

Simple Linear Regression using R

All Models are Wrong, but some are Useful.
-George Box

Outline

- 1 Introduction to Linear Regression
- 2 Describing and Exploring the Data
- 3 Building a Linear Regression Model
- 4 Checking the Model Assumptions
- 5 Evaluating the Simple Linear Regression Model
- 6 Summary

Introduction to Linear Regression

Applications of Linear Regression

Applications of Linear Regression in real life.

- Business: Predict revenue based on the amount spent on advertisements.
- Agriculture: Explore the effect of water and fertilizer on a crop's yield.
- Biomedical research: Understand the relationship between cholesterol level and dosage of drug.
- Education: Predict a learner's final test score based on their pre-test scores.
- Sport: Predict an NBA player's points scored by the number of yoga sessions and weightlifting sessions they had.

Learning Objectives

In this video, you will learn to:

- Build a Simple Linear Regression using R.
- Interpret Confidence Intervals and the Prediction Interval.
- Evaluate the regression model using a training and a test dataset.
- Incorporate Data-Informed Decision Making (DIDM) framework into the Simple Linear Regression model.

Use Case Scenario: Advertising Dataset

Chairismatic

Specializes in producing ergonomic chairs that are comfortable, especially, for long hours of working from home.

Goal

To provide insights on how to improve the sales revenue of ergonomic chairs and to allocate advertising budget appropriately.



Describing and Exploring the Data

Data-Informed Decision Making (DIDM) Framework

Recall from Data Literacy Program Basic,



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

Ask – Formulate Focused Questions

1 Ask

- Is there an association between advertising expenditure and sales revenue?
- What is the strength of the association?
- How accurately can they use advertising expenditure to predict the sales revenue?

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(Sales.Revenue,  
  Advertising.Expenditure)
```

| | Sales.Revenue | Advertising.Expenditure |
|---|---------------|-------------------------|
| 1 | 22.1 | 337.1 |
| 2 | 10.4 | 128.9 |
| 3 | 12.0 | 132.4 |
| 4 | 16.5 | 251.3 |
| 5 | 17.9 | 250.0 |
| 6 | 7.2 | 132.6 |

- Note that the Sales Revenue are in units of thousand ('000).
- As advertising expenditure and sales revenue are continuous, a Simple Linear Regression model can be used.

DIDM Framework



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

Analyse – Critically Appraise and Analyse the Data

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

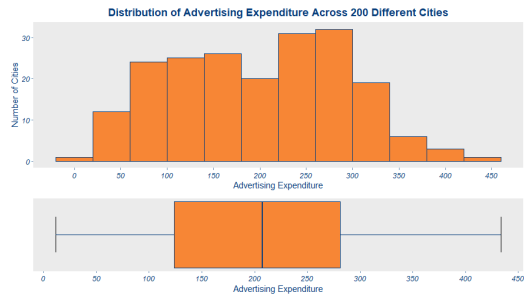
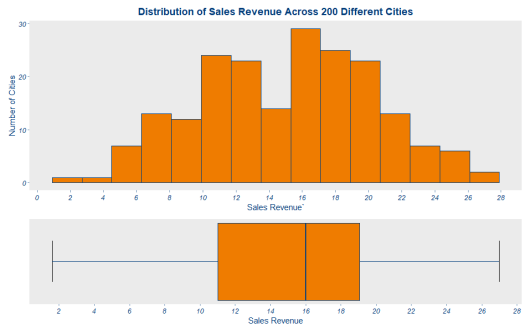


Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Analyse – Critically Appraise and Analyse the Data

Data Exploration

Check the distribution of each variables.



Analyse – Critically Appraise and Analyse the Data

Data Exploration (cont'd)

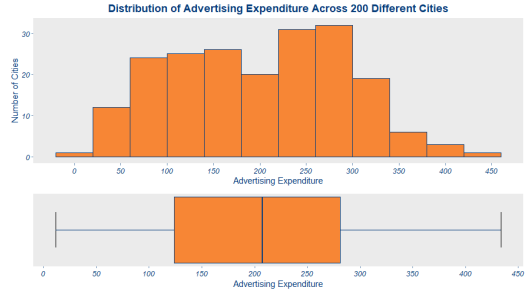
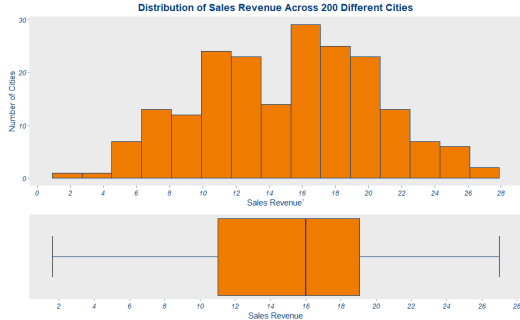
Check the skewness of the distribution

- $-0.5 \leq \text{skewness} \leq 0.5$: The distribution is approximately symmetric.
- $-1 \leq \text{skewness} < -0.5$ or $0.5 < \text{skewness} \leq 1$: The distribution is moderately skewed.
- $\text{skewness} < -1$ or $\text{skewness} > 1$: The distribution is highly skewed.

Analyse – Critically Appraise and Analyse the Data

Data Exploration (cont'd)

Check the distribution of each variables.



- `skewness(df.advert$Sales.
Revenue)`

- `skewness(df.advert$
Advertising.Expenditure)`

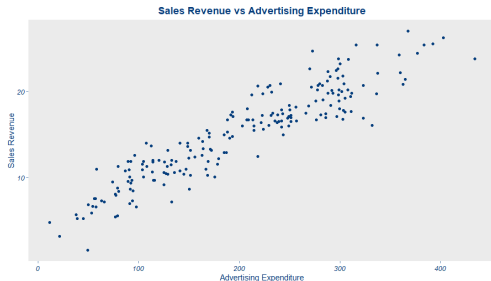
- `[1] 0.04874815`

- `[1] -0.07263683`

Analyse – Critically Appraise and Analyse the Data

cont'd

- Check whether the relationship between the two variables, Advertising Expenditure and Sales Revenue, is linear.



- Based on the scatterplot, there is a linear relationship between Advertising Expenditure and Sales Revenue.
- Linear correlation coefficient R code:

```
cor(df.advert$Sales.Revenue, df.advert$Advertising.Expenditure)
```

```
[1] 0.924917
```

Building a Linear Regression Model

What is a Linear Regression Model?

Linear Regression Model

Statistical approach for predicting the continuous response variable, Y , on the basis of one or more predictors, X .

- When we only have one predictor (one X -variable), we call our model a Simple Linear Regression model.
- When we have more than one predictor (multiple X -variables), we call our model a Multiple Linear Regression model.
- Mathematically, we can write the Simple Linear Regression model as:

$$Y = \beta_0 + \beta_1 X$$

- For our advertising dataset,

$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{Advertising Expenditure}$$

Splitting the Dataset into Training and Test Datasets

Training Set

Used to fit the parameters of the linear regression model.

Test Set

Used to evaluate the performance of the model.

- For Advertising Data, we split the data randomly into 80–20 ratio, resulting in 160 training observations vs. 40 test observations.

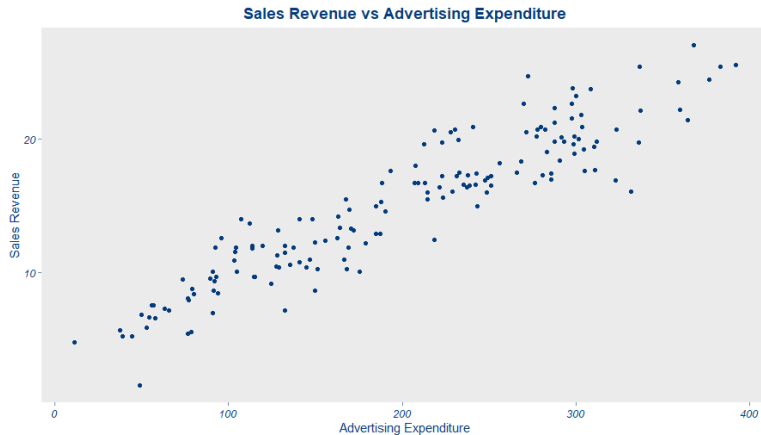
```
set.seed(10)
dt = sort(sample(nrow(df.advert), nrow(df.advert)*.8))
train<-df.advert[dt,]
test<-df.advert[-dt,]
```

Checking the Model Assumptions

Assumptions of the Simple Linear Regression

Assumptions check

- 1 Observations are independent from each other.
- 2 The relationship between the predictor variable X and the response variable Y is linear.



Assumptions of the Simple Linear Regression

Assumptions Check (cont'd)

- 3 The residuals are normally distributed.
- 4 The residuals are evenly scattered, and they do not change with the value of the predictor variable.
- Recall that

$$\text{Residuals} = \text{Actual value} - \text{Predicted value}$$

Fitting the Simple Linear Regression on the Training Dataset

- Simple Linear Regression model using R programming:

```
lm.slr <- lm(Sales.Revenue ~ Advertising.Expenditure, train)
```

- Residuals function in R code:

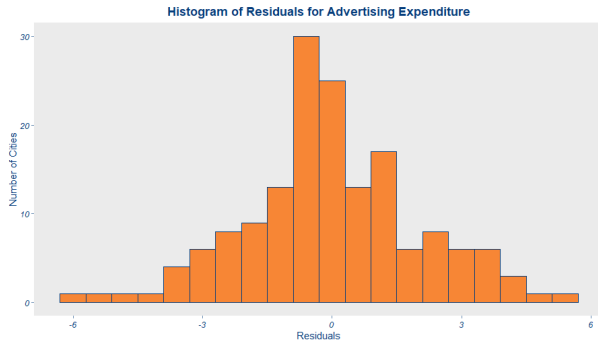
```
residuals <- resid(lm.slr)  
head(residuals)
```

| 1 | 2 | 3 | 4 | 6 | 7 |
|------------|------------|-----------|------------|------------|-----------|
| -0.4224779 | -0.8044735 | 0.6052623 | -1.3582858 | -4.2056099 | 1.4163808 |

Residual Assumption Check on the Training Dataset

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

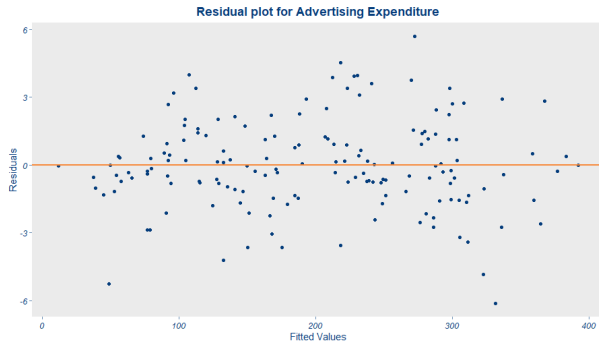
```
ggplot(lm.slr, aes(x=residuals)) +  
  geom_histogram(binwidth = 0.6, color="#003D7C",  
    fill="#EF7C00") +  
  nus_theme() +  
  labs(x="Residuals", y="Number of Cities", title =  
    "Histogram of Residuals for Advertising  
    Expenditure")
```



Residual Assumption Check on the Training Dataset

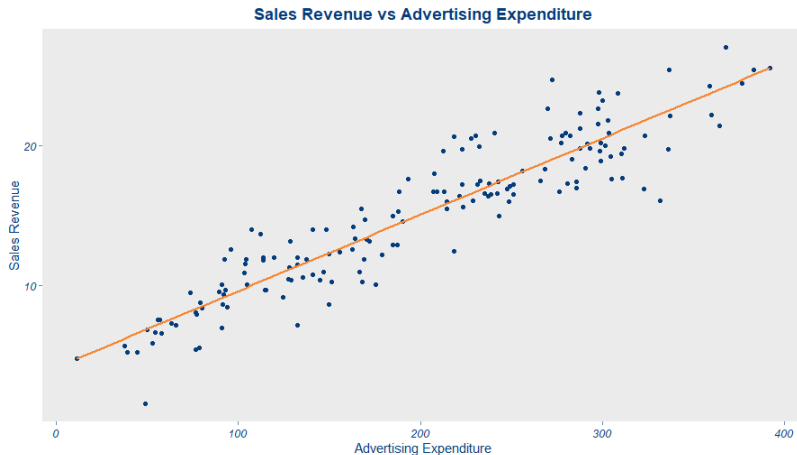
- To check assumption 4 (Residuals are evenly scattered)

```
ggplot(lm.slr, aes(x=
  Advertising.Expenditure, y=
  residuals)) +
  geom_point(color="#003D7C")
  +
  geom_abline(slope=0, color
    ="#EF7C00", size=1) +
  nus_theme() +
  labs(x="Fitted Values", y="
    Residuals", title="
    Residual plot for
    Advertising Expenditure
    ")
```



Fitting of the Best Fit Line using Least Sum of Squares Error

We use β_0 and β_1 to construct the best fit line, how do we get these coefficients' estimates?



Fitting of the Best Fit Line using Least Sum of Squares Error

cont'd

- Least Squares Method

- ▶ Chooses estimates of β_0 and β_1 that minimizes the residual sum of squares (RSS),

$$\begin{aligned}\text{RSS} &= (\text{Residual}_1)^2 + (\text{Residual}_2)^2 + (\text{Residual}_3)^2 + \dots + (\text{Residual}_n)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2\end{aligned}$$

- ▶ where (Y_i) is the actual Y and (\hat{Y}_i) is the predicted Y .

- What's next?

- ▶ The Goodness of Fit of the model.
- ▶ The statistical significance of the overall model and the statistical significance of our model coefficients, β_0 and β_1 .

Evaluating a Simple Linear Regression Model

Goodness of Fit

- Formula to calculate R^2 is

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS = Total Sum of Squares or the total variance of the variable Y which can be thought of as the amount of variability in the variable Y **before** performing the regression.
- RSS = Residual Sum of Squares which measures the variability of Y that is left unexplained **after** performing the regression.

Goodness of Fit

cont'd

- R^2 measures the proportion of the variability in Y that can be explained by X.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- R^2 value ranges from 0 to 1.
 - ▶ $R^2 = 0$: The regression model does not explain the variability in Y.
 - ▶ $R^2 = 1$: The regression model can *perfectly* explain all the variability in Y.
- As a rule of thumb, an $R^2 \geq 0.7$ is interpreted as an acceptable model.

Goodness of Fit

cont'd

- Use the `summary()` function.

```
summary(lm.slr)
```

```
Call:
lm(formula = Sales.Revenue ~ Advertising.Expenditure, data = train)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|---------|---------|---------|--------|--------|
| | -6.1235 | -1.0901 | -0.0961 | 1.1477 | 5.6893 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|------------|
| (Intercept) | 4.19731 | 0.37910 | 11.07 | <2e-16 *** |
| Advertising.Expenditure | 0.05436 | 0.00173 | 31.42 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.997 on 158 degrees of freedom

Multiple R-squared: 0.862, Adjusted R-squared: 0.8611

F-statistic: 987.1 on 1 and 158 DF, p-value: < 2.2e-16

Goodness of Fit

cont'd

- Use the `summary()` function.

```
summary(lm.slr)
```

```
Call:
lm(formula = Sales.Revenue ~ Advertising.Expenditure, data = train)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -6.1235 | -1.0901 | -0.0961 | 1.1477 | 5.6893 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|------------|
| (Intercept) | 4.19731 | 0.37910 | 11.07 | <2e-16 *** |
| Advertising.Expenditure | 0.05436 | 0.00173 | 31.42 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.997 on 158 degrees of freedom

Multiple R-squared: 0.862, Adjusted R-squared: 0.8611

F-statistic: 987.1 on 1 and 158 DF, p-value: < 2.2e-16

Hypothesis Testing on Simple Linear Regression

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

```
Call:
lm(formula = Sales.Revenue ~ Advertising.Expenditure, data = train)
```

```
Residuals:
```

| | Min | 1Q | Median | 3Q | Max |
|--|---------|---------|---------|--------|--------|
| | -6.1235 | -1.0901 | -0.0961 | 1.1477 | 5.6893 |

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|------------|
| (Intercept) | 4.19731 | 0.37910 | 11.07 | <2e-16 *** |
| Advertising.Expenditure | 0.05436 | 0.00173 | 31.42 | <2e-16 *** |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.997 on 158 degrees of freedom
```

```
Multiple R-squared:  0.862,    Adjusted R-squared:  0.8611
```

```
F-statistic: 987.1 on 1 and 158 DF,  p-value: < 2.2e-16
```


Hypothesis Testing on Simple Linear Regression

Using t-test

- $H_0 : \beta_1 = 0.$
- $H_1 : \beta_1 \neq 0.$

```
Call:
lm(formula = Sales.Revenue ~ Advertising.Expenditure, data = train)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.1235 -1.0901 -0.0961  1.1477  5.6893
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.19731    0.37910   11.07  <2e-16 ***
Advertising.Expenditure 0.05436    0.00173   31.42  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.997 on 158 degrees of freedom
Multiple R-squared:  0.862,    Adjusted R-squared:  0.8611
F-statistic: 987.1 on 1 and 158 DF,  p-value: < 2.2e-16
```

Hypothesis Testing on Simple Linear Regression

Using t-test (cont'd)

- $H_0 : \beta_0 = 0.$
- $H_1 : \beta_0 \neq 0.$

Call:

```
lm(formula = Sales.Revenue ~ Advertising.Expenditure, data = train)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -6.1235 | -1.0901 | -0.0961 | 1.1477 | 5.6893 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|------------|
| (Intercept) | 4.19731 | 0.37910 | 11.07 | <2e-16 *** |
| Advertising.Expenditure | 0.05436 | 0.00173 | 31.42 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.997 on 158 degrees of freedom

Multiple R-squared: 0.862, Adjusted R-squared: 0.8611

F-statistic: 987.1 on 1 and 158 DF, p-value: < 2.2e-16

- Since all the p-values $< \alpha$, there is sufficient evidence to reject all the Null Hypotheses at the 5% level of statistical significance.

Using the Simple Linear Regression Model to Predict the Sales Revenue

- Recall our regression equation (obtained from `summary()` function):

$$\text{Sales Revenue} = 4.197 + 0.054 \times \text{Advertising Expenditure}$$

- where $\beta_0 = 4.197$ and $\beta_1 = 0.054$.
- If the company spends \$199 on advertisement, we can estimate sales revenue to be about \$14,943.
- If the company spends \$200 on advertisement, we can estimate sales revenue to be about \$14,997.
- When advertising expenditure increases by \$1, on average, the estimate of the sales revenue increases by \$54.
- Extrapolating the values beyond our data range can be misleading as we cannot assume that the same linear trend will remain beyond the range of the X values from our training dataset.

Interpreting the Confidence Intervals of the Regression Coefficients

- Consider the regression model

$$\text{Sales Revenue} = 4.197 + 0.054 \times \text{Advertising Expenditure}$$

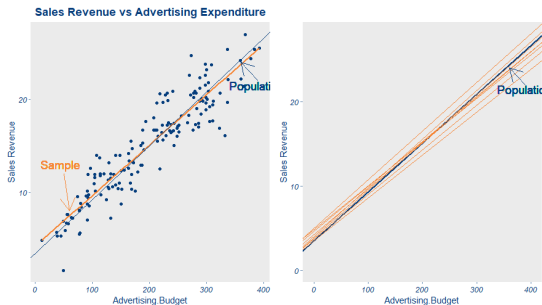
- where $\beta_0 = 4.197$ and $\beta_1 = 0.054$.
- The actual parameters might be some other values and they are unknown to us as we do not have the full population data.

Interpreting the Confidence Intervals of the Regression Coefficients

cont'd

- Suppose the actual regression model for the population data is

$$\text{Sales Revenue} = 3.45 + 0.058 \times \text{Advertising Expenditure}$$



- β_0 difference (population - sample) = $3.45 - 4.20 = -0.85$.
- β_1 difference (population - sample) = $0.058 - 0.054 = 0.004$.

The Confidence Intervals of the Regression Coefficients

cont'd

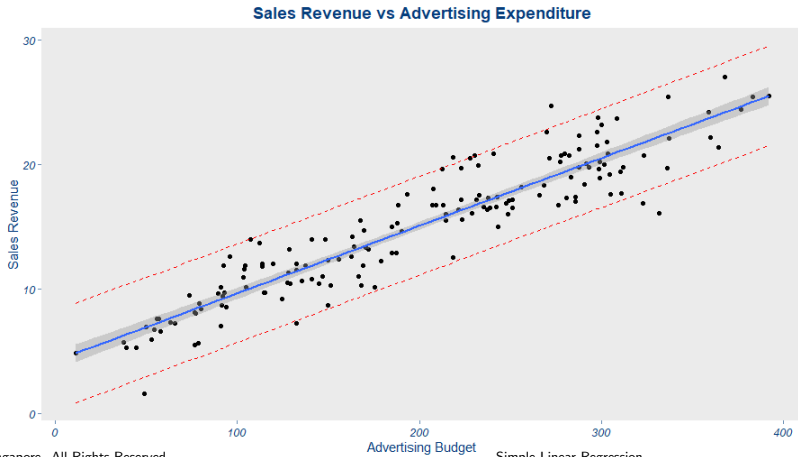
- `confint(lm.slr, level=0.95)`

| | 2.5 % | 97.5 % |
|-------------------------|------------|------------|
| (Intercept) | 3.44855426 | 4.94607211 |
| Advertising.Expenditure | 0.05094385 | 0.05777858 |

- The 95% CI for β_1 is [0.051, 0.058].
- The 95% CI for β_0 is [3.449, 4.946].
- It can be used to test the accuracy of each coefficient estimate.
- A smaller confidence interval would have a more accurate population coefficient estimate.

Confidence Interval & Prediction Interval of the Y variable

- Prediction interval predicts **the range of values on which a future individual observation would fall**.
- Confidence interval shows the likely range of values computed from samples of data associated with some statistical parameter of the data, such as the population mean.



Confidence Interval & Prediction Interval of the Y variable

cont'd

- To calculate the confidence interval for the average sales revenue for markets that spent \$200 on advertisements:

```
predict(lm.slr, data.frame(Advertising.Expenditure=200), interval="confidence")
```

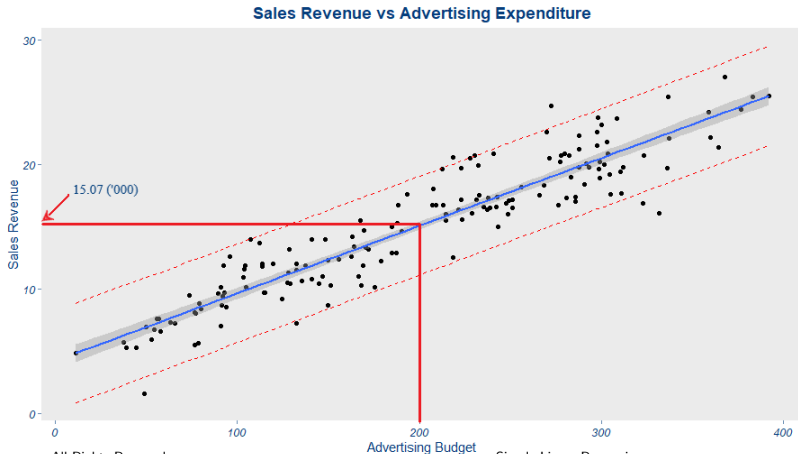
- ▶ The 95% CI for the mean sales revenue when the advertisement expenditure is \$200 is [14.758, 15.381] ('000).
- To calculate the prediction interval for the average sales revenue for markets that spent \$200 on advertisements:

```
predict(lm.slr, data.frame(Advertising.Expenditure=200), interval="prediction")
```

- ▶ The 95% PI for the sales revenue of a particular market that is spending \$200 on advertisement is [11.112, 19.027] ('000).

Confidence Interval & Prediction Interval of the Y variable

- Prediction interval predicts **the range of values on which a future individual observation would fall**.
- Confidence interval shows the likely range of values computed from samples of data associated with some statistical parameter of the data, such as the population mean.



Evaluating the Regression Model using the Test Dataset

1 Mean Squared Error & Mean Absolute Error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \text{or} \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

- Mean Squared Error (MSE) measures the average squared of the errors while the MAE measures the mean absolute error.

```
mse_data <- mean((actual - predicted)^2)
mae_data <- mean(abs(actual - predicted))
```

- A large MSE implies that, on average, the predicted value differs greatly from the actual value and the model has a low predictive power.

Evaluating the Regression Model using the Test Dataset

cont'd

2 Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

- ▶ Root Mean Squared Error (RMSE) measures the deviance of the predicted value from the best fit line.

```
rmse_data <- sqrt(mse_data)
```

- ▶ A small RMSE implies that the predicted value are close to the actual values.

Evaluating the Regression Model using the Test Dataset

cont'd

3 Mean Absolute Percentage Error

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{|Y_i|} \times 100$$

- ▶ Mean Absolute Percentage Error (MAPE) measures a simple average of absolute percentage errors.

```
mape_data <- mean(abs((actual - predicted)/actual))*100
```

- ▶ Commonly used in industries as it is easy to interpret and explain.

Summary of the Regression Model Evaluation

cont'd

Summary table:

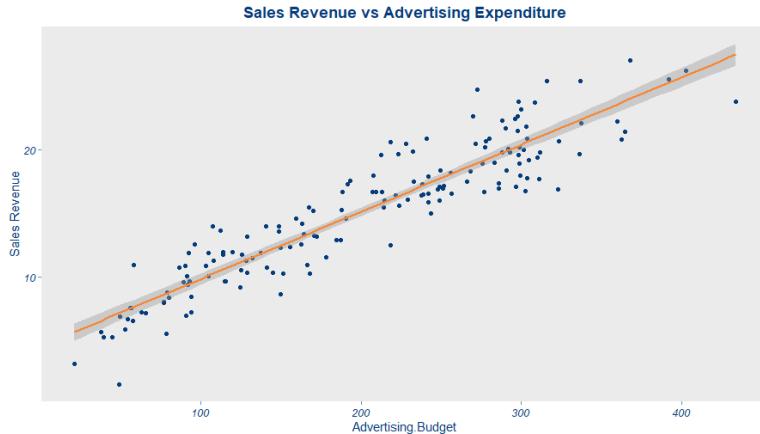
| | Training Data | Test Data |
|------|---------------|-----------|
| MSE | 3.940 | 4.457 |
| MAE | 1.596 | 1.773 |
| RMSE | 1.985 | 2.111 |
| MAPE | 12.663 | 13.721 |

- When comparing the MSE, MAE, RMSE, MAPE values of the test and training data, we **would** expect that the error for the test data to be higher.
- If our model is a good model, we **should** expect to see the errors from the training and test datasets to be very similar.

The Finalised Simple Linear Regression Model

- The DLP team now has a Simple Linear Regression model of

$$\text{Sales Revenue} = 4.197 + 0.054 \times \text{Advertising Expenditure}$$

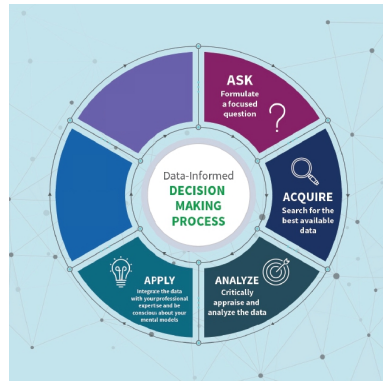


Apply – Integrating the Model with Professional Expertise

4 Apply

- The company could allocate more budget to advertising since there is a positive correlation between advertising expenditure and sales.
- The company could estimate the expected sales revenue and the prediction interval for a certain amount of advertising expenditure based on the regression model.

DIDM Framework



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

Announce – Decide and Communicate

5 Announce

- 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.

DIDM Framework



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

Assess – Monitor the Outcome

6 Assess

- 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>

Summary








Recap on Building a Simple Linear Regression model

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

Up next:

- Building and evaluating the Multiple Linear Regression model.

References

-  Zach, “4 examples of using linear regression in real life,” May 2020.
-  “Data-informed decision making framework.”
-  Tung.M.Phung, “Confidence intervals for linear regression coefficients,” Nov 2020.
-  Zach, “Confidence interval vs. prediction interval: What’s the difference?,” Aug 2021.
-  Kassambara, Suraj, David@choicemaster.org, Kassambara, Genghiskhan, and Raul, “Predict in r: Model predictions and confidence intervals,” Mar 2018.
-  A. T. Arnholt, “Machine learning with caret in r.”
-  Zach, “How to calculate mean absolute percentage error (mape) in excel,” May 2021.