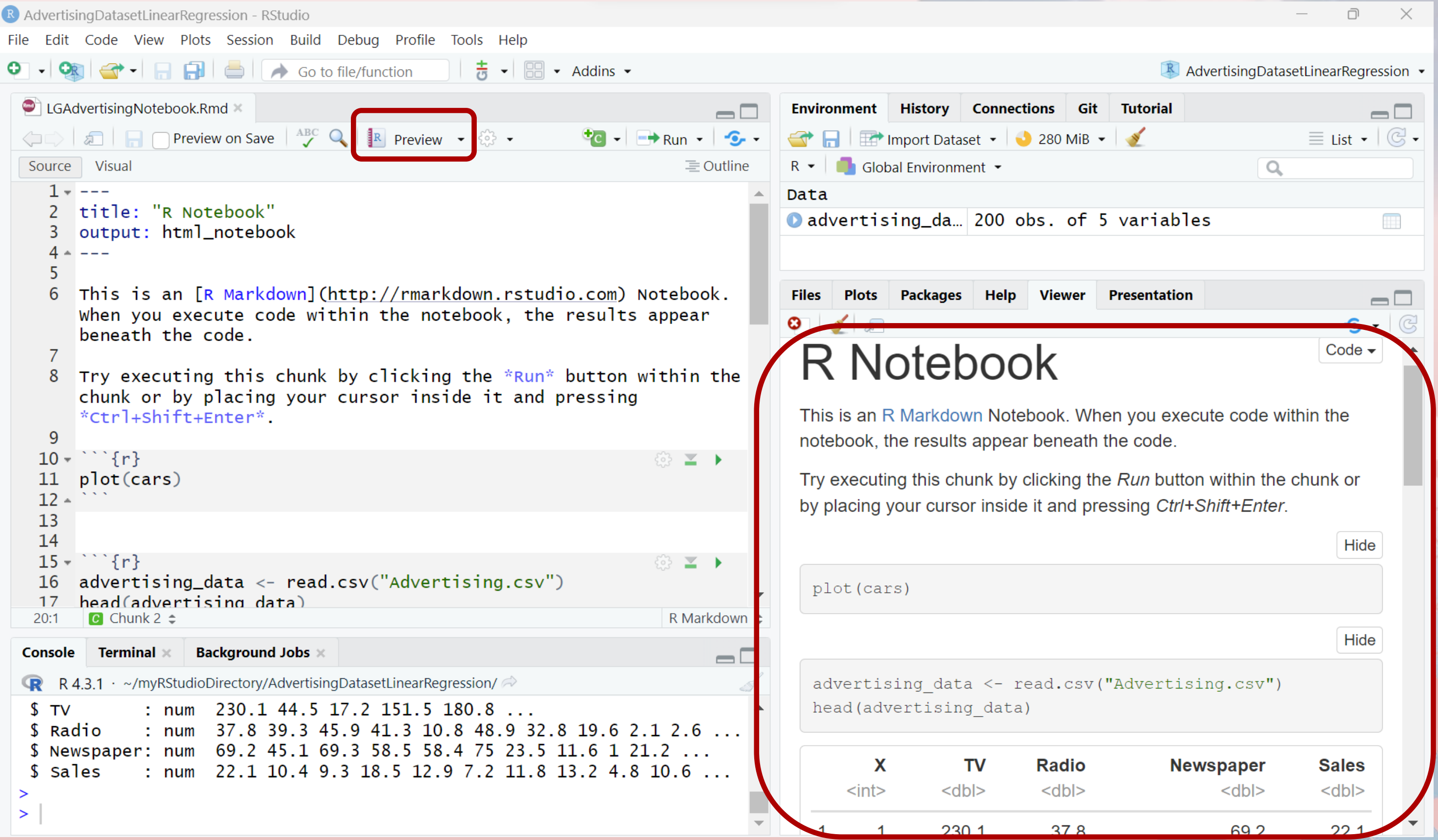
https://www.kaggle.com/datasets/bumba5341/advertisingcsv/



	A	B	C	D	E
1		TV	Radio	Newspaper	Sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6
13	12	214.7	24	4	17.4
14	13	23.8	35.1	65.9	9.2
15	14	97.5	7.6	7.2	9.7
16	15	204.1	32.9	46	19
17	16	195.4	47.7	52.9	22.4
18	17	67.8	36.6	114	12.5
19	18	281.4	39.6	55.8	24.4
20	19	69.2	20.5	18.3	11.3
21	20	147.3	23.9	19.1	14.6
22	21	218.4	27.7	53.4	18
23	22	237.4	5.1	23.5	12.5

Exercise on Linear Regression [5/8]

Click on Preview; then R Notebook (HTML) will appear on the right-side pane


AdvertisingDatasetLinearRegression - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
AdvertisingDatasetLinearRegression
LGAdvertisingNotebook.Rmd
Preview on Save
Preview
Run
Source
Visual
Outline
20:1
Chunk 2
R Markdown
Environment History Connections Git Tutorial
Import Dataset
280 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Code
Hide
Hide
Console Terminal Background Jobs
R 4.3.1 · ~/myRStudioDirectory/AdvertisingDatasetLinearRegression/

~~~~
---
title: "R Notebook"
output: html_notebook
---

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook.
When you execute code within the notebook, the results appear
beneath the code.

Try executing this chunk by clicking the *Run* button within the
chunk or by placing your cursor inside it and pressing
*Ctrl+Shift+Enter*.

```{r}
plot(cars)
```


```{r}
advertising_data <- read.csv("Advertising.csv")
head(advertising_data)
~~~~

```
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
>
> 
```

Data

advertising_da... 200 obs. of 5 variables

R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
plot(cars)
```

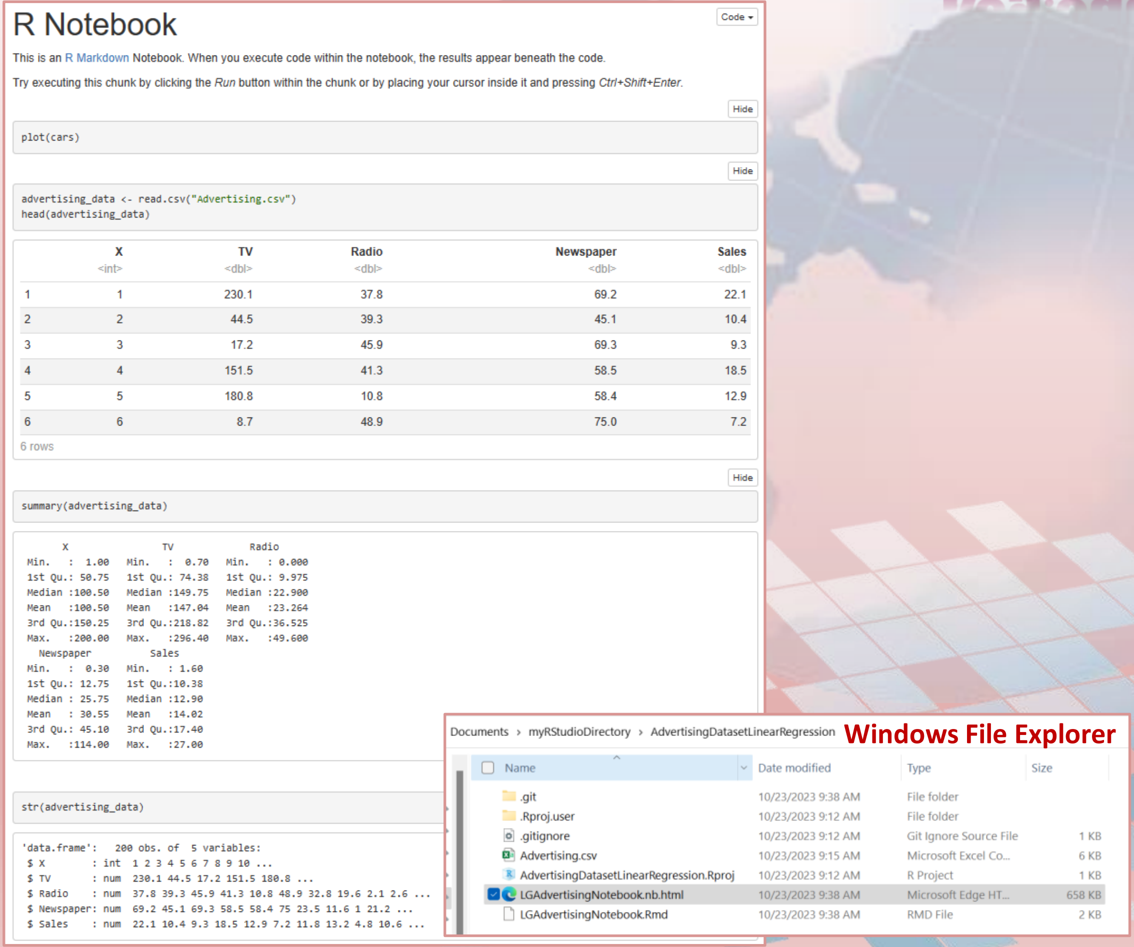
```
advertising_data <- read.csv("Advertising.csv")
head(advertising_data)
```

	X <int>	TV <dbl>	Radio <dbl>	Newspaper <dbl>	Sales <dbl>
1	1	230.1	37.8	69.2	22.1

Exercise on Linear Regression [6/8]

kcg.edu

From Windows File Explorer, you can open R Notebook (HTML) in your browser; refresh every time you make changes by executing commands in RStudio



R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
plot(cars)
```

```
advertising_data <- read.csv("Advertising.csv")
head(advertising_data)
```

	X	TV	Radio	Newspaper	Sales
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75.0	7.2

6 rows

```
summary(advertising_data)
```

```
       X                TV             Radio       
 Min.   :  1.00   Min.   :  0.70   Min.   : 0.000  
 1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.: 9.975  
 Median :100.50   Median :149.75   Median :22.900  
 Mean   :100.50   Mean   :147.04   Mean   :23.264  
 3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.525  
 Max.   :200.00   Max.   :296.40   Max.   :49.600  
   Newspaper          Sales      
 Min.   :  0.30   Min.   : 1.60  
 1st Qu.: 12.75   1st Qu.:10.38  
 Median : 25.75   Median :12.90  
 Mean   : 30.55   Mean   :14.02  
 3rd Qu.: 45.10   3rd Qu.:17.40  
 Max.   :114.00   Max.   :27.00  
```

```
str(advertising_data)
```

```
'data.frame':   200 obs. of  5 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

Windows File Explorer

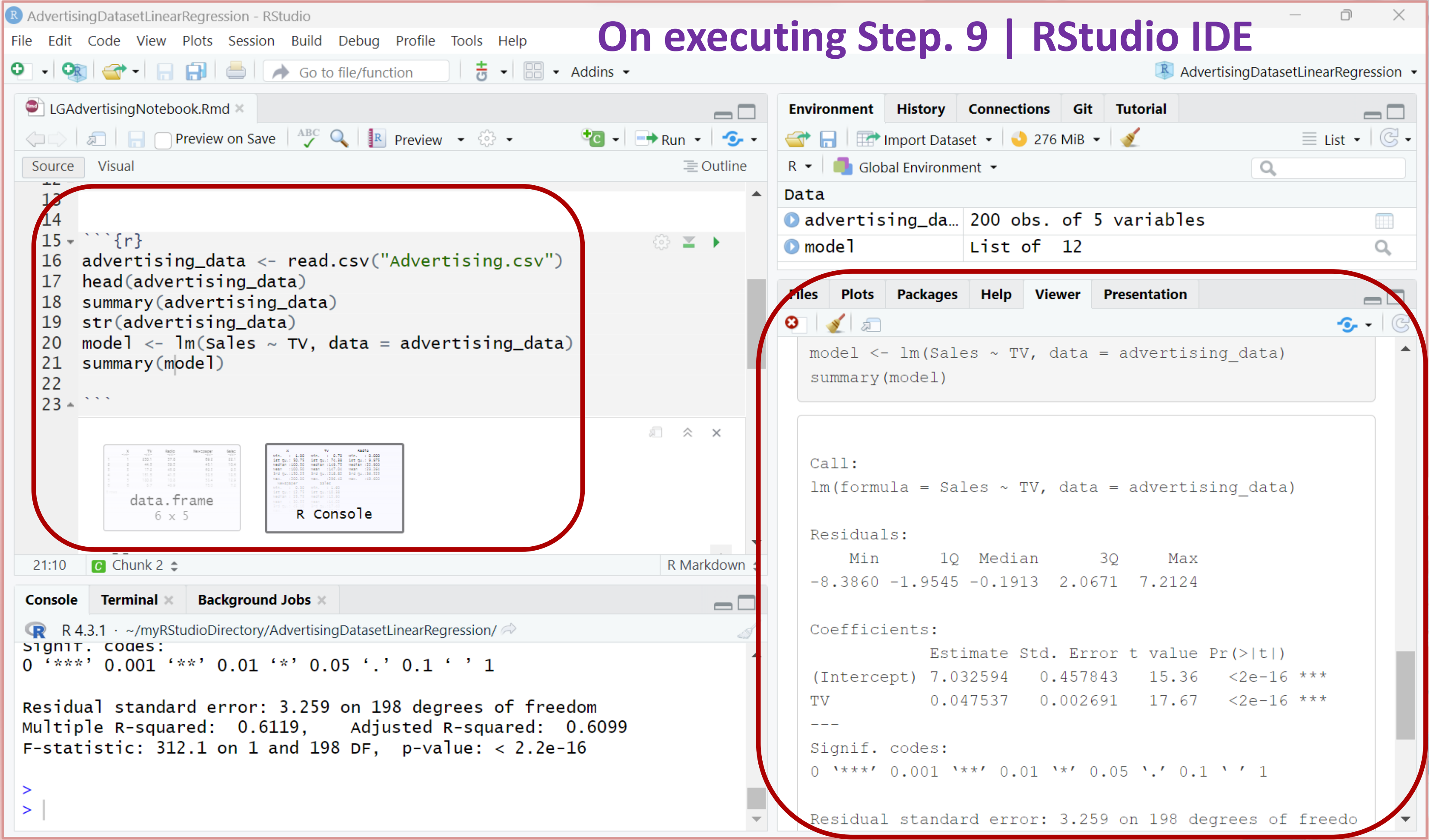
Documents › myRStudioDirectory › AdvertisingDatasetLinearRegression

□ Name	Date modified	Type	Size
.git	10/23/2023 9:38 AM	File folder	
.Rproj.user	10/23/2023 9:12 AM	File folder	
.gitignore	10/23/2023 9:12 AM	Git Ignore Source File	1 KB
Advertising.csv	10/23/2023 9:15 AM	Microsoft Excel Co...	6 KB
AdvertisingDatasetLinearRegression.Rproj	10/23/2023 9:12 AM	R Project	1 KB
☑ LGAdvertisingNotebook.nb.html	10/23/2023 9:38 AM	Microsoft Edge HT...	658 KB
LGAdvertisingNotebook.Rmd	10/23/2023 9:38 AM	RMD File	2 KB

Exercise on Linear Regression [7/8]

kcg.edu

On executing Step. 9 | RStudio IDE



~~~
```{r}
advertising_data <- read.csv("Advertising.csv")
head(advertising_data)
summary(advertising_data)
str(advertising_data)
model <- lm(Sales ~ TV, data = advertising_data)
summary(model)

```
~~~

Data

advertising_da...	200 obs. of 5 variables
model	List of 12

Console:

```
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: < 2.2e-16

>
>
```

Viewer:

```
model <- lm(Sales ~ TV, data = advertising_data)
summary(model)
```

```
Call:
lm(formula = Sales ~ TV, data = advertising_data)

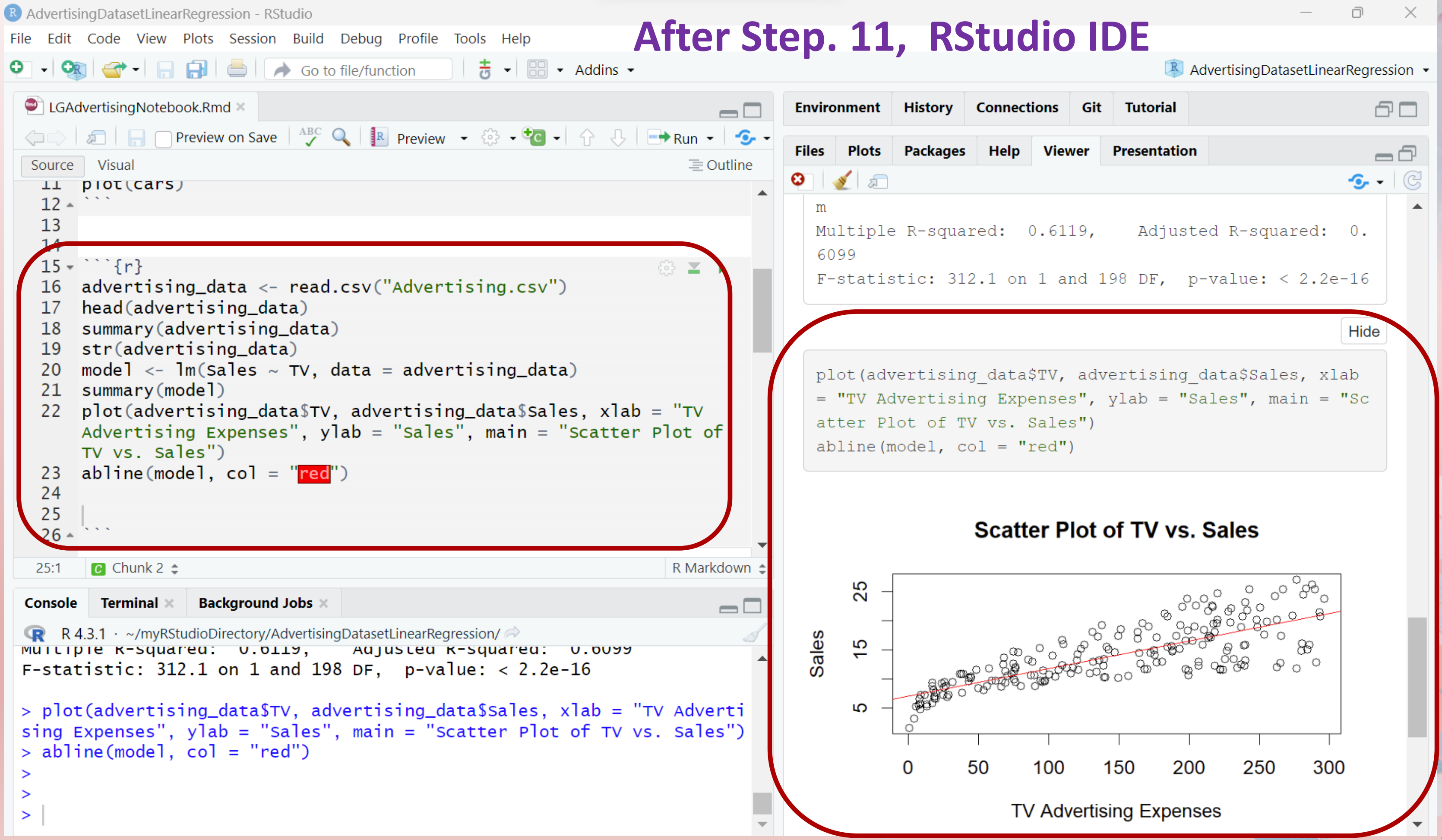
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedo
```

Exercise on Linear Regression [8/8]

After Step. 11, RStudio IDE


AdvertisingDatasetLinearRegression - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
AdvertisingDatasetLinearRegression
LGAdvertisingNotebook.Rmd
Preview on Save
Preview
Run
Source
Visual
Outline
25:1
Chunk 2
R Markdown
Console
Terminal
Background Jobs
R 4.3.1 · ~/myRStudioDirectory/AdvertisingDatasetLinearRegression/
Environment
History
Connections
Git
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation
Hide
Scatter Plot of TV vs. Sales
Sales
5
15
25
0
50
100
150
200
250
300
TV Advertising Expenses

```
plot(cars)
```

```{r}
advertising_data <- read.csv("Advertising.csv")
head(advertising_data)
summary(advertising_data)
str(advertising_data)
model <- lm(Sales ~ TV, data = advertising_data)
summary(model)
plot(advertising_data$TV, advertising_data$Sales, xlab = "TV
Advertising Expenses", ylab = "Sales", main = "Scatter Plot of
TV vs. Sales")
abline(model, col = "red")


```
```

```
m
Multiple R-squared:  0.6119,	Adjusted R-squared:  0.
6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
plot(advertising_data$TV, advertising_data$Sales, xlab
= "TV Advertising Expenses", ylab = "Sales", main = "Sc
atter Plot of TV vs. Sales")
abline(model, col = "red")
```

```
Multiple R-squared:  0.6119,	Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16

> plot(advertising_data$TV, advertising_data$Sales, xlab = "TV Adverti
sing Expenses", ylab = "Sales", main = "Scatter Plot of TV vs. Sales")
> abline(model, col = "red")
>
>
> 
```
```

# Homework Activity: Further Practice

kcg.edu

**To practice linear regression analysis in RStudio and understand how to interpret the results of a regression model; repeat analysis for "Radio" and "Newspaper" advertising expenses to see how these relate to "Sales".**

**Analyze the findings obtained in your R Notebook HTML format using the guidance provided in Appendix (Next Slide)**

# Appendix (related to Point 10 of Slide 9)

## Step 10. Interpret results from the regression summary, paying attention to coefficients, p-values, and R-squared.

Interpreting the results from a regression summary in R, focusing on coefficients, p-values, and R-squared, is essential to understand the relationships within your dataset.
Here's how to interpret these key elements:

### 1. Significance (importance) of Coefficients:

- In a linear regression summary, you will see a table that includes coefficients for each predictor variable (independent variable) in your model. These coefficients represent change in dependent variable (in your case, "Sales") for a one-unit change in predictor variable, while holding other predictors constant.
- Coefficients are typically labeled as "Estimate," and they represent the slope of regression line. For instance, if coefficient for "TV" is 0.0475, it means that for every additional unit of TV advertising expenses, Sales is expected to increase by 0.0475 units, assuming all other factors remain constant.

### 2. P-values: Validating Null/Alternative Hypothesis

- P-values associated with each coefficient test the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (usually less than 0.05) indicates that predictor variable is statistically significant in predicting the dependent variable.
- If the p-value is less than your chosen significance level (commonly 0.05), you can reject the null hypothesis and conclude that predictor has a statistically significant effect on the dependent variable.

### 3. R-squared ($R^2$):

- R-squared is a measure of how well the independent variables (predictors) explain the variability in dependent variable. It is a value between 0 and 1, where a higher value indicates a better fit of model to the data.
- An R-squared value of 0 means that the model does not explain any of the variability in the dependent variable, while an R-squared value of 1 means that the model explains all the variability.
- In the context of linear regression, R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. For example, if R-squared is 0.75, it means that 75% of the variance in Sales is explained by the predictor variables in your model.

In a nutshell, when interpreting the regression summary:

- Look at coefficients to understand the direction and strength of relationship between predictor variables and dependent variable.
- Examine p-values to determine whether predictor variables are statistically significant.
- Assess R-squared to gauge how well your model explains the variability in dependent variable.

A high R-squared, significant coefficients, and their directions can help you draw meaningful conclusions about the relationships within your data and predictive power of your model.

# Practice Class Activity: Linear Regression using Some Synthetic Data [1/2]

**This code generates synthetic data, fits a linear regression model, provides a summary of the model, plots data and regression line, creates a residual plot, calculates Mean Squared Error (MSE), and finally, offers a brief analysis of the outcome.**

**Step 1: Generate synthetic data**

```
set.seed(123)  # Set a seed for reproducibility
x <- 1:20
y <- 2 * x + rnorm(20, mean = 0, sd = 2)  # Generate synthetic data with some noise
```

Step 2: Perform linear regression

```
lm_model <- lm(y ~ x)
```

Step 3: Create a summary of regression modeling

```
summary(lm_model)
```

Step 4: Plot data and regression line

```
plot(x, y, main = "Linear Regression Example", xlab = "X", ylab = "Y")
abline(lm_model, col = "red")
```

Practice Class Activity: Linear Regression using Some Synthetic Data [2/2]

kcg.edu

Step 5: Create a residual plot

```
par(mfrow = c(1, 2))
plot(lm_model, which = 1)  # Residuals vs Fitted
plot(lm_model, which = 2)  # Normal Q-Q Plot
```

Step 6: Calculate the Mean Squared Error (MSE)

```
predicted <- predict(lm_model)
mse <- mean((y - predicted)^2)
cat("Mean Squared Error (MSE):", mse, "\n")
```

Step 7: Analyze the outcome

You can analyze the coefficient estimates, p-values, and R-squared from summary output.

Additionally, MSE provides a measure of the model's goodness of fit where lower values indicate better fit.