MODULE
ECONOMETRICS 1

Arranged by:

MAHRUS LUTFI ADI KURNIAWAN

ECONOMICS DEVELOPMENT STUDY PROGRAM
FACULTY OF ECONOMICS AND BUSINESS
AHMAD DAHLAN UNIVERSITY

2021/2022

INTRODUCTION

The preparation of the Econometrics 1 module is basically to help with the literature needs of students of the Faculty of Economics and Business. Besides being intended to help students understand Econometrics 1, this module can be used to study other courses related to economics. For this reason, this module explains various materials about what regression analysis, polynomial equations, log-linear models, variable indicators, and difference estimators is.

Hopefully this module can provide broader knowledge to the reader. Although this module has many drawbacks. The author needs constructive criticism and suggestions. Thank You.

Yogyakarta, March 2022

The Writer

CONTENTS

A. Polynomial Equations

A polynomial equation is an equation where a polynomial is set equal to zero. i.e., it is an equation formed with variables, non-negative integer exponents, and coefficients together with operations and an equal sign. It has different exponents.

Formula:

$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$

Coefficient β_3 is constant, does not depend on the level ADVERT, does not capture diminishing returns in reality.

Solution:

$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$

$$\left.\frac{\Delta E(SALES)}{\Delta ADVERT}\right| (PRICE\text{ held constant}) = \frac{\partial E(SALES)}{\partial ADVERT} = \beta_3 + 2\beta_4 ADVERT$$

The number implies the marginal effect of advertising on sales.

Example:

$$\widehat{SALES} = 109.72 - 7.640\, PRICE + 12.151\, ADVERT - 2.768\, ADVERT^2$$

(se) (6.80) (1.046) (3.556) (0.941)

And the marginal effect of advertising on sales will be:

$$\widehat{\frac{\Delta E(SALES)}{\Delta ADVERT}} = 12.151 - 5.536\, ADVERT$$

How does it imply? When ADVERT = 0.5, the marginal effect of advertising on sales is 9.38. When ADVERT = 2, the marginal effect is 1.08.

Interaction variables:

We wish to study the effect of income and age on an individual's expenditure on pizza take a random sample of 40 individuals, age 18 and older, and record their annual expenditure on pizza.

$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + e$

And the estimation will be:

$$\widehat{PIZZA} = 342.88 - 7.576\, AGE + 1.832\, INCOME$$

Implication: regardless of age, an increase in income should lead to an increase in pizza spending. However, as a person gets older, the marginal propensity to buy pizza decreases so income is age dependent.

Effect of one variable is modified by another → interaction variable, example:

$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4 (AGE \times INCOME) + e$

The estimation model:

$\widehat{PIZZA} = 161.47 - 2.977 AGE + 6.980 INCOME - 0.1232 (AGE \times INCOME)$

Hence, marginal effect of age upon pizza expenditure:

$$\frac{\partial \widehat{E(PIZZA)}}{\partial AGE} = b_2 + b_4 INCOME$$

$$= -2.977 - 0.1232\ INCOME$$

$$= \begin{cases} -6.06 \ for\ INCOME = 25 \\ -14.07 \ for\ INCOME = 90 \end{cases}$$

So, individual with \$25k income will reduce pizza expenditures by \$6.06, whereas individual with \$90k income will reduce pizza expenditures by \$14.07.

B. Log-Linear Model

A log-linear model is a mathematical model that takes the form of a function whose logarithm equals a linear combination of the parameters of the model, which makes it possible to apply (possibly multivariate) linear regression.

Formula:

Wage depends on years of education and years of experience:

$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$

If we believe the effect of an extra year of experience on wages will depend on the level of education, then:

$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 (EDUC \times EXPER) + e$

$\widehat{\ln(WAGE)} = 1.392 + 0.09494 EDUC + 0.00633 EXPER - 0.0000364 (EDUC \times EXPER)$

The greater the number of years of education (experience), the less valuable is an extra year of experience (education). For a person with 8 years education, how much increase in wages for additional years of experience? If 16 years?

Formula:

$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE$

$$\ln(WAGE) = \begin{cases} \beta 1 + \beta 2EDUC & MALES\ (FEMALE = 0) \\ (\beta 1 + \delta) + \beta 2EDUC & FEMALES\ (FEMALE = 1) \end{cases}$$

That dependent variable is ln (*WAGE*), does that have an effect?

Rough Calculation:

$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \delta$

$\widehat{\ln(WAGE)} = 1.6539 + 0.0962EDUC - 0.2432FEMALE$

We estimate 24.32% differential between male and female wages.

Exact Calculation:

$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \ln\left(\frac{WAGE\ females}{WAGE\ males}\right) = \delta$

$\frac{WAGE\ females}{WAGE\ males} = e^{\delta}$

$\frac{WAGE\ females}{WAGE\ males} - \frac{WAGE\ males}{WAGE\ males} = \frac{WAGE\ females - WAGE\ males}{WAGE\ males} =$

$e^{\delta} - 1$

$= 100\ (e^{\delta} - 1)\ \% = (100\ (e^{-0.2432} - 1)\ \% = -21.59\%$

The percentage difference between wages of females and males is 21.59%.

C. Measuring Goodness of Fit

Coefficient of determination R^2 is a measure of the proportion of variation in the dependent variable that is explained by variation in the explanatory variable, how well the estimated regression fits the data.

$$\mathbf{R^2 = SSR/SST = 1- (SSE/SST)}$$

note:

SSR : variation in y explained by the model

SST : total variation in y about its mean

SSE : portion of the variation in y that is not explained by the model

$$\mathbf{SST = (N - 1)\ s^2_y}$$

Example:

R^2 = 0.448 → 44.8% variation in variable y is explained by the variation in x_1 and x_2 → 55.2% in our sample is left unexplained due to variation in the error term. If model does not contain an intercept parameter, R^2 is no longer appropriate → SST ≠ SSR + SSE → better not to report R^2.

D. Indicator variables

Indicator variables are used to account for qualitative factors in econometrics models, called as dummy, binary, or dichotomous → take just two values, one or zero, to indicate the presence or absence or to indicate a condition is true or false, or numeric variable for qualitative characteristic, ex: D = 1 if characteristic is present, or 0 if it is not present.

Formula:

$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e$

The price of a house is explained as a function of the value of land and square foot of living area (SQFT).

$$E\,(PRICE) = \begin{cases} (\beta 1 + \delta) + \beta 2SQFT & when\ D = 1 \\ \beta 1 + \beta 2SQFT & when\ D = 0 \end{cases}$$

In the desirable neighborhood D = 1, & the intercept of the regression function is (β_1+δ). In other areas the regression function intercept is β_1, assuming that $\delta > 0$. The interpretation of δ is a location premium → intercept dummy variable. If $\delta = 0$ → no location premium.

Choosing the Reference Group:

$$LD = \begin{cases} 1 & if\ property\ is\ not\ in\ the\ desirable\ neighborhood \\ 0 & if\ property\ is\ in\ the\ desirable\ neighborhood \end{cases}$$

The indicator variable is the opposite from *D*, and $LD = 1 - D$, if we include in the model:

$PRICE = \beta_1 + \delta D + \lambda\, LD + \beta_2 SQFT + e$

In this model, $D + LD = 1$ → exact collinearity → least square estimator is not defined → dummy variable trap.

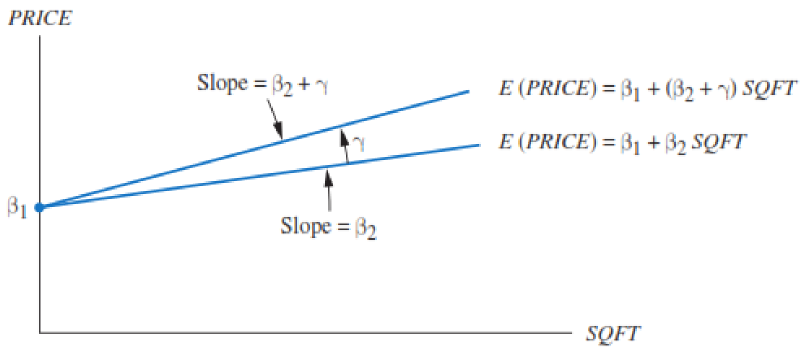
Slope Indicator Variables:

$PRICE = \beta_1 + \beta_2 SQFT + \Upsilon\,(SQFT \text{ x } D) + e$

(SQFT x D) is a interaction variable, capture the interaction effect of location & size on house price, slope dummy variable, allows for a change in slope of the relationship, taking value equal to SQFT for houses in the desirable neighbourhood, and zero for homes in other neighbourhoods.

$E\,(PRICE) = \beta_1 + \beta_2 SQFT + \Upsilon\,(SQFT \text{ x } D) = LD =$

$$\begin{cases} \beta 1 + (\beta 2 + \Upsilon) SQFT & when D = 1 \\ \beta 1 + \beta 2 SQFT & when D = 0 \end{cases}$$



Picture 1. ***Slope Indicator Variables***

An Example:

A real estate economist collects information on 1000 house price sales from university town and in a neighbourhood about three miles from university.

$PRICE = \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \Upsilon\,(SQFT \text{ x } UTOWN) + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e$

AGE (in years), location (*UTOWN* – 1 for homes near the university, 0 otherwise), whether the house has a pool (*POOL* – 1 if present, 0 otherwise), and whether the house gas a fireplace (*FPLACE* = 1 if present, 0 otherwise).

The estimated regression function for the houses near the university:

$\widehat{PRICE} = (24.5 + 27.453) + (7.6122 + 1.2994)\ SQFT - 0.1901 AGE + 4.3772 POOL + 1.6492 FPLACE$

$= 51.953 + 8.9116 SQFT - 0.1901 AGE + 4.3772 POOL + 1.6492 FPLACE$

For houses in other areas,

$\widehat{PRICE} = 24.5 + 7.6122\ SQFT - 0.1901 AGE + 4.3772 POOL + 1.6492 FPLACE$

Estimate the regression results!

E. Interaction Between Qualitative Factors

We want to know whether wage determination is not discriminatory.

WAGE $= \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \Upsilon (BLACK \times FEMLAE) + e$

$$E(WAGE) = \begin{cases} \beta 1 + \beta 2 EDUC & WHITE-MALE \\ (\beta 1 + \delta 1) + \beta 2 EDUC & BLACK-MALE \\ (\beta 1 + \delta 2) + \beta 2 EDUC & WHITE-FEMALE \\ (\beta 1 + \delta 1 + \delta 2 + \Upsilon) + \beta 2 EDUC & BLACK-FEMALE \end{cases}$$

White males are the reference group, black = 0 and female = 0.

F. Qualitative Factors with Several Categories

An example is the variable region of the country in our wage equation: northeast, mid-west, south and west.

WAGE $= \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 \mathrm{WEST} + e$

The sum of regional indicator variables will be 1 → exact collinearity → dummy variable trap → least squares estimation fails → solution: reference group.

$$E(WAGE) = \begin{cases} (\beta 1 + \delta 3) + \beta 2 EDUC & WEST \\ (\beta 1 + \delta 2) + \beta 2 EDUC & MIDWEST \\ (\beta 1 + \delta 1) + \beta 2 EDUC & SOUTH \\ \beta 1 + \beta 2 EDUC & NORTHEAST \end{cases}$$

G. Testing the Equivalence of Two Regressions

Regression functions for the house prices in two locations.

PRICE $= \beta_1 + \delta D + \beta_2 SQFT + \Upsilon (SQFT \times D) + e$

$$E(PRICE) = \begin{cases} \alpha 1 + \alpha 2\, SQFT & D = 1 \\ \beta 1 + \beta 2\, SQFT & D = 0 \end{cases}$$

Estimating separate regressions for each neighbourhood → Chow Test for the equivalence of two regressions.

WAGE $= \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \Upsilon (BLACK \times FEMALE) + e$

Are there differences between the wage regressions for the south and the rest of the country?

WAGE $= \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \Upsilon (BLACK \times FEMALE)$

$+ \Theta_1 SOUTH + \Theta_2(EDUC \times SOUTH) + \Theta_3(BLACK \times SOUTH) + \Theta_4 (FEMALE \times SOUTH) + \Theta_5(BLACK \times FEMALE \times SOUTH) + e$

$$E(WAGE) = \begin{cases} \beta1 + \beta2 EDUC + \delta1 BLACK + \delta2\, FEMALE + \Upsilon\,(BLACK \times FEMALE) & SOUTH = 0 \\ (\beta1 + \theta1) + (\beta1 + \theta2) EDUC + (\delta1 + \theta3) BLACK + (\delta2 + \theta4) FEMALE + (\Upsilon + \theta5)\,(BLACK \times FEMALE) & SOUTH = 1 \end{cases}$$

$$\mathbf{H_0: \Theta_1 = \Theta_2 = \Theta_3 = \Theta_4 = \Theta_5 = 0}$$

If rejected, so some difference in the wage equation in the southern & rest of the country.

$$F = \frac{(\boldsymbol{SSEr - SSEu})/\boldsymbol{J}}{\boldsymbol{SSEu}/\,(\boldsymbol{N - K})}$$

SSE_U is obtained from the full model, SSE_R from model with SOUTH = 0, J is for added variables, N is for total observations, K is for total parameter in full model. When F-stat < F-tab → fail to reject H_0 → wage equation is the same in the southern region and rest of the country.

H. Controlling for Time

Seasonal indicators:

1. Summer → outdoor cooking on BBQ grills → sales of Royal Oak charcoal briquettes (dependent variable).
2. Explanatory variables: price of Royal Oak, price of competitive brands, price of complementary goods, advertising.
3. Monthly indicator variables, AUG = 1 if August, AUG = 0 otherwise.
4. Seasonal indicator variable, SUMMER = 1 if June, July, or August; SUMMER = 0 otherwise.

Regime effects:

Economic regime: set of structural economic conditions for certain period → may behave differently, ex: investment tax credit (enacted in 1962, suspended in 1966, reinstated in 1970, and eliminated in 1986).

$$ITC_t = \begin{cases} 1 & \text{if } t = 1962 - 1965, 1970 - 1986 \\ 0 & \text{otherwise} \end{cases}$$

$$INV_t = \beta_1 + \delta_1 ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$$

If tax credit was successful, $\delta > 0$.

I. The Linear Probability Model

Problem:

1. Usual error term assumptions cannot hold → *y* only takes two values, so do the error term → distribution of errors is not bell-shaped curve.
2. The error is not homoscedastic → variance estimator is incorrect.
3. Predicted values can fall outside the (0,1) interval → interpretation as probabilities does not make sense.
4. Advantage of simplicity → good estimates of marginal effects of changes in explanatory variables on choice probability *p*.

A marketing example:

$$COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$$

$E\ (\widehat{COKE}) = p_{COKE} = 0.8902 - 0.4009PRATIO + 0.0772DISP\_COKE - 0.1657DISP\_PEPSI$

If coke is 10% more expensive than Pepsi → probability of purchasing Coke reduced by 0.04. A store display for Coke is estimated to increase the probability of purchasing Coke by 0.077. A Pepsi display is estimated to reduce the probability of purchasing Coke by 0.166.

J. Treatment Effects

1. Those who had not gone to the hospital vs those who had been to the hospital.
2. Post hoc, ergo propter hoc → one event's preceding another does not mean the first cause of the second.
3. Those who had been in a hospital are less healthy but does not imply that going to the hospital causes less healthy.
4. Selection bias: some people chose (self-selected) to go to the hospital & the others did not.
5. When membership in treated group is determined by choice → not random sample.

6. Selection bias is an issue when asking how much participation in job training program increase wage? If voluntary, greater proportion of less skilled workers taking advantage.
7. Causal effect/treatment effect: measuring the effect of a new type of fertilizer on rice production → randomly assign identical rice fields to be treated with a new fertilizer (treatment group), others being treated with existing product (control group) → comparing both → randomized controlled experiment → preventing selection bias.

K. Difference Estimator

Simple regression model with explanatory variable as dummy variable.

$$d_i = \begin{cases} 1 & \text{individual in treatment group} \\ 0 & \text{individual in control group} \end{cases}$$

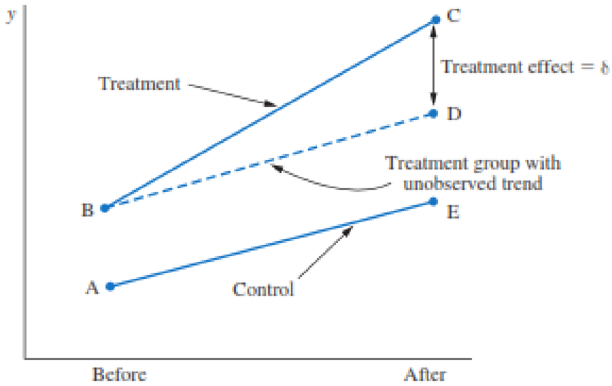
$y_i = \beta_1 + \beta_2 d_i + e_i$, $i = 1, \ldots, N$

$$E(y_i) = \begin{cases} \beta 1 + \beta 2 & \text{if in treatment group}, di = 1 \\ \beta 1 & \text{if in control group}, di = 0 \end{cases}$$

Those who had not gone to the hospital (control group) had average health score of 3.93, those who had been to the hospital (treatment group) had average health score of 3.21 → bias because the pre-existing health conditions for the treated group are poorer than the control group → solution: randomly assign.

The differences in differences estimator:

1. Randomized controlled experiments: expensive and involving human subjects.
2. Natural experiments (quasi experiments): approximate that would happen in randomized controlled experiment → before & after data.
3. Treatment group: affected by policy, control group: unaffected by policy → examine any change & compare it.



Picture 2. The differences in differences estimator

Note:

1. Before the policy → treatment group: B, after the policy: C.
2. Before the policy → control group: A, after the policy: E.
3. BD: what we imagine the treatment group growth in the absence of policy change.

$$\boldsymbol{\delta = (C - E) - (B - A)}$$

δ: differences in differences → estimator of treatment effect.

Formula:

$y_{it} = \beta_1 + \beta_2 TREAT_i + \beta_3 AFTER_t + \delta\,(TREAT_i \text{ x } AFTER_t) + e_{it}$

$$E\,(y_{it}) = \begin{cases} \beta 1 & TREAT = 0, AFTER = 0\ [Control\ before = A] \\ \beta 1 + \beta 2 & TREAT = 1, AFTER = 0\ [Treatment\ before = B] \\ \beta 1 + \beta 3 & TREAT = 0, AFTER = 1\ [Control\ after = B] \\ \beta 1 + \beta 2 + \beta 3 + & TREAT = 0, AFTER = 1\ [Treatment\ after = B] \end{cases}$$

$\delta \;= (C - E) - (B - A) = [(\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3)] - [(\beta_1 + \beta_2) - \beta_1]$

$\delta \;= [(b_1 + b_2 + b_3 + \delta) - (b_1 + b_3)] - [(b_1 + b_2) - b_1]$

$\;\;= (y_{treatment,\ After}) - (y_{control,\ After}) - (y_{treatment,\ Before}) - (y_{control,\ Before})$

Estimating the Effect of Minimum Wage Change:

Card & Krueger (1994) collected data on 410 fast food restaurants in New Jersey (treatment group) & Pennsylvania (control group), before period Feb 1992 and after period Nov 1992 → estimate the effect of treatment: raising minimum wage in New Jersey on employment in New Jersey → no significant reduction. In Pennsylvania, employment fell during Feb-Nov. Recall: minimum wage level was changed in New Jersey so that employment levels in Pennsylvania were not affected.

The differences in differences estimate of the change in employment due to change in minimum wage:

δ = ($FTE_{NJ\ After} - FTE_{PA\ After}$) – ($FTE_{NJ\ Before} - FTE_{PA\ Before}$)

= (21.0274 – 21.1656) – (20.4394 – 23.3312)

= 2.7536

Based on DD using sample means, employment increased by 2.75 employees during the period of increased New Jersey minimum wage → positive effect.

FTE: employment, treatment variable → NJ = 1 if New Jersey & 0 if Pennsylvania, time indicator → D = 1 if from November & 0 if from February, then the regression:

$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta\ (NJ_i \times D_t) + e_{it}$

Using 794 observations, with $\alpha = 0.05$ → not significant → we cannot conclude that the increase in minimum wage in New Jersey reduced employment at New Jersey.

L. Regression

Example:

Table 1.

Restaurant	**Number of Consumer**	**Business Ranking**	**Restaurant Annual Revenue**	**Predicted value (ŷ =125,29 + 14,2 + 22,81)**	**Residual Value**
1	20,8	3	527,1	489,08	38,02
2	27,5	2	548,7		
3	32,3	6	767,2		
4	37,2	5	722,9		
5	39,6	8	826,3		
6	45,1	3	810,5		
7	49,9	9	1040,7		
8	55,4	5	1033,6		
9	61,7	4	1090,3		
10	64,6	7	1235,8		
SSE = $(38,02)^2$ + ... + ... + ... + ... =					

We can write the regression equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

become,

$$= 125{,}29 + 14{,}2 + 22{,}81$$

We can call the above equation the "predictive value of the regression equation" (least square prediction equation). To get the residual value, you must get the "predicted" value of each restaurant, in the following way:

$$\hat{y} = 125{,}29 + 14{,}2(20{,}8) + 22{,}81(3)$$

$$= 489{,}08$$

Since the annual revenue value of restaurant 1 is 527.1, the residual value of restaurant 1 is:

$$y - \hat{y} = 527{,}1 - 489{,}08 = 38{,}02$$

If the residual values of 10 restaurants have been obtained, and the total results are added up, we will get the Sum of Squared Residual (SSE) value. After getting the SSE value, the next step is to find the mean square error (MSE) and standard error (SE) values. We can write the formula for MSE as follows:

$$s^2 = \frac{SSE}{n - (k + 1)}$$

the model applies utility to equation (*k*) where k is the number of independent variables, and (*n*) is the number of restaurants. In the standard error equation, we can write as follows:

$$s = \sqrt{MSE}$$

Conduct R^2 and Adjusted R^2:

In simple regression, we have performed calculations related to R-squared and Adjusted R-Squared. In multiple linear regression it is not much different from what we have done in simple regression. We can write the formula as follows:

$$\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 = \sum (\hat{y}_i - \bar{y})^2$$

With the data we use, we can write it as follows: (Looking back at the simple regression notes), we can call the total variation below as a whole.

$$\sum (y_i - \bar{y})^2 = (527{,}1 - 860{,}31)^2 + (548{,}7 - 860{,}31)^2 + \cdots \dots \dots \dots \dots \dots \dots$$

$$+ (1235{,}8 - 860{,}31)^2 = \mathbf{495776{,}51}$$

The SSE value can also be called the unexplained variation, so to calculate the R-squared value, we can write the equation:

$$Explained\ variation = Total\ variation - Unexplained\ variation$$

Then the R-squared formula is:

$$R^2 = \frac{Explained\ variation}{Total\ Variation}$$

The equation for the multiple regression correlation value is as follows:

$$r = \sqrt{R^2}$$

While the equation for the adjusted R-squared value is as follows:

$$\check{R}^2 = (R^2 - \frac{k}{n-1})(\frac{n-1}{n-(k+1)})$$

where k = number of independent variables, n = number of restaurants, and R^2 = is the value of R-squared. The next step is to determine the f-count value, with the equation we can write as follows:

$$F = \frac{(Explained\ variation)/k}{(Unexplained\ variation)/(n-(k+1))}$$

How to read from each result obtained is the same as what we have learned in simple regression. (Look again at the simple regression notes).

REFERENCE :

1. Wooldridge, J, M. 2010. Introductory Econometrics: A Modern Approach. WILEY.
2. Dougherty, C. 2001. Introduction to Econometrics, Third Edition. Oxford.
3. Martin, V, L. Hurn, S. dan Harris, D. 2011. Econometric Modelling With Time Series, Specification, Estimation and Testing. WILEY.