Linear Regression : Multiple Linear Regression

Learning Goals

- Why multiple regressors?
- Data Visualization: Scatterplot matrix
- Correlation matrix
- Multiple Linear regression model
- Ordinary Least Squares
- Interpretation of coefficient estimates
- Basic tests
- Assumptions

Model and Assumptions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + \varepsilon$$

Assumption

- **Linearity (Assumptions about the form of the model):**
  - Linear in parameters
- **Assumptions about the errors:**
  - IID Normal (independently and identically distributed)
  - Zero mean
  - Constant variance (Homoscedasticity)
  - Independent of each other
- **Assumptions about the predictors:**
  - Non-random
  - Measured without error
  - Linearly independent of each other
- **Assumptions about the observations:**
  - Equally reliable

The Cars Data

DATA : CARS , 81 observations

VOL = cubic feet of cab space

HP = engine horsepower

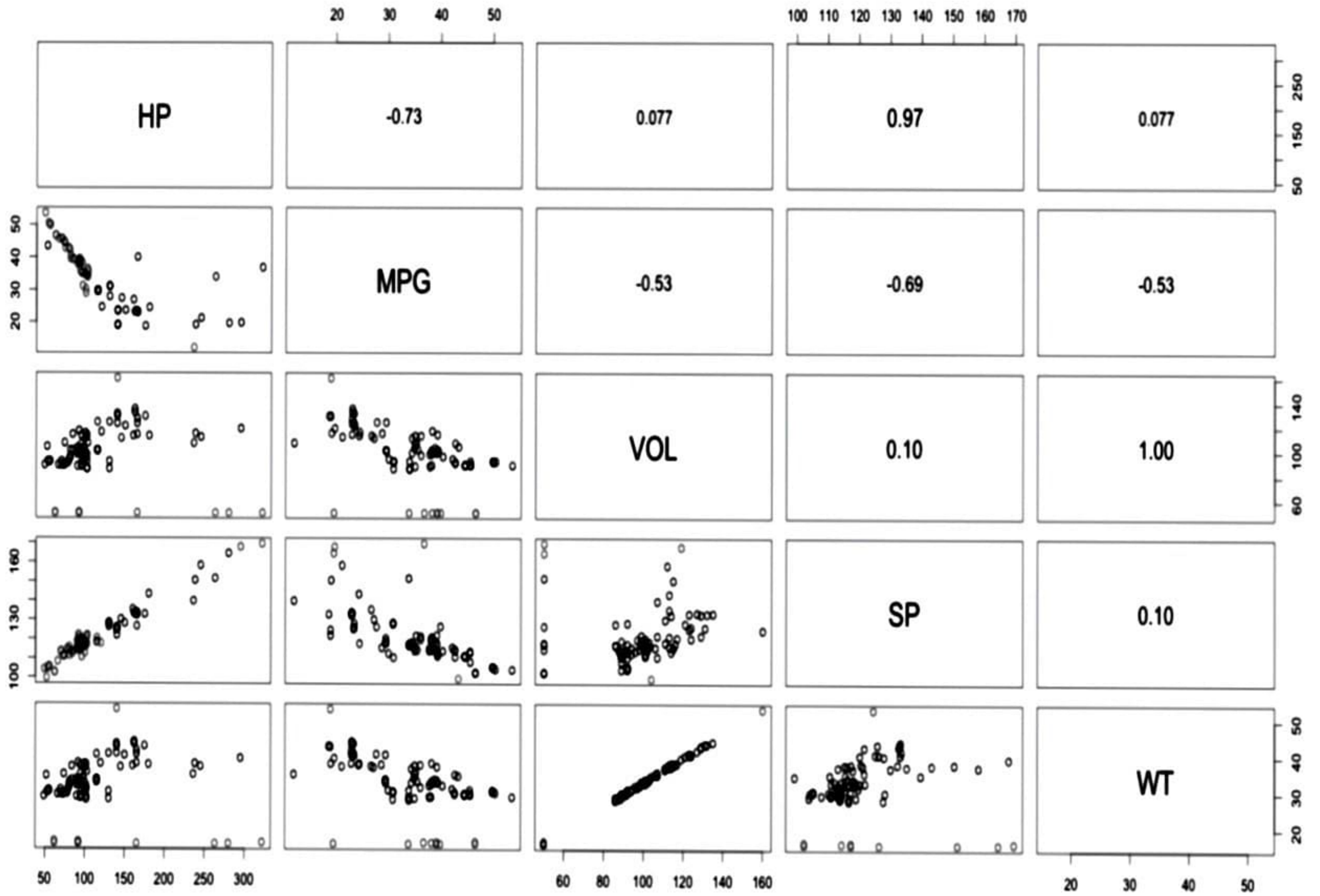
MPG = average miles per gallon

SP = top speed, miles per hour

WT = vehicle weight, hundreds of pounds

Our interest is to model the MPG of a car based on the other variables

Scatter Plot Matrix with Correlation Coefficients



Model Validation Techniques

Collinearity

Learning Goals

- What is Collinearity?
- Ill-effects of Collinearity
- Detection
  - Correlation Matrix
  - VIF
- Remedies
  - Subset selection
  - Best subset
  - Criteria for best subset
    - R^2, Adj. R^2, AIC

Detection of Collinearity: Methods for measuring Collinearity

- Correlation Matrix (Cars Data)

	HP	MPG	VOL	SP	WT
HP	1	-0.7250383	0.07745947	0.9738481	0.07651307
MPG	-0.72503835	1	-0.52905658	-0.6871246	-0.52675909
VOL	0.07745947	-0.5290566	1	0.10217	0.99920308
SP	**0.97384807**	-0.6871246	0.10217001	1	0.10243919

- Variance Inflation Factor (VIF)

Collinearity: Remedies

The next question would be to check which pair to include (VOL, SP) , (VOL, HP), (WT, SP) or (WT, HP)

- Subset Selection
- Best Subset
  - Based on R^2
  - Based on AIC

AIC: $2p-2\log(n[\log(2\pi)$

Model Validation Techniques

Residuals: $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$

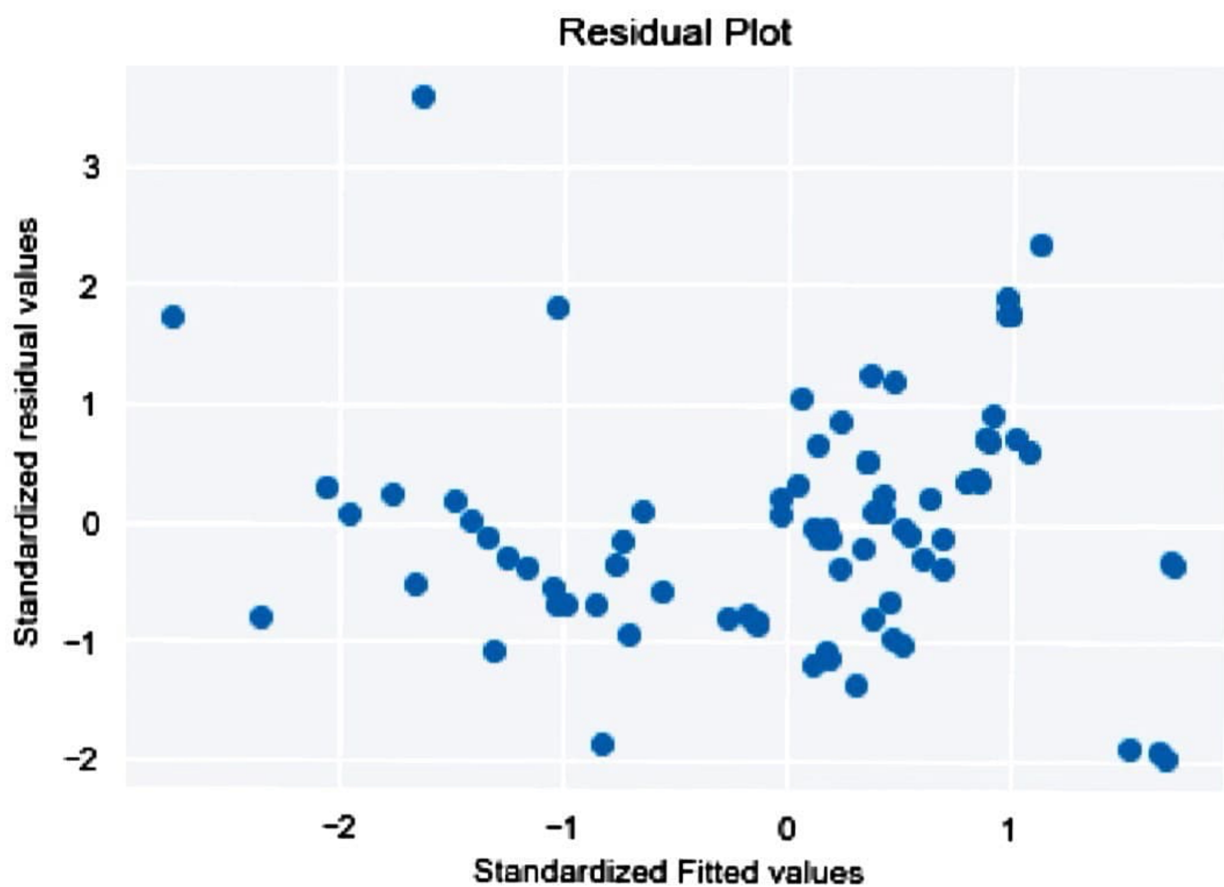
e_i vs Yi cap plot : will be used to check for linear relation, constant variance

If relation is nonlinear, U-shaped pattern appears
If error variance is non constant, funnel shaped pattern appears
If assumptions are met, random cloud of points appears

e_i vs X_i plot : will be used to check for linear relation, constant variance

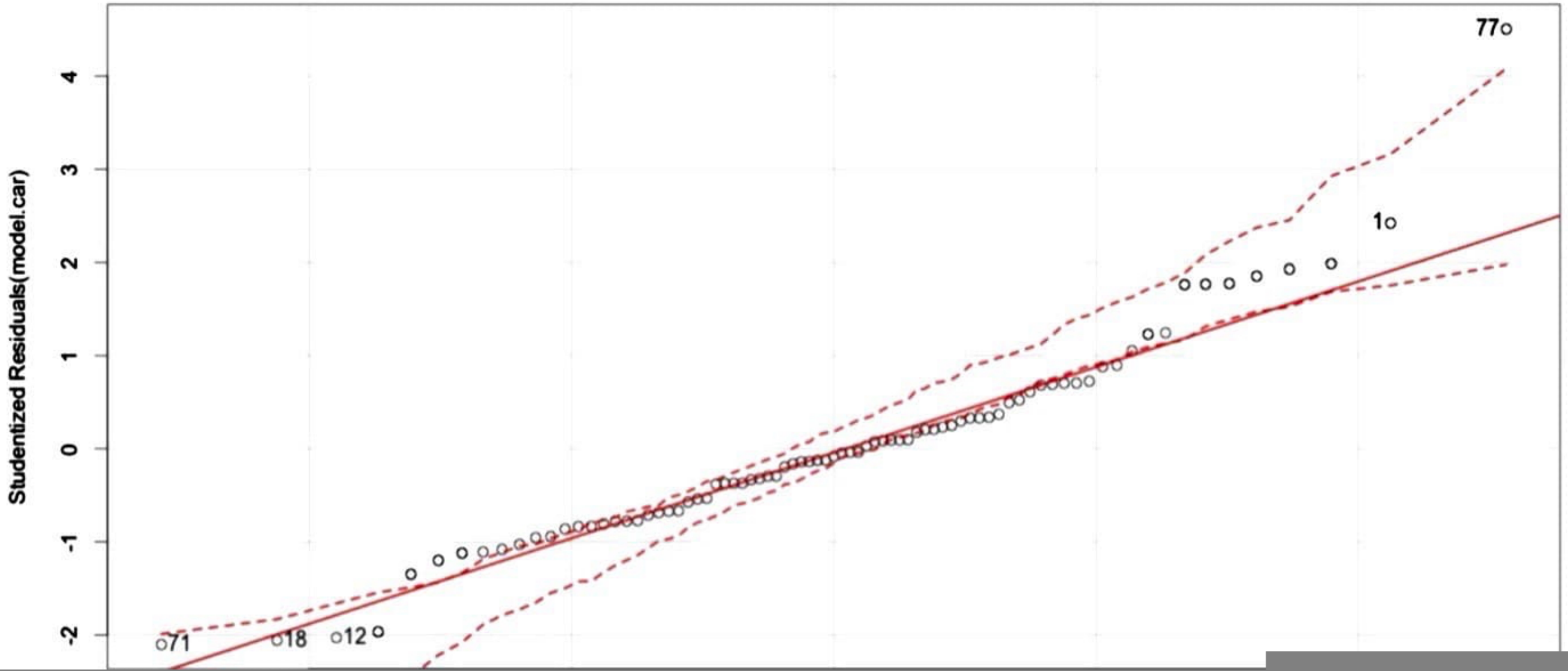
If relation is nonlinear, U-shaped pattern appears
If error variance is non constant, funnel shaped pattern appears
If assumptions are met, random cloud of points appears

Residual Plots: Fitted vs. Residuals

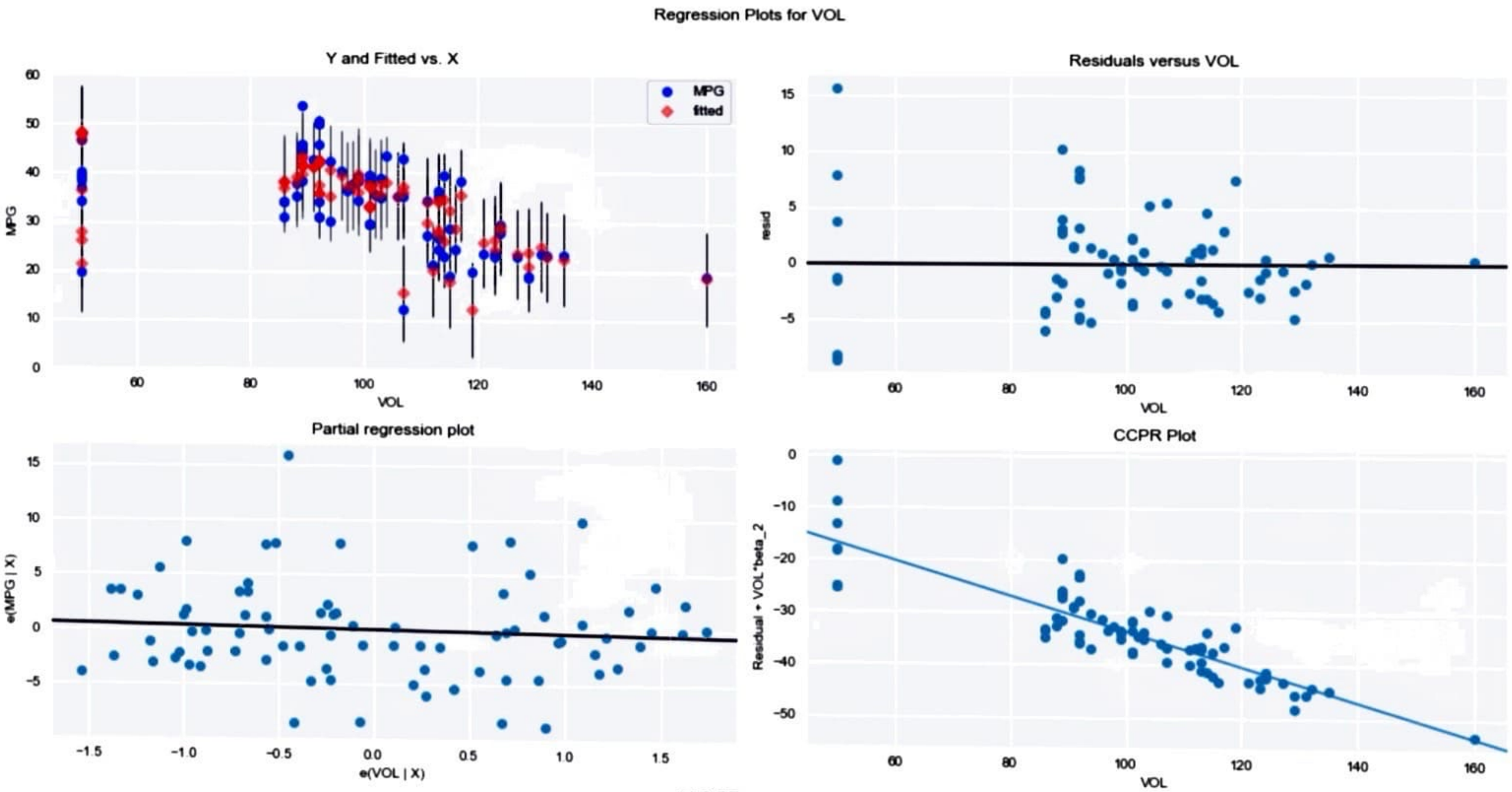


at to look for: No patterns, no problems. Model is good if residuals m

Checking for Normality: QQ-Plots



Residual Plots: Regressors vs. Residuals



Component and Component Plus Residual (CCPR)

What to look for: No patterns, no problems.

Model deletion Diagnostics

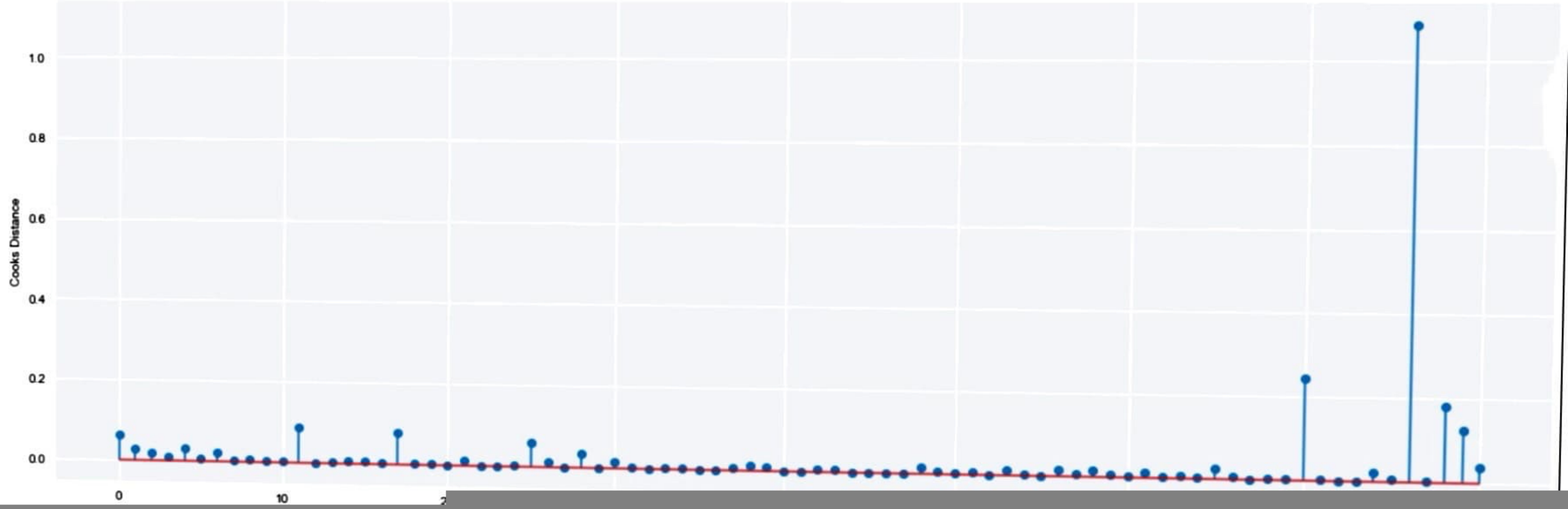
Model deletion Diagnostics

 # *Cook's distance* measures the difference between the **regression coefficients** obtained from **the full data and the regression coefficients obtained by deleting the ith observation**, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation.

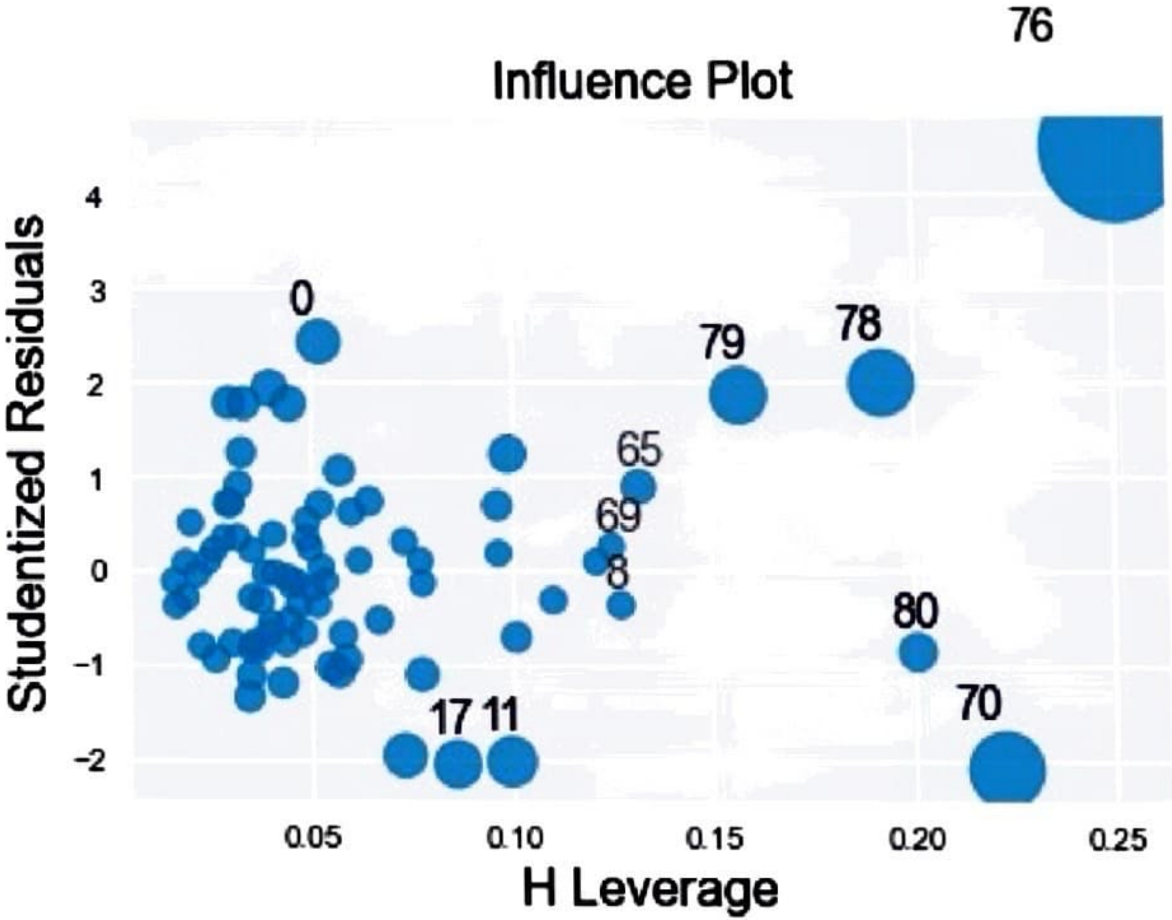
 # *Hat-points/ Leverage value / Influence* of an observation measures the influence of that observation on the overall fit of the regression function

Leverage value of more than $3(k + 1)/n$ is treated as highly influential

Diagnostics Plot : Cook's Distance



High Influence points



Leverage values of more than
3*(k + 1)/n are treated as highly influential observations.

Improve the Model

1. Deleting the 70 and 76th Observation : Check the model accuracy and variable significance
2. Discard the variable which are involved in the multicollinearity