ISSN:3030-332X
Impact factor: 9,493
International journal of scientific researchers

Multiple Regression Analysis Problem

Barakayeva Mavjuda
Teacher of the Department of Mathematics at National Pedagogical University of Uzbekistan Named after Nizami
E-mail: barakayevamavjuda0106@gmail.com
doi: https://doi.org/10.5281/zenodo.20570532

Abstract: This article investigates the application of multiple regression analysis to examine the relationship between milk sales volume and price under different operating conditions. A simple regression model is first estimated to evaluate the impact of price on sales volume using the ordinary least squares method. The statistical adequacy of the model is assessed through the coefficient of determination, t-statistics, and Fisher's F-test. To account for the effect of a strike period, a dummy variable is incorporated into the regression model. The results show that the inclusion of the dummy variable substantially improves the explanatory power and statistical significance of the model. The findings confirm the significant influence of the strike factor on milk sales and demonstrate the importance of incorporating qualitative variables into econometric models.
Keywords: multiple regression analysis, dummy variables, ordinary least squares, coefficient of determination, Fisher's F-test, econometric modeling.

Considering various aspects of multiple regression analysis, we have identified several problems that frequently arise in its practical application. It should be noted that regression models are sufficiently flexible tools that make it possible, among other things, to assess the influence of qualitative characteristics (such as occupation, gender, presence of children, etc.) on the variable under study. This is achieved by introducing into the set of regressors the so-called dummy variables, which usually take the values of 1 or 0 depending on the presence or absence of the corresponding characteristic in a particular observation. From a formal point of view, dummy variables do not differ from other regressors. However, special attention should be paid to their proper use and to the correct interpretation of their estimated coefficients.

Example. A certain company is engaged in the sale of milk. The table presents the monthly sales volumes Q (thousand liters) at different prices P (UZS per liter). During the fifth, sixth, and seventh months, a strike occurred at one of the company's enterprises.

Month	**Q**	**P**	**Month**	**Q**	**P**
1	98	10,0	8	113	13,0
2	100	11,0	9	116	13,0
3	103	12,5	10	118	13,8
4	105	12,5	11	121	14,2
5	80	14,6	12	123	14,4
6	87	14,6	13	126	15,0
7	94	14,9	14	128	16,1

Using regressions of Q on P, determine:

a) whether there was a shift in the intercept (constant term) during the strike compared with the normal operating period;

b) whether there was a shift in both the intercept and the slope coefficient for P.

Solution. Let us estimate the parameters of the multiple regression model. We determine the vector of estimated regression coefficients. According to the ordinary least squares method, the vector Y(X) is obtained from the following expression:

$\widehat{\beta}_{OLS} = Y(X) = (X^TX)^{-1} \cdot X^TY$

where: X is the price (UZS per liter); Y is the sales volume (thousand liters).

№	X	Y	Z
1	10,0	98	0
2	11,0	100	0
3	12,5	103	0
4	12,5	105	0
5	14,6	80	1
6	14,6	87	1
7	14,9	94	1
8	13,0	113	0
9	13,0	116	0
10	13,8	118	0
11	14,2	121	0
12	14,4	123	0
13	15,0	126	0
14	16,1	128	0

A unit (constant) column is appended to the matrix of explanatory variables X_j:

Matrix X = $\begin{pmatrix} 1 & 10 \\ 1 & 11 \\ 1 & 12,5 \\ 1 & 12,5 \\ 1 & 14,6 \\ 1 & 14,6 \\ 1 & 14,9 \\ 1 & 13 \\ 1 & 13 \\ 1 & 13,8 \\ 1 & 14,2 \\ 1 & 14,4 \\ 1 & 15 \\ 1 & 16.1 \end{pmatrix}$; Matrix Y = $\begin{pmatrix} 98 \\ 100 \\ 103 \\ 105 \\ 80 \\ 87 \\ 94 \\ 113 \\ 116 \\ 118 \\ 121 \\ 123 \\ 126 \\ 128 \end{pmatrix}$

Matrix X^T = $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 11 & 12,5 & 12,5 & 14,6 & 14,6 & 14,9 & 13 & 13 & 13,8 & 14,2 & 14,4 & 15 & 16,1 \end{pmatrix}$

Multiplying the matrices, we obtain: $(X^TX) = \begin{pmatrix} 14 & 189,6 \\ 189,6 & 2603,48 \end{pmatrix}$

Find the inverse of the matrix $(X^TX)^{-1}$: $(X^TX)^{-1} = \begin{pmatrix} 5,201 & -0,379 \\ -0,379 & 0,028 \end{pmatrix}$

Multiplying the matrices, we obtain: $(X^TY) = \begin{pmatrix} 1512 \\ 20564,4 \end{pmatrix}$

The estimated regression coefficient vector is given by:

$Y(X) = \begin{pmatrix} 5,201 & -0,379 \\ -0,379 & 0,028 \end{pmatrix} \cdot \begin{pmatrix} 1512 \\ 20564,4 \end{pmatrix} = \begin{pmatrix} 74,819 \\ 2,45 \end{pmatrix}$

The estimated regression equation is:

$$\mathbf{Y(X) = 74,8192 + 2,4501 \cdot X} \qquad \mathbf{(1)}$$

To obtain an unbiased estimate of the variance, we perform the following calculations:

Unbiased error (residual): ε = Y–Y(x), where Y(x) represents the estimated values obtained from the regression model.

J	X	Y	Y(x)	ε = Y - Y(x)	ε^2	(Y-Ycp)²	|ε : Y|
1	10,0	98	99,32	-1,32	1,742	100	0,0135
2	11,0	100	101,77	-1,77	3,132	64	0,0177
3	12,5	103	105,445	-2,445	5,978	25	0,0237
4	12,5	105	105,445	-0,445	0,198	9	0,00424
5	14,6	80	110,59	-30,59	935,752	784	0,382
6	14,6	87	110,59	-23,59	556,491	441	0,271
7	14,9	94	111,325	-17,325	300,158	196	0,184
8	13,0	113	106,67	6,33	40,069	25	0,056
9	13,0	116	106,67	9,33	87,049	64	0,0804
10	13,8	118	108,63	9,37	87,797	100	0,0794

1	14,2	121	109,61	11,39	129,731	169	0,0941
2	14,4	123	110,1	12,9	166,409	225	0,105
3	15,0	126	111,57	14,43	208,223	324	0,115
4	16,1	128	114,265	13,735	188,646	400	0,107
Σ	**189,6**	**1512**			**2711,375**	**2926**	**1,534**
Σ/n	**13,5429**	**108**					

Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{\sum |\varepsilon : Y|}{n} \cdot 100\% = \frac{1{,}534}{14} \cdot 100\% \approx 10{,}95\%$$

The variance estimate is: $s_e^2=(Y-Y(X))^T(Y-Y(X))=2711{,}375$

The unbiased estimate of the variance is:

$$s^2 = \frac{1}{n-m-1} \cdot s_e^2 = \frac{1}{14-1-1} \cdot 2711{,}375 = 225{,}9479$$

The estimate of the standard deviation (standard error of the estimate of Y) is:

$$S = \sqrt{s^2} = \sqrt{225{,}9479} \approx 15{,}032$$

Next, we determine the estimated covariance matrix of the coefficient vector:

$\widehat{V}(\widehat{\beta}_{OLS}) = S^2 \bullet (X^TX)^{-1}$:

$$\widehat{V}(\widehat{\beta}_{OLS}) = 225{,}9479 \cdot \begin{pmatrix} 5{,}201 & -0{,}379 \\ -0{,}379 & 0{,}028 \end{pmatrix} = \begin{pmatrix} 1175{,}186 & -85{,}584 \\ -85{,}584 & 6{,}319 \end{pmatrix}$$

The variances of the model parameters are determined by the relation

$S^2_i = K_{ii}$,

that is, by the elements located on the main diagonal:

$$S_{b0} = \sqrt{1175{,}186} \approx 34{,}281$$

$$S_{b1} = \sqrt{6{,}319} \approx 2{,}514$$

The degree of the combined influence of the explanatory variables on the dependent variable is assessed by the multiple correlation coefficient, which is calculated as follows:

$$R = \sqrt{1 - \frac{s_e^2}{\sum (y_i - \bar{y})^2}} = \sqrt{1 - \frac{2711{,}375}{2926}} = \sqrt{0{,}07335} \approx 0{,}27083$$

The relationship between the dependent variable Y and the explanatory variable(s) X is weak.

Coefficient of determination:

$R^2= 0{,}27083^2 = 0{,}07335$

The quantity $v = n - m - 1$ is referred to as the number of degrees of freedom. When estimating a multiple linear regression model, it is generally accepted that, to ensure statistical reliability, the number of observations should be at least three times greater than the number of parameters being estimated.

t-statistic:

T_{crit} (n-m-1;α/2) = (12 ; 0,025) = 2,179

$$t_i = \frac{b_i}{S_{bi}}$$

$$|t_0| = \left|\frac{74{,}819}{34{,}281}\right| \approx 2{,}183 > 2{,}179$$

The regression coefficient b_0 is statistically significant.

$$|t_1| = \left|\frac{2{,}45}{2{,}514}\right| \approx 0{,}975 < 2{,}179$$

The statistical significance of the regression coefficient b_1 is not confirmed.

F-statistic (Fisher's F-test)

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = \frac{0{,}07335}{1-\ 0{,}07335} \cdot \frac{14-1-1}{1} \approx 0{,}95$$

The critical (tabulated) value for the degrees of freedom

$k_1=1$, $k_2=n–m–1=14–1–1=12$,

$F_{crit}\ (1;12) = 4{,}75$

Since the calculated value of the F-statistic is less than its critical value

$F_{obs} < F_{crit}$

the coefficient of determination is not statistically significant, and the regression model cannot be considered statistically reliable.

An increase in the price of milk by 1 UZS per liter leads, on average, to an increase in sales volume by 2.45 thousand liters.

The statistical significance of the regression equation was tested using the coefficient of determination and Fisher's F-test. It was found that, in the situation under study, 7.34% of the total variability of Y is explained by changes in the explanatory variable(s) X_j.

It was also established that the model parameters are not statistically significant.

The obtained regression equation does not take into account the influence of the factor "strike."

To account for this effect, a dummy (binary) variable $\mathbf{Z_1}$, is introduced into the regression model, where

$$z_{i1} = \begin{cases} 1, & if\ a\ strike\ occurred, \\ 0, & under\ normal\ operating\ conditions. \end{cases}$$

By appending the column vector Z to the matrix of explanatory variables X, we obtain the augmented design matrix X*

Matrix $X^{*T} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 11 & 12{,}5 & 12{,}5 & 14{,}6 & 14{,}6 & 14{,}9 & 13 & 13 & 13{,}8 & 14{,}2 & 14{,}4 & 15 & 16{,}1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Multiplying the matrices, we obtain: $(X^{*T} \cdot X^*) = \begin{pmatrix} 14 & 189{,}6 & 3 \\ 189{,}6 & 2603{,}48 & 44{,}1 \\ 3 & 44{,}1 & 3 \end{pmatrix}$

Next, we find the inverse of the matrix: $(X^{*T} \cdot X^*)^{-1}$:

$(X^{*T} \cdot X^*)^{-1} = \begin{pmatrix} 5{,}801 & -0{,}432 & 0{,}545 \\ -0{,}432 & 0{,}028 & -0{,}0481 \\ 0{,}545 & -0{,}0481 & 0{,}495 \end{pmatrix}$

Multiplying the matrices, we obtain: $(X^{*T} \cdot Y) = \begin{pmatrix} 1512 \\ 20564{,}4 \\ 261 \end{pmatrix}$

The estimated regression coefficient vector is given by:

$$Y(X,Z) = \begin{pmatrix} 5{,}801 & -0{,}432 & 0{,}545 \\ -0{,}432 & 0{,}028 & -0{,}0481 \\ 0{,}545 & -0{,}0481 & 0{,}495 \end{pmatrix} \cdot \begin{pmatrix} 1512 \\ 20564{,}4 \\ 261 \end{pmatrix} = \begin{pmatrix} 35{,}861 \\ 5{,}887 \\ -35{,}397 \end{pmatrix}$$

The estimated regression equation is:

$$\mathbf{Y(X,Z)=35{,}8612+5{,}8868 \cdot X-35{,}3969 \cdot Z} \qquad \mathbf{(2)}$$

To obtain an unbiased estimate of the variance, we perform the following calculations.

Residual (unbiased error): $\varepsilon = Y–Y(x)$, where Y(x) denotes the values estimated from the regression model.

J	X	Z	Y	Y(X,Z)	ε = Y - Y(x,z)	ε^2	(Y-Ycp)²	|ε : Y|
1	10,0	0	98	94,729	3,271	10,699	100	0,033
2	11,0	0	100	100,616	-0,616	0,379	64	0,006

3	12,5	0	103	109,446	-6,446	41,551	25	0,063
4	12,5	0	105	109,446	-4,446	19,767	9	0,042
5	14,6	1	80	86,4113	-6,411	41,105	784	0,080
6	14,6	1	87	86,4113	0,589	0,347	441	0,007
7	14,9	1	94	88,1774	5,823	33,903	196	0,062
8	13,0	0	113	112,389	0,611	0,373	25	0,005
9	13,0	0	116	112,389	3,611	13,037	64	0,031
10	13,8	0	118	117,099	0,901	0,812	100	0,008
11	14,2	0	121	119,454	1,546	2,392	169	0,013
12	14,4	0	123	120,631	2,369	5,613	225	0,019
13	15,0	0	126	124,163	1,837	3,375	324	0,015
14	16,1	0	128	130,638	-2,638	6,961	400	0,021
Σ	**189,6**	**3**	**1512**	**1512**		**180,313**	**2926**	**0,405**
Σ/n	**13,5429**		**108**	**108**				

Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{\sum |\varepsilon: Y|}{n} \cdot 100\% = \frac{0,405}{14} \cdot 100\% \approx 2,89\%$$

The estimated variance is given by: $s_e^2=(Y-Y(X))^T(Y-Y(X))=180,313$

The unbiased variance estimate is given by:

$$s^2 = \frac{1}{n-m-1} \cdot s_e^2 = \frac{1}{14-2-1} \cdot 180,313 = 16,3921$$

The estimated standard deviation (standard error of the estimate of Y) is:

$$S = \sqrt{s^2} = \sqrt{16,3921} \approx 4,049$$

Next, we estimate the covariance matrix of the coefficient vector:

$\widehat{V}(\widehat{\beta}_{OLS}) = S^2 \bullet (X^{*T} \cdot X^*)^{-1}$:

$$\widehat{V}(\widehat{\beta}_{OLS}) = 16,3921 \cdot \begin{pmatrix} 5,801 & -0,432 & 0,545 \\ -0,432 & 0,028 & -0,0481 \\ 0,545 & -0,0481 & 0,495 \end{pmatrix} = \begin{pmatrix} 95,087 & -7,076 & 8,931 \\ -7,076 & 0,535 & -0,788 \\ 8,931 & -0,788 & 8,114 \end{pmatrix}$$

The variances of the model parameters are determined by the relation

$S^2_i = K_{ii}$,

that is, they are the elements located on the main diagonal of the covariance matrix:

$$S_{b0} = \sqrt{95,087} \approx 9,751$$

$$S_{b1} = \sqrt{0,535} \approx 0,731$$

$$S_{b2} = \sqrt{8,114} \approx 2,849$$

The degree of the combined effect of the explanatory variables on the dependent variable is measured by the multiple correlation coefficient, which is given by:

$$R = \sqrt{1 - \frac{s_e^2}{\sum (y_i - \bar{y})^2}} = \sqrt{1 - \frac{180,313}{2926}} = \sqrt{0,938} \approx 0,9687$$

A strong relationship exists between the dependent variable Y and the explanatory variable(s) X.

Coefficient of determination: $R^2 = 0,9687^2 = 0,9384$

A more reliable assessment is provided by the adjusted coefficient of determination

$$\overline{R}^2 = 1 - (1 - 0,9384) \cdot \frac{14 - 1}{14 - 2 - 1} \approx 0,927$$

The quantity $v = n - m - 1$ is called the number of degrees of freedom. In multiple linear regression analysis, it is generally accepted that, to ensure statistical reliability, the number of observations should be at least three times greater than the number of estimated parameters.

t-statistic:

T_{crit} (n-m-1;α/2) = (11 ; 0,025) = 2,201

$$t_i = \frac{b_i}{S_{bi}}$$

$$|t_0| = \left|\frac{35,861}{9,751}\right| \approx 3,678 > 2,201$$

The regression coefficient b_0 is statistically significant.

$$|t_1| = \left|\frac{5,887}{0,731}\right| \approx 8,049 > 2,201$$

The regression coefficient b_1 is statistically significant.

$$|t_2| = \left|\frac{-35,397}{2,849}\right| \approx 12,426 > 2,201$$

The regression coefficient b_2 is statistically significant.

F-statistic. Fisher's criterion:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0,9384}{1 - 0,9384} \cdot \frac{14 - 2 - 1}{2} \approx 83,751$$

The critical (tabulated) value for the degrees of freedom

k_1=2, k_2=n–m–1=14–2–1=11, is F_{crit} (2;11) = 3,98

Since the observed value satisfies $F_{obs} > F_{crit}$, the coefficient of determination is statistically significant, and the regression equation is statistically reliable. In other words, the regression coefficients (b_i) are jointly significant.

Thus, the empirical results indicate that the “strike” factor has a significant effect. Hence, there are grounds to conclude that the regression relationship between milk sales volume and price differs significantly under normal operating conditions and during periods of strike activity.

References

1. Wooldridge, J.M. (2020). Introductory Econometrics: A Modern Approach (7th ed.). Boston: Cengage Learning.
2. Gujarati, D.N., & Porter, D.C. (2009). Basic Econometrics (5th ed.). New York: McGraw-Hill Education.
3. Greene, W.H. (2018). Econometric Analysis (8th ed.). New York: Pearson Education.
4. Stock, J.H., & Watson, M.W. (2020). Introduction to Econometrics (4th ed.). New York: Pearson.
5. Montgomery, D.C., Peck, E.A., & Vining, G.G. (2021). Introduction to Linear Regression Analysis (6th ed.). Hoboken, NJ: Wiley.
6. Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models (5th ed.). New York: McGraw-Hill.
7. Draper, N.R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). New York: Wiley.
8. Kennedy, P. (2008). A Guide to Econometrics (6th ed.). Malden, MA: Blackwell Publishing.

9. Berndt, E.R. (2005). The Practice of Econometrics: Classic and Contemporary. Moscow: UNITY-DANA.
10. Cramer, H. (1975). Mathematical Methods of Statistics. Moscow: Mir.
11. Maddala, G.S. (2001). Introduction to Econometrics (3rd ed.). Chichester: John Wiley & Sons.
12. Fukushige, M. (2025). Variable Selection and Variable Integration for Categorical Dummy Variables. SN Operations Research Forum, 6(1).
13. Harismahyanti, A., & Nur'eni, N. (2025). Analyzing the Impact of Inflation, Exports and Unemployment on Economic Growth in Indonesia: A Fixed Effects Least Squares Dummy Variable Panel Regression Approach. Priviet Social Sciences Journal, 5(8), 338–348.
14. Hidayaturrahman, U., Basirun, B., Ayundasari, D.S., & Alayyubi, M.Z. (2025). Application of Dummy Regression to Estimate the Income. Jurnal Statistika, 18(1), 789–797.
15. Bahri, B., Jayadi, M., Triana, D., Fatmawati, R., & Fadiah, I.A. (2025). Analysis of Structural Break Using Dummy Variable Regression on Industrial Stock Returns in Indonesia Stock Exchange. American Journal of Open Research, 4(9), 294–301.
16. Te Grotenhuis, M., & Thijs, P. (2015). Dummy Variables and Their Interactions in Regression Analysis. Sociological Methods & Research.
17. Crown, W.H. (2010). There's a Reason They Call Them Dummy Variables. Health Services Research, 45(5), 1440–1443.
18. Athey, S., & Imbens, G. (2016). The State of Applied Econometrics: Causality and Policy Evaluation. Journal of Economic Perspectives, 31(2), 3–32.
19. Roth, J., Sant'Anna, P.H.C., Bilinski, A., & Poe, J. (2022). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. Journal of Econometrics, 235(2), 2218–2244.
20. Zaman, A., Rousseeuw, P.J., & Orhan, M. (2017). Econometric Applications of High-Breakdown Robust Regression Techniques. Economics Bulletin, 37(3), 1733–1748.