॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Symbiosis International (Deemed University)

(Established under section 3 of the UGC Act, 1956)

Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category – I by UGC Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Assignment No. 07

Subject: Data Science Lab

Name of Student	**Achyut Shukla**
PRN No.	**20070122005**
Branch	CS
Class	**A**
Academic Year & Semester	2023-24, VII
Date of Performance	**29th August, 2023**
Title of Lab Assignment	Multiple Regression Model Development

Theory: Multiple Linear Regression

Multiple linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and multiple independent variables (features). It is an extension of simple linear regression, where we consider more than one predictor variable to make predictions. In the context of your assignment, multiple linear regression aims to predict a continuous target variable based on one or more input features.

Model Representation:
In multiple linear regression, the relationship between the dependent variable Y and the independent variables X1, X2, ..., Xn is represented as:

$$Y=\beta0+\beta1 * X1+\beta2 * X2+...+\beta n * Xn+\varepsilon$$

Here:

Y is the dependent variable (the variable you want to predict). X1,
X2, ..., Xn are the independent variables (features).
β0 is the intercept, representing the value of Y when all independent variables are zero.

β1, β2, ..., βn are the coefficients, representing the change in Y for a unit change in each respective independent variable. ε represents the error term, accounting for the variability in Y that is not explained by the independent variables.

Model Training:

The goal of training a multiple linear regression model is to find the values of the coefficients (β0, β1, β2, ..., βn) that minimize the sum of squared residuals. Residuals are the differences between the actual target values and the predicted values by the model.

This minimization is typically achieved using methods like ordinary least squares (OLS) or gradient descent. The coefficients are estimated to fit the best linear relationship between the features and the target.

Assumptions of Multiple Linear Regression:

Linearity: The relationship between the dependent and independent variables is linear.
Independence: The errors (residuals) are independent of each other.
Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
No multicollinearity: The independent variables are not highly correlated with each other. Normally Distributed Errors: The errors follow a normal distribution.

Model Evaluation:

To evaluate the performance of a multiple linear regression model, several metrics can be used, including:

Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.
Root Mean Squared Error (RMSE): The square root of MSE, providing a measure in the same units as the target variable.
R-squared (R^2): Represents the proportion of the variance in the target variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
Feature Selection:
Feature selection is an important step in multiple linear regression. It involves choosing the most relevant independent variables to include in the model. Techniques like forward selection, backward elimination, or Lasso regression can be used for feature selection.

In summary, multiple linear regression is a powerful tool for modeling relationships between multiple predictors and a continuous target variable. Understanding its theory, assumptions, and evaluation metrics is crucial for accurate model building and interpretation.

Code:

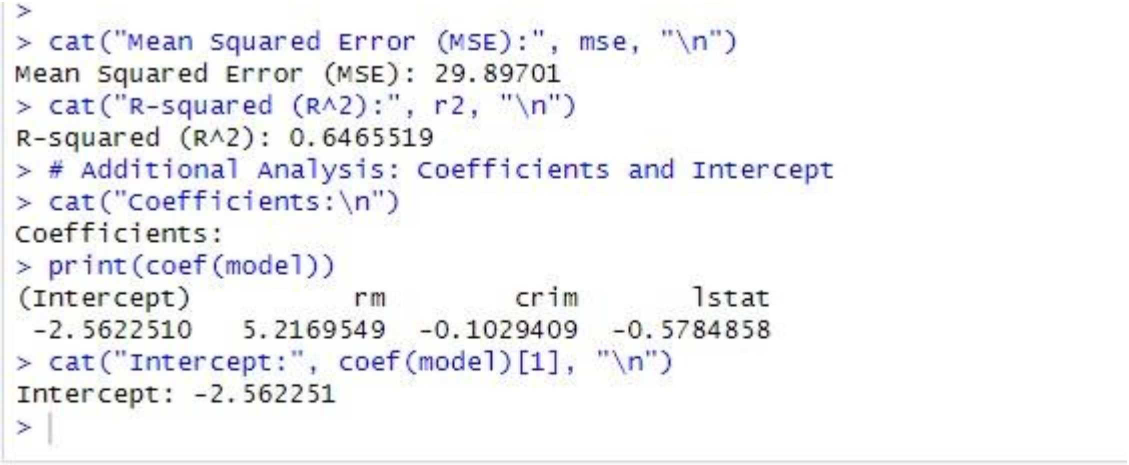
```
 # Step 1: Load the dataset
 library(MASS)  # Load the MASS library for the Boston dataset data(Boston)
 # Load the Boston Housing dataset

 # Step 2: Exploratory Data Analysis (EDA) and Visualization #
 Display basic statistics of the dataset
```

```
 summary(Boston)

 # Visualize the relationships between variables
 plot(Boston$rm, Boston$medv, xlab = "Average Number of Rooms (RM)",
     ylab = "Median Home Value (MEDV)", main = "Relationship between RM and MEDV")
 # Step 3: Assign Target Variable and Feature Selection
 X <- Boston[, c("rm", "crim", "lstat")]  # Example features, choose relevant features from your dataset
 y <- Boston$medv
 # Step 4: Model Selection - Multiple Linear Regression
 # Step 5: Model Training - Fit the model model
 <- lm(y ~ ., data = data.frame(y = y, X))
 # Step 6: Model Evaluation - Using appropriate metrics
 y_pred <- predict(model, newdata = data.frame(X)) mse
 <- mean((y - y_pred)^2)
 r2 <- 1 - mse / var(y)
 cat("Mean Squared Error (MSE):", mse, "\n") cat("R-squared
 (R^2):", r2, "\n")

 # Step 7: Model Improvement (if needed) - Fine-tune hyperparameters

 # Additional Analysis: Coefficients and Intercept
cat("Coefficients:\n") print(coef(model))
cat("Intercept:", coef(model)[1], "\n")
```

Output:

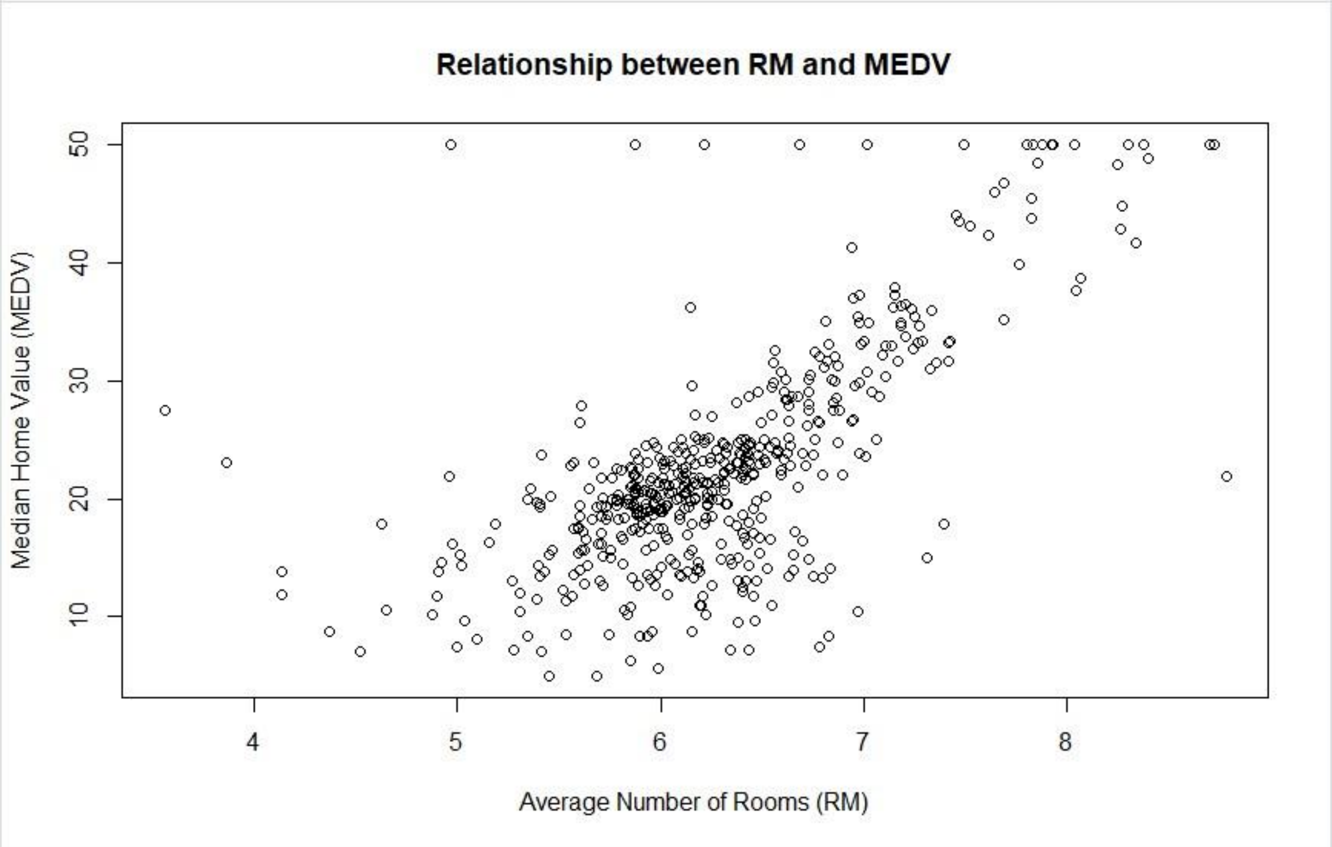
Answer:

Model Evaluation:



```
>
> cat("Mean Squared Error (MSE):", mse, "\n")
Mean Squared Error (MSE): 29.89701
> cat("R-squared (R^2):", r2, "\n")
R-squared (R^2): 0.6465519
> # Additional Analysis: Coefficients and Intercept
> cat("Coefficients:\n")
Coefficients:
> print(coef(model))
(Intercept)          rm        crim       lstat
 -2.5622510   5.2169549  -0.1029409  -0.5784858
> cat("Intercept:", coef(model)[1], "\n")
Intercept: -2.562251
> |
```

Plot:



Conclusion: *We've learnt how to develop a multiple regression model.*