Examples on the use of Dummy Variable:

when the explanatory variable has a qualitative nature

EXAMPLE 9.1 Public School Teachers' Salaries by Geographical Region

Table 9.1 Average Salary of Public School Teachers by State, 2005-2006

	Salary	Spending	D_2	D_3		Salary	Spending	D_2	D_3
Connecticut	60,822	12,436	1	0	Georgia	49,905	8,534	0	1
Illinois	58,246	9,275	1	0	Kentucky	43,646	8,300	0	1
Indiana	47,831	8,935	1	0	Louisiana	42,816	8,519	0	1
Iowa	43,130	7,807	1	0	Maryland	56,927	9,771	0	1
Kansas	43,334	8,373	1	0	Mississippi	40,182	7,215	0	1
Maine	41,596	11,285	1	0	North Carolina	46,410	7,675	0	1
Massachusetts	58,624	12,596	1	0	Oklahoma	42,379	6,944	0	1
Michigan	54,895	9,880	1	0	South Carolina	44,133	8,377	0	1
Minnesota	49,634	9,675	1	0	Tennessee	43,816	6,979	0	1
Missouri	41,839	7,840	1	0	Texas	44,897	7,547	0	1
Nebraska	42,044	7,900	1	0	Virginia	44,727	9,275	0	1
New Hampshire	46,527	10,206	1	0	West Virginia	40,531	9,886	0	1
New Jersey	59,920	13,781	1	0	Alaska	54,658	10,171	0	0
New York	58,537	13,551	1	0	Arizona	45,941	5,585	0	0
North Dakota	38,822	7,807	1	0	California	63,640	8,486	0	0
Ohio	51,937	10,034	1	0	Colorado	45,833	8,861	0	0
Pennsylvania	54,970	10,711	1	0	Hawaii	51,922	9,879	0	0
Rhode Island	55,956	11,089	1	0	Idaho	42,798	7,042	0	0
South Dakota	35,378	7,911	1	0	Montana	41,225	8,361	0	0
Vermont	48,370	12,475	1	0	Nevada	45,342	6,755	0	0
Wisconsin	47,901	9,965	1	0	New Mexico	42,780	8,622	0	0
Alabama	43,389	7,706	0	1	Oregon	50,911	8,649	0	0
Arkansas	44,245	8,402	0	1	Utah	40,566	5,347	0	0
Delaware	54,680	12,036	0	1	Washington, D.C.	47,882	7,958	0	0
District of Columbia	59,000	15,508	0	1	Wyoming	50,692	11,596	0	0
Florida	45,308	7,762	0	1					

Note: $D_2 = 1$ for states in the Northeast and North Central; 0 otherwise.
$D_3 = 1$ for states in the South; 0 otherwise.

Source: National Educational Association, as reported in 2007.

These 51 areas can be classified into 3 geographical regions:

1.West (13 states)

2.North (21 states)

3.South (17 states)

We may want to find out if the average annual salary of public school teachers differs among the three geographical regions of the United States.

Consider the following model:

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

where Y_i = (average) salary of public school teacher in state i

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

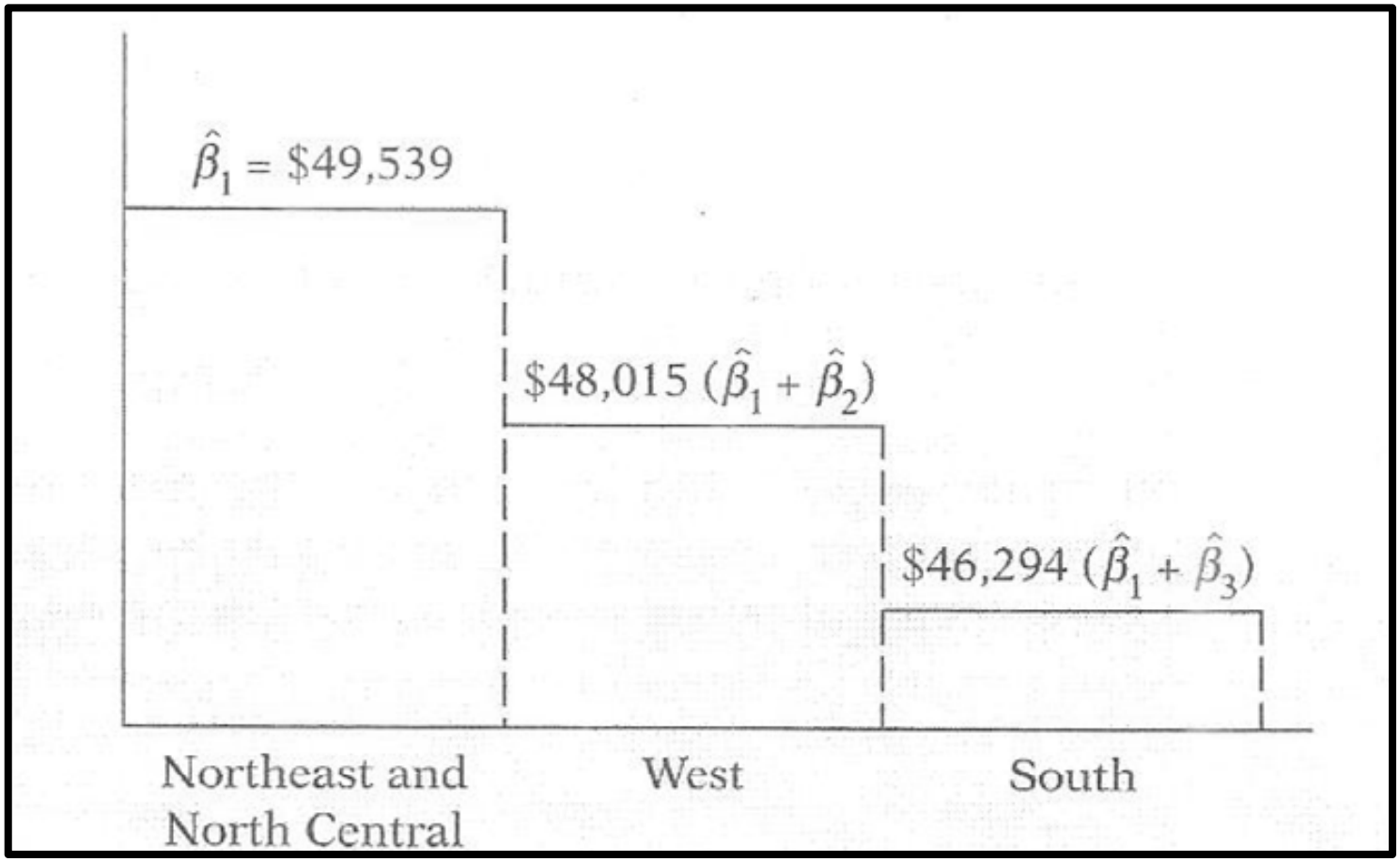
$= 0$ otherwise (i.e., in other regions of the country)

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

$= 0$ otherwise (i.e., in other regions of the country)

EXAMPLE 9.1 Public School Teachers' Salaries by Geographical Region (Continued)

Figure 9.1 Average Salary (in dollars) of public school teachers in three regions.



EXAMPLE 9.1 Public School Teachers' Salaries by Geographical Region (Continued)

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2 \qquad \textbf{(9.2.2)}$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3 \qquad \textbf{(9.2.3)}$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1 \qquad \textbf{(9.2.4)}$$

$$\begin{aligned}
\hat{Y}_i &= 48{,}014.615 \quad + 1{,}524.099 D_{2i} - 1{,}721.027 D_{3i} \\
\text{se} &= (1857.204) \quad (2363.139) \quad (2467.151) \\
t &= \quad (25.853) \quad (0.645) \quad (-0.698) \\
&\quad (0.0000)^* \quad (0.5220)^* \quad (0.4888)^* \quad R^2 = 0.0440
\end{aligned} \qquad \textbf{(9.2.5)}$$

EXAMPLE 9.2 Hourly Wages in Relation to Marital Status and Region of Residence

From a sample of 528 persons in May 1985, the following regression results were obtained.[8]

$$\begin{aligned} \hat{Y}_i &= 8.8148 + 1.0997D_{2i} - 1.6729D_{3i} \\ \text{se} &= (0.4015) \quad (0.4642) \quad (0.4854) \\ t &= (21.9528) \quad (2.3688) \quad (-3.4462) \\ &\quad (0.0000)^* \quad (0.0182)^* \quad (0.0006)^* \end{aligned} \tag{9.3.1}$$

where Y = hourly wage ($) $R^2 = 0.0322$

D_2 = married status; 1 = married, 0 = otherwise

D_3 = region of residence; 1 = South, 0 = otherwise

and * denotes the p values.

In this example we have two qualitative regressors, each with two categories. Hence we have assigned a single dummy variable for each category.

Which is the benchmark category here? Obviously, it is unmarried, non-South residence. In other words, unmarried persons who do not live in the South are the omitted category. Therefore, all comparisons are made in relation to this group. The mean hourly wage in this benchmark is about $8.81. Compared with this, the average hourly wage of those who are married is higher by about $1.10, for an actual average wage of $9.91 (= 8.81 + 1.10). By contrast, for those who live in the South, the average hourly wage is lower by about $1.67, for an actual average hourly wage of $7.14.

Are the preceding average hourly wages statistically different compared to the base category? They are, for all the differential intercepts are statistically significant, as their p values are quite low.

The point to note about this example is this: *Once you go beyond one qualitative variable, you have to pay close attention to the category that is treated as the base category, since all comparisons are made in relation to that category. This is especially important when you have several qualitative regressors, each with several categories.* But the mechanics of introducing several qualitative variables should be clear by now.

EXAMPLE 9.3 Teachers' Salary in Relation to Region and Spending on Public School per Pupil

To motivate the analysis, let us reconsider Example 9.1 by maintaining that the average salary of public school teachers may not be different in the three regions if we take into account any variables that cannot be standardized across the regions. Consider, for example, the variable *expenditure on public schools by local authorities,* as public education is primarily a local and state question. To see if this is the case, we develop the following model:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i \qquad \textbf{(9.4.1)}$$

where 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

The data on X are given in Table 9.1. Keep in mind that we are treating the West as the benchmark category. Also, note that besides the two qualitative regressors, we have a quantitative variable, X, which in the context of the ANCOVA models is known as a **covariate**, as noted earlier.

From the data in Table 9.1, the results of the model (9.4.1) are as follows:

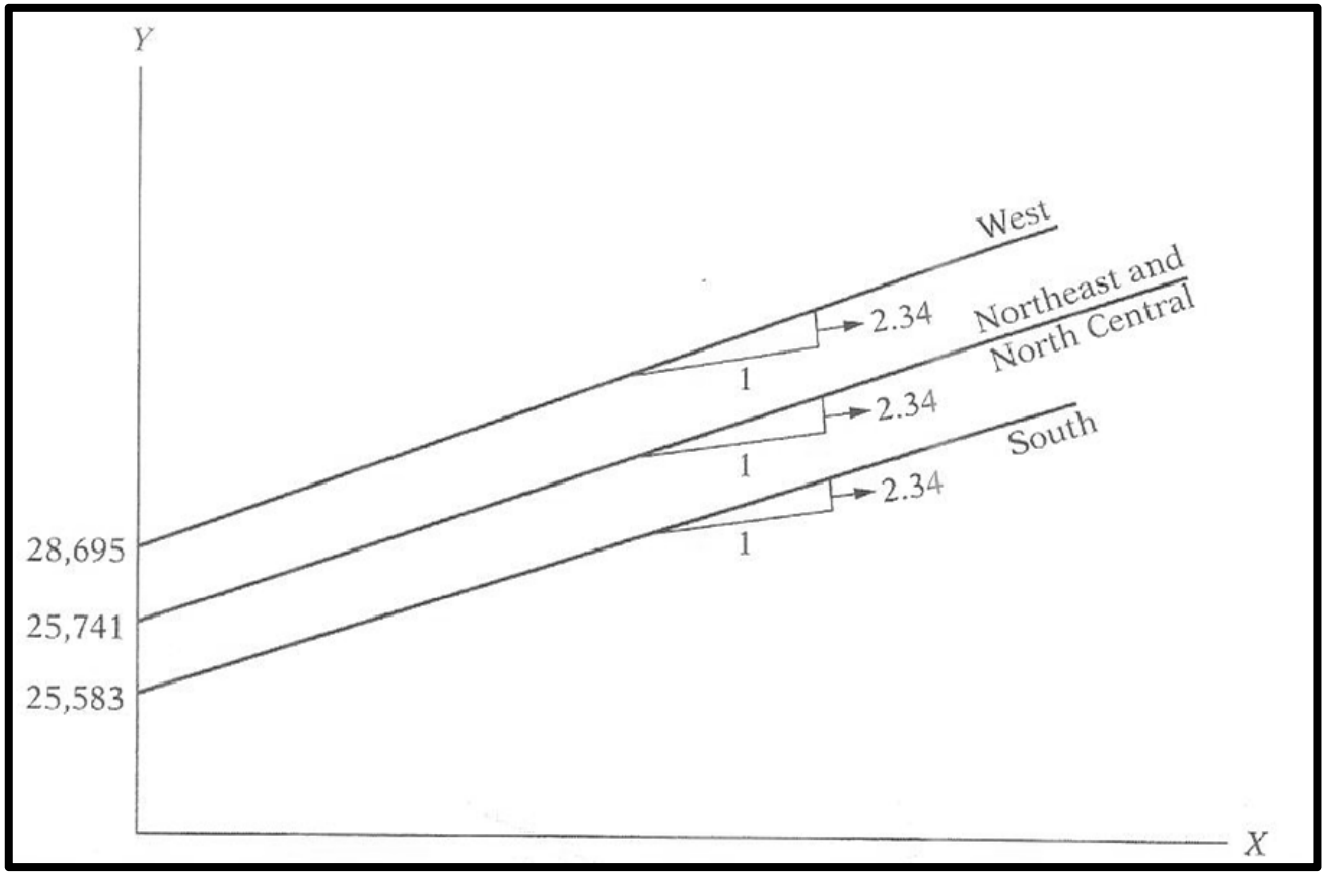
$$\begin{aligned} \hat{Y}_i &= 28{,}694.918 \quad - 2{,}954.127 D_{2i} - 3{,}112.194 D_{3i} + \quad 2.3404 X_i \\ \text{se} &= \quad (3262.521) \quad (1862.576) \quad (1819.873) \quad (0.3592) \\ t &= \quad (8.795)^* \quad (-1.586)^{**} \quad (-1.710)^{**} \quad (6.515)^* \end{aligned} \qquad \textbf{(9.4.2)}$$

$$R^2 = 0.4977$$

where * indicates p values less than 5 percent, and ** indicates p values greater than 5 percent.

EXAMPLE 9.3 Teachers' Salary in Relation to Region and Spending on Public School per Pupil

Figure 9.2 Public School Teacher's Salary (y) in Relation to per Pupil Expenditure on Education (x)



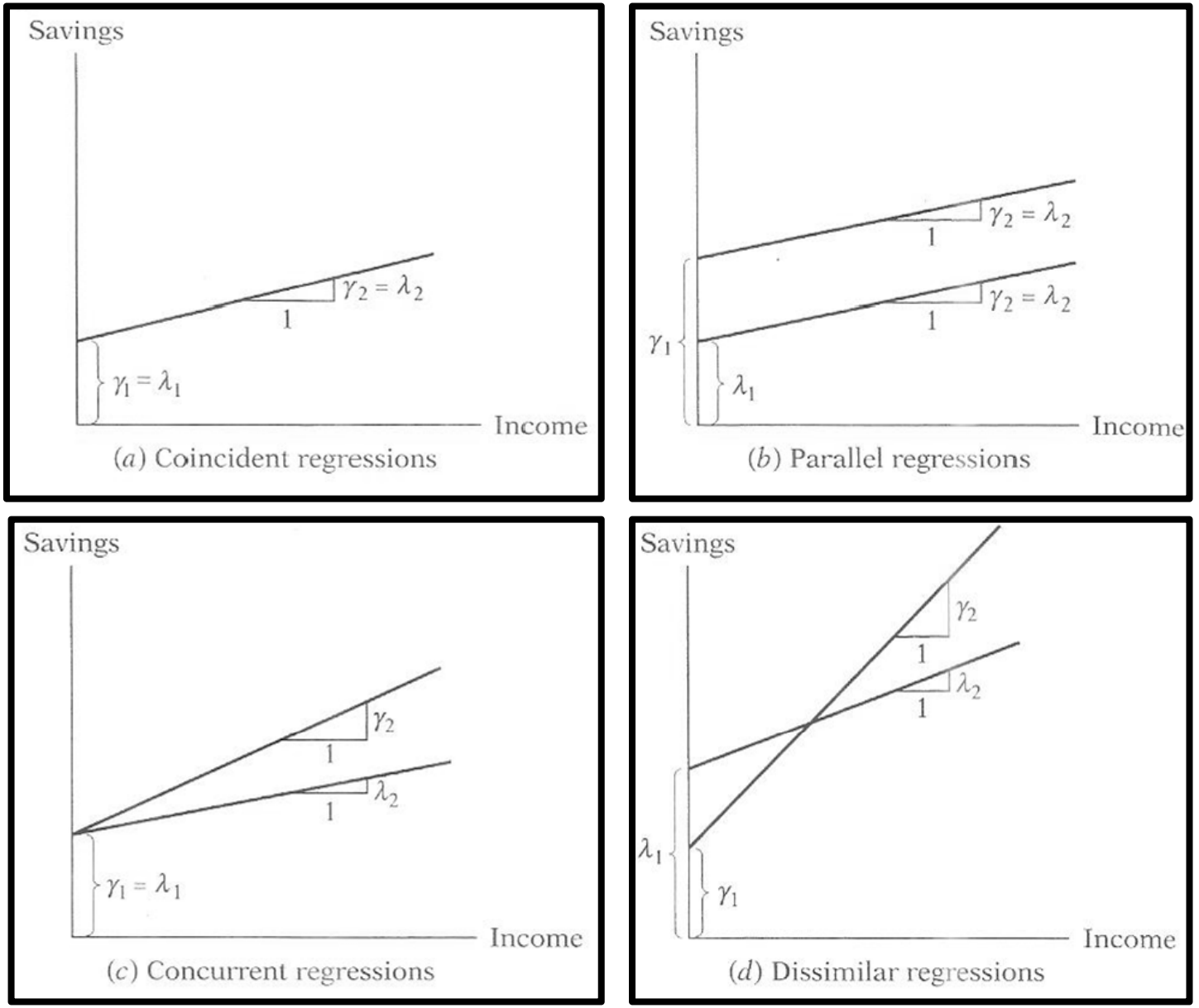
The Dummy Variable Alternative to the Chow Test

Table 9.2
Savings and Income Data, U.S., 1970-1995

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

The Dummy Variable Alternative to the Chow Test

Figure 9.3 Plausible Savings-Income Regression



The Dummy Variable Alternative to the Chow Test

The source of difference, if any, can be pinned down by pooling all the observations (26 in all) and running just one multiple regression as shown below:[10]

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \qquad (9.5.1)$$

where Y = savings
X = income
t = time
D = 1 for observations in 1982–1995
= 0, otherwise (i.e., for observations in 1970–1981)

Table 9.2 shows the structure of the data matrix.

To see the implications of Eq. (9.5.1), and, assuming, as usual, that $E(u_i) = 0$, we obtain:

Mean savings function for 1970–1981:

$$E(Y_t \mid D_t = 0, X_t) = \alpha_1 + \beta_1 X_t \qquad (9.5.2)$$

Mean savings function for 1982–1995:

$$E(Y_t \mid D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t \qquad (9.5.3)$$

The reader will notice that these are the same functions as Eqs. (8.7.1) and (8.7.2), with $\lambda_1 = \alpha_1$, $\lambda_2 = \beta_1$, $\gamma_1 = (\alpha_1 + \alpha_2)$, and $\gamma_2 = (\beta_1 + \beta_2)$. Therefore, estimating Eq. (9.5.1) is equivalent to estimating the two individual savings functions in Eqs. (8.7.1) and (8.7.2).

EXAMPLE 9.4 Structural Differences in the U.S. Savings-Income Regression

Before we proceed further, let us first present the regression results of model (9.5.1) applied to the U.S. savings–income data.

$$\begin{aligned} \hat{Y}_t &= \quad 1.0161 \quad + 152.4786 D_t + \ 0.0803 X_t - \ 0.0655 (D_t X_t) \\ \text{se} &= (20.1648) \quad\quad (33.0824) \quad\quad (0.0144) \quad\quad (0.0159) \\ t &= \ (0.0504)^{**} \quad\quad (4.6090)^{*} \quad\quad (5.5413)^{*} \quad (-4.0963)^{*} \end{aligned} \tag{9.5.4}$$

$$R^2 = 0.8819$$

where * indicates p values less than 5 percent and ** indicates p values greater than 5 percent.

Savings–income regression, 1970–1981:

$$\hat{Y}_t = 1.0161 + 0.0803 X_t \tag{9.5.5}$$

Savings–income regression, 1982–1995:

$$\begin{aligned} \hat{Y}_t &= \quad (1.0161 + 152.4786) + (0.0803 - 0.0655) X_t \\ &= 153.4947 + \quad 0.0148 X_t \end{aligned} \tag{9.5.6}$$

Interaction Effects Using Dummy Variables

Dummy variables are a flexible tool that can handle a variety of interesting problems. To see this, consider the following model:

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \qquad (9.6.1)$$

where 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

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i \qquad (9.6.2)$$

where the variables are as defined for model (9.6.1).

From Eq. (9.6.2), we obtain:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i \qquad (9.6.3)$$

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

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

EXAMPLE 9.5
Average Hourly Earnings in Relation to Education, Gender, and Race

Let us first present the regression results based on model (9.6.1). Using the data that were used to estimate regression (9.3.1), we obtained the following results:

$$\hat{Y}_i = -0.2610 - 2.3606D_{2i} - 1.7327D_{3i} + 0.8028X_i \quad \textbf{(9.6.4)}$$
$$t = (-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*} \quad (9.9094)^{*}$$
$$R^2 = 0.2032 \qquad n = 528$$

where * indicates p values less than 5 percent and ** indicates p values greater than 5 percent.

The reader can check that the differential intercept coefficients are statistically significant, that they have the expected signs (why?), and that education has a strong positive effect on hourly wage, an unsurprising finding.

As Eq. (9.6.4) shows, *ceteris paribus,* the average hourly earnings of females are lower by about $2.36, and the average hourly earnings of nonwhite non-Hispanic workers are also lower by about $1.73.

We now consider the results of model (9.6.2), which includes the interaction dummy.

$$\hat{Y}_i = -0.26100 - 2.3606D_{2i} - 1.7327D_{3i} + 2.1289D_{2i}D_{3i} + 0.8028X_i \quad \textbf{(9.6.5)}$$
$$t = (-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*} \quad (1.7420)^{**} \quad (9.9095)^{**}$$
$$R^2 = 0.2032 \qquad n = 528$$

where * indicates p values less than 5 percent and ** indicates p values greater than 5 percent.

As you can see, the two additive dummies are still statistically significant, but the interactive dummy is not at the conventional 5 percent level; the actual p value of the interaction dummy is about the 8 percent level. If you think this is a low enough probability, then the results of Eq. (9.6.5) can be interpreted as follows: Holding the level of education constant, if you add the three dummy coefficients you will obtain: -1.964 ($= -2.3605 - 1.7327 + 2.1289$), which means that mean hourly wages of nonwhite/non-Hispanic female workers is lower by about $1.96, which is between the value of -2.3605 (gender difference alone) and -1.7327 (race difference alone).

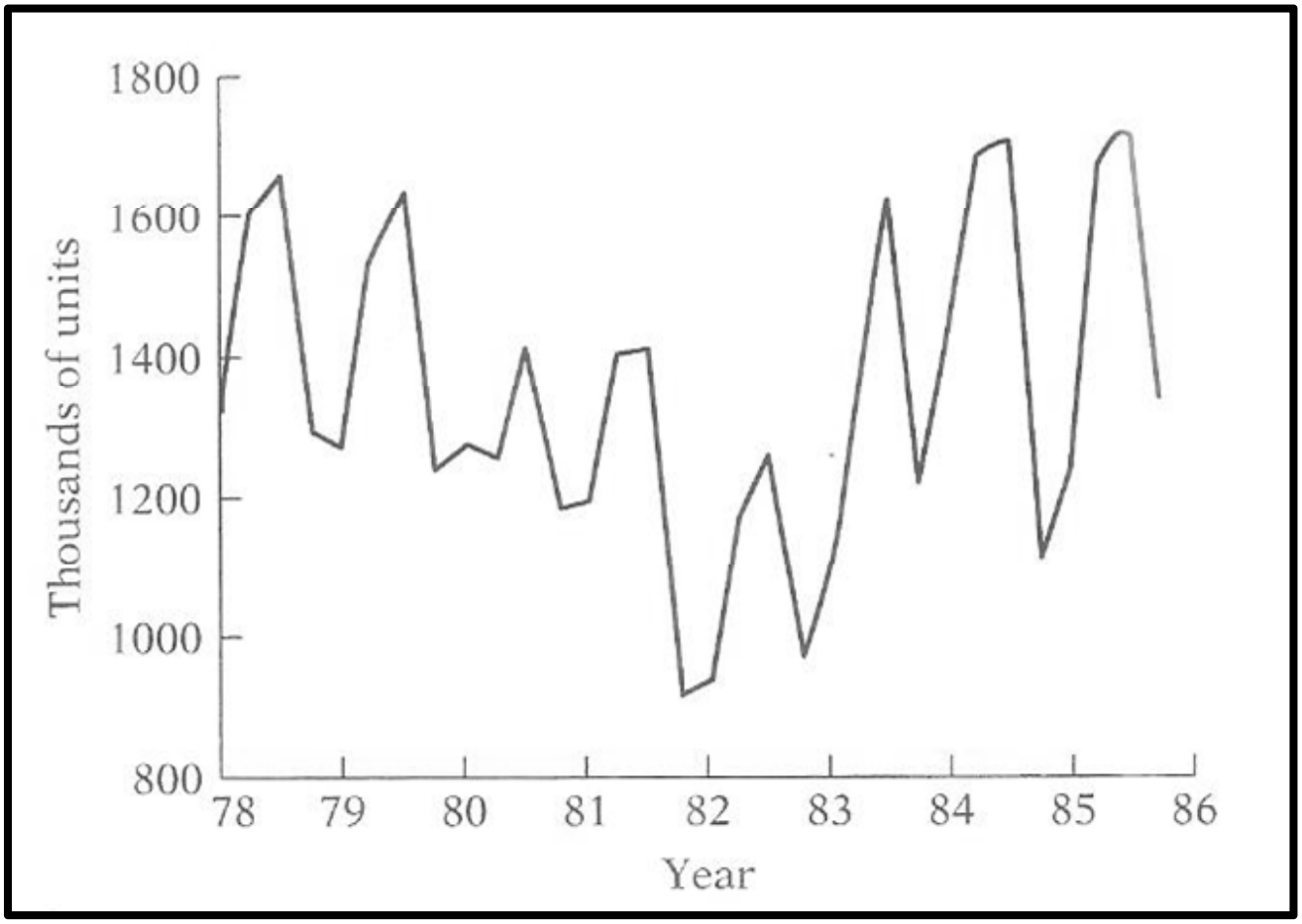
EXAMPLE 9.6 Seasonality in Refrigerator Sales

Table 9.4 U.S. Refrigerator Sales (thousands), 1978-1985 (quarterly)

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.

Figure 9.4 Sales of Refrigerators 1978-1985 (quarterly)



EXAMPLE 9.6 Seasonality in Refrigerator Sales

$$\hat{Y}_t = 1{,}222.125D_{1t} + 1{,}467.500D_{2t} + 1{,}569.750D_{3t} + 1{,}160.000D_{4t}$$
$$t = \quad (20.3720) \qquad (24.4622) \qquad (26.1666) \qquad (19.3364) \qquad \textbf{(9.7.2)}$$
$$R^2 = 0.5317$$

$$\hat{Y}_t = 1{,}222.1250 + 245.3750D_{2t} + 347.6250D_{3t} - 62.1250D_{4t}$$
$$t = \quad (20.3720)^* \qquad (2.8922)^* \qquad (4.0974)^* \qquad (-0.7322)^{**} \qquad \textbf{(9.7.3)}$$
$$R^2 = 0.5318$$

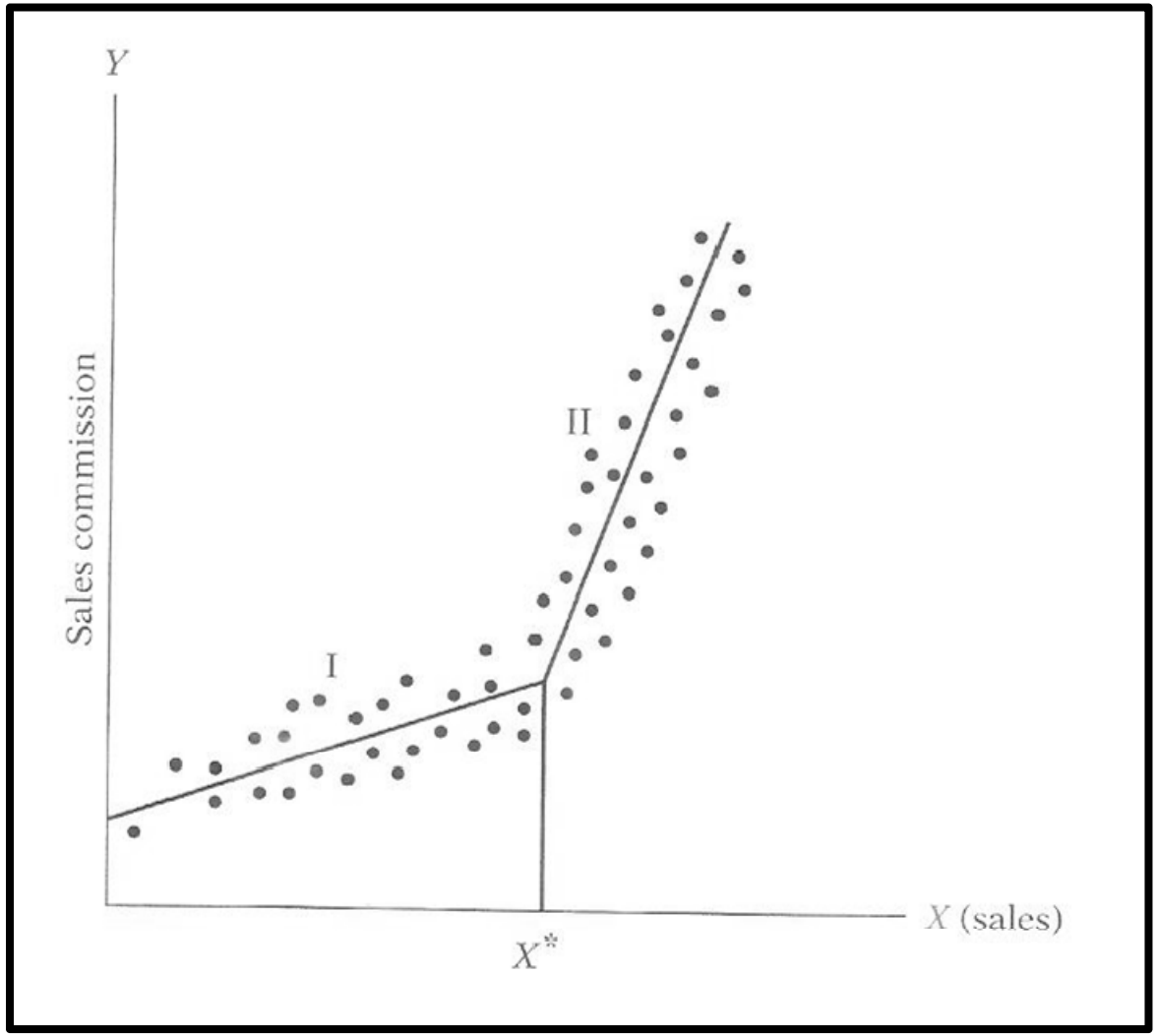
where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

$$\hat{Y}_t = 456.2440 + 242.4976D_{2t} + 325.2643D_{3t} - 86.0804D_{4t} + 2.7734X_t$$
$$t = \quad (2.5593)^* \qquad (3.6951)^* \qquad (4.9421)^* \qquad (-1.3073)^{**} \qquad (4.4496)^* \qquad \textbf{(9.7.4)}$$
$$R^2 = 0.7298$$

where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

Piecewise Linear Regression

Figure 9.5 Hypothetical relationship between sales commission and sale volume.



$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i \qquad (9.8.1)$$

where Y_i = sales commission
X_i = volume of sales generated by the sales person
X^* = threshold value of sales also known as a **knot** (known in advance)[17]
$D = 1$ if $X_i > X^*$
$= 0$ if $X_i < X^*$

Assuming $E(u_i) = 0$, we see at once that

$$E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i \qquad (9.8.2)$$

which gives the mean sales commission up to the target level X^* and

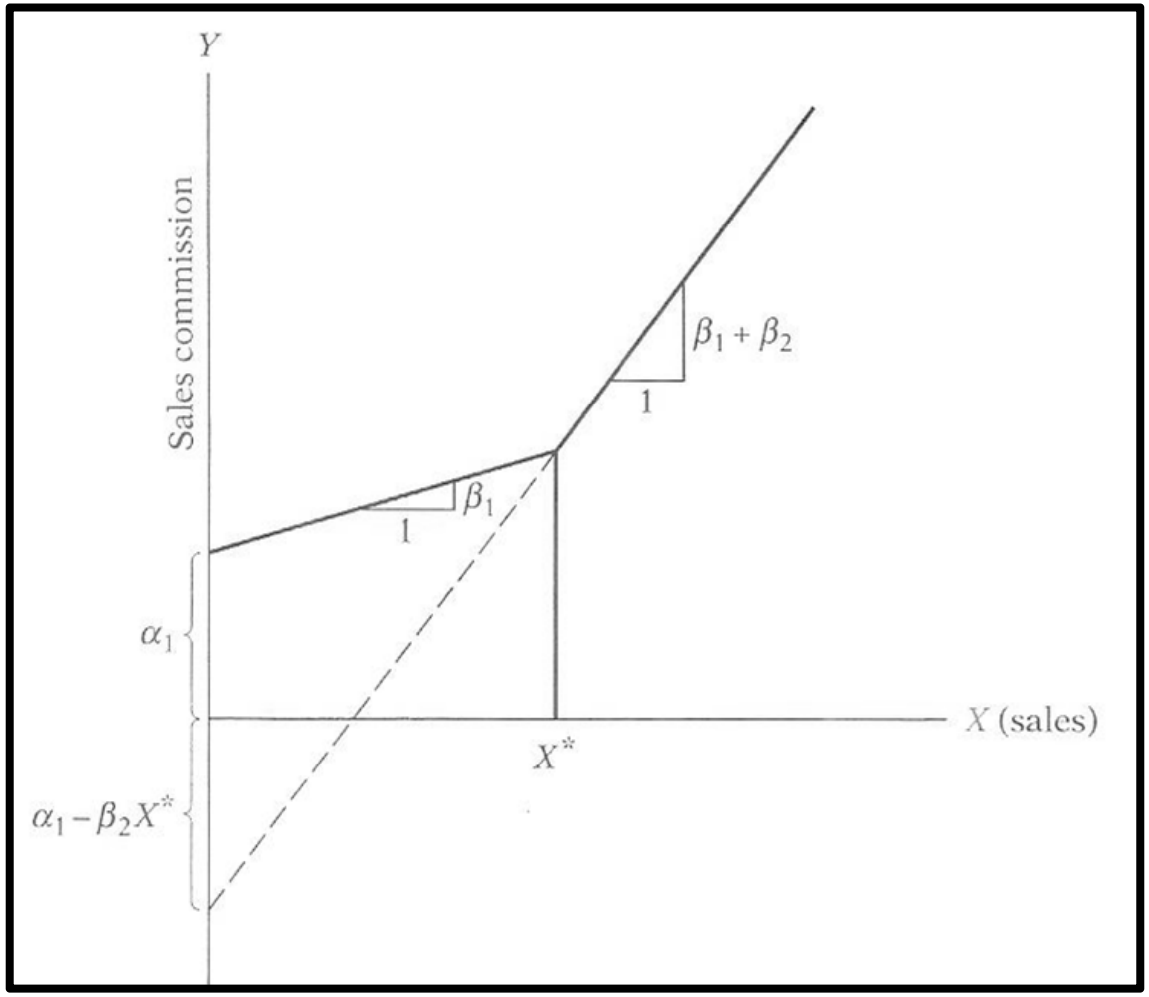
$$E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i \qquad (9.8.3)$$

which gives the mean sales commission beyond the target level X^*.

Thus, β_1 gives the slope of the regression line in segment I, and $\beta_1 + \beta_2$ gives the slope of the regression line in segment II of the piecewise linear regression shown in Figure 9.5. A test of the hypothesis that there is no break in the regression at the threshold value X^* can be conducted easily by noting the statistical significance of the estimated differential slope coefficient $\hat{\beta}_2$ (see Figure 9.6).

Incidentally, the piecewise linear regression we have just discussed is an example of a more general class of functions known as **spline functions.**[18]

Figure 9.6 Parameters of the piecewise linear regression



EXAMPLE 9.7 Total Cost 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

EXAMPLE 9.7 Total Cost in Relation to Output (Continued)

As an example of the application of the piecewise linear regression, consider the hypothetical total cost–total output data given in Table 9.6. We are told that the total cost may change its slope at the output level of 5,500 units.

Letting Y in Eq. (9.8.4) represent total cost and X total output, we obtain the following results:

$$\begin{aligned} \hat{Y}_i &= -145.72 \quad + 0.2791X_i + 0.0945(X_i - X_i^*)D_i \\ t &= \quad (-0.8245) \quad (6.0669) \quad (1.1447) \\ & \qquad R^2 = 0.9737 \qquad X^* = 5{,}500 \end{aligned} \tag{9.8.4}$$

As these results show, 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.

EXAMPLE 9.8 Logarithm of Hourly Wages in Relation to Gender

To illustrate Eq. (9.10.1), we use the data that underlie Example 9.2. The regression results based on 528 observations are as follows:

$$\widehat{\ln Y_i} = 2.1763 - 0.2437D_i \quad \textbf{(9.10.4)}$$
$$t = (72.2943)^* \quad (-5.5048)^*$$
$$R^2 = 0.0544$$

where * indicates p values are practically zero.

Taking the 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. Thus, the female workers' median hourly earnings are lower by about 21.94 percent compared to their male counterparts [(8.8136 − 6.8796)/8.8136].

Interestingly, we can obtain semielasticity for a dummy regressor directly by the device suggested by Halvorsen and Palmquist.[19] *Take the antilog (to base e) of the estimated dummy coefficient and subtract 1 from it and multiply the difference by 100.* (For the underlying logic, see Appendix 9.A.1.) Therefore, if you 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 ($D = 1$) median salary is lower than that of her male counterpart by about 21.63 percent, the same as we obtained previously, save the rounding errors.