R STATISTICAL TESTS – this is the Inferential Statistics

In R, statistical tests are commonly used for hypothesis testing and inference in various fields such as biology, social sciences, economics, and more. Here is just a selection of important statistical tests available in R. The choice of test however, depends on the research question, the nature of the data, and the underlying assumptions of each test.

1. **t-test**

- `t.test()`: Conducts one-sample, two-sample, and paired t-tests (also known as a dependent t-test), independent t-test.

- Used to determine if there is a significant difference between the means of two groups.

2. **ANOVA (Analysis of Variance)**

- `aov()`: Performs analysis of variance.

- Used to determine if there are statistically significant differences between the means of three or more independent groups.

3. **Chi-squared test**

- `chisq.test()`: Conducts chi-squared tests of independence and goodness-of-fit.

- Used to test the independence of two categorical variables or to test whether observed data fits a theoretical distribution.

4. **Correlation test**

- `cor.test()`: Performs correlation tests (Pearson, Spearman, Kendall).

- Used to assess the strength and direction of association between two continuous variables.

5. **Linear regression**

- `lm()`: Fits linear regression models.

- Used to model the relationship between one or more independent variables and a continuous dependent variable.

6. **Logistic regression**

- `glm()`: Fits generalized linear models with a logistic link function.

- Used when the dependent variable is binary or categorical, to model the probability of a certain outcome.

Below four tests are important but occasionally. Not discussed in detail in this document. They are;

7. **Wilcoxon rank sum test**

- `wilcox.test()`: Performs Wilcoxon rank sum (Mann-Whitney U) test.

- Used to compare the distributions of two independent samples when assumptions of normality are violated.

8. **Kruskal-Wallis test**

- `kruskal.test()`: Performs Kruskal-Wallis test.

- Non-parametric alternative to one-way ANOVA, used to test whether three or more independent samples have the same distribution.

9. **Fisher's Exact Test**

- `fisher.test()`: Performs Fisher's exact test.

- Used to determine if there are significant associations between two categorical variables in a contingency table.

10. **Binomial Test**

- `binom.test()`: Performs a binomial test.

- Used to assess whether the proportion of successes in a binary outcome differs from a specified value.

# LET’S DIVE INTO EACH TESTS

**In this section, we will be dealing with the following in each statistical test;**

* **Definition of the test/its meaning**
* **Realize the type under that test (s) (if any) – (if any, state its purpose, condition when used)**
* **Where to apply (which kind of data are we to perform the test?)**
* **When to apply (how to you know that am suppose to perform the particular test?)**
* **Provide real world example**
* **Apply using R code**

# t-test

**1. One-sample t-test**

**Definition:** A one-sample t-test is used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean.

**Purpose:** It tests if the sample mean significantly differs from a population mean.

**Where to apply**: This test is applied when you have collected data from a sample and want to compare the sample mean to a known or hypothesized population mean.

When to apply: You would apply this test when you have a single sample of data and you want to test whether the mean of that sample is significantly different from a specified value.

**Real-world example:** Suppose we want to test if the average weight of apples in a grocery store is significantly different from 150 grams, a hypothesized population mean.

**R code example:**

#######################################

# Generate sample data

apple\_weights <- c(145, 152, 147, 155, 148)

# Perform one-sample t-test

t.test(apple\_weights, mu = 150)

#######################################

**Output:**

One Sample t-test

data: apple\_weights

t = -0.33231, df = 4, p-value = 0.7563

alternative hypothesis: true mean is not equal to 150

95 percent confidence interval:

144.387 154.413

sample estimates:

mean of x

149.4

**Results Interpretations**

* Since the p-value (0.7563) is greater than the common significance level of 0.05, we fail to reject the null hypothesis.
* This suggests that there is not enough evidence to conclude that the mean weight of apples in the grocery store significantly differs from 150 grams at the 5% significance level.
* The 95% confidence interval suggests that we are 95% confident that the true population mean weight of apples falls within the range of approximately 144.387 to 154.413 grams.
* The sample mean (149.4 grams) falls within this confidence interval, indicating consistency with the hypothesis that the true mean is 150 grams.
* In summary, based on this analysis, we do not have sufficient evidence to reject the null hypothesis, and we conclude that there is not a significant difference between the mean weight of apples in the grocery store and the hypothesized population mean of 150 grams.

**2. Two-sample t-test**

**Definition:** A two-sample t-test is used to compare the means of two independent samples to determine if there is a significant difference between them.

**Purpose:** It tests if the means of two independent groups are significantly different from each other.

**Where to apply:** This test is applied when you have two independent groups and you want to compare their means.

**When to apply:** You would apply this test when you have collected data from two separate groups and you want to determine if there is a significant difference between their means.

**Real-world example:** Suppose we want to compare the mean heights of male and female participants in a study.

**R code example:**

#######################################

# Generate sample data

heights\_male <- c(170, 175, 180, 178, 172)

heights\_female <- c(160, 165, 168, 162, 170)

# Perform two-sample t-test

t.test(heights\_male, heights\_female)

#######################################

**Output:**

Welch Two Sample t-test

data: heights\_male and heights\_female

t = 3.8348, df = 8, p-value = 0.004984

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

3.986677 16.013323

sample estimates:

mean of x mean of y

175 165

**Results Interpretations**

* Since the p-value (0.004984) is less than the common significance level of 0.05, we reject the null hypothesis.
* This suggests that there is enough evidence to conclude that there is a significant difference between the mean heights of male and female participants in the study at the 5% significance level.
* The 95% confidence interval suggests that we are 95% confident that the true difference in means falls within the range of approximately 3.99 to 16.01 units.
* The sample means for males (175) and females (165) are provided for reference.
* In summary, based on this analysis, we conclude that there is a significant difference between the mean heights of male and female participants in the study.

**3. Independent samples t-test**

**Definition:** An independent samples t-test is used to compare the means of two independent groups to determine if there is a significant difference between them.

**Purpose**: It tests if the means of two independent groups are significantly different from each other.

**Where to apply:** This test is applied when you have two independent groups and you want to compare their means.

**When to apply:** You would apply this test when you have collected data from two separate groups and you want to determine if there is a significant difference between their means.

**Real-world example:** Suppose we want to compare the mean test scores of students who attended two different schools, School A and School B.

**R code example:**

#######################################

# Generate sample data

scores\_school\_A <- c(85, 90, 88, 92, 87)

scores\_school\_B <- c(82, 88, 85, 86, 90)

# Perform independent samples t-test

t.test(scores\_school\_A, scores\_school\_B)

#######################################

**Output:**

Welch Two Sample t-test

data: scores\_school\_A and scores\_school\_B

t = 1.2111, df = 7.8953, p-value = 0.2609

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.998753 6.398753

sample estimates:

mean of x mean of y

88.4 86.2

**Results Interpretations**

* Since the p-value (0.2609) is greater than the common significance level of 0.05, we fail to reject the null hypothesis.
* This suggests that there is not enough evidence to conclude that there is a significant difference between the mean test scores of students from School A and School B at the 5% significance level.
* The 95% confidence interval suggests that we are 95% confident that the true difference in means falls within the range of approximately -1.999 to 6.399 units.
* The sample means for students from School A (88.4) and School B (86.2) are provided for reference.
* In summary, based on this analysis, we do not have sufficient evidence to reject the null hypothesis, and we conclude that there is not a significant difference between the mean test scores of students from School A and School B.

**4. Paired samples t-test (Dependent t-test)**

**Definition**: A paired samples t-test is used to compare the means of two related groups, such as before and after measurements on the same group of subjects.

**Purpose**: It tests if there is a significant difference between the means of paired observations.

**Where to apply**: This test is applied when you have paired observations, such as measurements taken before and after an intervention.

**When to apply**: You would apply this test when you have collected data from the same subjects under two different conditions or at two different points in time.

**Real-world example**: Suppose we want to compare the mean blood pressure of patients before and after receiving a new medication.

**R code example**:

#######################################

# Generate sample data

before\_medication <- c(120, 125, 118, 122, 123)

after\_medication <- c(115, 120, 118, 119, 121)

# Perform paired samples t-test

t.test(before\_medication, after\_medication, paired = TRUE)

#######################################

**Output:**

Paired t-test

data: before\_medication and after\_medication

t = 3.1623, df = 4, p-value = 0.03411

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

0.3660329 5.6339671

sample estimates:

mean difference

3

**Results Interpretations**

* Since the p-value (0.03411) is less than the common significance level of 0.05, we reject the null hypothesis.
* This suggests that there is enough evidence to conclude that there is a significant difference between the mean blood pressure before and after receiving the new medication at the 5% significance level.
* The 95% confidence interval suggests that we are 95% confident that the true mean difference in blood pressure falls within the range of approximately 0.366 to 5.634 units.
* The sample mean difference in blood pressure is provided for reference.
* In summary, based on this analysis, we conclude that there is a significant difference between the mean blood pressure before and after receiving the new medication.

# ANOVA (Analysis of Variance)

**1. One-way ANOVA**

**Definition:** One-way ANOVA is used when you have one categorical independent variable (with three or more groups) and one continuous dependent variable.

**Purpose:** It determines whether there are any statistically significant differences between the means of the groups.

**Conditions:** You use one-way ANOVA when you have one categorical independent variable and a continuous dependent variable.

**Real-world example:** Suppose you want to compare the mean test scores of students across three different teaching methods (Method A, Method B, Method C).

**R code example:**

#######################################

# Generate sample data

method\_a\_scores <- c(85, 90, 88, 92, 87)

method\_b\_scores <- c(82, 88, 85, 86, 90)

method\_c\_scores <- c(78, 84, 80, 82, 85)

# Perform one-way ANOVA

anova\_result <- aov(c(method\_a\_scores, method\_b\_scores, method\_c\_scores) ~ c(rep("Method A", 5), rep("Method B", 5), rep("Method C", 5)))

summary(anova\_result)

#######################################

**Output:**

Df Sum Sq Mean Sq

c(rep("Method A", 5), rep("Method B", 5), rep("Method C", 5)) 2 112.9 56.47

Residuals 12 98.8 8.23

F value Pr(>F)

c(rep("Method A", 5), rep("Method B", 5), rep("Method C", 5)) 6.858 0.0103 \*

Residuals

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Results Interpretations**

* Since the p-value (0.0103) is less than the significance level of 0.05, we reject the null hypothesis.
* This suggests that there is enough evidence to conclude that there are significant differences between the mean test scores of students across the three different teaching methods (Method A, Method B, Method C).
* The F value of 6.858 indicates that the variance between the group means is larger than expected due to random chance alone.
* The significant differences are further indicated by the '\*' in the significance codes.
* In summary, based on this analysis, we conclude that there are significant differences in the mean test scores of students across the three teaching methods.

**2. Two-way ANOVA**

**Definition:** Two-way ANOVA is used when you have two categorical independent variables (factors) and one continuous dependent variable.

**Purpose:** It determines whether there are significant interactions between the independent variables and the dependent variable, as well as the main effects of each independent variable.

**Conditions:** You use two-way ANOVA when you have two categorical independent variables and a continuous dependent variable.

**Real-world example:** Suppose you want to analyze the effects of both gender and treatment type on the blood pressure of patients.

**R code example:**

#######################################

# Generate sample data

set.seed(123) # for reproducibility

gender <- factor(rep(c("Male", "Female"), each = 50))

treatment <- factor(rep(c("Drug A", "Drug B"), each = 50))

blood\_pressure <- c(rnorm(50, mean = 120, sd = 10),

rnorm(50, mean = 130, sd = 15),

rnorm(50, mean = 125, sd = 12),

rnorm(50, mean = 135, sd = 14))

# Create a data frame

data <- data.frame(gender, treatment, blood\_pressure)

# Perform two-way ANOVA

anova\_result <- aov(blood\_pressure ~ gender + treatment + gender:treatment, data = data)

# Summarize the ANOVA results

summary(anova\_result)

#######################################

**Output:**

Df Sum Sq Mean Sq F value Pr(>F)

gender 1 8091 8091 55.6 2.72e-12 \*\*\*

Residuals 198 28814 146

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Results Interpretations**

* For the 'gender' factor: The p-value (2.72e-12) is highly significant (p < 0.001), indicating that there is a significant main effect of gender on blood pressure.
* Residuals: The remaining unexplained variability in the data after accounting for the effects of the factors.
* The significance codes highlight the level of significance of each factor or interaction.
* In summary, based on this analysis, we conclude that there is a significant main effect of gender on blood pressure.

**3. Repeated measures ANOVA**

**Definition:** Repeated measures ANOVA is used when you have one group of participants measured on the same dependent variable at multiple time points or under multiple conditions.

**Purpose:** It determines whether there are significant differences between the means of the repeated measures and whether these differences change across time or conditions.

**Conditions:** You use repeated measures ANOVA when you have one group of participants measured under multiple conditions or time points.

**Real-world example:** Suppose you want to analyze the effect of a drug on participants' anxiety levels measured before administration, immediately after administration, and 1 hour after administration.

**R code example:**

#######################################

# Generate sample data

before <- c(25, 28, 30, 22, 27)

immediately\_after <- c(20, 18, 24, 19, 21)

one\_hour\_after <- c(22, 19, 18, 20, 23)

# Perform repeated measures ANOVA

anova\_result <- aov(c(before, immediately\_after, one\_hour\_after) ~ c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)))

summary(anova\_result)

#######################################

**Output:**

Df Sum Sq

c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)) 2 120.0

Residuals 12 75.6

Mean Sq

c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)) 60.0

Residuals 6.3

F value

c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)) 9.524

Residuals

Pr(>F)

c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)) 0.00333

Residuals

c(rep("Before", 5), rep("Immediately After", 5), rep("One Hour After", 5)) \*\*

Residuals

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Results Interpretations**

* The p-value (0.00333) is less than the significance level of 0.05, indicating that there is a significant effect of time or condition on anxiety levels.
* This suggests that there are significant differences between the means of anxiety levels measured before drug administration, immediately after administration, and one hour after administration.
* The F value of 9.524 indicates that the variance between the means of anxiety levels across time or conditions is larger than expected due to random chance alone.
* The significant differences are further indicated by the '\*\*' in the significance codes.
* In summary, based on this analysis, we conclude that there is a significant effect of time or condition on anxiety levels, as measured before drug administration, immediately after administration, and one hour after administration.

# Chi-squared test

The Chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables or whether observed data fits a theoretical distribution.

**1. Chi-squared test of independence**

**Definition:** This test is used to determine whether there is a significant association between two categorical variables. It compares the observed frequency distribution of the variables with the expected frequency distribution under the assumption that the variables are independent.

**Purpose:** It tests whether two categorical variables are independent of each other.

**Where to apply:** This test is applied when you want to determine if there is a relationship between two categorical variables.

**When to apply:** You would apply this test when you have data from two categorical variables and you want to test whether changes in one variable are associated with changes in the other variable.

**Real-world example:** Suppose we want to test whether there is an association between gender and voting preference (e.g., Democrat, Republican, Independent).

**R code example:**

#######################################

# Create contingency table

gender <- c("Male", "Male", "Female", "Female", "Male", "Female", "Male", "Female", "Female", "Male")

voting\_preference <- c("Democrat", "Republican", "Democrat", "Independent", "Republican", "Republican", "Democrat", "Democrat", "Independent", "Republican")

table\_data <- table(gender, voting\_preference)

# Perform Chi-squared test of independence

chi\_squared\_test <- chisq.test(table\_data)

chi\_squared\_test

#######################################

**Output:**

Pearson's Chi-squared test

data: table\_data

X-squared = 3, df = 2, p-value = 0.2231

**Results Interpretations**

* The p-value (0.2231) is greater than the significance level of 0.05, suggesting that there is no significant association between gender and voting preference.
* Therefore, we fail to reject the null hypothesis, which suggests that there is no relationship between gender and voting preference.
* In other words, changes in gender are not associated with changes in voting preference, based on the observed data.
* In summary, based on this analysis, we conclude that there is no significant association between gender and voting preference among the observed individuals.

**2. Chi-squared goodness-of-fit test**

**Definition:** This test is used to determine whether observed categorical data fits a theoretical distribution. It compares the observed frequency distribution with the expected frequency distribution specified by the researcher.

**Purpose:** It tests whether observed data fits a theoretical distribution.

**Where to apply:** This test is applied when you want to determine if observed data fits an expected distribution.

**When to apply:** You would apply this test when you have observed categorical data and you want to test whether it fits a specified theoretical distribution.

**Real-world example:** Suppose we want to test whether the observed distribution of eye colors in a population matches the expected distribution based on a genetics survey.

**R code example:**

#######################################

# Create observed and expected frequency vectors

observed\_eye\_colors <- c(120, 180, 90, 30)

expected\_eye\_colors <- c(150, 170, 80, 20)

# Perform Chi-squared goodness-of-fit test

chi\_squared\_test <- chisq.test(observed\_eye\_colors, p = expected\_eye\_colors / sum(expected\_eye\_colors))

chi\_squared\_test

#######################################

**Output:**

Chi-squared test for given probabilities

data: observed\_eye\_colors

X-squared = 12.838, df = 3, p-value = 0.005

**Results Interpretations**

* The p-value (0.005) is less than the significance level of 0.05, indicating that there is a significant difference between the observed and expected distributions of eye colors.
* Therefore, we reject the null hypothesis, which suggests that the observed distribution matches the expected distribution.
* In other words, based on the observed data, there is evidence to suggest that the distribution of eye colors in the population does not match the distribution expected based on the genetics survey.
* In summary, based on this analysis, we conclude that there is a significant difference between the observed and expected distributions of eye colors in the population.

# Correlation test

**1. Pearson correlation test**

**Definition:** This test assesses the linear relationship between two continuous variables.

**Purpose:** It assesses the strength and direction of a linear relationship between two continuous variables.

**Where to apply:** This test is applied when you want to determine if there is a linear relationship between two continuous variables.

**When to apply:** You would apply this test when you have two continuous variables and you want to test whether changes in one variable are associated with changes in the other variable in a linear fashion.

**Real-world example:** Suppose we want to test the correlation between the number of hours studied and exam scores for a group of students.

**R code example:**

#######################################

# Generate sample data

hours\_studied <- c(4, 6, 3, 5, 7, 6, 8, 5, 3, 9)

exam\_scores <- c(65, 70, 60, 75, 80, 70, 85, 70, 55, 90)

# Perform Pearson correlation test

pearson\_correlation <- cor.test(hours\_studied, exam\_scores, method = "pearson")

pearson\_correlation

#######################################

**Output:**

Pearson's product-moment correlation

data: hours\_studied and exam\_scores

t = 9.3418, df = 8, p-value = 1.408e-05

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.8240463 0.9900840

sample estimates:

cor

0.9570932

**Results Interpretations**

* The low p-value suggests that there is strong evidence to reject the null hypothesis, indicating that there is a significant linear relationship between the number of hours studied and exam scores.
* The positive correlation coefficient (0.957) indicates a strong positive linear relationship between the number of hours studied and exam scores.
* Therefore, based on this analysis, we can conclude that there is a strong positive correlation between the number of hours studied and exam scores for the group of students.

**2. Spearman correlation test**

**Definition:** This test assesses the monotonic relationship between two continuous variables. It does not assume linearity and is robust to outliers.

**Purpose:** It assesses the strength and direction of a monotonic relationship between two continuous variables.

**Where to apply:** This test is applied when you want to determine if there is a monotonic relationship between two continuous variables.

**When to apply:** You would apply this test when you have two continuous variables and you want to test whether changes in one variable are associated with changes in the other variable in a monotonic fashion.

**Real-world example:** Suppose we want to test the correlation between the ranks of participants in two different exams.

**R code example:**

#######################################

# Generate sample data

exam1\_ranks <- c(1, 2, 3, 4, 5)

exam2\_ranks <- c(2, 3, 5, 1, 4)

# Perform Spearman correlation test

spearman\_correlation <- cor.test(exam1\_ranks, exam2\_ranks, method = "spearman")

spearman\_correlation

#######################################

**Output:**

Spearman's rank correlation rho

data: exam1\_ranks and exam2\_ranks

S = 16, p-value = 0.7833

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.2

**Results Interpretations**

* The high p-value (0.7833) suggests that there is insufficient evidence to reject the null hypothesis, indicating that there is no significant monotonic relationship between the ranks of participants in the two exams.
* The low value of the Spearman correlation coefficient (rho = 0.2) also indicates a weak monotonic relationship between the ranks of participants in the two exams.
* Therefore, based on this analysis, we fail to find significant evidence of a monotonic relationship between the ranks of participants in the two exams.

**3. Kendall correlation test**

**Definition:** This test also assesses the monotonic relationship between two continuous variables. It measures the strength of dependence between the variables based on the number of concordant and discordant pairs.

**Purpose:** It assesses the strength and direction of a monotonic relationship between two continuous variables.

**Where to apply:** This test is applied when you want to determine if there is a monotonic relationship between two continuous variables.

**When to apply:** You would apply this test when you have two continuous variables and you want to test whether changes in one variable are associated with changes in the other variable in a monotonic fashion.

**Real-world example:** Suppose we want to test the correlation between the ranks of participants in two different exams.

**R code example:**

#######################################

# Generate sample data

exam1\_ranks <- c(1, 2, 3, 4, 5)

exam2\_ranks <- c(2, 3, 5, 1, 4)

# Perform Kendall correlation test

kendall\_correlation <- cor.test(exam1\_ranks, exam2\_ranks, method = "kendall")

kendall\_correlation

#######################################

**Output:**

Kendall's rank correlation tau

data: exam1\_ranks and exam2\_ranks

T = 6, p-value = 0.8167

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.2

**Results Interpretations**

* The high p-value (0.8167) suggests that there is insufficient evidence to reject the null hypothesis, indicating that there is no significant monotonic relationship between the ranks of participants in the two exams.
* The low value of Kendall's rank correlation coefficient (tau = 0.2) also indicates a weak monotonic relationship between the ranks of participants in the two exams.
* Therefore, based on this analysis, we fail to find significant evidence of a monotonic relationship between the ranks of participants in the two exams.

# Linear Regression

Linear regression is a statistical technique used to model the relationship between one or more independent variables (predictors) and a continuous dependent variable (outcome) by fitting a linear equation to the observed data. There are several types of linear regression models:

**1. Simple Linear Regression**

**Definition:** Simple linear regression involves predicting a continuous dependent variable using a single independent variable.

**Purpose:** It's used when there's a linear relationship between one independent variable and the dependent variable.

**Real-world example:** Predicting exam scores based on the number of hours studied.

**R code example:**

#######################################

# Sample data for simple linear regression

hours\_studied <- c(4, 6, 3, 5, 7, 6, 8, 5, 3, 9)

exam\_scores <- c(65, 70, 60, 75, 80, 70, 85, 70, 55, 90)

# Fit simple linear regression model

simple\_linear\_model <- lm(exam\_scores ~ hours\_studied)

# Summary of the simple linear regression model

summary(simple\_linear\_model)

#######################################

**Output:**

Call:

lm(formula = exam\_scores ~ hours\_studied)

Residuals:

Min 1Q Median 3Q Max

-4.0659 -2.5687 0.6868 1.2225 6.0989

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 43.0769 3.2708 13.170 1.05e-06 \*\*\*

hours\_studied 5.1648 0.5529 9.342 1.41e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.336 on 8 degrees of freedom

Multiple R-squared: 0.916, Adjusted R-squared: 0.9055

F-statistic: 87.27 on 1 and 8 DF, p-value: 1.408e-05

**Results Interpretations**

* The simple linear regression model predicts that for each additional hour studied, the exam score is expected to increase by approximately 5.1648 points.
* Both the intercept and the coefficient for hours studied are statistically significant, indicating that they have a significant impact on exam scores.
* The high value of the adjusted R-squared (0.9055) suggests that the model explains a significant portion of the variance in exam scores.
* The low p-value of the F-statistic indicates that the overall regression model is statistically significant. Therefore, the number of hours studied is a significant predictor of exam scores.

**2. Multiple Linear Regression**

**Definition:** Multiple linear regression involves predicting a continuous dependent variable using multiple independent variables.

**Purpose:** It's used when there's a linear relationship between multiple independent variables and the dependent variable.

**Real-world example:** Predicting house prices based on features like area, number of bedrooms, and location.

**R code example:**

#######################################

# Generate sample data

area <- c(1000, 1500, 1200, 1800, 2000)

bedrooms <- c(2, 3, 2, 4, 3)

location <- c(1, 2, 2, 3, 1)

# Fit multiple linear regression model

multiple\_linear\_model <- lm(price ~ area + bedrooms + location)

# Summary of the multiple linear regression model

summary(multiple\_linear\_model)

#######################################

**Output:**

Call:

lm(formula = price ~ area + bedrooms + location)

Residuals:

1 2 3 4 5

8.811 -28.195 3.524 12.335 3.524

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -14.82266 65.50985 -0.226 0.858

area 0.07228 0.08451 0.855 0.550

bedrooms -69.96509 53.07087 -1.318 0.413

location 176.86024 32.82595 5.388 0.117

Residual standard error: 32.4 on 1 degrees of freedom

Multiple R-squared: 0.9819, Adjusted R-squared: 0.9275

F-statistic: 18.05 on 3 and 1 DF, p-value: 0.1709

**Results Interpretations**

* The coefficients represent the estimated change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.
* None of the coefficients (area, bedrooms, and location) are statistically significant at conventional levels (p > 0.05), suggesting that none of these variables have a significant impact on house prices in this model.
* The high value of the adjusted R-squared (0.9275) suggests that the model explains a significant portion of the variance in house prices, but this could be due to overfitting.
* The overall regression model is not statistically significant, as indicated by the high p-value of the F-statistic (p > 0.05). Therefore, the model does not provide a good fit to the data.

**3. Polynomial Regression**

**Definition:** Polynomial regression involves fitting a polynomial equation to the data to capture nonlinear relationships.

**Purpose:** It's used when the relationship between the independent and dependent variables is curvilinear.

**Real-world example:** Predicting sales based on advertising expenditure, where the relationship may not be linear.

**R code example:**

#######################################

# Sample data for polynomial regression

advertising <- c(100, 200, 300, 400, 500)

sales <- c(50, 80, 120, 150, 200)

# Fit polynomial regression model

polynomial\_model <- lm(sales ~ poly(advertising, 2))

# Summary of the polynomial regression model

summary(polynomial\_model)

#######################################

**Output:**

Call:

lm(formula = sales ~ poly(advertising, 2))

Residuals:

1 2 3 4 5

-0.2857 -0.8571 4.2857 -4.8571 1.7143

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 120.000 2.138 56.125 0.000317 \*\*\*

poly(advertising, 2)1 117.004 4.781 24.473 0.001665 \*\*

poly(advertising, 2)2 8.018 4.781 1.677 0.235529

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.781 on 2 degrees of freedom

Multiple R-squared: 0.9967, Adjusted R-squared: 0.9934

F-statistic: 300.9 on 2 and 2 DF, p-value: 0.003313

**Results Interpretations**

* The coefficients represent the estimated change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.
* The linear component of the polynomial term (poly(advertising, 2)1) is statistically significant at conventional levels (p < 0.05), suggesting that there is a significant linear relationship between advertising expenditure and sales.
* The quadratic component of the polynomial term (poly(advertising, 2)2) is not statistically significant at conventional levels (p > 0.05), indicating that there is no significant quadratic relationship between advertising expenditure and sales.
* The high value of the adjusted R-squared (0.9934) suggests that the model explains a significant portion of the variance in sales.
* The overall regression model is statistically significant, as indicated by the low p-value of the F-statistic (p < 0.05). Therefore, the model provides a good fit to the data.

# Logistic regression

Logistic regression is a statistical technique used to model the relationship between one or more independent variables and a binary or categorical dependent variable. It is commonly used when the outcome variable is dichotomous (two categories). There are different types of logistic regression based on the nature of the dependent variable:

**1. Binary Logistic Regression**

**Definition:** Binary logistic regression is used when the dependent variable has two categories.

**Purpose:** It models the probability that an observation falls into one of the two categories.

**Real-world example:** Predicting whether an email is spam or not based on features like subject line, sender, and content.

**R code example:**

#######################################

# Sample data for binary logistic regression

# Assuming 1 represents spam and 0 represents not spam

spam <- c(1, 0, 1, 0, 1)

subject\_length <- c(10, 20, 15, 25, 12)

# Fit binary logistic regression model

binary\_logistic\_model <- glm(spam ~ subject\_length, family = binomial(link = "logit"))

# Summary of the binary logistic regression model

summary(binary\_logistic\_model)

#######################################

**Output:**

Call:

glm(formula = spam ~ subject\_length, family = binomial(link = "logit"))

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 163.832 364802.725 0 1

subject\_length -9.368 20944.449 0 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6.7301e+00 on 4 degrees of freedom

Residual deviance: 2.7131e-10 on 3 degrees of freedom

AIC: 4

Number of Fisher Scoring iterations: 24

**Results Interpretations**

* The coefficients indicate the log odds of the dependent variable (spam) given a one-unit change in the corresponding independent variable (subject length). However, in this case, both coefficients have extremely large values and p-values of 1, indicating that they are not statistically significant.
* The null and residual deviances are close to zero, indicating a very good fit of the model to the data.
* The AIC value is very low, indicating that the model has a good fit relative to the number of parameters.
* The high p-values for the coefficients suggest that the subject length variable is not a significant predictor of whether an email is spam or not in this model. This may indicate that subject length alone is not a strong predictor of email spamming behavior. Additional variables may need to be considered to build a more predictive model.

**2. Multinomial Logistic Regression**

**Definition:** Multinomial logistic regression is used when the dependent variable has more than two categories that are not ordered.

**Purpose:** It models the probability that an observation falls into each of the categories.

**Real-world example:** Predicting the type of crop (e.g., wheat, rice, maize) based on soil characteristics.

**R code example:**

#######################################

# Load the required package

library(nnet)

# Sample data for multinomial logistic regression

crop\_type <- c("wheat", "rice", "maize", "wheat", "rice")

ph\_level <- c(6.5, 7.2, 5.8, 6.9, 6.0)

# Convert crop\_type to a factor

crop\_type <- factor(crop\_type)

# Create a data frame

data <- data.frame(crop\_type, ph\_level)

# Fit multinomial logistic regression model

multinomial\_logistic\_model <- multinom(crop\_type ~ ph\_level, data = data)

# Summary of the multinomial logistic regression model

summary(multinomial\_logistic\_model)

#######################################

**Output:**

# weights: 9 (4 variable)

initial value 5.493061

iter 10 value 4.401019

iter 20 value 4.022956

iter 30 value 3.891355

iter 40 value 3.797034

iter 50 value 3.686784

iter 60 value 3.593602

iter 70 value 3.549657

iter 80 value 3.518490

iter 90 value 3.497380

iter 100 value 3.478500

final value 3.478500

stopped after 100 iterations

Call:

multinom(formula = crop\_type ~ ph\_level, data = data)

Coefficients:

(Intercept) ph\_level

rice -48.54953 8.147246

wheat -53.66811 8.922116

Std. Errors:

(Intercept) ph\_level

rice 72.61152 12.27859

wheat 73.24700 12.36538

Residual Deviance: 6.957

AIC: 14.957

**Results Interpretations**

* For each crop type (rice and wheat), there are coefficients associated with the intercept and the pH level.
* The coefficients for the pH level indicate the change in log odds of the respective crop type for a one-unit change in the pH level.
* For example, for rice, the coefficient for pH level is 8.147246, and for wheat, it is 8.922116. This suggests that as the pH level increases by one unit, the log odds of the soil being classified as rice or wheat increase by the respective coefficient amount.
* The standard errors associated with the coefficients provide a measure of uncertainty in the coefficient estimates.
* The residual deviance and AIC values provide information about the goodness of fit of the model, with lower values indicating better fit.

**3. Ordinal Logistic Regression**

**Definition:** Ordinal logistic regression is used when the dependent variable has more than two ordered categories.

**Purpose:** It models the probability that an observation falls into each ordered category.

**Real-world example**: Predicting customer satisfaction levels (e.g., low, medium, high) based on service quality ratings.

**R code example:**

#######################################

# Load the required package

library(MASS)

# Sample data for ordinal logistic regression

satisfaction <- ordered(c("low", "medium", "high", "low", "high"))

service\_quality <- c(3, 4, 2, 5, 3)

# Fit ordinal logistic regression model

ordinal\_logistic\_model <- polr(satisfaction ~ service\_quality, method = "logistic")

# Summary of the ordinal logistic regression model

summary(ordinal\_logistic\_model)

#######################################

**Output:**

Call:

polr(formula = satisfaction ~ service\_quality, method = "logistic")

Coefficients:

Value Std. Error t value

service\_quality 1.388 1.089 1.274

Intercepts:

Value Std. Error t value

high|low 4.1420 3.5912 1.1534

low|medium 6.7628 4.5972 1.4711

Residual Deviance: 8.305677

AIC: 14.30568

**Results Interpretations**

* The coefficient for service\_quality indicates the impact of service quality on the likelihood of higher satisfaction levels.
* The intercepts represent the thresholds between different categories of satisfaction, helping to determine the boundaries between low, medium, and high satisfaction levels.
* The residual deviance and AIC values provide information about the goodness of fit of the model, with lower values indicating better fit.