Lab 02: Bivariate regression and Transformation

**Due date:** Wednesday, Feb 12, 2025 submitted as Word document to Canvas ***Lab02***  link

This lab counts 9 % toward your total grade.

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

1. Explore bivariate regression
2. Confidence interval
3. Accuracy of the model
4. Variable Transformation

**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: Bivariate regression (2.5 pts)

The **MASS** library contains the **Boston** data set, which records the attribute information about house in suburbs of Boston. We will use **rm** (average number rooms per house) to predict **medv** (median value of owner-occupied homes in $1000s).

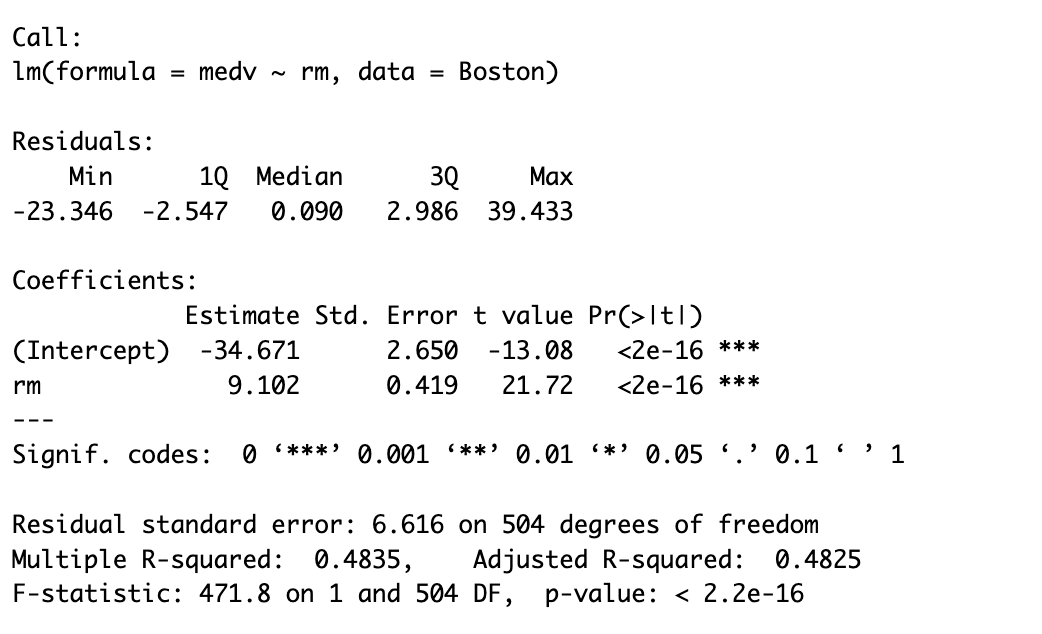
1. Using the **lm()** function to fit a bivariate linear regression model, with **medv** as the dependent variable and **rm** as the independent variable. (0.5 pts)

data('Boston')

boston\_model = lm(medv~rm, data = Boston)

1. Get detailed information about the linear model you constructed in 1.a using **summary()**. Interpret the intercept, slope and . (1 pts)

summary(boston\_model)



1. Compute the 95% confidence interval for the estimated regression parameters. Does the conclusion align with the results obtained from the t-test in part 1.b? yes or no, please interpret. (1pts)

confint(boston\_model)

A black numbers and text

AI-generated content may be incorrect.

Yes, the conclusion align with the results obtained from the t-test in part 1.b. As summary shows that the t-test checking whether rm has a significant impact on medv. Low p- value and confidence interval without 0 reveals that rm is statistically significant.

# Task 2: Bivariate Regression Model and Variable Transformation (6.5 pts)

The **UN** dataset from the **carData** package contains various global development indicators. We will analyze the relationship between **infant mortality rate (infantMortality) (**dependent variable) and **GDP per capita (**ppgdp**)** as (independent variable).

1. Remove NA value in UN dataset using **na.omit(UN).** (0.5 pts)

data('UN')

par(mfrow = c(1,1))

new\_un = na.omit(UN)

1. Create a scatterplot of **infant mortality rate (infantMortality)** versus **GDP per capita (ppgdp) using car::scatterplot()**. By visually inspecting the box plots and the LOESS curve, determine whether data transformation is advisable for dependent variable and independent variable. (1pts)

car::scatterplot(infantMortality~ppgdp, data = new\_un,xlab = 'Gross Domestic Product per Capita',

ylab = 'Infant Mortality Rate (per 1000 births)',

regLine = list(col = 'darkgreen'),

smooth = list(col.smooth = "red"))

A graph with numbers and lines

AI-generated content may be incorrect.

Yes, it is data transformation is advisable for dependent variable and independent variables because:

Dependent and independent variable have skewness based on boxplot. There is a non-linear relationship between GDP per capita (independent variable) and infant mortality rate (dependent variable). Log transformation might help linearize the relationship for better model fitting.

1. If a transformation is needed for the independent variable, find the using:

**Box-Cox transformation (summary (car::powerTransform(lm(*varName*~1))))** Please justify whether log-transformation () should be used or if is more appropriate. (1pts)

p1 = powerTransform(ppgdp~1, data = new\_un, family = 'bcPower')

summary(p1)

A screenshot of a computer code

AI-generated content may be incorrect.

Test for if transformation is needed has p-value < 2.22e-16, so we need to to use transformaton. Test for if transformation parameter equal to 0 (log transformation)., p-value of 0.68866. It is statistically non-significant. So we failed to reject the null-hypothesis (transformation parameter is equal to 0), so We should use log transformation (transformation parameter equal to 0)

1. Transformed the independent variable using , , . (1pts)
   1. Construct the histogram for the transformed distribution with different value.

par(mfrow = c(1,3))

lambda\_1 = car::bcPower(new\_un$ppgdp, lambda= 1)

lambda\_optimal = car::bcPower(new\_un$ppgdp, lambda=0.019)

lambda\_negative = car::bcPower(new\_un$ppgdp, lambda=-1)

hist(lambda\_1,breaks = 12,main = 'lambda = 0',xlab = 'x')

hist(lambda\_optimal,breaks = 12,main = 'lambda = 0.019',xlab = 'x')

hist(lambda\_negative,breaks = 12,main = 'lambda = -1',xlab = 'x')

A graph of a number of columns

AI-generated content may be incorrect.

As we can see from the difffernt values of lambda (-1, 0.019, 1), there is an adjustment in the skewness. When , no transformation is applied, preserving the original distribution. When the transformation effectively adjusted the skewness, bringing the distribution closer to normaity. However, when The transformation over-adjusted the distribution.

* 1. Evaluate the skewness and test whether the variables are approximately normal distribution.

shapiro.test(lambda\_1)

shapiro.test(lambda\_optimal)

shapiro.test(lambda\_negative)

A screenshot of a computer code

AI-generated content may be incorrect.

1. None of the three distributions follow a normal distribution. However, the transformation with the optimal lambda value brings the distribution closer to normality compared to the other two, as indicated by the larger p-value.If a transformation appears necessary for dependent variable, find the . (1pts)

p2 = powerTransform(infantMortality~log(ppgdp), data = new\_un, family = 'bcPower')

summary(p2)

A white paper with black text

AI-generated content may be incorrect.

1. Create a scatterplot with transformed variables **using car::scatterplot()**. Visually inspecting the box plots and the LOESS curve and describe how the transformation affects the relationship compared to the scatterplot in Task 2.b. (1pts)

car::scatterplot(log(infantMortality)~log(ppgdp), data = new\_un,xlab = 'Gross Domestic Product per Capita',

ylab = 'Infant Mortality Rate (per 1000 births)',

regLine = list(col = 'darkgreen'),

smooth = list(col.smooth = "red"))

A graph of a graph with blue and red dots

AI-generated content may be incorrect.

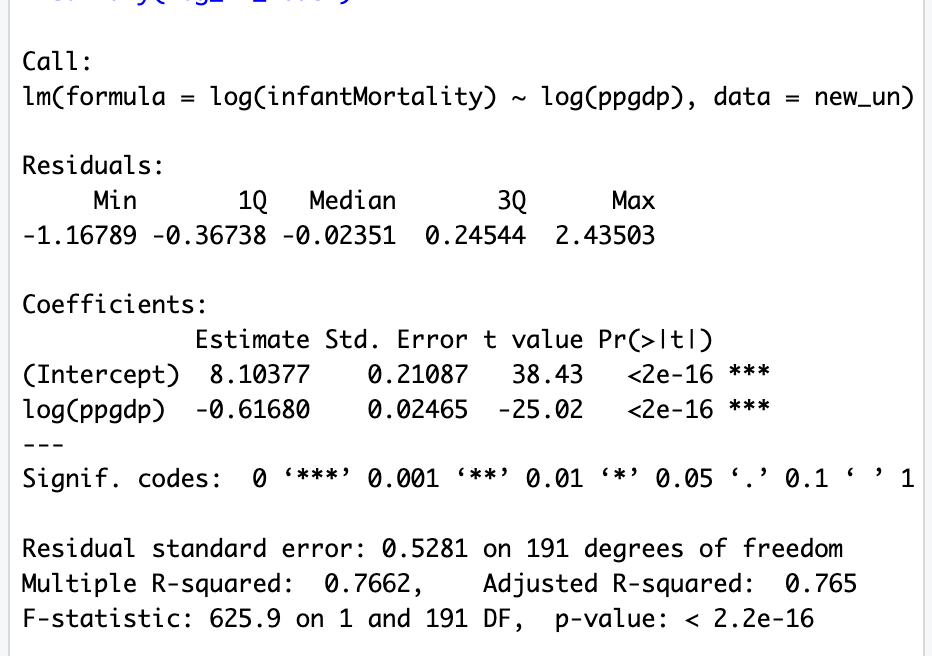
Box plots and the LOESS curve was non- linear and skewed. Transformation has improved the linear relationship and stabilizes the variance.

GDP per capita was decreasing with the increase in low infant mortality rate (per 1000 births) and flattens for higher GDP. After transformation, the spread of points is approximately aligned with straight-line trend.

1. Estimate the bivariate regression model using the transformed variables and interpret the estimated coefficients. (1pts)

log\_lm\_model = lm(log(infantMortality)~log(ppgdp), data = new\_un)

summary(log\_lm\_model)



Estimated regression model is:

log(infantMortality)=8.10377−0.61680×log(ppgdp)

According to the model, predicted log of infant mortality will be 8.10377, when Gross domestic product per capita is 0.

The negative value of slope= -0.61680, indicates that there is a negative relationship between GDP per capita and infant mortality. With 1% increase in GDP, there is 0.61680% decrease in infant mortality rate.