**Subset Selection Techniques Using the Hitters Dataset**

**Best Subset Selection**

1. Loading and Preparing Data:

- The `Hitters` dataset is loaded from a CSV file.

- Missing values in the `Salary` column are removed using `na.omit()`, which drops rows with any missing values.

2. Performing Best Subset Selection:

- The `regsubsets()` function from the `leaps` package is used to perform best subset selection. This method identifies the best subset of predictors for each possible model size (i.e., number of predictors).

- The model is fitted with `nvmax = 19`, which considers all possible subsets up to 19 predictors.

3. Analyzing Results:

- The `summary()` function extracts important statistics from the model, including R-squared (`rsq`), adjusted R-squared (`adjr2`), Cp statistic (`cp`), and Bayesian Information Criterion (`bic`).

- Various plots are created to visualize these statistics:

- RSS Plot: Shows how the residual sum of squares (RSS) changes with the number of variables.

- Adjusted R-Squared Plot: Highlights the model with the highest adjusted R-squared, indicating the best trade-off between model complexity and fit.

- Cp Plot: Shows the Cp statistic, which helps in identifying models with minimal prediction error.

- BIC Plot: Uses BIC to penalize models for complexity, favoring simpler models.

- The model with 11 variables is highlighted as the one with the highest adjusted R-squared. Models with 10 variables (Cp) and 6 variables (BIC) are also identified as optimal based on these criteria.

4. Plotting Best Models:

- `plot(regfit.full, scale = "r2")` and similar functions visualize the best models according to different criteria like R-squared, adjusted R-squared, Cp, and BIC.

5. Extracting Coefficients:

- The coefficients of the best model with 6 predictors (based on BIC) are extracted using `coef(regfit.full, 6)`.

**Forward and Backward Stepwise Selection**

1. Forward Stepwise Selection:

- The `regsubsets()` function is applied again, but this time using the `method = "forward"` argument. Forward selection starts with no predictors and adds them one by one, choosing the one that provides the best improvement at each step.

2. Backward Stepwise Selection:

- Similarly, `regsubsets()` is used with `method = "backward"`. Backward selection starts with all predictors and removes them one by one, dropping the least significant predictor at each step.

3. Comparing Models:

- The coefficients of the best 7-variable models obtained from best subset selection, forward stepwise selection, and backward stepwise selection are compared.

**Choosing Among Models Using the Validation-Set Approach and Cross-Validation**

1. Validation Set Approach:

- The dataset is split into training and test sets using random sampling (`train` and `test`).

- A best subset selection model is fit to the training data.

- The model’s performance is evaluated on the test set by computing the mean squared error (MSE) for different model sizes (1 to 19 variables).

- The best model is identified based on the lowest MSE.

2. Cross-Validation:

- 10-fold cross-validation is implemented to evaluate model performance.

- The dataset is split into 10 parts (`folds`). For each part, the model is trained on the other 9 parts and tested on the remaining one.

- This process is repeated for each fold, and the MSE is averaged across all folds to determine the best model.

- A plot of the cross-validation errors helps in identifying the optimal number of predictors.

3. Final Model Selection:

- The final best model with 10 predictors is chosen based on cross-validation, and the coefficients of this model are extracted.

**Summary**

- The code covers three key methods of subset selection: Best Subset Selection, Forward Stepwise Selection, and Backward Stepwise Selection.

- It evaluates models using different criteria like Adjusted R-Squared, Cp, BIC, and Cross-Validation

- The final selection is based on cross-validation, which provides a robust estimate of model performance on unseen data.

This workflow is a comprehensive example of model selection in regression, providing insights into choosing the most suitable set of predictors for a given problem.