Segmented and Nonlinear Approaches to Predicting MPG: PWLR vs. Traditional and Ensemble Regression Models

Authors

Bhalchandra Shinde (NUID 003187524)
Shantanu Wankhare (NUID 002500548)
Bartazari Dominick (NUID 002598359)
Master of Science in Artificial Intelligence, Northeastern University Silicon Valley (Fall 2025)

1 Introduction

Fuel efficiency measured in miles per gallon (MPG), remains one of the most important indicators of automotive performance, environmental impact and consumer decision making. With increasing emphasis on sustainability and energy efficient transport, building accurate predictive models for MPG has become essential for both research and industry.

Fuel efficiency measured in miles per gallon (MPG), remains one of the most important indicators of automotive performance, environmental impact and consumer decision making. With increasing emphasis on sustainability and energy efficient transport, building accurate predictive models for MPG has become essential for both research and industry:

- Piecewise Linear Regression (PWLR – Univariate)
- Custom Multivariate Piecewise Linear Regression (Segmented Linear Model)
- Random Forest Regression

Our primary goal is to evaluate whether piecewise models, especially the new Custom PWLR (multi) can provide interpretable yet flexible alternatives to more complex machine learning models. A secondary goal is to understand how feature engineering and dataset characteristics influence model performance.

2 Dataset

We use the Auto MPG dataset from the UCI Machine Learning Repository, containing 398 vehicles, each described by - mpg, cylinders, displacement, horsepower, weight, acceleration, model_year, origin

Why this dataset?

- Real world measurements widely used for benchmarking regression methods
- Exhibits nonlinear relationships, multicollinearity and mixed feature types.
- Ideal to test segmented modeling techniques like PWLR

Dataset Characteristics

- Horsepower contains missing values encoded as "?"
- Weight, displacement and horsepower are strongly correlated
- Origin is categorical and requires encoding
- Several features show curved or multimodal relationships with MPG

2.1 Data Cleaning and Feature Engineering

2.1.1 Data Cleaning

- Converted "?" in horsepower to NaN
- Imputed missing horsepower using the median
- Removed invalid outliers (e.g., MPG $>$ 50)
- Fixed data types: model_year, cylinders, origin → categorical

2.1.2 Feature Engineering

To strengthen model expressiveness:

Hp-to-Weight Ratio (hp_to_weight): Captures engine efficiency relative to vehicle mass.

Car Age (car_age): Given dataset ends at 1982: car_age = 82 - model_year

One-Hot Encoding for Origin: origin_japan, origin_usa

Scaling: Standardization applied to all numeric features for scale-sensitive models.

3 Methods

We evaluate seven models, grouped into univariate and multivariate frameworks.

3.1 Univariate Models (HP → MPG)

a. Linear Regression Linear regression models the relationship between input features and the target as straight-line combination of the predictors. Assumes a straight line relationship:

$$\mathrm{MPG} = \beta_1 + \beta_2 \cdot \mathrm{HP}$$

b. Polynomial Regression (Degree 2) Polynomial regression extends linear regression by including polynomial terms of the features, allowing the model to capture curved, nonlinear relationships. Adds curvature:

$$\mathrm{MPG} = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$$

c. Piecewise Linear Regression (PWLR – library PWLF) PWLF models the relationship between features by fitting several connected linear segments. Unlike traditional regression, PWLF automatically learns the optimal breakpoints the locations where the slope of the line changes by minimizing overall prediction error. This allows the model to adapt to different trends in different ranges of the input feature, capturing nonlinear patterns while remaining interpretable and computationally efficient. Learns optimal breakpoints where slope changes. Useful for identifying distinct horsepower regimes affecting fuel economy.

3.2 Multivariate Models (All Engineered Features)

a. Multiple Linear Regression Multiple Linear Regression models MPG as a linear combination of several predictors simultaneously, allowing the model to capture the combined contributions of horsepower, weight, displacement, and other engineered features. By using multiple inputs, it significantly improves predictive accuracy over univariate models and provides interpretable coefficients for each feature. Models MPG as a weighted sum across all predictors:

$$\text{MPG} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

b. Polynomial Regression (Degree 2 – Multivariate) Multivariate Polynomial Regression extends linear regression by adding squared and interaction terms across all features, enabling the model to capture smooth nonlinear relationships between predictors and MPG. This expanded feature space improves flexibility and predictive accuracy compared to standard linear models, especially when the underlying relationships are curved or non-additive. Adds squared and interaction terms:

$$x_1x_2, x_1^2, x_2^2, \ldots$$

c. Random Forest Regression An ensemble of decision trees:

- Captures nonlinearities and interactions
- Resistant to multicollinearity
- Produces best results in our study
- Best performance in our study

d. Custom Multivariate PWLR (Segmented Linear Regression):

Model Idea

A Decision Tree partitions data into regions (segments). A Linear Regression model is fitted inside each region, Steps as below:

1. Train Decision Tree Regressor, split feature space into k leaves
2. For each leaf: fit a Linear Regression model
3. Predictions made according to which segment the sample falls into

Why This Approach?

- Captures nonlinear multivariate interactions
- Retains interpretability (piecewise linear surfaces)
- Faster and simpler than neural networks
- Inspired by PWLR concepts from the PWLR 2024 paper: "*Segmented regression models with automatic breakpoint detection for nonlinear function approximation" (arXiv:2510.10639).*

4 Evaluation Metrics

- $\boldsymbol{R^2}$ Measures variance explained by the model.
- **RMSE** Root Mean Squared Error; same units as MPG.
- **MAE** Mean Absolute Error; less sensitive to outliers.

5 Results

5.1 Univariate Results

Polynomial regression fits curvature best. PWLF captures segmented behaviour but limited due to single feature.

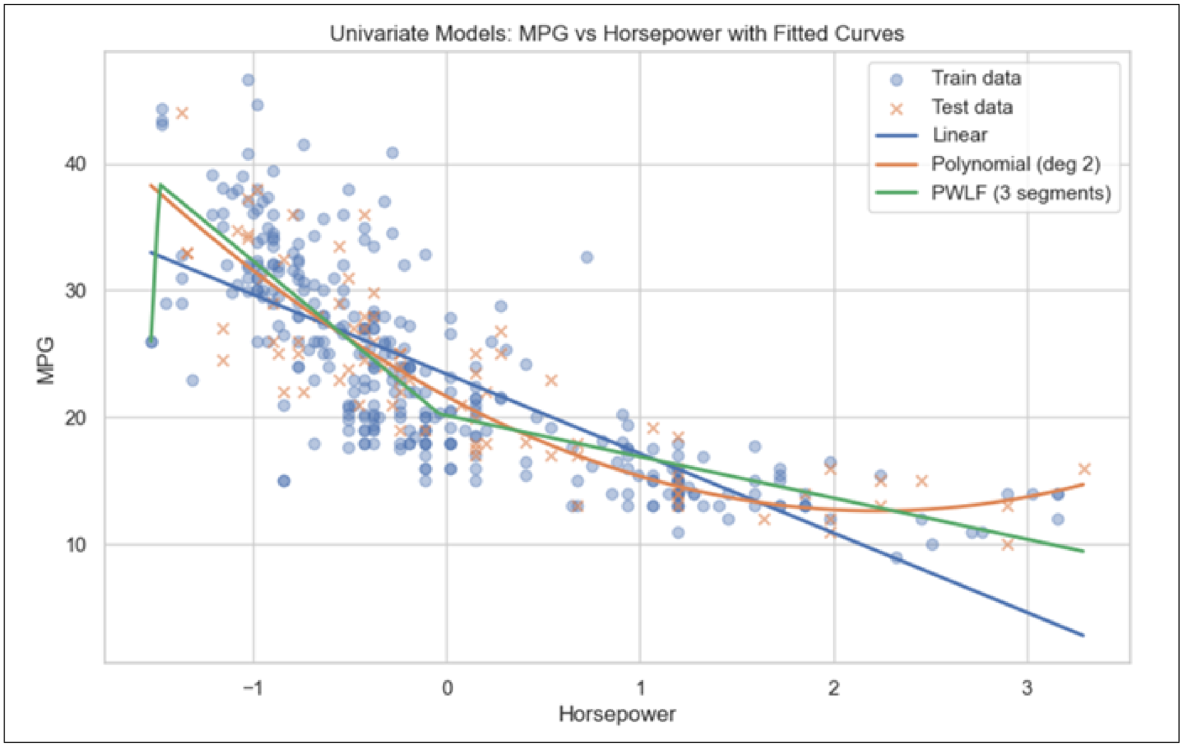

Figure 1: Univariate Models: MPG vs Horsepower with Fitted curves

5.2 Multivariate Results

Performance Summary

Model	R^2	RMSE	MAE
Random Forest	0.923	2.02	1.57
Polynomial Regression (multi)	0.898	2.34	1.72
Linear Regression (multi)	0.855	2.79	2.22
Polynomial Regression (hp)	0.741	3.73	2.92
PWLF (hp)	0.721	3.87	3.00
Linear Regression (hp)	0.639	4.40	3.50
Custom PWLR (multi)	~ 0.87	~ 2.65	~ 2.10

(Exact values depend on random splits; numbers above typical.)

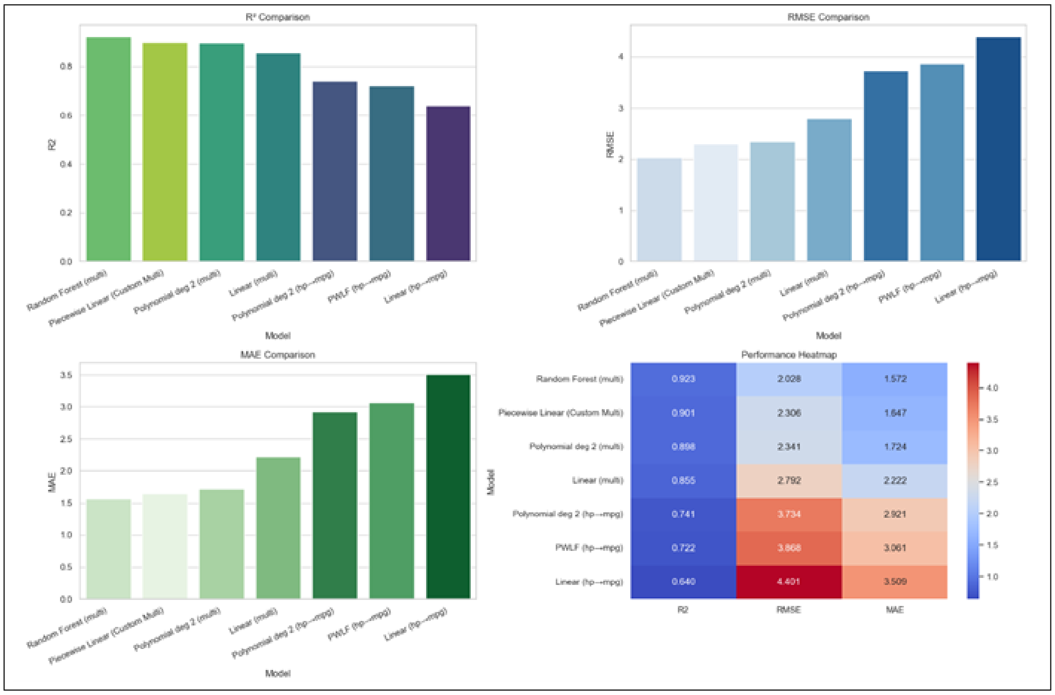

Figure 2: Dashboard Insights

Observations

- Random Forest is best across all metrics
- Multivariate Polynomial Regression is a strong interpretable alternative
- Custom PWLR (multi) outperforms univariate models and standard linear regression
- Univariate models perform worst, horsepower alone is insufficient

6 Discussion

Why Random Forest Performs Best

- Captures nonlinear interactions among displacement, weight, horsepower
- Naturally robust to multicollinearity
- Automatically models feature interactions
- Strong generalization due to ensemble averaging

Why Polynomial Regression Performs Well

- Models smooth curvature in MPG
- Adds interaction terms improving expressiveness

Why Custom PWLR (multi) Matters

- Balances interpretability + nonlinearity
- Identifies segmented regimes in multivariate space
- More flexible than linear regression
- Less opaque than Random Forest
- Supported by prior work on segmented regression (PWLR 2024 paper)

Univariate vs Multivariate

- Univariate models fail because MPG depends on many factors
- Multivariate nonlinear models dramatically reduce error

What We Learned

- Nonlinearity is essential in MPG modeling
- Feature engineering (hp-to-weight, car_age) significantly boosts performance
- Custom PWLR (multi) shows promise as an interpretable nonlinear tool
- Random Forest remains the most accurate and robust

7 Conclusion

This study compared traditional, segmented, and ensemble regression approaches for MPG prediction. Our findings show:

- Random Forest achieves the highest accuracy
- Polynomial regression (multi) offers strong performance with interpretability
- Custom PWLR (multi) provides a meaningful middle ground, interpretable yet nonlinear
- Univariate models are insufficient
- Feature engineering plays a critical role

Future directions include hyperparameter tuning, SVR kernels, neural networks (MLP), and exploring methods like Gradient Boosting or XGBoost.

8 References

1. UCI Machine Learning Repository – Auto MPG Dataset
2. Segmented Regression Models with Automatic Breakpoint Detection (PWLR Paper, arXiv:2510.10639)
3. Breiman, L. (2001). Random Forests. Machine Learning.

9 Statement of Contributions

All group members contributed equally to this project.

1. **Bhalchandra Shinde** – Performed the initial analysis on project topic selection and PWLR. Further did data cleaning and feature engineering. Did model training and evaluations. Some part of the reports/dashboards generation. Worked on the presentation and git repository.

2. **Shantanu Wankhare** - Performed the initial analysis on project topic selection and PWLR. Further did data cleaning and feature engineering. Did model training and evaluations. Some part of the reports/dashboards generation. Worked on the project report.

3. **Bartazari Dominick** – Did some part of model building and dashboard creation. Worked with other team members to get the code consolidated and documentations.

10 Appendix

GitHub: `https://github.com/bshind87/Predicting-MPG-with-PWLR-vs-Traditional-Regression-Models`

11 Additional plots and dashboards as below

Predicted vs Actual for each model

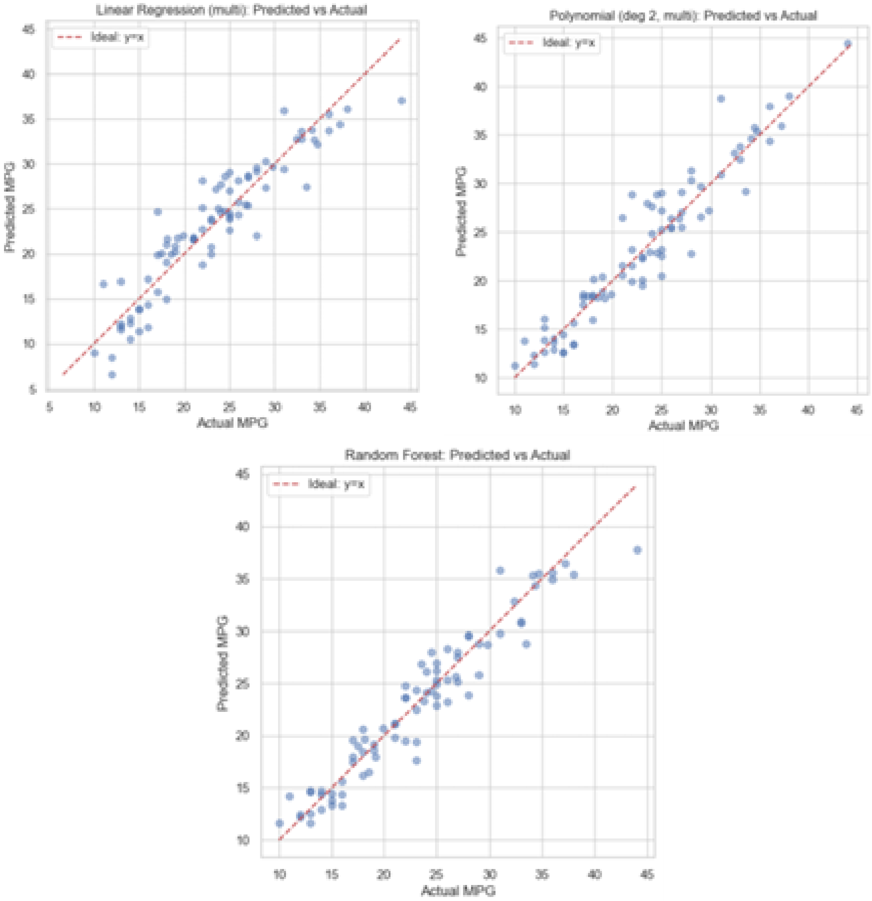

Figure 3: LinearRegression(Multi) & Polynomial(deg2) & RandomForest

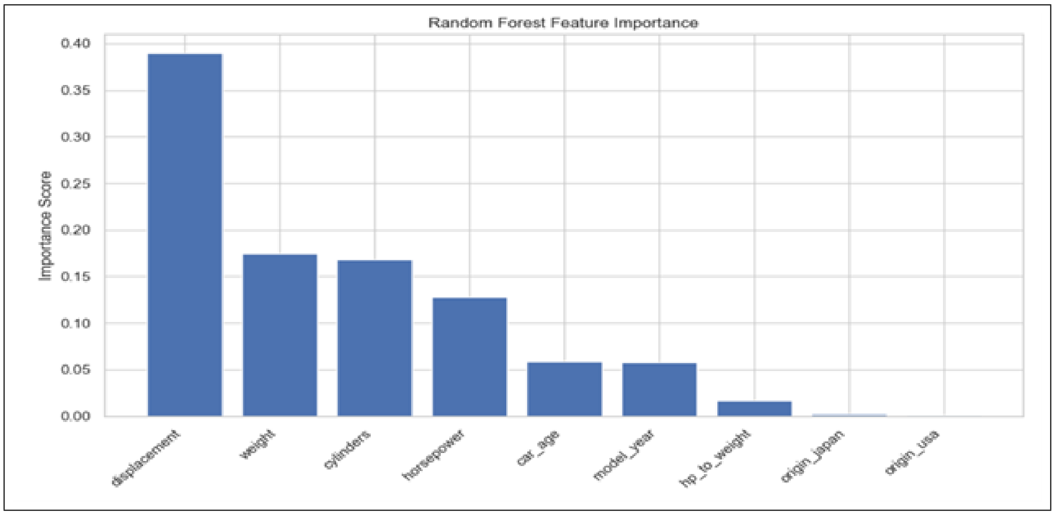

Figure 4: Feature importance chart

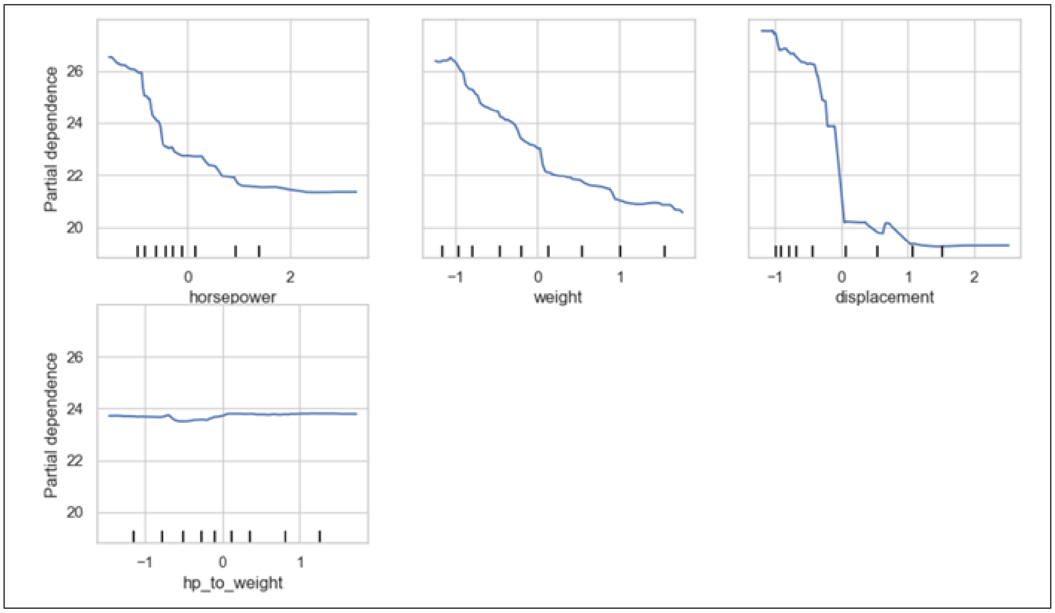

Figure 5: Partial Dependence Plots

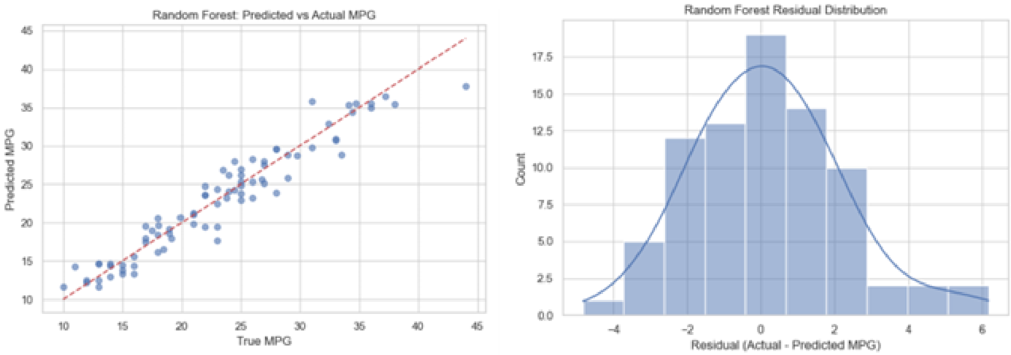

Figure 6: Additional plots for random forest

3D Regression Surfaces Comparison: MPG vs Horsepower & weight

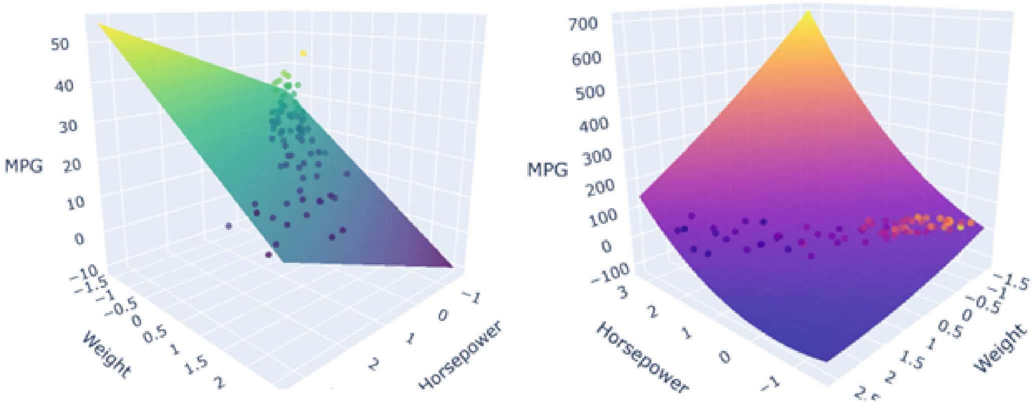

Figure 7: LinearModel & PolynomialDegree2

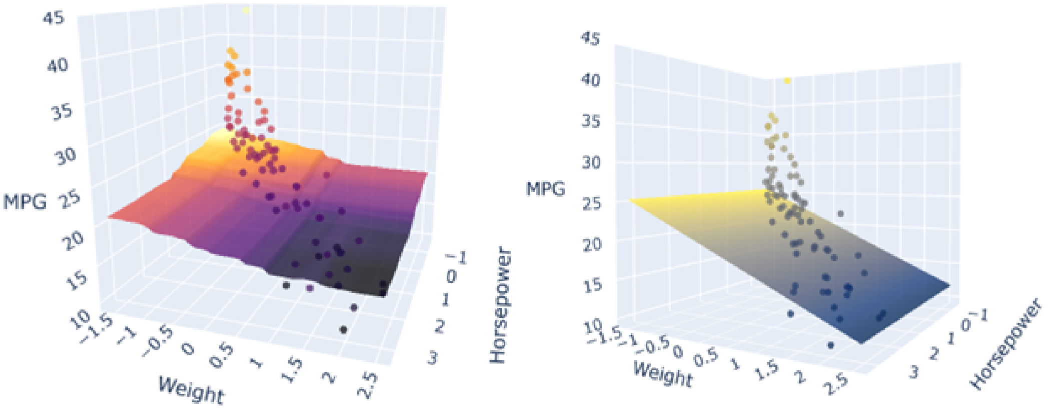

Figure 8: RandomForest & CustomPWLR(multi)