

Multiple Regression Analysis

DR. HEMAL PANDYA

Introduction

If a single independent variable and a single dependent variable is used to explain variations, then the model is known as a simple regression model.

If multiple independent variables are used to explain the variation in a single dependent variable, it is called multiple regression model .

Multiple Regression is an appropriate method of analysis when the research problem involves a single metric dependent variable presumed to be related to two or more metric or non-metric independent variables.

Population Simple Linear Regression Function

The population regression model:

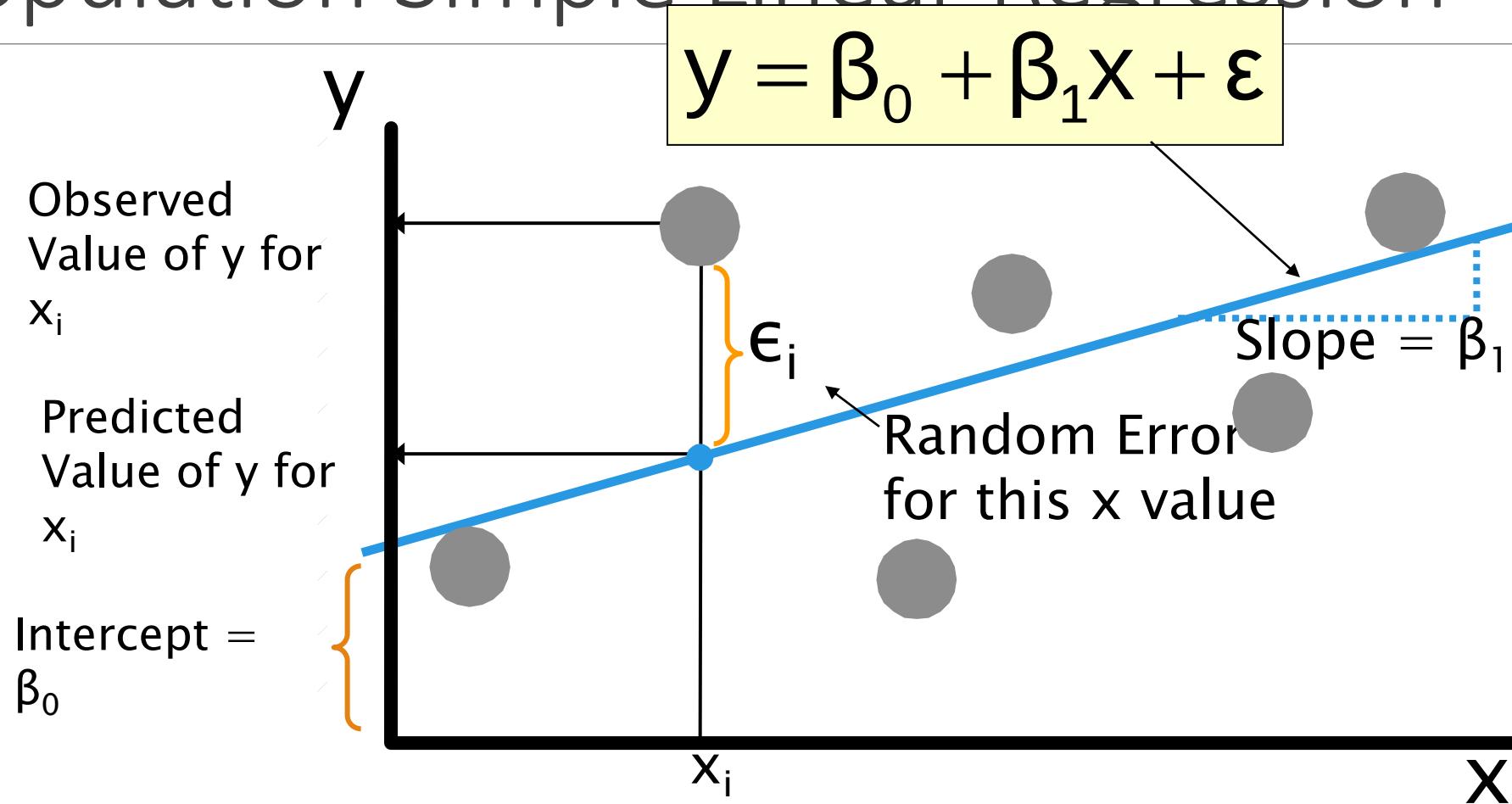
The diagram illustrates the population simple linear regression function $y = \beta_0 + \beta_1 X + \epsilon$. The equation is enclosed in a yellow box. Labels with arrows point to each term: 'Dependent Variable' points to y , 'Population intercept' points to β_0 , 'Population Slope Coefficient' points to β_1 , 'Independent Variable' points to X , and 'Random Error term, or residual' points to ϵ . Below the equation, two curly braces group the terms: the first brace under $\beta_0 + \beta_1 X$ is labeled 'Linear component', and the second brace under ϵ is labeled 'Random Error component'.

$$y = \beta_0 + \beta_1 X + \epsilon$$

Labels and components:

- Dependent Variable: y
- Population intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X
- Random Error term, or residual: ϵ
- Linear component: $\beta_0 + \beta_1 X$
- Random Error component: ϵ

Population Simple Linear Regression



Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated
(or
predicted) y
value

Estimate of
the
regression
intercept

Estimate of the
regression
slope

Independent
variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Interpretation of the Slope and the Intercept

b_0 is the estimated average value of y when the value of x is zero

b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Least Squares Criterion

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

The Least Squares Equation

The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b_1 = r \frac{s_y}{s_x}$$

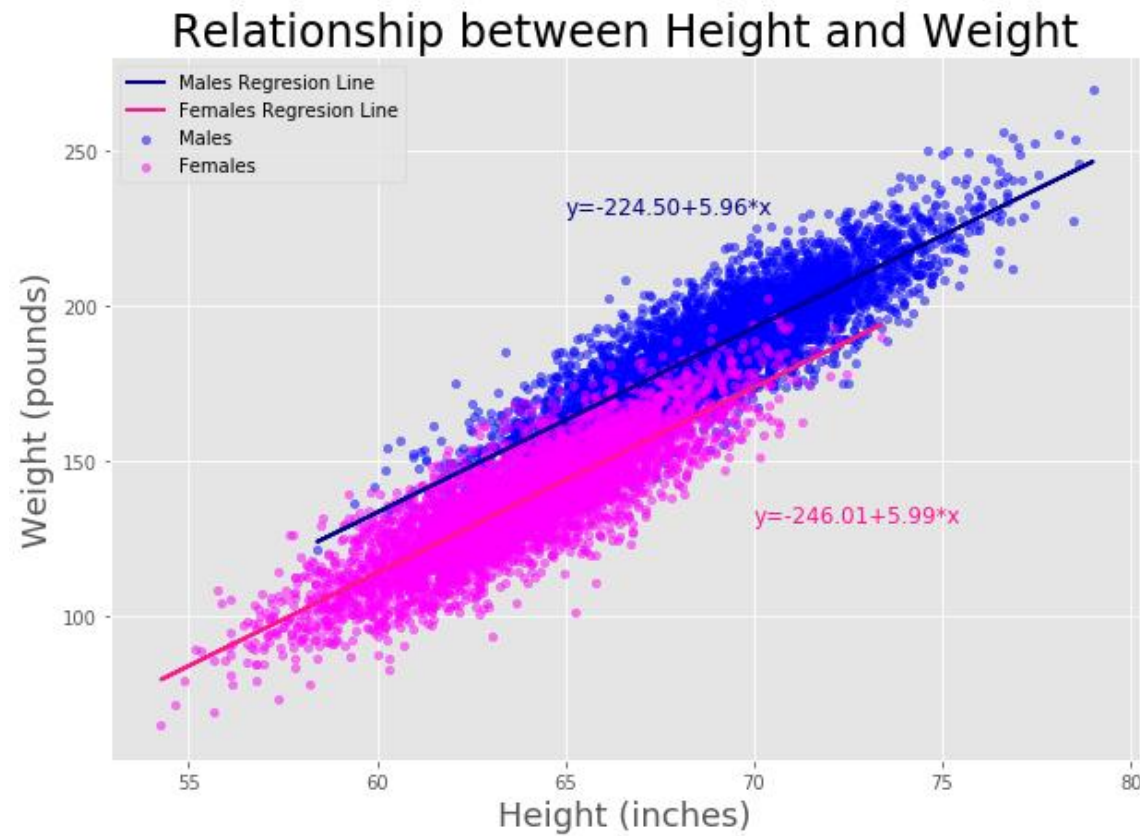
algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Simple Linear Reression



Introduction

Multiple Regression helps to predict the changes in dependent variable in response to changes in independent variables. This objective is most often achieved through the rule of least squares. The main objective of regression analysis is to explain the variations in one variable (dependent variable), based on the variations in one or more variables (independent variables). The form of regression equation could be linear or non-linear.

Multiple Linear Regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e_i$$

β_0 is the intercept term and

β_i is the slope of Y on dimension X_i

- $\beta_1, \beta_2, \dots, \beta_n$ called “partial” regression coefficients

The magnitudes (and even signs) of $\beta_1, \beta_2, \dots, \beta_n$ depend on which other variables are included in the multiple regression model might not agree in magnitude (or even sign) with the bivariate correlation coefficient r between X_i and Y

Flow chart

Stage 1

Research Problem

Select objectives:

- Prediction
- Explanation

Select dependent and independent variables

Stage 2

Research Design Issues:

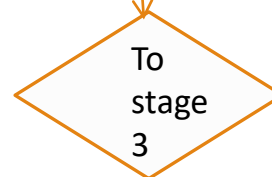
Obtain an adequate sample size to ensure:

- Statistical power
- Generalizability

Creating Additional Variables

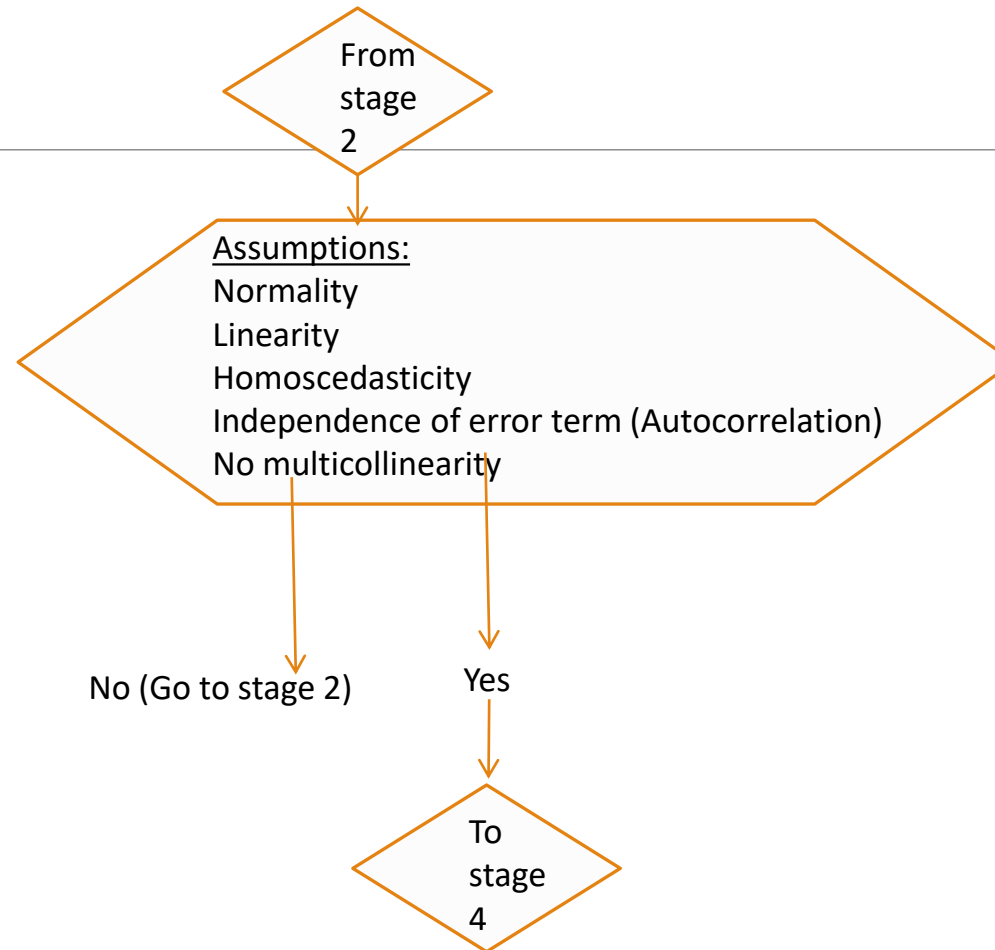
Transformation to meet assumptions

Dummy variables for use of non- metric
variables



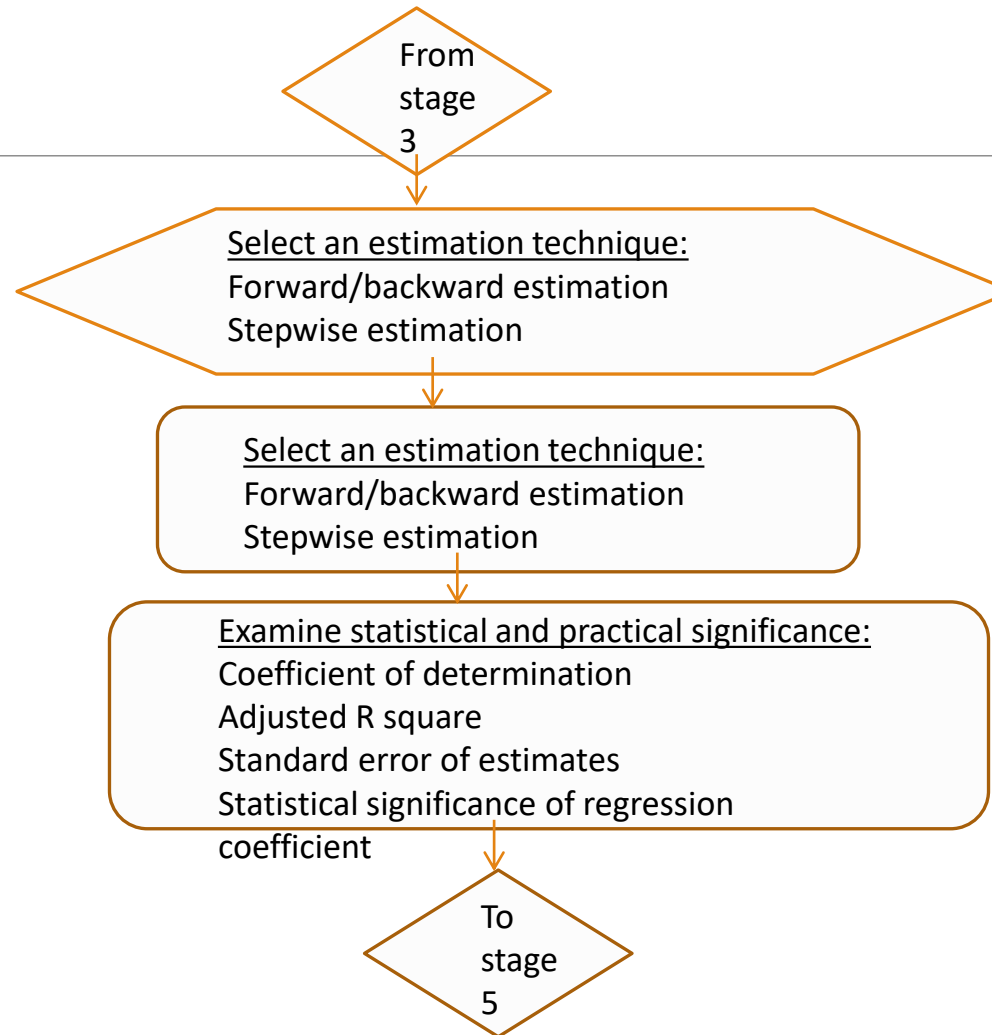
Flow chart

Stage 3



Flow chart

Stage 4



Flow chart

Stage 5

From
stage
4

Interpret the regression variable:

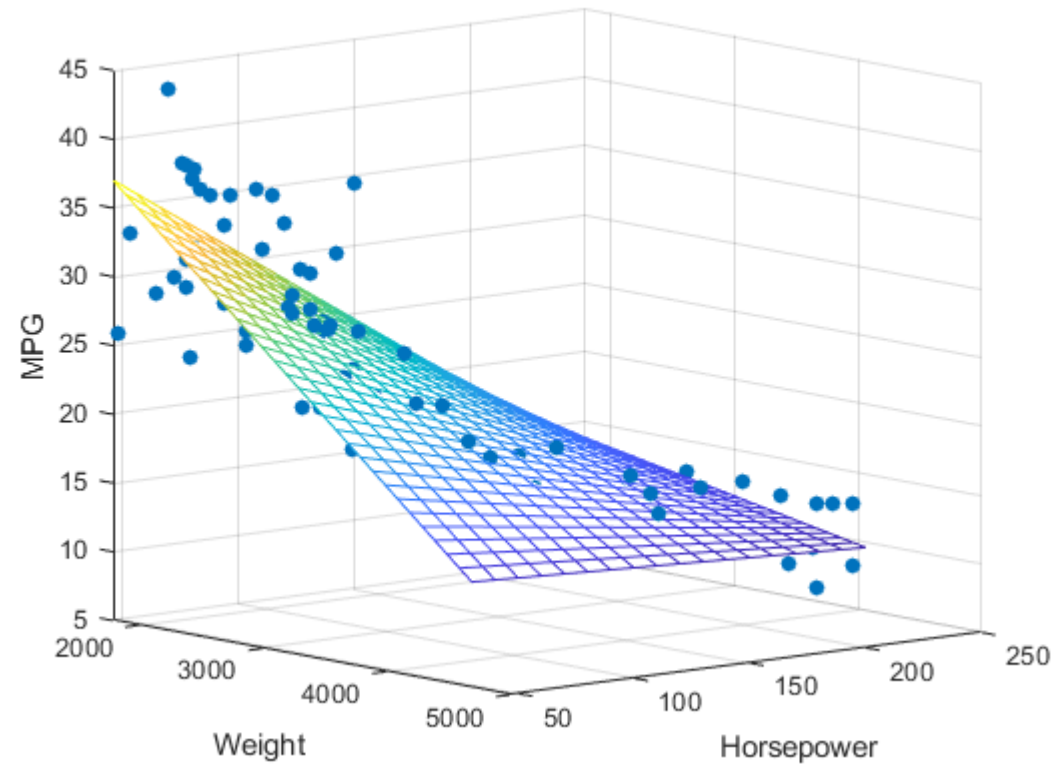
- Evaluate the prediction equation with the regression coefficient
- Evaluate the relative importance of independent variables with the beta coefficients
- Assess Multicollinearity

Stage 6

Validate the results:

Split-sample analysis

Visualization of Multiple Regression



Residuals in Multiple Regression Model

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$
$$e = y - \hat{y}$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Here $\hat{y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2X_{2i} - \dots - \hat{\beta}_kX_{ki})^2$$

A residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model.

The residual sum of squares is one of many statistical properties enjoying a renaissance in financial markets. Ideally, the sum of squared residuals should be a smaller or lower value in any regression model.

Assumptions of Multiple Linear Regression

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

- Linear relationship
- Multivariate Normality
- No or little Multicollinearity
- No Auto-correlation
- Homoscedasticity

Properties of a Good Estimator

- Unbiasedness
- Consistency
- Sufficiency
- Efficiency

Gauss Markov Theorem

Gauss–Markov theorem, states that in a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, the **Best Linear Unbiased Estimator (BLUE)** of the coefficients is given by the ordinary least squares (OLS) estimator.

Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators.

The **Gauss–Markov** assumptions concern the set of error random variables, ε_i :

- They have mean zero: $E[\varepsilon_i]=0$
- They are **homoscedastic**, that is all have the same finite variance:
- $\text{Var}(\varepsilon_i)=\sigma^2$
- Distinct error terms are uncorrelated : $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$. **(No Autocorrelation)**
- The explanatory variables are uncorrelated with each other. **(No Multicollinearity)**
- The model is Completely Specified. **(No Specification Bias)**
- The model is exactly identified **(No Identification Bias)**

Case Study:

A marketing manufacturer and a marketer of electric motors would like to build a regression model consisting of five or six independent variables, to predict sales. Past data has been collected for 15 sales territories, on sales and six different independent variables. Build a regression model and recommend whether or not it should be used by the company.

Variables under Study:

Dependent variable	
Y	Sales in RS. Lakh in the territory
Independent variable	
X ₁	Market Potential(in RS. Lakh
X ₂	No. of dealers
X ₃	No. of salespeople
X ₄	Index of computer activity (on 5 point scale)(1= low, 5=high level)
X ₅	No. of service people
X ₆	No. of existing customers

Data

	1 SALES	2 POTENTL	3 DEALERS	4 PEOPLE	5 COMPET	6 SERVICE	7 CUSTOM
1	5	25	1	6	5	2	20
2	60	150	12	30	4	5	50
3	20	45	5	15	3	2	25
4	11	30	2	10	3	2	20
5	45	75	12	20	2	4	30
6	6	10	3	8	2	3	16
7	15	29	5	18	4	5	30
8	22	43	7	16	3	6	40
9	29	70	4	15	2	5	39
10	3	40	1	6	5	2	5
11	16	40	4	11	4	2	17
12	8	25	2	9	3	3	10
13	18	32	7	14	3	4	31
14	23	73	10	10	4	3	43
15	81	150	15	35	4	7	70

SPSS Commands for Multiple Regression

Type the data along with the variable labels and the value labels in an SPSS file, and to get the output for a regression problem, follow the directions:

1. Click on ANALYSE at the SPSS menu bar.
2. Click on REGRESSION, followed by LINEAR.
3. In the dialogue box which appears, select a dependent variable by clicking on the arrow leading to the dependent box after highlighting the appropriate variable from the list of the variables on the left side.
4. Select the independent variables to be included in the regression model in the same way, transferring them from left side to the right side box by clicking on the arrow leading to the box called independent variables or independents.

SPSS Commands for Multiple Regression

5. In the same dialogue box, select the METHOD. Choose:

- ENTER as the method if you want all independent variables to be included in the model.
- STEPWISE if you want to use forward stepwise regression.
- BACKWARD if you want to use a backward stepwise regression.

6. Select OPTIONS if you want additional output options, select the ones you want, and click CONTINUE.

7. Select PLOTS if you want to see some plots such as residual plots, select those you want, and click CONTINUE.

8. Click OK from the main dialogue box to get the REGRESSION Output.

Testing for Normality of Variables

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SALES	.254	15	.010	.819	15	.007
PONTL	.267	15	.005	.783	15	.002
DEALER	.190	15	.151	.905	15	.115
PEOPLE	.179	15	.200*	.863	15	.026
SERVICE	.192	15	.143	.885	15	.056
CUSTOM	.137	15	.200*	.954	15	.596

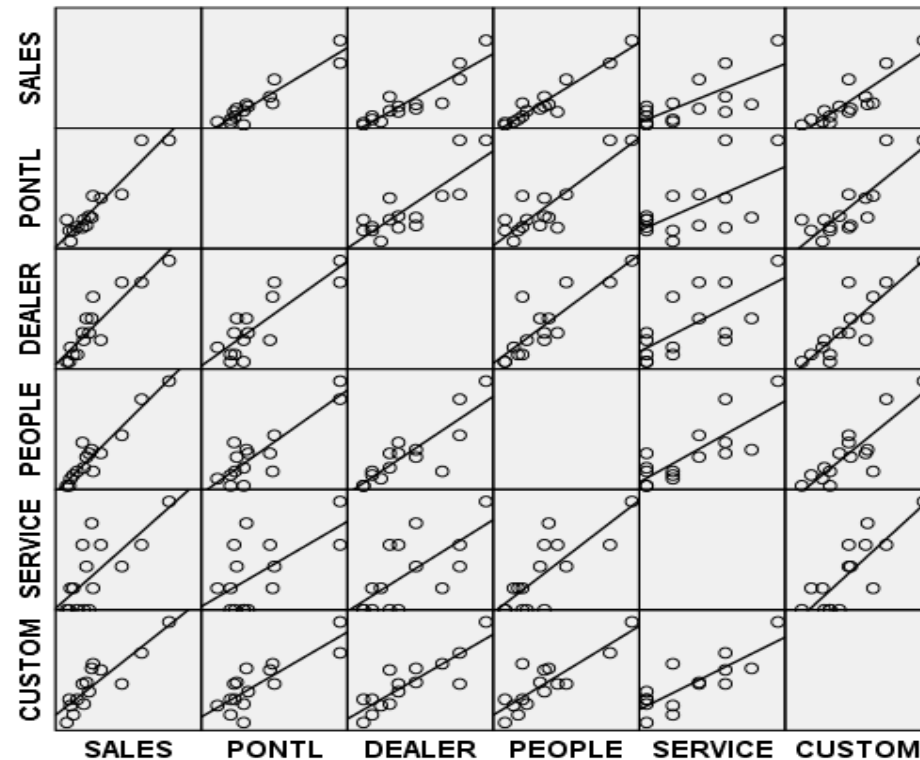
SPSS Commands for Testing Linearity

Assumption #1: The relationship between the IVs and the DV is linear.

- To produce a scatterplot, **CLICK** on the **Graphs** menu option and **SELECT Chart Builder**
- To produce a scatterplot, **SELECT** the **Scatter/Dot** option from the Gallery options in the bottom half of the dialog box. Then drag and drop the **Scatterplot Matrix icon for one or** into the **Chart Preview Window**.
- Next, we need to tell SPSS what to draw. To do this, drag and drop the DV onto the graph's Y-Axis and one of the IVs onto the graph's X-Axis.
- **Click OK**
- You will get the Scatter Plot Matrix in the Output Sheet
- Select it with a **DOUBLE CLICK** and Insert Straight Lines on each Scatter Plot of the Matrix from the new window and close that new Window

You will get the following Output. Scatterplots show that this assumption had been met (although you would need to formally test each IV yourself).

Testing for Linearity



Testing for Multicollinearity

Assumption #2: There is no multicollinearity in your data. This is essentially the assumption that your predictors are not too highly correlated with one another.

- To test this assumption, go to ANALYZE> Regression>Linear
- Insert the Dependent and Independent Variables in their respective Dialogue Boxes.
- **SELECT Statistics>Collinearity Diagnostics.**
- **Press Continue**

SPSS OUTPUT: Correlation Matrix

Correlation coefficients and significance value

	SALES	POTENTL	DEALERS	PEOPLE	COMPET	SERVICE	CUSTOM
SALES	1	0.945	0.908	0.953	-0.046	0.726	0.878
Sig.	.	(0)	(0)	(0)	(0.436)	(0.001)	(0)
POTENTL	0.945	1	0.837	0.877	0.14	0.613	0.831
Sig.	(0)	.	(0)	(0)	(0.309)	(0.008)	(0)
DEALERS	0.908	0.837	1	0.855	-0.082	0.685	0.86
Sig.	(0)	(0)	.	(0)	(0.385)	(0.002)	(0)
PEOPLE	.953	0.877	0.855	1	-0.036	0.794	0.854
Sig.	(0)	(0)	(0)	.	(0.449)	(0)	(0)
COMPET	-0.046	0.14	-0.082	-0.036	1	-0.178	-0.015
Sig.	(0.436)	(0.309)	(0.385)	(0.449)	.	(0.263)	(0.479)
SERVICE	0.726	0.613	0.685	0.794	-0.178	1	0.818
Sig.	(0.001)	(0.008)	(0.002)	(0)	(0.263)	.	(0)
CUSTOM	0.878	0.831	0.86	0.854	-0.015	0.818	1
Sig.	(0)	(0)	(0)	(0)	(0.479)	(0)	.

Zero-Order Correlations

First, a zero-order correlation simply refers to the correlation between two variables (i.e., the independent and dependent variable) without controlling for the influence of any other variables. Essentially, this means that a zero-order correlation is the same thing as a Pearson correlation. So why are we discussing the zero-order correlation here? When conducting an analysis with more than two variables (i.e., multiple independent variables or control variables), it may be of interest to know the simple bivariable relationships between the variables to get a better sense of what happens when you begin to control for other variables. This is why SPSS gives you the option to report zero-order correlations when running a multiple linear regression analysis.

Checking for Multicollinearity: Zero Order Correlation

The coefficient of correlation r as a measure of the degree of linear association between two variables. For the three-variable regression model we can compute three correlation coefficients: r_{12} (correlation between Y and X_2), r_{13} (correlation coefficient between Y and X_3), and r_{23} (correlation coefficient between X_2 and X_3); notice that we are letting the subscript 1 represent Y for notational convenience. These correlation coefficients are called **gross** or **simple correlation coefficients**, or **correlation coefficients of zero order**.

Checking for Multicollinearity

The first assumption we can test is that the predictors (or IVs) are not too highly correlated. We can do this in two ways. First, we need to look at the Correlations table. Correlations of more than 0.8 may be problematic. If this happens, consider removing one of your IVs.

Further we can examine Part and Partial Correlations. The partial and Part Correlation coefficients are all less than 0.8 indicating moderate Multicollinearity amongst the explanatory variables.

Coefficients ^a						
Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	PONTL	.945	.766	.181	.158	6.348
	DEALER	.908	.489	.085	.218	4.582
	PEOPLE	.953	.672	.138	.115	8.684
	COMPUTER	-.046	-.437	-.074	.800	1.251
	SERVICE	.726	-.064	-.010	.328	3.044
a. Dependent Variable: SALES						

Checking for Multicollinearity: Partial Correlations

Next, a partial correlation is the correlation between an independent variable and a dependent variable after controlling for the influence of other variables on both the independent and dependent variable. In a partial correlation, the influence of the control variables on both the independent and dependent variables are taken into account.

$r_{12.3}$ = partial correlation coefficient between Y and X_2 , holding X_3 constant

$r_{13.2}$ = partial correlation coefficient between Y and X_3 , holding X_2 constant

$r_{23.1}$ = partial correlation coefficient between X_2 and X_3 , holding Y constant

Checking for Multicollinearity: Partial Correlations

These partial correlations can be easily obtained from the simple or zero-order, correlation coefficients as follows (for proofs, see the exercises)²²:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (7.11.1)$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \quad (7.11.2)$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}} \quad (7.11.3)$$

The partial correlations given in Eqs. (7.11.1) to (7.11.3) are called **first-order correlation coefficients**. By *order* we mean the number of secondary subscripts. Thus $r_{12.34}$ would be the correlation coefficient of order two, $r_{12.345}$ would be the correlation coefficient of order three, and so on. As noted previously, r_{12} , r_{13} , and so on are called *simple* or *zero-order correlations*. The interpretation of, say, $r_{12.34}$ is that it gives the coefficient of correlation between Y and X_2 , holding X_3 and X_4 constant.

Checking for Multicollinearity: Partial Correlations

3. Examination of partial correlations. Because of the problem just mentioned in relying on zero-order correlations, Farrar and Glauber have suggested that one should look at the partial correlation coefficients.¹⁹ Thus, in the regression of Y on X_2 , X_3 , and X_4 , a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}$, $r^2_{13.24}$, and $r^2_{14.23}$ are comparatively low may suggest that the variables X_2 , X_3 , and X_4 are highly intercorrelated and that at least one of these variables is superfluous.

Checking for Multicollinearity: Part or Semi-Partial Correlations

This brings us to the part correlation, which is sometimes referred to as the “semi-partial” correlation. Like the partial correlation, the part correlation is the correlation between two variables (independent and dependent) after controlling for one or more other variables. However, for the part correlation, only the influence of the control variables on the independent variable is taken into account. In other words, the part correlation does not control for the influence of the confounding variables on the dependent variable. You might wonder why you would only want to control for effects on the independent variable and not the dependent variable? The primary reason for conducting **the part correlation would be to see how much unique variance the independent variable explains in relation to the total variance in the dependent variable**, rather than just the variance unaccounted for by the control variables.

Checking for Multicollinearity: Partial & Part Correlations

In general, If all the Partial and Part Correlation Coefficients are greater than 0.7 and they are significant, then there is an indication of the existence of Multicollinearity. Either of the two variables which have very high partial or part coefficients must be deleted from the model.

SPSS OUTPUT: Collinearity Diagnostics

- Tolerance is measure of co-linearity and multicollinearity.
- The tolerance of variable 'i' (TOL i) is $(1 - (R_i)^2)$.
- **TOL** should be Greater than **0.2**
- The range of tolerance depends on the situation.
- Variance inflation factor(VIF) is directly related to tolerance value $VIF\ i = (1 / TOL\ i)$
- Large VIF indicates a high degree of co-linearity or multicollinearity among the independent variables.
- Range of **VIF** is zero to infinity. It should be **Less than 10**. Greater than 10 indicates severe Multicollinearity

Testing for Autocorrelation

Assumption #3: The values of the residuals are independent. This is basically the same as saying that we need our observations (or individual data points) to be independent from one another (or uncorrelated).

- We can test this assumption using the **Durbin-Watson** statistic, so **SELECT** this option.
- **CLICK Continue** to continue.
- To test the next assumption, **CLICK** on the **Plots** option in the main Regression Dialog box.

Testing for Autocorrelation: Model Summary

Model Summary ^b									
Model	R	R Square	Adjusted R Square	Durbin-Watson	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.989 ^a	.977	.960	1.909	.977	57.133	6	8	.000

- Durbin-Watson Statistic measures autocorrelation.
- $D \approx 2(1 - \hat{\rho})$
- If strong positive autocorrelation then $\hat{\rho} = 1$ and $DW = 0$
- If strong negative autocorrelation then $\hat{\rho} = -1$ and $DW = 4$
- if no autocorrelation then $\hat{\rho} = 0$ and $DW = 2$
- So the best can hope for is a DW of 2.
- Here, the value of Durbin-Watson is nearer to 2, it implies that there is no autocorrelation in the model.

Testing Homoscedasticity

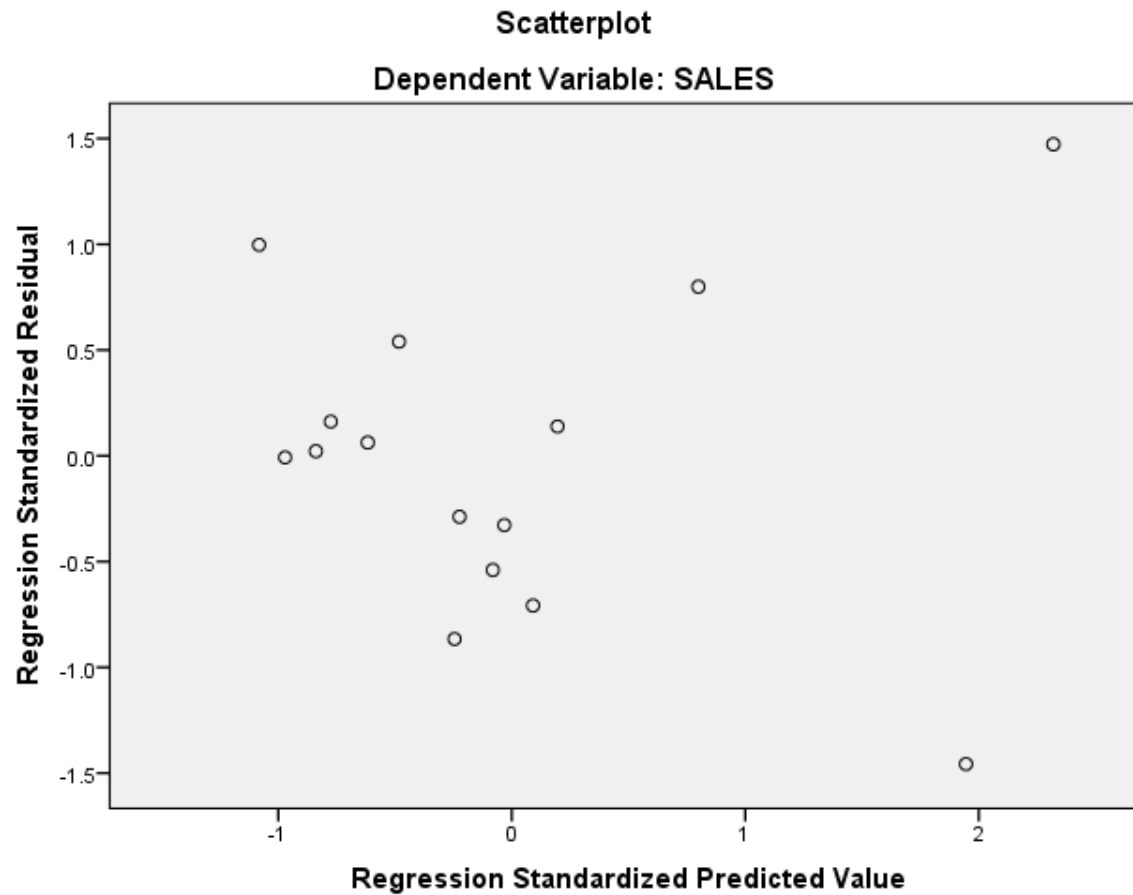
Assumption #4: The variance of the residuals is constant.

This is called *homoscedasticity*, and is the assumption that the variation in the residuals (or amount of error in the model) is similar at each point across the model. In other words, the spread of the residuals should be fairly constant at each point of the predictor variables (or across the linear model). We can get an idea of this by looking at our original scatterplot... but to properly test this, we need to ask SPSS to produce a special scatterplot for us that includes the whole model (and not just the individual predictors).

To test the 4th assumption, we need to plot the standardized values our model would predict, against the standardized residuals as obtained.

To do this, first **CLICK** on the **ZPRED** variable and **MOVE** it across to the **X-axis**. Next, **SELECT** the **ZRESID** variable and **MOVE** it across to the **Y-axis**.

Testing for Homoscedasticity



Testing for Homoscedasticity

- This graph plots the standardized values our model would predict, against the standardized residuals obtained.
- As the predicted values increase (along the X-axis), the variation in the residuals should be roughly similar. If everything is ok, this should look like a random array of dots.
- If the graph looks like a funnel shape, then it is likely that this assumption has been violated.

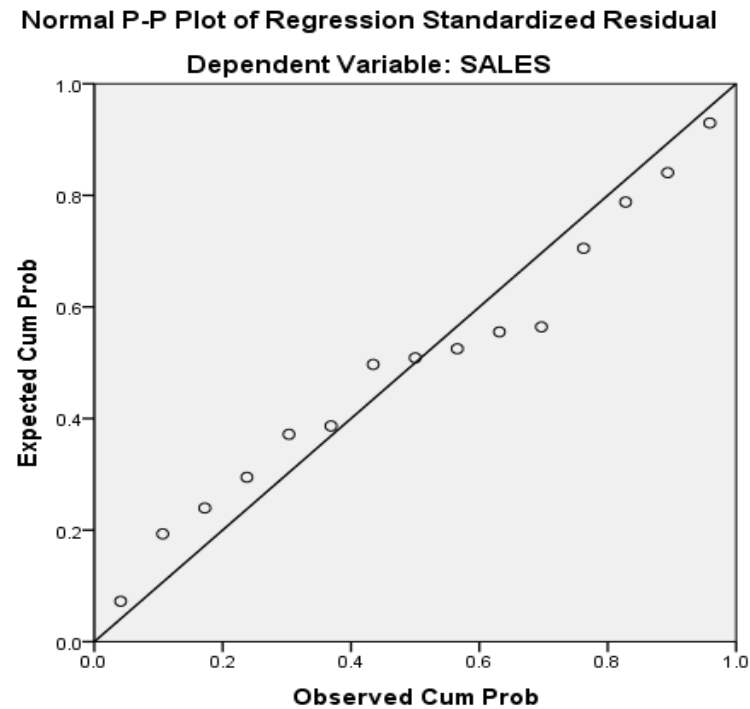
Testing Normality of Residuals

Assumption #5: The values of the residuals are normally distributed. This assumption can be tested by looking at the distribution of residuals. We can do this by **CHECKING** the **Normal probability plot** option.

Next, **SELECT Continue**

In our case, the P-P plot for the model suggested that the assumption of normality of the residuals may have been violated. However, as only extreme deviations from normality are likely to have a significant impact on your findings, the results are probably still valid.

Normality of Residuals



Testing for Outliers

Assumption #6: There are no influential cases biasing your model.

Significant outliers and influential data points can place undue influence on your model, making it less representative of your data as a whole. To identify any particularly influential data points, first **CLICK** the **SAVE** option in the main Regression dialog box.

You can test for influential cases using **Cook's Distance**.

SELECT the **Cook's** option now to do this.

Then **CLICK** on **Continue**

And finally **CLICK** on **OK** in the main Regression dialog box to run the analysis.

SPSS now produces both the results of the multiple regression, and the output for assumption testing.

Testing for Outliers

Assumption #6: There are no influential cases biasing your model.

Cook's Distance values were all under 1, suggesting individual cases were not unduly influencing the model.

The values of Cook's Distance will be displayed in the DATA VIEW of SPSS. You will note a new variable (Column) will be created displaying these values for all the observations in your SPSS Data Sheet.

SPSS CORE OUTPUT : Descriptive Statistics

Descriptive Statistics				
	Mean	Std. Deviation	N	CV
SALES	24.1333	21.98008	15	91.07781
POTENTL	55.8000	42.54275	15	76.24149
DEALERS	6.0000	4.40779	15	73.46317
PEOPLE	14.8667	8.33981	15	56.09725
COMPET	3.4000	.98561	15	28.98853
SERVICE	3.6667	1.63299	15	44.53569
CUSTOM	29.7333	16.82883	15	56.59927

Lowest C.V. indicates that particular variable is most consistent.

Variable 4 the index of computer activity is having lowest c.v., thus it is the most consistent variable.

SPSS CORE OUTPUT:

Regression Coefficients

Coefficients ^a						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.173	5.813		-.546	.600
	POTENTL	.227	.075	.439	3.040	.016
	DEALERS	.819	.631	.164	1.298	.230
	PEOPLE	1.091	.418	.414	2.609	.031
	COMPET	-1.893	1.340	-.085	-1.413	.195
	SERVICE	-.549	1.568	-.041	-.350	.735
	CUSTOM	.066	.195	.050	.338	.744

- The STANDARDISED COEFFICIENTS indicates variation in dependent variable explained by each independent variable.(consider absolute value) when intercept is zero.
- UNSTANDARDISED COEFFICIENTS are coefficients of independent variable in regression equation.
- Here we can see that highest variation is explained by COMPET i.e. 189.3% and lowest variation is explained by CUSTOM i.e. 6.6%.

Testing the Significance of Coefficient Estimates

the estimators $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$ are themselves normally distributed with means equal to True $\beta_1, \beta_2, \dots, \beta_n$, hence we can test the significance of these coefficient estimates using one- sample *t*-test as follows:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}$$
$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)}$$
$$t = \frac{\hat{\beta}_3 - \beta_3}{\text{se}(\hat{\beta}_3)}$$

Coefficient of Determination: R-Square and Adj. R-Square

$$\text{TSS} = \text{ESS} + \text{RSS},$$

$$\begin{aligned} R^2 &= \frac{\text{ESS}}{\text{TSS}} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \end{aligned}$$

An important property of R^2 is that it is a nondecreasing function of the number of explanatory variables or regressors present in the model; as the number of regressors increases, R^2 almost invariably increases and never decreases. Stated differently, an additional X variable will not decrease R^2 .

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2 / (n - k)}{\sum y_i^2 / (n - 1)}$$

where k = the number of parameters in the model *including the intercept term*. (In the three-variable regression, $k = 3$. Why?) The R^2 thus defined is known as the **adjusted R^2** , denoted by $\bar{R}^2$. The term *adjusted* means adjusted for the df associated with the sums of squares entering into (7.8.1): $\sum \hat{u}_i^2$ has $n - k$ df in a model involving k parameters, which include

... it is good practice to use $\bar{R}^2$ rather than R^2 because R^2 tends to give an overly optimistic picture of the fit of the regression, particularly when the number of explanatory variables is not very small compared with the number of observations.¹¹

SPSS CORE OUTPUT: Model Summary

Model Summary ^b									
Model	R	R Square	Adjusted R Square	Durbin-Watson	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.989 ^a	.977	.960	1.909	.977	57.133	6	8	.000

- Here, $r = 0.989$ which indicates a strong uphill (positive) linear relationship.
- R square is 0.977 i.e. 97.7 % variation in dependent variable has been explained by independent variables and remaining 2.7% is due to error.

- Adjusted R square is 0.960

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

[NOTE:

- Use the adjusted R-square to compare models with different numbers of predictors
- Use the predicted R-square to determine how well the model predicts new observations and whether the model is too complicated.]

SPSS OUTPUT: ANOVA

Testing the Overall Significance of a Multiple Regression: The F Test

Decision Rule. Given the k -variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

(i.e., all slope coefficients are simultaneously zero) versus

H_1 : Not all slope coefficients are simultaneously zero

compute

$$F = \frac{\text{ESS}/\text{df}}{\text{RSS}/\text{df}} = \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \quad (8.5.7)$$

If $F > F_\alpha(k-1, n-k)$, reject H_0 ; otherwise you do not reject it, where $F_\alpha(k-1, n-k)$ is the *critical* F value at the α level of significance and $(k-1)$ numerator df and $(n-k)$ denominator df. Alternatively, if the p value of F obtained from (8.5.7) is sufficiently low, one can reject H_0 .

SPSS OUTPUT: ANOVA TABLE

ANOVA TABLE FOR THE THREE-VARIABLE REGRESSION

Source of variation	SS	df	MSS
Due to regression (ESS)	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{2}$
Due to residual (RSS)	$\sum \hat{u}_i^2$	$n - 3$	$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 3}$
Total	$\sum y_i^2$	$n - 1$	

ANOVA: F-Test

$$TSS = ESS + RSS,$$

$$F = \frac{\text{MSS of ESS}}{\text{MSS of RSS}}$$

$$= \frac{ESS/df}{RSS/df}$$

$$F = \frac{ESS/(k-1)}{RSS/(n-k)}$$

SPSS OUTPUT: ANOVA

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6609.485	6	1101.581	57.133	.000 ^b
	Residual	154.249	8	19.281		
	Total	6763.733	14			

Here, the significance value is 0.000 i.e. it is less than 0.05(level of significance).

If the p-value is less than or equal to significance level, NULL HYPOTHESIS is rejected and concluded that there is a significance difference between means.

Dealing with Econometric Problems Associated with OLS

- Non-zero Expectation of Residuals
- Multicollinearity
- Heteroscedasticity
- Autocorrelation
- Specification Bias
- Identification Bias
- Non-Normality of Residuals

DETECTION OF MULTICOLLINEARITY

1. High R^2 but few significant t ratios.

2. High pair-wise correlations among regressors:

Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem.

3. Examination of partial correlations: Values of partial correlations greater than 0.7

DETECTION OF MULTICOLLINEARITY

4. **Auxiliary Regressions:** one way of finding out which X variable is related to other X variables is to regress each X_i on the remaining X variables and compute the corresponding R^2 , which we designate as R_i^2 ; each one of these regressions is called an **Auxiliary Regression**, auxiliary to the main regression of Y on the X 's.

NOTE: SAS uses *Eigen values* and the *condition index* to diagnose multicollinearity. The value of Condition Index less than 30 is acceptable to support the Null Hypothesis of No Multicollinearity.

The **condition number k** defined as:

$$CI = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

Consequences of Multicollinearity

1. Although BLUE, the OLS estimators have large variances and covariance, making precise estimation difficult.

➤ The speed with which variances and covariance increase can be seen with the **variance-inflating factor (VIF)**, which is defined as-

$$VIF = \frac{1}{(1 - r_{23}^2)}$$

■ VIF shows how the variance of an estimator is *inflated* by the presence of multicollinearity

2. Because of consequence 1 (VIF > 10), the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” (i.e., the true population coefficient is zero) more readily.

■ In cases of high collinearity the estimated standard errors increase dramatically, thereby making the calculated values of t-test smaller. Therefore, in such cases, one will increasingly accept the null hypothesis that the relevant true population value is zero.

CONSEQUENCES OF MULTICOLLINEARITY

3. Also because of consequence 1, the t ratio of one or more coefficients tends to be statistically insignificant. As noted, this is the “classic” symptom of multicollinearity.

- If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero.

4. Although the t ratio of one or more coefficients is statistically insignificant, R^2 , the overall measure of goodness of fit, can be very high.

- In cases of high collinearity, it is possible to find, as we have just noted, that one or more of the partial slope coefficients are individually statistically insignificant on the basis of the t test.
- Yet the R^2 in such situations may be so high, say, in excess of 0.9, that on the basis of the F test one can convincingly reject the hypothesis.
- This is one of the signals of multicollinearity—insignificant t values but a high overall R^2 (and a significant F value).

Consequences of Multicollinearity

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

6. Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

Multicollinearity-Remedial Measures

Do nothing or

follow some rules of thumb.

Use of A priori information

Combining cross-sectional and time series data.

Dropping a variable(s) and specification bias.

Transformation of variables

Additional or new data.

Reducing collinearity in polynomial regressions

Other methods of remedying multicollinearity i.e. factor analysis or principal components or ridge regression

CONSEQUENCES of Autocorrelation

1. **CONSEQUENCES OF USING OLS:** in the presence of autocorrelation the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed, but they are no longer efficient.

2. **THE BLUE ESTIMATOR IN THE PRESENCE OF AUTOCORRELATION**

DETECTING AUTOCORRELATION

I. Graphical Method

Use a graph of residuals versus data order (1, 2, 3, 4, n) to visually inspect residuals for autocorrelation.

A positive autocorrelation is identified by a clustering of residuals with the same sign.

A negative autocorrelation is identified by fast changes in the signs of consecutive residuals.

II. The Runs Test

III. Durbin–Watson *d* Test

Use the Durbin-Watson statistic to test for the presence of autocorrelation.

The test is based on an assumption that errors are generated by a first-order autoregressive process.

Autocorrelation-REMEDIAL MEASURES

1. Try to find out if the autocorrelation is pure autocorrelation and not the result of mis-specification of the model. sometimes we observe patterns in residuals because the model is mis-specified that is, it has excluded some important variables—or because its functional form is incorrect.
- 2.If it is pure autocorrelation, one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (pure) autocorrelation.
3. In large samples, we can use the **Newey–West method to obtain** standard errors of OLS estimators that are corrected for autocorrelation. This method is actually an extension of White's heteroscedasticity - consistent standard errors method.
4. In some situations we can continue to use the OLS method.

Consequences of heteroscedasticity

➤ OLS ESTIMATION IN THE PRESENCE OF HETEROSCEDASTICITY:

$\hat{\beta}$ is *still linear unbiased and consistent but, it is no longer best* and does not possess the minimum variance o.

DETECTION OF HETEROSCEDASTICITY

1. Informal Methods

Nature of the Problem

Graphical Method

2. Formal Methods

Park Test

Glejser Test

Spearman's Rank Correlation Test.

Goldfeld-Quandt Test

Breusch–Pagan–Godfrey Test

White's General Heteroscedasticity Test

REMEDIAL MEASURES-Heteroscedasticity

1. When σ_i^2 Is Known: *The Method of Weighted Least Squares.*
2. When σ_i^2 Is Not Known: White's Heteroscedasticity-Consistent Variances and Standard Errors

THANK YOU