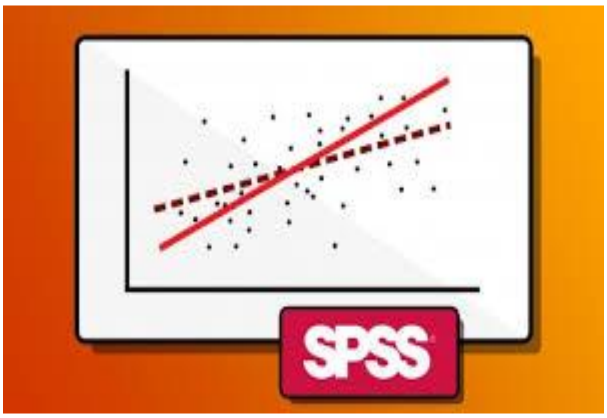

REGRESSION
ANALYSIS
ASSUMPTIONS
THEOREM
OBSERVATION
VECTOR
SET
DERIVATION
ASYMPTOTIC
DIAGNOSTIC
MATRIX
PARAMETRIC
PREDICT
METHODS
NUMBER
AGGREGATED
ELECTROMECHANICAL
INFORMATION
ESTIMATION
DATA
VARIANCE
RESIDUAL
OBSERVED


SPSS

CASE STUDY

REGRESSION ANALYSIS (RA)

Due Date: **9th September, 2015**

Date of Submission: **8th September, 2015**

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

Contents

About RA:

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling 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 (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

Classical assumptions for Regression Analysis:

- Sample is representative of the population for the inference prediction.
- Normality of Errors: Error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error.
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- Homoscedasticity: The variance of the error is constant across observations If not, weighted least squares or other methods might instead be used.

Dataset: “Mtcars”

Aim: To explore the relationship of various variables designed to analyze the performance of cars on “Miles per gallon” (mpg).

a) Model Building

The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as **simple linear regression**. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as multiple linear regression, also known as **multivariable linear regression**.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	carb, am, vs, drat, qsec, gear, disp, hp, wt, cyl[b]	.	Enter

a. Dependent Variable: mpg

b. All requested variables entered.

- R^2 is **coefficient of determination** indicates how well data fit a statistical model.
- **Adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.934[a]	.873	.813	2.6070

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

Here, We have R^2= 0.873 and adj R^2= 0.813. Since R^2 is close to 1 we can say that model is a good fit.

Test for Overall Regression

Under normality assumption for the error terms, significance overall regression can be tested using an F-test. The procedure is as follows.

H_0: $\beta_1=\beta_2=\ldots..=\beta_{10}$

H_1: $\beta_i \neq 0$ for at least one i. (i=1,2,3,.........,10)

Test-statistic:

F=MSreg/MSres

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	983.318	10	98.332	14.468	.000[b]
	Residual	142.729	21	6.797		
	Total	1126.047	31			

a. Dependent Variable: mpg

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

Inference:

Since p-value for overall regression is less than 0.05, so we reject our Null-Hypothesis and say that βi≠0 for at least one i.

Test for Individual Regressors

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.163	21.599		.934	.361
	cyl	.517	.974	.153	.531	.601
	disp	-.007	.011	-.134	-.607	.550
	hp	-.031	.020	-.355	-1.594	.126
	drat	1.550	1.696	.138	.914	.371
	wt	-2.192	1.165	-.290	-1.882	.074
	qsec	.101	.837	.026	.121	.905
	vs	.176	2.429	.015	.073	.943
	am	.810	2.082	.067	.389	.701
	gear	1.307	1.524	.160	.858	.401
	carb	-1.154	.528	-.251	-2.185	.040

a. Dependent Variable: mpg

Under normality assumption for the error terms, significance of individual parameters can be tested using t-test.

H0: $\beta_i=0$

H1: $\beta_i \neq 0$

Test-Statistic: $t=(\hat{\beta})/(S.E(\hat{\beta}))$

Inference:

Since p-value for all the individual regressors are greater than 0.05, so we fail to reject our Null-Hypothesis and conclude that $\beta_i=0$. Only carb is coming out to be significant.

FITTED MODEL:

mpg= 20.163 +(0.517**cyl**)+(-0.007**disp**)+(-0.031**hp**)+(1.550**drat**)+(-2.192**wt**)+(0.101**qsec**)+(0.176**vs**)

+ (0.810**am**)+(1.307**gear**)+(-1.154**carb**).

b) MULTICOLLINEARITY

Problem of multicollinearity

- Multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables.
- When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients.
- It will result in incorrect conclusions about the relationship between outcome variable and predictor variables.

Consequences of high multicollinearity:

- Increased standard error of estimates of the β's (decreased reliability).
- Often confusing and misleading results.

Detection of Multicollinearity:

1) Examination of Correlation Matrix:

The easiest way to measure the extent of multicollinearity is simply to look at the matrix of correlations between the individual variables.

- Large correlation coefficients in the correlation matrix of predictor variables indicate multicollinearity.(Taking threshold for significant (absolute) correlation to be 0.75)

- If there is a multicollinearity between any two predictor variables, then the correlation coefficient between these two variables will be near to unity.

Correlations

			cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Spearman's rho	cyl	Correlation Coefficient	1.000	.928**	.902**	-.679**	.794**	-.535**	-.814**	-.522**	-.564**	.559**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.002	.000	.002	.001	.001
		N	32	32	32	32	32	32	32	32	32	32
	disp	Correlation Coefficient	.928**	1.000	.851**	-.684**	.776**	-.444*	-.724**	-.624**	-.594**	.537**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.011	.000	.000	.000	.002
		N	32	32	32	32	32	32	32	32	32	32
	hp	Correlation Coefficient	.902**	.851**	1.000	-.520**	.679**	-.646**	-.752**	-.362*	-.331	.686**
		Sig. (2-tailed)	.000	.000	.	.002	.000	.000	.000	.042	.064	.000
		N	32	32	32	32	32	32	32	32	32	32
	drat	Correlation Coefficient	-.679**	-.684**	-.520**	1.000	-.697**	.058	.447*	.687**	.745**	-.122
		Sig. (2-tailed)	.000	.000	.002	.	.000	.754	.010	.000	.000	.505
		N	32	32	32	32	32	32	32	32	32	32
	wt	Correlation Coefficient	.794**	.776**	.679**	-.697**	1.000	-.233	-.505**	-.710**	-.598**	.372*
		Sig. (2-tailed)	.000	.000	.000	.000	.	.200	.003	.000	.000	.036
		N	32	32	32	32	32	32	32	32	32	32
	qsec	Correlation Coefficient	-.535**	-.444*	-.646**	.058	-.233	1.000	.771**	-.162	-.181	-.602**
		Sig. (2-tailed)	.002	.011	.000	.754	.200	.	.000	.376	.323	.000
		N	32	32	32	32	32	32	32	32	32	32
	vs	Correlation Coefficient	-.814**	-.724**	-.752**	.447*	-.505**	.771**	1.000	.168	.283	-.620**
		Sig. (2-tailed)	.000	.000	.000	.010	.003	.000	.	.357	.117	.000
		N	32	32	32	32	32	32	32	32	32	32
	am	Correlation Coefficient	-.522**	-.624**	-.362*	.687**	-.710**	-.162	.168	1.000	.808**	-.136
		Sig. (2-tailed)	.002	.000	.042	.000	.000	.376	.357	.	.000	.458
		N	32	32	32	32	32	32	32	32	32	32
	gear	Correlation Coefficient	-.564**	-.594**	-.331	.745**	-.598**	-.181	.283	.808**	1.000	.028
		Sig. (2-tailed)	.001	.000	.064	.000	.000	.323	.117	.000	.	.880
		N	32	32	32	32	32	32	32	32	32	32
	carb	Correlation Coefficient	.559**	.537**	.686**	-.122	.372*	-.602**	-.620**	-.136	.028	1.000
		Sig. (2-tailed)	.001	.002	.000	.505	.036	.000	.000	.458	.880	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Inference:

Since value of absolute:

a) Correlation of cyl with disp, hp, drat, wt, vs is greater than 0.75, so we say cyl has strong linear relationship with each of respective variables.

b) Correlation between factor disp and cyl, hp, wt, vs is greater than 0.75, so we say disp has strong linear relationship with each of respective variables.

c) Correlation between factor hp and cyl, disp, vs is greater than 0.75, so we say hp has strong linear relationship with each of respective variables.

d) Correlation between factor drat and gear is greater than 0.75, so we say drat has strong linear relationship with each other.

e) Correlation between factor wt and cyl, disp, am is greater than 0.75, so we say wt has strong linear relationship with each of respective variables.

f) Correlation between factor qsec and vs is greater than 0.75, so we say qsec has strong linear relationship with each other.

g) Correlation between factor vs and cyl, disp, hp, qsec is greater than 0.75, so we say vs has strong linear relationship with each of respective variables.

h) Correlation between factor am and wt, gear is greater than 0.75, so we say am has strong linear relationship with each of respective variables.

i) Correlation between factor gear and drat, am is greater than 0.75, so we say gear has strong linear relationship with each of repective variables.

j) Correlation between factor carb and other variables are less than threshold value. So we conclude it has not much strong relationship with other variables.

Inference: We conclude that "*cyl"* has stronger relationship with most of the variables. So, if infer that we

drop cyl multicollinearity could be reduced.

2) Variance Inflation Factor:

- The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an ordinary least- squares regression analysis.
- The VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity.
- Rule of Thumb: If any of the VIF values exceeds 5 or 10, it implies that the associated regression coefficients are poorly estimated because of multicollinearity.

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	cyl	.072	13.801
	disp	.123	8.104
	hp	.122	8.206
	drat	.267	3.750
	wt	.255	3.925
	qsec	.134	7.486
	vs	.146	6.837
	am	.203	4.921
	gear	.173	5.767
	carb	.459	2.180

a. Dependent Variable: mpg

(Taking threshold for VIF to be 10)

Inference:

Since VIF of cyl is greater than 10, we say that it has strong linear relationship with all other regressors.

So, we drop cyl and run regression again we get the following results:

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	carb, am, hp, vs, drat, wt, gear, qsec, disp[b]	.	Enter

a. Dependent Variable: mpg

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.934[a]	.872	.819	2.5641

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

b. Dependent Variable: mpg

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	981.401	9	109.045	16.585	.000[b]
	Residual	144.647	22	6.575		
	Total	1126.047	31			

a. Dependent Variable: mpg

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	25.093	19.182		1.308	.204		
	disp	-.005	.010	-.110	-.519	.609	.129	7.770
	hp	-.027	.018	-.311	-1.529	.141	.141	7.092
	drat	1.336	1.620	.119	.825	.418	.283	3.538
	wt	-2.059	1.119	-.272	-1.840	.079	.267	3.743
	qsec	.037	.814	.009	.045	.965	.136	7.327
	vs	-.227	2.269	-.019	-.100	.921	.162	6.169
	am	.812	2.047	.067	.396	.696	.203	4.921
	gear	1.038	1.414	.127	.734	.470	.195	5.130
	carb	-1.110	.513	-.241	-2.164	.042	.470	2.127

a. Dependent Variable: mpg

Inference:

We can conclude now that problem of multicollinearity has been resolved since VIF of all the variables are less than 10.

c) **Parsimonious Modelling or Model Selection**

Forward selection:

This approach builds the model starting with no variables in the model and adds useful variables one by one.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	disp	.	Forward (Criterion: Probability-of-F-to-enter <= .150)
2	wt	.	Forward (Criterion: Probability-of-F-to-enter <= .150)
3	carb	.	Forward (Criterion: Probability-of-F-to-enter <= .150)

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	.895[b]	.801	.787	2.7800
3	.917[c]	.842	.825	2.5240

a. Predictors: (Constant), disp

b. Predictors: (Constant), disp, wt

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

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	808.888	1	808.888	76.513	.000[b]
	Residual	317.159	30	10.572		
	Total	1126.047	31			
2	Regression	901.925	2	450.962	58.352	.000[c]
	Residual	224.122	29	7.728		
	Total	1126.047	31			
3	Regression	947.666	3	315.889	49.584	.000[d]
	Residual	178.381	28	6.371		
	Total	1126.047	31			

a. Dependent Variable: mpg

b. Predictors: (Constant), disp

c. Predictors: (Constant), disp, wt

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

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	29.600	1.230		24.070	.000		
	disp	-.041	.005	-.848	-8.747	.000	1.000	1.000
2	(Constant)	36.112	2.151		16.786	.000		
	disp	-.025	.006	-.512	-4.019	.000	.423	2.364
	wt	-3.345	.964	-.442	-3.470	.002	.423	2.364
3	(Constant)	37.157	1.992		18.655	.000		
	disp	-.022	.006	-.453	-3.848	.001	.408	2.449
	wt	-3.015	.884	-.398	-3.411	.002	.415	2.410
	carb	-1.029	.384	-.223	-2.680	.012	.814	1.228

a. Dependent Variable: mpg

Inference:

Using method of forward selection we get the following regression model:

mpg = 37.157-0.022 disp-3.015wt-1.029 carb

Backward elimination

Instead of starting with no variables in the model, start with all predictor variable in the model and remove unhelpful variables from the model one by one.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	carb, am, hp, vs, drat, wt, gear, qsec, disp[b]	.	Enter
2	.	qsec	Backward (criterion: Probability of F-to-remove >= .200).
3	.	vs	Backward (criterion: Probability of F-to-remove >= .200).
4	.	am	Backward (criterion: Probability of F-to-remove >= .200).
5	.	disp	Backward (criterion: Probability of F-to-remove >= .200).
6	.	drat	Backward (criterion: Probability of F-to-remove >= .200).

a. Dependent Variable: mpg

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.934[a]	.872	.819	2.5641
2	.934[b]	.872	.827	2.5079
3	.934[c]	.871	.834	2.4556
4	.933[d]	.870	.839	2.4193
5	.932[e]	.868	.843	2.3875
6	.929[f]	.862	.842	2.3976

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

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

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

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

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

f. Predictors: (Constant), carb, hp, wt, gear

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	981.401	9	109.045	16.585	.000[b]
	Residual	144.647	22	6.575		
	Total	1126.047	31			
2	Regression	981.387	8	122.673	19.504	.000[c]
	Residual	144.660	23	6.290		
	Total	1126.047	31			
3	Regression	981.329	7	140.190	23.249	.000[d]
	Residual	144.718	24	6.030		
	Total	1126.047	31			
4	Regression	979.720	6	163.287	27.898	.000[e]
	Residual	146.327	25	5.853		
	Total	1126.047	31			
5	Regression	977.841	5	195.568	34.309	.000[f]
	Residual	148.206	26	5.700		
	Total	1126.047	31			
6	Regression	970.843	4	242.711	42.223	.000[g]
	Residual	155.204	27	5.748		
	Total	1126.047	31			

a. Dependent Variable: mpg

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

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

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

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

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

g. Predictors: (Constant), carb, hp, wt, gear

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	25.093	19.182		1.308	.204		
	disp	-.005	.010	-.110	-.519	.609	.129	7.770
	hp	-.027	.018	-.311	-1.529	.141	.141	7.092
	drat	1.336	1.620	.119	.825	.418	.283	3.538
	wt	-2.059	1.119	-.272	-1.840	.079	.267	3.743
	qsec	.037	.814	.009	.045	.965	.136	7.327
	vs	-.227	2.269	-.019	-.100	.921	.162	6.169
	am	.812	2.047	.067	.396	.696	.203	4.921
	gear	1.038	1.414	.127	.734	.470	.195	5.130
	carb	-1.110	.513	-.241	-2.164	.042	.470	2.127
2	(Constant)	25.885	7.421		3.488	.002		
	disp	-.005	.010	-.109	-.529	.602	.132	7.598
	hp	-.028	.014	-.316	-1.927	.066	.207	4.823
	drat	1.321	1.550	.117	.852	.403	.296	3.384
	wt	-2.052	1.083	-.271	-1.895	.071	.273	3.662
	vs	-.158	1.643	-.013	-.096	.924	.296	3.378
	am	.825	1.980	.068	.417	.681	.208	4.813
	gear	1.012	1.261	.124	.803	.430	.235	4.263
	carb	-1.111	.502	-.241	-2.213	.037	.470	2.127
3	(Constant)	25.803	7.219		3.574	.002		
	disp	-.005	.009	-.103	-.537	.596	.146	6.848
	hp	-.028	.014	-.314	-1.970	.060	.210	4.757
	drat	1.292	1.488	.115	.868	.394	.307	3.254
	wt	-2.047	1.059	-.270	-1.933	.065	.274	3.654
	am	.907	1.755	.075	.516	.610	.254	3.943
	gear	.991	1.215	.121	.816	.423	.242	4.128
	carb	-1.090	.444	-.237	-2.456	.022	.576	1.735
4	(Constant)	25.238	7.030		3.590	.001		
	disp	-.005	.009	-.107	-.567	.576	.146	6.837
	hp	-.026	.014	-.301	-1.939	.064	.216	4.622
	drat	1.410	1.448	.125	.974	.340	.315	3.177
	wt	-2.267	.955	-.300	-2.374	.026	.327	3.062
	gear	1.288	1.054	.158	1.222	.233	.312	3.202
	carb	-1.108	.436	-.241	-2.541	.018	.580	1.724
5	(Constant)	23.549	6.283		3.748	.001		
	hp	-.032	.009	-.364	-3.467	.002	.458	2.182
	drat	1.558	1.406	.138	1.108	.278	.325	3.074
	wt	-2.349	.931	-.310	-2.522	.018	.334	2.992
	gear	1.589	.897	.195	1.771	.088	.419	2.384
	carb	-1.138	.427	-.247	-2.663	.013	.588	1.699
6	(Constant)	28.621	4.323		6.621	.000		
	hp	-.035	.009	-.399	-3.962	.000	.503	1.988
	wt	-2.727	.870	-.360	-3.133	.004	.386	2.591
	gear	2.107	.770	.258	2.737	.011	.575	1.739
	carb	-1.049	.421	-.228	-2.490	.019	.610	1.640

a. Dependent Variable: mpg

Inference:

Using method of Backward elimination, we get the following regression model:

mpg = 28.261-0.035hp-2.727wt+2.107gear-1.409carb

Stepwise Selection

This approach combines both forward selection and backward deletion. It allows variable added early on to be dropped out and variables that are dropped at one point to be added back in.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	disp	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
2	wt	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
3	carb	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).

a. Dependent Variable: mpg

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.718	.709	3.2515
2	.895[b]	.801	.787	2.7800
3	.917[c]	.842	.825	2.5240

a. Predictors: (Constant), disp

b. Predictors: (Constant), disp, wt

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

d. Dependent Variable: mpg

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	808.888	1	808.888	76.513	.000[b]
	Residual	317.159	30	10.572		
	Total	1126.047	31			
2	Regression	901.925	2	450.962	58.352	.000[c]
	Residual	224.122	29	7.728		
	Total	1126.047	31			
3	Regression	947.666	3	315.889	49.584	.000[d]
	Residual	178.381	28	6.371		
	Total	1126.047	31			

a. Dependent Variable: mpg

b. Predictors: (Constant), disp

c. Predictors: (Constant), disp, wt

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.600	1.230		24.070	.000		
	disp	-.041	.005	-.848	-8.747	.000	1.000	1.000
2	(Constant)	36.112	2.151		16.786	.000		
	disp	-.025	.006	-.512	-4.019	.000	.423	2.364
	wt	-3.345	.964	-.442	-3.470	.002	.423	2.364
3	(Constant)	37.157	1.992		18.655	.000		
	disp	-.022	.006	-.453	-3.848	.001	.408	2.449
	wt	-3.015	.884	-.398	-3.411	.002	.415	2.410
	carb	-1.029	.384	-.223	-2.680	.012	.814	1.228

a. Dependent Variable: mpg

Inference:

Using method of Stepwise Selection, we get the following regression model:

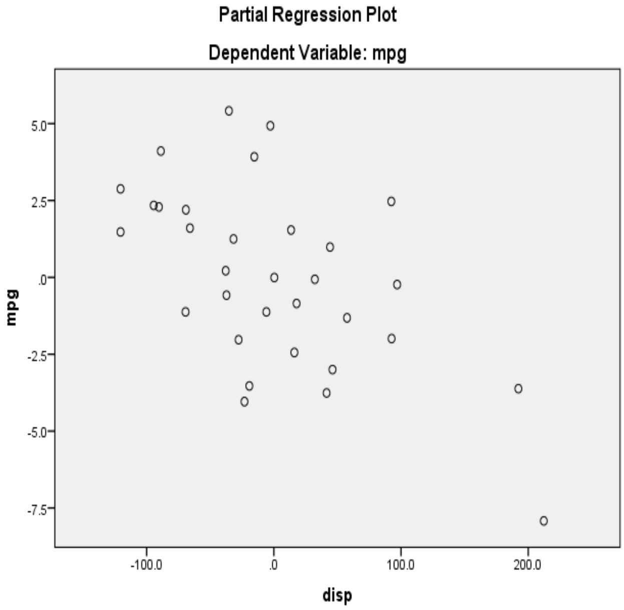
mpg = 37.157-.022disp-3.015wt-1.029carb

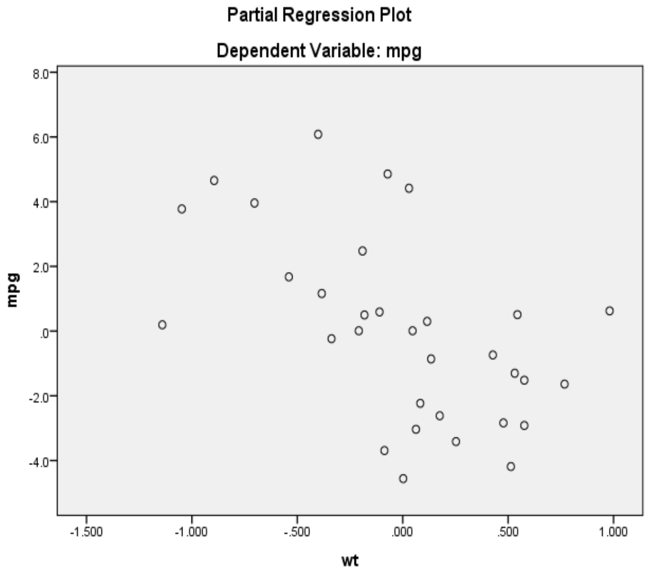
Actions:

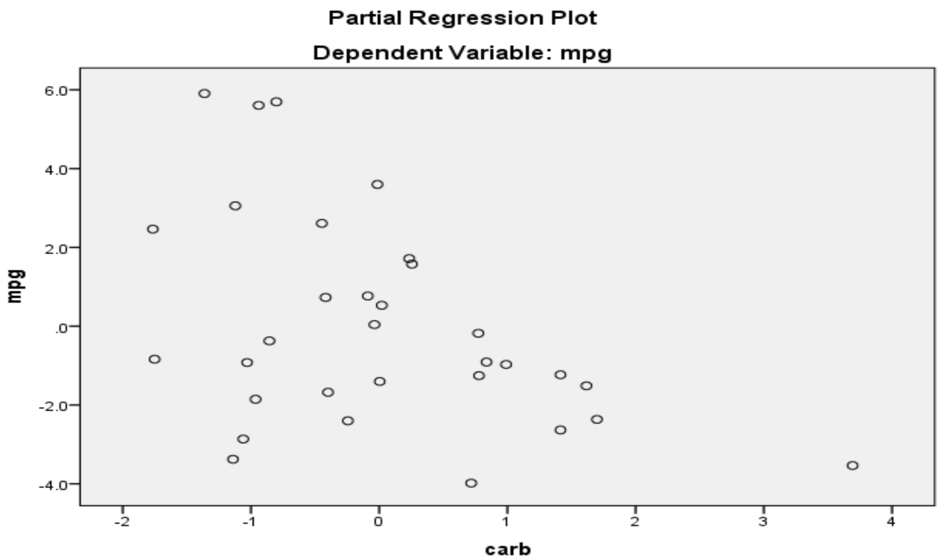
- Stepwise procedures are relatively cheap (less variables).
- Stepwise methods use a restricted search through the space of potential models and use a dubious hypothesis testing based method for choosing between models.
  Hence we adopt stepwise selection process.

d) Validation of Assumptions And Residual Analysis

1) Linearity Of Regression







Inference:

Only mpg and disp are almost linear but we observe that other two plots (mpg-wt and mpg-carb) are not forming a straight line so they are non-linear which indicates the need to transform and re-build the model.

2) Test for Autocorrelation

Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).

Following is the rule:

- if 1≤ DW ≤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[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.848[a]	.718	.709	3.2515	
2	.895[b]	.801	.787	2.7800	
3	.917[c]	.842	.825	2.5240	2.125

a. Predictors: (Constant), disp

b. Predictors: (Constant), disp, wt

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

d. Dependent Variable: mpg

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	6.816	29.029	20.091	5.5290	32
Residual	-4.5506	4.8706	.0000	2.3988	32
Std. Predicted Value	-2.401	1.617	.000	1.000	32
Std. Residual	-1.803	1.930	.000	.950	32

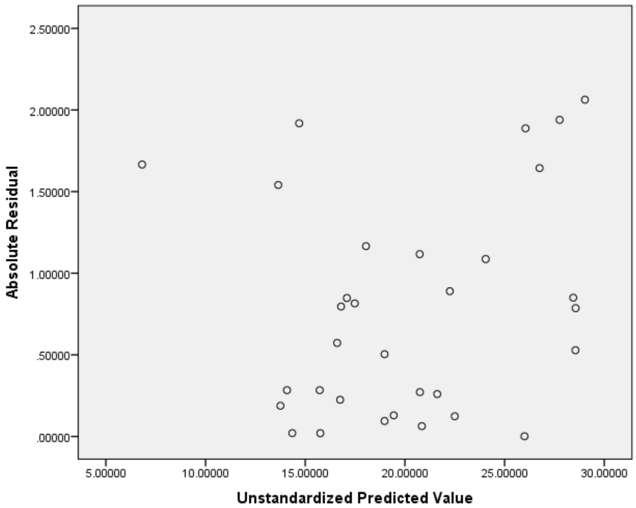
a. Dependent Variable: mpg

Inference:

Since model’s Durbin Watson statistic is 2.125 implies there is an evidence of no autocorrelation.

3) Test for Hetereoscedasticity

Heteroscedasticity may occur when some variables are skewed and others are not. Thus, checking that your data are normally distributed should cut down on the problem of heteroscedasticity.



It can be infer from the scatter plot that there is absolutely no linear relationship between absolute residuals and predicted values. The points are completely scattered in the plot hence we can say that the model is not heteroscedastic in nature.

Nonparametric Correlations

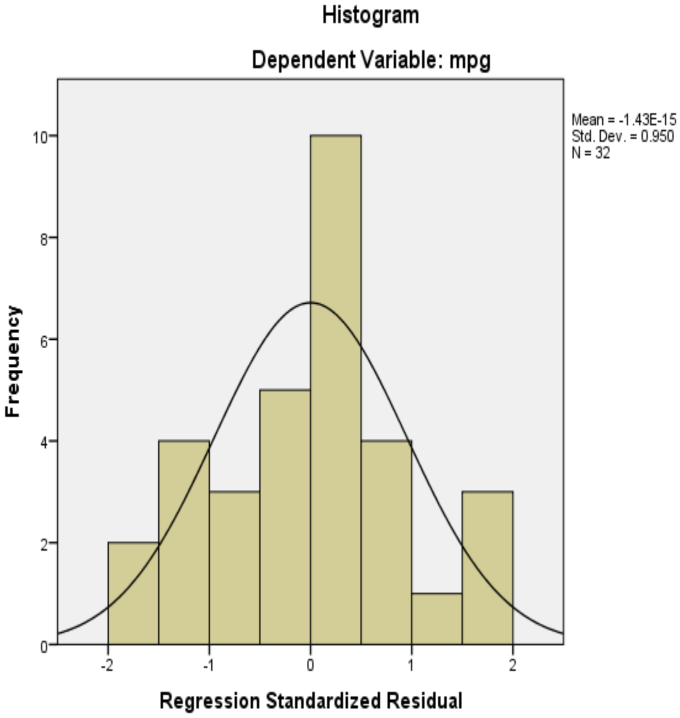
Correlations

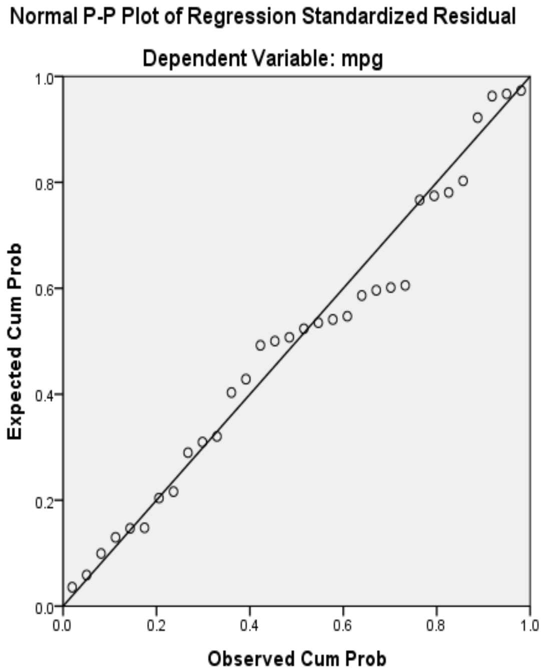
			Absolute Residual	Unstandardized Predicted Value
Spearman's rho	Absolute Residual	Correlation Coefficient	1.000	.188
		Sig. (2-tailed)	.	.303
		N	32	32
	Unstandardized Predicted Value	Correlation Coefficient	.188	1.000
		Sig. (2-tailed)	.303	.
		N	32	32

Since Spearman's rank correlation between predicted response and absolute residual is not significant. We fail to reject the hypothesis that correlation is not significant which further implies it is homoscedastic in nature.

Inference: It is homoscedastic in nature.

4) Normality of Errors





Ho :Errors are normally distributed.

$H1$: Errors are not close to normal distribution

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.139	32	.119	.967	32	.430

a. Lilliefors Significance Correction

Inference:

a) Using histogram we can see that errors are almost normally distributed since it is forming a symmetric line.

b) Using P-P plot we can see that errors lie around a straight line so errors are normally distributed.

c) Using K-S and S-W tests, we conclude that errors are normally distributed .Since p-values are greater than 0.05, we accept our null hypothesis that errors are normally distributed.
This implies errors are normally distributed.

5) Detection of Outliers

S.no	Unstandardized Predicted Value	Unstandardized Residual	Studentized Residual	Absolute Stundetized Residual	Centered Leverage Value
1	21.61853	-0.61853	-0.25974	0.25974	0.07862
2	20.84969	0.15031	0.06294	0.06294	0.07357
3	26.75431	-3.95431	-1.64388	1.64388	0.06049
4	20.75175	0.64825	0.27146	0.27146	0.07366
5	16.79797	1.90203	0.79601	0.79601	0.07256
6	20.73996	-2.63996	-1.11689	1.11689	0.09179
7	14.34879	-0.04879	-0.0202	0.0202	0.05344
8	22.25014	2.14986	0.89001	0.89001	0.05288
9	22.5007	0.2993	0.12392	0.12392	0.05304
10	18.97878	0.22122	0.09455	0.09455	0.10947
11	18.97878	-1.17878	-0.50381	0.50381	0.10947
12	15.72456	0.67544	0.28339	0.28339	0.07709
13	16.74968	0.55032	0.2247	0.2247	0.02727
14	16.59893	-1.39893	-0.5729	0.5729	0.03282
15	6.81644	3.58356	1.66596	1.66596	0.24246
16	13.64455	-3.24455	-1.54084	1.54084	0.27276
17	14.08509	0.61491	0.28366	0.28366	0.23115
18	27.76151	4.63849	1.93962	1.93962	0.07105
19	28.56279	1.83721	0.78598	0.78598	0.11112
20	29.02942	4.87058	2.06287	**2.06287**	**0.09371**
21	26.0506	-4.5506	-1.88747	1.88747	0.05635
22	17.48191	-1.98191	-0.81504	0.81504	0.0406
23	18.04657	-2.84657	-1.16572	1.16572	0.03278
24	13.75499	-0.45499	-0.18786	0.18786	0.048
25	14.69579	4.50421	1.91845	1.91845	0.10349
26	28.5539	-1.2539	-0.52824	0.52824	0.08429
27	25.99747	0.00253	0.00104	0.00104	0.0443
28	28.443	1.957	0.85008	0.85008	0.13686
29	15.75305	0.04695	0.01975	0.01975	0.08189
30	19.43945	0.26055	0.1294	0.1294	0.33236
31	17.08845	-2.08845	-0.8481	0.8481	0.01692
32	24.05241	-2.65241	-1.08677	1.08677	0.03374

If an observation has leverage more than 2p/n, where n is the no. of observations and p is the no. of variables, then ,It would be influential.

Rule:

- The observations corresponding to which the absolute value of studentized residuals lie beyond 3 can surely be taken as outliers.
- If observations corresponding to which the absolute value of studentized residuals lies between 2 and 3 have an levereage value greater than 2p/n , then those observations are also taken to be outlier.

Inference:

- We observe from above table that no value of absolute residual is > 3 implies there is no influential observation.

- Observation corresponding to which the absolute value of studentized residuals lies between 2 and 3 is 20th observation which is marked bold in above table. Its leverage is 0.09371< 0.1875 (=2p/n) which implies it is not an outlier.

  [Here p=3(no.of variables in the model) , n=32(no.of observations)]

Thus, it can be concluded that on using stepwise selection method, we end up with no outliers.

Sign- off note:

Where this can go next?

The most valuable (and correct) use of regression is in making predictions. Though, only a small minority of regression exercises end up by making a prediction, however.