Chapter 4. Indicator Variables.

4.1 : Intro

Indicator or dummy variables are variables that take only 2 values : 0 and 1. Normally 1 represents the presence of some attribute and 0 its absence.

4.2 Simple Application.

Consider the simplest case where we have a single independent variable x_{i1}, which is a dummy.

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \quad (*).$$

where $x_{i1} = \begin{cases} 0 & \text{when } i = 1, \dots, n_1. \\ 1 & \text{when } i = n_1+1, \dots n. \end{cases}$

and ε_i's are iid $N(0, \sigma^2)$

Let $\mu_1 = \beta_0$, $\mu_2 = \beta_0 + \beta_1$.

$(*)$ becomes $y_i = \begin{cases} \mu_1 + \varepsilon_i & , i = 1, \dots, n_1 \\ \mu_2 + \varepsilon_i & , i = n_1+1, \dots n. \end{cases}$

This is the model for the two-sample testing problem.

$$H: \mu_1 = \mu_2 \quad \text{vs} \quad A: \mu_1 \neq \mu_2.$$

The equivalent test would consist of testing

$$H: \beta_1 = 0 \quad \text{vs} \quad A: \beta_1 \neq 0$$

Notice that we have two means but only one indicator variable, the parameter of which is a difference of means.

Suppose we used a second indicator x_{i2}

$x_{i2} = 1$ if $i = 1, \ldots, n_1$ 0 otherwise.

$$y = X\beta + \varepsilon.$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1_{n_1} & 0 & 1_{n_1} \\ 1_{n-n_1} & 1_{n-n_1} & 0 \end{bmatrix}$$

$$X = \begin{array}{c} n_1 \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right. \\ n-n_1 \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right. \end{array} \left[\begin{array}{ccc} 1 & & 1 \\ 1 & 0 & 1 \\ \vdots & & \vdots \\ 1 & & 1 \\ \hline 1 & 1 & \\ \vdots & \vdots & 0 \\ 1 & 1 & \end{array} \right] \left. \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \right\} n$$

Which such a design matrix X can be used. $X'X$ is not non-singular and parameter estimates $\hat{\beta}$ must be based on a generalized inverse $(X'X)^-$ of $(X'X)$. Therefore $\hat{\beta} = (X'X)^- X'y$ is not unique. But testing of $H: \mu_1 = \mu_2$ vs $A: \mu_1 \neq \mu_2$ is still possible.

4.3 Polychotomous variables.

Variables taking a finite number of values — but more than two - may be called polychotomous variables. Such polychotomous variables are sometimes called factors and their values are called levels.

Consider a case where we have a single factor with p levels.

For level 1, let there be n_1 observations $y_1, \dots y_{n_1}$. For level 2, let there be $n_2 - n_1$ observations.

$$y_i = \begin{cases} \mu_1 + \varepsilon_i & i = 1, \dots, n_1 \\ \mu_2 + \varepsilon_i, & i = n_1 + 1, \dots n_2 \\ \vdots & \\ \mu_p + \varepsilon_i, & i = n_{p-1} + 1, \dots n_p \end{cases} \quad (*).$$

$\varepsilon_i \sim N(0, \sigma^2)$

Notice that we have $N_1 = n_1$ observations with mean μ_1. $N_2 = n_2 - n_1$, with mean μ_2.

$\cdots$ $N_p = n_p - n_{p-1}$. with mean μ_p.

If we wished to see if our polychotomous variables affected the dependent variable at all, we would test:

$H: \mu_1 = \mu_2 = \cdots = \mu_p$

$A: \mu_i \neq \mu_j$ for at least one pair i, j with $i \neq j$

Let

$$\boxed{\beta_0 = \mu_1, \quad \beta_j = \mu_{j+1} - \mu_j \quad \text{for } j = 1, 2, \cdots p-1.} \quad **$$

$$x_{ij} = \begin{cases} 1 & \text{if } i = n_j + 1, \cdots n_{j+1}. \\ 0 & \text{otherwise.} \end{cases}$$

Then $\circledast$ becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i$$

This is a multiple regression model.

Notice that the reparameterization ** to convert ⊛ into a regression model format is far from unique.

Notice also that here too, as in Section 4.2, and for much the same reasons we have one fewer indicator variable than the number of means. If we had more than one factor, then for each factor we would usually need one less indicator variable than number of levels. Obviously, the hypothesis (4.5) is equivalent to the hypothesis $\beta_j = 0$ for $j = 1, \ldots, p-1$.

Service	Psychometric Scores (QUAL)
Public	61.59, 79.19, 68.89, 72.16, 70.66, 63.17, 53.66, 68.69, 68.75, 60.52, 68.01, 73.06, 55.93, 74.88, 62.55, 69.90, 66.61, 63.80, 45.83, 64.48, 58.11, 73.24, 73.24, 69.94
Private Non-profit	76.77, 68.33, 72.29, 69.48, 59.26, 67.16, 71.83, 64.63, 78.31, 61.48
Private	71.77, 82.92, 72.26, 71.75, 67.95, 71.90

EXHIBIT 4.3: Measures of Quality for Agencies Delivering Transportation for the Elderly and Handicapped

SOURCE: Slightly modified version of data supplied by Ms. Claire McKnight of the Department of Civil Engineering, City University of New York.

Example 4.2

Transportation services for the elderly and handicapped are provided by public, private not-for-profit and private for-profit agencies (although in each case, financial support is mainly through public funding). To see if the quality of the services provided under the three types of ownership was essentially the same, a scale measuring quality was constructed using psychometric methods from results of questionnaires administered to users of such services. Each of several services in the State of Illinois was scored using this scale. Exhibit 4.3 shows the score for each agency.

QUAL	X_1	X_2	QUAL	X_1	X_2
61.59	0	0	58.11	0	0
79.19	0	0	73.23	0	0
68.89	0	0	73.12	0	0
72.16	0	0	69.94	0	0
70.66	0	0	76.77	1	0
63.17	0	0	68.33	1	0
53.70	0	0	72.29	1	0
68.69	0	0	69.48	1	0
68.75	0	0	59.26	1	0
60.52	0	0	67.16	1	0
68.01	0	0	71.89	1	0
73.62	0	0	64.63	1	0
55.93	0	0	78.31	1	0
74.88	0	0	61.48	1	0
62.58	0	0	71.77	1	1
69.90	0	0	82.92	1	1
66.61	0	0	72.26	1	1
63.80	0	0	71.75	1	1
45.83	0	0	67.95	1	1
65.48	0	0	71.90	1	1

EXHIBIT 4.4: Values of x_{i1}'s and x_{i2}'s and Corresponding Values of QUAL

The dependent variable QUAL and the independent variables X_1 and X_2 are shown in Exhibit 4.4. Notice that the definition of the independent variables is slightly different from that given by (4.6), although the latter would have worked about as well. Here it made sense to first distinguish between private and public and then between for-profit and not-for-profit. Portions of the output from a least squares package are shown in Exhibit 4.5. Since we wish to test the hypothesis that coefficients of $X1$ and $X2$ are both zero against the alternative that at least one is not equal to zero, the value of the appropriate statistic is the F-value 2.51, which shows that we can reject the hypothesis at a 10 per cent level but not at a 5 per cent level. We also see that the least squares estimate for the mean level of quality of public services (since this level corresponds to $X1 = 0$ and $X2 = 0$) is about 66.18. For private non-profit systems the estimated mean quality index rises by about 2.78 and the quality index for for-profit organizations rises an additional 4.13. However, neither factor is significant at any reasonable level.

Given the nature of the results obtained, one might be tempted to conjecture that if more privately run services were represented in the data set, stronger results might have been obtained. If this problem had been brought to us by a client, we would then have recommended that they increase the size of the data-set. While on the subject of making recommendations to clients, we would also suggest that the client look into the possibility of finding other independent variables (e.g., was the driver a volunteer? Was the transportation service the main business of the provider? etc.), which by reducing s might help achieve significance. ■

Source	DF	Sum of Squares	Mean Square	F value	p-value
MODEL	2	243.81	121.91	2.511	0.0950
ERROR	37	1796.58	48.56		
C. TOTAL	39	2040.40			

Variable	b_j	s.e.(b_j)	$t(b_j)$	p-value
Intercept	66.18	1.422	46.5	0.0001
x_1	2.78	2.623	1.060	0.2963
x_2	4.13	3.598	1.148	0.2583

$$R^2 = .1195 \quad R_a^2 = .0719 \quad s = 6.968$$

EXHIBIT 4.5: Analysis of Variance Table and Parameter Estimates for Quality Data

4.4 Continuous and Indicator Variables

Mixing continuous and dichotomous or polychotomous independent variables presents no particular problems. In the case of a polychotomous variable, one simply converts it into a set of indicator variables and adds them to the variable list.

Example 4.3

The house-price data of Exhibit 2.2, p. 32, were collected from three neighborhoods or zones; call them A, B and C. For these three levels we need to use two dummy variables. We chose

$$L1 = \begin{cases} 1 & \text{if property is in zone A} \\ 0 & \text{otherwise} \end{cases}$$

$$L2 = \begin{cases} 1 & \text{if property is in zone B} \\ 0 & \text{otherwise.} \end{cases}$$

Obviously, if $L1 = 0$ and $L2 = 0$, the property is in C. Data for $L1$ and $L2$ are also presented in Exhibit 2.2. A portion of the output using these variables is given in Exhibit 4.6. As the output shows, if two identical houses were in zones A and C, the former would cost an estimated $2700 more and a property in zone B would cost $5700 more than an identical one in zone C. Notice that simply comparing the means of house values in two

Variable	b_j	s.e.(b_j)	$t(b_j)$	p-value
Intercept	16.964	4.985	3.403	0.0039
FLR	0.017	.0032	5.241	0.0001
RMS	3.140	1.583	1.984	0.0659
BDR	-6.702	1.807	-3.708	0.0021
BTH	2.466	2.462	1.002	0.3323
GAR	2.253	1.451	1.553	0.1412
LOT	0.288	0.127	2.258	0.0393
FP	5.612	3.059	1.835	0.0865
ST	10.017	2.318	4.320	0.0006
L1	2.692	2.867	0.939	0.3626
L2	5.692	2.689	2.117	0.0514

$$R^2 = .9258 \quad R_a^2 = .8764 \quad s = 4.442$$

EXHIBIT 4.6: Output for House Price Data When L1 and L2 Are Included

areas would give us a comparison of house prices in the areas, not the price difference between identical houses. The two comparisons would be quite different if, say, on the average, houses in one of the two areas were much larger than in the other. For this reason, had we included only $L1$ and $L2$ in the model, and no other variables, the meaning of the coefficients would be quite different.

If we wished to test if location affects property values, we would test the hypothesis that the coefficients of $L1$ and $L2$ are both zero against the alternative that at least one of the coefficients is non-zero. The value of the F test statistic turns out to be 2.343 for which the p-value is .13. ■

4.5 Broken Line Regression

Exhibit 4.9 illustrates a plot of points which would appear to require two lines rather than a single straight line. It is not particularly difficult to fit such a 'broken line' regression. Let us assume the break occurs at the known value x of the independent variable and define

$$\delta_i = \begin{cases} 1 & \text{if } x_{i1} > x \\ 0 & \text{if } x_{i1} \le x. \end{cases}$$

Then the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - x)\delta_i + \epsilon_i \tag{4.8}$$

suffices, as can be readily verified. Situations when x is treated as an unknown can be handled using nonlinear regression (see Appendix C, particularly Example C.4, p. 313).

Obs	Country	LIFE	INC	Obs	Country	LIFE	INC
1	AUSTRALIA	71.0	3426	52	CAMEROON	41.0	165
2	AUSTRIA	70.4	3350	53	CONGO	41.0	281
3	BELGIUM	70.6	3346	54	EGYPT	52.7	210
4	CANADA	72.0	4751	55	EL SALVADOR	58.5	319
5	DENMARK	73.3	5029	56	GHANA	37.1	217
6	FINLAND	69.8	3312	57	HONDURAS	49.0	284
7	FRANCE	72.3	3403	58	IVORY COAST	35.0	387
8	WEST GERMANY	70.3	5040	59	JORDAN	52.3	334
9	IRELAND	70.7	2009	60	SOUTH KOREA	61.9	344
10	ITALY	70.6	2298	61	LIBERIA	44.9	197
11	JAPAN	73.2	3292	62	MOROCCO	50.5	279
12	NETHERLANDS	73.8	4103	63	PAPUA	46.8	477
13	NEW ZEALAND	71.1	3723	64	PARAGUAY	59.4	347
14	NORWAY	73.9	4102	65	PHILLIPPINES	51.1	230
15	PORTUGAL	68.1	956	66	SYRIA	52.8	334
16	SWEDEN	74.7	5596	67	THAILAND	56.1	210
17	SWITZERLAND	72.1	2963	68	TURKEY	53.7	435
18	BRITAIN	72.0	2503	69	SOUTH VIETNAM	50.0	130
19	UNITED STATES	71.3	5523	70	AFGHANISTAN	37.5	83
20	ALGERIA	50.7	430	71	BURMA	42.3	73
21	ECUADOR	52.3	360	72	BURUNDI	36.7	68
22	INDONESIA	47.5	110	73	CAMBODIA	43.7	123
23	IRAN	50.0	1280	74	CENTRAL AFRICAN REPUBLIC	34.5	122
24	IRAQ	51.6	560				
25	LIBYA	52.1	3010	75	CHAD	32.0	70
26	NIGERIA	36.9	180	76	DAHOMEY	37.3	81
27	SAUDI ARABIA	42.3	1530	77	ETHIOPIA	38.5	79
28	VENEZUELA	66.4	1240	78	GUINEA	27.0	79
29	ARGENTINA	67.1	1191	79	HAITI	32.6	100
30	BRAZIL	60.7	425	80	INDIA	41.2	93
31	CHILE	63.2	590	81	KENYA	49.0	169
32	COLOMBIA	45.1	426	82	LAOS	47.5	71
33	COSTA RICA	63.3	725	83	MADAGASCAR	36.0	120
34	DOMINICAN REP.	57.9	406	84	MALAWI	38.5	130
35	GREECE	69.1	1760	85	MALI	37.2	50
36	GUATEMALA	49.0	302	86	MAURITANIA	41.0	174
37	ISRAEL	71.4	2526	87	NEPAL	40.6	90
38	JAMAICA	64.6	727	88	NIGER	41.0	70
39	MALAYSIA	56.0	295	89	PAKISTAN	51.2	102
40	MEXICO	61.4	684	90	RWANDA	41.0	61
41	NICARAGUA	49.9	507	91	SIERRA LEONE	41.0	148
42	PANAMA	59.2	754	92	SOMALIA	38.5	85
43	PERU	54.0	334	93	SRI LANKA	65.8	162
44	SINGAPORE	67.5	1268	94	SUDAN	47.6	125
45	SPAIN	69.1	1256	95	TANZANIA	40.5	120
46	TRINIDAD	64.2	732	96	TOGO	35.0	160
47	TUNISIA	51.7	434	97	UGANDA	47.5	134
48	URUGUAY	68.5	799	98	UPPER VOLTA	31.6	62
49	YUGOSLAVIA	67.7	406	99	SOUTH YEMEN	42.3	96
50	ZAMBIA	43.5	310	100	YEMEN	42.3	77
51	BOLIVIA	49.7	193	101	ZAIRE	38.8	118

EXHIBIT 4.7: Data on Per-Capita Income (in Dollars) and Life Expectancy
SOURCE: Leinhardt and Wasserman (1979), from the *New York Times* (September, 28, 1975, p. E-3). Reproduced with the permission of the *New York Times*.

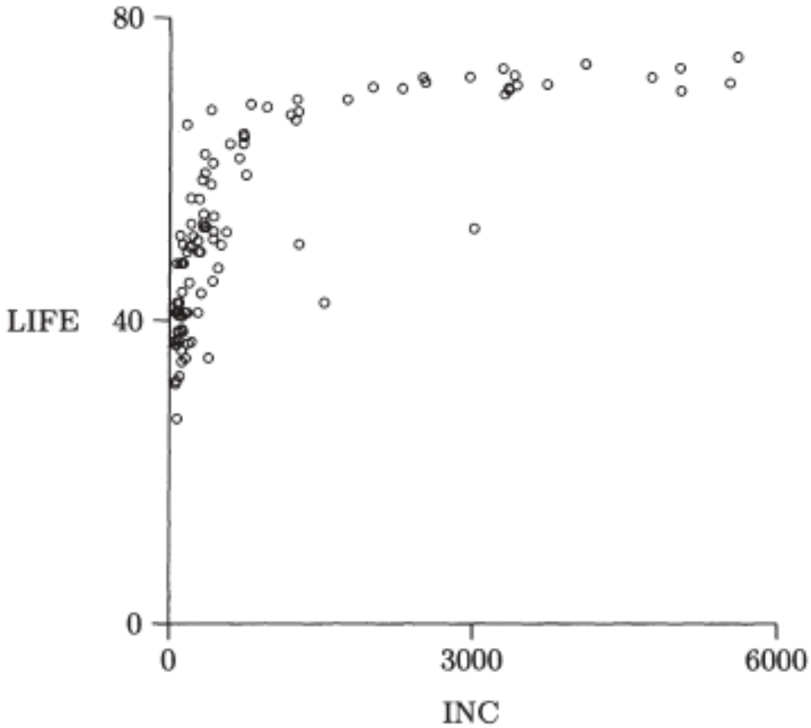

EXHIBIT 4.8: Plot of Life Expectancy Against Per-Capita Income

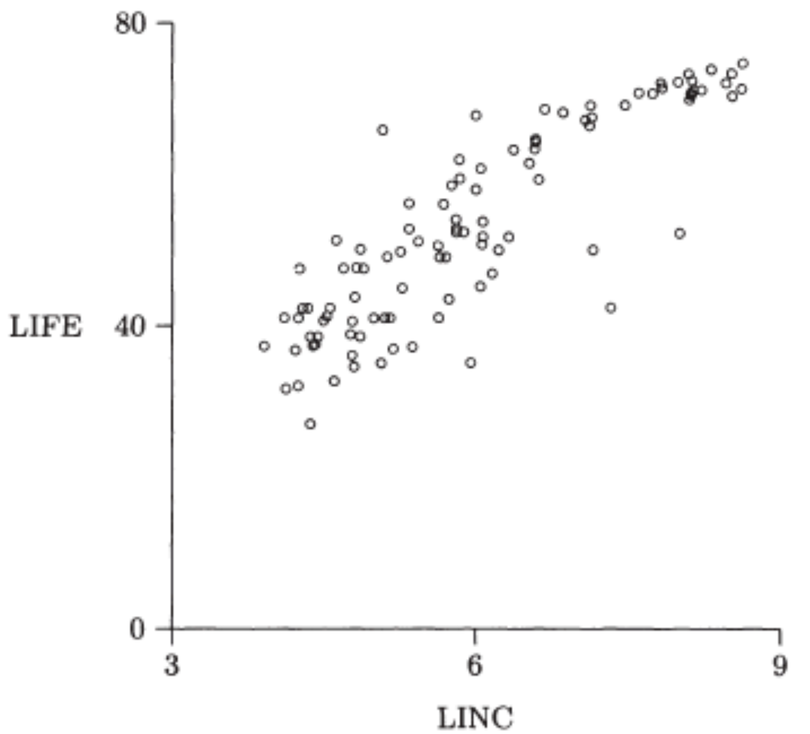

EXHIBIT 4.9: Plot of Life Expectancy Against Log of Per-Capita Income

Example 4.4

Usually poorer countries (i.e., those with lower per capita incomes) have lower life expectancies than richer countries. Exhibit 4.7 gives life expectancies (LIFE) and per capita incomes (INC) in 1974 dollars for 101 countries in the early 70's. Exhibit 4.8 shows a plot which is difficult to read. Taking logarithms of income 'spreads out' the low income points and (see Exhibit 4.9) we discern a pattern that seems to consist of two separate lines: one for the poorer countries, where LIFE increases rapidly with LINC (= log(INC)), and another for the richer countries, where the rate of growth of life expectancy with LINC is much smaller. Therefore, we fitted an equation of the form (4.8) with $\delta_i = 1$ if LINC > 7 and $\delta_i = 0$ otherwise, and obtained

$$\begin{array}{ccccccc} \text{LIFE} & = & -2.40 & + & 9.39\ \text{LINC} & - & 3.36\,[\delta_i(\text{LINC} - 7)] \\ & & (4.73) & & (.859) & & (2.42) \end{array} \tag{4.9}$$

$$(R^2 = .752, s = 6.65)$$

where, as before, the parenthetic quantities are standard errors. The 7 was found by inspecting Exhibit 4.9. We shall return to this example in future chapters. ■

4.6 Indicators as Dependent Variables

While it is not desirable to use dichotomous dependent variables in a linear least squares analysis (typically logit, probit or contingency table analysis is used for this purpose), if we are willing to aggregate our data, least squares analysis may still be used. The example below illustrates such a case. Another case is illustrated in Chapter 9.

Example 4.5

An interesting problem for political scientists is to determine how a particular group of people might have voted for a particular candidate. Typically such assessments are made using exit polls. However, with adequate data, regression procedures might be used to obtain estimates.

Consider the data of Exhibit 4.10 in which the columns Garcia, Martinez and Yanez give the total votes for each of those candidates. (Note that votes for the three candidates may not add to the total turnout because of write-in votes, spoilt ballots, etc.) Let p_L be the probability that a Latino casts a valid vote for (say) Garcia and p_N the probability that a non-Latino casts a valid vote for him. If LATV_i and NONLV_i are, respectively, the total Latino and non-Latino votes cast in each precinct i, the expected number of votes for Garcia is

$$p_L \, \text{LATV}_i + p_N \, \text{NONLV}_i.$$

Since we have the total vote count for Garcia, p_L and p_N can be readily estimated by least squares and we obtain

$$\begin{array}{rcccl} \text{GARCIA} & = & .37\ \text{LATV} & + & .64\ \text{NONLV} \\ & & (.043) & & (.052) \end{array}$$

$$(R^2 = .979, s = 18.9).$$

Therefore, we estimate that roughly 37 per cent of the Latinos voted for Garcia and about 64 per cent of the others voted for him. ∎

Variables such as all those in Exhibit 4.10 will be called *counted* variables since they are obtained by counting. We might prefer to use as dependent variable the proportion of all voters who voted for Garcia. Such a variable will be called a *proportion of counts*. Both counted variables and proportions of counts usually require special care, as we shall see in Chapters 6 and 9.

Pr.	LATV	NONLV	TURNOUT	GARCIA	MARTINEZ	YANEZ
1	114	78	192	95	59	15
2	143	100	243	120	74	41
3	105	91	196	120	58	18
4	176	97	273	138	71	26
5	169	141	310	143	85	48
6	190	110	300	158	97	29
7	1	305	306	206	15	11
8	190	132	322	128	125	43
9	120	62	182	79	70	27
10	186	224	410	169	158	49
11	152	85	237	105	81	24
12	164	89	253	124	60	29
13	168	64	232	111	89	13
14	75	157	232	143	27	25
15	177	60	237	98	87	21
16	140	121	261	128	92	40
17	178	115	293	150	66	52
18	157	85	242	108	78	31
19	76	124	200	124	24	14
20	120	59	179	73	70	11
21	84	65	149	52	65	12
22	119	92	211	123	55	15
23	172	144	316	136	127	30
24	87	59	146	118	21	7
25	134	59	193	114	55	20
26	137	60	197	83	67	39
27	167	131	298	147	112	42

EXHIBIT 4.10: Votes from Chicago's Twenty-Second Ward by Precinct (Pr.)
SOURCE: Ray Flores, The Latino Institute, Chicago.