**Day\_3\_Session\_01**

**Module VI: Regression Analysis**

### Simple Linear regression and diagnostic checks

In this example, we'll use the "hprice1" dataset and create a simple linear regression model to predict housing prices based on the number of bedrooms.

**## Before running Regression Analysis**

## Import house price data set

hprice <- read.csv("hprice.csv", header = T, stringsAsFactors = F)

glimpse(hprice1)

missing\_glimpse(hprice1)

ff\_glimpse(hprice1)

head(hprice1)

names(hprice)

## Descriptive analysis

hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

describe(fast = T)

## Correlation table

hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

correlate() %>% shave() %>%

fashion()

## Correlation plot

hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

correlate() %>% shave() %>%

rplot(colours = c("red","green"))

## Correlation table with p values

hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

corr.test()

## Correlation heat map

cor\_mat <- hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

cor()

# Create a correlation heatmap

corrplot(cor\_mat,

method = "color", # Use color to represent correlations

type = "full", # Display the lower triangle of the matrix

tl.cex = 0.7, # Adjust text size for variable labels

tl.col = "black" # Set variable label color to black

)

hprice %>%

select(price,assess,bdrms,lotsize,sqrft,colonial) %>%

ggpairs()

### Variable specific statistics

##preliminary tests to make sure that the test assumptions are met

##Violin plot Using lessR package

Plot(price, data = hprice)

Plot(price, data = hprice, by1 = colonial)

###density plot

ggplot(hprice, aes(x = price)) +

geom\_density(fill = "chartreuse2", alpha = 0.3) +

labs(x = "Price",y = "Density")+

theme\_minimal()

**## Simple Regression Model**

Import hprice data set

# Fit a simple linear regression model

lm\_model <- lm(price ~ bdrms, data = hprice)

# Summarize the regression results

summary(lm\_model)

**## Diagnostic analysis: Method I**

# Method I

library(mosaic)

library(lattice)

qqmath(~resid(lm\_model))

qqmath(~resid(lm\_model), type = c("p","r"))

plot(lm\_model , which = c(1,2))

#Create density plot of residuals

plot(density(resid(lm\_model)), col="blue", lwd=3)

shapiro.test(resid(lm\_model))

# To test whether the error terms for each group have similar variance (homogeneity of variance),

plot(lm\_model$residuals~lm\_model$fitted.values, col="red")

lines(lowess(lm\_model$fitted.values,lm\_model$residuals), col="blue", lwd=2)

**## Diagnostic analysis: Method II**

# Method II

res <- resid(lm\_model) # get list of residuals

# produce a residual vs fitted plot for visulaizting heteroscedasticity

#produce residual vs. fitted plot

plot(fitted(lm\_model), res)

#add a horizontal line at 0

abline(0,0)

#create Q-Q plot for residuals

qqnorm(res)

#add a straight diagonal line to the plot

qqline(res)

#Create density plot of residuals

plot(density(res))

**## Diagnostic analysis: Method 3**

## Diagnostic analysis

# Load necessary libraries

library(ggplot2)

library(broom)

# Create a scatter plot with the regression line

ggplot(data = hprice1, aes(x = bdrms, y = price)) +

geom\_point() + # Add points

geom\_smooth(method = "lm", # Add linear regression line

formula = y ~ x, # Customize the formula if needed

color = "blue", # Line color

se = FALSE) + # Don't show confidence intervals

labs(

title = "Scatter Plot with Regression Line",

x = "Bed Rooms",

y = "House Price"

)

# Create a data frame with model residuals and fitted values

lm\_data <- augment(lm\_model)

# Create a scatter plot of observed vs. predicted values with the regression line

ggplot(lm\_data, aes(x = .fitted, y = .resid)) +

geom\_point() +

geom\_hline(yintercept = 0, color = "red", linetype = "dashed") +

labs(

title = "Residual Plot",

x = "Fitted Values",

y = "Residuals"

)

# Create a histogram of residuals

ggplot(lm\_data, aes(x = .resid)) +

geom\_histogram(binwidth = 2, fill = "blue", color = "black") +

labs(

title = "Histogram of Residuals",

x = "Residuals",

y = "Frequency"

)

# Create a Q-Q plot to check for normality of residuals

ggplot(lm\_data, aes(sample = .resid)) +

geom\_qq() +

geom\_qq\_line(color = "blue") +

labs(

title = "Q-Q Plot of Residuals",

x = "Theoretical Quantiles",

y = "Sample Quantiles"

)

**## Export the summary result: Simple linear regression**

## Export the summary result

library(stargazer)

# Export as a txt file

stargazer(lm\_model, type = "text", out = "E:/R\_workshop\_SUST/regression\_summary.txt")

# Export for LaTeX

stargazer(lm\_model, type = "latex", out = "E:/R\_workshop\_SUST/regression\_summary.tex")

# Export as html

stargazer(lm\_model, type = "html", out = "E:/R\_workshop\_SUST/regression\_summary.html")

**### Multiple regression model, linear restrictions, hypothesis testing**

We perform a multiple linear regression analysis using the "hprice" dataset

## Regression model (multiple)

# Fit a multiple linear regression model

lm\_model2 <- lm(price ~ bdrms+bdrms+lotsize+sqrft, data = hprice)

# Summarize the regression results

summary(lm\_model2)

## Multicollinearity test

vif(lm\_model2)

# Fit a another linear regression model

lm\_model3 <- lm(price ~ assess+bdrms+bdrms+lotsize+sqrft, data = hprice)

# Summarize the regression results

summary(lm\_model3)

## Multicollinearity test

vif(lm\_model3)

**## Hypothesis testing with linear restrictions**

We set a null hypothesis that the variables of lotsize and sqrft have no significant impact on the price. We will perform an F-test to test the hypothesis.

So, the null hypothesis is:

H0= lotsize=sqrft= 0

## We can examine the hypothesis and F test using car package

### Hypothesis test

# Fit a another linear regression model

un\_mod <- lm(price ~ assess+bdrms+bdrms+lotsize+sqrft, data = hprice)

myH0 <- c("lotsize","sqrft")

linearHypothesis(un\_mod, myH0)

**## Reporting regression results as we often see in empirical papers/published article**

library(stargazer)

stargazer(list(m1, m2,m3),type="text",keep.stat=c("n","rsq"), out = "E:/R\_workshop\_SUST/reg\_models.txt")

**Module VII: Logistic Regression**

We will use the "wage1" dataset to perform logistic regression and show the details.

# Load the wooldridge package

library(wooldridge)

library(tidyverse)

# Load the wage1 dataset

data("wage1")

# View the structure of the dataset

str(wage1)

# Create a new variable "high\_wage" based on the condition

wage1$high\_wage <- ifelse(wage1$wage > 5, 1, 0)

table(wage1$high\_wage)

theme\_set(theme\_minimal())

ggplot(wage1,aes(x=educ,y=high\_wage))+

geom\_jitter(height = 0.05,

alpha=0.1)

# Perform logistic regression

model1 <- glm(high\_wage ~ educ, data = wage1, family = binomial)

# Display the logistic regression summary

summary(model1)

ggplot(wage1,aes(x=educ,y=high\_wage))+

geom\_jitter(height = 0.05,

alpha=0.1)+

geom\_smooth(method = "glm",

method.args=list(family="binomial"),

se=F)

**### Using lessR package**

## Using lessR package

library(lessR)

Logit(high\_wage ~ educ, data = wage1)

## Multiple logistic regression

Logit(high\_wage ~ educ+exper+tenure, data = wage1)

**Part IV: T-Test**

**One Sample T-Test**

A one-sample t-test is used to determine whether the mean of a single sample differs from a known or hypothesized population mean. Let's test whether the mean sepal length of the iris dataset significantly differs from a hypothesized population mean of 5.8.

Set up your hypotheses.

Null Hypothesis (H0): The mean sepal length in the iris dataset is equal to 5.8.

Alternative Hypothesis (Ha): The mean sepal length in the iris dataset is not equal to 5.8.

# Load the iris dataset

data(iris)

# Perform the one-sample t-test

result\_one\_sample <- t.test(iris$Sepal.Length, mu = 5.8)

# View the test result

result\_one\_sample

##Compare the p-value to your chosen significance level

#to determine whether to reject the null hypothesis.

**Two Independent Sample T-Test**

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

Let's test whether there is a significant difference in sepal length (Sepal.Length) between two species of iris flowers: Setosa and Versicolor.

Set up your hypotheses.

Null Hypothesis (H0): The mean sepal length of Setosa and Versicolor iris flowers is equal.

Alternative Hypothesis (Ha): The mean sepal length of Setosa and Versicolor iris flowers is not equal.

# Filter data for Setosa and Versicolor species

setosa\_sepal\_length <- iris$Sepal.Length[iris$Species == "setosa"]

versicolor\_sepal\_length <- iris$Sepal.Length[iris$Species == "versicolor"]

# Perform the two-sample t-test

result\_two\_sample <- t.test(setosa\_sepal\_length, versicolor\_sepal\_length)

# View the test result

result\_two\_sample

**Paired sample T-Test**

We'll use the ToothGrowth dataset, which contains data on the length of tooth growth in guinea pigs under different supplement conditions.

Let's say we want to test if there is a significant difference in tooth growth between two delivery methods of vitamin C supplements: OJ (orange juice) and VC (ascorbic acid).

Null Hypothesis (H0): The mean tooth growth after using OJ is equal to the mean tooth growth after using VC.

Alternative Hypothesis (Ha): The mean tooth growth after using OJ is not equal to the mean tooth growth after using VC.

# Load the ToothGrowth dataset

data(ToothGrowth)

# View the first few rows of the dataset

head(ToothGrowth)

# Summary statistics of the dataset

summary(ToothGrowth)

# Structure of the dataset

str(ToothGrowth)

#Prepare the data

# Create vectors for OJ and VC

oj\_tooth\_length <- ToothGrowth$len[ToothGrowth$supp == "OJ"]

vc\_tooth\_length <- ToothGrowth$len[ToothGrowth$supp == "VC"]

# Perform the paired sample t-test

result\_pt <- t.test(oj\_tooth\_length, vc\_tooth\_length, paired = TRUE)

# View the test result

result\_pt

**Chi-squared Goodness fit test**

It is primarily used to assess if a sample data set follows a particular theoretical distribution or if there is a significant association between two categorical variables. For example:

- Checking if the distribution of blood types in a sample population follows the expected distribution (e.g., 40% Type A, 30% Type B, 20% Type AB, 10% Type O).

- Testing whether the frequency of certain traits in a population (e.g., eye color) matches expected Mendelian ratios.

- Investigating whether there is a significant association between gender and a particular medical condition (e.g., heart disease) by examining a contingency table of gender (male/female) and medical condition (present/absent).

- Studying whether there is a significant relationship between a person's education level and their voting preferences in an election.

Example: Let's say, using a hypothetical dataset, we want to test whether the distribution of political party affiliations in a sample of voters matches an expected distribution.

Null Hypothesis (H0): The observed distribution of voter party affiliations matches the expected distribution.

Alternative Hypothesis (Ha): The observed distribution of voter party affiliations is different from the expected distribution.

# Create a hypothetical dataset of voter party affiliations

voter\_data <- c("Republican", "Democrat", "Independent", "Republican", "Republican","Democrat", "Independent", "Independent", "Democrat", "Independent")

# Create a table of observed frequencies

observed <- table(voter\_data)

# Expected frequencies based on the assumed distribution

expected <- c(0.4 \* length(voter\_data), 0.4 \* length(voter\_data), 0.2 \* length(voter\_data))

# Calculate the Chi-squared statistic

chi\_squared <- sum((observed - expected)^2 / expected)

# Calculate the p-value

p\_value <- 1 - pchisq(chi\_squared, 2)

# Set the significance level

alpha <- 0.05

# Make a decision

if (p\_value < alpha) {

cat("Reject the null hypothesis. The observed distribution is different from the expected distribution.")

} else {

cat("Fail to reject the null hypothesis. The observed distribution is consistent with the expected distribution.")

}

# Display the Chi-squared statistic and p-value

cat("\nChi-squared statistic:", chi\_squared)

cat("\np-value:", p\_value)

**Chi-squared test of independence**

This test is used to determine whether there is a significant association or relationship between two categorical variables. In other words, it assesses whether the occurrence of one variable is independent of the occurrence of the other variable. It is often applied to contingency tables.

Example 1: We want to test if there is any significant statistical association between employee onboarding status and their turnover status.

# Import Employee Turnover data

emp <- read.csv("EmployeeTurnover.csv")

## construct a contingency table (cross tabulation)

tb1 <- xtabs(~Onboarding+Turnover, data = emp)

print(tb1)

## Chi-square test of independence

#Pearson's Chi-squared test #(2x2)

chisq.test(tb1)

# with p-value = 0.000001476, there is a significant statistical association

prop.table(tb1,2)

## Phi coefficient or pearson corr coeff

library(psych)

phi(tb1)

## Examine the observed and expected value

chisq <- chisq.test(tb1)

chisq$observed

chisq$expected

chisq$residuals ## pearson residuals

Example 2: We use housetasks data set. The data is a contingency table containing 13 housetasks and their distribution in the couple. Test the hypothesis whether wife’s housing tasks are independent of husband’s tasks at .05 significance level.

# Import the data

file\_path <- "http://www.sthda.com/sthda/RDoc/data/housetasks.txt"

housetasks <- read.delim(file\_path, row.names = 1)

head(housetasks)

#Contingency table can be visualized using

#the function balloonplot() [in gplots package].

library("gplots")

# 1. convert the data as a table

dt <- as.table(as.matrix(housetasks))

# 2. Graph

balloonplot(t(dt), main ="housetasks", xlab ="", ylab="",

label = FALSE, show.margins = FALSE)

#It’s also possible to visualize a contingency table as a mosaic plot.

library("graphics")

mosaicplot(dt, shade = TRUE, las=2,

main = "housetasks")

chisq <- chisq.test(housetasks)

chisq

As the p-value 0.0000003 is smaller than the .05 significance level, we do reject the null hypothesis that the wife’s housekeeping tasks habit is independent of her husband’s tasks.

**##Cronbach Alpha/** **McDonald's ω (omega) is a measure of internal consistency reliability.**

Internal consistency, often measured using Cronbach's Alpha, is a statistical measure of how well the items within a scale or questionnaire correlate with each other. If Alpha is close to 1, it indicates high internal consistency, suggesting that the items in your scale are strongly correlated with each other. McDonald's ω (omega) is a measure of internal consistency reliability, similar to Cronbach's Alpha.

If Alpha is closer to 0, it suggests lower internal consistency, indicating that the items in your scale may not be measuring the same underlying construct consistently.

Typically, a Cronbach's Alpha of 0.70 or higher is considered acceptable for most research purposes, but the threshold can vary depending on the context and the specific field of study.

##Import Job satisfaction data

job <- read.csv("JobSatisfaction.csv")

library(jmv)

reliability(

data = job,

vars = c("TurnInt1", "TurnInt2","TurnInt3"),

omegaScale = TRUE,

meanScale = TRUE,

sdScale = TRUE,

corPlot = TRUE,

alphaItems = TRUE,

omegaItems = TRUE,

meanItems = TRUE,

sdItems = TRUE,

itemRestCor = TRUE)

reliability(

data = job,

vars = c("Engage1","Engage2","Engage3","Engage4"),

omegaScale = TRUE,

meanScale = TRUE,

sdScale = TRUE,

corPlot = TRUE,

alphaItems = TRUE,

omegaItems = TRUE,

meanItems = TRUE,

sdItems = TRUE,

itemRestCor = TRUE)

**Survival Analysis**

We will use the classic “Survival from Malignant Melanoma” dataset. The data consist of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark, during the period 1962 to 1977. We are interested in the association between tumour ulceration and survival after surgery.

**Part V: ANOVA**

**One way ANOVA**

We'll use the InsectSprays dataset, which contains data on the effectiveness of different insect sprays. This dataset contains information about the count of insects killed by various insect sprays.

# Load the InsectSprays dataset

data(InsectSprays)

# View the first few rows of the dataset

head(InsectSprays)

Next, perform a one-way ANOVA to determine if there are significant differences in insect kill counts among different sprays. In this analysis, the dependent variable is the insect count, and the independent variable is the type of insect spray. Also, we may want to perform post-hoc tests (e.g., Tukey's HSD or pairwise t-tests) to identify which specific groups differ from each other.

# Perform the one-way ANOVA

anova\_result <- aov(count ~ spray, data = InsectSprays)

# Summarize the ANOVA results

summary(anova\_result)

#Post-Hoc Tests

InsectSprays %>%

aov(count ~ spray, data = .) %>%

TukeyHSD() #%>%

plot()

**Two-way ANOVA**

We'll use the ToothGrowth dataset, which contains data related to the effect of two supplements on tooth growth. This dataset contains information about the length of tooth growth in guinea pigs exposed to different supplements and doses. We perform a two-way ANOVA to determine if there are significant effects of both the "supplement" and "dose" factors on tooth growth. In this analysis, "len" (tooth length) is the dependent variable, and "supp" (supplement type) and "dose" (dose level) are the independent variables. We also want to perform post-hoc tests to further investigate the differences between groups or levels.

# Load the ToothGrowth dataset

data(ToothGrowth)

# View the first few rows of the dataset

head(ToothGrowth)

# Perform the two-way ANOVA

anova\_result <- aov(len ~ supp \* dose, data = ToothGrowth)

# Summarize the ANOVA results

summary(anova\_result)

#Post-Hoc Tests

ToothGrowth %>%

aov(len ~ supp \* dose, data = .) %>%

TukeyHSD() %>%

plot()

## Converting categorical values into labels

# Load the dplyr library if not already loaded

library(dplyr)

##using affairs data

# Create the new column 'marriage' with recoded values

affairs <- affairs %>%

mutate(marriage = case\_when(

ratemarr %in% c(1, 2) ~ "unhappy",

ratemarr == 3 ~ "average",

ratemarr %in% c(4, 5) ~ "happy",

TRUE ~ as.character(ratemarr) # Keep other values as is

))

Ref:

* Hadley Wickham’s Advanced R book
* Roger Peng’s R Programming for Data Science book
* DataCamp’s Intermediate R course
* Coursera’s R Programming course
* https://ggplot2-book.org/getting-started.html
* Data Carpentry (http://datacarpentry.org/), data camp, data quest, Kaggle
* Harvard Chan Bioinformatics Core (HBC) under the open access terms of the Creative Commons Attribution license (CC BY 4.0),
* The Book of R: A First Course in Programming and Statistics by Tilman M. Davies
* Equitable Equations (<https://www.youtube.com/watch?v=TuMjC0HFF3c&list=PLKBUk9FL4nBYpUKszG4edyAiM9aeTT1yv>)
* MarinStatsLectures-R Programming & Statistics (<https://www.youtube.com/@marinstatlectures/playlists>)