# Model 1: Baseline Linear Regression

The first model employs a basic linear regression using selected features, with categorical variables (gender and subspecies) converted via one-hot encoding. This serves as a starting point to establish baseline performance.

The model achieves a Train R² of 0.508 and a Test R² of 0.381, meaning it captures about 50.8% of the variance in the training data but drops to only 38.1% on the test set. This discrepancy suggests moderate overfitting, where the model performs significantly better on training data than unseen data.

The Train MSE is 452.431, while the Test MSE is 370.73. Interestingly, the test error is slightly lower than the training error, which may be due to sampling variation or the presence of outliers in the training set. Nevertheless, both MSE values are relatively high, indicating considerable prediction error overall.

Residual analysis supports these findings. The train residuals vs. predicted plot reveals a strong curve, suggesting that the model cannot capture nonlinear trends in the data.

The best combination of features found was 'WHT', 'AGE', 'DSI'.

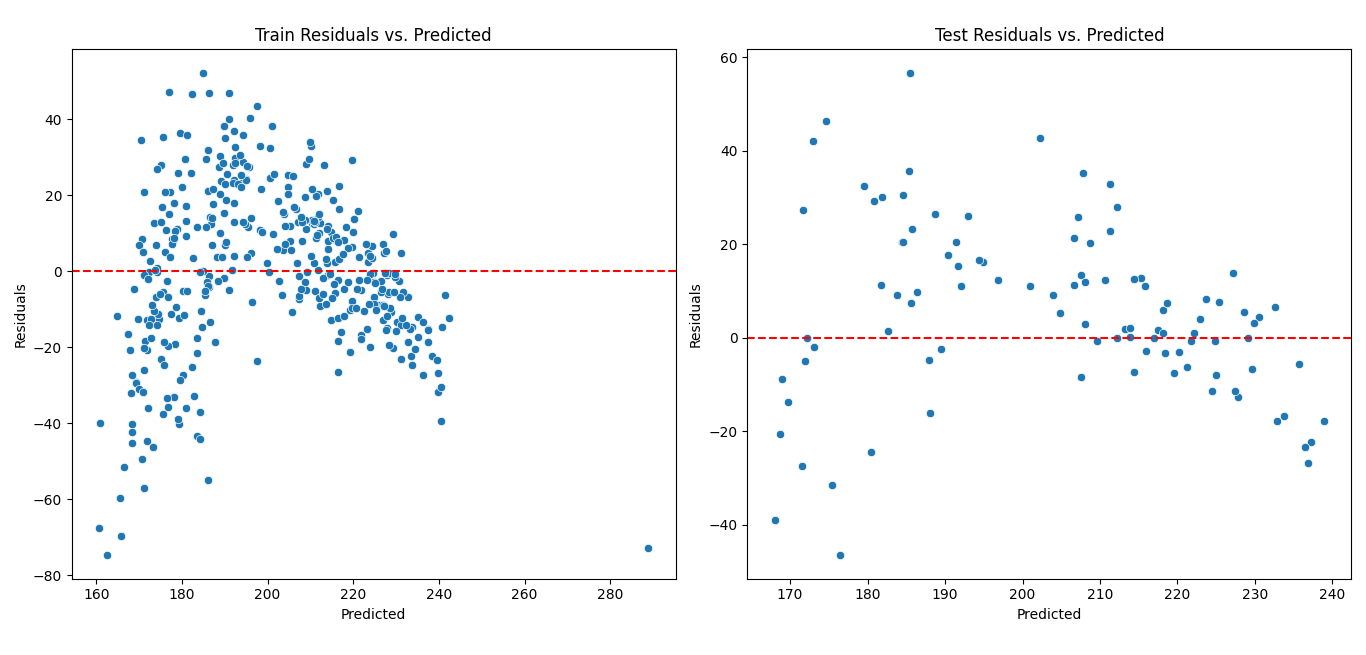
*Summary of Results:*

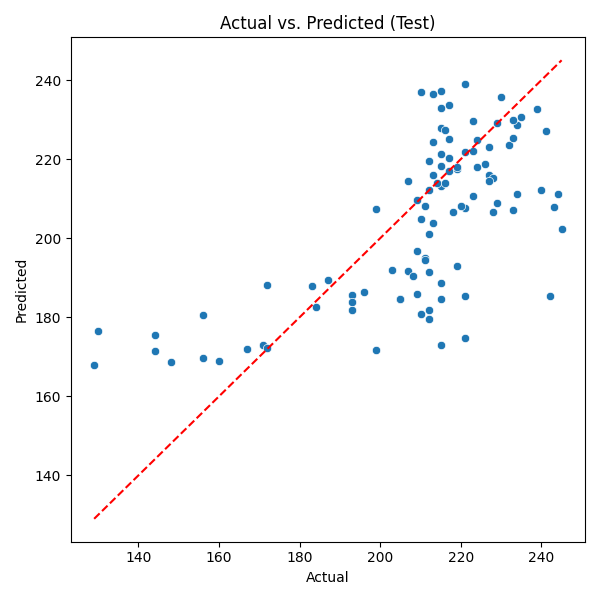
Train R²: 0.508

Test R²: 0.381

Train MSE: 452.431

Test MSE: 370.73



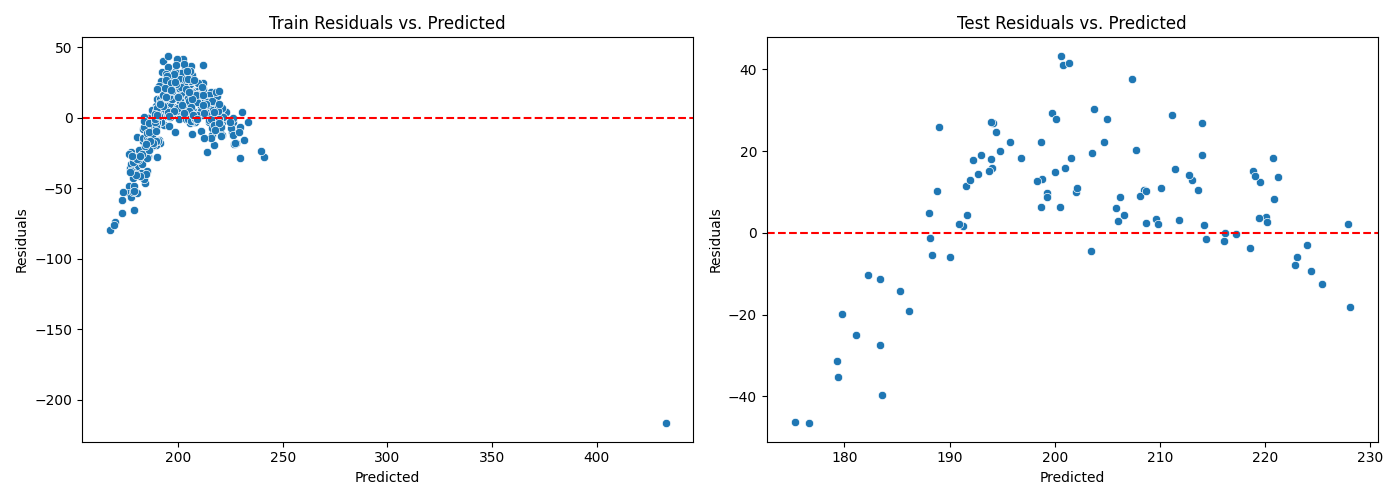


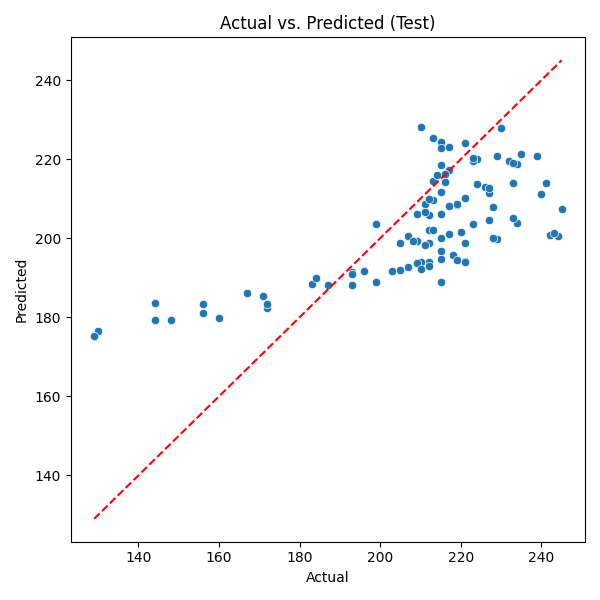
*‘WHT’,’AGE’,’DSI’ Results:*

Train R²: 0.361

Test R²: 0.407

Train MSE: 588.157

Test MSE: 355.145



# Model 2: Preprocessing

Building upon the first model, the second version integrates key preprocessing techniques. First, outliers are removed by filtering observations with z-scores greater than 3, helping reduce noise in the training data. Second, MinMax scaling is applied to standardize the feature range, ensuring that all numerical variables contribute equally to model learning.

These enhancements lead to a Train R² of 0.567 and a Test R² of 0.468, showing improved generalization and reduced overfitting compared to the first model. The gap between training and test performance narrows, and both R² values are higher.

The Train MSE drops to 355.956, and Test MSE improves to 318.552, reflecting lower average prediction errors. The residual plots show a similar curvature as before, indicating that linear regression still struggles with underlying non-linear relationships. However, the residuals are more evenly spread around zero, suggesting better model consistency.

The Actual vs. Predicted plot shows points more closely aligned along the diagonal, confirming improved predictive alignment.

The best combination of features found was 'WHT', 'DSI', "SUS\_Grauer's", 'GND\_Male'.

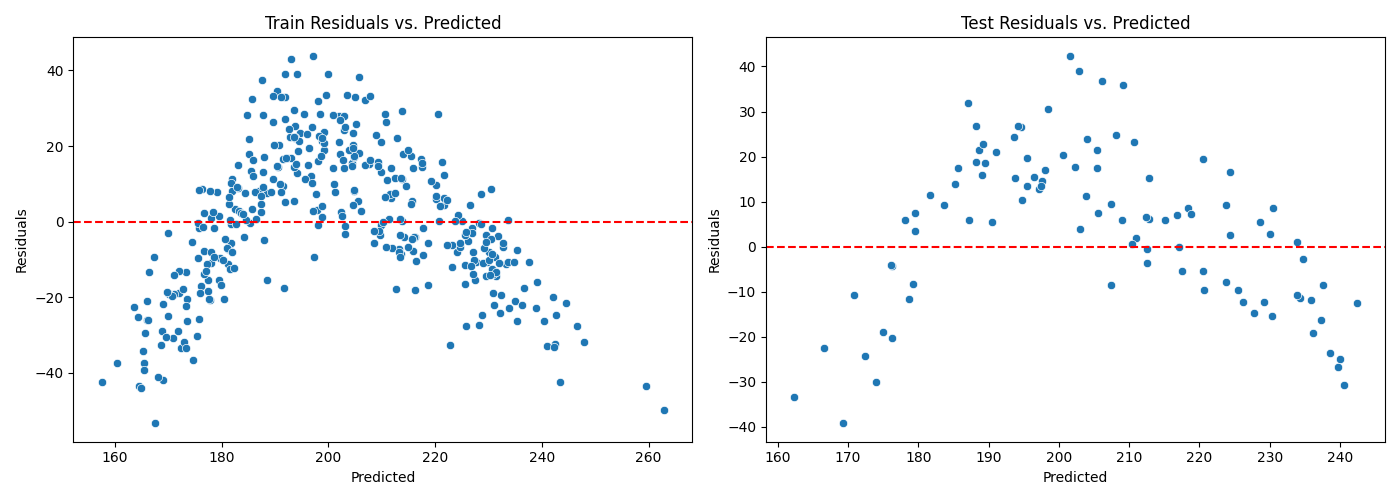
*Summary of Results:*

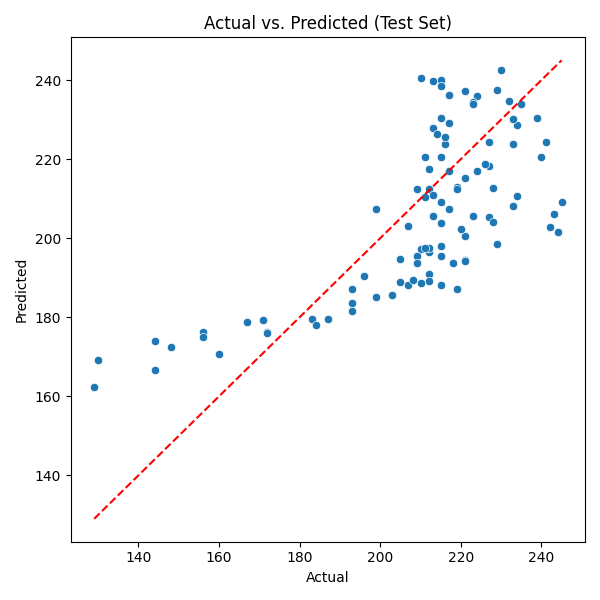
Train R²: 0.567

Test R²: 0.468

Train MSE: 355.956

Test MSE: 318.552





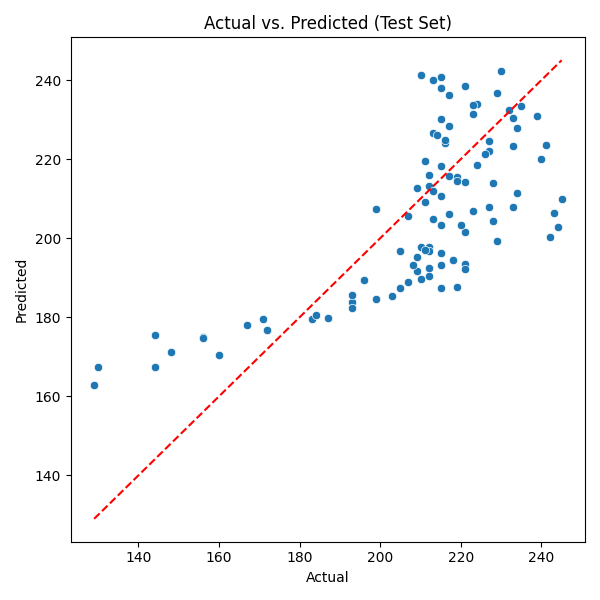
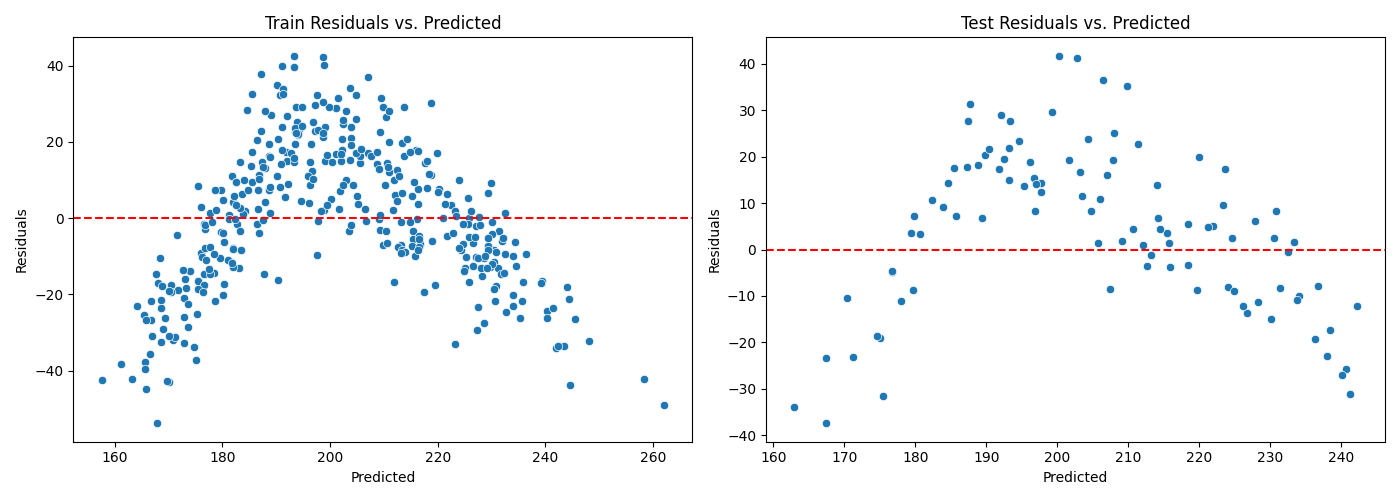
*'WHT', 'DSI', "SUS\_Grauer's", 'GND\_Male' Results:*

Train R²: 0.565

Test R²: 0.477

Train MSE: 357.274

Test MSE: 313.43



# Model 3: Polynomial Features and Lasso Regularization

To address the model's inability to capture non-linear relationships, the third model introduces polynomial features. This allows it to learn interactions and curved patterns in the data. Lasso regression is employed to regularize the model and prevent overfitting. The best alpha value (penalty strength) is selected via cross-validation. Feature selection is then performed using SelectFromModel to retain only impactful features.

The result is a significant improvement. The model achieves a Train R² of 0.869 and a Test R² of 0.791, capturing a much larger proportion of the variance. The Train MSE is reduced to 107.517, and the Test MSE is 125.129, reflecting a good balance between fit and generalization.

The residual plots show a mostly random distribution around zero for both sets. The Actual vs. Predicted plot shows tight alignment with the ideal diagonal, especially in the upper value range.

*Summary of Results:*

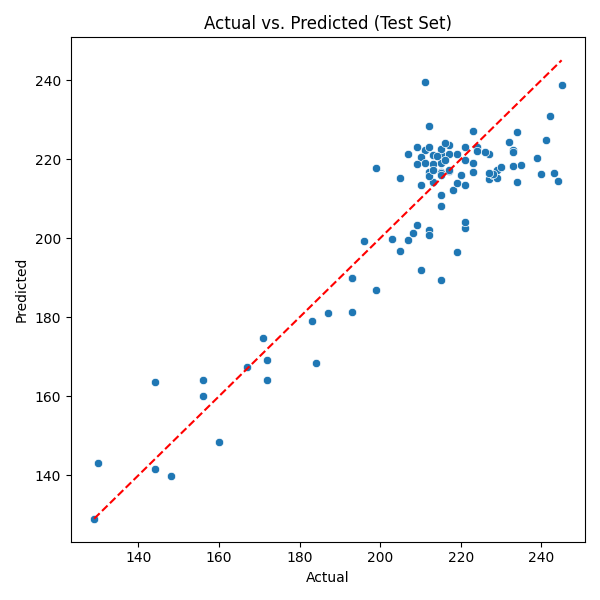
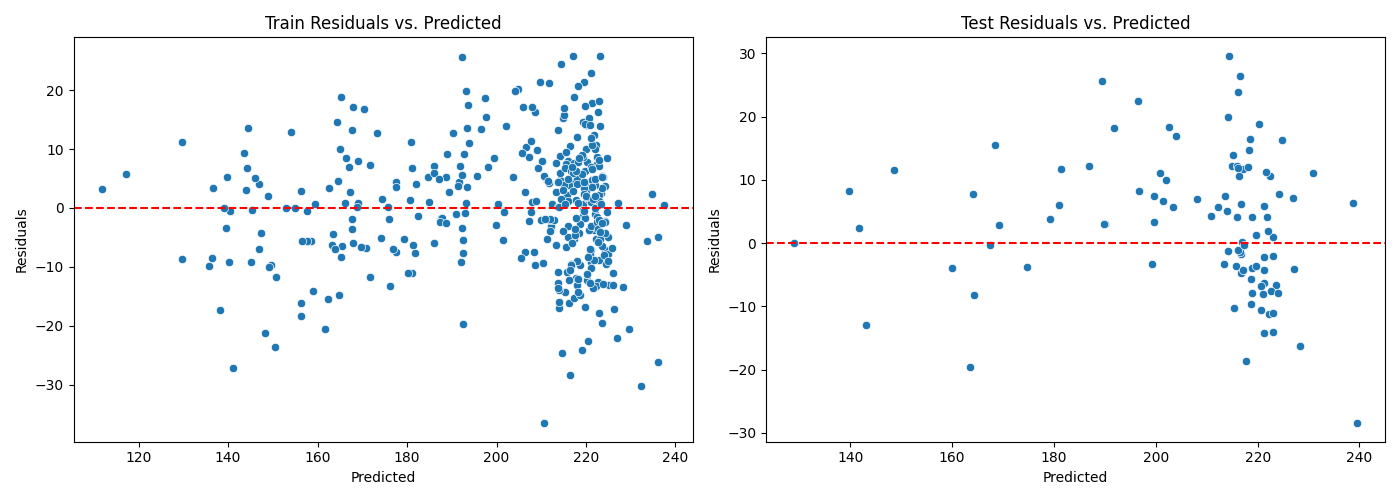
Train R²: 0.869

Test R²: 0.791

Train MSE: 107.517

Test MSE: 125.129

Best alpha: 0.01



# Model 4: Automated Lasso with Feature Selection

The final model refines the third by automating the alpha tuning process using LassoCV, which selects the best penalty term via cross-validation. The model then uses SelectFromModel again to retain only the most meaningful features. This process reduces the number of features from 54 to 32, simplifying the model without reducing its accuracy.

Performance remains stable: the Train R² is 0.869 and Test R² is 0.791, identical to the previous model. The Train MSE slightly increases to 108.044, while Test MSE remains at 125.129, confirming that generalization has not been compromised by reducing the feature set.

Visually, the residuals remain well-distributed with no major patterns, and the Actual vs. Predicted plot still shows strong agreement with the reference line. The reduced feature set enhances model interpretability and robustness, making this version more efficient and elegant.

*Summary of Results:*

Best alpha (LassoCV): 0.0117

Features before selection: 54

Features after selection: 32

Train R²: 0.869

Test R²: 0.791

Train MSE: 108.044

Test MSE: 125.129

