International Journal of
Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2018; 3(3): 36-47
© 2018 Stats & Maths
www.mathsjournal.com
Received: 06-03-2018
Accepted: 07-04-2018

M Rajani
Department of Statistics,
S.V. University, Tirupati,
Andhra Pradesh, India

S Damodharan
Department of Statistics,
S.V. University, Tirupati,
Andhra Pradesh, India

K Murali
Department of Statistics,
S.V. University, Tirupati,
Andhra Pradesh, India

M Subbarayudu
Department of Statistics,
S.V. University, Tirupati,
Andhra Pradesh, India

Correspondence
M Rajani
Department of Statistics,
S.V. University, Tirupati,
Andhra Pradesh, India

Statistical modeling: Analysis of population with different categories

M Rajani, S Damodharan, K Murali and M Subbarayudu

Abstract
Regression models with dummy and effect coding methods are used to analyze the census population (total, male, female) of India with three age groups and in four southern states of India also analyzed the sex ratio/ urban rural ration in four southern states of India.

Keywords: Dummy coding, effect coding, regression

Introduction

The purpose of this section is to analyze the Indian census population (1901-2011) with three age groups and in four southern states of India. Also analyzed the sex ratio / urban rural ratio in four southern states of India. Here we consider age groups and states as qualitative variables.

Qualitative variables usually indicate the presence or absence of a quality or an attribute such as male or female, black or white, catholic or non catholic, citizen or non citizen. A regression model which contains explanatory variables that are exclusively qualitative in nature is known as analysis of variance (ANOVA) model and a model which contains the explanatory variables as an admixture of both quantitative and qualitative variables is known as analysis of covariance (ANCOVA) model.

The use of categorical variables as independent variables in the regression model involves the application of coding methods. Coding methods refer to ways in which membership in a given group can be represented in mutually exclusive and exhaustive manner. Any qualitative variable with k categories or classes can be represented by creating (k-1) dummy variables that takes on numerical values. This process involves assigning one numerical value which is called a code to all subjects of a particular group and a different numerical value to all those of the other groups to convert the qualitative variables into quantitative variables to run the regression.

Aim

(i) To find the relation between population (total, male, female) on different age groups in India.

(ii) To find the relation between population (total, male, female) on four southern states of India.

(iii) To find the relation between sex ratio / urban-rural ratio on four southern states of India.

(iv) To test the significant difference in population (total, male, female) in India between census years and between age groups in India.

(v) To test the significant difference in population (total, male, female) between census years and between four southern states of India (Andhra Pradesh, Tamil Nadu, Kerala and Karnataka).

(vi) To test significant difference in sex ratio / urban-rural ratio between census years and between four southern states of India (Andhra Pradesh, Tamil Nadu, Kerala and Karnataka).

Methodology

To achieve the aims (i), (ii), & (iii), we apply the statistical regression model with categorical variables as independent variables using dummy and effect coding methods. To achieve the aims (iv), (v) & (vi) we apply two way analysis of variance.

Dummy coding

Dummy variables are used as a classifying device in that they divide the entire sample into various groups based on qualities and implicitly allow to run the individual regressions for each sub group. Dummy coding method represents a group membership with dummy variables that take on values 0 and 1. That is membership in a particular group is coded as one where as non membership in a group is coded as zero. However assignment of 1 and 0 values to categories is arbitrary. The category to which the value zero is assigned is often referred as base, bench mark, control, comparison or omitted category. The number of dummy variables must be less than the number of categories or classifications of each qualitative variable to avoid dummy variable trap. The coefficients attached to the dummy variables must always be interpreted in relation to the base or reference group that assign value zero. If a model has several qualitative variables with several classes introduction of dummy variables can be consuming a large number of degrees of freedom. Dummy explanatory variables are denoted by the symbol ‘D’ rather than by the usual symbol ‘X’ in regression model to emphasize that we are dealing with qualitative variables.

The regression model involving one qualitative variable as independent variable with k categories or classes, using dummy coding can be represented as

$$Y_{ij} = \alpha_0 + \sum_{j=1}^{k-1} \alpha_j D_{ij} + \epsilon_{ij} \qquad (1)$$

Where

Y_{ij} = The score on the dependent variable for subject i in group j

α_0 = The intercept that represents the mean of the group coded 0 on all the dummy variables

k = number of categories or classifications of dummy independent variable

α_j = The regression coefficient associated with j^{th} group, it represents the difference between the mean of the group coded 1 on the corresponding dummy variable and the mean of the group coded 0 on all the dummy variables

D_{ij} = The numerical value of dummy variable assigned to subject i in the j^{th} group

ϵ_{ij} = The error associated with i^{th} subject in the j^{th} group

Effect coding

Effect coding is appropriate where each group is compared with entire set of groups rather than with a reference group. That is effect coding is useful in testing the effect of a treatment defined as the deviation between treatment mean and the grand mean. In effect coding the dummy variables takes the values 1, 0 and -1. The coding method used for effect coding is similar to that used for dummy coding except for the way in which the reference group is identified. Using dummy coding the reference group is coded as 0 but in the effect coding it is coded as -1. A regression model involving one qualitative variable as independent variable with k categories or classes using effect coding can be represented as

$$Y_{ij} = \beta_0 + \sum_{j=1}^{k-1} \beta_j E_{ij} + \epsilon_{ij} \qquad (2)$$

where

Y_{ij}= The score on the dependent variable Y for subject i in group j

β_0 = The intercept that represents the grand mean of the dependent variable for all groups

k = number of categories of the dummy independent variable

β_j = Regression coefficient associated with j^{th} group, it represents the difference between the mean of the group coded 1 on the corresponding dummy variable and the grand mean of all groups

E_{ij} = The numerical value of dummy variable assigned to subject i in the j^{th} group

ϵ_{ij} = The error associated with the i^{th} subject in the j^{th} group

R^2 can be interpreted in terms of the proportion of variance in the dependent variable that is accounted for by the categorical independent variable.

Example for dummy coding and effect coding

Hussain Alkharusi (2012) [1] explained the use of categorical variables in regression analysis through dummy and effect coding methods with an example. In dummy coding method it is observed that the test of significance of a given regression coefficient is equivalent to a test of difference between the mean of the group associated with the regression coefficient and the mean of the reference group. In effect coding method each group is compared with the entire set of groups rather than with a reference group. Here the test for significance of regression coefficient is equivalent to testing the significance of the treatment effect. The two methods, dummy coding and effect coding gives entirely different interpretations regarding regression coefficients. But in both the methods the values of R^2 is same.

In studying the relationship between k groups of independent categorical variable and the corresponding scores of dependent variable, (k-1) dummy variables D_1, D_2, ….., D_{k-1} are needed to run the regression with dummy coding method. Similarly (k-1) dummy variables E_1, E_2, …..E_{k-1} are needed to run the regression with effect coding method. Table (4.4.1) shows the dummy coding and effect coding of independent categorical variable. In dummy coding method the subjects in the 1st group have been coded as 1 for D_1 and 0 for D_2, D_3…., D_{k-1}; those in the 2nd group have been coded as 1 for D_2 and 0 for others; those in the 3rd group have been coded as 1 for D_3 and 0 for others ; ..….; those in the kth group have been coded as 0 for all D_1, D_2, ….., D_{k-1}. And in effect coding method scores of first group receives a 1 for E_1 and 0 for others; scores of second group receive a 1 for E_2 and 0 for others; scores of third group receives a 1 for E_3 and 0 for others; …. ; scores of kth group receives -1 for all E_1, E_2, ….., E_{k-1}.

Table 1: Example for dummy and effect coding for the data of k groups

Groups of categorical independent variable	Observations on dependent variable	Dummy Coding						Effect Coding					
		D_1	D_2	.	.	.	D_{k-1}	E_1	E_2	.	.	.	E_{k-1}
1st group	y_{11}	1	0				0	1	0				0
	y_{21}	1	0				0	1	0				0
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	$y_{n_1 1}$	1	0				0	1	0				0
2nd group	y_{12}	0	1				0	0	1				0
	y_{22}	0	1				0	0	1				0
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	$y_{n_2 2}$	0	1				0	0	1				0
.	.		.						.				
.	.		.						.				
.	.		.						.				
kth group	y_{1k}	0	0				0	-1	-1				-1
	y_{2k}	0	0				0	-1	-1				-1
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	.	.	.				.	.	.				.
	$y_{n_k k}$	0	0				0	-1	-1				-1

Empirical Analysis

For the propose of empirical analysis, the population data with different categories namely total, male, female, urban, rural related to India as well as four southern states of India namely Andhra Pradesh, Tamil Nadu, Kerala and Karnataka for twelve census years (1901 – 2011) also the population of India in three age group intervals 0 – 14, 15 – 59 and 60+ is extracted from census of India, Office of Registrar general,India.(Source:http://www.censusindia.gov.in/2011census/PCA/A2_Data_Tables/00%20A%202-India.pdf).

The data along with dummy and effect coding methods is presented in tables (2) – (4) and (5) – (7) respectively to find the relation between population (total, male, female) on three different age groups 0 – 14, 15 – 59 and 60 + of India; tables (8) – (10) and (11) – (13) to find the relation between population (total, male, female) on four southern states of India; tables (14), (15) / (16), (17) to find the relation between sex ratio / urban rural ratio on four southern states of India. Note that, since there are three age groups we consider two dummy variables and there are four southern states, we consider three dummy variables as independent variables. Later regression equations are presented for each of the tables from (2) - (17) and the corresponding interpretations are drawn.

Dummy coding data for three age groups

Table 2

Age group	Total pop	D_1	D_2
0-14	920.22	1	0
	970.55	1	0
	985.17	1	0
	1115.92	1	0
	1220.47	1	0
	1354.09	1	0
	1805.24	1	0
	2296.79	1	0
	2712.82	1	0
	3157.15	1	0
	3692.71	1	0
	3571.18	1	0
15-59	1344.58	0	1
	1421.79	0	1
	1397.34	0	1
	1562.29	0	1
	1813.18	0	1
	2054.60	0	1
	2341.10	0	1
	2861.40	0	1
	3662.65	0	1
	4697.63	0	1
	5945.37	0	1
	7566.06	0	1
60+	121.58	0	0
	131.09	0	0
	133.20	0	0
	114.38	0	0
	156.14	0	0
	205.82	0	0
	250.36	0	0
	328.90	0	0
	430.50	0	0
	575.57	0	0
	709.74	0	0
	968.46	0	0

Table 3

Age group	Male pop	D_1	D_2
0-14	473.50	1	0
	498.15	1	0
	506.49	1	0
	571.72	1	0
	623.66	1	0
	688.32	1	0
	925.53	1	0
	1190.17	1	0
	1399.35	1	0
	1392.77	1	0
	1905.13	1	0
	1869.37	1	0
15-59	678.84	0	1
	724.12	0	1
	714.74	0	1
	801.84	0	1
	933.03	0	1
	1064.94	0	1
	1212.91	0	1
	1482.74	0	1
	1883.46	0	1
	2434.05	0	1
	3086.53	0	1
	3875.82	0	1
60+	55.56	0	0
	61.63	0	0
	64.28	0	0
	55.74	0	0
	80.21	0	0
	102.04	0	0
	124.46	0	0
	167.59	0	0
	219.09	0	0
	298.76	0	0
	351.23	0	0
	479.80	0	0

Table 4

Age group	Female pop	D_1	D_2
0-14	445.97	1	0
	471.34	1	0
	478.80	1	0
	543.16	1	0
	594.01	1	0
	665.37	1	0
	877.31	1	0
	1106.62	1	0
	1313.20	1	0
	1522.40	1	0
	1742.54	1	0
	1691.85	1	0
15-59	663.08	0	1
	697.72	0	1
	681.37	0	1
	757.71	0	1
	877.09	0	1
	988.40	0	1
	1128.58	0	1
	1376.01	0	1
	1778.43	0	1
	2259.18	0	1
	2864.52	0	1
	3689.17	0	1
60+	64.55	0	0
	68.04	0	0
	67.52	0	0
	57.03	0	0
	75.80	0	0
	101.82	0	0
	123.51	0	0
	158.47	0	0
	207.87	0	0
	276.80	0	0
	352.48	0	0
	493.46	0	0

Effect coding data for three age groups

Table 5

Age group	Total pop	E_1	E_2
0-14	920.22	1	0
	970.55	1	0
	985.17	1	0
	1115.92	1	0
	1220.47	1	0
	1354.09	1	0
	1805.24	1	0
	2296.79	1	0
	2712.82	1	0
	3157.15	1	0
	3692.71	1	0
	3571.18	1	0
15-59	1344.58	0	1
	1421.79	0	1
	1397.34	0	1
	1562.29	0	1
	1813.18	0	1
	2054.6	0	1
	2341.1	0	1
	2861.4	0	1
	3662.65	0	1
	4697.63	0	1
	5945.37	0	1
	7566.06	0	1
60+	121.58	-1	-1
	131.09	-1	-1
	133.20	-1	-1
	114.38	-1	-1
	156.14	-1	-1
	205.82	-1	-1
	250.36	-1	-1
	328.90	-1	-1
	430.50	-1	-1
	575.57	-1	-1
	709.74	-1	-1
	968.46	-1	-1

Table 6

Age group	Male pop	E_1	E_2
0-14	473.5	1	0
	498.15	1	0
	506.49	1	0
	571.72	1	0
	623.66	1	0
	688.32	1	0
	925.53	1	0
	1190.17	1	0
	1399.35	1	0
	1392.77	1	0
	1905.13	1	0
	1869.37	1	0
15-59	678.84	0	1
	724.12	0	1
	714.74	0	1
	801.84	0	1
	933.03	0	1
	1064.94	0	1
	1212.91	0	1
	1482.74	0	1
	1883.46	0	1
	2434.05	0	1
	3086.53	0	1
	3875.82	0	1
60+	55.56	-1	-1
	61.63	-1	-1
	64.28	-1	-1
	55.74	-1	-1
	80.21	-1	-1
	102.04	-1	-1
	124.46	-1	-1
	167.59	-1	-1
	219.09	-1	-1
	298.76	-1	-1
	351.23	-1	-1
	479.80	-1	-1

Table 7

Age group	Female pop	E_1	E_2
0-14	445.97	1	0
	471.34	1	0
	478.8	1	0
	543.16	1	0
	594.01	1	0
	665.37	1	0
	877.31	1	0
	1106.62	1	0
	1313.2	1	0
	1522.4	1	0
	1742.54	1	0
	1691.85	1	0
15-59	663.08	0	1
	697.72	0	1
	681.37	0	1
	757.71	0	1
	877.09	0	1
	988.40	0	1
	1128.58	0	1
	1376.01	0	1
	1778.43	0	1
	2259.18	0	1
	2864.52	0	1
	3689.17	0	1
60+	64.55	-1	-1
	68.04	-1	-1
	67.52	-1	-1
	57.03	-1	-1
	75.80	-1	-1
	101.82	-1	-1
	123.51	-1	-1
	158.47	-1	-1
	207.87	-1	-1
	276.80	-1	-1
	352.48	-1	-1
	493.46	-1	-1

Dummy coding data for four states

Table 8

States	Total pop	D_1	D_2	D_3
AP	190.66	1	0	0
	214.47	1	0	0
	214.20	1	0	0
	242.03	1	0	0
	272.89	1	0	0
	311.15	1	0	0
	359.83	1	0	0
	435.03	1	0	0
	535.51	1	0	0
	665.08	1	0	0
	762.10	1	0	0
	845.80	1	0	0
TN	192.53	0	1	0
	209.03	0	1	0
	216.29	0	1	0
	234.72	0	1	0
	262.68	0	1	0
	301.19	0	1	0
	336.87	0	1	0
	411.99	0	1	0
	484.08	0	1	0
	558.59	0	1	0
	624.06	0	1	0
	721.47	0	1	0
Kerala	63.90	0	0	1
	71.50	0	0	1
	78.00	0	0	1
	95.10	0	0	1
	110.30	0	0	1
	135.50	0	0	1
	169.00	0	0	1
	213.50	0	0	1
	254.50	0	0	1
	290.90	0	0	1
	318.40	0	0	1
	333.90	0	0	1
Karnat	130.55	0	0	0
	135.25	0	0	0
	133.78	0	0	0
	146.33	0	0	0
	162.55	0	0	0
	194.02	0	0	0
	235.87	0	0	0
	292.99	0	0	0
	371.36	0	0	0
	449.77	0	0	0
	528.51	0	0	0
	610.95	0	0	0

Table 9

States	Male pop	D_1	D_2	D_3
AP	96.07	1	0	0
	107.72	1	0	0
	107.49	1	0	0
	121.84	1	0	0
	137.82	1	0	0
	157.43	1	0	0
	181.62	1	0	0
	220.09	1	0	0
	271.09	1	0	0
	337.25	1	0	0
	385.27	1	0	0
	424.42	1	0	0
TN	94.19	0	1	0
	102.37	0	1	0
	106.59	0	1	0
	115.78	0	1	0
	130.57	0	1	0
	150.04	0	1	0
	169.11	0	1	0
	208.28	0	1	0
	244.88	0	1	0
	282.99	0	1	0
	314.01	0	1	0
	361.38	0	1	0
Kerala	31.90	0	0	1
	35.60	0	0	1
	38.80	0	0	1
	47.00	0	0	1
	54.40	0	0	1
	66.80	0	0	1
	83.60	0	0	1
	105.90	0	0	1
	124.90	0	0	1
	142.80	0	0	1
	154.70	0	0	1
	168.60	0	0	1
Karnat	65.82	0	0	0
	68.28	0	0	0
	67.94	0	0	0
	74.45	0	0	0
	82.94	0	0	0
	98.67	0	0	0
	120.41	0	0	0
	149.72	0	0	0
	189.23	0	0	0
	229.52	0	0	0
	268.99	0	0	0
	309.67	0	0	0

Table 10

States	Female pop	D_1	D_2	D_3
AP	94.59	1	0	0
	106.75	1	0	0
	106.71	1	0	0
	120.20	1	0	0
	135.07	1	0	0
	155.17	1	0	0
	178.22	1	0	0
	214.94	1	0	0
	264.41	1	0	0
	327.83	1	0	0
	376.83	1	0	0
	421.39	1	0	0
TN	98.33	0	1	0
	106.66	0	1	0
	109.69	0	1	0
	118.94	0	1	0
	132.11	0	1	0
	151.15	0	1	0
	167.76	0	1	0
	203.71	0	1	0
	239.20	0	1	0
	275.60	0	1	0
	310.05	0	1	0
	360.09	0	1	0
Kerala	32.00	0	0	1
	35.90	0	0	1
	39.20	0	0	1
	48.00	0	0	1
	55.90	0	0	1
	68.70	0	0	1
	85.40	0	0	1
	107.60	0	0	1
	129.20	0	0	1
	148.10	0	0	1
	163.70	0	0	1
	177.00	0	0	1
Karnat	94.73	0	0	0
	66.97	0	0	0
	65.84	0	0	0
	71.88	0	0	0
	79.61	0	0	0
	95.35	0	0	0
	115.46	0	0	0
	143.27	0	0	0
	182.13	0	0	0
	220.25	0	0	0
	259.52	0	0	0
	301.29	0	0	0

Effect coding data for four states

Table 11

States	Total pop	E_1	E_2	E_3
AP	190.66	1	0	0
	214.47	1	0	0
	214.20	1	0	0
	242.03	1	0	0
	272.89	1	0	0
	311.15	1	0	0
	359.83	1	0	0
	435.03	1	0	0
	535.51	1	0	0
	665.08	1	0	0
	762.10	1	0	0
	845.80	1	0	0
TN	192.53	0	1	0
	209.03	0	1	0
	216.29	0	1	0
	234.72	0	1	0
	262.68	0	1	0
	301.19	0	1	0
	336.87	0	1	0
	411.99	0	1	0
	484.08	0	1	0
	558.59	0	1	0
	624.06	0	1	0
	721.47	0	1	0
Kerala	63.90	0	0	1
	71.50	0	0	1
	78.00	0	0	1
	95.10	0	0	1
	110.30	0	0	1
	135.50	0	0	1
	169.00	0	0	1
	213.50	0	0	1
	254.50	0	0	1
	290.90	0	0	1
	318.40	0	0	1
	333.90	0	0	1
Karnat	130.55	-1	-1	-1
	135.25	-1	-1	-1
	133.78	-1	-1	-1
	146.33	-1	-1	-1
	162.55	-1	-1	-1
	194.02	-1	-1	-1
	235.87	-1	-1	-1
	292.99	-1	-1	-1
	371.36	-1	-1	-1
	449.77	-1	-1	-1
	528.51	-1	-1	-1
	610.95	-1	-1	-1

Table 12

States	Male pop	E_1	E_2	E_3
AP	96.07	1	0	0
	107.72	1	0	0
	107.49	1	0	0
	121.84	1	0	0
	137.82	1	0	0
	157.43	1	0	0
	181.62	1	0	0
	220.09	1	0	0
	271.09	1	0	0
	337.25	1	0	0
	385.27	1	0	0
	424.42	1	0	0
TN	94.19	0	1	0
	102.37	0	1	0
	106.59	0	1	0
	115.78	0	1	0
	130.57	0	1	0
	150.04	0	1	0
	169.11	0	1	0
	208.28	0	1	0
	244.88	0	1	0
	282.99	0	1	0
	314.01	0	1	0
	361.38	0	1	0
Kerala	31.90	0	0	1
	35.60	0	0	1
	38.80	0	0	1
	47.00	0	0	1
	54.40	0	0	1
	66.80	0	0	1
	83.60	0	0	1
	105.90	0	0	1
	124.90	0	0	1
	142.80	0	0	1
	154.70	0	0	1
	168.60	0	0	1
Karnat	65.82	-1	-1	-1
	68.28	-1	-1	-1
	67.94	-1	-1	-1
	74.45	-1	-1	-1
	82.94	-1	-1	-1
	98.67	-1	-1	-1
	120.41	-1	-1	-1
	149.72	-1	-1	-1
	189.23	-1	-1	-1
	229.52	-1	-1	-1
	268.99	-1	-1	-1
	309.67	-1	-1	-1

Table 13

States	Female pop	E_1	E_2	E_3
AP	94.59	1	0	0
	106.75	1	0	0
	106.71	1	0	0
	120.20	1	0	0
	135.07	1	0	0
	155.17	1	0	0
	178.22	1	0	0
	214.94	1	0	0
	264.41	1	0	0
	327.83	1	0	0
	376.83	1	0	0
	421.39	1	0	0
TN	98.33	0	1	0
	106.66	0	1	0
	109.69	0	1	0
	118.94	0	1	0
	132.11	0	1	0
	151.15	0	1	0
	167.76	0	1	0
	203.71	0	1	0
	239.2	0	1	0
	275.6	0	1	0
	310.05	0	1	0
	360.09	0	1	0
Kerala	32.00	0	0	1
	35.90	0	0	1
	39.20	0	0	1
	48.00	0	0	1
	55.90	0	0	1
	68.70	0	0	1
	85.40	0	0	1
	107.60	0	0	1
	129.20	0	0	1
	148.10	0	0	1
	163.70	0	0	1
	177.00	0	0	1
Karnat	94.73	-1	-1	-1
	66.97	-1	-1	-1
	65.84	-1	-1	-1
	71.88	-1	-1	-1
	79.61	-1	-1	-1
	95.35	-1	-1	-1
	115.46	-1	-1	-1
	143.27	-1	-1	-1
	182.13	-1	-1	-1
	220.25	-1	-1	-1
	259.52	-1	-1	-1
	301.29	-1	-1	-1

Dummy coding data for sex ratio / urban-rural ratio

Table 14

States	Sex Ratio	D1	D2	D3
AP	985	1	0	0
	992	1	0	0
	993	1	0	0
	987	1	0	0
	980	1	0	0
	986	1	0	0
	981	1	0	0
	977	1	0	0
	975	1	0	0
	972	1	0	0
	978	1	0	0
	993	1	0	0
TN	1044	0	1	0
	1042	0	1	0
	1029	0	1	0
	1027	0	1	0
	1012	0	1	0
	1007	0	1	0
	992	0	1	0
	978	0	1	0
	977	0	1	0
	974	0	1	0
	987	0	1	0
	995	0	1	0
Kerala	1004	0	0	1
	1008	0	0	1
	1011	0	0	1
	1021	0	0	1
	1027	0	0	1
	1028	0	0	1
	1022	0	0	1
	1032	0	0	1
	1016	0	0	1
	1036	0	0	1
	1058	0	0	1
	1084	0	0	1
Karnataka	983	0	0	0
	981	0	0	0
	969	0	0	0
	965	0	0	0
	960	0	0	0
	966	0	0	0
	959	0	0	0
	957	0	0	0
	963	0	0	0
	960	0	0	0
	965	0	0	0
	968	0	0	0

Table 15

States	U-R Ratio	D1	D2	D3
AP	107	1	0	0
	112	1	0	0
	114	1	0	0
	125	1	0	0
	155	1	0	0
	345	1	0	0
	211	1	0	0
	239	1	0	0
	304	1	0	0
	368	1	0	0
	377	1	0	0
	504	1	0	0
TN	165	0	1	0
	177	0	1	0
	178	0	1	0
	220	0	1	0
	245	0	1	0
	322	0	1	0
	364	0	1	0
	434	0	1	0
	492	0	1	0
	519	0	1	0
	787	0	1	0
	938	0	1	0
Kerala	76	0	0	1
	80	0	0	1
	96	0	0	1
	107	0	0	1
	122	0	0	1
	156	0	0	1
	178	0	0	1
	244	0	0	1
	231	0	0	1
	359	0	0	1
	351	0	0	1
	913	0	0	1
Karnataka	144	0	0	0
	131	0	0	0
	160	0	0	0
	181	0	0	0
	205	0	0	0
	298	0	0	0
	287	0	0	0
	321	0	0	0
	406	0	0	0
	448	0	0	0
	515	0	0	0
	631	0	0	0

Effect coding data for sex ratio / urban-rural ratio

Table 16

States	F-M Ratio	E1	E2	E3
AP	985	1	0	0
	992	1	0	0
	993	1	0	0
	987	1	0	0
	980	1	0	0
	986	1	0	0
	981	1	0	0
	977	1	0	0
	975	1	0	0
	972	1	0	0
	978	1	0	0
	993	1	0	0
TN	1044	0	1	0
	1042	0	1	0
	1029	0	1	0
	1027	0	1	0
	1012	0	1	0
	1007	0	1	0
	992	0	1	0
	978	0	1	0
	977	0	1	0
	974	0	1	0
	987	0	1	0
	995	0	1	0
Kerala	1004	0	0	1
	1008	0	0	1
	1011	0	0	1
	1021	0	0	1
	1027	0	0	1
	1028	0	0	1
	1022	0	0	1
	1032	0	0	1
	1016	0	0	1
	1036	0	0	1
	1058	0	0	1
	1084	0	0	1
Karnataka	983	-1	-1	-1
	981	-1	-1	-1
	969	-1	-1	-1
	965	-1	-1	-1
	960	-1	-1	-1
	966	-1	-1	-1
	959	-1	-1	-1
	957	-1	-1	-1
	963	-1	-1	-1
	960	-1	-1	-1
	965	-1	-1	-1
	968	-1	-1	-1

Table 17

States	U-R Ratio	E1	E2	E3
AP	107	1	0	0
	112	1	0	0
	114	1	0	0
	125	1	0	0
	155	1	0	0
	345	1	0	0
	211	1	0	0
	239	1	0	0
	304	1	0	0
	368	1	0	0
	377	1	0	0
	504	1	0	0
TN	165	0	1	0
	177	0	1	0
	178	0	1	0
	220	0	1	0
	245	0	1	0
	322	0	1	0
	364	0	1	0
	434	0	1	0
	492	0	1	0
	519	0	1	0
	787	0	1	0
	938	0	1	0
Kerala	76	0	0	1
	80	0	0	1
	96	0	0	1
	107	0	0	1
	122	0	0	1
	156	0	0	1
	178	0	0	1
	244	0	0	1
	231	0	0	1
	359	0	0	1
	351	0	0	1
	913	0	0	1
Karnataka	144	-1	-1	-1
	131	-1	-1	-1
	160	-1	-1	-1
	181	-1	-1	-1
	205	-1	-1	-1
	298	-1	-1	-1
	287	-1	-1	-1
	321	-1	-1	-1
	406	-1	-1	-1
	448	-1	-1	-1
	515	-1	-1	-1
	631	-1	-1	-1

i) Regression of the population (Total/Male/Female) on three different age groups in India during census data (1901-2011):

Dummy coding method

For total population [from the data of table (2)]

$$\hat{Y} = 343.8112 + 1639.7151\ D_1 + 2711.8547\ D_2;\ R^2 = 0.4329 \tag{3}$$

[0.8935] [3.0131] [4.9833]; F = 12.5979
(0.3781) (0.0049) (0.0000)

For male population [from the data of table (3)]

$$\hat{Y} = 171.6992 + 831.9808\ D_1 + 1402.7192\ D_2;\ R^2 = 0.4364 \tag{4}$$

[0.8700] [2.9808] [5.0256]; F = 12.7743
(0.3906) (0.0054) (0.0000)

For female population [from the data of table (4)]

$$\hat{Y} = 170.6125 + 783.7683\ D_1 + 1309.4925\ D_2;\ R^2 = 0.4333 \tag{5}$$

[0.9197] [2.9873] [4.9911]; F = 12.6170
(0.3644) (0.0053) (0.0000)

Figures in squares brackets and in parenthesis of equations respectively indicates t-values and p-values.

From equation (3), the intercept 343.8112 represents the mean of the age group 60 +. The regression coefficient associated with D_1 (1639.7151) indicates that the mean of age group 0 − 14 is 1639.7151 points greater than that of age group 60 +. This difference is statistically significant. Also the regression coefficient associated with D_2 (2711.8547) indicates that the mean of age group 15 −59 is 2711.8547 points greater than that of age group 60 +. This difference is statistically significant. Since R^2 is significant, we say that proportion of variance in total population accounted for by the age group is statistically significant. Similarly we may write the interpretation of equation (4) and equation (5).

Effect coding method

For total population [from the data of table (5)]

$$\hat{Y} = 1794.3344 + 189.1918\ E_1 + 1261.3314\ E_2;\ R^2 = 0.4329 \tag{6}$$

[8.0766] [0.6022] [4.0146]; F = 12.5979
(0.0000) (0.5512) (0.0003)

For male population [from the data of table (6)]

$$\hat{Y} = 916.5992 + 87.0808\ E_1 + 657.8192\ E_2;\ R^2 = 0.4364 \tag{7}$$

[8.0440] [0.5404] [4.0821]; F = 12.7743
(0.0000) (0.5926) (0.0003)

For female population [from the data of table (7)]

$$\hat{Y} = 868.3661 + 86.0147\ E_1 + 611.7389\ E_2;\ R^2 = 0.4333 \tag{8}$$

[8.1073] [0.5678] [4.0385]; F = 12.6170
(0.0000) (0.5740) (0.0003)

From equation (6), the intercept 1794.3344 represents the grand mean of all age groups. The regression coefficient associated with E_1 (189.1918) indicates that the mean of age group 0 −14 is 189.1918 points greater than the grand mean of all age groups. This difference is statistically in significant. Also the regression coefficient associated with E_2 (1261.3314) indicates that the mean of age group 15 −59 is 1261.3314 points greater than the grand mean of all age groups. This difference is statistically significant. Since R^2 is significant, we say that proportion of variance in total population accounted for by the age groups is statistically significant. Similarly we may write the interpretation of equation (7) and equation (8).

ii) Regression of the population (Total/Male/Female) on four southern states of India for census data (1901-2011):

Dummy coding method

For total population [from the data of table (8)]

$$\hat{Y} = 282.6608 + 138.0683\ D_1 + 96.7975\ D_2 - 104.7858\ D_3;\ R^2 = 0.2372 \tag{9}$$

[5.3972] [-0.7841] [1.1399] [-0.8351]; F = 4.5605
(0.0000) (0.4372) (0.2605) (0.4082)

For male population [from the data of table (9)]

$$\hat{Y} = 143.8033 + 68.5392\ D_1 + 46.2125\ D_2 - 55.8867\ D_3;\ R^2 = 0.2390 \tag{10}$$

[5.6072] [1.8897] [1.2742] [-1.5409]; F = 4.6069
(0.0000) (0.0654) (0.2093) (0.1305)

For female population [from the data of table (10)]

$$\hat{Y} = 141.3583 + 67.1508\ D_1 + 48.0825\ D_2 - 50.4667\ D_3;\ R^2 = 0.2323 \tag{11}$$

[5.6549] [1.8995] [1.3601] [-1.4275]; F = 4.4378
(0.0000) (0.0641) (0.1807) (0.1605)

From equation (9), the intercept 282.6608 represents the mean of the Karnataka. The regression coefficient associated with D_1 (138.0683) indicates that the mean of Andhra Pradesh is 138.0683 points greater than that of Karnataka. The regression coefficient associated with D_2 (96.7975) indicates that the mean of Tamil Nadu is 96.7975 points greater than that of Karnataka. Also the regression coefficient associated with D_3 (-104.7858) indicates that the mean of Kerala is 104.7858 points lesser than that of Karnataka. All the three regression coefficients are not significant, but R^2 is significant. Similarly we may write the interpretation of equation (10) and equation (11).

Effect coding method

For total population [from the data of table (11)]

$$\hat{Y} = 315.1808 + 105.5483\ E_1 + 64.2775\ E_2 - 137.3058\ E_3;\ R^2 = 0.2372 \tag{12}$$

[12.4307] [2.4034] [1.4636] [-3.1265]; F = 4.5605
(0.0000) (0.0205) (0.1504) (0.0031)

For male population [from the data of table (12)]

$$\hat{Y} = 158.5196 + 53.8229\ E_1 + 31.4963\ E_2 - 70.6029\ E_3;\ R^2 = 0.2390 \tag{13}$$

[12.3620] [2.4233] [1.4181] [-3.1788]; F = 4.6069
(0.0000) (0.0196) (0.1632) (0.0027)

For male population [from the data of table (13)]

$$\hat{Y} = 157.5500 + 50.9592\ E_1 + 31.8908\ E_2 - 66.6583\ E_3;\ R^2 = 0.2323 \tag{14}$$

[12.6052] [2.3539] [1.4731] [-3.0791]; F = 4.4378
(0.0000) (0.0231) (0.1478) (0.0036)

From equation (12), the intercept 315.1808 represents the grand mean of all states. The regression coefficient associated with E_1 (105.5483) indicates that the mean of Andhra Pradesh is 105.5483 points greater than the grand mean of all states. The regression coefficient associated with E_2 (64.2775) indicates that the mean of Tamil Nadu is 64.2775 points greater than the grand mean of all states. Also the regression coefficient associated with E_3 (-137.3058) indicates that the mean of Kerala is 137.3058 points lesser than the grand mean of all states. The regression coefficients of E_1, E_3 and R^2 are statistically significant, but not the regression coefficients of E_2. Similarly we can interpret the equation (13) and equation (14).

iii) Regression of sex ratio/urban rural ratio on four southern states of India for census data (1901-2011): Dummy coding method

For sex ratio [from the data of table (14)]

$\hat{Y}$ = 966.3333 + 16.9167 D_1 + 39.0000 D_2 + 62.5833 D_3; R^2 = 0.6552 (15)
[187.8212] [2.3250] [5.3600] [8.6012]; F = 27.8709
(0.0000) (0.0247) (0.0000) (0.0000)

For urban rural ratio [from the data of table (15)]

$\hat{Y}$ = 310.6108 - 63.8150 D_1 + 92.7700 D_2 - 67.9700 D_3; R^2 = 0.1040 (16)
[5.3972] [-0.7841] [1.1399] [-0.8351]; F = 1.7027
(0.0000) (0.4372) (0.2605) (0.4082)

From equation (15), the intercept 966.3333 represents the mean of the state Karnataka. The regression coefficient associated with D_1 (16.9167) indicates that the mean of Andhra Pradesh is 16.9167 points greater than that of Karnataka. This difference is statistically significant. Also the regression coefficient associated with D_2 (39.0000) indicates that the mean of Tamil Nadu is 39.0000 points greater than that of Karnataka. This difference is statistically significant. And also the regression coefficient associated with D_3 (62.5833) indicates that the mean of Kerala is 62.5833 points greater than that of Karnataka. This difference is statistically significant. Since R^2 is significant, we say that proportion of variance in sex ratio accounted for by the four states is statistically significant.

From equation (16), the intercept 310.6108 represents the mean of the state Karnataka. The regression coefficient associated with D_1 (- 63.8150) indicates that the mean of Andhra Pradesh is 63.8150 points lesser than that of Karnataka. The regression coefficient associated with D_2 (92.7700) indicates that the mean of Tamil Nadu is 92.7700 points greater than that of Karnataka. And the regression coefficient associated with D_3 (- 67.9700) indicates that the mean of Kerala is 67.9700 points lesser than that of Karnataka. R^2 and all the three regression coefficients are statistically insignificant.

Effect coding method

For sex ratio [from the data of table (16)]

$\hat{Y}$ = 995.9583 −12.7083 E_1 + 9.3750 E_2 + 32.9583 E_3; R^2 = 0.6552 (17)
[387.1586] [-2.8522] [2.1041] [7.3969]; F = 27.8709
(0.0000) (0.0066) (0.0411) (0.0000)

For urban rural ratio [from the data of table (17)]

$\hat{Y}$ = 300.8571 - 54.0613 E_1 + 102.5238 E_2 - 58.2163 E_3; R^2 = 0.1040 (18)
[10.4555] [-1.0847] [2.0571] [-1.1681]; F = 1.7027
(0.0000) (0.2840) (0.0456) (0.2491)

From equation (17), the intercept 995.9583 represents the grand mean of all states. The regression coefficient associated with E_1 (-12.7083) indicates that the mean of Andhra Pradesh is 12.7083 points lesser than the grand mean of all states. This difference is statistically significant. Also the regression coefficient associated with E_2 (9.3750) indicates that the mean of Tamil Nadu is 9.3750 points greater than the mean of all states. This difference is statistically significant. And also the regression coefficient associated with E_3 (32.9583) indicates that the mean of Kerala is 32.9583 points greater than the grand mean of all states. This difference is statistically significant. Since R^2 is significant, we say that proportion of variance in sex ratio accounted for by the four states is statistically significant.

From equation (18), the intercept 300.8571 represents the grand mean of all states. The regression coefficient associated with E_1 (- 54.0613) indicates that the mean of Andhra Pradesh is 54.0613 points lesser than the grand mean of all states. The regression coefficient associated with E_2 (102.5238) indicates that the mean of Tamil Nadu is 102.5238 points greater than the grand mean of all states. The regression coefficient associated with E_3 (-58.2163) indicates that the mean of Kerala is 58.2163 points lesser than the grand mean of all states. The regression coefficients of E_2 is significant but not the regression coefficients of E_1, E_3 and R^2.

From Two Way ANOVA

(1) Regarding India with respect to total population, male population and female population it is observed that there is significant difference both in census years and in three age groups.

(2) Regarding total population, male population and female population it is observed that there is significant difference both in census years and in four southern states of India.

(3) Regarding sex ratio it is observed that there is no significant difference between census years, but there is significant difference between four states Andhra Pradesh, Tamil Nadu, Kerala and Karnataka.

(4) Regarding urban rural ratio it is observed that there is significant difference both in census years and in four southern states of India.

Conclusions

From dummy and effect coding regression methods the following conclusions are drawn:

(i). Regarding population in different age groups

The average no. of total population in age group 60+ is 348.8112 lakhs. The average no. of total population in age groups 0-14 and 15 – 59 respectively given by 1639.7151, 2711.8547 lakhs greater than that of age group 60 +.Average no of total population of all age groups is 1794.3344 lakhs. The average no. of total population in age groups 0-14 and 15-59 respectively given by 189.1918, 1261.3314 lakhs greater than that of all age groups.

The average no. of male population in age group 60+ is 171.6992 lakhs. The average no. of male population in age groups 0 – 14 and 15 – 59 respectively given by 831.9808, 1402.7192 lakhs greater than that of age group 60 +. The average no. of male population of all age groups is 916.5992 lakhs. The average no of male population in age groups 0 – 14 and 15 – 59 respectively given by 87.0808, 657.8192 lakhs greater than that of all age groups.

The average no of female population in age group 60+ is 170.6125 lakhs. The average no of female population in age group 0 – 14 and 15 – 59 respectively given by 783.7683, 1309.4925 lakhs greater than that of age group 60 +.The average no. of female population of all age groups is 868.3661lakhs. The average no of female population in age groups 0 – 14 and 15 – 59 respectively given by 86.0147, 611.7389 lakhs greater than that of all age groups.

(ii). Regarding population in four southern states of India

The average no. of total population in Karnataka is 282.668 lakhs. The average no. of total population in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 138.0683, 96.7975,-104.7858 lakhs greater than that of Karnataka. Average no. of total population of four states is 315.1808 lakhs. The average no. of total population in Andhra Pradesh, Tamil Nadu, Kerala respectively give by 105.5483, 64.2775, -137.3058 lakhs greater than the average no. of total population of four states.

The average no. of male population in Karnataka is 143.8033 lakhs. The average no. of male population in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 68.5392, 46.2125, -55.8865 lakhs greater than that of Karnataka. The average no. of male population of four states is 158.5196 lakhs. The average no. of male population in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 53.8229, 31.4963, -70.6029 lakhs greater than the average no. of total population of four states.

The average no. of female population in Karnataka is 141.3583 lakhs. The average no. of female population in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 67.1508, 48.0825, -50.4667 lakhs greater than that of Karnataka. The average no. of female population of four states is 157.5500 lakhs. The average no. of female population in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 50.9592, 31.8908, -60.6583 lakhs is greater than the average no. of total population of four states.

(iii). Regarding sex ratio / urban rural ratio in four southern states of India

The average sex ratio in Karnataka is 966.3333. The average sex ratio in Andhra Pradesh, Tamil Nadu, Kerala respectively given by 16.9167, 39, 62.5833 greater than that of Karnataka. The average sex ratio of four states is 955.9583. The average sex ratio in Andhra Pradesh, Tamil Nadu, Kerala respectively given by -12.7083, 9.3750, 32.9583 greater than the average sex ratio of four states.

The average urban rural ratio in Karnataka is 310.6108. The average urban rural ratio in Andhra Pradesh, Tamil Nadu, Kerala, respectively given by - 63.8150, 92.77, - 67.97 greater than that of Karnataka. The average urban rural ratio of four states is 300.8571. The average urban rural ratio in Andhra Pradesh, Tamil Nadu, Kerala respectively given by -54.0613, 102.5238, -58.2163 greater than the average of four states.

From two way analysis of variance it is concluded that

(1) There is significant difference both in census years and in age groups with respect to total population, male population and female population in India.

(2) There is significant difference both in census years and in four southern states of India with respect to total population, male population and female population.

(3) There is no significant difference between census years (1901 – 2015), but there is significant difference between four southern states of India with respect to sex ratio.

(4) There is significant difference both in census years and in four southern states of India with respect to urban rural ratio.

Appendix

Table 1: Population of India in lakhs (age groups, total, male, female)

Census Year	Total			Male			Female		
	0-14	15-59	60+	0-14	15-59	60+	0-14	15-59	60+
1901	920.22	1344.58	121.58	473.50	678.84	55.56	445.97	663.08	64.55
1911	970.55	1421.79	131.09	498.15	724.12	61.63	471.34	697.72	68.04
1921	985.17	1397.34	133.20	506.49	714.74	64.28	478.80	681.37	67.52
1931	1115.92	1562.29	114.38	571.72	801.84	55.74	543.16	757.71	57.03
1941	1220.47	1813.18	156.14	623.66	933.03	80.21	594.01	877.09	75.80
1951	1354.09	2054.60	205.82	688.32	1064.94	102.04	665.37	988.40	101.82
1961	1805.24	2341.10	250.36	925.53	1212.91	124.46	877.31	1128.58	123.51
1971	2296.79	2861.40	328.90	1190.17	1482.74	167.59	1106.62	1376.01	158.47
1981	2712.82	3662.65	430.50	1399.35	1883.46	219.09	1313.20	1778.43	207.87
1991	3157.15	4697.63	575.57	1392.77	2434.05	298.76	1522.40	2259.18	276.80
2001	3692.71	5945.37	709.74	1905.13	3086.53	351.23	1742.54	2864.52	352.48
2011	3571.18	7566.06	968.46	1869.37	3875.82	479.80	1691.85	3689.17	493.46

Census Year	Andhra Pradesh					Tamil Nadu				
	Total Pop	Male	Female	Rural	Urban	Total Pop	Male	Female	Rural	Urban
1901	190.66	96.07	94.59	172.26	18.40	192.53	94.19	98.33	165.28	27.24
1911	214.47	107.72	106.75	192.82	21.65	209.03	102.37	106.66	177.53	31.49
1921	214.20	107.49	106.71	192.33	21.87	216.29	106.59	109.69	182.00	32.48
1931	242.03	121.84	120.20	215.09	26.94	234.72	115.78	118.94	192.42	42.30
1941	272.89	137.82	135.07	236.23	36.66	262.68	130.57	132.11	210.94	51.74
1951	311.15	157.43	155.17	156.95	54.20	301.19	150.04	151.15	2227.86	73.34
1961	359.83	181.62	178.22	297.09	62.75	336.87	169.11	167.76	246.96	89.91
1971	435.03	220.09	214.94	351.00	84.03	411.99	208.28	203.71	287.34	124.65
1981	535.51	271.09	264.41	410.62	124.88	484.08	244.88	239.20	324.56	159.52
1991	665.08	337.25	327.83	486.21	178.87	558.59	282.99	275.60	367.81	190.78
2001	762.10	385.27	376.83	552.24	208.09	624.06	314.01	310.05	349.22	274.84
2011	845.80	424.42	421.39	563.12	283.54	721.47	361.38	360.09	372.30	349.17

Census Year	Kerala					Karnataka				
	Total Pop	**Male**	**Female**	**Rural**	**Urban**	**Total Pop**	**Male**	**Female**	**Rural**	**Urban**
1901	63.90	31.90	32.00	59.40	4.50	130.55	65.82	94.73	114.12	16.43
1911	71.50	35.60	35.90	66.20	5.30	135.25	68.28	66.97	119.55	15.71
1921	78.00	38.80	39.20	71.20	6.80	133.78	67.94	65.84	115.32	18.46
1931	95.10	47.00	48.00	85.90	9.20	146.33	74.45	71.88	123.87	22.46
1941	110.30	54.40	55.90	98.30	12.00	162.55	82.94	79.61	134.93	27.63
1951	135.50	66.80	68.70	117.20	18.30	194.02	98.67	95.35	149.48	44.53
1961	169.00	83.60	85.40	143.50	25.50	235.87	120.41	115.46	183.20	52.66
1971	213.50	105.90	107.60	178.80	43.70	292.99	149.72	143.27	221.77	71.22
1981	254.50	124.90	129.20	206.80	47.70	371.36	189.23	182.13	264.06	107.30
1991	290.90	142.80	148.10	214.10	76.80	449.77	229.52	220.25	310.69	139.08
2001	318.40	154.70	163.70	235.70	82.70	528.51	268.99	259.52	348.14	179.20
2011	333.90	168.60	177.00	174.56	159.32	610.95	309.67	301.29	374.69	236.26

References

1. Husain Alkharusin. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding, International Journal of Education. 2012; 4(2):202-210.
2. Norman R Draper, Harry Smith. Applied regression analysis, Third Edition, Wiley India Pvt. Ltd, 2011.