***Case Study-Linear Regression Analysis***

***Submitted by: T. Sujaya***

**LINEAR REGRESSION ANALYSIS**

Regression Analysis deals with studying relationship from a set of variables to another variable, where the former is called the set of independent variables or regressors or predictors and the latter is called the dependent variable or regressed variable or response. In fact we are trying to quantify 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.

**DATA SOURCE:**  We have taken 'mtcars' dataset from R (We have exported it to SPSS using function write.csv (mtcars, ”mtcars.sav”).The data source is Henderson and Velleman (1981), building multiple regression models interactively. Biometrics, 37,391 to 441. The data was extracted from the 1974 Motor Trend US Magazine and comprises fuel consumption and ten aspects of automobile design and performance for 32 automobiles(1973-74 models). It's a data frame with 32 observations on 11 variables.

Table 1: Description of Variables in Our Dataset

|  |  |  |
| --- | --- | --- |
| Serial No. | Variable Name | Type |
| 1 | Miles /(US) gallon (mpg) | Continuous |
| 2 | Number of cylinders (cyl) | Discrete |
| 3 | Displacement (disp) | Continuous |
| 4 | Gross horsepower (hp) | Continuous |
| 5 | Rear axle ratio (drat) | Continuous |
| 6 | Weight (wt) | Continuous |
| 7 | (1/4)th mile time(qsec) | Continuous |
| 8 | V/S (vs) | Discrete |
| 9 | Transmission (0=automatic,1=manual)(am) | Nominal |
| 10 | Number of forward gears(gear) | Discrete |
| 11 | Number of carburetors(carb) | Discrete |

**INTRODUCTION**

Linear Regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted as X. The case of one explanatory variable is called simple linear regression. It is the most basic and commonly used predictive analysis.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.

With any data set there are a number of lines that could be used to summarise the general trend and so we need a way to decide which of many possible lines to choose .For the sake of doing accurate conclusions we want to fit a model that best describes a data. There are several ways to fit a straight line to the data we have collected. We use a mathematical technique to establish the line that best describes the data collected. This method is called the method of least squares (OLS).The line of best fit results in the least amount of difference between the observed data points and the line

A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

**Multiple linear regression** attempts to model the relationship between two or more explanatory variables and a response variable by fitting a **linear** equation to observed data. Every value of the independent variable X is associated with a value of the dependent variable y. The following is a multiple linear regression model:

**y=bo+b1X1+b2X2+........bpXp +ei (1)**

where bo=intercept or regression constant

b1=partial regression coefficient of the first predictor(X1)

ei=difference between predicted and observed value of y for the ith subject

**Assumptions**

When we choose to analyze our data using multiple regression part of the process involves checking to make sure that the data we want to analyze can actually be analyzed using multiple regression. If our data satisfies the given assumptions then multiple regression will give a valid result. The assumptions are:

1)The regression model is linear in terms of parameters ,it may or may not be linear in variables y and the Xs. For example the regression model in (1) is linear.

2) The dependent variable should be measured on a continuous scale(i.e. it is either an interval or ratio variable)

3) There should be independence of observations, i.e. there is no correlation between 2 error terms. There is no autocorrelation.cov(eiej|X)=0 **( i ≠j)**

If there is **autocorrelation** present in the data the following **consequences** follow:

a) The OLS (Ordinary Least Squares) estimators are still unbiased and consistent.

b) They are still normally distributed in large samples.

c) But they are no longer efficient. They are no longer BLUE. In most cases OLS standard errors are underestimated which means the estimated t values are inflated giving the appearance that a coefficient is more significant than it actually may be.

d) As a result the hypothesis testing procedure becomes suspect, since the estimated standard errors may not be reliable, even asymptotically. In consequence the usual t and F test become invalid.

We use Durbin -Watson statistic. to test for autocorrelation. 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

4) The variance of each ei given the value of X is constant or homoscedastic (homo means equal and scedastic means variance).That is

var(ei|X)=**σ2**

There are various reasons for heteroscedasticity which are as follows:

a) Presence of outliers in the data

b) Incorrect functional form of regression model.

c)Incorrect transformation of data

d) Mixing observations with different measures of scale.

**Consequences**

a) Heteroscedasticity does not alter the unbiasedness and consistency properties of OLS estimators.

b) But OLS estimators are no longer having minimum variance or are efficient, i.e. they are no longer BLUE(Best Linear Unbiased Estimators), they are simply LUE(Linear Unbiased Estimators).

c) As a result the t and F tests may not be reliable resulting in erroneous conclusions regarding the statistical significance of the estimated regression coefficients.

**To check for heteroscedasticity**

a)We 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.

b)Test if Spearman’s rank correlation between predicted response and absolute residual is significant, then we conclude there is heteroscedasticity, otherwise homoscedasticity.

5)There is no perfect linear relationships among the X variables. This is the assumption of no **multicollinearity**. If there is just one perfect linear relationship between two or more regressors ,it is called collinearity but if there is more than one perfect linear relationship ,it is called multicollinearity. It can also be defined as a high degree of correlation (linear dependence) among several independent variables. It commonly occurs when a large number of variables are incorporated in a regression model. Perfect multicollinearity is a rarity, it violates the assumption that X matrix is full ranked ,making OLS impossible.

Symptoms of multicollinearity may be observed in situations where:

a)small changes in the data produce wide swings in parameter estimates.

b)coefficients may have very high standard errors and low significance levels even though they are jointly sufficient and the R2 for regression is very high

c) Coefficients may have the wrong sign or implausible magnitude.

**Consequences**

a)OLS estimators are still BLUE ,but they have large variances and covariances making precise estimation difficult.

b) As a result the confidence intervals tend to be wider .Therefore we may not reject the "zero null hypothesis"(i.e. the true population coefficient is zero)

c) Because of a) the t ratios of one or more coefficients tend to be statistically insignificant.

d) Even though some regression coefficients may be statistically insignificant, the R2 value may be very high.

e)The OLS estimators and their standard errors can be sensitive to small changes in the data.

f) Adding a collinear variable to the chosen regression model can alter the coefficient values of the other variables in the model.

This should not be surprising, because if two variables are highly collinear it is very difficult to isolate the impact of each variable separately on the regressand.

Detection and Removal of Multicollinearity can be done using Correlation Analysis and Variance Inflation Factors (VIFs).The **Variance Inflation Factor** (VIF) is an indicator of the effect that the other independent variables have on the standard error of a regression coefficient. A large VIF indicates high multicollinearity. We take threshold for significant (absolute) correlation to be 0*.*75 and that for VIF to be 10. VIF shows how multicollinearity has increased the instability of the coefficient estimates. Tolerance is defined as reciprocal of VIF. Tolerance is also defined as 1-r 2 where r2 is coefficient of determination. Tolerance less than 0.1 indicates significant multicollinearity. Multicollinearity can be detected by using the correlation analysis that is we compute a spearman rank correlation matrix for all the variables. Then we eliminate variables which are highly correlated. If a variable is common between two pairs it is eliminated first. This is not a reliable or proper method as we are dropping the variables only on the basis of its correlation coefficient. It may lead to specification bias.

6) There needs to be a **linear relationship** between

a) The dependent variable and each of your independent variables and

b) The dependent variable and the independent variables collectively.

We create scatter plots and partial regression plots in SPSS and then visually inspect them to check for linearity. If the relationship displayed in your scatter plots and partial regression plots are not linear, we will have to either run a non-linear regression analysis or "transform" your data, which we can do using SPSS Statistics.

7) There should be **no significant outliers**, **high leverage points** or **highly influential points**. Outliers, leverage and influential points are different terms used to represent observations in your data set that are in some way unusual when we wish to perform a multiple regression analysis. These different classifications of **unusual points** reflect the different impact they have on the regression line. An observation can be classified as more than one type of unusual point. However, all these points can have a very negative effect on the regression equation that is used to predict the value of the dependent variable based on the independent variables. This can change the output that SPSS Statistics produces and reduce the predictive accuracy of your results as well as the statistical significance. Fortunately, when using SPSS Statistics to run multiple regression on the data, we can detect possible outliers, high leverage points and highly influential points.

8) Finally, we need to check that the **residuals (errors)** 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; or

(b) a Normal Q-Q Plot of the studentized residuals.

**R2 or measure of goodness of fit of the estimated regression**

R2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2 **coefficient of determination** is a statistical measure of how well the regression line approximates the real data points. An R2 of 1 indicates that the regression line perfectly fits the data.

The **coefficient of determination**, denoted ***R*2** or ***r*2** , is a number that indicates how well data fit a statistical model – sometimes simply a line or a curve. It is a **statistic** used in the context of statistical models whose main purpose is either the **prediction** of future outcomes or the testing of **hypothesis**, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.

**Adjusted R2**

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.  
The reason this is important is because R-squared can be changed by adding more and more independent variables, irrespective of how well they are correlated to the dependent variable. This isn't a desirable property of a goodness-of-fit statistic. Conversely, adjusted R-squared provides an adjustment to the R-squared statistic such that an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. That is the desired property of a goodness-of-fit statistic.  
In the case of a linear regression with more than one variable: adjusted R-squared. For a single independent variable model, both statistics are interchangeable.

**Parsimonious Modelling**

**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.

Variable selection is intended to select the best subset of predictors. It is done because:

1. We want to explain the data in the simplest way .Redundant predictors should be removed. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is the best.

2. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.

3. Collinearity is caused by having too many variables trying to do the same job.

4. Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

**Backward Elimination**

This is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.

1. Start with all the predictors in the model

2. Remove the predictor with highest p-value greater than a*crit*

3. Refit the model and go to 2

4. Stop when all p-values are less than a*crit* .

The a*crit* is sometimes called the .p-to-remove. and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

**Forward Selection**

The simplest data-driven model building approach is called *forward selection*. In this approach, one adds 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 it's P-value is below some pre-set level. It is customary to set this value above the conventional .05 level at say .10 or .15, because of the exploratory nature of this method.

Thus we begin with a model including the variable that is most significant in the initial analysis, and continue adding variables until none of remaining variables are "significant" when added to the model. Note that this multiple use of hypothesis testing means that the real type I error rate for a variable (i.e. the chance of including it in the model given it isn't really necessary), does not equal the critical level we choose. Continue until no new predictors can be added. In fact, because of the complexity that arises from the complex nature of the procedure, it is essentially impossible to control error rates and this procedure must be viewed as exploratory.

**Stepwise Regression**

The basic procedure is as follows:

1)First we select the Z most correlated with y (suppose it is Z1) and find the first order linear regression equation **Y^=f(Z1).**We check if this variable is significant. If it is not, we quit and adopt the model **Y=Y bar** as best, otherwise we search for the second predictor variable to enter regression. We examine the partial correlation coefficients of all the predictors not in regression at this stage namely Zj,**j≠1** with Y; i.e. Y and Zjare both adjusted for their straight line relationships with Z1,and the correlation between these adjusted variables is calculated for all **j ≠ 1.**Mathematically this is equivalent to finding the correlations between the correlations between (1) the residuals from the regression Y cap=f(Z1) and (2) the residuals from each of the j regressions Zj cap=fj(Z1).The Zj with the highest partial correlation coefficient with Y is now selected ,(suppose this is Z2) and a second regression equation is Y cap=f(Z1,Z2) is fitted.

2)The overall regression is checked for significance, the improvement in value of R2 is noted ,and the partial F values for both variables now in the equation (not just the one recently entered) are examined. The lower of these two partial Fs are then compared with an appropriate F percentage point and the corresponding predictor variable is retained in the equation or rejected according to whether the test is significant or not. This testing of the least useful predictor currently in the equation is carried out at every stage of the stepwise procedure . A predictor that may have been the best entry candidate at an earlier stage may at a later stage be superfluous because of the relationships between it and the other variables now in the regression.

The following three stepwise methods are available in SPSS. The method used by SPSS is explained below:

* **Stepwise** Based on the p-value of F (probability of F), SPSS starts by entering the variable with the smallest p-value; at the next step again the variable (from the list of variables not yet in the equation) with the smallest p-value for F and so on. Variables already in the equation are removed if their p-value becomes larger than the default limit due to the inclusion of another variable. The method terminates when no more variables are eligible for inclusion or removal. This methods is based on both probability-to-enter (PIN) and probability to remove (POUT) (or alternatively FIN and FOUT).
* **Backward Elimination**: First all variables are entered into the equation and then sequentially removed. For each step SPSS provides statistics, namely R2. At each step, the largest probability of F is removed (if the value is larger than POUT. Alternatively FOUT can be specified as a criterion.
* **Forward selection**: at each step the variable not yet in the equation with the smallest probability pf F is entered. as long as the value is smaller than PIN. Alternatively you can use the value of F by specifying FIN on /CRITERIA. The procedure stops when there are no variables that meet the entry criterion.

**Outliers**

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 it’s 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 it’s 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. We study some common methods:

**1. Leverage Value:** Leverage of an observation is calculated as the corresponding element of the hat matrix of regression:

***li* = (*X*T(*X*T*X*)−1*X*)*i,i***

If an observation has leverage more than 2*p/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***.

**2. Studentized Residual:** Studentized residual is a scaled residual which is calculated as follows:

***ri* = *ei/ √(MSRes*(1 − *li*))**

If *abs*(*ri*) *>* 3 then *i* th observation is an influential observation.

Other methods for outlier detection include ***Cook’s Distance*, *DFFITS* and *DFBETAS*.**

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

**Cook s Distance**

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 and

k is the number of independent variables.

To identify the influential cases with large Cook's distances, we sort the data set by the Cook's distance variable, 'coo\_1' that SPSS created in the data set.

**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 we pre process the data using Exploratory Data Analysis(EDA).But we build a regression model using the original data (i.e. before pre processing).The estimates,R2 value and diagnostics are obtained as follows:

We set the following hypothesis:(F test)

**Ho**: The dependent variable is not significantly explained by the independent variables.

**H1:**The dependent variable is significantly explained by the independent variables.

and (t-test)

Ho:β=0 Individual regressors are insignificant.

H1:β≠0 Individual regressors are significant

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.1 Variables Entered/Removed** | | | |
| Model | Variables Entered | Variables Removed | Method |
| 1 | carb, am, vs, drat, qsec, gear, disp, hp, wt, cyla | . | 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 1.2 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 | .932a | .869 | .807 | 2.6502 | .869 | 13.932 | 10 | 21 | .000 |
| a. Predictors: (Constant), carb, am, vs, drat, qsec, gear, disp, hp, wt, cyl | | | | | | | | | |

The Model Summary Table shows the multiple Correlation Coefficient (R) using all the predictors simultaneously, is 0.932. R2 is 0.869 and adjusted R2 is 0.807 meaning that 80.7% variance in mpg is explained by the 10 independent predictors. The standard error of estimate is a measure of accuracy of the prediction. As p value<0.05 the combination of these variables significantly predicts the dependent variable(mpg).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.3 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 | | | | | | |

ANOVA Table tests the significance of the regression model. We can see from the table that Sig(p value)=0.00.As p<0.05 our predictors are significantly better than would be expected by chance. The regression line predicted by the independent variables explains a significant amount of variance in the dependent variable.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.4 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 | | | | | | | | |

We observe that cyl, disp, wt are showing VIF s higher than 10 ,which indicates high multicollinearity. 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).As these are standardised we can compare them. We see that largest influence on mpg is from weight(wt) .The unstandardised coefficients B column ,gives us the coefficients of the independent variables in the regression equation.

**y=12.303 -.111b1+.013b2--.021b3+ .787b4 -3.715 b5+ .821b6+ .318b7+ 2.520b8+ .655b9 -.199b10**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.5 Collinearity Diagnosticsa** | | | | | | | | | | | | | | |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | | | | | | | | |
| (Constant) | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
| 1 | 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 |
| a. Dependent Variable: mpg | | | | | | | | | | | | | | |

This table tells how each variable is contributing to collinearity. Condition number for 8,9,10,11 variable is above 30 which indicates severe multicollinearity.

Then we build a linear regression model for the processed data. We are testing overall significance by using F test and individual significance of regressors using t test. We set the following hypothesis:

**Ho**: The dependent variable is not significantly explained by the independent variables.

**H1:**The dependent variable is significantly explained by the independent variables.

and

Ho:β=0 individual regressor is insignificant.

H1:β≠0 Individual regressors are significant.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.6 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 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 1.7 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 | | | | | | | | | |

The Model Summary Table shows the multiple Correlation Coefficient (R) using all the predictors simultaneously, is 0.929. R2 is 0.862 and adjusted R2 is 0.797 meaning that 79.7% variance in mpg is explained by the 10 independent predictors. The standard error of estimate is a measure of accuracy of the prediction. As p value<0.05 the combination of these variables significantly predicts the dependent variable(mpg).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.8 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 | | | | | | | | | | | | | | |
| ANOVA Table tests the significance of the regression model. We can see from the table that Sig(p value)=0.00.As p<0.05 our predictors are significantly better than would be expected by chance. The regression line predicted by the independent variables explains a significant amount of variance in the dependent variable. So overall regression is significant. | | | | | | | | | | | | | | |
| **Table 1.9 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 | | | | | | | | | | | | | | | |

We observe that only cyl, is showing VIF higher than 10 ,which indicates high multicollinearity. 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).As these are standardised we can compare them. We see that largest influence on mpg is from displacement(disp) .We get shaky results as beta coefficient for hp is 0 but its p value is significant .The unstandardized coefficients B column ,gives us the coefficients of the independent variables in the regression equation.

**y=20.693 +.756b1-.027b2+1.944b3- .927b4 -1.272b5+ 0.928b6- 3.445b7+ 0.0b8+ .322b9 -1.462b10**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.10 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 | | | | | | | | | | | | | | |

This table tells how each variable is contributing to collinearity. It gives amount of variance in proportions of each variable in the model.

Then we use correlation analysis approach for detection of multicollinearity. Here we notice that (cyl,vs),(cyl,wt),(cyl,disp),(cyl,hp),(disp,hp),(am,gear),(vs,qsec) are having high correlations (above 0.75)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.11 Correlations** | | | | | | | | | | | |
|  |  | 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). | | | | | | | | | | | |

After removing cyl, hp ,gear and vs we get the table below with correlations below threshold value of 0.75.

Table 1.12

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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). | | | | | | | | |

This method is not reliable as correlation value is not a strong basis for removing variables. Inspection of the correlation matrix for high pairwise correlations is not sufficient sine multicollinearity can exist with no pairwise correlations being high.

**Using VIF Approach**

We remove cyl after we notice that its VIF is very high.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.13 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. | | | |

We build a linear regression model without cyl. The default method is enter method. It's also called forced entry method.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.14 Model Summaryb** | | | | | | | | | |
| 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 | | | | | | | | | |

The Model Summary Table shows the multiple Correlation Coefficient (R) using all the predictors simultaneously, is 0.927. R2 is 0.859 and adjusted R2 is 0.802 meaning that 80.2% variance in mpg is explained by the 9 independent predictors. As p value<0.05 the combination of these variables significantly predicts the dependent variable(mpg).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.15 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 | | | | | | |

ANOVA Table tests the significance of the regression model. We can see from the table that Sig(p value)=0.00.As p<0.05 our predictors are significantly better than would be expected by chance. The regression line predicted by the independent variables explains a significant amount of variance in the dependent variable. So overall regression is significant

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.16 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 | | | | | | | | |

We observe that all the VIF s are under 10 which means that multicollinearity is under check and it does not pose a serious problem .The unstandardized coefficients B column ,gives us the coefficients of the independent variables in the regression equation. The β values tell us to what degree each predictor affects the outcome if the effects of all other predictors are held constant. 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.

**y=26.253 -.022b1+1.656b2-1.287b3- .753b4 +0.607b5- 2.820b6+ 0.000214b7+ 0.211b8-1.380b9**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.17 Collinearity 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 can see the amount of variance explained by each variable in the model.

After detection and removal of multicollinearity ,we use parsimonious modelling to obtain the best model.

First we use the method of forward selection.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.18 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 | | | |

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 specific criterion. The current model is compared to the model when the concerned predictor.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.19 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) | | | | | | | | | |

The model summary tells us whether the model is successful in predicting miles per gallon(mpg).When one regressor is (disp) is used to predict mpg the value of R is 0.848 and adjusted R square (the estimate of proportion of variance accounted for by regression) is 70.9%.With R now at 0.898 and adjusted R2 up from 70.9% to 80.6%,the addition of 1 more variable has slightly improved the predictive power. But addition of

1 more variable has increased R2 to 85%.The predictive power has significantly improved after adding 1 more variable carb. As p value<0.05 in all the 3 models ,the individual regressors are significant.

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 addition of the new predictors (model 2) causes R2 to increase and this change in the amount of variance that can be explained gives rise to an increase in the F ratio of 0.00 which is again significant. The change statistics therefore tell us about the difference made by adding new predictors to the model. The p value increases to 0.001<0.05 which implies that overall regression is significant in second 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. The p value (0.007<0.05) which implies the overall regression is significant.

As p value <0.05 in all the 3 models ,we may reject our null hypothesis and conclude that the overall regression is significant in all the 3 models. Miles per gallon (mpg) can be explained significantly by the predictors.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.20 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 | | | | | | |

ANOVA tests whether a model is significantly better at predicting the outcome .Specifically ,the F ratio

represents the ratio of the 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 the inaccuracy within the model then the value of F will be greater than 1.For the initial model the F ratio is 14.918 which is very unlikely to have happened by chance (p<0.05).For the second model value of F is higher (17.545) which is also highly significant. The initial model significantly improved our ability to predict the outcome variable ,but that the new model (with extra predictors) is even better(because F ratio is more significant).Finally we notice that the 7 th model is having a very high F value of 52.989 which is very significant at 5% level of significance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.21 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 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.For example:β =0.005 in the first model indicates that as displacement increases by 1 standard deviation miles per gallon increases by 0.005 standard deviations.

**The regression models are;**

**1)y=29.6-0.041b1**

**2)y=36.979-0.029b1-3.431b2**

**3)y=38.207-0.025b1-3.182b2-1.075b3**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.22 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. In forward selection method, 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.

This information together tells us what will happen in the following step. For example consider step 1 which contains 8 excluded variables .Wt has the highest partial correlation(-0.556) and it is statistically significant thus it will be the variable entered in step 2.On the second step with 7 variables (of the 10 )excluded we see that wt with a statistically significant partial correlation of -4.80 wins the struggle for entry next. By the time we reach the 3 rd step there is no variable of the excluded set that has a statistically significant partial correlation for entry at step 4, thus stepwise ends after completing the 3 rd step.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.23 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 criterion but instead of starting the model with only 1 variable ,it begins the model with all predictors included. Then SPSS 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 predictor to be removed is the one which has the least impact on how the model fits the data.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.24 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 | | | | | | | | | |

This table provides an overview of the results. First model contains all the predictors in the regression

model. The adjusted R2 slightly changes from 80.2% to 81% ,meaning that removal of hp does not affect R2 much. After removal of some variables we end up with only 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 0.859 and this change in the amount of variance explained gives rise to an F ratio of 14.918 which is significant with a probability less than 0.05.The addition of the new predictors (model 2) causes R2 to remain constant and this change in the amount of variance that can be explained gives rise to a decrease in the F ratio of 0.00 which is again significant. The change statistics therefore tell us about the difference made by adding new predictors to the model. The p value increases to 0.99>0.05 which implies that overall regression is insignificant in second model. Finally model 7 causes the R2 to change to 0.85 that is 85% of the variation in the model is explained by the model. The p value (0.276>0.05) which implies the overall regression is insignificant.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1.25 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 | | | | | | |

ANOVA tests whether a model is significantly better at predicting the outcome .Specifically ,the F ratio

represents the ratio of the 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 the inaccuracy within the model then the value of F will be greater than 1.For the initial model the F ratio is 14.918 which is very unlikely to have happened by chance (p<0.05).For the second model value of F is higher (17.545) which is also highly significant. The initial model significantly improved our ability to predict the outcome variable ,but that the new model (with extra predictors) is even better(because F ratio is more significant).Finally we notice that the 7 th model is having a very high F value of 52.989 which is very significant at 5% level of significance.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.26 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 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 unstandardised 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 1.27 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.

Then we use stepwise regression method to obtain a parsimonious model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.28 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 1.29 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) | | | | | | | | | |

This table provides an overview of the results. 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 addition of the new predictors (model 2) causes R2 to change and this change in the amount of variance that can be explained gives rise to an increase in the F ratio to 0.01 which is again significant. 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 1.30 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 1.31 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 1.32 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 | | | | | | |

This table contains summaries of the variables that SPSS is considering entering into the model. The summary gives an estimate of each predictors β value, if it was entered into the equation at this point and calculates the 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.

This information together tells us what will happen in the following step. For example consider step 1 which contains 8 excluded variables . Wt has the highest partial correlation(-0.556) and it is statistically significant thus it will be the variable entered in step 2.On the second step with 7 variables ( of the 10 )excluded we see that wt with a statistically significant partial correlation of -4.80 wins the struggle for entry next. By the time we reach the 3 rd step there is no variable of the excluded set that has a statistically significant partial correlation for entry at step 4, thus stepwise ends after completing the 3 rd step.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.33 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 1.34 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 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. For exploratory data analysis this is the best possible method. Stepwise regression is an automated tool used in the exploratory stages of model building to identify a useful subset of predictors.

Now we build a regression model with the 3 best predictors (carb, wt ,disp)

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.35 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 1.36 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 1.37 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 1.38 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 1.39 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.

To test the normality of residuals we must look at the histogram and normal probability plot of the data.

The histogram should look like a normal distribution (a bell shaped curve) .For the mtcars data the distribution is roughly normal.(although there is a slight deficiency of residuals near 0.

Fig 1.1

It shows the data is approximately normally distributed as it resembles a bell shape.

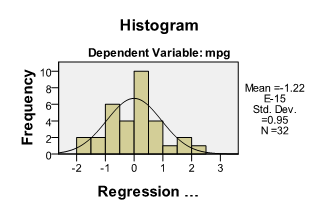
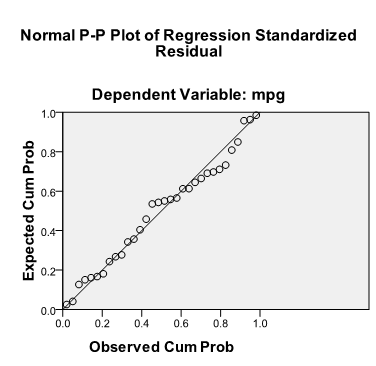


Fig 1.2



The data fits well around the straight line which implies the model is a good fit. The data is normally distributed as data points stay close to the normal line.

| **Table 1.40: Tests of Normality** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | Df | Sig. | Statistic | df | Sig. |
| SMEAN(qsec) | .078 | 32 | .200\* | .971 | 32 | .514 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

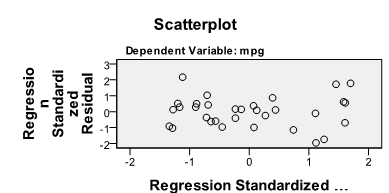
We set the following hypothesis:

Ho-The observed data fits the normal distribution.

H1-The observed 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 observed data fits the normal distribution.

Fig 1.3

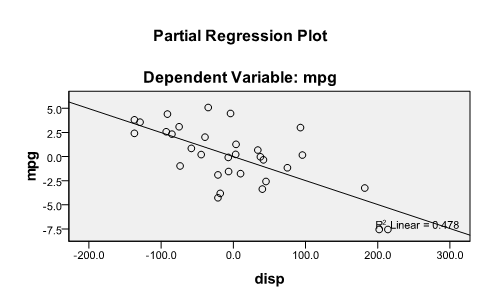


Our residuals are scattered ,it indicates constant variance assumption is not violated.

As the dots are scattered ,it indicates data meet the assumption of errors being normally distributed and the variances of the residuals being constant.

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



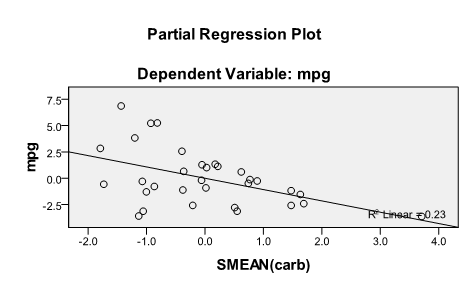
For displacement the partial plot shows a strong negative relationship with mpg. The gradient of the line is β for disp variable in the model. There are no outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe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

Fig 1.5



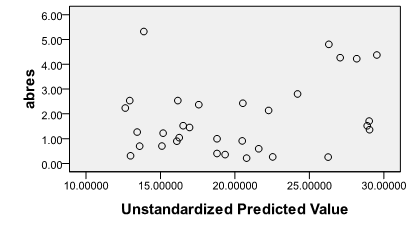
For weight the partial plot shows a negative relationship with mpg. The gradient of the line is β for wt variable in the model. There are no outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe that it 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



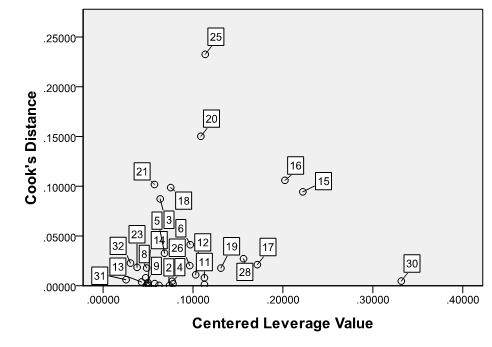
For carburettors the partial plot shows a strong negative relationship with mpg. The gradient of the line is β for carb variable in the model. There are no outliers and the cloud of dots is evenly placed out around the line indicating homoscedasticity. As the regression line fits the data points we observe 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 1.41 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 1.42 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**

1)Internet 2)SPSS for intermediate statistics (second edition)-Use and Interpretation by Nancy L Leech, Karen Barrett, George Morgan 3)Discovering Statistics using SPSS for Windows by Andy Field 4) SPSS Explained by Perry R. Hinton, Isabella McMurray, Charlotte Brownlow 5) Abhishek Sir 's additional notes