STAT 6210 Homework 7 Report
Xue Ming Wang (Vivian)
G20580112

1. Data set

The dataset used for this report was collected from a study on the analgesic effects of treatments for neuralgia in elderly patients. The dataset consists of five variables, including treatment, sex, age, duration, and pain. In particular, the variable Treatment compares the effectiveness of two treatments (A, B) and a placebo (P), while the response variable Pain indicates whether the patient reported pain or not. Other explanatory variables include age, sex, and duration of the complaint before treatment initiation. The purpose of this report is to provide insights into the effectiveness of the different treatments and to analyze the influence of the explanatory variables on the patients' pain levels.

2. LOESS Procedure

First, we perform the LOESS method to examine the correlation between the response and Age and Duration separately.

2.1 Age

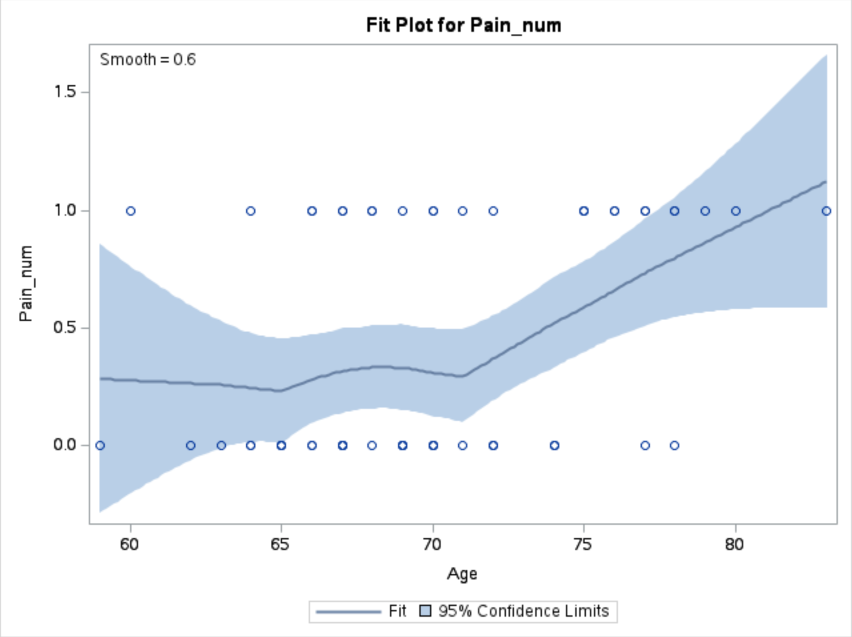

Figure 1. LOESS Procedure for Age

In Figure 1, a LOESS plot depicts the relationship between Age, the explanatory variable, and Pain, the response variable, which indicates whether a patient reported pain or not (with 1 indicating Yes and 0 indicating No). The plot displays a smooth curve with 95% confidence intervals. The LOESS curve illustrates a nonlinear relationship between Age and Pain, indicating a slight positive association between the two variables. The curve appears to be relatively flat for patients under 65 years old but increases significantly for those over 65, suggesting that elderly patients are more likely to report pain than their younger counterparts. The shaded areas around the curve represent the confidence intervals, reflecting the uncertainty in the estimated relationship between Age and Pain. The wider the confidence intervals, the greater the uncertainty in the estimates. The plot shows relatively wide confidence intervals, particularly for patients over the age of 70, indicating considerable variability in the relationship between Age and Pain.

2.2 Duration

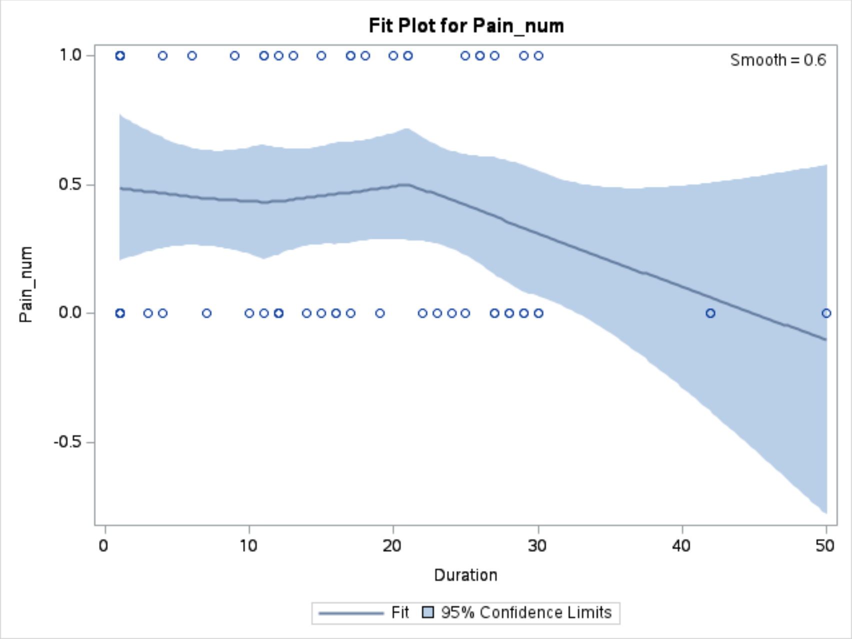

Figure 2. LOESS Procedure for Duration

Figure 2 displays a LOESS plot depicting the relationship between Pain and Duration, accompanied by a smooth curve and 95% confidence intervals. The curve represents a nonlinear relationship between Pain and Duration, with Pain showing an initial increase as Duration remains relatively constant at around 22, followed by a decrease as Duration reaches 50. The narrower confidence interval compared to the first plot indicates a higher degree of certainty in the estimate. However, further analysis is necessary to confirm the consistency of the trend between the response variable (Pain) and the explanatory variables (Age/Duration).

3. Simple LOGISTIC Procedure

3.1 Age

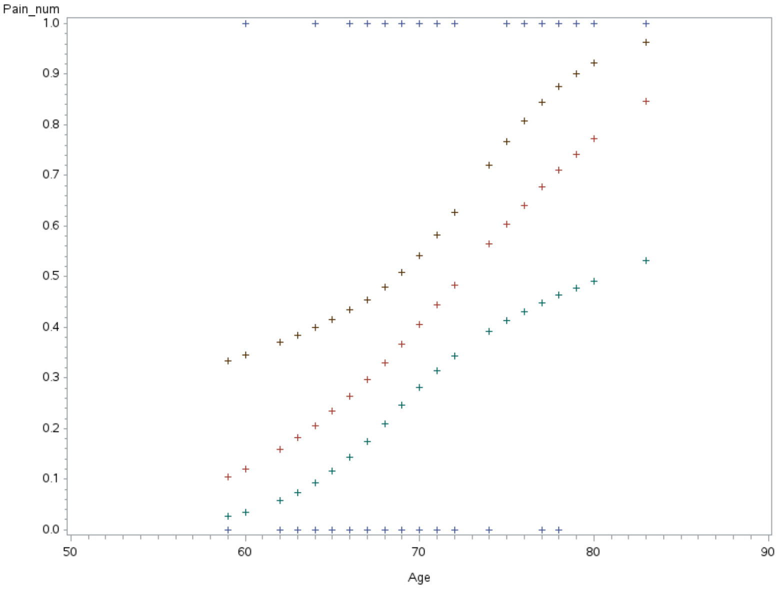

Figure 3. LOGISTIC Procedure for Age

Figure 3 showed the logistic regression analysis of the relationship between the pain response and the explanatory variables, Age. We see that there is a slight positive relationship between Age and Pain, which suggests that older patients may be more likely to report pain than younger patients. The confidence intervals are relatively wide, indicating some uncertainty in the estimates.

3.2 Duration

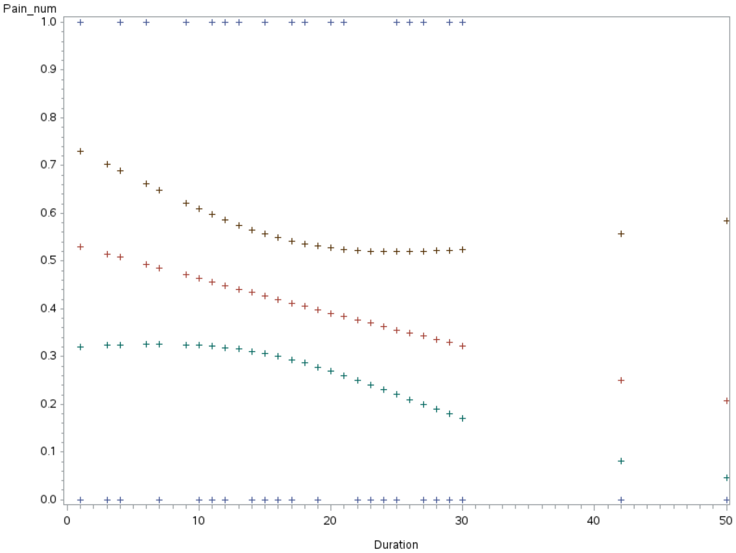

Figure 4. LOGISTIC Procedure for Duration

Figure 4 showed the logistic regression analysis of the relationship between the pain response and the explanatory variables Duration. We see a clear negative relationship between Duration and Pain, which suggests that patients who have had neuralgia for longer are less likely to report pain than those who have had it for a shorter time. The confidence intervals are narrower in this plot, indicating more precise estimates.

3.3 Comparing Simple Logistic Regression and Local Regression Analysis

After analyzing the relationship between Age and Duration with Pain using both LOESS and logistic regression models, we found that the trend in the probability of Pain is consistent for Age, but not for Duration. Both the LOESS and logistic regression models show an increasing probability of Pain with increasing Age, with the LOESS curve showing more variability and capturing potential non-linearities in the relationship while the logistic regression curve assumes a linear relationship. However, for Duration, the LOESS plot indicates a non-linear relationship with a flat trend to around duration 22, where it starts to decrease afterwards, while the logistic regression plot shows a clear negative linear relationship, with the probability of Pain decreasing from 0 to 50 for Duration.

Therefore, the trend is consistent for Age, but not for Duration. However, to gain a more comprehensive understanding of the relationship between Pain and its explanatory variables, we plan to conduct further analysis using all the relevant explanatory variables.

4. Full Logistic Regression Model

Parameter	Estimate	Pr > ChiSq
Intercept	19.0804	0.0049
Treatment A	0.8772	0.0963
Treatment B	1.4246	0.0183
Sex F	0.9118	0.0213
Age	-0.2650	0.0057

Table 1. Analysis Of Maximum Likelihood Parameter Estimates

Based on the results shown in Table 1, we have developed a logistic regression model to predict pain in patients with neuralgia. The model equation is $\log(p/1-p) = 19.0804 + 0.8772$ (Treatment A) + 1.4246 (Treatment B) + 0.9118 (Sex F) - 0.2650 (Age). The backward elimination procedure revealed that Treatment B, Sex F, and Age are significant predictors of pain, while Treatment A is not. This indicates that patients who receive Treatment B are more likely to experience pain compared to those who do not receive this treatment. Additionally, female patients are more likely to experience pain compared to male patients, while older patients are less likely to experience pain than younger patients.

The Hosmer-Lemeshow goodness-of-fit test yielded a p-value of 0.4055, suggesting that the model fits the data well and is not significantly different from the observed data. However, we need to assess the predictive power of the model to determine if it is optimized for prediction.

5. Predictive Power

5.1 Boxplot

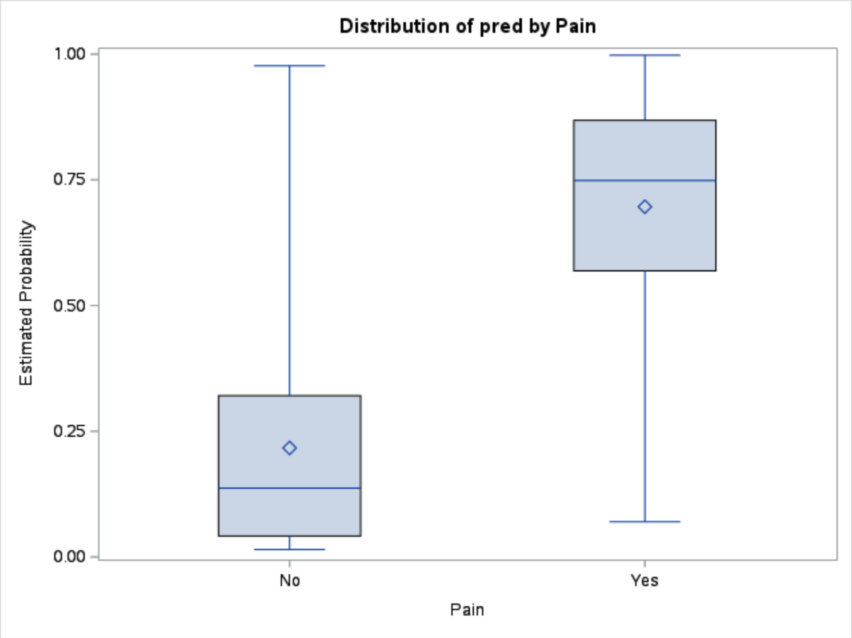

Figure 5. Boxplot for Predicted Probability vs. Response

The boxplot in figure 5 illustrates the distribution of predicted probabilities across each level of the response variable Pain. For the response variable "Pain = No", the median of the predicted probabilities is approximately 0.2, and the interquartile range (IQR) is relatively narrow. Conversely, for the response variable "Pain = Yes", the median of the predicted probabilities is around 0.8, and the IQR is wider. The boxplot suggests that the predicted probabilities for "Pain=No" are generally lower with a smaller range compared to the predicted probabilities for "Pain=Yes", which are higher with a wider range. These results indicate that the logistic regression model can differentiate between the two levels of the response variable with reasonable accuracy, implying strong predictive power.

5.2 ROC Curves

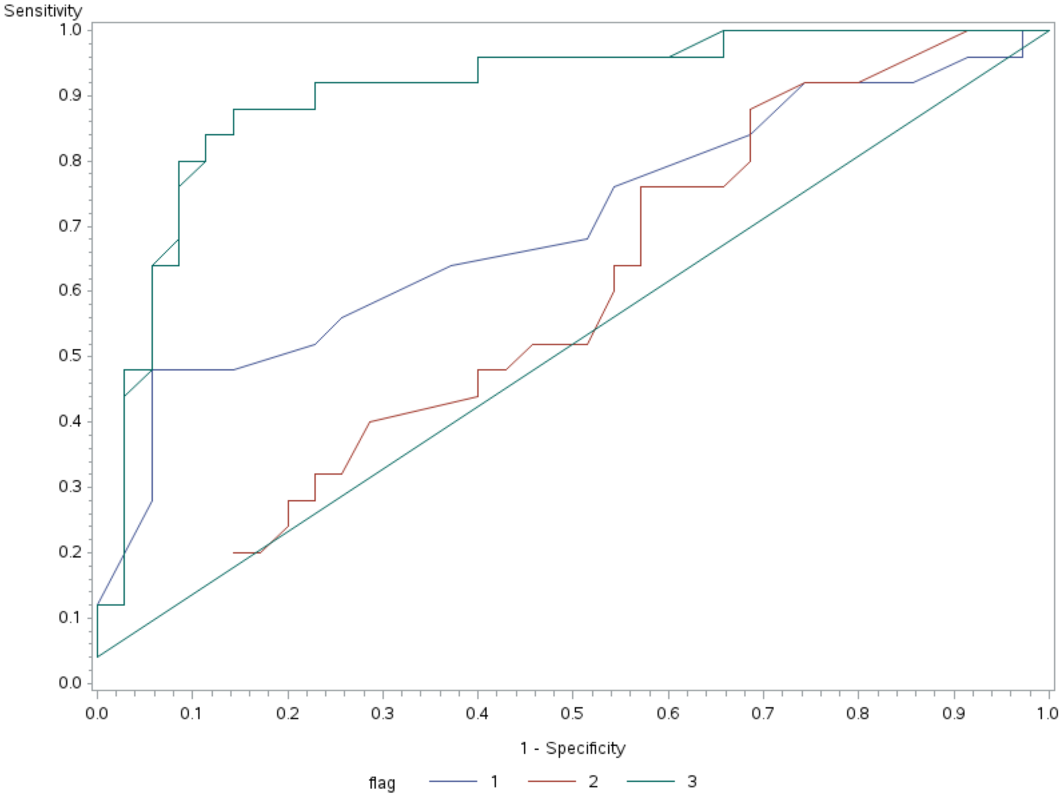

Figure 6. Logistic Model vs. Age Only vs. Duration Only in ROC Curves

The sensitivity (true positive rate) is plotted on the y-axis, and the x-axis shows 1-specificity (false positive rate) in the graph (Figure 6) displaying three ROC curves. Each curve represents one of the three models, with different colors and line patterns used to distinguish them. ROC curve 1 corresponds to Duration only, ROC curve 2 represents Age only, and ROC curve 3 corresponds to the logistic regression model with all explanatory variables.

Upon comparing the three curves, it is evident that the logistic regression model with all explanatory variables has the highest area under the curve (AUC = 0.9057, which is above 0.9 and can be considered as excellent), indicating the best overall performance in predicting Pain. Conversely, the model with age only has the lowest AUC, indicating the worst performance. The model with duration only has an intermediate AUC, indicating moderate performance. This plot demonstrates that the optimized model, which is the logistic regression model with all explanatory variables, has better overall performance in correctly classifying the outcome variable (Pain) in comparison to the models with only Age or Duration.

6. Model Diagnostics

6.1 Index Plot: Residual vs. Observation Number

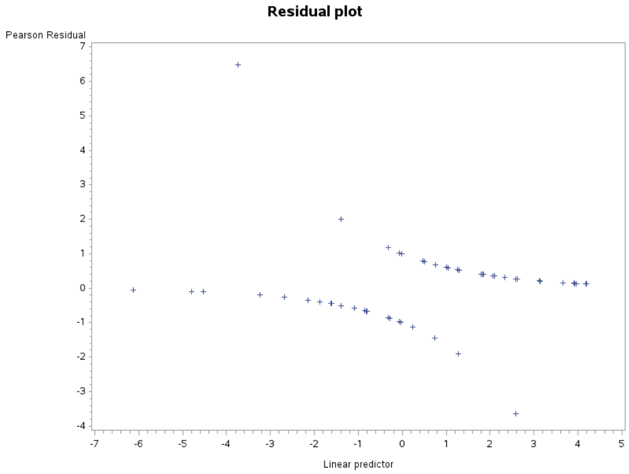

Figure 7. Index Plot: Residual vs. Observation Number

The purpose of the index plot in Figure 7 is to identify any unusual observations or patterns in the residuals. The plot shows that the residuals are distributed randomly around the zero line and have few outliers, indicating a good fit of the model to the data.

6.2 Plot of Residual vs. Predicted Probability

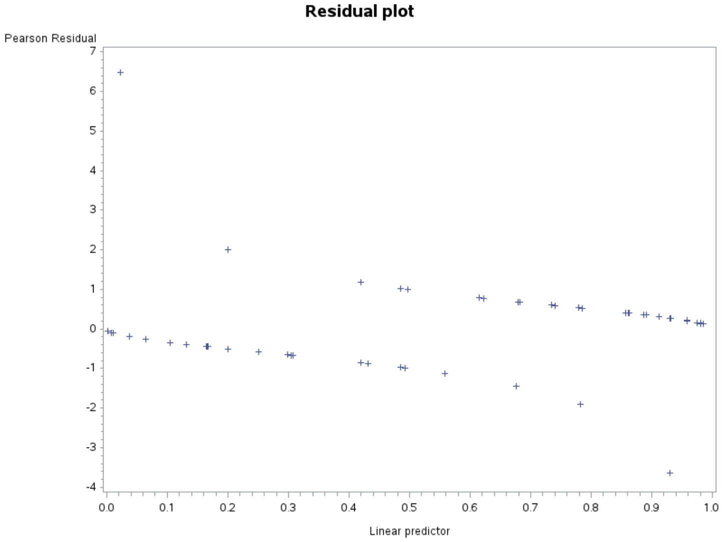

Figure 8. Plot of Residual vs. Predicted Probability

Figure 8 displays the scatter plot of residuals vs. predicted probability, which helps detect any systematic patterns in the residuals across different predicted probabilities. As shown in the plot, there is no clear evidence of any systematic patterns, indicating that the model is a good fit for the data.

6.3 Plot of Residual vs. Each Explanatory Variable

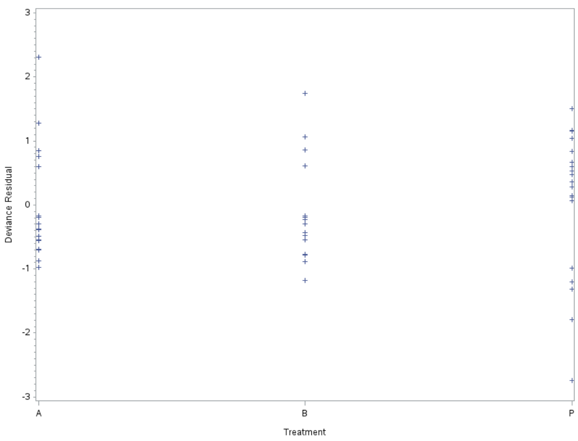

Figure 9. Plot of Residual vs. Treatment

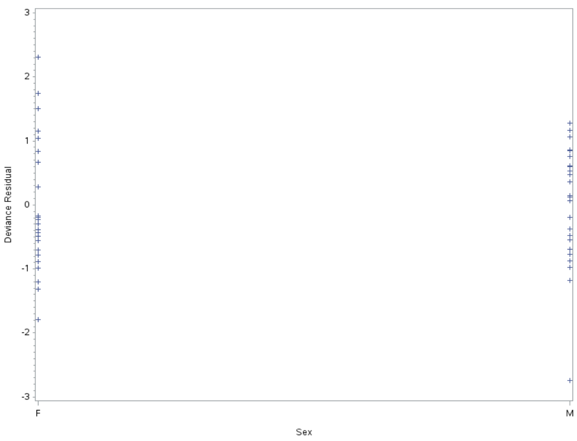

Figure 10. Plot of Residual vs. Sex

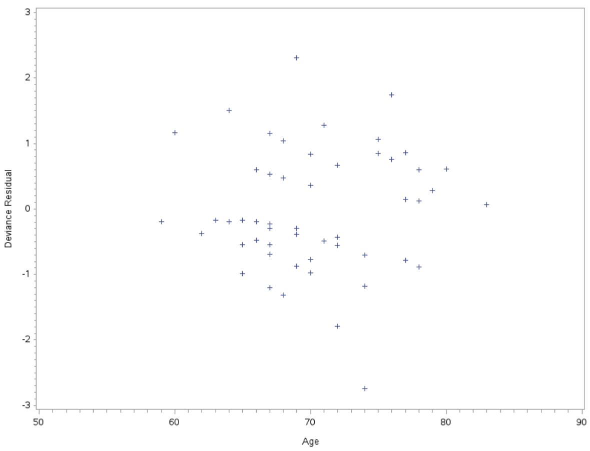

Figure 11. Plot of Residual vs. Age

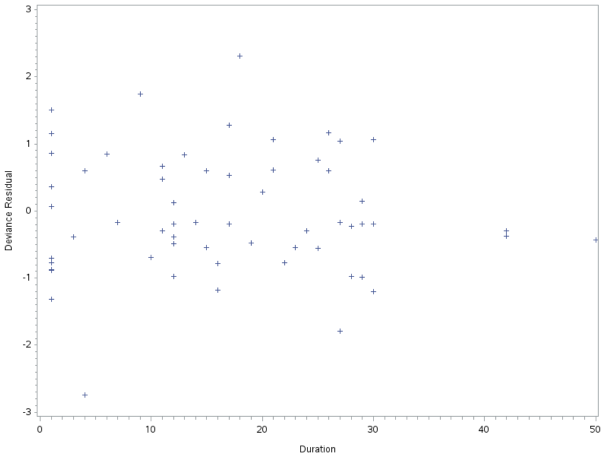

Figure 12. Plot of Residual vs. Duration

The set of scatter plots above shows the residuals plotted against each explanatory variable. These plots are used to identify any patterns or trends in the residuals that may be associated with the values of each explanatory variable. In an ideal scenario, the residuals should be randomly dispersed around zero for all levels of each explanatory variable.

The residual vs. Treatment plot (figure 9) demonstrates that there is no significant pattern in the residuals across the different levels of Treatment, indicating that Treatment is a reliable predictor of Pain. The residual vs. Sex plot (figure 10) displays a similar pattern as Treatment, suggesting that Sex is also a good predictor for Pain. The residual vs. Age plot (figure 11) reveals that the residuals are randomly dispersed around the zero line, implying that Age is also a reliable predictor of Pain. Additionally, the residual vs. Duration plot (figure 12) shows that the residuals are randomly dispersed around the zero line, implying that Duration is also a good predictor for Pain.

Overall, there is no clear evidence of any patterns or trends in the residuals for any of the explanatory variables, indicating that the model is an appropriate fit for the data.

7. Conclusion

Overall, the relationship between Pain and Age is consistent between the simple logistic regression and local regression analyses. However, for Duration, the trends observed in the two analyses are not consistent.

Based on the analysis above, the logistic regression model with all explanatory variables is the optimizal model for predicting pain in patients with neuralgia. The backward elimination procedure revealed that Treatment B, Sex F, and Age are significant predictors of pain, while Treatment A and Duration are not. The Hosmer-Lemeshow goodness-of-fit test indicated that the model fits the data well and is not significantly different from the observed data. The boxplot and ROC curves suggest that the logistic regression model has strong predictive power and better overall performance in correctly classifying the outcome variable (Pain) compared to the models with only Age or Duration.

The model diagnostics include three plots to evaluate the appropriateness of the model assumptions for the data. The index plot and the scatter plot of residuals vs. predicted probability both suggest that the model is a good fit for the data, with no unusual patterns or systematic errors in the residuals. The scatter plots of residuals vs. each explanatory variable also show that there are no clear patterns or trends in the residuals, indicating that the model assumptions are appropriate for the data. Based on the model diagnostics, it can be concluded that the model assumptions are appropriate for the data. Therefore, the logistic regression model with all explanatory variables is a good parsimonious model with good predictive power and meets the assumptions of logistic regression.