DUMMY VARIABLE REGRESSION MODELS

- The Nature of Dummy Variables
- Caution in the use of Dummy Variable
- ANOVA
- The Dummy Variable Alternative to Chow Test
- Interaction Effects Using Dummy Variables
- The Use of Dummy Variable in Seasonal Analysis

THE NATURE OF DUMMY VARIABLES

- Qualitative variables or nominal scale variables
- E.g. religion, sex, nationality, geographical region, etc.
- Variables that assume such 0 and 1 values

DUMMY VARIABLE REGRESSION MODELS

$$Y_i = \beta_1 + \beta_2 X_i + \alpha D_i + u_i$$

$$Y_i = \beta_1 + \beta_2 X_i + \alpha D_i + u_i$$

$Y_i =$ total consumption (thousand baht)

$X_i =$ total income (thousand baht)

$D_i = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$

$u_i =$ residual term

$$E(Y_i \mid D_i = 0) = \beta_1 + \beta_2 X_i + \alpha(0) = \beta_1 + \beta_2 X_i$$

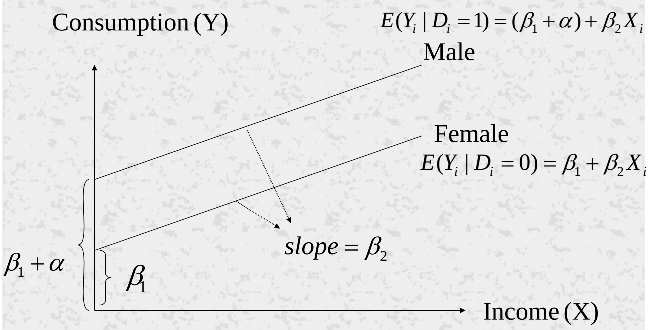
$$E(Y_i \mid D_i = 1) = \beta_1 + \beta_2 X_i + \alpha(1) = (\beta_1 + \alpha) + \beta_2 X_i$$

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = \alpha$$

The difference between mean total consumption of male and female is equal α

$\alpha > 0$ when the mean total consumption of male is more than female

$\alpha < 0$ when the mean total consumption of male is less than female



$$H_0 : \alpha = 0$$
$$H_1 : \alpha \neq 0$$

$$\hat{Y}_i = 2.34 + 0.83 X_i + 0.44 D_i$$
$$t^* = (2.18)(10.44) \quad (5.26)$$

CAUTION IN THE USE OF DUMMY VARIABLE

- If a qualitative variable has m categories, introduce only (m-1) dummy variables
- The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category**
- The intercept value represents the mean value of the benchmark category

- The coefficient attached to the dummy variables in

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

are known as the **differential intercept coefficients**

- If a qualitative variable has more than one category, the choice of the benchmark category is strictly up to the researcher

- **Dummy variable trap**

There is a way to circumvent this trap by introducing as many dummy variables as the number of categories of that variable, provided we do not introduce the intercept in such a model

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

We do not fall into the dummy variable trap, as there is no longer perfect collinearity

- Which is a better method of introducing a dummy variable:

(1) introduce a dummy for each category and omit the intercept term or

(2) include the intercept term and introduce only (m-1) dummies, where m is the number of categories of the dummy variable?

REGRESSION WITH A MIXTURE OF QUANTITATIVE AND QUALITATIVE REGRESSORS: THE ANCOVA MODELS

- Example 9.3 Teacher's salary in relation to region and spending on public school per pupil

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$$

$Y_i =$ average annual salary of public school teachers in state ($)

$X_i =$ spending on public school per pupil ($)

$D_{2i} = 1,$ if the state is in the Northeast or North Central

$= 0,$ otherwise

$D_{3i} = 1,$ if the state is in the South

$= 0,$ otherwise

Source	SS	df	MS
Model	1.1204e+09	3	373457549
Residual	1.1309e+09	47	24062057.9
Total	2.2513e+09	50	45025787.3

Number of obs =	51
F(3, 47) =	15.52
Prob > F =	0.0000
R-squared =	0.4977
Adj R-squared =	0.4656
Root MSE =	4905.3

salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
spending	2.340429	.359225	6.52	0.000	1.617761	3.063096
d2	-2954.127	1862.576	-1.59	0.119	-6701.146	792.8921
d3	-3112.195	1819.873	-1.71	0.094	-6773.306	548.9165
_cons	28694.92	3262.521	8.80	0.000	22131.57	35258.26

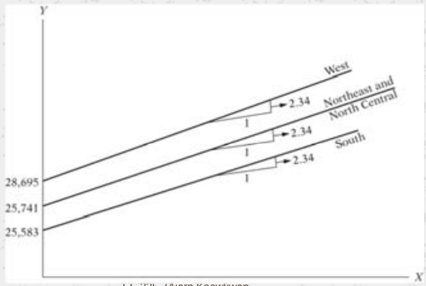
$$\hat{Y}_i = 28{,}694.918 - 2{,}954.127 D_{2i} - 3{,}112.194 D_{3i} + 2.3404 X_i$$
$$se = (3262.521) \quad (1862.576) \quad (1819.873) \quad (0.3592)$$
$$t = (8.795)^{*} \quad (-1.586)^{**} \quad (-1.710)^{**} \quad (6.515)^{*}$$
$$R^2 = 0.4977$$

where * indicates p values less than 5 percent

** indicates p values greater than 5 percent

As these results suggest, ceteris paribus: as public expenditure goes up by a dollar, on average, a public school teacher's salary goes up by about $2.34

- Controlling for spending on education, we now see that the differential intercept coefficient is not significant for either the Northeast and North Central region or for the South.

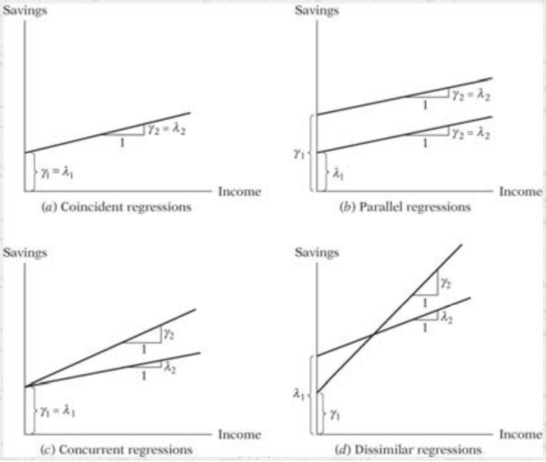


THE DUMMY VARIABLE ALTERNATIVE TO CHOW TEST

Chow Test – examine the structural stability of a regression model

- **Coincident regressions** – both the intercept and the slope coefficients are the same in the two regressions
- **Parallel regressions**- only the intercepts in the two regressions are different but the slopes are the same

- **Concurrent regressions** – the intercepts in the two regressions are the same, but the slopes are different
- **Dissimilar regressions** – both the intercepts and slopes in the two regressions are different



(a) Coincident regressions (b) Parallel regressions (c) Concurrent regressions (d) Dissimilar regressions

EXAMPLE

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

Y = savings

X = income

t = time

D =1 for observations in 1982-1995

=0, otherwise

TABLE 9.2
Savings and Income Data, United States, 1970–1995

Source: *Economic Report of the President*, 1997, Table B-28, p. 332.

Observation	Savings	Income	Dum
1970	61	727.1	0
1971	68.6	790.2	0
1972	63.6	855.3	0
1973	89.6	965	0
1974	97.6	1054.2	0
1975	104.4	1159.2	0
1976	96.4	1273	0
1977	92.5	1401.4	0
1978	112.6	1580.1	0
1979	130.1	1769.5	0
1980	161.8	1973.3	0
1981	199.1	2200.2	0
1982	205.5	2347.3	1
1983	167	2522.4	1
1984	235.7	2810	1
1985	206.2	3002	1
1986	196.5	3187.6	1
1987	168.4	3363.1	1
1988	189.1	3640.8	1
1989	187.8	3894.5	1
1990	208.7	4166.8	1
1991	246.4	4343.7	1
1992	272.6	4613.7	1
1993	214.4	4790.2	1
1994	189.4	5021.7	1
1995	249.3	5320.8	1

Note: Dum = 1 for observations beginning in 1982; 0 otherwise. Savings and income figures are in billions of dollars.

$$E(u_i) = 0$$

Mean savings function for 1970-1981

$$E(Y_t \mid D_t = 0, X_t) = \alpha_1 + \beta_1 X_t$$

Mean savings function for 1982-1995

$$E(Y_t \mid D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t$$


```
      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  3,    22) =   54.78
       Model |  88079.8327     3  29359.9442           Prob > F      =  0.0000
    Residual |  11790.2539    22  535.920634           R-squared     =  0.8819
-------------+------------------------------           Adj R-squared =  0.8658
       Total |  99870.0867    25  3994.80347           Root MSE      =   23.15

------------------------------------------------------------------------------
     savings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0803319   .0144968     5.54   0.000     .0502673    .1103964
         dum |   152.4786   33.08237     4.61   0.000     83.86992    221.0872
      incdum |  -.0654694   .0159824    -4.10   0.000     -.098615   -.0323239
       _cons |   1.016115   20.16483     0.05   0.960    -40.80319    42.83542
------------------------------------------------------------------------------
```


$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

$$\hat{Y}_t = 1.0161 + 152.4786 D_t + 0.0803 X_t + 0.0655 (D_t X_t)$$

$$se = (20.1648)\ (33.0824)\quad (0.0144)\quad (0.0159)$$

$$t = (0.0504)^{**} (4.6090)\quad (5.5413)^{*}\quad (-4.0963)^{*}$$

$$R^2 = 0.8891$$

where * indicates p values less than 5 percent

** indicates p values greater than 5 percent

Savings function for 1970-1981

$$\hat{Y}_t = 1.0161 + 0.0803 X_t$$

Savings function for 1982-1995

$$\hat{Y}_t = (1.0161 + 152.4786) + (0.0803 - 0.0655) X_t$$
$$= 153.4947 + 0.0148 X_t$$

- The Chow test does not explicitly tell us which coefficient, intercept, or slope is different or whether both are different in two periods.
- That is, one can obtain a significant Chow test because the slope only is different or the intercept only is different or both are different

INTERACTION EFFECTS USING DUMMY VARIABLES

- Interaction between the two qualitative variables

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$$

Y = hourly wage in dollars

X = education (years of schooling)

D_2 = 1 if female, 0 otherwise

D_3 = 1 if nonwhite and non-Hispanic, 0 otherwise

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i$$

which is the mean hourly wage function for female nonwhite/ non-Hispanic workers

α_2 = differential effect of being a female

α_3 = differential effect of being a nonwhite/non-Hispanic

α_4 = differential effect of being a female nonwhite/non-Hispanic

THE USE OF DUMMY VARIABLE IN SEASONAL ANALYSIS

- **Deseasonalization** or **Seasonal Adjustment** - the process of removing the seasonal component from a time series
- **Important Economic time series** such as the unemployment rate, the consumer price index (CPI), the producer price index (PPI), etcs.

EXAMPLE

TABLE 9.3
Quarterly Data on Appliance Sales (in thousands) and Expenditure on Durable Goods (1978–I to 1985–IV)

Source: *Business Statistics and Survey of Current Business*, Department of Commerce (various issues).

DISH	DISP	FRIG	WASH	DUR	DISH	DISP	FRIG	WASH	DUR
841	798	1317	1271	252.6	480	706	943	1036	247.7
957	837	1615	1295	272.4	530	582	1175	1019	249.1
999	821	1662	1313	270.9	557	659	1269	1047	251.8
960	858	1295	1150	273.9	602	837	973	918	262
894	837	1271	1289	268.9	658	867	1102	1137	263.3
851	838	1555	1245	262.9	749	860	1344	1167	280
863	832	1639	1270	270.9	827	918	1641	1230	288.5
878	818	1238	1103	263.4	858	1017	1225	1081	300.5
792	868	1277	1273	260.6	808	1063	1429	1326	312.6
589	623	1258	1031	231.9	840	955	1699	1228	322.5
657	662	1417	1143	242.7	893	973	1749	1297	324.3
699	822	1185	1101	248.6	950	1096	1117	1198	333.1
675	871	1196	1181	258.7	838	1086	1242	1292	344.8
652	791	1410	1116	248.4	884	990	1684	1342	350.3
628	759	1417	1190	255.5	905	1028	1764	1323	369.1
529	734	919	1125	240.4	909	1003	1328	1274	356.4

Note: DISH = dishwashers; DISP = garbage disposers; FRIG = refrigerators; WASH = washing machines; DUR = durable goods expenditure, billions of 1982 dollars.

SALES OF REFRIGERATORS OVER THE SAMPLE PERIOD

$$Y_i = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t$$

Y_i = Sales of refrigerators (in thousands)

$D = \begin{cases} 0 & \text{otherwise} \\ 1 & \text{the relevant quarter} \end{cases}$

$$\hat{Y}_t = 1,222.125 D_{1t} + 1,467.500 D_{2t} + 1,569.750 D_{3t} + 1,160.000 D_{4t}$$

$$t = (20.3720) \quad (24.4622) \quad (26.1666) \quad (19.3364)$$

$$R^2 = 0.5317$$

The average sales of refrigerators (in thousands of units in each season (i.e., quarter)

The average sales of refrigerators in the first quarter in thousands of units, is about 1,222

The average sales of refrigerators in the second quarter in thousands of units, is about 1,468

The average sales of refrigerators in the third quarter in thousands of units, is about 1,570

The average sales of refrigerators in the fourth quarter in thousands of units, is about 1,160

TABLE 9.4
U.S. Refrigerator Sales (thousands), 1978–1985 (quarterly)

Source: Business Statistics and Survey of Current Business, Department of Commerce (various issues).

FRIG	DUR	D_2	D_3	D_4	FRIG	DUR	D_2	D_3	D_4
1317	252.6	0	0	0	943	247.7	0	0	0
1615	272.4	1	0	0	1175	249.1	1	0	0
1662	270.9	0	1	0	1269	251.8	0	1	0
1295	273.9	0	0	1	973	262.0	0	0	1
1271	268.9	0	0	0	1102	263.3	0	0	0
1555	262.9	1	0	0	1344	280.0	1	0	0
1639	270.9	0	1	0	1641	288.5	0	1	0
1238	263.4	0	0	1	1225	300.5	0	0	1
1277	260.6	0	0	0	1429	312.6	0	0	0
1258	231.9	1	0	0	1699	322.5	1	0	0
1417	242.7	0	1	0	1749	324.3	0	1	0
1185	248.6	0	0	1	1117	333.1	0	0	1
1196	258.7	0	0	0	1242	344.8	0	0	0
1410	248.4	1	0	0	1684	350.3	1	0	0
1417	255.5	0	1	0	1764	369.1	0	1	0
919	240.4	0	0	1	1328	356.4	0	0	1

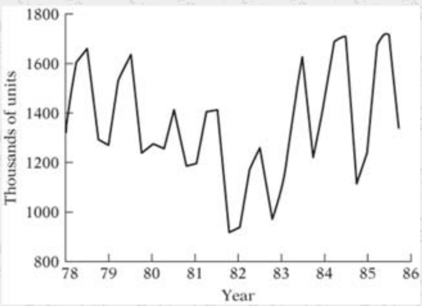
Note: FRIG = refrigerator sales, thousands.
DUR = durable goods expenditure, billions of 1982 dollars.
D_2 = 1 in the second quarter, 0 otherwise.
D_3 = 1 in the third quarter, 0 otherwise.
D_4 = 1 in the fourth quarter, 0 otherwise.

$$\hat{Y}_t = 1,222.1250 + 245.3750 D_{2t} + 347.6250 D_{3t} - 62.1250 D_{4t}$$

$$t = (20.3720)^* \quad (2.8922)^* \quad (4.0974)^* \quad (-0.7322)^{**}$$

$$R^2 = 0.5318$$

p values <5% ** p values >5%



$$Y_i = \beta_1 + \beta_2 X_{2i} + \alpha Sex_i + \gamma_1 Edu_{1i} + \gamma_2 Edu_{2i} + \lambda_1 (Sex_i Edu_{1i}) + \lambda_2 (Sex_i Edu_{2i}) + u_i$$

Y_i = Consumption expenditure (thousand baht)

X_i = Income (thousand baht)

$Sex_i = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$

$Educ_{1i} = \begin{cases} 1 & \text{Bachelor degree} \\ 0 & \text{others} \end{cases}$

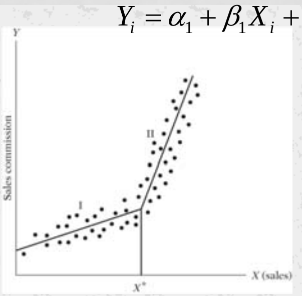
$Educ_{2i} = \begin{cases} 1 & \text{Master degree or higher} \\ 0 & \text{others} \end{cases}$

$$E(Y_i \mid Sex_i = 0, Edu_{1i} = 0, Edu_{2i} = 0) = \beta_1 + \beta_2 X_i$$
$$E(Y_i \mid Sex_i = 0, Edu_{1i} = 1, Edu_{2i} = 0) = (\beta_1 + \gamma_1) + \beta_2 X_i$$
$$E(Y_i \mid Sex_i = 0, Edu_{1i} = 0, Edu_{2i} = 1) = (\beta_1 + \gamma_2) + \beta_2 X_i$$
$$E(Y_i \mid Sex_i = 1, Edu_{1i} = 0, Edu_{2i} = 0) = (\beta_1 + \alpha) + \beta_2 X_i$$
$$E(Y_i \mid Sex_i = 1, Edu_{1i} = 1, Edu_{2i} = 0) = (\beta_1 + \alpha + \gamma_1 + \lambda_1) + \beta_2 X_i$$
$$E(Y_i \mid Sex_i = 1, Edu_{1i} = 0, Edu_{2i} = 1) = (\beta_1 + \alpha + \gamma_2 + \lambda_2) + \beta_2 X_i$$

$$E(Y_i \mid Sex_i = 0, Edu_{1i} = 1, Edu_{2i} = 0) - E(Y_i \mid Sex_i = 0, Edu_{1i} = 0, Edu_{2i} = 0) = \gamma_1$$

PIECEWISE LINEAR REGRESSION

$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$$

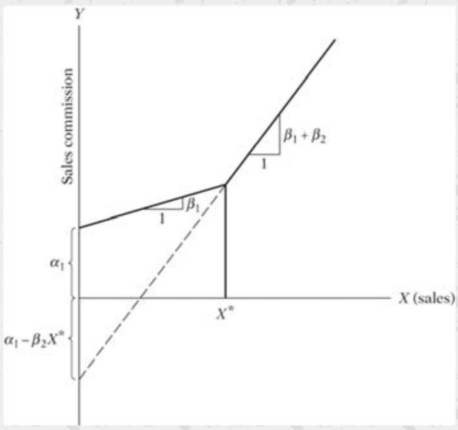


$$\begin{aligned} \text{where } Y_i &= \text{sales commission} \\ X_i &= \text{volume of sales generated by the sales person} \\ X^* &= \text{threshold value of sales also known as a knot} \\ D_i &= 1 \text{ if } X_i > X^* \\ &= 0 \text{ if } X_i < X^* \end{aligned}$$

Assuming $E(u_i) = 0$

$$E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i$$

$$E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i$$



EXAMPLE

Total Costs in Relation to Output

TABLE 9.6
Hypothetical Data on Output and Total Cost

Total Cost, Dollars	Output, Units
256	1,000
414	2,000
634	3,000
778	4,000
1,003	5,000
1,839	6,000
2,081	7,000
2,423	8,000
2,734	9,000
2,914	10,000

Source	SS	df	MS
Model	8832644.9	2	4416322.45
Residual	238521.502	7	34074.5002
Total	9071166.4	9	1007907.38

Number of obs =	10
F(2, 7) =	129.61
Prob > F =	0.0000
R-squared =	0.9737
Adj R-squared =	0.9662
Root MSE =	184.59

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.2791258	.0460081	6.07	0.001	.1703338	.3879177
xd	.0945	.0825524	1.14	0.290	-.1007054	.2897054
_cons	-145.7167	176.7341	-0.82	0.437	-563.6265	272.1932

We are told that the total cost may change its slope at the output level of 5,500 units

$$\hat{Y}_i = -145.72 + 0.2791X^* + 0.0945(X_i - X_i^*)D_i$$
$$t = (-0.8245)\ (6.0669)\qquad(1.1447)$$
$$R^2 = 0.9737$$
$$X^* = 5,500$$

- The marginal cost of production is about 28 cents per unit and although it is about 37 cents (28+9) for output over 5,500 units, the difference between the two is not statistically significant because the dummy variable is not significant at, say, the 5 percent level.
- For all practical purposes, then, one can regress total cost on total output, dropping the dummy variable

SOME TECHNICAL ASPECT OF THE DUMMY VARIABLE TECHNIQUE

- The interpretation of Dummy variables in semilogarithmic regressions

$$\ln Y_i = \beta_1 + \beta_2 D_i + u_i$$
$$Y = hourly\ \ wage\ \ rate(\$)$$
$$D = 1\ \ for\ \ female\ and\ \ 0\ \ for\ \ male$$

Wage function for male workers:

$$E(\ln Y_i \mid D_i = 0) = \beta_1$$

Wage function for female workers:

$$E(\ln Y_i \mid D_i = 1) = \beta_1 + \beta_2$$

When we take antilog of β_1, this represents the median hourly wages of male workers

When we take antilog of $(\beta_1 + \beta_2)$, we obtain the median hourly wages of female workers

LOGARITHM OF HOURLY WAGES IN RELATION TO GENDER

$$\widehat{lnY_i} = 2.1763 - 0.2437D_i$$
$$t = (72.2943)(-5.5048)$$

$$R^2 = 0.0544$$

- Taking antilog of 2.1763, we find $8.8136, which is the median hourly earnings of male workers, and taking the antilog of (2.1763-0.2437=1.92857), we obtain $6.8796, which is the median hourly earnings of female workers

- We can obtain semielasticity for a dummy regressor

  Take the antilog (to base e) of the estimated dummy coefficient and subtract 1 from it and multiply the difference by 100.

  Take the antilog of -0.2437, you will obtain 0.78366. Subtracting 1 from this gives -0.2163. After multiplying this by 100, we get -21.63 percent, suggesting that a female worker's median salary is lower than that of her male counterpart by about 21.63%

SOURCE

Gujarati, D.N. (2009) Basic Econometrics. 5th ed. Singapore, McGraw-Hill.