Multiple Linear Regression using R

Life is ten percent what you experience and ninety percent how you respond to it.
-Dorothy M. Neddermeyer

Outline

Introduction to Multiple Linear Regression

Learning Objectives

In this video, you will learn to:

- Build a Multiple Linear Regression using R.
- Understand why the Multiple Linear Regression Model is usually more suitable to use than the Simple Linear Regression Model.
- Evaluate the regression model and understand how to interpret the outputs to make inferences, predictions and data-driven recommendations.

Simple Linear Regression and Multiple Linear Regression in a Nutshell

	Simple Linear Regression	Multiple Linear Regression
Linear Model	$Y = \beta_0 + \beta_1 X$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$
Assumptions	1. Each observation is independent from the others. 2. The relationship between the predictor variable X and the response variable Y is linear. 3. The residuals are normally distributed. 4. The residuals are evenly scattered around the center line of zero, with no obvious pattern across all values of the predictor variable. In other words, the distribution of the residuals does not change with the value of the predictor variable.	1. Each observation is independent from the others. 2. The relationships between the predictor variables X and the response variable Y are linear. 3. There is no multicollinearity between the predictor variables. 4. The residuals are normally distributed. 5. The residuals are evenly scattered around the center line of zero, with no obvious pattern across all values of the predictor variable. In other words, the distribution of the residuals does not change with the value of the predictor variable.

Simple Linear Regression and Multiple Linear Regression in a Nutshell

	Simple Linear Regression	Multiple Linear Regression
Linear Model	$Y = \beta_0 + \beta_1 X$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$
Assumptions	1. Each observation is independent from the others. 2. The relationship between the predictor variable X and the response variable Y is linear. 3. The residuals are normally distributed. 4. The residuals are evenly scattered around the center line of zero, with no obvious pattern across all values of the predictor variable. In other words, the distribution of the residuals does not change with the value of the predictor variable.	1. Each observation is independent from the others. 2. The relationships between the predictor variables X and the response variable Y are linear. **3. There is no multicollinearity between the predictor variables.** 4. The residuals are normally distributed. 5. The residuals are evenly scattered around the center line of zero, with no obvious pattern across all values of the predictor variable. In other words, the distribution of the residuals does not change with the value of the predictor variable.

Case Scenario Revisit: Advertising Dataset

Action Plans from the Previous Video

- The DLP Team can evaluate how accurate the model has been in predicting the sales revenue by comparing the predicted sales revenue to the actual sales revenue.
- The DLP team should announce their results to the company to let the company decide on the next step they should take.
- The DLP Team could explore further on what kind of advertisement improves the sales revenue and whether there are other significant factors that could help them to improve their sales revenue prediction.
- The DLP Team should remember to constantly update their data and forecasts as old models may become inaccurate over time and irrelevant.

DIDM Framework

Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Describing and Exploring the Data

Ask – Formulate Focused Questions

1. **Ask**

- Based on the Simple Linear Regression model results, the team had decided to explore further on how different types of advertising media impact the sales revenue:
  - Would looking at advertising expenditure for different media separately better predict sales revenue?
  - Would spending \$50,000 on television advertisements and \$50,000 on radio advertisements result in more sales revenue than allocating \$100,000 to either television or radio only?
- The DLP Team could choose to create 3 Simple Linear Regression models:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{TV Advertisement}$$
$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{Radio Advertisement}$$
$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{Newspaper Advertisement}$$

DIDM Framework

Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Ask – Formulate Focused Questions

cont'd

① **Ask**

- The Multiple Linear Regression equation with the 3 predictor variables may be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Or can be formulated as:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper}$$

DIDM Framework

Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Acquire – Obtain the Best Available Data

2 **Acquire**

- From R code:

```
df.advert <- read.csv("Advertising.csv
    ", header = TRUE)
head(df.advert) %>% select(1:4)
```

```
     TV Radio Newspaper Sales.Revenue
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          12.0
4 151.5  41.3      58.5          16.5
5 180.8  10.8      58.4          17.9
6   8.7  48.9      75.0           7.2
```

- Take note that the Sales Revenue are in units of thousand ('000).

DIDM Framework

Source: https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions

Analyse – Critically Appraise and Analyse the Data

cont'd

3. **Analyse**

- Check whether there are any missing cells or duplicates in our data by using the following code:

```
sum(is.na(df.advert))
```

```
[1] 0
```

```
sum(duplicated(df.advert))
```

```
[1] 0
```

- To check the number of cities in our dataset, we use:

```
nrow(df.advert)
```

- [1] 200

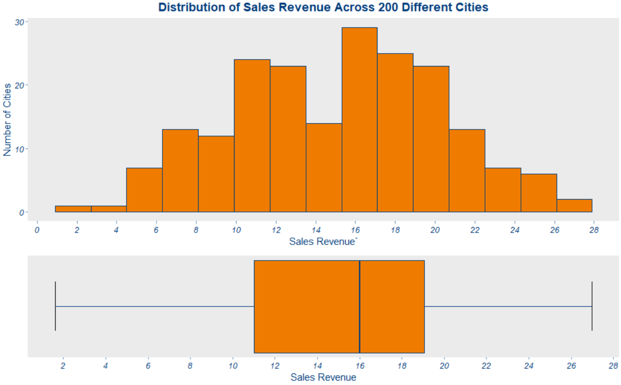
DIDM Framework

Source: https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions

Analyse – Critically Appraise and Analyse the Data

Data Exploration

Check the distribution of the dependent response variable.



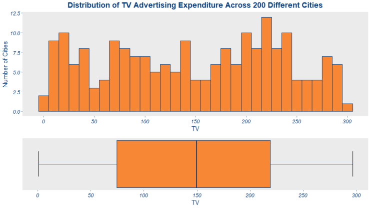
- `skewness(df.advert$Sales.Revenue)`
- `[1] -0.07263683`

Analyse – Critically Appraise and Analyse the Data

Data Exploration (cont'd)

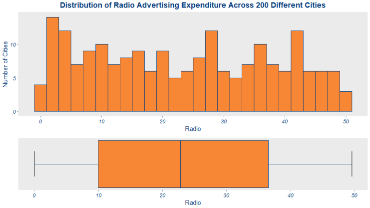
Check the distribution for each of the independent predictor variables.

TV



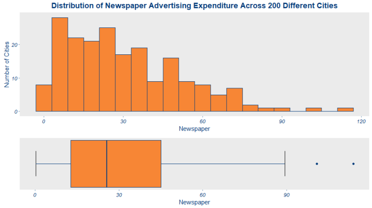
- `skewness(df.advert$TV)`
- [1] -0.06880905

Radio



- `skewness(df.advert$Radio)`
- [1] 0.09276672

Newspaper



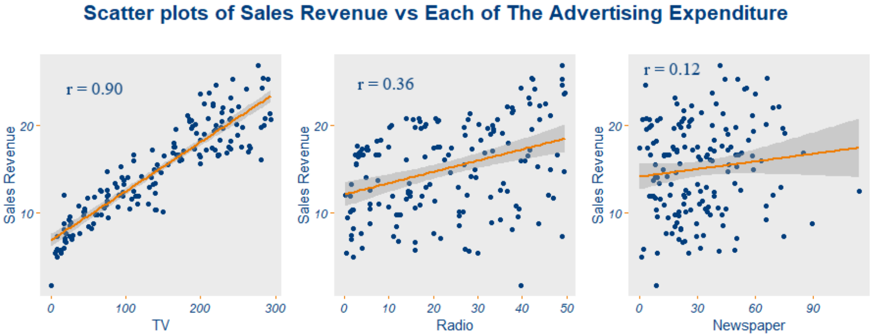
- `skewness(df.advert$Newspaper)`
- [1] 0.8813443

Checking the Model Assumptions

Assumptions of the Multiple Linear Regression

Assumptions check

1. Each observation is independent from the others.
2. The relationships between the predictor variables X and the response variable Y are linear.



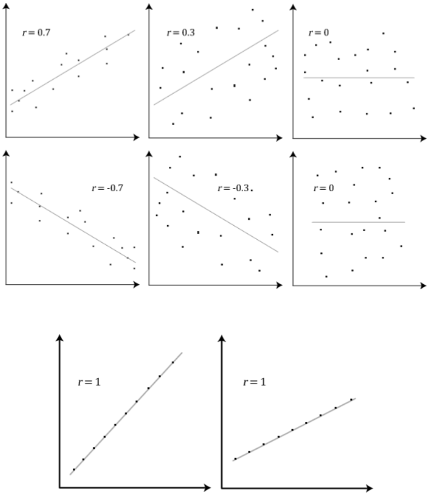
Recap on Correlation Coefficient from DLP Basic

Correlation Coefficient, r

- r ranges between -1 and 1.
- $r > 0$: **Positive** linear correlation.
- $r < 0$: **Negative** linear correlation.
- $r = 0$: **No** linear correlation.

Strength of Correlation Coefficient

- $-1 < r < -0.7$ or $0.7 < r < 1$: **Strong** Correlation.
- $-0.7 \leqslant r < -0.3$ or $0.3 < r \leqslant 0.7$: **Moderate** Correlation.
- $-0.3 \leqslant r < 0$ or $0 < r \leqslant 0.3$: **Weak** Correlation.

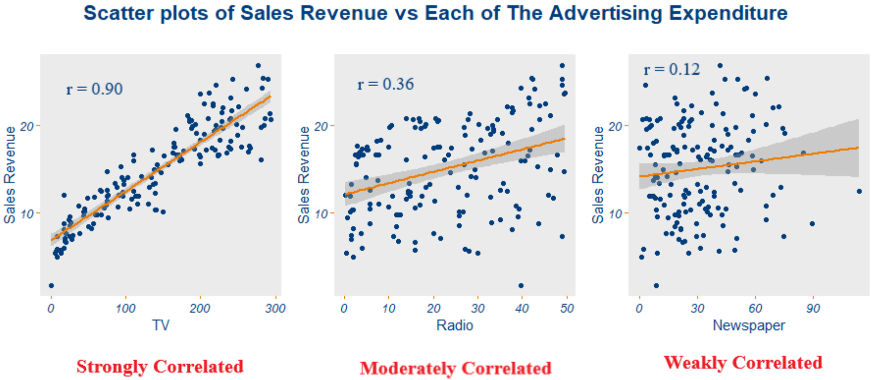


Source: `https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php`

Assumptions of the Multiple Linear Regression

Assumptions check

1. Each observation is independent from the others.
2. The relationships between the predictor variables X and the response variable Y are linear.



Assumptions of the Multiple Linear Regression

Assumptions Check (cont'd)

3. There is no multicollinearity between the predictor variables.

   - ```
     correlation <- cor(df.advert[,1:3])
     correlation
     ```

     ```
                      TV      Radio  Newspaper
     TV        1.00000000 0.05480866 0.05664787
     Radio     0.05480866 1.00000000 0.35410375
     Newspaper 0.05664787 0.35410375 1.00000000
     ```

4. The residuals are normally distributed.
5. Homoscedasticity, which means the variance of the residuals are the same across all values of the predictor variables.
   - Plot the standardised residuals versus the predicted values in a scatterplot.
   - Standardised residuals is useful because the raw residual may have non-constant variance.
     ```

Building a Multiple Linear Regression Model

Splitting the Dataset into Training and Test Datasets

- Similar to our previous video, we split the data randomly into 80–20 ratio, resulting in 160 training observations vs. 40 test observations.
- ```
  set.seed(10)
  dt = sort(sample(nrow(df.advert), nrow(df.advert)*.8))
  train<-df.advert[dt,]
  test<-df.advert[-dt,]
  ```
  ```

Fitting the Multiple Linear Regression on the Training Dataset

- Multiple Linear Regression model using R programming:

```
lm1 <- lm(Sales.Revenue ~ TV + Radio + Newspaper, train)
```

- Parameters are estimated using the same approach used in the Simple Linear Regression, which is by minimising the residuals sum or squares.

$$
\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \ldots - \beta_p X_{ip})^2
\end{aligned}
$$

- where (Y_i) is the actual Y, $(\hat{Y}_i)$ is the predicted Y and β_0, β_1, β_2 are the coefficient estimates.

Fitting the Multiple Linear Regression on the Training Dataset

cont'd

- Use the `summary()` function.

```
summary(lm1)
```

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

Fitting the Multiple Linear Regression on the Training Dataset

cont'd

- Use the summary() function.

```
summary(lm1)
```

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

- where the intercept coefficient, $\beta_0 = 4.5596$, TV coefficient, $\beta_1 = 0.0553$, Radio coefficient, $\beta_2 = 0.1041$, Newspaper coefficient, $\beta_3 = 0.0022$.

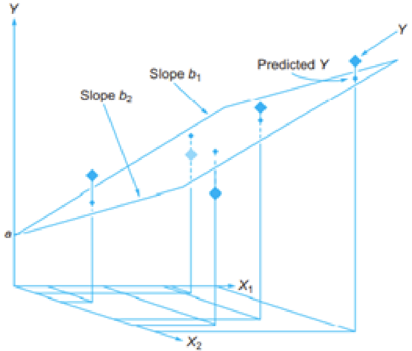
Evaluating a Multiple Linear Regression Model

Goodness of Fit – Adjusted R-squared

- Recall that R^2 is defined as
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$
- where RSS is the residual sum of squares, and TSS is the total sum of squares.
- Since RSS will always decrease as more predictor variables are added to the model, the R^2 will always increase as more predictor variables are added.
- The formula for the adjust R^2 is
$$\text{Adjusted}R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$
- where n is the number of observations and d is the number of variables.

Plotting the Relationship in Multiple Linear Regression

- Unlike Simple Linear Regression, we cannot always plot the relationship in a Multiple Linear Regression using a scatterplot.
- When there is only one predictor (like in Simple Linear Regression), we can plot the relationship between the predictor as a line on a flat 2-dimensional space.
- When there are two predictors, we have a flat prediction plane on a 3-dimensional space as shown:



Hypothesis Testing on Multiple Linear Regression

F-test

- H_0 : The model with no predictor variable fits the data as good as the current regression model ($\beta_1 = \beta_2 = \beta_3 = 0$).
- H_1 : The current regression model fits the data better than the model with no predictor variable (At least one of $\beta_1, \beta_2, \beta_3 \neq 0$).

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

Hypothesis Testing on Multiple Linear Regression

Using t-test

- $H_0 : \beta_j = 0.$
- $H_1 : \beta_j \neq 0.$
- With j representing the jth regression coefficient in the model.

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

Hypothesis Testing on Multiple Linear Regression

Using t-test

- $H_0 : \beta_j = 0$.
- $H_1 : \beta_j \neq 0$.
- With j representing the jth regression coefficient in the model.

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

Hypothesis Testing on Multiple Linear Regression

Using t-test

- $H_0 : \beta_j = 0.$
- $H_1 : \beta_j \neq 0.$
- With j representing the jth regression coefficient in the model.

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1395 -0.8170  0.0001  0.8485  3.7259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.559627   0.351193  12.983   <2e-16 ***
TV          0.055332   0.001573  35.180   <2e-16 ***
Radio       0.104089   0.009629  10.810   <2e-16 ***
Newspaper   0.002205   0.006829   0.323    0.747    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 156 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.9011 
F-statistic: 483.9 on 3 and 156 DF,  p-value: < 2.2e-16
```

Choosing the Best Multiple Linear Regression Model

$$\text{Sales Revenue} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper}$$

- Drop the Newspaper predictor variable and re-run the model:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio}$$

- ```
  lm2 <- lm(Sales.Revenue ~ TV + Radio, train)
  summary(lm2)
  ```

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2110 -0.8043  0.0230  0.8430  3.7284 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.604900   0.321058   14.34   <2e-16 ***
TV          0.055316   0.001568   35.29   <2e-16 ***
Radio       0.105236   0.008924   11.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 157 degrees of freedom
Multiple R-squared:  0.9029,	Adjusted R-squared:  0.9017 
F-statistic:   730 on 2 and 157 DF,  p-value: < 2.2e-16
```
```

Choosing the Best Multiple Linear Regression Model

$$\text{Sales Revenue} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper}$$

- Drop the Newspaper predictor variable and re-run the model:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio}$$

- ```
  lm2 <- lm(Sales.Revenue ~ TV + Radio, train)
  summary(lm2)
  ```

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2110 -0.8043  0.0230  0.8430  3.7284 

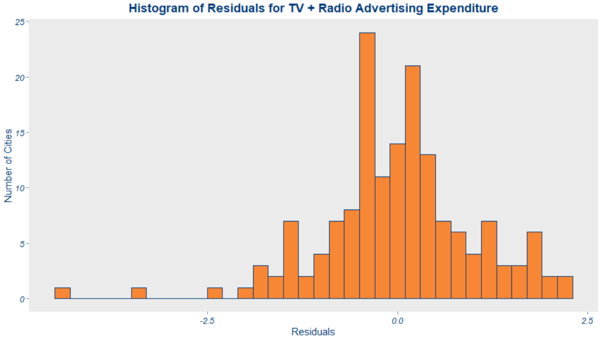
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.604900   0.321058   14.34   <2e-16 ***
TV          0.055316   0.001568   35.29   <2e-16 ***
Radio       0.105236   0.008924   11.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 157 degrees of freedom
Multiple R-squared:  0.9029,	Adjusted R-squared:  0.9017 
F-statistic:   730 on 2 and 157 DF,  p-value: < 2.2e-16
```
```

Residual Assumption Check on the Training Dataset

- To check assumption 3 (Residuals must be normally distributed)

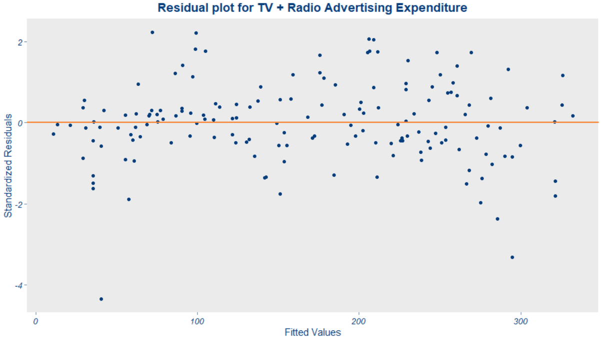
```
ggplot(lm2, aes(x=residuals2))
   +
    geom_histogram(binwidth
       =0.2, color="#003D7C",
       fill="#EF7C00") +
    nus_theme() +
    labs(x="Residuals",y="
       Number of Cities", title
       ="Histogram of Residuals
        for TV + Radio
       Advertising Expenditure
       ")
```



Residual Assumption Check on the Training Dataset

- To check assumption 4 (Homoscedasticity – Residuals are evenly scattered)

```
ggplot(lm2, aes(x=TV+Radio,y=
    residuals2)) +
     geom_point(color="#003D7C")
        +
     geom_abline(slope=0,color
        ="#EF7C00") +
     nus_theme() +
     labs(x="Fitted Values",y="
        Standardized Residuals",
        title="Residual plot for
         TV + Radio Advertising
        Expenditure")
```



Interpreting the Multiple Linear Regression Equation

- The regression equation obtained from `summary()` function:

$$\text{Sales Revenue} = 4.605 + 0.055 \times \text{TV} + 0.105 \times \text{Radio}$$

- where $\beta_0 = 4.605, \beta_1 = 0.055$ and $\beta_2 = 0.105$.
- If the company spends \$100 in TV and \$5 in radio, we can estimate sales revenue to be about \$10,630 (Hint: substitute the TV and Radio with the values given).
- If the company spends \$200 in TV and \$5 in radio, we can estimate sales revenue to be about \$16,130 (Hint: substitute the TV and Radio with the values given).
- If the company increases spending on TV by \$100 while holding radio constant, we can expect an average increase in sales revenue as much as \$5,500.

Confidence Interval & Prediction Interval of the Y variable

- To spend \$150 for advertising on TV and \$30 on Radio in each cities,

```
predict(lm2, data.frame(TV=150, Radio=30), interval="confidence")
```

  - The 95% CI for the mean sales revenue when the TV and Radio are \$150 and \$30 respectively is $[15.767, 16.351]$ ('000).

- To spend \$150 for advertising on TV and \$30 on Radio in a particular city,

```
predict(lm2, data.frame(TV=150, Radio=30), interval="prediction")
```

  - The 95% PI for the sales revenue in a particular city that is spending \$150 and \$30 on TV and Radio advertisement respectively is $[12.726, 19.392]$ ('000).

Evaluating the Regression Model using the Test Dataset

1. MSE: measures the average squared of the errors.
2. MAE: measures the mean absolute error.
3. RMSE: measures the deviance of the predicted value form the best fit line.
4. MAPE: measures the average of absolute percentage errors.

- The values can be obtained by using these codes in R:

```
mse_data <- mean((actual - predicted)^2)
mae_data <- mean(abs(actual - predicted))
rmse_data <- sqrt(mse_data)
mape_data <- mean(abs((actual - predicted)/actual))*100
```

Summary of the Regression Model Evaluation

Summary table:

	Training Data	Test Data
MSE	2.772406	2.487013
MAE	1.234186	1.275588
RMSE	1.665054	1.577027
MAPE	11.43309	10.53649

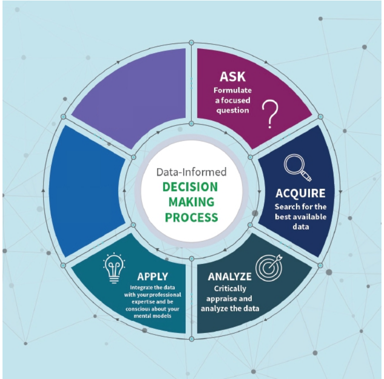
- The values between training set and test set are very similar to each other.
- We can therefore conclude that our training model is good and there is no overfitting.

Apply – Integrating the Model with Professional Expertise

4. **Apply**

- The company could decide how much budget to allocate to TV and Radio advertisements.
- The regression coefficients seem to suggest that within the range of available data, there is likely more increase in the Sales Revenue with every $ spent on Radio advertisements as compared to TV advertisements (Hint: Compare the value of β_1 and β_2).
- The company should not allocate budget for Newspaper advertisements as Newspaper is not a significant predictor for the Sales Revenue.

DIDM Framework



Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Announce – Decide and Communicate

5. **Announce**

- After applying some changes to their advertisement expenditure budget based on the model results, the DLP team could help Chairismatic to evaluate how accurate the predictions of the model were so that Chairismatic can announce the result to the entire company.

DIDM Framework

Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Assess – Monitor the Outcome

6 **Assess**

- The Chairismatic company's leaders together with the DLP team could assess whether the model is good enough to rely on or whether the model could still be further improved.
- Investigate whether there are other factors related to the Sales Revenue that may improve the accuracy of the Sales Revenue predictions.

DIDM Framework

Source: https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions

Ask – Formulate Focused Questions Revisit

- Q: Should we use the formula Sales Revenue $= \beta_0 + \beta_1$TV $+ \beta_2$Radio $+ \beta_3$Newspaper for prediction?
  - The finalised Multiple Linear Regression model that can be used to predict Sales Revenue is:

$$\text{Sales Revenue} = 4.605 + 0.055 \times \text{TV} + 0.105 \times \text{Radio}$$

  - Compared to the previous Simple Linear Regression model which only looked at the **total** advertising expenditure, this model helps us to understand better how to allocate the advertising budget across the different media types.
- Q: Would looking at advertising expenditure for different media separately better predict sales revenue?
  - What we know is that TV and Radio both increases Sales Revenue, but the regression coefficients suggest they have different impact on sales revenue.
- Q: Would spending $50,000 on television advertisements and $50,000 on radio advertisements result in more sales revenue than allocating $100,000 to either television or radio only?
  - Based on the regression equation, it seems that allocating $100,000 on radio advertisements only results in the most increase of sales revenue (Hint: Substitute the value into the finalised regression equation).

Summary

Extend the concepts of the Simple Linear Regression to the Multiple Linear Regression model

- Split dataset into training and test sets.
- Check assumptions.
- Evaluate the Multiple Linear Regression model.
- Using the application of DIDM framework.

Up next

Discuss the pitfalls and problems encountered in Multiple Linear Regression.

References

- C. Dissertation, “Assumptions of multiple linear regression,” Aug 2021.
- “Standardized residual definition,” Jul 2021.