Review Article

F-Test of Overall Significance in Regression Analysis Simplified

Onchiri Sureiman, Callen Moraa Mangera[1]

Department of Educational Planning and Management, Masinde Muliro University and Technology, Kakamega, [1]Department of Physiotherapy, The Nairobi Hospital, Nairobi, Kenya

Abstract

Regression analysis is using the relationship between a known value and an unknown variable to estimate the unknown one. Here, an estimate of the dependent variable is made corresponding to given values of independent variables by placing the relationship between the variables in the form of a regression line. To determine how well the regression line obtained fits the given data points, F-test of overall significance is conducted. The issues involved in the F-test of overall significance are many and mathematics involved is rigorous, especially when more than two variables are involved. This study describes in details how the test can be conducted and finally gives the simplified approach of test using an online calculator.

Keywords: F-test, hypothesis testing, online calculator, regression

Introduction

The term "regression" was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or "regress" toward the mean height of the population. He designated the word regression as name of the process of predicting one variable from another variable.[1] Then came the term "multiple regression" to describe the process by which several variables are used to predict one another.[2] The F-Test of overall significance in regression is a test of whether or not your linear regression model provides a better fit to a dataset than a model with no predictor variables.

Assumptions Underlying F-test of Overall Significance in Regression Analysis

The main assumptions include:

Linearity

Linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots.[3]

Normality

The linear regression analysis requires all variables to be multivariate normal.[4] This assumption can best be checked with a histogram or a Q-Q-Plot. There are also a variety of statistical tests for normality, including the Kolmogorov–Smirnov test, the Shapiro–Wilk test, the Jarque–Bera test, and the Anderson–Darling test.[5] When the data are not normally distributed a nonlinear transformation (e.g., log-transformation) might fix this issue.

Multicollinearity

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.[3] Multicollinearity may be tested with three central criteria:

i. Tolerance – The tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 - R^2$ for these first step regression analysis.[6] With $T < 0.1$ there might be

Access this article online

Quick Response Code:

Website: www.j-pcs.org

DOI: 10.4103/jpcs.jpcs_18_20

Address for correspondence: Mr. Onchiri Sureiman, Department of Educational Planning and Management, Masinde Muliro University and Technology, Kakamega, Kenya.
E-mail: sureimanonchiri@gmail.com

Submitted: 04-Mar-2020 **Revised:** 30-Apr-2020
Accepted: 28-Jun-2020 **Published:** 27-Aug-2020

How to cite this article: Sureiman O, Mangera CM. F-test of overall significance in regression analysis simplified. J Pract Cardiovasc Sci 2020;6:116-22.

multicollinearity in the data and with T < 0.01 there certainly is

ii. Correlation matrix – When computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1[7]

iii. Variance Inflation Factor (VIF) – The VIF of the linear regression is defined as VIF = 1/T. With VIF >5, there is an indication that multicollinearity may be present; with VIF >10, there is certainly multicollinearity among the variables.[3] The simplest way to address the problem is to remove independent variables with high VIF values.

Homoscedasticity

The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line). The Goldfeld–Quandt, Breush–Pagan, Park and White's tests can also be used to test for heteroscedasticity.[8]

How to Interpret the F-statistic

The F-statistic is calculated as regression MS/residual MS. This statistic indicates whether the regression model provides a better fit to the data than a model that contains no independent variables. In essence, it tests if the regression model as a whole is useful. If the $P <$ the significance level, there is sufficient evidence to conclude that the regression model fits the data better than the model with no predictor variables. This finding is good because it means that the predictor variables in the model actually improve the fit of the model. In general, if none of the predictor variables in the model are statistically significant, the overall F statistic is also not statistically significant.

Illustrative Examples on Determining F-test of Overall Significance in Regression Analysis

This tutorial walks through examples of a regression analysis using two methods (manual and online calculator) providing an in-depth explanation of how to read and interpret the output of a regression table.

Example 1

In estimating output (Y) of physiotherapist from a knowledge of his/her test score on the aptitude test (X_1) and years of experience (X_2) in a hospital, the Table 1 summarizes the findings of the study.

$H_0: Y = b_0$

$H_1: Y = b_0 + b_1X_1 + b_2X_2$

Table 1: Test scores, experience, and output of physiotherapist

X_1	X_2	Y
160	5.5	32
80	6.0	15
112	9.5	30
185	5.0	34
152	8.0	35
90	3.0	10
170	9.0	39
140	5.0	26
115	0.5	11
150	1.5	23

Test the following hypotheses at α=0.05

Table 2: Obtaining regression equation

Y	X_1	X_2	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
32	160	5.5	5120	176	880	25600	30.25
15	80	6.0	1200	90	480	6400	36
30	112	9.5	3360	285	1064	12544	90.25
34	185	5.0	6290	170	925	34225	25
35	152	8.0	5320	280	1216	23104	64
10	90	3.0	900	30	270	8100	9
39	170	9.0	6630	351	1530	28900	81
26	140	5.0	3640	130	700	19600	25
11	115	0.5	1265	5.5	57.5	13225	0.25
23	150	1.5	3450	34.5	225	22500	2.25
255	1354	53	37175	1552	7347.5	194128	363

Table 3: Calculation of total, explained, and unexplained variation

Y	X_1	X_2	Y_c	$(Y-\bar{Y})^2$	$Y-Y_c$	$(Y-Y_c)^2$	$(Y_c-\bar{Y})^2$	Std residual
32	160	5.5	31.119	42.25	0.881	0.776	31.575	0.780
15	80	6.0	15.146	110.25	−0.146	0.022	107.214	−0129
30	112	9.5	28.933	20.25	1.067	1.138	11.786	0.945
34	185	5.0	35.424	72.25	−1.424	2.027	98.479	−1.260
35	152	8.0	34.421	90.25	0.579	0.336	79.576	0.513
10	90	3.0	11.269	240.25	−1.269	1.610	202.526	−1.123
39	170	9.0	40.239	182.25	−1.239	1.536	217.238	−1.097
26	140	5.0	25.876	0.25	0.123	0.0153	0.141	0.110
11	115	0.5	11.574	210.25	−0.574	0.330	193.922	−0.509
23	150	1.5	21.000	6.25	2.000	4.001	20.253	1.771
255	1354	53		974.5		11.791	962.710	

Y_c: Predicted Y, $Y-Y_c$: Residual

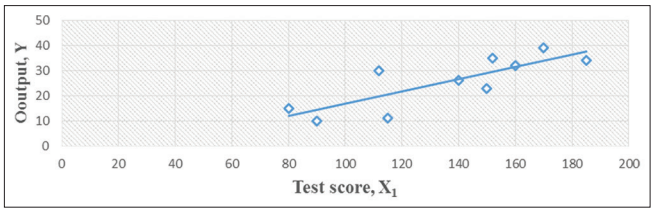

Figure 1: Output against test score.

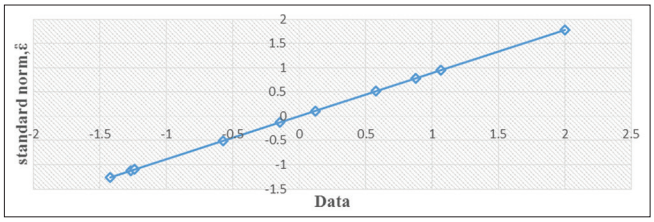

Figure 3: Residuals: QQ plot.

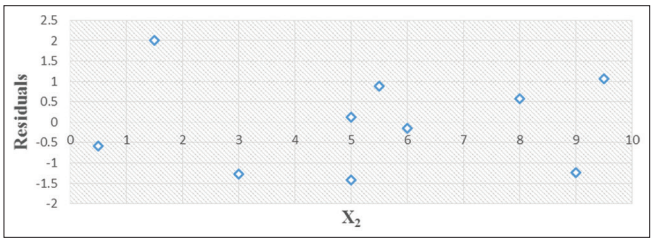

Figure 5: X2 residuals plot.

Manual Computation of F-Test of Overall Significance in Regression Analysis

Obtaining the regression equation

The given data are reproduced in Table 2. Table 2 also shows other inputs required for obtaining the regression equation.

$$\bar{X}_1 = \frac{\sum x_1}{n} = \frac{255}{10} = 25.5, \bar{X}_2 = \frac{\sum x_2}{n} = \frac{1354}{10}$$

$$= 135.4, \bar{Y}_1 = \frac{\sum Y}{n} = \frac{53}{10} = 5.3$$

The general form of multiple equation applicable in this case is:

$$Y = b_0 + b_1X_1 + b_2X_2$$

Moreover, the required normal equations to find the values of b_0, b_1, and b_2 can be written as under:

$$\sum Y = n\, b_0 + b_1 \sum X_1 + b_2 \sum X_2 \quad (1)$$

$$\sum X_1 Y = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 \quad (2)$$

$$\sum Y X_2 = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 \quad (3)$$

Accordingly, the three equations are:

$$255 = 10\, b_0 + 1354\, b_1 + 53\, b_2$$

$$37175 = 1354\, b_0 + 194128\, b_1 + 7374.5\, b_2$$

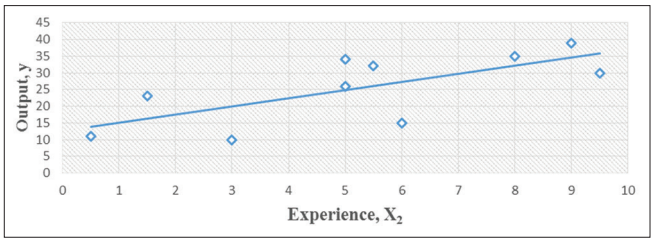

Figure 2: Output against experience.

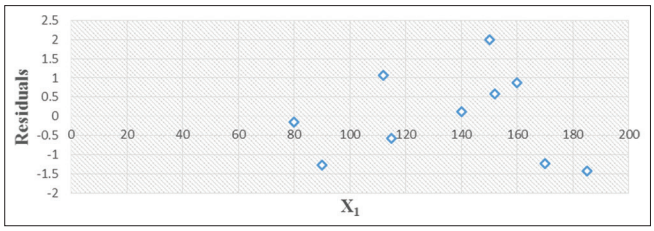

Figure 4: X1 residuals plot.

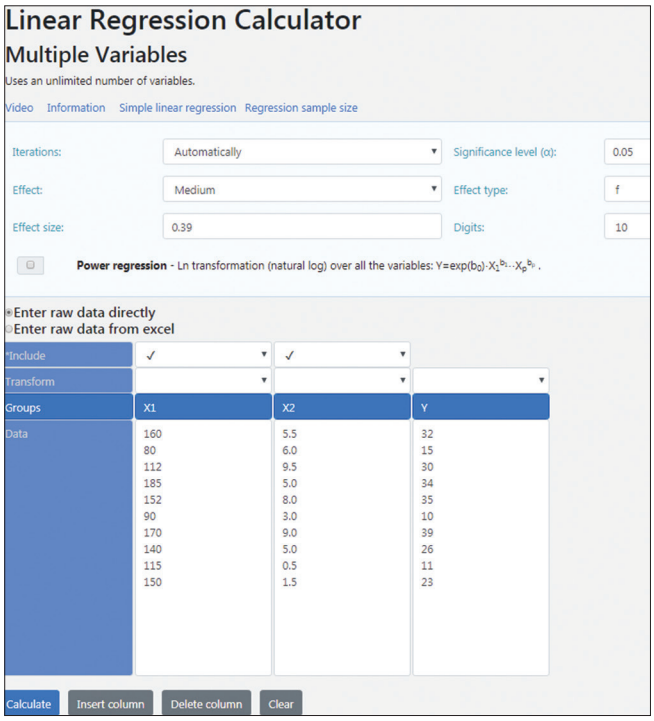

Figure 6: Setting up the data in the table of an online calculator.

$$1552 = 53\, b_0 + 7347.5\, b_1 + 363\, b_2$$

Solving the three equations simultaneously, we obtain $b_0 = -13.824567$, $b_1 = 0.212167$, and $b_2 = 1.999461$. Thus, the regression equation of Y on X_1 and X_2 is: $Y_C = -13.824567 + 0.212167\, X_2 + 1.999461\, X_2$.

Calculation of R and F-ratios

To determine the R and F statistic, we need to calculate total, explained and unexplained variation as shown in Table 3.

$$\text{Total variation (sum of squares total, SST)} = \sum (Y - \bar{Y})^2 = 974.5$$

Explained variation (sum of square regression, SSR) $= \sum (Y_c - \bar{y})^2 = 962.710$

Unexplained variation (sum of squares error, SSE) $= \sum (Y - \bar{Y_c})^2 = 11.791$

Table 4: Excerpts from significance points of the variance-ratio "F"

n_2	$P=0.05$ n_1				
	1	2	3	4	6
6	5.99	5.14	4.76	4.53	4.28
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48

Table 5: Correlation matrix

	Y	X_1	X_2
Y	1.000	0.814	0.709
X_1	0.814	1.000	0.181
X_2	0.709	0.181	1.000

$R\text{square}(R^2) = SSR/SST = 962.710/974.5 = 0.988,\ R = 0.984$

$\text{Mean square regression}\ (MS_R) = SSR/df = 962.710/2 = 481.355$

$\text{Mean square error}\ (MS_E) = SSE/df = 11.791/7 = 1.684$

$F = MS_R/MS_E = 481.355/1.684 = 285.775$

Goodness of fit

The F table value [Table 4] corresponding with degree of freedom $n_1 = 2$ and $n_2 = 7$ is 4.74. Since 285.775>4.74, we ignore the null hypothesis and conclude that $Y \neq b_0$ or $Y = b_0 + b_1X_1 + b_2X_2$.

Validity checking

a. Linearity: The relationship between the Y and X_1 variables is linear [Figure 1] as well as the relationship between the Y and X_2 variables [Figure 2]

b. Normality
QQ-Plot illustrates [Figure 3] that all variables to be multivariate normal

c. Multicollinearity
Tolerance = $1\text{-}R^2$ = 1-0.987902 = 0.012098. With 0.012098> 0.01 but 0.012098 <0.1, there might be

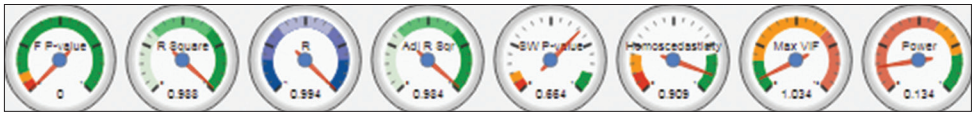

Figure 7: Regression statistics.

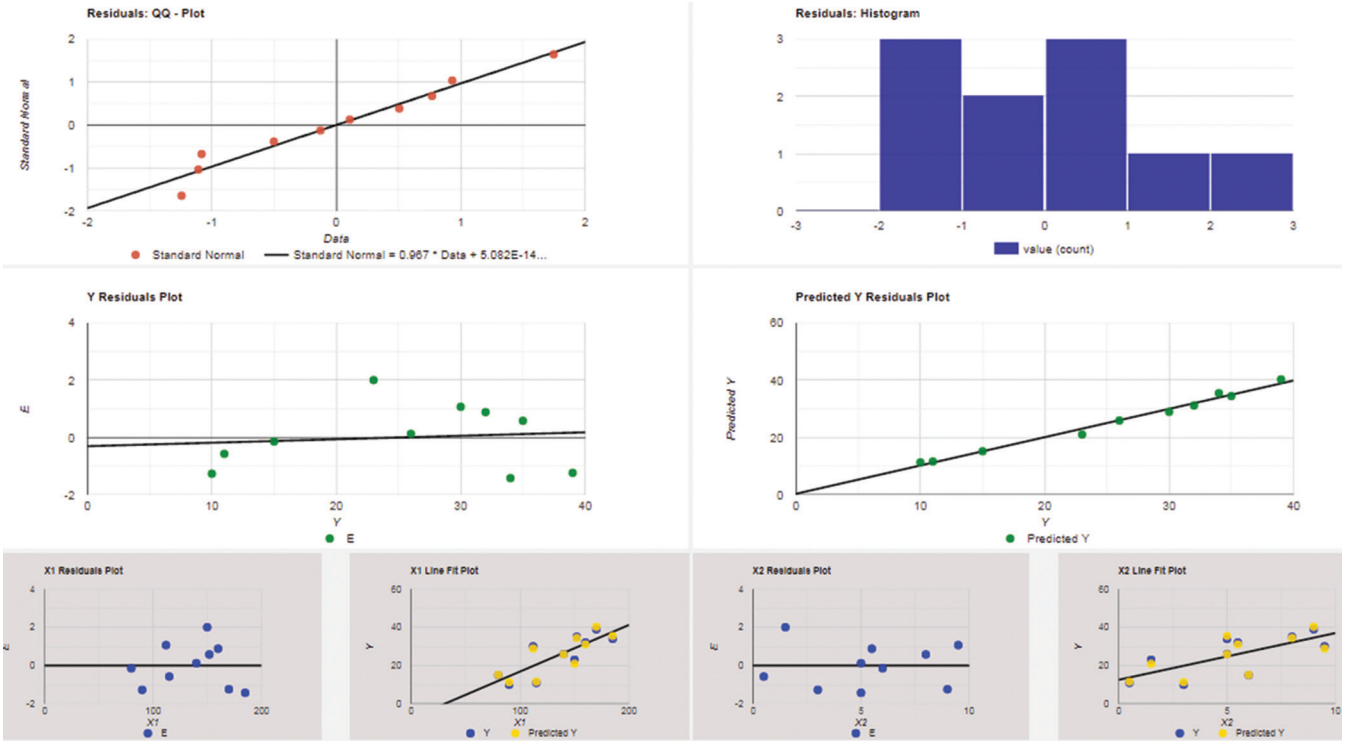

Figure 8: Residual plots.

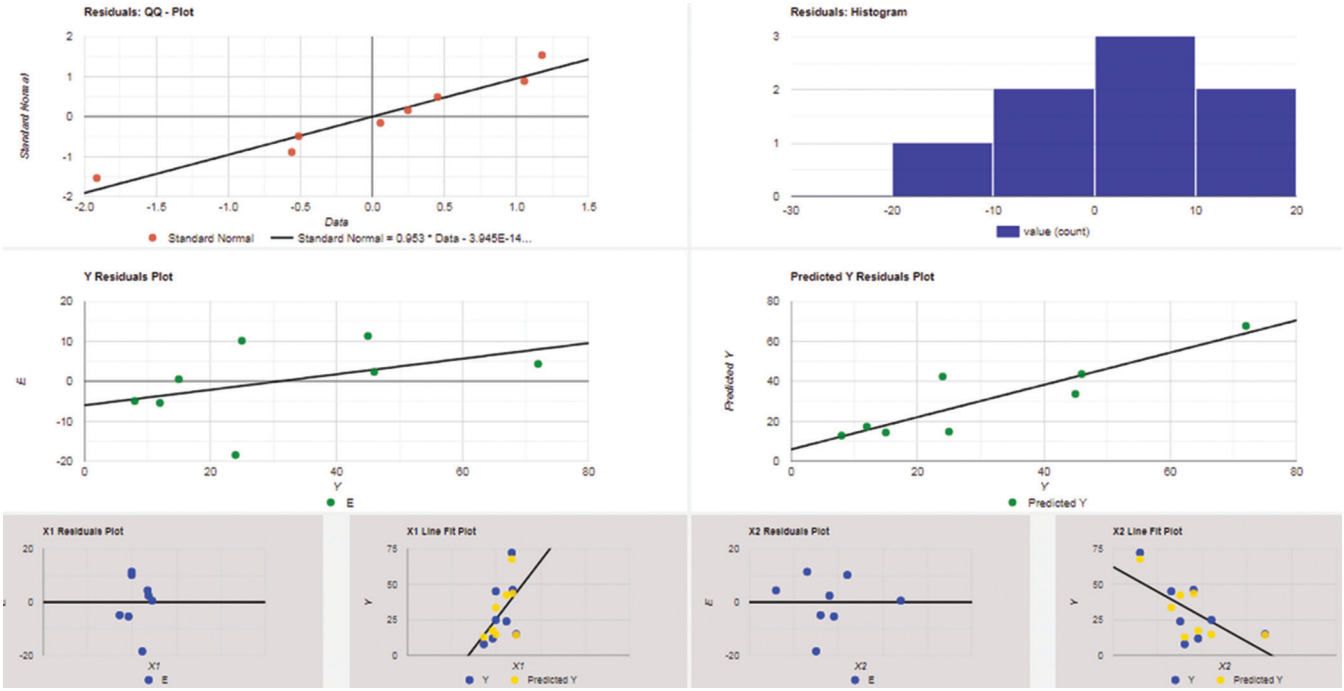

Figure 9: Residual plots.

multicollinearity in the data.

d. Homoscedasticity-homogeneity of variance.

The data are homoscedastic since the residuals are equal across the regression line [Figures 4 and 5].

Using an Online Linear Regression Analysis Calculator (SIMPLIFIED Method)

The F-test of overall significance in regression analysis can be done through online calculators which are easily available in internet. For use friendly online calculator, you may visit this uniform locator http://www.statskingdom.com/410 multi_linear_regression.htm.

In the software, it is really easy to conduct an F-test and most of the assumptions are preloaded. The calculator uses variables transformations, calculates the Linear equation, R, *P* value, outliers and the adjusted Fisher-Pearson coefficient of skewness. After checking the residuals' normality, multicollinearity, homoscedasticity, and priori power, the program interprets the results. Then, it draws a histogram, a residuals QQ-plot, a correlation matrix, a residuals x-plot and a distribution chart. You may transform the variables exclude any predictor or run backward stepwise selection automatically based on the predictor's *P* value.

The basic step for using and online calculator is to correctly fill in you data into it [Figure 6]. For instance, in the above example, we have to fill in the data in the columns of an online calculator. Click the calculate button.

Figure 10: Regression statistics.

Summary Output

The output of the F-test is summarized below by the regression equation, regression statistics [Figure 7], correlation matrix [Table 5], ANOVA [Table 6], coefficient table iteration I [Table 7], and residual graphs [Figure 8].

Regression equation is $Y = -13.825 + 0.212\ X_1 + 1.999\ X_2$

Validity Checking

a. Residual Normality: Linear regression assumes normality for residual errors. Shapiro–Wilk $P = 0.664$ [Figure 7]. It is assumed that the data are normally distributed

b. Homoscedasticity-Homogeneity of Variance: The White test $P = 0.909$ [Figure 7]. It is assumed that the variance is homogeneous

c. Multicollinearity-Intercorrelations among the Predictors: There is no multicollinearity concern as all the VIF values are smaller than 2.5 [Tables 6 and 7]

d. Priori power-of the entire model (2 predictors): The priori power should be calculated before running the regression. Although the power is low: 0.134 [Figure 7], we reject the H_0.

Table 6: ANOVA table

Source	DF	SS	MS	*F* statistic	*P*
Regression (between $\hat{y}_i$ and y_i)	2	962.710	481.355	285.802	1.95e-7
Residual (between y_i and $\hat{y}_i$)	7	11.790	1.684		
Total (between y_i and y_i)	9	974.500	108.278		

SS: Sum of squares, Df: Degrees of freedom, MS: Mean square

Table 7: Coefficient Table Iteration 1 (adjusted R^2=0.984)

	Coefficient	SE	*t*-statistic	Lower $t_{0.025(7)}$	Upper $t_{0.975(7)}$	Stand coefficient	*P*	VIF
B	−13.825	1.795	−7.701	−18.069	−9.580	0.00	0.000116	
X_1	0.212	0.0127	16.759	0.182	0.242	0.708	6.59e-7	1.034
X_2	1.999	0.146	13.728	1.655	2.344	0.580	0.00000257	1.034

VIF: Variance inflation factor, SE: Standard error

Table 8: Data on use of hypotensive drugs

Y	X_1	X_2
2.45	84	15
1.72	66	8
2.37	68	46
2.23	65	24
1.92	69	12
1.99	72	25
1.99	63	45
2.35	56	72

Test the following hypotheses at α=0.05

Table 9: Correlation matrix

	Y	X_1	X_2
Y	1.000	0.483	−0.645
X_1	0.483	1.000	0.200
X_2	−0.645	0.200	1.000

Interpretation of the Output

Y and X relationship

R square (R^2) equals 0.988. It means that the predictors (Xi) explain 98.8% of the variance of Y. Adjusted R square equals 0.984. The coefficient of multiple correlation® equals 0.994. It means that there is a very strong direct relationship between the predicted data (ŷ) and the observed data (y).

Goodness of fit

Right-tailed F test is used to check if the entire regression model is statistically significant. From Table 6, F (1, 7) = 285.802, P = 1.94764e-7. Since $P < \alpha$ (0.05), we reject the H_0. The linear regression model, $Y = b_0 + b_1X_1 + b_2X_2$, provides a better fit than the model without the independent variables resulting in, $Y = b_0$.

As shown in Table 5, P value for X_1 = 6.59e-7 and for X_2 = 0.00000257. All the independent variables (Xi) are significant since $P < \alpha$ (0.05). The Y-intercept (b): Two-tailed, T = −7.701131, P = 0.000116139 [Table 7]. Hence, b is significantly different from zero.

Example 2

The data in Table 8 are taken from a clinical trial to compare two hypotensive drugs used to lower the blood pressure during operations. The dependent variable, y, is the recovery time (in minutes) elapsing between the time at which the drug was discontinued and the time at which the systolic blood pressure had returned to 100 mmHg. The two predictors are quantity of drugs used in mg (x_1) and mean level of systolic blood pressure during hypotension in mmHg (x_2).

H_0: $Y = b_0$

H_1: $Y = b_0 + b_1X_1 + b_2X_2$

Using an Online Linear Regression Analysis Calculator (Simplified Method)

To analyze the relationship between quantity of drugs used and mean level of systolic blood pressure during hypotension, we run a multiple linear regression using quantity of drugs used and mean level of systolic blood pressure during hypotension taken as the predictor variables and recovery time as the response variable. The output of the F-test is summarized below by the regression equation, residual plots [Figure 9], correlation matrix [Table 9], ANOVA [Table 10] coefficient table iteration I [Table 11], and Regression statistics [Figure 10].

Regression equation is predicted $Y = 58.603 + 53.688\ X_1 - 2.091\ X_2$.

Validity Checking

a. Residual Normality: Linear regression assumes normality for residual errors. Shapiro–Wilk P = 0.638 [Figure 10]. It is assumed that the data are normally distributed
b. Homoscedasticity-Homogeneity of Variance: The White test P value [Figure 10] equals 0.567 (F = 0.637). It is assumed that the variance is homogeneous
c. ©Multicollinearity-Intercorrelations among the Predictors. There is no multicollinearity concern as all the VIF values are smaller than 2.5 [Table 11]

Table 10: ANOVA table

Source	DF	SS	MS	*F* statistic	*P*
Regression (between $\hat{y}_i$ and y_i)	2	2686.086	1343.043	10.382	0.0166
Residual (between y_i and $\hat{y}_i$)	5	646.789	129.358		
Total (between y_i and y_i)	7	3332.875	476.125		

SS: Sum of squares, Df: Degrees of freedom, MS: Mean square

Table 11: Coefficient table iteration 1 (adjusted R^2=0.728)

	Coefficient	SE	*t*-statistic	Lower $t_{0.025\ (5)}$	Upper $t_{0.975\ (5)}$	Stand coefficient	*P*	VIF
b	58.603	46.328	1.265	−60.488	177.695	0.00	0.262	
X_1	53.688	16.938	3.170	10.148	97.227	0.637	0.0248	1.042
X_2	−2.091	0.544	−3.842	−3.491	−0.692	−0.773	0.0121	1.042

SS: Sum of squares, Df: Degrees of freedom, MS: Mean square, VIF: Variance inflation factor, SE: Standard error

d. Priori power-of the Entire Model (2 Predictors): Although the power is low: 0.106 [Figure 10], we reject the $H_{0.}$

The power to prove each predictor significance is always lower than the power of the entire model.

Interpretation of the Output

Y and X relationship

R square (R^2) equals 0.806**.** It means that the predictors (X_i) explain 80.6% of the variance of Y. Adjusted R square equals 0.728. The coefficient of multiple correlation ® equals 0.898. It means that there is a very strong direct relationship between the predicted data (ŷ) and the observed data (y).

Goodness of fit

Right-tailed F test is used to check if the entire regression model is statistically significant. From Table 10, $F_{(1,\ 5)}$ = 10.382, P = 0.0166. Since $P < \alpha$ (0.05), we reject the H_0.The linear regression model, $Y = b_0 + b_1X_1 + b_2X_2$, provides a better fit than the model without the independent variables resulting in, $Y = b_0$.

As shown in Table 11, P value for X_1 = 0.0248 and for X_2 = 0.0121. All the independent variables (Xi) are significant since P values< α (0.05). The Y-intercept (b): Two-tailed, T = 1.265, P = 0.262 [Table 11]. Hence, b is not significantly different from zero. It is still most likely recommended not to force b to be zero.

What Does an F-test of Overall Significance Test Tell and what it Does Not

The F statistic represents the ratio of the variance explained by the regression model (regression mean square) to the not explained variance (residuals mean square). It can be calculated easily using an online calculator in comparison to the manual approach. The F-test of overall significance tests whether all of the predictor variables are jointly significant while the *t*-test of significance for each individual predictor variable merely tests whether each predictor variable is individually significant. Thus, the F-test determines whether or not all of the predictor variables are jointly significant. It is possible that each predictor variable is not significant and yet the F-test says that all of the predictor variables combined are jointly significant.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Kothari CN. Quantitative Techniques. 3rd ed. New Delhi: UBS Publishers' DistributorsPut LTD; 2007.
2. Vohra ND. Quantitative Techniques in Management. 3rd ed. New Delhi: Tata McGraw-Hill Publishing Company Limited; 2007.
3. Armitage P, Berry G, Mathews, JN. Statistical Methods in Medical Research. 4th ed. Massachusetts: Blackwell Science; 2002.
4. Sullivan LS. Essentials of Biostatistics Workbook. 2nd ed. London: Jones and Bartlett Learning; 2003.
5. Ogunleye LI, Oyejola BA, Obisesan KO. Comparison of some common tests for normality. Int J Probabil Statist 2018;7:5, 130-7.
6. Whetherill GB, Duncombe P, Kenward M, Kollerstrom J, Paul SR, Vowden BJ, *et al*. Regression Analysis with Applications. London: Chapman and Hall; 1986.
7. Harris M, Taylor G. Medical Statistics Made Easy. New Yolk: Springer-Verlag; 2003.
8. Su H, Berenson ML. Comparing tests of homoscedasticity in simple linear regression. JSM Math Stat 2017;4:1017.