**Student: Seif Kungulio**

**Date: 04/06/2025**

**Subject: Project 4**

**Class: DSCI 512**

**Section: 01W**

**Instructor: Dr. Nengbing Tao**

**File Name: Project4\_Kungulio\_Seif.docx**

1. Read the dataset in Boston.csv into R. Call the loaded data Boston. Make sure that you have the directory set to the correct location for the data.

A screenshot of a computer program

AI-generated content may be incorrect.

1. The response is nox and the predictor is dis. Use the poly() function to fit a cubic polynomial regression to predict nox using dis. Report the regression output.

A screenshot of a computer program

AI-generated content may be incorrect.

A cubic polynomial regression model effectively describes the complex, non-linear relationship between distance to employment centers **(dis)** and nitrogen oxide concentration **(nox)** in Boston. The model exhibits a strong fit **(R-squared ~71.48%)**, significant coefficients, and accurate predictions, demonstrating that “dis” is a powerful predictor of “nox”.

1. Your assistant data scientist, Tom Johnson, is considering predicting “nox” using “dis” as a predictor. He proposes models from degree 5, degree 4, and degree 3, and degree 2 polynomial regression. Please perform cross-validation using caret package to select the optimal degree for the polynomial and justify your answer.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer error

AI-generated content may be incorrect.

A degree 3 polynomial regression model, validated through 10-fold cross-validation, is **the optimal choice** for predicting nitrogen oxide concentration **(nox)** from distance to employment centers **(dis)** in the Boston dataset. It yielded the **lowest RMSE**, indicating **the best fit**, while higher-degree models showed overfitting and lower-degree models underfitting. This suggests a cubic relationship between “dis” and “nox”, and the cross-validation assures the model's reliability for unseen data.

1. Tom just took the DSCI 512. You recommend that he perform the following GAM analysis.
   1. Predict “nox” using a smoothing spline of degree 3 in “dis” and a smoothing spline of degree 2 in “medv”.

A screenshot of a computer program

AI-generated content may be incorrect.

A Generalized Additive Model (GAM) model effectively predicts nitrogen oxide **(nox)** concentrations in Boston using distance to employment centers **(dis)** and median home values **(medv)**. Both predictors show significant, non-linear relationships with “nox”, with “dis” having a stronger influence than “medv”. The GAM's fit is satisfactory, demonstrating the necessity of non-linear smooth functions over linear models for accurately representing these relationships.

* 1. Predict “nox” using a smoothing spline of degree 2 in “dis” and a smoothing spline of degree 1 in “medv”.

A screenshot of a computer program

AI-generated content may be incorrect.

This GAM model effectively predicts nitric oxide concentration **(nox)** in the Boston housing dataset using non-linear relationships with weighted distances to employment centers **(dis)** and median home values **(medv)**. Both predictors show significant, complex, curved effects on “nox”, and the model significantly reduces deviance compared to a null model, indicating a good fit. Residual analysis supports the model's robustness and predictive power.

* 1. Perform anova analysis. Recommend the best model and justify your answer.

A screenshot of a computer program

AI-generated content may be incorrect.

The ANOVA analysis compares two models, Model 1 and Model 2, to determine which provides a better fit for the data. Model 2, with slightly fewer residual degrees of freedom (500 vs. 502), has a lower residual deviance (1.7112) compared to Model 1 (1.8363), indicating a better fit. The additional complexity in Model 2, such as more smoothing terms for the variables “dis” and “medv”, contributes to this improvement. The Chi-Square test shows a significant p-value (1.151e-08), confirming that Model 2's added complexity leads to a significantly better fit.

In conclusion, **Model 2 is the preferred model**, as it significantly improves the fit over Model 1, making it the more suitable choice for the data.