Case Study – Linear Regression Analysis

Submitted by: Vishal Kumar

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:

Regression Analysis

In statistics, regression analysis is a statistical process for estimating the relationships among variables. More specifically, it helps to 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. Commonly, it estimates the conditional expectation of the dependent variable given the independent variables i.e. the average value of the dependent variable when the independent variables are fixed. The estimation target is a function of the independent variables called the regression function.

Regression analysis is also used to understand which among the independent variables are related to the dependent variable and to explore the forms of these relationships. Regression analysis can also be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships as correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data.

Assumptions in Linear Regression Model:

1. The regression model is linear in terms of parameter(s).
2. X_i and U_i are uncorrelated.
3. $E(U_i) = 0$ and $V(U_i) = \sigma^2$ for all i=1, 2, 3, . . ., n.
4. There does not exist any relationship between any independent variables i.e. no Multicollinearity.
5. Error variance is constant for all values of i=1, 2, 3. . ., n i.e. no Heteroscedasticity.
6. Errors are uncorrelated i.e. no Autocorrelation.
7. Functional form of the model is correct i.e. no Specification Bias.
8. Errors are normally distributed as $N(0, \sigma^2)$.

Pre-Processing Data

Before moving ahead with regression analysis, we have tried to explore the data using exploratory techniques. We have found some outliers and treated them as follows:

1. In Horse Power(hp), 31st observation was an outlier and has been replaced with series mean.
2. In Weight(wt), 16th and 17th observation were outliers and have been replaced with mean of 2 nearby points. After the values were replaced, we found a new outlier, but now we have replaced the new outlier with the maximum value in the data apart from the outlier.
3. In ¼ Time(qsec), 9th observation was an outlier and has been replaced with mean of 2 nearby points.

No missing value has been observed in our dataset. No outliers were detected now at this point.

Model Building

Fitting a Linear Regression Model

In linear regression we try to model the relationship between a dependent and one or more independent variable. Using the data points we try to find a function based on independent variables that are used to estimate dependent variable. Linear model means the model is linear in terms of the parameters. The functional form of linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots\ldots\ldots + \beta_n X_n + \varepsilon_i \qquad \text{where } \varepsilon_i \sim N(0, \sigma^2)$$

Where Y_I is the dependent variable

X_i are the independent variable

β_i's are the unknown parameters

$\epsilon_\text{I's}$ are the error component

Using the data we estimate β_i's, by the method of ordinary least square. The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter β. The estimators are unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors.

Our fitted model is:

$\widehat{mpg}$ = 20.869 + 0.998(cyl) - 0.012(disp) – 0.027(hp) + 1.455(drat) – 3.316(wt) + 0.107(qsec) + 0.281(vs) – 0.188(am) + 1.584(gear) – 1.197(carb)

We observe that displacement, horse power, weight, automatic and no. of carburettor affect the dependent variable inversely.

Testing the Significance of Individual Regression

The t-test is used to conduct hypothesis tests on the regression coefficients obtained in simple linear regression.

$$H_0: \beta_i = 0 \qquad \text{Vs} \qquad H_1: \beta_i \neq 0$$

Under H$_0$, the Test Statistics is given by:

$$T = \frac{\widehat{\beta}_i}{\sqrt{SE(\widehat{\beta}_i)}} \sim t_{n-k-1}$$

where K=no. of predictors in the model.

Testing criteria: If calculated t > tabulated t, we reject the hypothesis at α% level of significance and conclude that regression coefficients are significant. Or else we can use p-value to test the significance, if p-value is greater than 0.05, then we fail to reject the null hypothesis.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.869	22.755		.917	.369
	cyl	.998	1.002	.296	.996	.330
	disp	-.012	.011	-.240	-1.108	.280
	hp	-.027	.020	-.267	-1.379	.182
	drat	1.455	1.610	.129	.904	.376
	wt	-3.316	1.575	-.394	-2.105	.047
	qsec	.107	.838	.027	.128	.900
	vs	.281	2.146	.024	.131	.897
	am	-.188	2.156	-.016	-.087	.931
	gear	1.584	1.589	.194	.997	.330
	carb	-1.197	.507	-.321	-2.360	.028

Conclusion: As we can see most of the regressors have p-value greater than 0.05 i.e. not significant. Only few are significant. In the next section we will check the overall significance of the model.

Testing the Significance of Overall Regression

The overall significance of the regression can be tested with the ratio of the explained to the unexplained variance. This follows an F distribution with k-1 and n-k degrees of freedom, where n is number of observations and k is number of parameters estimated:

$$F_{k-1,n-k} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

If the calculated F ratio exceeds the tabular value of F at the specified level of significance and degrees of freedom, the hypothesis is accepted that the regression parameters are not all equal to zero and that R^2 is significantly different from zero. A "high" value for F statistic suggests a significant relationship between the dependent and independent, leading to the rejection of the null hypothesis that the coefficients of all explanatory variables are jointly zero.

$$H_0: \beta_1 = \beta_2 = \ldots \beta_k = 0 \qquad \text{Vs} \qquad H_1: not\ all\ \beta_i = 0$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	990.622	10	99.062	15.361	.000[b]
	Residual	135.425	21	6.449		
	Total	1126.047	31			

a. Dependent Variable: mpg

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

Conclusion: Since p-value is 0.00 < 0.05, hence we reject H_0 at 5% level of significance and conclude that there is a significant relation between the regressand and the regressors.

R Square and Adjusted R Square

R^2 measures the linear relationship between the independent variables and the dependent variable. It is defined as

$$R^2 = 1 - \frac{SSE}{SST}$$

which is the sum of squared errors divided by the total sum of squares. SST = SSE + SSR which are the total error and total sum of the regression squares. As independent variables are added SSR will continue to rise (and since SST is fixed) SSE will go down and R^2 will continually rise irrespective of how valuable the variables you added are.

The Adjusted R^2 is attempting to account for statistical shrinkage. Models with tons of predictors tend to perform better in sample than when tested out of sample. The adjusted R^2 "penalizes" you for adding the extra predictor variables that don't improve the existing model. It can be helpful in model selection. Adjusted R^2 will equal R^2 for one predictor variable. As you add variables, it will be smaller than R^2.

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.938[a]	.880	.822	2.5395

Conclusion: We see R^2 and Adjusted R^2 are quite high which means that model is a good fit. But there is further scope for improvisation.

After model building we see that the overall regression is significant but the individual regression coefficients are not coming out to be significant, this is possibly an indication of Multicollinearity. We will deal with this problem in the next section.

Multicollinearity

Multicollinearity is a state of very high inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

Causes of Multicollinearity:

- It is caused by an inaccurate use of dummy variables.
- It is caused by the inclusion of a variable which is computed from other variables in the data set.
- Multicollinearity can also result from the repetition of the same kind of variable.
- Generally occurs when the variables are highly correlated to each other.

Consequences of Multicollinearity:

- The partial regression coefficient due to multicollinearity may not be estimated precisely. The standard errors are likely to be high.
- Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
- Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

In the presence of high multicollinearity, the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data under study.

Detection and Removal of Multicollinearity using Correlation Analysis

The first step in detecting multicollinearity is to examine the correlation among the independent variables. We do this by looking at the correlation matrix. We take correlation value of 0.75 or more to be significant in either direction. Table given below shows the pairwise correlation between the independent variables:

	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
cyl	1.000	.928	.898	-.679	.825	-.558	-.814	-.522	-.564	.580
disp	.928	1.000	.865	-.684	.806	-.463	-.724	-.624	-.594	.540
hp	.898	.865	1.000	-.539	.697	-.619	-.752	-.445	-.437	.674
drat	-.679	-.684	-.539	1.000	-.729	.079	.447	.687	.745	-.125
wt	.825	.806	.697	-.729	1.000	-.246	-.533	-.724	-.637	.416
qsec	-.558	-.463	-.619	.079	-.246	1.000	.792	-.162	-.164	-.671
vs	-.814	-.724	-.752	.447	-.533	.792	1.000	.168	.283	-.634
am	-.522	-.624	-.445	.687	-.724	-.162	.168	1.000	.808	-.064
gear	-.564	-.594	-.437	.745	-.637	-.164	.283	.808	1.000	.115
carb	.580	.540	.674	-.125	.416	-.671	-.634	-.064	.115	1.000

Conclusion: Pair of variables for which correlation is high are suspected to be linearly associated. Hence from the table given above we highlight the correlation which are significant i.e. greater than 0.75. We also notice that cylinder variable seems to be correlated with 4 other variables, this variable might be collinear and in the next section we will use more sophisticated technique to assess multicollinearity.

Detection and Removal of Multicollinearity using Variance Inflation Factors (VIFs)

Variance inflation factor is a measure of degree to which variance of an ordinary least square estimate is inflated by collinearity. VIF for the kth predictor is given by:

$$VIF = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 value obtained by regressing the kth predictor on the remaining predictors. VIF greater than 10 is considered to be significant. Table given below shows VIF values for various predictors.

Model	Collinearity Statistics	
	Tolerance	VIF
cyl	.065	15.388
disp	.122	8.219
hp	.153	6.523
drat	.281	3.564
wt	.163	6.126
qsec	.126	7.942
vs	.178	5.622
am	.180	5.564
gear	.151	6.608
carb	.310	3.228

a. Dependent Variable: mpg

Conclusion

From the table we see that variance inflation factor for cylinder is greater than 10, hence we suspect that cylinder exhibit high correlation with other predictors. Hence we will remove this variable and fit our model again.

Regression after removing Cylinder variable from the model:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935^a	.874	.823	2.5390

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

Conclusion: Comparing new R^2 and adjusted R^2 with our previous model, we observe that R^2 has not decreased significantly, while adjusted R^2 is same.

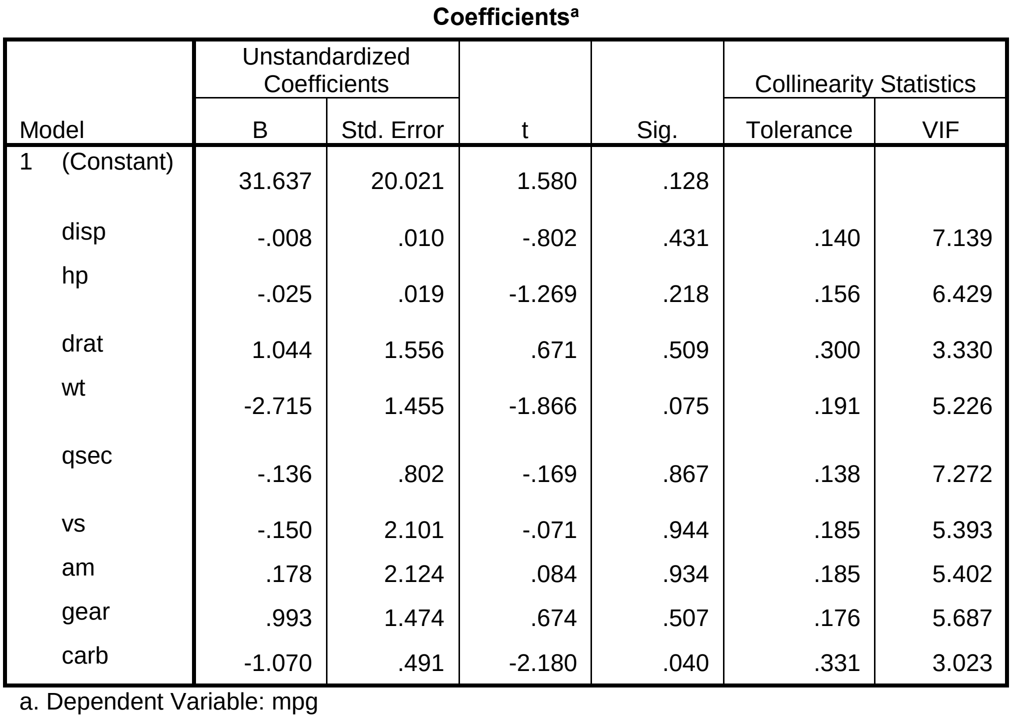
ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	984.222	9	109.358	16.964	.000^b
	Residual	141.825	22	6.447		
	Total	1126.047	31			

a. Dependent Variable: mpg

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

Conclusion: Over all regression is significant as the p-value is less than 0.05.



Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	31.637	20.021	1.580	.128		
	disp	-.008	.010	-.802	.431	.140	7.139
	hp	-.025	.019	-1.269	.218	.156	6.429
	drat	1.044	1.556	.671	.509	.300	3.330
	wt	-2.715	1.455	-1.866	.075	.191	5.226
	qsec	-.136	.802	-.169	.867	.138	7.272
	vs	-.150	2.101	-.071	.944	.185	5.393
	am	.178	2.124	.084	.934	.185	5.402
	gear	.993	1.474	.674	.507	.176	5.687
	carb	-1.070	.491	-2.180	.040	.331	3.023

a. Dependent Variable: mpg

Conclusion: We see that overall regression coefficients are still not coming out to be significant. But all the VIF's are now under 10. We don't have enough evidence to suspect multicollinearity now.

Our fitted model is:

$$\widehat{mpg} = 31.637 - 0.008(disp) - 0.025(hp) + 1.044(drat) - 2.715(wt) - 0.136(qsec) - 0.15(vs) - 0.178(am) + 0.993(gear) - 1.07(carb)$$

Parsimonious Modelling

Parsimonious means "mean" or "tight-fitted" normally, but in statistics it refers to simplicity. Therefore a Parsimonious model refers to the "simplest plausible model with the fewest possible number of variables".

Forward Selection

The simplest data-driven model building approach is called forward selection. In this approach, we add variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as its P-value is below some pre-set level.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	wt		Forward (Criterion: Probability-of-F-to-enter <= .150)
2	hp		Forward (Criterion: Probability-of-F-to-enter <= .150)
3	carb		Forward (Criterion: Probability-of-F-to-enter <= .150)
4	gear		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	.853[a]	.728	.719	3.1960
2	.913[b]	.834	.822	2.5402
3	.920[c]	.846	.830	2.4850
4	.930[d]	.864	.844	2.3784

a. Predictors: (Constant), wt

b. Predictors: (Constant), wt, hp

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

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

Conclusion: When we compare R^2 for model 4 to our previous model, we observe that R^2 has decreased but adjusted R^2 has increased, this means our new model explains more variability. Overall regression is significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41.934	2.503		16.753	.000
	wt	-7.175	.801	-.853	-8.958	.000
2	(Constant)	40.800	2.007		20.330	.000
	wt	-4.769	.847	-.567	-5.628	.000
	hp	-.044	.010	-.433	-4.300	.000
3	(Constant)	40.992	1.967		20.837	.000
	wt	-4.664	.832	-.555	-5.607	.000
	hp	-.038	.011	-.372	-3.485	.002
	carb	-.495	.326	-.133	-1.517	.140
4	(Constant)	31.151	5.542		5.621	.000
	wt	-3.391	1.044	-.403	-3.249	.003
	hp	-.032	.011	-.317	-2.989	.006
	carb	-1.059	.432	-.284	-2.451	.021
	gear	1.836	.972	.225	1.888	.070

a. Dependent Variable: mpg

Conclusion: Our fitted model is:

$$\widehat{mpg} = 31.151 - 0.32(hp) - 3.391(wt) + 1.836(gear) - 1.059(carb)$$

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 selection. Under this approach, one starts with fitting a model with all the variables of interest. Then the least significant variable is dropped, so long as it is not significant at our chosen critical level. We continue by successively re-fitting reduced models and applying the same rule until all remaining variables are statistically significant.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	disp, carb, am, drat, vs, qsec, wt, gear, hp[b]		Enter
2		vs	Backward (criterion: Probability of F-to-remove >= .200).
3		am	Backward (criterion: Probability of F-to-remove >= .200).
4		qsec	Backward (criterion: Probability of F-to-remove >= .200).
5		drat	Backward (criterion: Probability of F-to-remove >= .200).
6		disp	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	.935[a]	.874	.823	2.5390
2	.935[b]	.874	.830	2.4835
3	.935[c]	.874	.837	2.4318
4	.935[d]	.873	.843	2.3886
5	.933[e]	.871	.846	2.3663
6	.930[f]	.864	.844	2.3784

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

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

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

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

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

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

Conclusion

When we compare R^2 for model 6 to our previous model, we observe that R^2 has decreased but adjusted R^2 has increased, this means our new model explains more variability. Over all regression is significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	31.637	20.021		1.580	.128
	qsec	-.136	.802	-.035	-.169	.867
	gear	.993	1.474	.122	.674	.507
	am	.178	2.124	.015	.084	.934
	drat	1.044	1.556	.093	.671	.509
	carb	-1.070	.491	-.287	-2.180	.040
	vs	-.150	2.101	-.013	-.071	.944
	wt	-2.715	1.455	-.323	-1.866	.075
	hp	-.025	.019	-.243	-1.269	.218
	disp	-.008	.010	-.162	-.802	.431
2	(Constant)	32.356	16.927		1.911	.068
	qsec	-.171	.614	-.044	-.279	.783
	gear	.950	1.315	.116	.723	.477
	am	.219	1.997	.018	.110	.913
	drat	1.027	1.504	.091	.683	.502
	carb	-1.067	.478	-.286	-2.233	.036
	wt	-2.705	1.417	-.322	-1.909	.069
	hp	-.025	.019	-.245	-1.318	.201
	disp	-.008	.009	-.159	-.825	.418
3	(Constant)	32.942	15.730		2.094	.047
	qsec	-.194	.565	-.049	-.343	.734
	gear	.992	1.232	.121	.806	.428
	drat	1.037	1.470	.092	.706	.487
	carb	-1.062	.466	-.285	-2.279	.032

	wt	-2.788	1.174	-.332	-2.374	.026
	hp	-.025	.018	-.248	-1.378	.181
	disp	-.008	.009	-.159	-.843	.407
4	(Constant)	28.304	7.920		3.574	.001
	gear	1.170	1.098	.143	1.066	.297
	drat	1.039	1.443	.092	.720	.478
	carb	-1.016	.439	-.272	-2.317	.029
	wt	-2.819	1.150	-.335	-2.451	.022
	hp	-.022	.015	-.216	-1.426	.166
	disp	-.008	.009	-.155	-.837	.411
5	(Constant)	32.293	5.606		5.761	.000
	gear	1.384	1.047	.169	1.322	.198
	carb	-.995	.433	-.267	-2.296	.030
	wt	-3.107	1.068	-.369	-2.909	.007
	hp	-.020	.015	-.199	-1.341	.192
	disp	-.010	.008	-.196	-1.130	.269
6	(Constant)	31.151	5.542		5.621	.000
	gear	1.836	.972	.225	1.888	.070
	carb	-1.059	.432	-.284	-2.451	.021
	wt	-3.391	1.044	-.403	-3.249	.003
	hp	-.032	.011	-.317	-2.989	.006

a. Dependent Variable: mpg

Conclusion: Our fitted model is:

$$\widehat{mpg} = 31.151 - 0.032(hp) - 3.391(wt) + 1.836(gear) - 1.059(carb)$$

Stepwise Selection

Stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps. Backward stepwise selection involves starting off in a backward approach and then potentially adding back variables if they later appear to be significant. 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.

Variables Entered/Removed[a]

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

a. Dependent Variable: mpg

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.853[a]	.728	.719	3.1960
2	.913[b]	.834	.822	2.5402
3	.920[c]	.846	.830	2.4850
4	.930[d]	.864	.844	2.3784

a. Predictors: (Constant), wt

b. Predictors: (Constant), wt, hp

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

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

Conclusion

When we compare R^2 for model 4 to our previous model, we observe that R^2 has decreased but adjusted R^2 has increased, this means our new model explains more variability. Over all regression is significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41.934	2.503		16.753	.000
	wt	-7.175	.801	-.853	-8.958	.000
2	(Constant)	40.800	2.007		20.330	.000
	wt	-4.769	.847	-.567	-5.628	.000
	hp	-.044	.010	-.433	-4.300	.000
3	(Constant)	40.992	1.967		20.837	.000
	wt	-4.664	.832	-.555	-5.607	.000
	hp	-.038	.011	-.372	-3.485	.002
	carb	-.495	.326	-.133	-1.517	.140
4	(Constant)	31.151	5.542		5.621	.000
	wt	-3.391	1.044	-.403	-3.249	.003
	hp	-.032	.011	-.317	-2.989	.006
	carb	-1.059	.432	-.284	-2.451	.021
	gear	1.836	.972	.225	1.888	.070

a. Dependent Variable: mpg

Conclusion

Our fitted model is:

$$\widehat{mpg} = 31.151 – 0.032(hp) – 3.391(wt) +1.836(gear) – 1.059(carb)$$

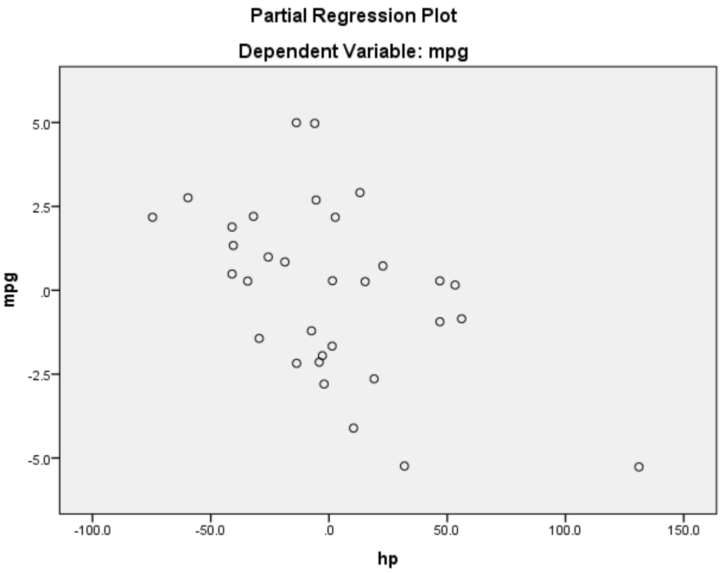
Validation of Assumptions and Residual Analysis

Linearity of Regression

The phrase linear in the expression Linear Regression refers to that response depends upon independent variables through a linear function of parameters. As the case when the response depends on independent variables through a non-linear function of independent variables we can include the variables in the model after some non-linear transformations. Here we are trying to test if a selected independent variable is included linearly or not. If it's non-linear then we first transform it and then rebuild the model. The technique which we use for this purpose is called Partial Regression Plots. The rule is that if a variable shows a linear relationship then it is to be included linearly otherwise it is to be included after a transformation. Under our model:

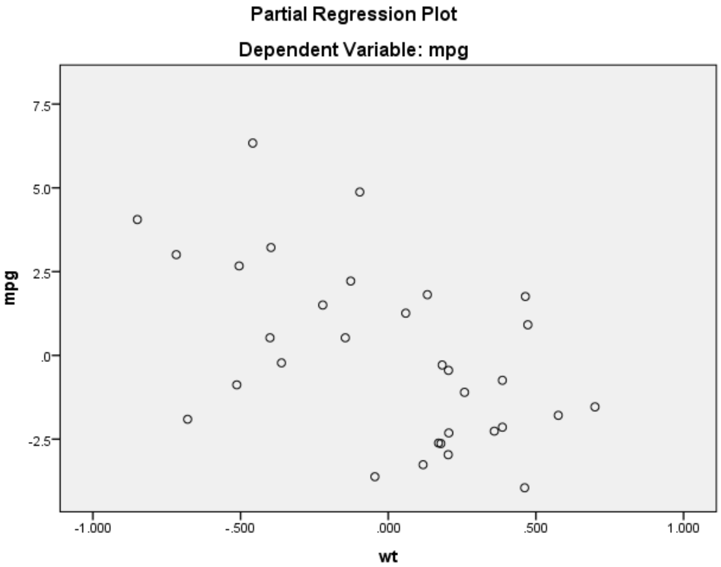
$$\widehat{mpg} = 31.151 - 0.032(hp) - 3.391(wt) + 1.836(gear) - 1.059(carb)$$

We make the following partial regression plot.



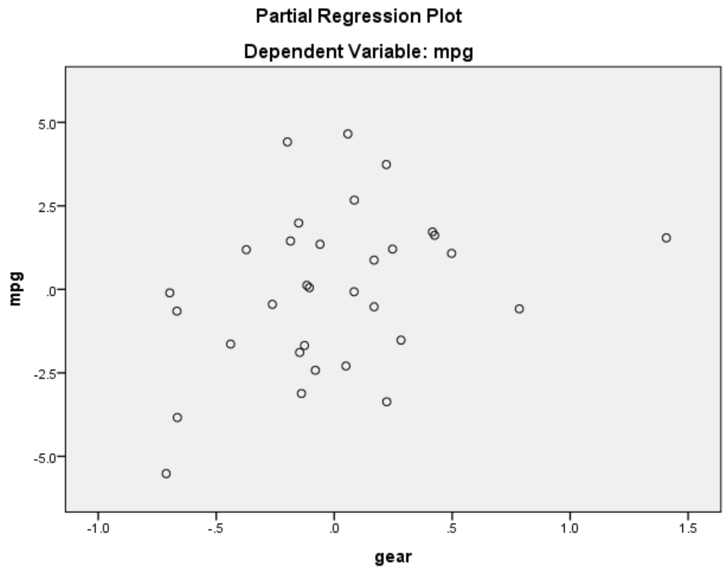
Conclusion

From the partial regression plot between "mpg" and "hp", we do not see any non-linear relationship between the two variables. Hence assumption of linearity holds here.



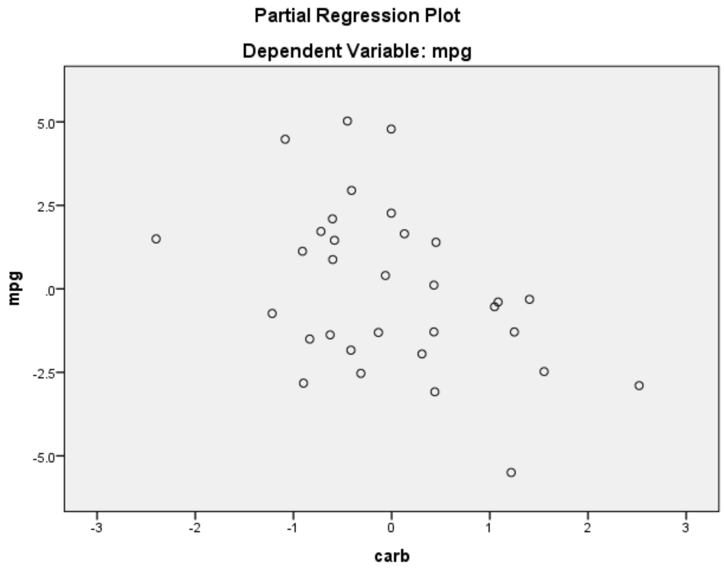
Conclusion

From the partial regression plot between "mpg" and "wt", we do not see any non-linear relationship between the two variables. Hence assumption of linearity holds here.



Conclusion

From the partial regression plot between "mpg" and "gear", we do not see any non-linear relationship between the two variables. Hence assumption of linearity holds here.



Conclusion

From the partial regression plot between "mpg" and "carb", we do not see any non-linear relationship between the two variables. Hence assumption of linearity holds here.

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. Presence of autocorrelation is detected by Durbin-Watson Test.

The Durbin–Watson statistic is a test statistic used to detect the presence of autocorrelation in the residuals (prediction errors) from a regression analysis, based on the assumption that the errors in regression model are generated by a first order autoregressive (AR(1)) process.

We set up the hypotheses as follows:

H_0: Absence of Autocorrelation Vs H_1: Presence of Autocorrelation

If e_t is the residual associated with the observation at time t, then the test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

where T is the number of observations. *d* is approximately equal to 2(1 − *r*), where *r* is the sample autocorrelation of the residuals. Decision rule based on the following number line:

Reject H_0 Positive correlation Exist.	No Autocorrelation Fail to reject H_0		Reject H_0 Negative correlation Exist
0	1	2	3 4

For our model:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.930[a]	.864	.844	2.3784	1.838

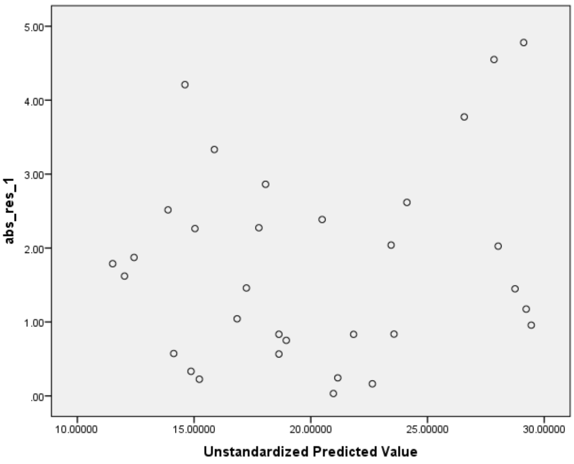
a. Predictors: (Constant), carb, gear, hp, wt

b. Dependent Variable: mpg

Conclusion: We observe for our model, value of d-Statistics is 1.836 ≈ 2, this indicates our model is free from autocorrelation.

Heteroscedasticity

Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it. Make an absolute residuals (y-axis) versus fitted (x-axis) plot if the plot doesn't exhibit any patterns, i.e. all the points are randomly scattered then there is no heteroscedasticity.



Conclusion

From the plot we see that points are scattered randomly, they don't exhibit any patterns, hence we don't have enough reason to believe presence of heteroscedasticity.

Presence of heteroscedasticity is also detected by a test based on Spearman's Rank Correlation. Spearman's rank correlation is computed between predicted dependent variable and absolute residual as follows:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)};$$

where $d_i = \hat{Y}_i - |\varepsilon_i|$

We set up the hypotheses as follows:

H_0: Absence of Heteroscedasticity
H_1: Presence of Heteroscedasticity
Define the t-statistic as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}}$$

If p value is < α/2 then reject the null hypothesis at α level of significance, i.e. Heteroscedasticity is present.

For the model under study, we have:

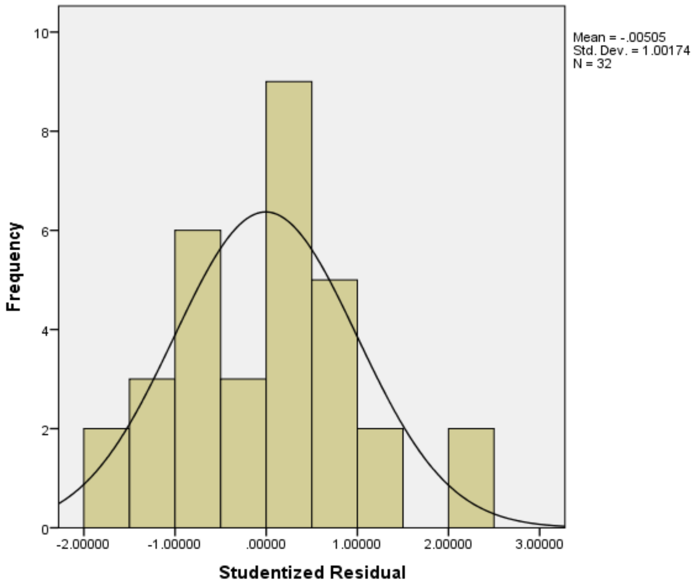
Spearman's rho		Unstandardized Predicted Value	Abs_res_1
Unstandardized Predicted Value	Correlation Coefficient	1.000	.055
	p-value		.763
abs_res_1	Correlation Coefficient	.055	1.000
	pvalue	.763	

Conclusion

We see that p-value is greater than 0.05, hence we fail to reject H_0 at 5% level of significance and conclude that Heteroscedasticity is absent from our model.

Normality of Errors

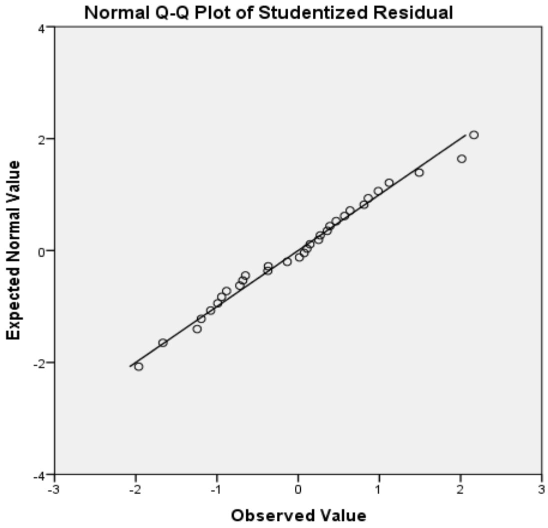
Histogram of Studentized Residuals: We can test the normality of error by plotting the frequency curve of the studentized residuals.



Conclusion

From the histogram, we see the curve is pretty much normal.

Q-Q Plots: We can test the normality of error by plotting quantiles of the studentized residuals vs standard normal distribution.



Conclusion

From the Q-Q plot, we see that the points are very close to the normal line. Hence we conclude that errors are normally distributed.

Outliers Detection

Outlier is simply an observation which is an aberration from the reaming observations. The outlier can have an atypical response and/or one or more independent variables' values. In this context two terms are important viz. leverage point and influential observations. An observation is said to be a leverage point if its remoteness doesn't affect the equation of the regression line, but the model summary statistics. An observation is said to be an influential observation if its remoteness does affect the equation of the regression line and the model summary statistics. The former can be viewed as less severe and the latter can be viewed as more severe. We want to detect such observations and delete them.

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.

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's a leverage point. It would be influential also if it has a significantly high value of Studentized Residual.

Index	Studentized Residual	Leverage Value
1	-0.37389	0.09222
2	0.0142	0.06543
3	-1.66585	0.06154
4	0.10819	0.06636
5	0.63811	0.04364
6	-1.07792	0.10271
7	0.86214	0.13457
8	0.38957	0.15368
9	0.07329	0.08606
10	0.25127	0.06993
11	-0.3696	0.06993
12	1.11985	0.07657
13	0.98566	0.03695
14	0.14506	0.03999
15	-0.719	0.07117
16	-1.96388	0.15643
17	0.27151	0.18069
18	2.01406	0.0667
19	0.5696	0.21784
20	2.16375	0.10611
21	-1.19419	0.11958
22	-0.99116	0.0381
23	-1.24402	0.03284
24	0.81012	0.10732
25	1.49022	0.08446
26	-0.64987	0.09005
27	-0.94376	0.15434
28	0.46735	0.22769
29	-0.68241	0.55546
30	0.35813	0.19072
31	-0.13538	0.47615
32	-0.88274	0.02475

Conclusion

There are no Studentized residuals that are greater than 3. Thus, we do not delete any observations.

In our model, p=4, n=32

So, 2p/n = 8/32 = 0.25

The observations corresponding to which the absolute value of studentized residuals lies between 2 and 3 are 18th and 20th. But their leverage values are less than 2p/n viz. 0.25. Hence they are not leverage points.

Hence after all model diagnosis our final fitted model is:

$$\widehat{mpg} = 31.151 - 0.032(hp) - 3.391(wt) + 1.836(gear) - 1.059(carb)$$