Lab 04: Regression Criticism

**Due date:** Thursday, March 14, 2025 submitted as Word document to Canvas ***Lab04*** link

This lab counts 9 % toward your total grade.

**Objectives:** In this lab, you will practice your skills in

1. Explore multiple regression
2. Residual plot
3. Multicollinearity
4. Outlier detection

**Format of answer:** Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included

# Task 1: Multiple regression (9 pts)

We will use the Boston dataset from the **MASS** package, which contains **506 observations and 15 variables** related to housing prices in the Boston suburbs. Data information please check this [link](https://search.r-project.org/CRAN/refmans/ISLR2/html/Boston.html) (Provide R code for each task and explain the result).

Total\_area is a simulated variable that is calculated based on **rm**.

1. Construct a multiple regression model using **medv** (median home value) as the dependent variable, with all other variables **except lstat** as independent variable (a total of 14 variables). (1 pt)

Boston = read.csv('Boston.csv')

var\_initial\_model <- Boston[ , !names(Boston) %in% "lstat"]

initial\_model = lm(medv~., data = var\_initial\_model)

summary(initial\_model)

A screenshot of a computer

AI-generated content may be incorrect.

1. Use Variance Inflation Factor (VIF) to detect multicollinearity among the predictors. (1 pt)

**vif(initial\_model)**

**A close up of numbers

AI-generated content may be incorrect.**

1. If multicollinearity is detected, remove **one highly collinear variable** based on domain knowledge and reasoning. Fit a new regression model (Model 2) without the problematic variable. (2 pts)

**As values of rm (314.326976)** and **total\_area (312.524907)** are above 10, they are considered as excessively high VIF values, suggesting severe multicollinearity. These variables are likely strongly correlated and could be problematic in the regression model.

var\_second\_model <- var\_initial\_model[ , !names(var\_initial\_model) %in% "total\_area"]

second\_model = lm(medv~., data = var\_second\_model)

summary(second\_model)

A screenshot of a computer

AI-generated content may be incorrect.

1. Compare Model 1 (from Task1. a) and Model 2 (from Task1. c) using: Standard errors and t-statistics. Please illustrate how multicollinearity impacts model estimates by analyzing changes in coefficient significance. (2 pts)

summary(initial\_model)

summary(second\_model)

vif\_initial\_model <- vif(initial\_model)

print(vif\_initial\_model)

vif\_second\_model <- vif(second\_model)

print(vif\_second\_model)

Some variables have decreased standard errors in Model 2, that makes the estimates more precise.The t-statistics have generally increased, indicating that some variables became more statistically significant after removingtotal\_aea. The significance of zn decreased, as seen in its t-value dropping from 2.669 to 2.573, meaning it is now less statistically significant. Some variables became slightly more significant, such as chas, nox, rm, and ptratio, as their t-values increased.

Multicollinearity is assessed using the Variance Inflation Factor (VIF). VIF for rm and total\_area are highin Model

1 , indicating moderate multicollinearity. After removing total\_area, the VIF values

decreased for rm in Model 2, meaning multicollinearity was reduced

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white background with black numbers

AI-generated content may be incorrect.

1. Check whether residuals exhibit constant variance (homoscedasticity) in Model 2. If heteroscedasticity is detected, suggest one potential remedies. (1 pt)

par(mfrow = c(1,1))

plot(initial\_model$fitted.values, initial\_model$residuals)

#add a horizontal line at 0

abline(0,0)

A graph of a number of dots

AI-generated content may be incorrect.

The Residuals vs. Fitted plot shows that residuals are not randomly scattered.There is a funnel shape (i.e., residuals spread out more for higher fitted values).This suggests heteroscedasticity (non-constant variance).

Possible Remedies

1. Log transformation is a powerful remedy for heteroscedasticity because it Stabilizes variance,
2. Reduces the effect of extreme values, Improves the linearity of relationships and normalizes skewed data. By applying log transformation to a dataset, F-statistic is 89.01 (higher than

before), indicating model 2 can explain more variable in the dependent variable.

1. Create Model 3 by adding **lstat** to Model 2. Use **ANOVA** to compare Model 2 (without lstat) and Model 3 (with lstat) to determine if lstat significantly improves model performance. (1 pt)

var\_third\_model = Boston[ , !names(Boston) %in% "total\_area"]

third\_model = lm(medv~., data = var\_third\_model)

anova(second\_model, third\_model)

lstat significantly improves the model\_third\_model because it reducesresidual sum of square (RSS) from 13490

to 11079, has a high F-statistic (107.06), and has an extremely low p-value (2.2e-16). This

confirms that model\_third\_model (with lstat) is statistically better than Model 2 (without lstat) and lstat

has a strong effect on medv.

A screenshot of a computer

AI-generated content may be incorrect.

1. Based on Model 3, using **car::influenceIndexPlot()** to investigate the presence of outliers or influential cases in the dataset. Explain the results and analyze the possibility that these observations might be influential. (1 pt)

car::influenceIndexPlot(third\_model)

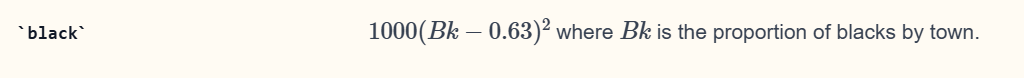
var\_third\_model[c(369,372,381,419), ]

A graph of a diagram

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



A graph of a line graph

AI-generated content may be incorrect.

Based on studentized residual (>3) and Bonferroin p-value (<0.05), Observations 369 and 372 are identified as outliers. When plotting the **X** variables for all observations where **medv = 50**, we found that observations **369** and **372** exhibit **higher house age, a higher proportion of non-retail business acres per town, and higher tax rates** compared to other observations with the same **medv** value. Notably, all three of these variables have a **negative relationship** with **medv**, indicating that higher values generally correspond to lower housing prices.

Therefore, among the observations where **medv = 50**, **observations 369 and 372** have less favorable housing attributes, which is why they are identified as **outliers**.

Based on hat-value Observations 381 and 419 have unusual x variables compared to others. Because they have extremely high crime rates (88.97, 73.53) and very low house prices (10.4, 8.8).