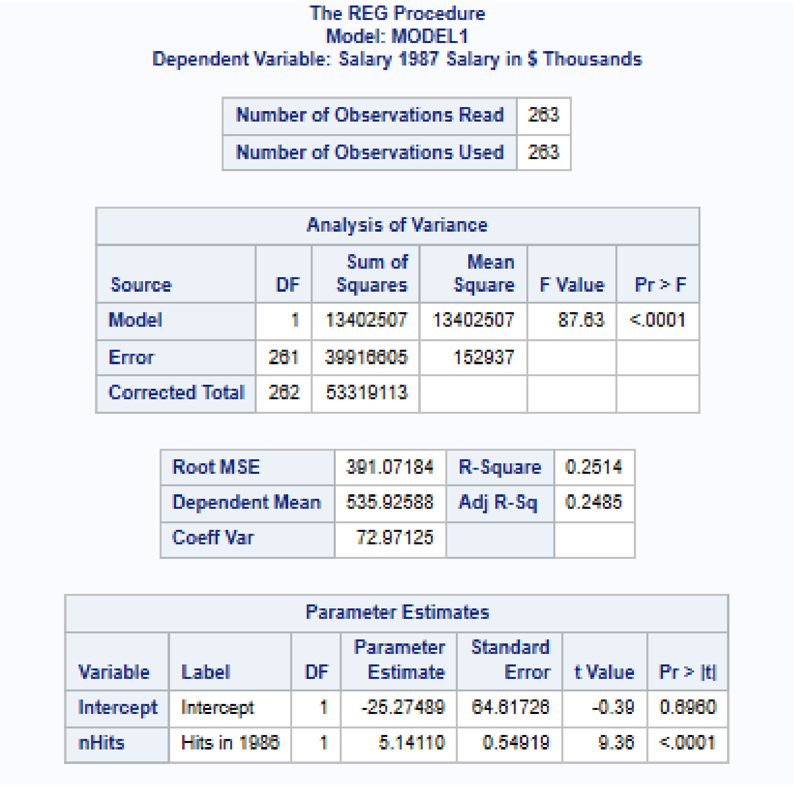
Aryan Sharma
SAS Assignment

Baseball Player Salary Modeling

Use SAS data set on Baseball, this data was briefly explored during the module for this unit, but you will explore it some more in this assignment. This data is in the SASHELP library and can be accessed and manipulated by creating a new data set and using the SET statement to set it equal to the Baseball data set. Note: When using the SET statement make sure you indicate that the Baseball data set is named SASHELP.Baseball so that SAS properly understands the reference.

Part A) Using one of the variables listed identify the regression model that explains the largest percentage of variation amongst the data for salary. Interpret the p-value and correlation coefficient of your model in context. Does this model violate any assumptions? Does this model appear valid? Justify your response

Result:



The REG Procedure
Model: MODEL1
Dependent Variable: Salary 1987 Salary in $ Thousands

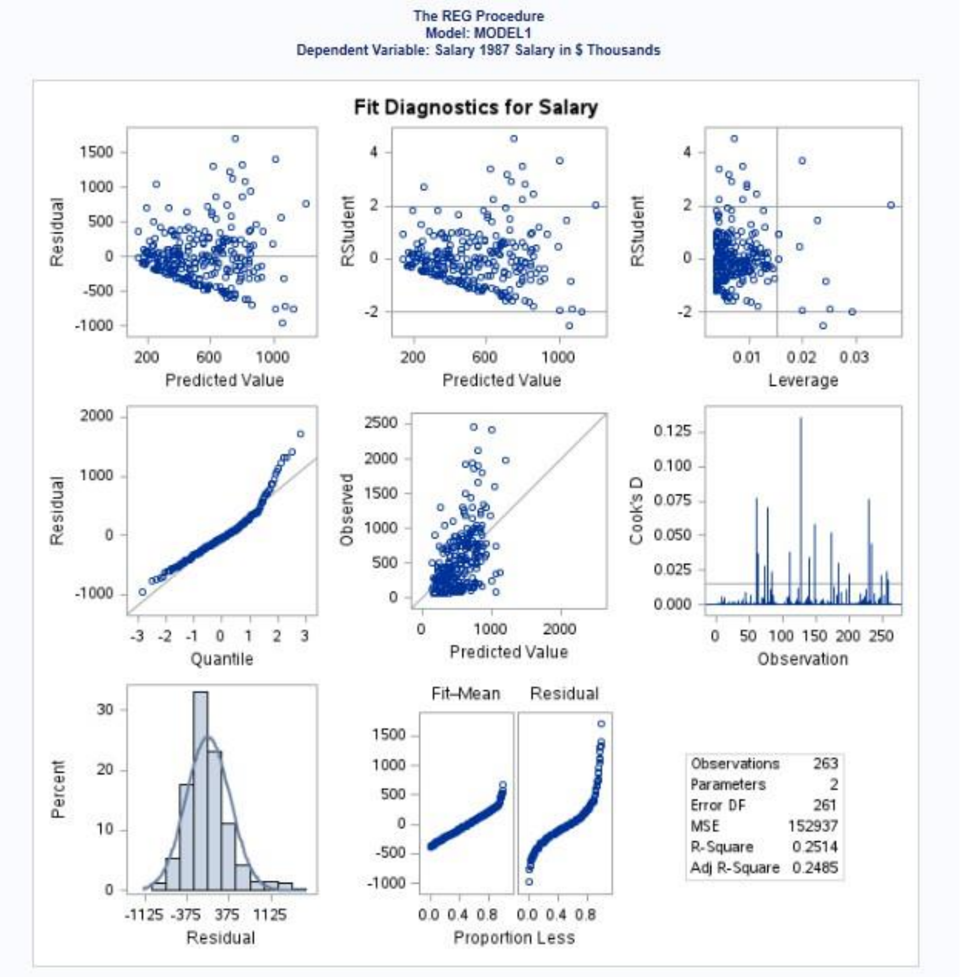
Number of Observations Read	263
Number of Observations Used	263

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13402507	13402507	87.63	<.0001
Error	261	39916605	152937		
Corrected Total	262	53319113			

Root MSE	391.07184	R-Square	0.2514
Dependent Mean	535.92588	Adj R-Sq	0.2485
Coeff Var	72.97125		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-25.27489	64.81726	-0.39	0.6960
nHits	Hits in 1986	1	5.14110	0.54919	9.36	<.0001

Aryan Sharma
SAS Assignment

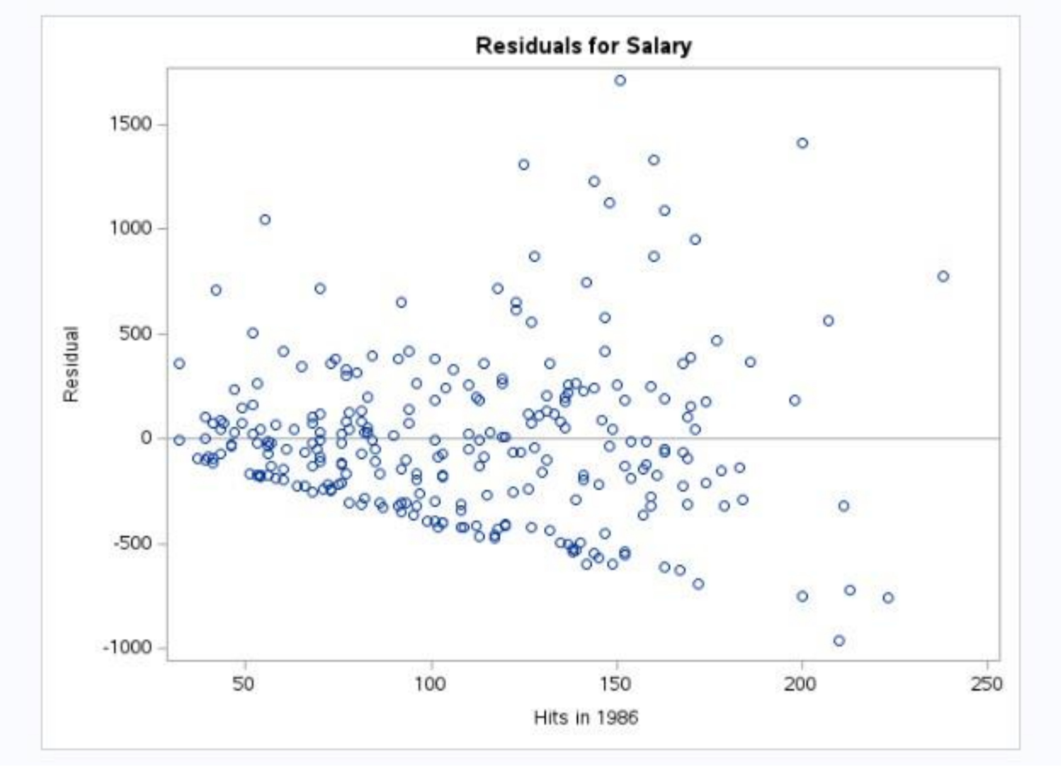


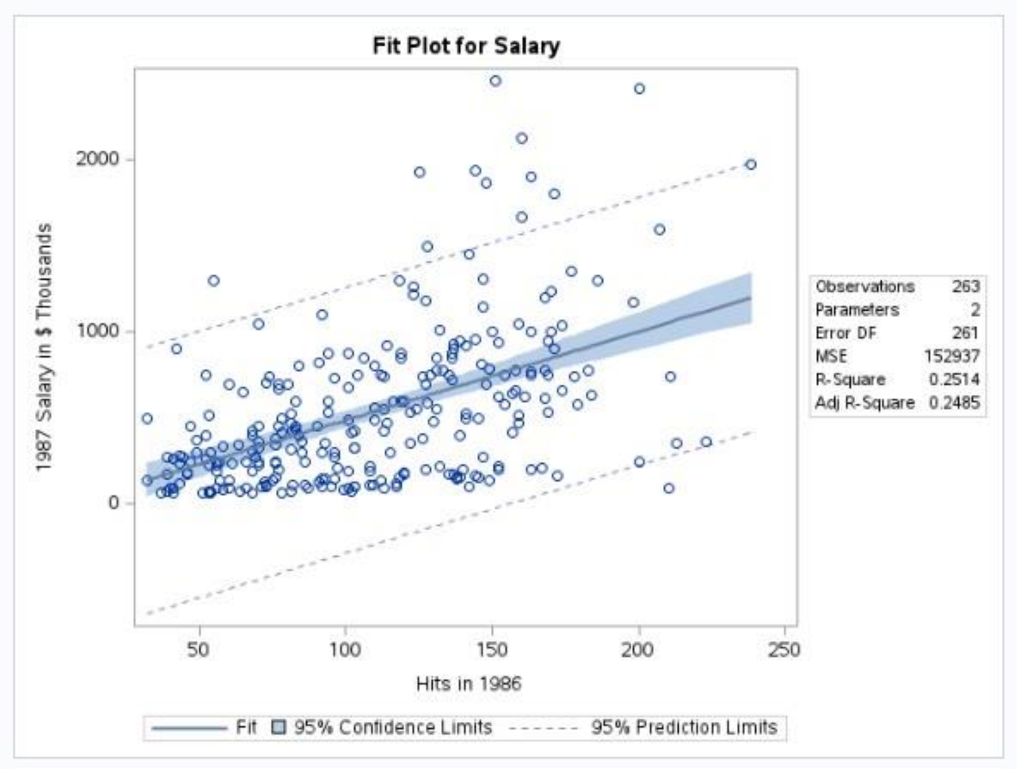
```
The REG Procedure
Model: MODEL1
Dependent Variable: Salary 1987 Salary in $ Thousands
```

Fit Diagnostics for Salary

```
Observations      263
Parameters          2
Error DF          261
MSE            152937
R-Square       0.2514
Adj R-Square   0.2485
```

Aryan Sharma
SAS Assignment





Observations	263
Parameters	2
Error DF	261
MSE	152937
R-Square	0.2514
Adj R-Square	0.2485

Aryan Sharma
SAS Assignment

Model Overview

We built a simple linear regression model to predict 1987 Salary (in $ thousands) using Hits in 1986 (nHits) as the sole predictor.

Model Equation:

Salary = -25.27 + 5.14 × nHits

Metric	Value
Observations	263
R-Square	0.2514
Adjusted R-Square	0.2485
MSE (Mean Squared Error)	152937
Root MSE	391.07
Coefficient (nHits)	5.14
P-Value (nHits)	< 0.0001

Interpretation of P-value and Correlation Coefficient

P-value for nHits: The p-value is less than 0.0001, indicating that Hits in 1986 is a statistically significant predictor of salary in 1987. We reject the null hypothesis and conclude that there is a linear relationship between hits and salary.

Correlation Coefficient (r): The square root of R-squared gives approximately r = sqrt(0.2514) ≈ 0.501. This indicates a moderate positive linear relationship between nHits and Salary. As a player achieves more hits in the 1986 season, their salary in 1987 tends to increase.

Model Assumptions and Diagnostics

1. Linearity

The scatter plot (Fit Plot for Salary) shows a linear trend between hits and salary. However, variability increases at higher values of nHits, suggesting potential heteroscedasticity.

2. Normality of Residuals

The histogram and Q-Q plot show roughly symmetric and bell-shaped residuals, with minor deviation in the tails. This assumption is reasonably met.

3. Homoscedasticity (Equal Variance of Residuals)

The residuals versus fitted values plot shows a funnel shape, where spread increases with higher fitted values. This suggests heteroscedasticity is present, violating one of the key linear regression assumptions.

4. Independence

No evidence of autocorrelation is visible from residual plots; this assumption appears to hold.

5. Influential Points

The Cook's Distance plot indicates some observations may have moderate influence but no extreme outliers dominate.

Conclusion: Model Validity

Strengths

- The model is statistically significant.
- Shows a moderate correlation.
- Provides a meaningful interpretation: on average, every additional hit in 1986 is associated with a roughly $5,140 increase in salary in 1987.

Limitations

- Low R-squared (0.2514) means that the model explains only about 25 percent of the variation in salary.
- Heteroscedasticity is present, which affects confidence in standard error estimates.

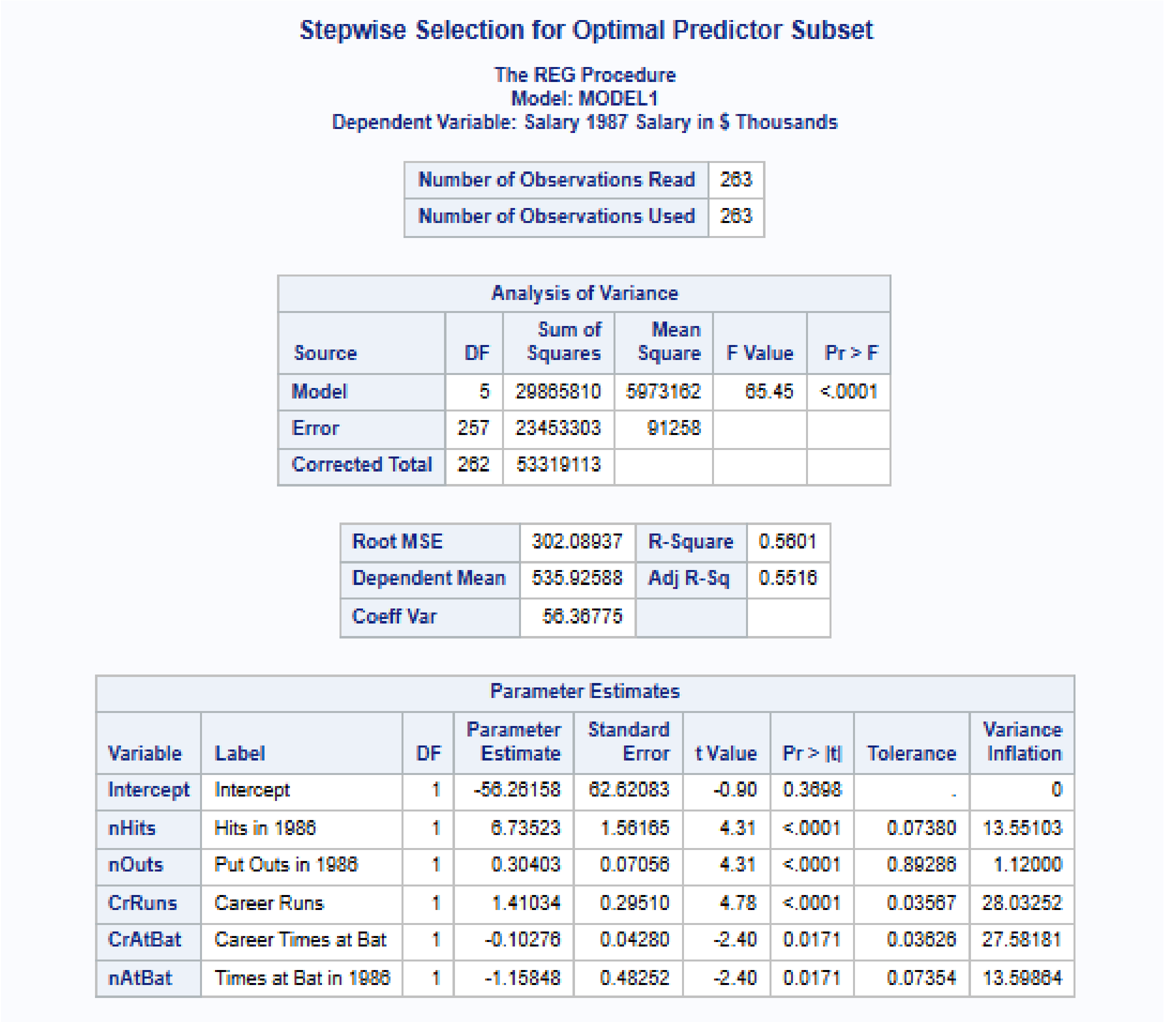
Final Verdict

The model is statistically valid for basic interpretation, but not ideal for prediction or inference without further refinement or inclusion of more variables. Consider transforming variables or using multiple regression for improved performance in future parts.

Aryan Sharma
SAS Assignment

Part B) Now selecting from all the variables listed create the best multivariable linear regression model to predict for salary. Interpret the p-value and correlation coefficients of your final model in context. Discuss your methodology for variable selection whether manual or otherwise and discuss the strength of your model. Does this model violate any assumptions? Does this model appear valid? Justify your response.

Result:



Stepwise Selection for Optimal Predictor Subset

The REG Procedure
Model: MODEL1
Dependent Variable: Salary 1987 Salary in $ Thousands

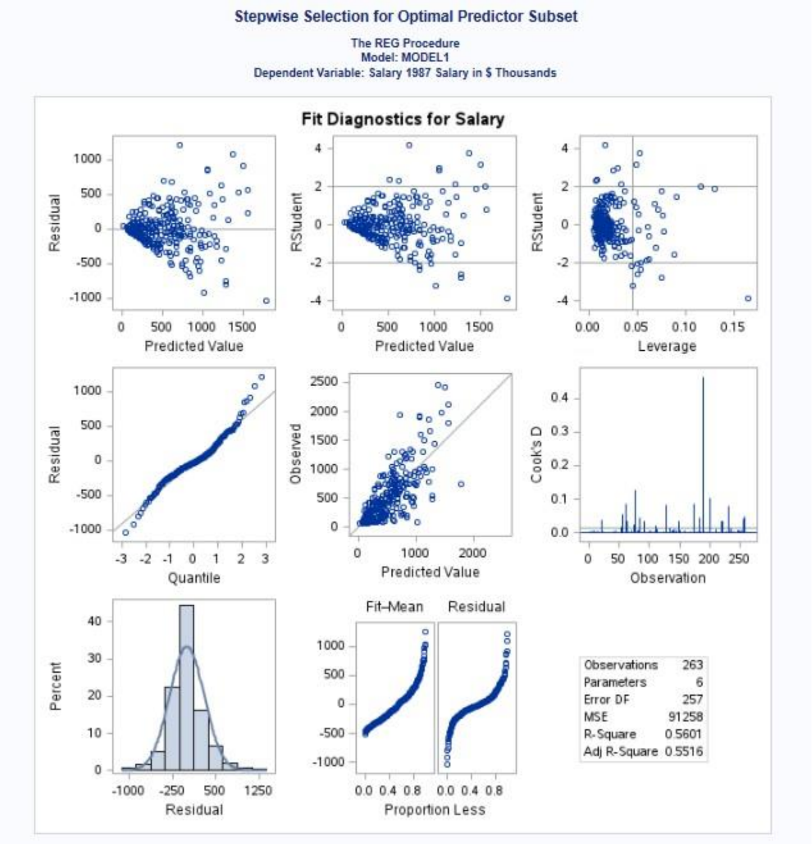
Number of Observations Read	263
Number of Observations Used	263

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	29865810	5973162	65.45	<.0001
Error	257	23453303	91258		
Corrected Total	262	53319113			

Root MSE	302.08937	R-Square	0.5601
Dependent Mean	535.92588	Adj R-Sq	0.5516
Coeff Var	56.36775		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-56.26158	62.62083	-0.90	0.3698	.	0
nHits	Hits in 1986	1	6.73523	1.56165	4.31	<.0001	0.07380	13.55103
nOuts	Put Outs in 1986	1	0.30403	0.07056	4.31	<.0001	0.89286	1.12000
CrRuns	Career Runs	1	1.41034	0.29510	4.78	<.0001	0.03567	28.03252
CrAtBat	Career Times at Bat	1	-0.10276	0.04280	-2.40	0.0171	0.03626	27.58181
nAtBat	Times at Bat in 1986	1	-1.15848	0.48252	-2.40	0.0171	0.07354	13.59864

Aryan Sharma
SAS Assignment

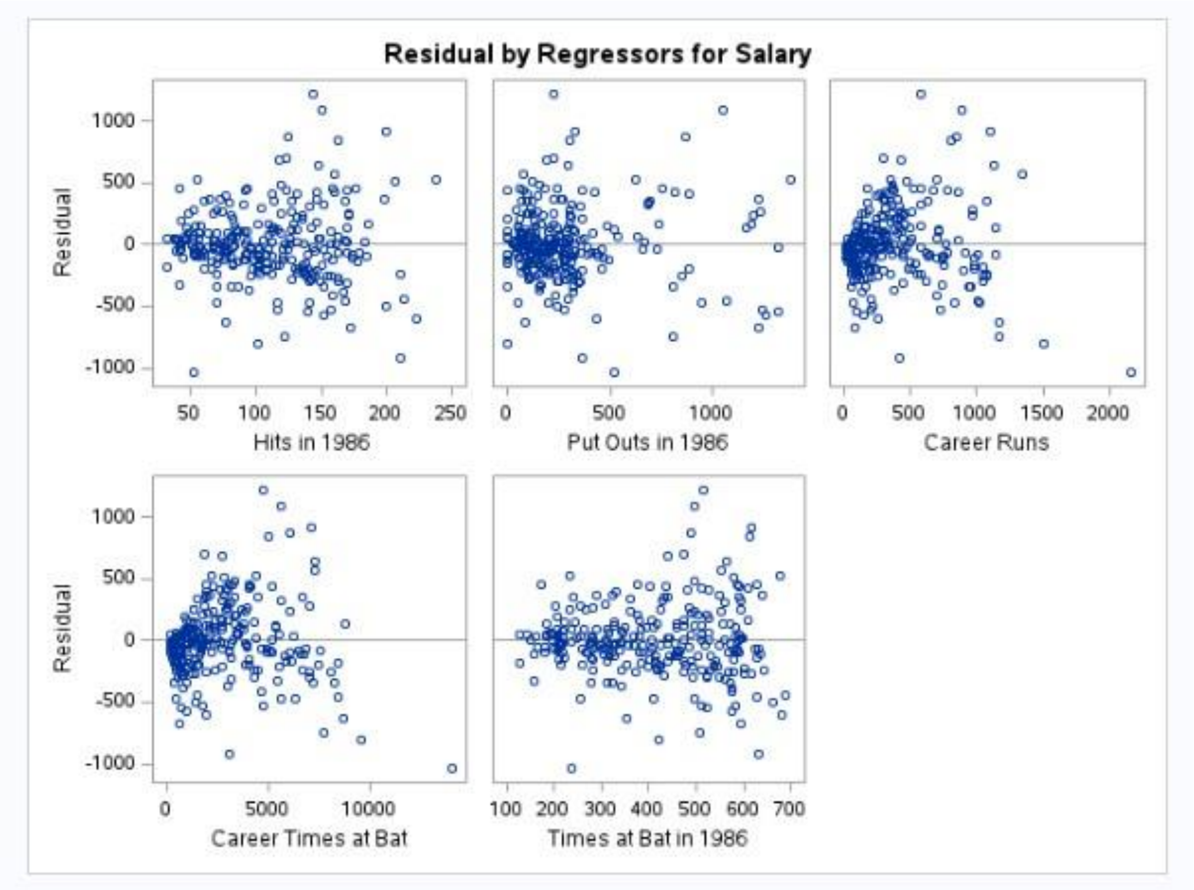


Stepwise Selection for Optimal Predictor Subset

The REG Procedure
Model: MODEL1
Dependent Variable: Salary 1987 Salary in $ Thousands

Fit Diagnostics for Salary

Observations	263
Parameters	6
Error DF	257
MSE	91258
R-Square	0.5601
Adj R-Square	0.5516



Residual by Regressors for Salary

Aryan Sharma
SAS Assignment

Model Summary and Interpretation

The final model includes the following predictors: nHits, nOuts, CrRuns, CrAtBat, and nAtBat. The overall model is statistically significant (F = 65.45, $p < 0.0001$), indicating that at least one predictor is significantly associated with the response variable.

R-squared = 0.5601 and Adjusted R-squared = 0.5516 indicate that the model explains approximately 56% of the variation in salaries, which is moderately strong for real-world data.

Interpretation of Coefficients and P-values

Significant predictors ($p < 0.05$):
- nHits (Estimate = 6.73523): Each additional hit in 1986 is associated with an increase in salary by approximately $6,735.
- nOuts (Estimate = 0.30403): Each additional put out in 1986 increases salary by approximately $304.
- CrRuns (Estimate = 1.41034): Each additional career run increases salary by approximately $1,410.

Non-significant predictors:
- CrAtBat and nAtBat have negative coefficients and p-values around 0.0171, slightly above typical thresholds, indicating weak significance.

Model Assumptions and Diagnostics

Assumption checks based on diagnostic plots:
- Linearity: Residual plots do not show strong patterns, supporting linear relationships.
- Independence: The data points appear randomly scattered, suggesting independence.
- Normality: The Q-Q plot shows that residuals mostly follow a normal distribution.
- Homoscedasticity: Some heteroscedasticity is visible, particularly at higher fitted values.
- Multicollinearity: CrRuns and CrAtBat show high Variance Inflation Factors (VIF > 10), indicating multicollinearity concerns.

Model Validity and Conclusion

The final model is statistically valid with moderate predictive power ($R^2 = 0.5601$). However, multicollinearity between certain career variables (e.g., CrRuns and CrAtBat) could inflate standard errors and impact interpretability. Despite minor violations in variance homogeneity, the model remains a reasonable tool for predicting player salaries. Further refinement may involve variable transformation or regularization techniques such as ridge regression to mitigate multicollinearity.