**Case Study-Linear Regression Analysis**

Submitted by: T. Vijaya

**INTRODUCTION**

Regression Analysis deals with studying relationship from a set of variables to another variable where the former is called a set of independent variables or regressors or predictors and the latter is called dependent variable or regressed variable or response. On the one hand, a single number, a correlation coefficient expresses the strength of the association between two variables. On the other there is a set of techniques known as regression which utilize the presence of an association between two variables to predict the values of one variable from those of another. We fit a predictive model to our data and use that model to predict values of the dependent variable from one or more independent variables. Simple regression seeks to predict an outcome from a single predictor whereas multiple regression seeks to predict an outcome from several predictors. In fact we are quantifying the effect of each regressors on the response. For example how different marketing drivers affect revenue of a particular company, how national income of countries are affected by different macroeconomic variables like unemployment, human development, agricultural and industrial growth.

A multiple linear regression is an equation in which each predictor variable has its own coefficient and the outcome variable is predicted from a combination of all the variables multiplied by their respective coefficients plus a residual term. It is expressed as follows:

Y=ß0+ß1X1+ß2X2…. + ßpXp +εi

Where Y is the outcome variable,ß1 is the coefficient of the first predictor(X1),intercept ß0 is the regression constant, εi is the difference between the predicted and the observed value of Y for the i th subject. The parameters ß1,ß2…,ßp are the partial regression coefficients .This equation is known as the multiple linear regression equation of Y upon X1,X2…,Xp. We seek to find the linear combination of predictors that correlate maximally with the outcome variable. The intercept is the distance from the origin to the point at which the line cuts the y-axis. In SPSS output the intercept b0 is referred as the constant.

***Fitting a Linear Regression Model***-In multiple linear regression (linear regression in SPSS) the outcome variable Y is predicted using the equation of straight line. Given that we have collected several values of Y and X1,…Xp, the unknown parameters in the equation can be calculated by fitting a model to the data(in this case a straight line) for which the sum of the squared differences between the line and the actual data points is minimized. This method is called the method of least squares.

***Assumptions of Multiple Linear Regression***-For testing hypotheses about the values of model parameters and to draw conclusions about a population based on a regression analysis, the following assumptions must be true: 1)**Normally distributed errors**-It is assumed that the residuals in the model are random, normally distributed, variables with a mean of 0.The differences between the model and the observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally.

2) **Homoscedasticity** - At each level of the predictor variable(s), the variance of the residual terms should be constant. They should have the same variance (homoscedasticity). An error term with non-constant variance is said to be heteroscedastic.

3)**Independent errors**-For any two observations the residual terms should be uncorrelated (or independent).This eventuality is sometimes described as a lack of autocorrelation.

4) **Independence**-It is assumed that all of the values of the outcome variable are independent (each value of outcome variable comes from a separate subject).

5) The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line.

***Model Parameters*-**The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However the ANOVA doesn’t tell us about the individual contribution of variables in the model. The values of ß1,….ßp represent the gradient of the regression line. Although this value is the slope of the regression line, it is more useful to think of this value as representing the change in the outcome associated with a unit change in the predictor. A bad model will have regression coefficients of zero for the predictors. A regression coefficient of zero means that the regression line is flat. If a variable significantly predicts an outcome, then it should have a ß value significantly different from zero. This hypothesis is tested using a t-test. The t-test is calculated by taking account of the standard error. We can use standard deviation of this distribution (known as standard error (s.e.)) as a measure of the similarity of ß values across samples. If the s.e. is very small, then it means that most samples have a ß value similar to the one in the sample collected. The t test tells us if the ß value is different from zero relative to the variation in ß values for similar samples. The ß values indicate to what degree each predictor affects the outcome if the effects of all other predictors are held constant. Standardized versions of the ß values are in many ways easier to interpret (coz they are are not dependent on units of measurement of the variables and are measured in standard deviation units and can be compared directly).They tell us the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor.

***R2 and adjusted R2***-When there are several predictors it does not make sense to look at the simple correlation coefficient and instead SPSS produces a multiple correlation coefficient (labeled R) .R is the correlation between the observed values of Y and the values of Y predicted by the multiple regression model .Large values of R represent a large correlation between the predicted and observed values of the outcome. R=1 represents a situation in which the model perfectly predicts the observed data.R2 is a measure of goodness of fit of the estimated regression.

In regression coefficient of determination (R2) is a statistical measure of how well the regression line approximates the real data points. It is the amount of variation in the outcome variable that is accounted for by the model. Adjusted R2 value indicates the loss of predictive power or shrinkage. Whereas R2 tells us how much of the variance in Y is accounted for by the regression model from our sample, the adjusted value tells us how much variance in Y would be accounted for if the model had been derived from the population from which the sample was taken. Adjusted R2 is an attempt to take account of the phenomenon of the R2 automatically and spuriously increasing when extra explanatory variables are added into the model. Unlike R2 adjusted R2 increases when a new explanatory is included only if the new explanatory improves the R2 more than would be expected by chance. Adjusted R2 does not have a same interpretation as R2 –while R2 is a measure of fit ,adjusted R2 is instead a comparative measure of suitability of alternative nested sets of explanators.

***Multicollinearity*** *is* a condition that can be extremely problematic and can lead to misleading and or inaccurate results. Multicollinearity (or collinearity) occurs when there are high inter-correlations among some set of the predictor variables in the regression model .If we have two predictors that are perfectly correlated, then the values of ß for each variable are interchangeable. As collinearity increases so do the standard errors of the ß coefficients, resulting in unstable predictor equations. This means that the estimated values of the regression coefficients (the ß values) will be unstable from sample to sample. High levels of collinearity increase the probability that a good predictor of the outcome will be found non-significant and rejected from the model. Multicollinearity between predictors makes it difficult to assess the individual importance of a predictor. If the predictors are highly correlated and each accounts for similar variance in the outcome, then how can we know which of the two is important? The model could include either one, interchangeably.

One way of identifying multicollinearity is to scan a correlation matrix of all of the predictor variables and see if any correlate very highly (means absolute correlation to be 0.75 and above).This is a good method but misses more subtle forms of multicollinearity. It may occur because several predictors, taken together, are related to some other predictors or set of predictors. For this reason it is important to test for multicollinearity when doing multiple regression. Another way is through the various collinearity diagnostics, one of which is the variance inflation factor (VIF) and the other one is Tolerance. The VIF indicates whether a predictor has a strong linear relationship with the other predictors.VIF is inversely related to the tolerance value. The threshold value is taken as 10.Tolerance is the percentage of the variance in a given predictor that cannot be explained by the other predictors. It is obtained by making each independent variable a dependent variable and regressing it against the remaining independent variables. When the tolerances are close to 0, there is high multicollinearity and the standard error of the regression coefficients will be inflated. Collinearity diagnostics measures how much regressors are related to other regressors and how this effects the stability and variance of the regression estimates. Eigen values provide an indication of how many distinct dimensions are there among the independent variables. When several Eigen values are close to 0, the variables are highly inter correlated .Small changes in the data values may lead to large changes in the estimates of the coefficients. Condition indices are the square roots of the ratios of the largest Eigen value to each successive Eigen value.A condition index greater than 30 suggests a serious problem with multicollinearity. Regression coefficient variance decomposition matrix shows the proportion of variance for each regression coefficient (and its associated variable) attributable to each condition index.

In order to examine collinearity, we first identify all condition indices above the threshold value of 30.Then for all condition indices exceeding the threshold,we identify variables with variance proportions above 0.90.A collinearity problem is indicated when a condition index identified as above the threshold value accounts for a substantial proportion of variance(0.90 or above)for two or more coefficients. Thus each row in the matrix with the proportions exceeding 0.90 for atleast two coefficients indicates significant correlations among the corresponding variables.

***Parsimonious Modeling***- Parsimonious means the simplest model/theory with the least assumptions and variables but with greatest explanatory power. A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. In Simultaneous/Forced entry (or Enter as it is known in SPSS) is a method in which all the relevant regressors are forced into the model simultaneously, so that the tests for each regression coefficient effectively put it at the end of the queue and test change in R2 in the presence of all other variables. The experimenter makes no decision about the order in which the variables are entered.

In (forward, backward and stepwise) methods the order in which predictors are entered into the model are based on a purely mathematical criterion. In the forward method, an initial model is defined that contains only the constant .The computer then searches for the predictor (out of the ones available) that best predicts the outcome variable-it does this by selecting the predictor that has the highest simple correlation with the outcome. If this predictor significantly improves the ability of the model to predict the outcome, then this predictor is retained in the model and the computer searches for the second predictor. The criterion used for selecting the second predictor is that it is the variable that has the largest semi partial correlation with the outcome. In statistical terms you can think of this like a partial correlation in that the computer correlates each of the predictors with the outcome while controlling for the effect of the first predictor. The reason that it is called a semi partial correlation is because the effects of the first predictor are partialled out of only the remaining predictors and are not controlled for in the outcome itself. This semi partial correlation gives a measure of how much ‘new variance’ in the outcome can be explained by each remaining predictor. The predictor that accounts for the most new variance is added to the model and if it makes a significant contribution to the predictive power of the model, it is retained and another predictor is considered.

The backward method is the opposite of the forward method in that the computer begins by placing all predictors in the model and then calculating the contribution of each one. Looking at the significance value of the t test for that predictor assesses the contribution of each predictor. This significance value is compared against a removal criterion (probability value for that test statistic).If a predictor meets the removal criterion it is removed from the model and the model is re-estimated for the remaining predictors. The contribution of the remaining predictors is then reassessed. The stepwise method in SPSS is the same as the forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor. As such the regression equation is constantly being reassessed to see whether any redundant predictors can be removed.

Validation of Assumptions and Residual Analysis

***Linearity of Regression*-**The term linear in linear regression refers to the response which depends upon independent variables through a linear function of parameters. When 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. Polynomial regression, log linear models are linear models only. We are interested in both things-non linearity in terms of parameters and non linearity in independent variables. We are trying to test if a selected independent variable is included linearly or not. 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.

Heteroscedasticity and autocorrelation

Heteroscedasticity is most commonly encountered in cross sectional data. There are various reasons for heteroscedasticity, such as the presence of outliers in the data, incorrect functional form of the regression model, incorrect transformation of data or mixing observations with different measures of scale. Autocorrelation is most commonly encountered in time series data.

Consequences of heteroscedasticity and autocorrelation -Does not alter the unbiasedness and consistency properties of least square estimators. They no longer are (BLUE-Best linear unbiased estimators).As a result t and F tests may not be reliable, resulting in erroneous conclusions regarding the statistical significance of the estimated regression coefficients.

We can check for heteroscedasticity by making a residuals vs fitted plot if the plot doesn’t exhibit any patterns i.e. all the points are randomly scattered then there is no heteroscedasticity. Test if Spearman’s rank correlation between predicted response and absolute residual is significant, if it is then there is heteroscedasticity.

We can check for autocorrelation by calculating the Durbin Watson (DW) test statistic. It tests for correlations between errors. Specifically it tests whether adjacent residuals are correlated. 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

***Normality of errors***-We need to check that the residuals are approximately normally distributed. Two common methods to check this assumption include using -a) a histogram (with a superimposed normal curve) and a Normal P-P plot. b) a Normal Q-Q plot of the studentized residuals.

***Outliers Detection***-An outlier is a case that differs substantially from the main trend of the data. Outliers can cause our model to be biased because they affect the values of the estimated regression coefficients. The outlier can have an atypical response and /or one or more independent variables values. In this context two terms are important –leverage point and influential observations. Leverage (sometimes called hat values) gauges the influence of the observed value of the outcome variable over predicted values. Leverage values lie between 0 and 1.However, cases with large leverage values will not necessarily have a large influence on the regression coefficients because they are measured on the outcome variables rather than the predictors. Leverage of an observation is calculated as the corresponding element of the hat matrix of regression: li= (Xt(XtX)-1X)i,i

If an observation has leverage more than 2p/n, where n is the number of observations and p is the number of variables, then it’s a leverage point. It would be influential also if it has a significantly high value of Studentized Residual. The differences between the values of the outcome predicted by the model and the values of the outcome observed in the sample are known as residuals. These residuals effectively represent the error present in the model. If model is a good fit residuals will be small. If any case stands out as having a large residual, then they could be outliers. Standardized residuals are the residuals divided by an estimate of their standard deviation. Studentized residual is the unstandardized residual divided by an estimate of its standard deviation that varies point by point. They provide a more precise estimate of the error variance of a specific case. It is calculated by using the formula:

**ri =ei /√MSRes(1-li).**

If abs (ri) >3 then ith observation is an influential observation. (Cook’s distance identifies cases that are influential or have a large effect on the regression solution and may be distorting the solution for the remaining cases in the analysis. While we cannot associate a probability with cook’s distance, we can identify problematic cases that have a score larger than the criteria computed. Using the formula: 4/(n-k-1) where n is the number of cases in the analysis, k is the number of independent variables .It is a measure of the overall influence of a case on the model.).If a point is a significant outlier on Y, but its cook’s distance <1, there is no need to delete that point since it does not have a large effect on the regression analysis. (A Common Rule)-The observations corresponding to which the absolute value of studentized residuals lie beyond 3 can surely be taken as outliers. If it lies between 2 and 3, we consider there leverage values. If the leverage is more than 2p/n, then those observations are taken to be outliers.

***Decision Rule*** –If p-value<0.05, we may reject the null hypothesis at 5% level of significance (here α=5). Consider the mtcars data set available in R. We have exported it to SPSS using function (write.csv (mtcars,”mtcars.sav”)).The data was extracted from the 1974 Motor Trend US Magazine and comprises gasoline mileage in miles per gallon and ten aspects of automobile design and performance for 32 automobiles(1973-74 models).It's a data frame with 32observations on 11 variables. The source of data is Henderson and Velleman (1981), building multiple regression models interactively.Biometrics,37, 391-441.

**Casewise diagnostics**-This option ,if selected, lists the observed value of the outcome, predicted value of the outcome, the difference between these values(the residual).

First of all we preprocess our data using Exploratory Data Analysis (E.D.A).After removing the outliers by treating them as missing observations, we proceed with linear regression. We build a linear regression model of mileage on rest of the attributes of the car. Given below are the number of variables and their respective symbols in our data set.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1.1: Description of Variables in Our Dataset   |  |  | | --- | --- | | S.No. | Variable Name | | 1 | Miles /(US) gallon (mpg) | | 2 | Number of cylinders (cyl) | | 3 | Displacement (disp) | | 4 | Gross horsepower (hp) | | 5 | Rear axle ratio (drat) | | 6 | Weight (wt) | | 7 | (1/4)th mile time(qsec) | | 8 | V/S(vs)(0=Vengine1=straightengine) | | 9 | Transmission (0=automatic,1=manual)(am) | | 10 | Number of forward gears(gear) | | 11 | Number of carburetors(carb) |   First of all we conduct linear regression analysis on the unprocessed mtcars data.  Table 1.2-Variables Entered/Removed  The table below shows the method of regression as Enter (simultaneous regression) and the dependent variable is mpg. In the variables entered column we have all the independent (predictor) variables | | | | |
| Model | Variables Entered | Variables Removed | Method |  |
| 1 | carb, am, vs, drat, qsec, gear, disp, hp, wt, cyla | . | Enter |

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R Square Change | F Change | df1 | df2 | Sig F change | a-dependent variable mpg |
| 1 | .932a | .869 | .807 | 2.6502 | .869 | 13.932 | 10 | 21 | 0.000 |

a)All requested variables entered

Table 1.3-Model Summary

|  |
| --- |
| 1) Large value of the multiple correlation coefficient R represents strong correlation between the predicted and observed values of the outcome.  2) R2 value is 0.869 which means that 86.9% of the variation in the outcome (mpg) is accounted by the predictors in the regression model of our sample.  3) Adjusted R2 is more conservative than R2.It gives us an idea of how well our model generalizes and indicates the loss of predictive power or shrinkage. The difference between adjusted R2 and R2 is 0.062(above 6.2%).This implies that if the model were derived from the population rather than a sample it would account for approximately 6.2 % less variance in the outcome.  4) The above model causes R2 to change from 0 to 0.869 and this change in the amount of variance explained gives rise to an F ratio of 13.932which is significant with a probability of 0.000. |

.

We are testing overall significance of the model by using F test and individual regressors significance by using t test

We set the following hypothesis (For F test)

Ho-Response variable is not significantly explained by the predictors

H1-Response variable is significantly explained by the predictors

| Table 1.4-ANOVAb | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
| 1 | Regression | 978.553 | 10 | 97.855 | 13.932 | .000a |
| Residual | 147.494 | 21 | 7.024 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), carb, am, vs, drat, qsec, gear, disp, hp, wt, cyl b. Dependent Variable: mpg | | | | | | |
| This table provides the results of a test of significance for R and R2 using the F statistic. In this analysis the p value(0.000) is well below (0.05) and therefore we can conclude that we can conclude that R ,R2 and adjusted R2 for the multiple regression conducted predicting mpg based on linear combination of predictors is statistically significant.  We set the following hypothesis (For t test) | | | | | | |

Ho-ß=0, i.e. Regressor variable does not significantly predict the outcome.

H1-ß≠0, i.e. Regressor variable significantly predicts the outcome.

| Table 1.5-Coefficientsa | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 12.303 | 18.718 |  | .657 | .518 |  |  |
| cyl | -.111 | 1.045 | -.033 | -.107 | .916 | .065 | 15.374 |
| Disp | .013 | .018 | .274 | .747 | .463 | .046 | 21.620 |
| hp | -.021 | .022 | -.244 | -.987 | .335 | .102 | 9.832 |
| Drat | .787 | 1.635 | .070 | .481 | .635 | .296 | 3.375 |
| wt | -3.715 | 1.894 | -.603 | -1.961 | .063 | .066 | 15.165 |
| Qsec | .821 | .731 | .243 | 1.123 | .274 | .133 | 7.528 |
| vs | .318 | 2.105 | .027 | .151 | .881 | .201 | 4.966 |
| am | 2.520 | 2.057 | .209 | 1.225 | .234 | .215 | 4.648 |
| Gear | .655 | 1.493 | .080 | .439 | .665 | .187 | 5.357 |
| carb | -.199 | .829 | -.053 | -.241 | .812 | .126 | 7.909 |
| a. Dependent Variable: mpg | | | | | | | | | |

The above table tells us about the parameters of the model. The first part of the table gives us estimates of ß values and these values indicate the individual contribution of each predictor to the model. .

2) Each of these ß values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the ß value differs significantly from zero. The t statistic tells whether a ß value is significantly different from zero. As the t test associated with a ß value is insignificant (p value>0.05) then the predictor variable does not significantly predict the model .The smaller the value of Sig (larger the magnitude of t) the greater the contribution of that predictor. As the standard error is very small (disp and hp), then it means that most samples have a ß value similar to the one in the sample collected

3) By looking at the last column of collinearity statistics we can draw inferences about multicollinearity. As the value of tolerance is nearer to 0 for three variables (cyl,disp,wt),it indicates high multicollinearity.This fact can be corroborated by using VIF. The VIF indicates whether a predictor has a strong linear relationship with the other predictors. The threshold value is taken as 10.VIF is above threshold of 10 for disp, cyl and wt.

| | Dimension | Eigenvalue | Condition Index | Table 1.6-Collinearity Diagnostics (Variance Proportions) | | | | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | (Constant) | cyl | Disp | hp | drat | Wt | qsec | vs | Am | gear | Carb | | 1 | 9.097 | 1.000 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | | 2 | 1.128 | 2.839 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .04 | .03 | .00 | .00 | | 3 | .564 | 4.016 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .06 | .12 | .00 | .00 | | 4 | .116 | 8.864 | .00 | .00 | .01 | .00 | .00 | .00 | .00 | .10 | .06 | .00 | .16 | | 5 | .048 | 13.722 | .00 | .00 | .04 | .10 | .01 | .00 | .00 | .34 | .18 | .01 | .02 | | 6 | .022 | 20.318 | .00 | .01 | .04 | .23 | .01 | .09 | .00 | .01 | .26 | .02 | .07 | | 7 | .010 | 30.682 | .00 | .24 | .08 | .01 | .05 | .00 | .00 | .11 | .26 | .21 | .03 | | 8 | .006 | 37.988 | .00 | .08 | .36 | .58 | .07 | .18 | .03 | .23 | .02 | .01 | .17 | | 9 | .006 | 39.155 | .00 | .03 | .00 | .06 | .55 | .00 | .00 | .01 | .00 | .48 | .01 | | 10 | .002 | 68.063 | .03 | .25 | .45 | .01 | .20 | .68 | .22 | .08 | .00 | .16 | .52 | | 11 | .000 | 143.421 | .97 | .38 | .01 | .00 | .12 | .05 | .75 | .03 | .06 | .12 | .00 | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

By observing the first column of eigen values we infer that for dimension (8,9,10,11) the variables are highly inter-correlated and their respective condition indices are above the threshold value of 30 indicating multicollinearity.

After removing the outliers by treating them as missing observations, we proceed with linear regression. We build a linear regression model of mileage on rest of the attributes of the car.

NOTE-We got outliers for wt,hp,qsec and carb. We have used the method of series mean to estimate missing observations.

| Table 1.7-Variables Entered/Removed | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp, cyla | . | Enter |
| a. All requested variables entered. | | | |

This table shows the method of regression as Enter (simultaneous regression) and the dependent variable is mpg.

In this method experimenter makes no decision about the order in which variables are entered.

| Table1.8-Model Summary | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .929a | .862 | .797 | 2.7172 | .862 | 13.152 | 10 | 21 | .000 |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp, cyl | | | | | | | | | |

1) For the given model, when all the variables are used as predictors, the multiple correlation coefficient is very high (0.929) indicating strong correlation between mpg and other variables (after removing the outliers). The next column gives us a value of R2 (0.859) which means that predictors account for 86.2% of the variation in mpg. When compared with previous model, there is an increase in R and R2.

2) The adjusted R2 indicates that we have a fairly good model explaining about 79.7% of the variance in mpg. The difference between R and R2 is (0.065) (about6. 5%).This shrinkage means that if the model were derived from the population rather than a sample it would account for approximately 6.5% less variance in the outcome.

3) The change statistics tell us that the model causes R2 to change from 0 to .862 and this change in amount of variance explained gives rise to an F ratio of 13.152 which is significant with a p value of 0.000.

| Table1.9-ANOVAb | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 971.001 | 10 | 97.100 | 13.152 | .000a |
| Residual | 155.046 | 21 | 7.383 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp, cyl | | | | | | |
| b. Dependent Variable: mpg | | | | | | |

The significance value in the ANOVA table when compared to a predetermined α, indicates whether changes in dependent variable values that accompany changes in independent variable values are significant. This table provides the results of a test of significance for R and R2 using the F statistic. In this analysis the p value(0.000) is well below (0.05) and therefore we can conclude that R ,R2 and adjusted R2 for the multiple regression conducted predicting mpg based on linear combination of predictors is statistically significant.

| Table 1.10-Coefficientsa | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 20.693 | 24.166 |  | .856 | .402 |  |  |
| Cyl | .756 | 1.099 | .224 | .688 | .499 | .062 | 16.188 |
| disp | -.027 | .013 | -.548 | -2.102 | .048 | .096 | 10.370 |
| drat | 1.944 | 1.752 | .172 | 1.110 | .280 | .271 | 3.685 |
| Vs | -.927 | 2.442 | -.078 | -.380 | .708 | .157 | 6.363 |
| Am | -1.272 | 2.484 | -.105 | -.512 | .614 | .155 | 6.451 |
| gear | .928 | 1.640 | .114 | .566 | .577 | .163 | 6.146 |
| SMEAN(wt) | -3.445 | 1.612 | -.394 | -2.137 | .045 | .193 | 5.194 |
| SMEAN(hp) | .000 | .024 | -.009 | -.037 | .971 | .116 | 8.603 |
| SMEAN(qsec) | .322 | .874 | .082 | .369 | .716 | .133 | 7.508 |
| SMEAN(carb) | -1.462 | .635 | -.317 | -2.302 | .032 | .345 | 2.900 |
| a. Dependent Variable: mpg | | | | | | | | |

1) Each of these ß values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the ß value differs significantly from zero. The t statistic tells whether a ß value is significantly different from zero. As the t test associated with a ß value is significant for (disp,wt and carb as p value<0.05) we conclude that these three predictor variables significantly predict the model .The smaller the value of Sig (larger the magnitude of t) the greater the contribution of that predictor.

3) By looking at the last column of collinearity statistics we can draw inferences about multicollinearity. As the value of tolerance is nearer to 0 for two variables (cyl, disp) ,it indicates high multicollinearity.This fact can be corroborated by using VIF. The VIF indicates whether a predictor has a strong linear relationship with the other predictors. The threshold value is taken as 10.VIF is above threshold of 10 for disp and cyl.

| Table1.11-Collinearity Diagnosticsa | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | | | | | | | | |
| (Constant) | cyl | Disp | drat | Vs | am | gear | SMEAN(wt) | SMEAN(hp) | SMEAN(qsec) | SMEAN(carb) |
| 1 | 1 | 9.168 | 1.000 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | 1.106 | 2.879 | .00 | .00 | .00 | .00 | .03 | .02 | .00 | .00 | .00 | .00 | .00 |
| 3 | .525 | 4.178 | .00 | .00 | .00 | .00 | .06 | .09 | .00 | .00 | .00 | .00 | .01 |
| 4 | .105 | 9.342 | .00 | .00 | .03 | .00 | .02 | .04 | .00 | .00 | .00 | .00 | .46 |
| 5 | .047 | 13.959 | .00 | .00 | .10 | .00 | .27 | .07 | .00 | .04 | .08 | .00 | .01 |
| 6 | .017 | 23.168 | .00 | .00 | .29 | .01 | .03 | .02 | .00 | .11 | .37 | .01 | .17 |
| 7 | .014 | 25.906 | .00 | .04 | .06 | .12 | .25 | .34 | .01 | .12 | .21 | .00 | .18 |
| 8 | .010 | 29.740 | .00 | .02 | .11 | .00 | .00 | .14 | .39 | .01 | .02 | .01 | .05 |
| 9 | .004 | 45.685 | .01 | .04 | .23 | .57 | .02 | .07 | .16 | .27 | .08 | .03 | .01 |
| 10 | .003 | 51.951 | .00 | .79 | .15 | .13 | .11 | .22 | .02 | .44 | .05 | .03 | .03 |
| 11 | .000 | 181.566 | .99 | .10 | .02 | .17 | .20 | .00 | .42 | .02 | .18 | .91 | .09 |
| a. Dependent Variable: mpg | | | | | | | | | | | | | | |

After removing the outliers we observe that the eigen values are nearer to zero for (9,10,11) respectively. Also the condition indices for these three dimensions is well above the threshold of 30 indicating multicollinearity.

| (Multicollinearity through correlation matrix)Table 1.12-Correlations(Spearman s rho) | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Cyl | Disp | drat | Vs | Am | gear | SMEAN(wt) | SMEAN(hp) | SMEAN(qsec) | SMEAN(carb) |
|  | Cyl | 1.000 | .928\*\* | -.679\*\* | -.814\*\* | -.522\*\* | -.564\*\* | .762\*\* | .898\*\* | -.535\*\* | .559\*\* |
| . | .000 | .000 | .000 | .002 | .001 | .000 | .000 | .002 | .001 |
| disp | .928\*\* | 1.000 | -.684\*\* | -.724\*\* | -.624\*\* | -.594\*\* | .706\*\* | .865\*\* | -.444\* | .537\*\* |
| .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | .011 | .002 |
| drat | -.679\*\* | -.684\*\* | 1.000 | .447\* | .687\*\* | .745\*\* | -.633\*\* | -.539\*\* | .058 | -.122 |
| .000 | .000 | . | .010 | .000 | .000 | .000 | .001 | .754 | .505 |
| Vs | -.814\*\* | -.724\*\* | .447\* | 1.000 | .168 | .283 | -.464\*\* | -.752\*\* | .771\*\* | -.620\*\* |
| .000 | .000 | .010 | . | .357 | .117 | .007 | .000 | .000 | .000 |
| Am | -.522\*\* | -.624\*\* | .687\*\* | .168 | 1.000 | .808\*\* | -.697\*\* | -.445\* | -.162 | -.136 |
| .002 | .000 | .000 | .357 | . | .000 | .000 | .011 | .376 | .458 |
| gear | -.564\*\* | -.594\*\* | .745\*\* | .283 | .808\*\* | 1.000 | -.559\*\* | -.437\* | -.181 | .028 |
| .001 | .000 | .000 | .117 | .000 | . | .001 | .012 | .323 | .880 |
| SMEAN(wt) | .762\*\* | .706\*\* | -.633\*\* | -.464\*\* | -.697\*\* | -.559\*\* | 1.000 | .621\*\* | -.258 | .312 |
| .000 | .000 | .000 | .007 | .000 | .001 | . | .000 | .154 | .082 |
| SMEAN(hp) | .898\*\* | .865\*\* | -.539\*\* | -.752\*\* | -.445\* | -.437\* | .621\*\* | 1.000 | -.598\*\* | .704\*\* |
| .000 | .000 | .001 | .000 | .011 | .012 | .000 | . | .000 | .000 |
| SMEAN(qsec) | -.535\*\* | -.444\* | .058 | .771\*\* | -.162 | -.181 | -.258 | -.598\*\* | 1.000 | -.602\*\* |
| .002 | .011 | .754 | .000 | .376 | .323 | .154 | .000 | . | .000 |
| SMEAN(carb) | .559\*\* | .537\*\* | -.122 | -.620\*\* | -.136 | .028 | .312 | .704\*\* | -.602\*\* | 1.000 |
| .001 | .002 | .505 | .000 | .458 | .880 | .082 | .000 | .000 | . |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | | | | | | | |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | | | | | | | | |

All those correlation coefficients whose value is more than the absolute threshold value of 0.75 have been chosen.

| Table 1.13-Correlations -Spearman's Rho | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | disp | Drat | Am | SMEAN(wt) | SMEAN(qsec) | SMEAN(carb) |
|  | Disp | R | 1.000 | -.684\*\* | -.624\*\* | .706\*\* | -.444\* | .537\*\* |
| Sig(2-tail) | . | .000 | .000 | .000 | .011 | .002 |
| Drat | R | -.684\*\* | 1.000 | .687\*\* | -.633\*\* | .058 | -.122 |
| Sig(2-tail) | .000 | . | .000 | .000 | .754 | .505 |
| Am | R | -.624\*\* | .687\*\* | 1.000 | -.697\*\* | -.162 | -.136 |
| Sig(2-tail) | .000 | .000 | . | .000 | .376 | .458 |
| SMEAN(wt) | R | .706\*\* | -.633\*\* | -.697\*\* | 1.000 | -.258 | .312 |
| Sig(2-tail) | .000 | .000 | .000 | . | .154 | .082 |
| SMEAN(qsec) | R | -.444\* | .058 | -.162 | -.258 | 1.000 | -.602\*\* |
| Sig(2-tail) | .011 | .754 | .376 | .154 | . | .000 |
| SMEAN(carb) | R | .537\*\* | -.122 | -.136 | .312 | -.602\*\* | 1.000 |
| Sig(2-tail) | .002 | .505 | .458 | .082 | .000 | . |
| \*\*Correlation is significant at the 0.01 level (2-tailed). | | | | | | | | |
| \*Correlation is significant at the 0.05 level (2-tailed). | | | | | | | | |

After removing cyl,vs,hp and gear all the correlations are below 0.5.Inspection of the correlation matrix for high pairwise correlations, is not sufficient since multicollinearity can exist with no pairwise correlations being high. So we remove those variables whose correlation value is more than 0.75.

| Table 1.14 Variables Entered/Removed | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1== | SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), dispa | . | Enter |
| a. All requested variables entered. | | | |

This table shows the method of regression as Enter and the dependent variable is mpg and we have 9 predictor variables .It means that we have entered all independent variables into the model simultaneously. We have removed Multicollinearity by removing the variable cyl.

Table 1.15 Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .927a | .859 | .802 | 2.6845 | .859 | 14.918 | 9 | 22 | .000 |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp | | | | | | | | | |
| b. Dependent Variable: mpg | | | | | | | | | |

For the given model, when all the variables except cyl are used as predictors, the multiple correlation coefficient is very high (0.927) indicating strong correlation between mpg and other variables. The next column gives us a value of R2 (0.859) which means that predictors account for 85.9% of the variation in mpg.

The adjusted R2 gives us some idea of how well our model generalizes and ideally its value should be same or very close to R2.The difference for the final model is (0.057) (about 5.7%).This shrinkage means that if the model were derived from the population rather than a sample it would account for approximately 5.7% less variance in the outcome.

The change statistics tell us the change in the F ratio resulting from each block of the hierarchy. So the model causes R2 to change from 0 to 0.859 and this change in amount of variance explained gives rise to an F ratio of 14.918 which is significant with a probability(p value=0.000) less than 0.05.

| Table 1.16 -ANOVAb | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
| 1 | Regression | 967.508 | 9 | 107.501 | 14.918 | .000a |
| Residual | 158.539 | 22 | 7.206 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp | | | | | | |
| b. Dependent Variable: mpg | | | | | | |

The next part of the output contains ANOVA that tests whether the model is significantly better at predicting the outcome than using the mean .F ratio is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model. If the improvement due to fitting the regression model is much greater than the inaccuracy within the model then the value of F is greater than one and SPSS calculates the exact probability of obtaining the value of F. The value of F is (14.918) which is highly significant (p=0.000<0.05).Hence we may reject our null hypothesis and conclude that response variable is significantly explained by the predictors.

| Table 1.17-Coefficientsa | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 26.253 | 22.499 |  | 1.167 | .256 |  |  |
| disp | -.022 | .011 | -.456 | -2.063 | .051 | .131 | 7.640 |
| drat | 1.656 | 1.681 | .147 | .985 | .335 | .288 | 3.474 |
| Vs | -1.287 | 2.357 | -.108 | -.546 | .591 | .165 | 6.071 |
| Am | -.753 | 2.338 | -.062 | -.322 | .751 | .171 | 5.855 |
| gear | .607 | 1.553 | .074 | .391 | .700 | .177 | 5.650 |
| SMEAN(wt) | -2.820 | 1.316 | -.323 | -2.143 | .043 | .282 | 3.545 |
| SMEAN(hp) | 2.143E-5 | .024 | .000 | .001 | .999 | .117 | 8.577 |
| SMEAN(qsec) | .211 | .848 | .054 | .249 | .806 | .138 | 7.251 |
| SMEAN(carb) | -1.380 | .616 | -.300 | -2.239 | .036 | .357 | 2.798 |
| a. Dependent Variable: mpg | | | | | | | | |

1) The above table tells us about the parameters of the model. The first part of the table gives us estimates

For these ß values and these values indicate the individual contribution of each predictor to the model. The ß values tell us about the relationship between mpg and each predictor. The ß values indicate to what degree each predictor affects the outcome if the effects of all other predictors are held constant.

2) Each of these ß values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the ß value differs significantly from zero. The t statistic tells whether a ß value is significantly different from zero. As the t test associated with a ß value is significant (in case of wt and carb p value<0.05) then the predictor is making significant contribution to the model .The smaller the value of Sig (larger the magnitude of t) the greater the contribution of that predictor.

3) (Standardized ß =-0.323 for wt) indicates that as wt increases by one standard deviation mpg decreases by 0.323 standard deviations .This interpretation is true only if the effects of other predictors in the model are held constant. (Standardized ß =-0.300 for carb) indicates that as carb increases by one standard deviation mpg decreases by 0.300 standard deviations .This interpretation is true only if the effects of other predictors in the model are held constant.

4)As the VIF values are below the threshold value of 10 it indicates absence of multicollinearity after removal of cyl.

| Table 1.18Collinearity Diagnosticsa | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | | | | | | | |
| (Constant) | disp | drat | vs | Am | gear | SMEAN(wt) | SMEAN(hp) | SMEAN(qsec) | SMEAN(carb) |
| 1 | 1 | 8.231 | 1.000 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | 1.053 | 2.796 | .00 | .00 | .00 | .03 | .03 | .00 | .00 | .00 | .00 | .00 |
| 3 | .525 | 3.959 | .00 | .00 | .00 | .07 | .10 | .00 | .00 | .00 | .00 | .01 |
| 4 | .101 | 9.026 | .00 | .05 | .00 | .02 | .04 | .00 | .00 | .00 | .00 | .48 |
| 5 | .046 | 13.313 | .00 | .12 | .01 | .32 | .08 | .01 | .06 | .08 | .00 | .01 |
| 6 | .017 | 22.022 | .00 | .39 | .00 | .01 | .00 | .00 | .15 | .43 | .00 | .22 |
| 7 | .012 | 25.892 | .00 | .19 | .14 | .19 | .27 | .01 | .37 | .17 | .00 | .09 |
| 8 | .010 | 29.147 | .00 | .03 | .01 | .01 | .46 | .48 | .07 | .00 | .01 | .12 |
| 9 | .004 | 43.577 | .01 | .21 | .71 | .05 | .02 | .13 | .19 | .10 | .05 | .00 |
| 10 | .000 | 163.899 | .99 | .00 | .13 | .31 | .00 | .38 | .15 | .22 | .93 | .07 |
| a. Dependent Variable: mpg | | | | | | | | | | | | | |

We first identify all condition indices above the threshold value of 30.Then for all condition indices exceeding the threshold, we identify variables with variance proportions above 0.90.A collinearity problem is indicated when a condition index identified as above the threshold value accounts for a substantial proportion of variance(0.90 or above)for two or more coefficients. Thus each row in the matrix with the proportions exceeding 0.90 for atleast two coefficients indicates significant correlations among the corresponding variables. (Ignoring the constant value of 0.99)Hence after observing the collinearity diagnostics table above we conclude that multicollinearity is not present in the data.

After detection and removal of multicollinearity we use parsimonious modeling to obtain the best model. First we use the method of forward selection.

| Table 1.19-Variables Entered/Removeda | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | Disp | . | Forward (Criterion: Probability-of-F-to-enter <= .150) |
| 2 | SMEAN(wt) | . | Forward (Criterion: Probability-of-F-to-enter <= .150) |
| 3 | SMEAN(carb) | . | Forward (Criterion: Probability-of-F-to-enter <= .150) |
| a. Dependent Variable: mpg | | | |

The method of regression used above is forward. When the forward method is employed SPSS begins with the model that includes the variable disp and then adds single predictors into the model based on entry and exit criterion. The current model is compared to the model when the concerned predictor is removed. If the removal of that predictor makes a significant difference to how well the model fits the observed data ,then the computer retains that predictor(because the model is better if the predictor is included).If the removal of the predictor makes little difference to the model ,then the computer rejects that predictor.

Table 1.20

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .848a | .718 | .709 | 3.2515 | .718 | 76.513 | 1 | 30 | .000 |
| 2 | .898b | .806 | .792 | 2.7481 | .087 | 12.998 | 1 | 29 | .001 |
| 3 | .922c | .850 | .834 | 2.4541 | .045 | 8.363 | 1 | 28 | .007 |

The next section of output describes the overall model (so it tells us how well the model predicts miles per gallon)

For model 1, when only one variable disp is used as a predictor, the simple correlation between disp and mpg is 0.848.The next column gives us a value of R2.For the first model its value is0.718 which means that displacement accounts for 71.8%of the variation in mpg. In the final model , its value has increased to 0.850.Therefore if displacement accounts for 71.8% ,we can tell that weight accounts for an additional 8.7 % and carburetors accounts for an additional 4.5 %.So the inclusion of the two new predictors has explained a small amount of variation in mpg.

The change statistics tell us the change in the F ratio resulting from each block of the hierarchy. So model 1 causes R2 to change from 0 to 0.718 and this change in amount of variance explained gives rise to an F ratio of 76.513 which is significant with a probability(p value=0.000) less than 0.05. Similarly for model 3 the addition of new predictor (carb in model 3) causes R2 to increase by 0.045 and this change in the amount of variance that can be explained gives rise to an increase in F ratio of 8.363,which is again significant (p value =0.007) less than 0.05.

| Table 1.21-ANOVAd | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
| 1 | Regression | 808.888 | 1 | 808.888 | 76.513 | .000a |
| Residual | 317.159 | 30 | 10.572 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 2 | Regression | 907.044 | 2 | 453.522 | 60.054 | .000b |
| Residual | 219.004 | 29 | 7.552 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 3 | Regression | 957.411 | 3 | 319.137 | 52.989 | .000c |
| Residual | 168.636 | 28 | 6.023 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), disp | | | | | | |
| b. Predictors: (Constant), disp, SMEAN(wt) | | | | | | |
| c. Predictors: (Constant), disp, SMEAN(wt), SMEAN(carb) | | | | | | |
| d. Dependent Variable: mpg | | | | | | |

The F ratio represents the ratio of improvement in prediction as a result of fitting the model relative to the inaccuracy that still exists in the model. For the initial model F ratio is 76.513 which is very unlikely to have happened by chance(p value <0.05).In the final model the value of F is even lower (52.989),which is also highly significant (p value <0.05). We infer that the initial model significantly improved our ability to predict the outcome variable as compared to the final model with extra predictors.

| Table 1.22-Coefficientsa | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 29.600 | 1.230 |  | 24.070 | .000 |
| disp | -.041 | .005 | -.848 | -8.747 | .000 |
| 2 | (Constant) | 36.979 | 2.296 |  | 16.109 | .000 |
| disp | -.029 | .005 | -.589 | -5.405 | .000 |
| SMEAN(wt) | -3.431 | .952 | -.393 | -3.605 | .001 |
| 3 | (Constant) | 38.207 | 2.094 |  | 18.250 | .000 |
| disp | -.025 | .005 | -.511 | -5.067 | .000 |
| SMEAN(wt) | -3.182 | .854 | -.364 | -3.726 | .001 |
| SMEAN(carb) | -1.075 | .372 | -.233 | -2.892 | .007 |
| a. Dependent Variable: mpg | | | | | | |

The ß values tell us about the relationships between mpg and each predictor. Each of these ß values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the ß value differs significantly from zero. The t statistic tells whether a ß value is significantly different from zero. As the t test associated with a ß value is significant then the predictor is making significant contribution to the model .The smaller the value of Sig (larger the magnitude of t) the greater the contribution of that predictor.

| Table 1.23-Excluded Variablesd | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Beta In | T | Sig. | Partial Correlation | Collinearity Statistics |
| Tolerance |
| 1 | drat | .160a | 1.169 | .252 | .212 | .496 |
| vs | .125a | .905 | .373 | .166 | .495 |
| am | .152a | 1.277 | .212 | .231 | .650 |
| gear | .014a | .115 | .909 | .021 | .691 |
| SMEAN(wt) | -.393a | -3.605 | .001 | -.556 | .565 |
| SMEAN(hp) | -.310a | -1.716 | .097 | -.304 | .271 |
| SMEAN(qsec) | .095a | .885 | .384 | .162 | .818 |
| SMEAN(carb) | -.264a | -2.732 | .011 | -.452 | .830 |
| 2 | drat | -.007b | -.054 | .957 | -.010 | .418 |
| vs | .099b | .843 | .406 | .157 | .493 |
| am | -.062b | -.511 | .613 | -.096 | .464 |
| gear | -.079b | -.774 | .445 | -.145 | .649 |
| SMEAN(hp) | -.253b | -1.647 | .111 | -.297 | .268 |
| SMEAN(qsec) | .091b | 1.006 | .323 | .187 | .818 |
| SMEAN(carb) | -.233b | -2.892 | .007 | -.480 | .821 |
| 3 | drat | .135c | 1.112 | .276 | .209 | .359 |
| vs | -.049c | -.414 | .682 | -.079 | .390 |
| am | .041c | .359 | .722 | .069 | .416 |
| gear | .061c | .586 | .563 | .112 | .498 |
| SMEAN(hp) | .004c | .020 | .984 | .004 | .163 |
| SMEAN(qsec) | -.046c | -.471 | .641 | -.090 | .589 |
| a. Predictors in the Model: (Constant), disp | | | | | | |
| b. Predictors in the Model: (Constant), disp, SMEAN(wt) | | | | | | |
| c. Predictors in the Model: (Constant), disp, SMEAN(wt), SMEAN(carb) | | | | | | |
| d. Dependent Variable: mpg | | | | | | |

At each stage of regression analysis SPSS provides a summary of variables that have not yet entered into the model.

| Table 1.24 Variables Entered/Removedb | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), dispa | . | Enter |
| 2 | . | SMEAN(hp) | Backward (criterion: Probability of F-to-remove >= .200). |
| 3 | . | SMEAN(qsec) | Backward (criterion: Probability of F-to-remove >= .200). |
| 4 | . | gear | Backward (criterion: Probability of F-to-remove >= .200). |
| 5 | . | am | Backward (criterion: Probability of F-to-remove >= .200). |
| 6 | . | vs | Backward (criterion: Probability of F-to-remove >= .200). |
| 7 | . | drat | Backward (criterion: Probability of F-to-remove >= .200). |
| a. All requested variables entered. | | | |
| b. Dependent Variable: mpg | | | |

This method uses the same removal criteria, but instead of starting the model with only one variable, it begins the model with all predictors included. The computer then tests whether any of these predictors can be removed from the model without having a substantial effect on how well the model fits the observed data. The first one to be removed has the least impact on how well the model fits the data.

| Table 1.25-Model Summary | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .927a | .859 | .802 | 2.6845 | .859 | 14.918 | 9 | 22 | .000 |
| 2 | .927b | .859 | .810 | 2.6255 | .000 | .000 | 1 | 22 | .999 |
| 3 | .927c | .859 | .817 | 2.5749 | .000 | .085 | 1 | 23 | .773 |
| 4 | .926d | .858 | .824 | 2.5284 | .000 | .105 | 1 | 24 | .748 |
| 5 | .926e | .858 | .830 | 2.4819 | .000 | .052 | 1 | 25 | .822 |
| 6 | .926f | .857 | .836 | 2.4438 | .000 | .177 | 1 | 26 | .678 |
| 7 | .922g | .850 | .834 | 2.4541 | -.007 | 1.237 | 1 | 27 | .276 |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp | | | | | | | | | |
| b. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, SMEAN(qsec), disp | | | | | | | | | |
| c. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, disp | | | | | | | | | |
| d. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), disp | | | | | | | | | |
| e. Predictors: (Constant), SMEAN(carb), vs, drat, SMEAN(wt), disp | | | | | | | | | |
| f. Predictors: (Constant), SMEAN(carb), drat, SMEAN(wt), disp | | | | | | | | | |
| g. Predictors: (Constant), SMEAN(carb), SMEAN(wt), disp | | | | | | | | | |

For model 1, when all variables are used as predictors, the multiple correlation coefficient is 0.927.In the final model it has reduced to 0.922.It implies there is a strong positive relationship .The change statistics tell us the change in the F ratio resulting from each block of the hierarchy. So model 1 causes R2 to change from 0 to 0.859 and this change in amount of variance explained gives rise to an F ratio of 14.918 which is significant with a probability(p value=0.000) less than 0.05. But it is insignificant for all other models. In the final model there are 3 predictors with R2 value 0.850 and F ratio has considerably decreased and is no longer significant (p value =0.276) greater than 0.05.

| Table 1.26-ANOVAh | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 967.508 | 9 | 107.501 | 14.918 | .000a |
| Residual | 158.539 | 22 | 7.206 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 2 | Regression | 967.508 | 8 | 120.938 | 17.545 | .000b |
| Residual | 158.539 | 23 | 6.893 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 3 | Regression | 966.919 | 7 | 138.131 | 20.833 | .000c |
| Residual | 159.128 | 24 | 6.630 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 4 | Regression | 966.221 | 6 | 161.037 | 25.189 | .000d |
| Residual | 159.826 | 25 | 6.393 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 5 | Regression | 965.889 | 5 | 193.178 | 31.360 | .000e |
| Residual | 160.158 | 26 | 6.160 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 6 | Regression | 964.801 | 4 | 241.200 | 40.388 | .000f |
| Residual | 161.246 | 27 | 5.972 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 7 | Regression | 957.411 | 3 | 319.137 | 52.989 | .000g |
| Residual | 168.636 | 28 | 6.023 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), SMEAN(hp), gear, SMEAN(qsec), disp | | | | | | |
| b. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, SMEAN(qsec), disp | | | | | | |
| c. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, disp | | | | | | |
| d. Predictors: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), disp | | | | | | |
| e. Predictors: (Constant), SMEAN(carb), vs, drat, SMEAN(wt), disp | | | | | | |
| f. Predictors: (Constant), SMEAN(carb), drat, SMEAN(wt), disp | | | | | | |
| g. Predictors: (Constant), SMEAN(carb), SMEAN(wt), disp | | | | | | |
| h. Dependent Variable: mpg | | | | | | |

The F ratio represents the ratio of improvement in prediction as a result of fitting the model relative to the inaccuracy that still exists in the model. If the improvement due to fitting the regression model is much greater than inaccuracy in the model then value of F>1.For the initial model F ratio is 14.918 which is very unlikely to have happened by chance(p value <0.05).In the final model the value of F is even higher (52.989),which is also highly significant (p value <0.05). We infer that the initial model significantly improved our ability to predict the outcome variable but the new model (final model) because the F ratio is more significant.

| Table 1.27-Coefficientsa | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
| B | Std. Error | Beta | Lower Bound | Upper Bound |
| 1 | (Constant) | 26.253 | 22.499 |  | 1.167 | .256 | -20.408 | 72.914 |
| disp | -.022 | .011 | -.456 | -2.063 | .051 | -.044 | .000 |
| drat | 1.656 | 1.681 | .147 | .985 | .335 | -1.829 | 5.142 |
| vs | -1.287 | 2.357 | -.108 | -.546 | .591 | -6.175 | 3.601 |
| am | -.753 | 2.338 | -.062 | -.322 | .751 | -5.601 | 4.096 |
| gear | .607 | 1.553 | .074 | .391 | .700 | -2.614 | 3.829 |
| SMEAN(wt) | -2.820 | 1.316 | -.323 | -2.143 | .043 | -5.549 | -.091 |
| SMEAN(hp) | 2.143E-5 | .024 | .000 | .001 | .999 | -.049 | .049 |
| SMEAN(qsec) | .211 | .848 | .054 | .249 | .806 | -1.548 | 1.970 |
| SMEAN(carb) | -1.380 | .616 | -.300 | -2.239 | .036 | -2.657 | -.102 |
| 2 | (Constant) | 26.261 | 19.797 |  | 1.327 | .198 | -14.691 | 67.214 |
| disp | -.022 | .008 | -.456 | -2.710 | .012 | -.039 | -.005 |
| drat | 1.656 | 1.644 | .147 | 1.008 | .324 | -1.744 | 5.057 |
| vs | -1.286 | 2.209 | -.108 | -.582 | .566 | -5.855 | 3.283 |
| am | -.752 | 2.280 | -.062 | -.330 | .744 | -5.469 | 3.964 |
| gear | .607 | 1.435 | .074 | .423 | .676 | -2.362 | 3.576 |
| SMEAN(wt) | -2.820 | 1.287 | -.323 | -2.192 | .039 | -5.483 | -.158 |
| SMEAN(qsec) | .210 | .720 | .053 | .292 | .773 | -1.280 | 1.701 |
| SMEAN(carb) | -1.379 | .513 | -.299 | -2.689 | .013 | -2.440 | -.318 |
| 3 | (Constant) | 31.549 | 7.868 |  | 4.010 | .001 | 15.311 | 47.787 |
| disp | -.023 | .008 | -.472 | -3.024 | .006 | -.039 | -.007 |
| drat | 1.554 | 1.575 | .138 | .987 | .334 | -1.697 | 4.805 |
| vs | -.911 | 1.761 | -.076 | -.517 | .610 | -4.546 | 2.725 |
| am | -.830 | 2.221 | -.069 | -.374 | .712 | -5.414 | 3.753 |
| gear | .393 | 1.211 | .048 | .325 | .748 | -2.106 | 2.893 |
| SMEAN(wt) | -2.933 | 1.204 | -.336 | -2.437 | .023 | -5.418 | -.449 |
| SMEAN(carb) | -1.388 | .502 | -.301 | -2.765 | .011 | -2.424 | -.352 |
| 4 | (Constant) | 32.066 | 7.565 |  | 4.239 | .000 | 16.486 | 47.647 |
| disp | -.023 | .007 | -.476 | -3.120 | .005 | -.038 | -.008 |
| drat | 1.651 | 1.519 | .146 | 1.087 | .287 | -1.477 | 4.779 |
| vs | -.798 | 1.696 | -.067 | -.471 | .642 | -4.291 | 2.695 |
| am | -.398 | 1.744 | -.033 | -.228 | .822 | -3.989 | 3.194 |
| SMEAN(wt) | -2.850 | 1.155 | -.326 | -2.468 | .021 | -5.228 | -.472 |
| SMEAN(carb) | -1.328 | .459 | -.288 | -2.897 | .008 | -2.273 | -.384 |
| 5 | (Constant) | 31.694 | 7.250 |  | 4.371 | .000 | 16.790 | 46.597 |
| disp | -.023 | .007 | -.463 | -3.330 | .003 | -.036 | -.009 |
| drat | 1.526 | 1.391 | .135 | 1.097 | .283 | -1.334 | 4.386 |
| vs | -.595 | 1.416 | -.050 | -.420 | .678 | -3.506 | 2.316 |
| SMEAN(wt) | -2.713 | .967 | -.310 | -2.805 | .009 | -4.700 | -.725 |
| SMEAN(carb) | -1.323 | .450 | -.287 | -2.943 | .007 | -2.247 | -.399 |
| 6 | (Constant) | 30.899 | 6.892 |  | 4.483 | .000 | 16.758 | 45.041 |
| disp | -.021 | .006 | -.436 | -3.602 | .001 | -.033 | -.009 |
| drat | 1.524 | 1.370 | .135 | 1.112 | .276 | -1.287 | 4.335 |
| SMEAN(wt) | -2.706 | .952 | -.310 | -2.843 | .008 | -4.660 | -.753 |
| SMEAN(carb) | -1.242 | .399 | -.270 | -3.109 | .004 | -2.061 | -.422 |
| 7 | (Constant) | 38.207 | 2.094 |  | 18.250 | .000 | 33.919 | 42.496 |
| disp | -.025 | .005 | -.511 | -5.067 | .000 | -.035 | -.015 |
| SMEAN(wt) | -3.182 | .854 | -.364 | -3.726 | .001 | -4.932 | -1.432 |
| SMEAN(carb) | -1.075 | .372 | -.233 | -2.892 | .007 | -1.836 | -.313 |
| a. Dependent Variable: mpg | | | | | | | | |

The final model obtained after doing backward regression has 3 predictors (disp,wt,carb) respectively.

The standardised beta coefficient column shows the contribution that an individual variable makes to the model. The beta weight is the average amount the dependent variable increases when the independent variable increases by one standard deviation (all other independent variables are held constant). The unstandardized coefficients B column ,gives us the coefficients of the independent variables in the regression equation. Each of these β values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the β values significantly differ from 0.

The final regression model is:

y=38.207-0.025b1-3.182b2-1.075b3

| ***Table 1.28-Excluded Variablesg*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics |
| Tolerance |
| 2 | SMEAN(hp) | .000a | .001 | .999 | .000 | .117 |
| 3 | SMEAN(hp) | -.029b | -.144 | .887 | -.030 | .155 |
| SMEAN(qsec) | .053b | .292 | .773 | .061 | .183 |
| 4 | SMEAN(hp) | -.032c | -.165 | .871 | -.034 | .155 |
| SMEAN(qsec) | .014c | .091 | .929 | .018 | .247 |
| gear | .048c | .325 | .748 | .066 | .268 |
| 5 | SMEAN(hp) | -.037d | -.192 | .849 | -.038 | .157 |
| SMEAN(qsec) | .025d | .186 | .854 | .037 | .311 |
| gear | .015d | .127 | .900 | .025 | .419 |
| am | -.033d | -.228 | .822 | -.046 | .272 |
| 6 | SMEAN(hp) | -.026e | -.142 | .888 | -.028 | .159 |
| SMEAN(qsec) | -.013e | -.123 | .903 | -.024 | .526 |
| gear | .023e | .203 | .841 | .040 | .433 |
| am | .003e | .023 | .982 | .005 | .376 |
| vs | -.050e | -.420 | .678 | -.082 | .390 |
| 7 | SMEAN(hp) | .004f | .020 | .984 | .004 | .163 |
| SMEAN(qsec) | -.046f | -.471 | .641 | -.090 | .589 |
| gear | .061f | .586 | .563 | .112 | .498 |
| am | .041f | .359 | .722 | .069 | .416 |
| vs | -.049f | -.414 | .682 | -.079 | .390 |
| drat | .135f | 1.112 | .276 | .209 | .359 |
| a. Predictors in the Model: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, SMEAN(qsec), disp | | | | | | |
| b. Predictors in the Model: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), gear, disp | | | | | | |
| c. Predictors in the Model: (Constant), SMEAN(carb), am, vs, drat, SMEAN(wt), disp | | | | | | |
| d. Predictors in the Model: (Constant), SMEAN(carb), vs, drat, SMEAN(wt), disp | | | | | | |
| e. Predictors in the Model: (Constant), SMEAN(carb), drat, SMEAN(wt), disp | | | | | | |
| f. Predictors in the Model: (Constant), SMEAN(carb), SMEAN(wt), disp | | | | | | |
| g. Dependent Variable: mpg | | | | | | |

The summary gives an estimate of each predictors β value, if it was removed from the equation at this point and calculates the t test for this value. In backward regression, SPSS should remove the predictor with the highest t statistic and will continue removing predictors until there are none left with t statistics that have significance values less than 0.05.The partial correlation also provides some indication as to what contribution (if any) an included predictor would make if it were removed from the model .This table gives information about the variables not in the regression equation at any point in time. We note that none of these variables were significant in the final model.

| Table 1.29-Variables Entered/Removeda | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | disp | . | Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200). |
| 2 | SMEAN(wt) | . | Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200). |
| 3 | SMEAN(carb) | . | Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200). |
| a. Dependent Variable: mpg | | | |

This table tracks variables which have been entered and removed at each step. We can see that disp, wt and carb have been entered into the model in 3 steps and no variable is removed.

| ***Table 1.30 -Model Summary*** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .848a | .718 | .709 | 3.2515 | .718 | 76.513 | 1 | 30 | .000 |
| 2 | .898b | .806 | .792 | 2.7481 | .087 | 12.998 | 1 | 29 | .001 |
| 3 | .922c | .850 | .834 | 2.4541 | .045 | 8.363 | 1 | 28 | .007 |
| a. Predictors: (Constant), disp | | | | | | | | | |
| b. Predictors: (Constant), disp, SMEAN(wt) | | | | | | | | | |
| c. Predictors: (Constant), disp, SMEAN(wt), SMEAN(carb) | | | | | | | | | |

First model contains only 1 predictor(disp) in the regression model. The multiple R tells us how strong the linear relationship is. As it is 0.84 it is moderately positive relationship ,for the 3 rd model its 0.922 which implies that it's an almost perfect positive relationship. The adjusted R2 changes from 70.9% to 79.2% ,meaning that addition of wt affects R2. After addition of some variables we end up with 3 predictors in our model. The change statistics tell us change in the F-ratio resulting from each block of the hierarchy. So model 1 causes R2 to change from 0 to 0.718 and this change in the amount of variance explained gives rise to an F ratio of 76.513 which is significant with a probability less than 0.05. The change statistics therefore tell us about the difference made by adding new predictors to the model. The p value increases to 0.007<0.05 which implies that overall regression is significant in third model. Finally model 3 causes the R2 to change to 0.85 that is 85% of the variation in the model is explained by the model.

| ***Table 1.31-ANOVAd*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 808.888 | 1 | 808.888 | 76.513 | .000a |
| Residual | 317.159 | 30 | 10.572 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 2 | Regression | 907.044 | 2 | 453.522 | 60.054 | .000b |
| Residual | 219.004 | 29 | 7.552 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| 3 | Regression | 957.411 | 3 | 319.137 | 52.989 | .000c |
| Residual | 168.636 | 28 | 6.023 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), disp | | | | | | |
| b. Predictors: (Constant), disp, SMEAN(wt) | | | | | | |
| c. Predictors: (Constant), disp, SMEAN(wt), SMEAN(carb) | | | | | | |
| d. Dependent Variable: mpg | | | | | | |

If the improvement due to fitting the regression model is much greater than the inaccuracy within the model then the value of F will be greater than 1.For the initial model the F ratio is 76.513 which is very unlikely to have happened by chance (p<0.05).For the second model value of F is lower (60.054) which is also highly significant. Finally we notice that the 3 rd model is having a very high F value of 52.989 (though less than 76.513)which is very significant at 5% level of significance.

| ***Table 1.32 -Coefficientsa*** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
| B | Std. Error | Beta | Lower Bound | Upper Bound |
| 1 | (Constant) | 29.600 | 1.230 |  | 24.070 | .000 | 27.088 | 32.111 |
| disp | -.041 | .005 | -.848 | -8.747 | .000 | -.051 | -.032 |
| 2 | (Constant) | 36.979 | 2.296 |  | 16.109 | .000 | 32.284 | 41.674 |
| disp | -.029 | .005 | -.589 | -5.405 | .000 | -.039 | -.018 |
| SMEAN(wt) | -3.431 | .952 | -.393 | -3.605 | .001 | -5.377 | -1.484 |
| 3 | (Constant) | 38.207 | 2.094 |  | 18.250 | .000 | 33.919 | 42.496 |
| disp | -.025 | .005 | -.511 | -5.067 | .000 | -.035 | -.015 |
| SMEAN(wt) | -3.182 | .854 | -.364 | -3.726 | .001 | -4.932 | -1.432 |
| SMEAN(carb) | -1.075 | .372 | -.233 | -2.892 | .007 | -1.836 | -.313 |
| a. Dependent Variable: mpg | | | | | | | | |

This table provides the details of the results. Both the raw and standardized regression coefficients are readjusted at each step to reflect the additional variables in the model. All p values are significant(in 3 models) hence we may accept our null hypothesis and conclude that individual regressors are significant. Each of these β values has an associated standard error indicating to what extent these values would vary across different samples and these standard errors are used to determine whether or not the β values significantly differ from 0.For example:β =-0.848 in the first model indicates that1 unit standard deviation change in disp is expected to result in a -0.848 standard deviation change in miles per gallon i.e. mpg decreases by 0.848 standard deviations.

The final regression model is:

y=38.207-0.025b1-3.182b2-1.075b3

| ***Table 1.33-Excluded Variablesd*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics |
| Tolerance |
| 1 | Drat | .160a | 1.169 | .252 | .212 | .496 |
| Vs | .125a | .905 | .373 | .166 | .495 |
| Am | .152a | 1.277 | .212 | .231 | .650 |
| Gear | .014a | .115 | .909 | .021 | .691 |
| SMEAN(wt) | -.393a | -3.605 | .001 | -.556 | .565 |
| SMEAN(hp) | -.310a | -1.716 | .097 | -.304 | .271 |
| SMEAN(qsec) | .095a | .885 | .384 | .162 | .818 |
| SMEAN(carb) | -.264a | -2.732 | .011 | -.452 | .830 |
| 2 | Drat | -.007b | -.054 | .957 | -.010 | .418 |
| Vs | .099b | .843 | .406 | .157 | .493 |
| Am | -.062b | -.511 | .613 | -.096 | .464 |
| Gear | -.079b | -.774 | .445 | -.145 | .649 |
| SMEAN(hp) | -.253b | -1.647 | .111 | -.297 | .268 |
| SMEAN(qsec) | .091b | 1.006 | .323 | .187 | .818 |
| SMEAN(carb) | -.233b | -2.892 | .007 | -.480 | .821 |
| 3 | Drat | .135c | 1.112 | .276 | .209 | .359 |
| Vs | -.049c | -.414 | .682 | -.079 | .390 |
| Am | .041c | .359 | .722 | .069 | .416 |
| Gear | .061c | .586 | .563 | .112 | .498 |
| SMEAN(hp) | .004c | .020 | .984 | .004 | .163 |
| SMEAN(qsec) | -.046c | -.471 | .641 | -.090 | .589 |
| a. Predictors in the Model: (Constant), disp | | | | | | |
| b. Predictors in the Model: (Constant), disp, SMEAN(wt) | | | | | | |
| c. Predictors in the Model: (Constant), disp, SMEAN(wt), SMEAN(carb) | | | | | | |
| d. Dependent Variable: mpg | | | | | | |

In stepwise regression this table contains summaries of the variables that SPSS is considering entering into the model. The summary gives an estimate of each predictors ß value if it was entered into the equation at this point and calculates a t test for this value. In stepwise regression, SPSS should enter the predictor with the highest t statistic and will continue entering predictors until there are none left with t statistics that have significance values less than 0.05.The partial correlation also provides some indication as to what contribution(if any) an excluded predictor would make if it were entered into the model.

| ***Table 1.34-Descriptive Statistics*** | | | |
| --- | --- | --- | --- |
|  | Mean | Std. Deviation | N |
| mpg | 20.091 | 6.0269 | 32 |
| disp | 230.722 | 123.9387 | 32 |
| SMEAN(wt) | 2.99769 | .689825 | 32 |
| SMEAN(carb) | 2.645 | 1.3087 | 32 |

This table gives the mean and standard deviation of the dependent variable (mpg) and 3 independent variables. We notice that displacement has the highest mean and standard deviation.

| ***Table 1.35-Correlations*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | mpg | disp | SMEAN(wt) | SMEAN(carb) |
| Pearson Correlation | mpg | 1.000 | -.848 | -.781 | -.568 |
| disp | -.848 | 1.000 | .659 | .413 |
| SMEAN(wt) | -.781 | .659 | 1.000 | .341 |
| SMEAN(carb) | -.568 | .413 | .341 | 1.000 |
| Sig. (1-tailed) | mpg | . | .000 | .000 | .000 |
| disp | .000 | . | .000 | .009 |
| SMEAN(wt) | .000 | .000 | . | .028 |
| SMEAN(carb) | .000 | .009 | .028 | . |
| N | mpg | 32 | 32 | 32 | 32 |
| disp | 32 | 32 | 32 | 32 |
| SMEAN(wt) | 32 | 32 | 32 | 32 |
| SMEAN(carb) | 32 | 32 | 32 | 32 |

We notice that disp and wt are highly correlated with mpg.

Stepwise method relies on the computer selecting variablesbased upon mathematical criteria.The models derived by computer often take advantage of random sampling variation and so decisions about which variables should be included will be based upon slight differences in their semi partial correlationStepwise regression is a combination of the forward and backward selection procedure. It is a modification of the forward selection so that after each step in which a variable is added all candidate variables are checked to see if their significance has been reduced below the specified tolerance level. This method is preferred because it's a combination of both forward selection and backward elimination. It checks the statistical significance of each variable in the best possible way.Stepwise regression is generally used in exploratory model building.

| ***Table 1.36-Variables Entered/Removed*** | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | SMEAN(carb), SMEAN(wt), dispa | . | Enter |
| a. All requested variables entered. | | | |

This table shows the number of variables entered for regression. The default method for multiple regression is the **Enter** method. This is also known as **direct** regression or **simultaneous** regression.

| ***Table 1.37-Model Summaryb*** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | | Durbin-Watson |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .922a | .850 | .834 | 2.4541 | .850 | 52.989 | 3 | 28 | .000 | 1.972 |
| a. Predictors: (Constant), SMEAN(carb), SMEAN(wt), disp | | | | | | | | | | |
| b. Dependent Variable: mpg | | | | | | | | | | |

As R value is 0.922 it implies the variables have a strong positive relationship. Adjusted R2 indicates

that we have a fairly good model as 83.4% variance can be predicted from the independent variables. The F statistic is high ,its 52.989 ( p value <0.05) which implies overall regression is significant. Our Durbin Watson statistic lies between 1 and 3 which shows autocorrelation is absent in the data. Our error terms are not correlated.

| ***Table 1.38-ANOVAb*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 957.411 | 3 | 319.137 | 52.989 | .000a |
| Residual | 168.636 | 28 | 6.023 |  |  |
| Total | 1126.047 | 31 |  |  |  |
| a. Predictors: (Constant), SMEAN(carb), SMEAN(wt), disp | | | | | | |
| b. Dependent Variable: mpg | | | | | | |

F value is also high which implies that the good model fit cannot be attributed to chance only. As p value<0.05 we may reject our null hypothesis and conclude that combination of these variables significantly predicts the dependent variable(mpg).Hence overall model is significant.

| ***Table 1.39-Coefficientsa*** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | | Collinearity Statistics | |
| B | Std. Error | Beta | Lower Bound | Upper Bound | Tolerance | VIF |
| 1 | (Constant) | 38.207 | 2.094 |  | 18.250 | .000 | 33.919 | 42.496 |  |  |
| disp | -.025 | .005 | -.511 | -5.067 | .000 | -.035 | -.015 | .525 | 1.903 |
| SMEAN(wt) | -3.182 | .854 | -.364 | -3.726 | .001 | -4.932 | -1.432 | .560 | 1.787 |
| SMEAN(carb) | -1.075 | .372 | -.233 | -2.892 | .007 | -1.836 | -.313 | .821 | 1.218 |
| a. Dependent Variable: mpg | | | | | | | | | | |

The standardised beta coefficient column shows the contribution that an individual variable makes to the model. The beta weight is the average amount the dependent variable increases when the independent variable increases by one standard deviation (all other independent variables are held constant). The unstandardized coefficients B column ,gives us the coefficients of the independent variables in the regression equation. As all p values are significant we may reject our null hypothesis and conclude that individual regressors are significant. All predictors are contributing significantly to the model.

The regression model is;

y=38.207-0.025b1-3.182b2-1.075b3

| ***Table 1.40-Collinearity Diagnosticsa*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Dimension | Eigen value | Condition Index | Variance Proportions | | | |
| (Constant) | disp | SMEAN(wt) | SMEAN(carb) |
| 1 | 1 | 3.740 | 1.000 | .00 | .01 | .00 | .01 |
| 2 | .124 | 5.481 | .03 | .60 | .00 | .32 |
| 3 | .119 | 5.612 | .11 | .05 | .03 | .67 |
| 4 | .017 | 14.758 | .85 | .34 | .97 | .00 |
| a. Dependent Variable: mpg | | | | | | | |

It measures how much regressors are related to other regressors and how this affects the stability and variance of the regression estimates. It gives the amount of variance proportions of each variable in the model.

| ***Table 1.41-Residuals Statisticsa*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Minimum | Maximum | Mean | Std. Deviation | N |
| Predicted Value | 12.637 | 29.526 | 20.091 | 5.5574 | 32 |
| Std. Predicted Value | -1.341 | 1.698 | .000 | 1.000 | 32 |
| Standard Error of Predicted Value | .585 | 1.478 | .845 | .199 | 32 |
| Adjusted Predicted Value | 12.957 | 29.262 | 20.078 | 5.5171 | 32 |
| Residual | -4.8034 | 5.3199 | .0000 | 2.3324 | 32 |
| Std. Residual | -1.957 | 2.168 | .000 | .950 | 32 |
| Stud. Residual | -2.050 | 2.344 | .003 | 1.015 | 32 |
| Deleted Residual | -5.2689 | 6.2206 | .0130 | 2.6642 | 32 |
| Stud. Deleted Residual | -2.183 | 2.568 | .007 | 1.056 | 32 |
| Mahal. Distance | .791 | 10.281 | 2.906 | 2.012 | 32 |
| Cook's Distance | .000 | .233 | .036 | .054 | 32 |
| Centered Leverage Value | .026 | .332 | .094 | .065 | 32 |
| a. Dependent Variable: mpg | | | | | |

It gives us information about residuals ,its mean and standard deviation. It also gives standardised ,studentised residuals and cooks distance.

Fig 1.1



It shows the data is approximately normally distributed ( a bell shaped curve).We have drawn a curve on the histogram to show the shape of the distribution..For the given data,the distribution is approximately normal (although there is a deviation in the residuals around zero)

Fig 1.2



The normal probability plot also shows up deviations from normality. The straight line in this plot represents a normal distribution, and the points represent the observed residuals. Therefore ,in a perfectly normally distributed data sets all points will lie on the line .This is pretty much what we see for the mtcars data.

| ***Table 1.42-Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| Standardized Residual | .101 | 32 | .200\* | .978 | 32 | .730 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

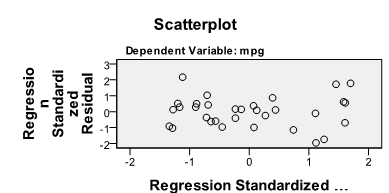
We set the following hypothesis:

Ho-The data fits the normal distribution.

H1-The data does not fit the normal distribution.

So from the table above we infer that the p value is 0.200(which is >0.05).Hence we may accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution

Fig 1.3



The graph looks like a random array of dots evenly dispersed around 0.As there is no sort of curve in this graph or any pattern we may conclude that data has not broken the assumption of normality. Note that the points are randomly and evenly dispersed throughout the plot. This pattern is indicative of the situation in which the assumptions of linearity and homoscedasticity have been met.

A final set of plots specified below are the partial plots. These plots are scatter plots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors.

Fig1.4



The partial plot shows a strong negative relationship between disp and mpg. The gradient of the line is β for disp variable in the model. There are no obvious outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe that it is linear with R2linear being around 0.478. As the variable shows near linear relationship we include it in our model.

Fig 1.5



The partial plot shows a strong negative relationship between wt and mpg. The gradient of the line is β for wt variable in the model. There are no outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe that it appears to be linear with R2linear being around 0.331.As the variable shows near linear relationship we include it in our model.

Fig 1.6



The partial plot shows a strong negative relationship between carb and mpg. The gradient of the line is β for carb variable in the model. There are no outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe that it appears to be linear with R2linear being around 0.23.As the variable shows near linear relationship we include it in our model.

Fig 1.7



As the dots are scattered ,it indicates data meet the assumption of errors being normally distributed and the variances of the residuals being constant. If the dots created a pattern this would indicate the residuals are not normally distributed, the residual is correlated with the independent variable and variances of the residual are not constant.

Fig 1.8



We notice that there is no observation which is very far from other observations. Hence we conclude there are no outliers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.43 Case Summariesa*** | | | | | | |
|  |  | Studentized Residual | Cook's Distance | Centered Leverage Value | Unstandardized Predicted Value | Unstandardized Residual |
| 1 | | -.25657 | .00201 | .07769 | 21.59436 | -.59436 |
| 2 | | .09349 | .00026 | .07377 | 20.78295 | .21705 |
| 3 | | -1.82681 | .08731 | .06349 | 27.06556 | -4.26556 |
| 4 | | .39300 | .00468 | .07677 | 20.48911 | .91089 |
| 5 | | 1.08939 | .03280 | .06830 | 16.16306 | 2.53694 |
| 6 | | -1.06033 | .04130 | .09686 | 20.52980 | -2.42980 |
| 7 | | .29753 | .00195 | .04980 | 13.60004 | .69996 |
| 8 | | .90873 | .01788 | .04845 | 22.26057 | 2.13943 |
| 9 | | .11274 | .00027 | .04792 | 22.53451 | .26549 |
| 10 | | .17783 | .00133 | .11264 | 18.79620 | .40380 |
| 11 | | -.43872 | .00809 | .11264 | 18.79620 | -.99620 |
| 12 | | .53571 | .01112 | .10294 | 15.17668 | 1.22332 |
| 13 | | .44101 | .00389 | .04282 | 16.25856 | 1.04144 |
| 14 | | -.38235 | .00323 | .04990 | 16.09946 | -.89946 |
| 15 | | -1.05500 | .09444 | .22215 | 12.63715 | -2.23715 |
| 16 | | -1.18001 | .10601 | .20220 | 12.93543 | -2.53543 |
| 17 | | .57840 | .02127 | .17150 | 13.43257 | 1.26743 |
| 18 | | 1.82090 | .09870 | .07515 | 28.17571 | 4.22429 |
| 19 | | .60675 | .01782 | .13098 | 29.03708 | 1.36292 |
| 20 | | 1.92183 | .15025 | .10870 | 29.52606 | 4.37394 |
| 21 | | -2.04993 | .10182 | .05711 | 26.30340 | -4.80340 |
| 22 | | -.61655 | .00810 | .04725 | 16.95249 | -1.45249 |
| 23 | | -1.00128 | .01858 | .03776 | 17.57096 | -2.37096 |
| 24 | | .13289 | .00045 | .06212 | 12.98947 | .31053 |
| 25 | | 2.34407 | .23255 | .11353 | 13.88006 | 5.31994 |
| 26 | | -.74662 | .02037 | .09628 | 29.01149 | -1.71149 |
| 27 | | -.10965 | .00027 | .05024 | 26.25790 | -.25790 |
| 28 | | .68734 | .02724 | .15613 | 28.87942 | 1.52058 |
| 29 | | .30025 | .00219 | .05734 | 15.09656 | .70344 |
| 30 | | .18350 | .00479 | .33165 | 19.34056 | .35944 |
| 31 | | -.63883 | .00614 | .02552 | 16.52262 | -1.52262 |
| 32 | | -1.17951 | .02285 | .03039 | 24.20402 | -2.80402 |
| Total | N | 32 | 32 | 32 | 32 | 32 |
| a. Limited to first 100 cases. | | | | | | |

1) SPSS produces a summary table for residual statistics and these should be examined for extreme cases.

We have computed leverage values in third column. We check if an observation has leverage more than 2p/n, where n is the number of observations and p is number of variables then it's a leverage point.

We observe that no observation has leverage more than (2p/n)=(8/32)=0.25 except 30 th observation,

hence it's a leverage point. It is not influential as it does not have a significantly high value of studentized residual. However cases with large leverage values will not necessarily have large influence on the regression coefficients because they are measured on the outcome variables rather than the predictors.

2) Now we will check the column of studentised residuals. As absolute value of no observation >3 hence we conclude that none of the observations are influential observations. ,that is none of them can be taken as outliers. But we have 2 observations corresponding to which the absolute value of studentized residuals lies between 2 and 3 i.e. (case no. 21,25).For such observations we consider the leverage values. As the leverage <0.25(=2p/n).So these observations can't be taken as outliers.

| ***Table 1.44-Correlations*** | | | | |
| --- | --- | --- | --- | --- |
|  |  |  | abres | Unstandardized Predicted Value |
| Spearman's rho | Abres | Correlation Coefficient | 1.000 | .172 |
| Sig. (2-tailed) | . | **.348** |
| N | 32 | 32 |
| Unstandardized Predicted Value | Correlation Coefficient | .172 | 1.000 |
| Sig. (2-tailed) | **.348** | . |
| N | 32 | 32 |

We set the hypothesis as follows:

Ho: There is no significant relationship between absolute residuals and predicted values.

H1:There is significant relationship between absolute residuals and predicted values.

As p value (0.348)>0.05 we may accept our null hypothesis and conclude that correlation is insignificant.

As Spearman’s rank correlation between predicted response and absolute residual is insignificant, then we conclude there is homoscedasticity.

**References**

* Discovering Statistics Using SPSS for Windows by Andy Field
* Fundamentals of Mathematical Statistics by S.C .Gupta and V.K. Kapoor
* Statistical computing by Debasis Kundu and Ayanendranath Basu
* SPSS for intermediate statistics by Nancy L. Leech, Karen C. Barrett, George A. Morgan
* Additional notes given by Abhishek K. Umrawal Sir
* Google has been used generously to verify facts.