Case Study – Linear Regression Analysis

Submitted by: Saikat Ghosh

Due Date: 9th September 2015

Date of Submission: 9th September 2015

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

Linear Regression:

In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables.

- **Assumptions for regression analysis :**

1. The sample is representative of the population for the inference prediction.
2. The error is a random variable with a mean of zero conditional on the explanatory variables.
3. The independent variables are measured with no error.
4. The independent variables (predictors) are linearly independent
5. The errors are uncorrelated.
6. The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

General Linear Model

In the more general multiple regression model, there are *p* independent variables:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable.

The least squares parameter estimates are obtained from *p* normal equations.

The residual can be written as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The **normal equations** are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i, \ \ j = 1, \dots, p.$$

In matrix notation, the normal equations are written as

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$

where the *ij* element of *X* is x_{ij}, the *i* element of the column vector *Y* is y_i, and the *j* element of $\hat{\beta}$ is $\hat{\beta}_j$.

Thus *X* is $n \times p$, *Y* is $n \times 1$, and $\hat{\beta}$ is $p \times 1$. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

1. Simple Model Building

i) Fitting a Linear Regression Model

Here we take "mpg" as dependent variable and rest all other variables as independent variables.

"mpg" = $\beta_0 + \beta_1$"cyl" + ... + β_p"carb" + ε

where $\beta_0, \beta_1, \ldots, \beta_p$ are termed as regression coefficients.

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
(Constant)	17.347	23.992	
cyl	.927	1.044	.275
disp	-.021	.011	-.436
hp	-.007	.022	-.065
drat	1.836	1.668	.163
wt	-3.780	1.633	-.437
qsec	.412	.870	.105
vs	-.822	2.293	-.069
am	-1.199	2.312	-.099
gear	1.387	1.606	.170
carb	-1.452	.611	-.320

In the above table regression coefficients β0, β1, ... , βp are represented by column "B" of "Unstandardized Coefficients".

ii) Testing the significance of Individual Parameters

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	17.347	23.992		.723	.478
cyl	.927	1.044	.275	.888	.385
disp	-.021	.011	-.436	-1.870	.076
hp	-.007	.022	-.065	-.297	.769
drat	1.836	1.668	.163	1.101	.284
wt	-3.780	1.633	-.437	-2.315	.031
qsec	.412	.870	.105	.474	.640
vs	-.822	2.293	-.069	-.359	.723
am	-1.199	2.312	-.099	-.519	.609
gear	1.387	1.606	.170	.863	.398
carb	-1.452	.611	-.320	-2.378	.027

H_0 : individual regressor is insignificant
H_1 : individual regressor is significant

From the above table, since p-value of all the parameters except "wt" & "carb" are greater than 0.05, hence we may conclude that all the individual parameters except "wt" & "carb" are insignificant at 5% l.o.s, i.e "wt" & "carb" are significant.

iii) Testing the significance of Overall Regression

ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	984.593	10	98.459	14.617	.000[b]
Residual	141.454	21	6.736		
Total	1126.047	31			

a. Dependent Variable: mpg
b. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp, qsec, cyl

H_0 : overall regression are insignificant
H_1 : overall regression are significant

From the above table, since p-value is less than 0.05, i.e we reject H_0. Hence we may conclude that the overall regression is significant at 5% l.o.s.

iv) R Square and Adjusted R Square

R-Square, also termed as coefficient of determination is a measure of goodness of model. It is defined as follows:

$$\boldsymbol{R^2 = \frac{SSReg}{SST} = 1 - \frac{SSRes}{SST}}$$

R^2 represents the proportion of variability explained by the model. Clearly $0 \le R^2 \le 1$. An adequate model is expected to have high R^2.

It can be proved that R^2 is an increasing function of the number of independent variables included in the model and hence it doesn't give true insight about goodness of the model. A refined measure of goodness which is free from this drawback, termed as adjusted R_2 is defined as follows.

$$\boldsymbol{R^2{}_{Adj} = 1 - \frac{SSRes/(n-p-1)}{SST/(p-1)}}$$

It can be observed that $\boldsymbol{-\infty \le R^2_{adj} \le R^2 \le 1}$.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935[a]	.874	.815	2.5954

a. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp, qsec, cyl

From the above table, since R_{adj} is quite high, hence we may conclude that the selected model is a good fit for the given data.
From the above hypothesis testing we see a high value of overall R^2.
We can also see that Adjusted R-square is always smaller than R-square.

2. Multicollinearity

i) Problem and its Consequences

The existence of near linear relationship among the explanatory variables is termed as **Multicollinearity**. In other words multicollinearity is a situation when one or more explanatory variables can be well expressed as a near linear combination of the other explanatory variables. Multicollinearity can arise due to several reasons like use of too many regressors, faulty data collection etc. In has been seen that presence of multicollinearity seriously weakens the results based on ordinary least squared technique. Following are the common ill consequences of this problem:

1. Variances of the ordinary least squares estimates of the regression coefficients are inflated.

2. Absolute values of the regression coefficients are very high.

3. Regressors which are expected to be important turn out to be insignificant.

4. The precision of the estimated regression coefficients decreases as more predictors are added to the model

5. The marginal contribution of any one predictor variable in reducing the error sum of squares depends on which other predictors are already in the model.

6. Hypothesis tests for $\beta_k = 0$ may yield different conclusions depending on which predictors are in the model.

ii) Detection and Removal of Multicollinearity

(i) Correlation Matrix

	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
cyl	1	.928**	.898**	-.679**	.825**	-.558**	-.814**	-.522**	-.564**	.570**
disp	.928**	1	.865**	-.684**	.783**	-.463**	-.724**	-.624**	-.594**	.540**
hp	.898**	.865**	1	-.539**	.690**	-.619**	-.752**	-.445*	-.437*	.690**
drat	-.679**	-.684**	-.539**	1	-.712**	0.079	.447*	.687**	.745**	-0.12
wt	.825**	.783**	.690**	-.712**	1	-0.269	-.532**	-.717**	-.628**	.384*
qsec	-.558**	-.463**	-.619**	0.079	-0.269	1	.792**	-0.162	-0.164	-.655**
vs	-.814**	-.724**	-.752**	.447*	-.532**	.792**	1	0.168	0.283	-.630**
am	-.522**	-.624**	-.445*	.687**	-.717**	-0.162	0.168	1	.808**	-0.09
gear	-.564**	-.594**	-.437*	.745**	-.628**	-0.164	0.283	.808**	1	0.085
carb	.570**	.540**	.690**	-0.12	.384*	-.655**	-.630**	-0.09	0.085	1

NOTE:

- As a rule-of-thumb if $|r_{ij}| > 0.75$ then Xi and Xj can be taken to have strong linear relationship.
- Deleting one of the regressors from all linearly related pairs will solve the problem.

From the correlation matrix, we observe that the following pairs of regressors have (absolute) correlation **greater than 0.75**

“Cyl” with (disp, hp, wt, vs)
“disp” with (hp, wt)
“hp” with “vs”
“qsec” with “vs”
“gear” with “am”

We can see that “cyl” has the stronger relationship with most of the regressors, So we can remove “Cyl”, “disp”, “hp”, “qsec”, “gear” to remove the “multicollinearity” from the model.
But we need to consider **“Variance Inflation Factor” (VIF)** test to infer about the “multicollinearity” because it gives better result.

(ii) Variance Inflation Factor (VIF) Approach

Let R_j^2 denote the coefficient of determination when Xj is regressed on all other predictor variables in the model.

Let $\mathbf{VIF_j = 1/ (1- R_j^2), \text{ for } j = 1,2, \ldots\ldots p-1}$

Clearly R2(j) and VIFj are positively related and hence VIFj is expected be high if jth is involved in multicollinearity.

As a rule-of-thumb if **VIFj > 10** then Xj can be taken to have strong linear relationship with the other regressors.

Model	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
cyl	.063	16.000
disp	.110	9.102
hp	.126	7.930
drat	.273	3.662
wt	.168	5.961
qsec	.122	8.198
vs	.163	6.145
am	.163	6.126
gear	.155	6.463
carb	.329	3.037

We see that the variable "cyl" is significant (i.e., VIFi > 10). Thus, to counter the problem of multicollinearity we remove “cyl”.

1. **Fitting a linear regression model (free from multicollinearity)**

- **Testing the significance of Individual Parameters**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	26.515	21.552		1.230	.232
disp	-.017	.010	-.342	-1.654	.112
hp	-.006	.022	-.057	-.263	.795
drat	1.464	1.607	.130	.911	.372
wt	-3.048	1.403	-.352	-2.173	.041
qsec	.185	.828	.047	.224	.825
vs	-1.163	2.250	-.097	-.517	.610
am	-.672	2.224	-.056	-.302	.765
gear	.900	1.503	.110	.599	.555
carb	-1.352	.597	-.298	-2.263	.034

H_0 : individual regressor is insignificant
H_1 : individual regressor is significant

From the above table, since p-value of all the parameters except “wt” & “carb” are greater than 0.05, hence we may conclude that all the individual parameters except “wt” & “carb” are insignificant at 5% l.o.s, i.e “wt” & “carb” are significant.

- **Testing the significance of Overall Regression**

ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	979.281	9	108.809	16.310	.000[b]
Residual	146.766	22	6.671		
Total	1126.047	31			

a. Dependent Variable: mpg
b. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp, qsec

H_0 : overall regression are insignificant
H_1 : overall regression are significant

From the above table, since p-value is less than 0.05, i.e we reject H_0. Hence we may conclude that the overall regression is significant at 5% l.o.s.

- **R Square and Adjusted R Square**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.933[a]	.870	.816	2.5829

a. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp, qsec

From the above table, since Radj is quite high, hence we may conclude that the selected model is a good fit for the given data.
From the above hypothesis testing we see a high value of overall R^2.
We can also see that Adjusted R-square is always smaller than R-square.
Value of R-square decreases because we remove "cyl"

Model	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
disp	.138	7.236
hp	.126	7.917
drat	.291	3.431
wt	.225	4.440
qsec	.133	7.491
vs	.167	5.973
am	.175	5.723
gear	.175	5.712
carb	.341	2.933

We can see that there is no multicollinearity related with the model because VIF of all the variables are less than 10.

From the two approaches used for removing the multicollinearity, we find that "Variance Inflation Factor (VIF)" approach is better than the R-square and Adjusted R-square approach. Hence, the **model under "Variance Inflation Factor (VIF)" approach is a better fit to the given data.**

2. Parsimonious Modeling or Model Selection

A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

a) Forward Selection

The simplest data-driven model building approach is called *forward selection*. The forward-selection technique begins with no variables in the model. Select the variable that has the highest R-Squared. At each step, select the variable that increases R-Squared the most. Stop adding variables when none of the remaining variables are significant. Note that once a variable enters the model, it cannot be deleted. Thus we begin with a model including the variable that is most significant in the initial analysis, and continue adding variables until none of remaining variables are "significant" when added to the model.

Coefficients[a]

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1 (Constant)	29.600	1.230		24.070	.000
disp	-.041	.005	-.848	-8.747	.000
2 (Constant)	37.863	2.325		16.287	.000
disp	-.025	.006	-.511	-4.386	.000
wt	-3.977	1.008	-.460	-3.945	.000
3 (Constant)	38.772	2.093		18.521	.000
disp	-.022	.005	-.446	-4.199	.000
wt	-3.589	.907	-.415	-3.956	.000
carb	-1.048	.358	-.231	-2.928	.007

a. Dependent Variable: mpg

Thus the model selected by *forward* selection method is:
"mpg" = β_0 + β_1"disp" + β_2"wt" + β_3"carb" + ε

Thus the model selected by *forward* selection method is:
"mpg" = 38.772 -.022 "disp" -3.589 "wt" - 1.048"carb" + ε

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.718	.709	3.2515
2	.904[b]	.817	.804	2.6677
3	.927[c]	.860	.845	2.3756

a. Predictors: (Constant), disp

b. Predictors: (Constant), disp, wt

c. Predictors: (Constant), disp, wt, carb

We can also see that the value of adjusted R-square is quite high. So, it is a good fit to the model.

b) Backward Elimination

Forward selection has drawbacks, including the fact that each addition of a new variable may render one or more of the already included variables non-significant. An alternate approach which avoids this is *backward elimination*. The backward elimination starts with all variables in the model. At each step, the variable that is the least significant is removed, so long as it is not significant at our chosen critical level. This process continues until no non-significant variables remain i.e., until all remaining variables are statistically significant.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	26.515	21.552		1.230	.232
	disp	-.017	.010	-.342	-1.654	.112
	hp	-.006	.022	-.057	-.263	.795
	drat	1.464	1.607	.130	.911	.372
	wt	-3.048	1.403	-.352	-2.173	.041
	qsec	.185	.828	.047	.224	.825
	vs	-1.163	2.250	-.097	-.517	.610
	am	-.672	2.224	-.056	-.302	.765
	gear	.900	1.503	.110	.599	.555
	carb	-1.352	.597	-.298	-2.263	.034
2	(Constant)	30.976	8.100		3.824	.001
	disp	-.016	.010	-.338	-1.675	.107
	hp	-.008	.019	-.082	-.445	.660
	drat	1.416	1.559	.126	.908	.373
	wt	-3.110	1.347	-.360	-2.309	.030
	vs	-.842	1.697	-.070	-.496	.625
	am	-.730	2.163	-.060	-.337	.739
	gear	.715	1.228	.088	.582	.566
	carb	-1.332	.579	-.294	-2.303	.031
3	(Constant)	30.782	7.929		3.882	.001
	disp	-.016	.010	-.327	-1.674	.107
	hp	-.009	.018	-.084	-.468	.644
	drat	1.349	1.518	.120	.889	.383
	wt	-2.869	1.120	-.332	-2.561	.017
	vs	-.565	1.458	-.047	-.387	.702
	gear	.477	.986	.058	.484	.633
	carb	-1.302	.561	-.287	-2.321	.029
4	(Constant)	30.131	7.616		3.956	.001
	disp	-.015	.009	-.307	-1.658	.110
	hp	-.008	.018	-.074	-.425	.675
	drat	1.289	1.484	.114	.868	.393
	wt	-2.889	1.100	-.334	-2.627	.014

Thus the model selected by *backward elimination* method is:
"mpg" = $\beta_0 + \beta_1$"disp" + β_2"wt" + β_3"carb" + ε

Thus the model selected by *backward elimination* method is:
"mpg" = 38.772 -.022 "disp" -3.589 "wt" -1.048"carb" + ε

	gear	.518	.964	.063	.537	.596
	carb	-1.241	.529	-.274	-2.345	.027
5	(Constant)	30.280	7.487		4.044	.000
	disp	-.018	.006	-.363	-2.844	.009
	drat	1.216	1.451	.108	.838	.409
	wt	-2.906	1.082	-.336	-2.687	.012
	gear	.523	.948	.064	.552	.586
	carb	-1.348	.457	-.298	-2.947	.007
6	(Constant)	31.656	6.968		4.543	.000
	disp	-.019	.006	-.383	-3.164	.004
	drat	1.460	1.364	.130	1.070	.294
	wt	-3.050	1.036	-.353	-2.945	.007
	carb	-1.223	.393	-.270	-3.114	.004
7	(Constant)	38.772	2.093		18.521	.000
	disp	-.022	.005	-.446	-4.199	.000
	wt	-3.589	.907	-.415	-3.956	.000
	carb	-1.048	.358	-.231	-2.928	.007

a. Dependent Variable: mpg

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.933[a]	.870	.816	2.5829
2	.932[b]	.869	.824	2.5290
3	.932[c]	.869	.830	2.4818
4	.932[d]	.868	.836	2.4393
5	.931[e]	.867	.841	2.4005
6	.930[f]	.865	.845	2.3694
7	.927[g]	.860	.845	2.3756

a. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp, qsec

b. Predictors: (Constant), carb, am, vs, drat, hp, gear, wt, disp

c. Predictors: (Constant), carb, vs, drat, hp, gear, wt, disp

d. Predictors: (Constant), carb, drat, hp, gear, wt, disp

e. Predictors: (Constant), carb, drat, gear, wt, disp

f. Predictors: (Constant), carb, drat, wt, disp

g. Predictors: (Constant), carb, wt, disp

We can also see that the value of adjusted R-square is quite high. So, it is a good fit to the model.

c) Stepwise Selection

Stepwise selection is a method that allows moving in either direction, dropping or adding variables at the various steps. The process is one of alternation between choosing the least significant variable to drop and then re-considering all dropped variables (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The second significance must be more stringent than the first.

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	29.600	1.230		24.070	.000
	disp	-.041	.005	-.848	-8.747	.000
2	(Constant)	37.863	2.325		16.287	.000
	disp	-.025	.006	-.511	-4.386	.000
	wt	-3.977	1.008	-.460	-3.945	.000
3	(Constant)	38.772	2.093		18.521	.000
	disp	-.022	.005	-.446	-4.199	.000
	wt	-3.589	.907	-.415	-3.956	.000
	carb	-1.048	.358	-.231	-2.928	.007

a. Dependent Variable: mpg

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.718	.709	3.2515
2	.904[b]	.817	.804	2.6677
3	.927[c]	.860	.845	2.3756

a. Predictors: (Constant), disp
b. Predictors: (Constant), disp, wt
c. Predictors: (Constant), disp, wt, carb

Thus the model selected by *stepwise* selection method is:
"mpg" = $\beta_0 + \beta_1$"disp" + β_2"wt" + β_3"carb" + ε

Thus the model selected by *stepwise* selection method is:
"mpg" = 38.772 -.022 "disp" -3.589 "wt" - 1.048"carb" + ε

We can also see that the value of adjusted R-square is quite high. So, it is a good fit to the model.

From the three approaches used for parsimonious modeling, we find that model selected by forward selection and stepwise selection methods are same. Also, the Adj. R2 for all the three models are equal but we use stepwise selection model to make our parsimonious model because this method performs both the technique of forward and backward simultaneously, i.e it is the combination of backward elimination and forward selection process.

So, our desired regression model is **"mpg" = $\beta_0 + \beta_1$"disp" + β_2"wt" + β_3"carb" + ε**
i.e "mpg" = 38.772 -.022 "disp" -3.589 "wt" -1.048"carb" + ε

3. Validation of Assumptions and Residual Analysis

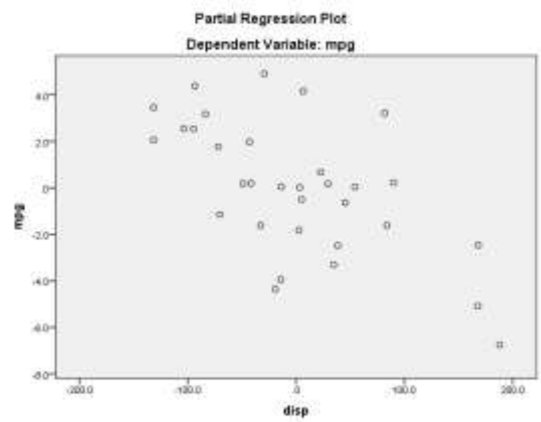
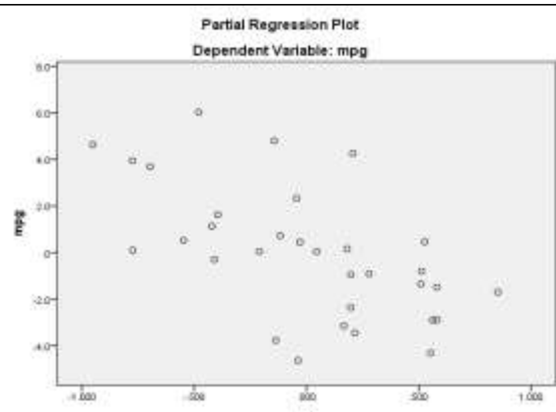
There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

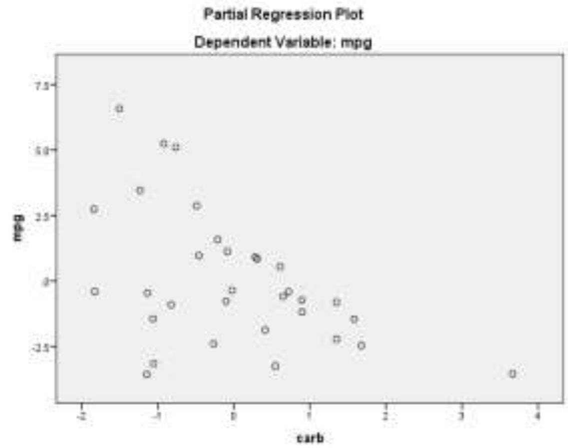
a) **linearity and additivity** of the relationship between dependent and independent variables.
b) **statistical independence** of the errors (in particular, no correlation between consecutive errors in the case of time series data)
c) **homoscedasticity** (constant variance) of the errors
d) **normality** of the error distribution.

If any of these assumptions is violated, then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be inefficient or seriously biased or misleading.

a) Linearity of Regression

Linearity of individual regressors in the complete model can be validated using the **partial regression plots**.

	From the plot we can see that “mpg” and “disp” is linearly related, so we can take “disp” in the model linearly.
	From the plot we can see that “mpg” and “wt” is linearly related, so we can take “wt” in the model linearly.

	From the plot we can see that “mpg” and “carb” is almost linearly related, so we can take “carb” in the model linearly.

b) Autocorrelation

When error terms for different sample observations are correlated to each other, i.e. there is a linear relationship among the error terms the situation is called as autocorrelation or serial correlation.

if $1 \leq DW \leq 3$ then there is no Autocorrelation,
if $0 < DW < 1$ then there is a positive autocorrelation, and
if $3 < DW < 4$ then there is a negative autocorrelation.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.927[a]	.860	.845	2.3756	2.032

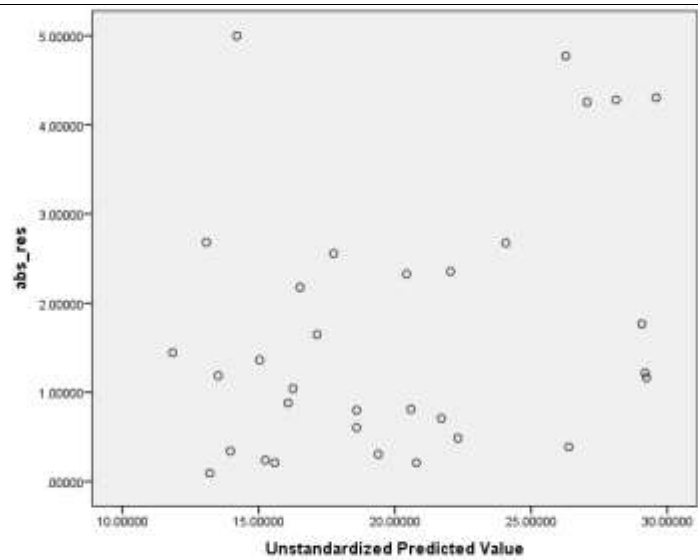
a. Predictors: (Constant), carb, wt, disp

b. Dependent Variable: mpg

Thus we have Durbin-Watson statistic D = 2.032 i.e, between 1 and 3. Hence we may conclude that the model selected by *stepwise* selection method does not have autocorrelation.

c) Heteroscedasticity

When error terms for different sample observations have same variances the situation is called as homoscedasticity. The opposite of homoscedasticity is called as hetroscedasticity.



The graph does not exhibit any pattern so we infer that there is no heteroscedasticity.

We are interested in detecting if there is a hetroscedasticity present in the model. This can also be done using a test based on **Spearman's Rank Correlation**

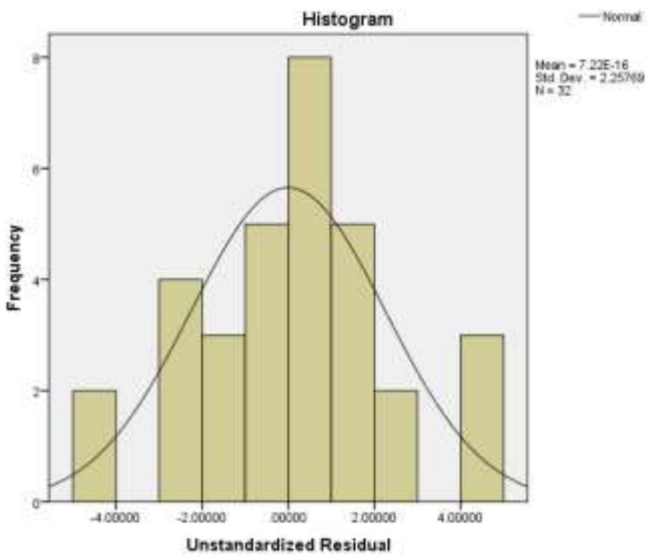
Correlations

			Unstandardized Predicted Value	abs_res
Spearman's rho	Unstandardized Predicted Value	Correlation Coefficient	1.000	.255
		Sig. (2-tailed)	.	.159
		N	32	32
	abs_res	Correlation Coefficient	.255	1.000
		Sig. (2-tailed)	.159	.
		N	32	32

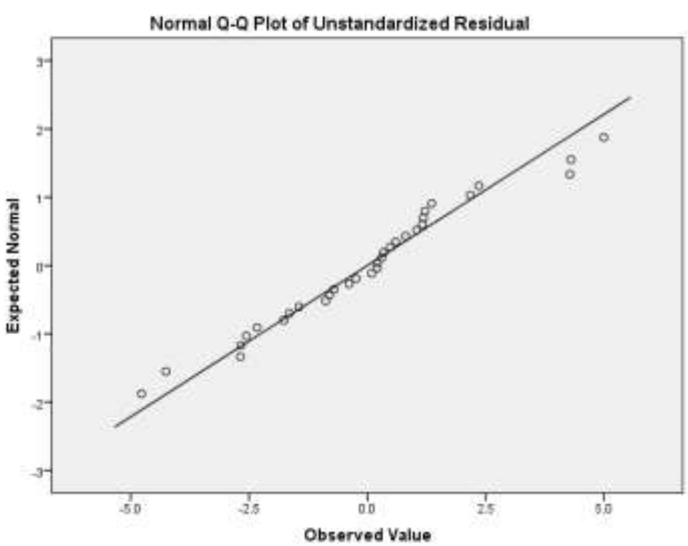
H_0 : correlation is insignificant i.e. heteroscedasticity is not present
H_1 : correlation is significant i.e. heteroscedasticity is present

The significant value is 0.152(>0.05) so we accept H_0 at 5% los i.e. there is no heteroscedasticity.

d) Normality of Errors



From the histogram with normal curve, we may conclude that the Residuals are almost normally distributed.



From Q-Q Plot we can see that residual is normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.117	32	.200*	.973	32	.585

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

H_0 : residual is normally distributed
H_1 : residual is not normally ditributed
From **Kolmogorov-Smirnov** Test with Sig. = 0.200 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., “residual” is normally distributed.

From **Shapiro-Wilk** Test with Sig. = 0.585(>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., “residual” is normally distributed.

➢ Influential Observations (Leverage and Outliers)

An **outlier** is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

A Common Rule: The observations corresponding to which the absolute value of studentized residuals lie beyond 3 can surely be taken as outliers. But the observations corresponding to which the absolute value of studentized residuals lies between 2 and 3 should also be dealt carefully. For such observations, we consider leverage values, if the leverage is more than $2p/n$, then those observations are also taken to be outliers.

Abs Studentized Residuals	Levarage Value	Abs SRE≥3	3 ≥ Abs SRE ≥ 2
0.31	0.07632	0	0
0.09	0.07021	0	0
1.88	0.06265	0	0
0.36	0.07932	0	0
0.97	0.07161	0	0
1.05	0.10606	0	0
0.15	0.04686	0	0
1.04	0.05739	0	0
0.21	0.05658	0	0
0.27	0.11457	0	0
0.36	0.11457	0	0
0.62	0.11052	0	0
0.46	0.04282	0	0
0.39	0.05066	0	0
0.68	0.167	0	0
1.29	0.19925	0	0
0.56	0.16559	0	0
1.9	0.07355	0	0
0.54	0.1381	0	0
1.95	0.10756	0	0
2.1	0.05704	0	1
0.72	0.04787	0	0
1.12	0.03851	0	0
0.04	0.05439	0	0
2.28	0.11373	0	1
0.8	0.09493	0	0
0.17	0.05229	0	0
0.57	0.17086	0	0
0.09	0.06611	0	0
0.16	0.32227	0	0
0.1	0.03796	0	0
1.16	0.03286	0	0

We find that no absolute Studentized Residual is greater than 3. Also, only the two values (denoted by 1) lies between 2 and 3 for which on comparing their leverage values with 3/16 turns out to be less implying that these observations also cannot be classified as outliers.

Hence there is no outlier present in the data.