Conducting and Presenting Multiple Linear Regression Analysis Using R

Maria Paula Rojas Conejo 400962585

Fresenius University of Applied Science

Data Science and Data Analytics (WS 2025/26)

Prof. Dr. Stephan Huber

2026-12-10

Author Note

Correspondence concerning this article should be addressed to Maria Paula Rojas Conejo 400962585, Email: rojas_conejo.maria@stud.hs-fresenius.de

Abstract

This document outlines the requirements for successfully contributing to the course through a presentation and a project report. It introduces the process, establishes the standards, and provides tips how students can effectively meet the expectations.

Conducting and Presenting Multiple Linear Regression Analysis Using R

Word count: 1460

1 Introduction

According to Alexopoulos (2010) when it comes to data analysis one of the main goals is to extract from raw information the accurate estimation. One of the common questions in data analysis is the statistical relationship between a response variable (Y) and explanatory variables (Xi). Alexopoulos (2010) continues to explain that an option to answer this question is to employ regression analysis in order to model its relationship. In this project we examine the fundamental principles of multiple linear regression, a widely used statistical method as the author mention. At the same time it is intended to clarify how the method works and when it is appropriate to use.

2 What is Multiple Linear Regression?

##Hello

The author Uyanık and Güler (2013) explains that regression analysis is used to determine the relationships between two or more variables that exhibit cause–effect patterns, and to make predictions based on those relationships. Similarly, Marill (2004) states that multiple linear regression is a generalization of simple linear regression in which more than one predictor variable is included. The formula is the following:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + u_i$$

Where:

- Y_i — dependent variable
- b_0 — intercept
- $b_1 \ldots b_n$ — regression coefficients
- $X_{1i} \ldots X_{ni}$ — independent variables
- u_i — disturbance error

3 Usefull aplications

Multiple linear regression is widely applied across numerous fields due to its ability to model relationships between multiple predictors and a continuous outcome. In business and economics for example, it is commonly used for demand forecasting as predicting sales based

on factors like price, advertising, etc. As well as for financial risk modeling and market research aimed at identifying the drivers of customer satisfaction or spending.

Moreover, social sciences and psychology rely heavily on multiple regression to explore how variables such as education, income, and demographics influence life satisfaction for example. Another example is to predict academic performance based on study habits and cognitive factors.

In medicine and public health, regression models can help predict patient outcomes from characteristics such as age, BMI, and biomarkers. Finally, in machine learning and data science, multiple linear regression serves both as a baseline predictive model for continuous outcomes and as a tool for feature importance analysis, often forming a foundational step before applying more advanced modeling techniques such as regularization or generalized linear models.

4 Assupmtions

According to Ernst and Albers (2017) there are 4 assupmtions in the linear regression model, violating these assumptions can create several kinds of problems. First, the resulting estimates may be biased, meaning they do not,reflect the true population values. Second, the estimators might become inconsistent, so increasing the sample size no longer estimates will approach the true parameters. Third, may lose its efficiency, meaning it can yield less precise estimates. Fourth, hypothesis tests and confidence intervals may become unreliable. Note that these assumptions specifically apply to estimation using the Ordinary Least Squares (OLS) method (a statistical method used to estimate the parameters of a linear regression model by minimizing the sum of the squared differences between the observed and predicted values), which is the default in many statistical software programs.

1. Linearity: The model assumes a linear and additive relationship between each predictor and the mean of the dependent variable, with errors averaging zero for any combination of predictors. Problems arise if relationships are nonlinear or if measurement error is present. Is important to mention that multicollinearity is not an assumption, but low overlap between predictors is desirable. (Ernst & Albers, 2017)
2. Normality: The model errors are assumed to follow a normal distribution centered at zero. (Ernst & Albers, 2017)
3. Homoscedasticity: The errors are expected to have constant variance across all values of the predictors. If the spread of errors changes (heteroscedasticity), the assumption is

violated. (Ernst & Albers, 2017)

4. Independence: Errors must be independent of each other, meaning their covariances are zero. This requires proper random sampling. Residual plots or autocorrelation checks help detect issues. (Ernst & Albers, 2017)

5 Multicollinearity

According to Daoud (2017) multicollinearity occurs when two or more predictor variables are highly correlated, which leads to an increase in the standard errors of the estimated coefficients. When standard errors become inflated, some or all coefficients may no longer appear significantly different from zero, meaning variables that should be statistically meaningful can incorrectly seem insignificant.

6 Linear Model or lm() in R

The Linear Model is is used to fit linear models, including multivariate ones. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance.

Basic Syntax

```
model <- lm(formula, data = dataset)
```

6.1 Preparing Data and Fitting the Model in R

R uses a concise formula syntax and the function lm() for linear models. Before fitting a model, it is good practice to inspect the data, check for missing values, and explore relationships.

```
# Load data
data(mtcars)
# Quick overview
str(mtcars)
summary(mtcars)
# Correlation matrix for numerical variables
cor(mtcars[, c("mpg", "wt", "hp", "cyl")])
# Simple scatterplot matrix
pairs(mtcars[, c("mpg", "wt", "hp", "cyl")])
```

6.2 Fitting a multiple linear regression with lm()

```
# Fit the model
model <- lm(mpg ~ wt + hp + cyl, data = mtcars)
# Basic summary output
summary(model)
```

To understand the output of this code, we need to understand the following concepts.

- Residuals: they are the difference between what the model predicts and what actually happened. Kim (2015) explains that residuals could reveal patterns in the data unexplained by the fitted model.
- Intercept: is the model's prediction of the outcome when all predictors are 0
- Estimated coefficients (Estimate): This tells you how much the outcome changes when a predictor increases by 1 unit, while the others stay the same.
- t-values and p-values: This is used to test whether the estimate is really different from zero and shows how likely it is that the result is just due to random chance.
- Residual standard error: This is the average size of prediction errors.
- Multiple R-squared: Measures the variation in the dependent variable the model explains.
- F-statistic: The F-test checks if the whole model (all predictors together) is useful.

At this stage, you have a working model, but it is not yet ready to report. Next you must check whether assumptions are reasonably satisfied.

6.3 Model Diagnostics in R

6.3.1 Residual plots and normality

```
# Basic diagnostic plots (4-panel layout)
par(mfrow = c(2, 2))
plot(model)
par(mfrow = c(1, 1))
```

Key plots:

- Residuals vs Fitted (model predicted values): Visualize whether the model is missing any non-linear relationships. When the residuals are evenly scattered around a flat horizontal line with no clear shape or trend, it suggests that the model fits well and that non-linear relationships are unlikely to be a problem. (Kim, 2015)
- Normal Q-Q: This plot shows if residuals are normally distributed. Points should follow the diagonal line if residuals are approximately normal. (Kim, 2015)
- Scale-Location: shows if residuals are spread equally along the ranges of predictors. This is how you can check homoscedasticity. It's good if you see a horizontal line with equally (randomly) spread points.(Kim, 2015)
- Residuals vs Leverage: it spots influential data points—cases that can strongly affect the regression results. The focus is on points that appear in the upper-right or lower-right areas of the plot or fall outside the dashed lines.(Kim, 2015)

6.3.2 Checking multicollinearity with VIF

The variance inflation factor (VIF) is a widely used measure; values above about 5–10 are often flagged as problematic.

```
# install.packages("car")  # run once if not installed
library(car)
vif(model)
```

6.4 Presenting Multiple Regression Results

The broom package converts R's model objects (like lm) into tidy data frames, making it easy to create tables and plots of results.

```
# install.packages("broom")  # run once
library(broom)
# Coefficient-level information
coef_table <- tidy(model, conf.int = TRUE)
coef_table
# Model-level summary (R^2, etc.)
model_fit <- glance(model)
model_fit
```

7 References

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, *14*(Suppl 1), 23.

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, *949*, 012009.

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ*, *5*, e3323.

Kim, B. (2015). *Understanding diagnostic plots for linear regression analysis*. University of Virginia Library; https://library.virginia.edu/data/articles/diagnostic-plots.

Marill, K. A. (2004). Advanced statistics: Linear regression, part II: Multiple linear regression. *Academic Emergency Medicine*, *11*(1), 94–102.

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234–240.

Appendix

Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published. I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

Checklist:

- ☒ The handout contains 3-5 pages of text.
- ☒ The submission contains the Quarto file of the handout.
- ☒ The submission contains the Quarto file of the presentation.
- ☒ The submission contains the HTML file of the handout.
- ☒ The submission contains the HTML file of the presentation.
- ☒ The submission contains the PDF file of the handout.
- ☒ The submission contains the PDF file of the presentation.
- ☒ The title page of the presentation and the handout contain personal details (name, email,matriculation number).
- ☒ The handout contains a bibliography, created using BibTeX with an APA citation style.
- ☒ Either the handout or the presentation contains R code that proofs the expertise in coding.
- ☒ The filled out Affidavit.
- ☒ The link to the presentation and the handout published on GitHub.

- ☒ In group work,each student's contribution is clearly defined and individual performance can be assessed using specified sections, page numbers or other objective criteria.

Maria Rojas, 10.12.2025, Köln, Germany