Chapter 8
Linear Regression with Categorical Predictors (60 Patients)

1 General Purpose

Variable restructuring is a valuable method for minimizing important biases in your everyday data analysis. In a study with a categorical predictor like races, the race values 1–4 have no incremental function, and, therefore, linear regression is not appropriate for assessing their effect on any outcome. Instead, restructuring the data for categorical predictors does the job.

2 Schematic Overview of Type of Data File

Outcome	predictor
.	race 1
.	race 1
.	race 2
.	race 3
.	race 4
.	race 3
.	race 1
.	race 2
.	race 4

T.J. Cleophas, A.H. Zwinderman, *SPSS for Starters and 2nd Levelers*,
DOI 10.1007/978-3-319-20600-4_8

After restructuring of the above data, the data file will look like underneath.

Outcome	race 1	race 2	race 3	race 4
.	yes	no	no	no
.	yes	no	no	no
.	no	yes	no	no
.	no	no	yes	no
.	no	no	no	yes
.	no	no	yes	no
.	yes	no	no	no
.	no	yes	no	no
.	no	no	no	yes

3 Primary Scientific Question

Linear regression is not appropriate for assessing categorical predictors. Can linear regression be appropriately used if the categorical predictors are restructured into multiple binary variables.

4 Data Example

In a study the scientific question was: does race have an effect on physical strength. The variable race has a categorical rather then linear pattern. The effects on physical strength (scores 0–100) were assessed in 60 subjects of different races (hispanics (1), blacks (2), asians (3),and whites (4)), ages (years), and genders (0 = female, 1 = male). The first 10 patients are in the table underneath.

patient number	physical strength	race	age	gender
1	70,00	1,00	35,00	1,00
2	77,00	1,00	55,00	0,00
3	66,00	1,00	70,00	1,00
4	59,00	1,00	55,00	0,00
5	71,00	1,00	45,00	1,00
6	72,00	1,00	47,00	1,00
7	45,00	1,00	75,00	0,00
8	85,00	1,00	83,00	1,00
9	70,00	1,00	35,00	1,00
10	77,00	1,00	49,00	1,00

The entire data file is in extras.springer.com, and is entitled “chapter8categorical-predictors”. Start by opening the data file in SPSS.

Command:
click race....click Edit....click Copy....click a new "var"....click Paste....highlight the values 2–4.. . .delete and replace with 0,00 values....perform the same procedure subsequently for the other races.

patient number	physical strength	race	age	gender	race 1 hispanics	race 2 blacks	race 3 asians	race 4 whites
1	70,00	1,00	35,00	1,00	1,00	0,00	0,00	0,00
2	77,00	1,00	55,00	0,00	1,00	0,00	0,00	0,00
3	66,00	1,00	70,00	1,00	1,00	0,00	0,00	0,00
4	59,00	1,00	55,00	0,00	1,00	0,00	0,00	0,00
5	71,00	1,00	45,00	1,00	1,00	0,00	0,00	0,00
6	72,00	1,00	47,00	1,00	1,00	0,00	0,00	0,00
7	45,00	1,00	75,00	0,00	1,00	0,00	0,00	0,00
8	85,00	1,00	83,00	1,00	1,00	0,00	0,00	0,00
9	70,00	1,00	35,00	1,00	1,00	0,00	0,00	0,00
10	77,00	1,00	49,00	1,00	1,00	0,00	0,00	0,00

The result is shown above. For the analysis we will use multiple linear regression. First the inadequate analysis.

5 Inadequate Linear Regression

For analysis the module Compare Means is required. It consists of the following statistical models:

Means,
One-Sample T-Test,
Independent-Samples T-Test,
Paired-Samples T-Test and
One Way ANOVA

Command:
Analyze. . ..Regression. . ..Linear. . ..Dependent: physical strength score. . ..Independent: race, age, gender. . ..OK.

The table shows that age and gender are significant predictors but race is not.

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	79,528	8,657		9,186	,000
	race	,511	1,454	,042	,351	,727
	age	−,242	,117	−,260	−2,071	,043
	gender	9,575	3,417	,349	2,802	,007

[a] Dependent variable: strengthscore

The variable race is analyzed as a stepwise rising function from 1 to 4, and the linear regression model assumes that the outcome variable will rise (or fall) simultaneously and linearly, but this needs not be necessarily so. Next a categorical analysis will be performed.

6 Multiple Linear Regression for Categorical Predictors

The above commands are given once more, but now the independent variables are entered slightly differently.

Command:
Analyze….Regression….Linear….Dependent: physical strength score….Independent: race 2, race 3, race 4, age, gender….click OK.

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	72,650	5,528		13,143	,000
	race2	17,424	3,074	,559	5,668	,000
	race3	−6,286	3,141	−,202	−2,001	,050
	race4	9,661	3,166	,310	3,051	,004
	age	−,140	,081	−,150	−1,716	,092
	gender	5,893	2,403	,215	2,452	,017

[a] Dependent variable: strengthscore

The above table shows that race 2–4 are significant predictors of physical strength.

The results can be interpreted as follows.

The underneath regression equation is used:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

$a =$ intercept
$b_1 =$ regression coefficient for blacks $(0 = \text{no}, 1 = \text{yes})$,
$b_2 =$ asians
$b_3 =$ whites
$b_4 =$ age
$b_5 =$ gender

If an individual is hispanic (race 1), then x_1, x_2, and x_3 will turn into 0, and the regression equation becomes $y = a + b_4x_4 + b_5x_5$.

In summary:

$$\begin{array}{ll} \text{if hispanic,} & y = a + b_4x_4 + b_5x_5. \\ \text{if black,} & y = a + b_1 + b_4x_4 + b_5x_5. \\ \text{if asian,} & y = a + b_2 + b_4x_4 + b_5x_5. \\ \text{if white,} & y = a + b_3 + b_4x_4 + b_5x_5. \end{array}$$

So, e.g., the best predicted physical strength score of a white male of 25 years of age would equal

$y = 72.65 + 9.66 - 0.14*25 + 5.89*1 = 84.7$ (on a linear scale from 0 to 100), ($* =$ sign of multiplication).

Compared to the presence of the hispanic race, the black and white races are significant positive predictors of physical strength ($p = 0.0001$ and 0.004 respectively), the asian race is a significant negative predictor ($p = 0.050$). All of these results are adjusted for age and gender, at least if we use $p = 0.10$ as criterion for statistical significance.

7 Conclusion

Multiple linear regression is adequate for testing categorical predictors after restructuring them into multiple binary variables. Also with a binary outcome variable categorical analysis of covariates is possible. Using logistic regression in SPSS is convenient for the purpose, we need not *manually* transform the quantitative estimator into a categorical one. For the analysis we apply the usual commands.

Command:
Analyze ....Regression....Binary logistic....Dependent variable.... Independent variables....then, open dialog box labeled Categorical Variables.... select the categorical variable and transfer it to the box Categorical Variables....then click Continue....click OK.

8 Note

More background, theoretical and mathematical information of categorical predictors is given in the Chap. 21, pp 243–252, in Statistics applied to clinical studies, Springer Heidelberg Germany, 2012, from the same authors.