Chapter 29
Various Methods for Analyzing Predictor Categories (60 and 30 Patients)

General Purpose

Categories unlike continuous data need not have stepping functions. In order to apply regression analysis for their analysis we need to recode them into multiple binary (dummy) variables. Particularly, if Gaussian distributions in the outcome are uncertain, automatic non-parametric testing is an adequate and very convenient modern alternative.

Specific Scientific Questions

1. Does race have an effect on physical strength (the variable race has a categorical rather than linear pattern).
2. Are the hours of sleep / levels of side effects different in categories treated with different sleeping pills.

Example 1

The effects on physical strength (scores 0–100) assessed in 60 subjects of different races (hispanics (1), blacks (2), asians (3), and whites (4)), and ages (years), are in the left three columns of the data file entitled “categoriesaspredictor”.

This chapter was previously published in “Machine learning in medicine-cookbook 2” as Chap. 5, 2014.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_29

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

Only the first 10 patients are displayed above. The entire data file in www.springer.com. For the analysis we will use multiple linear regression.

Command:

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

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	92,920	7,640		12,162	,000
	Race	−,330	1,505	−,027	−,219	,827
	Age	−,356	,116	−,383	−3,071	,003

[a] Dependent variable: strengthscore

The above table shows that age is a significant predictor but race is not. However, the analysis is not adequate, because the variable race is analyzed as a stepwise function from 1 to 4, and the linear regression model assumes that the outcome variable will rise (or fall) linearly, but, in the data given, this needs not be necessarily so. It may, therefore, be more safe to recode the stepping variable into the form of a categorical variable. The underneath data overview shows in the right 4 columns how it is manually done.

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

We, subsequently, use again linear regression, but now for categorical analysis of race.

Command:

click Transform….click Random Number Generators….click Set Starting Point…. click Fixed Value (2000000)….click OK….click Analyze….Regression ….Linear ….Dependent: physical strength score….Independent: race 1, race 3, race 4, age…. click Save….mark Unstandardized….in Export model information to XML (eXtended Markup Language) file: type "exportcategoriesaspredictor"….click Browse….File name: enter "exportcategoriesaspredictor"….click Continue….click OK.

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	97,270	4,509		21,572	,000
	Age	−,200	,081	−,215	−2,457	,017
	Race1	−17,483	3,211	−,560	−5,445	,000
	Race3	−25,670	3,224	−,823	−7,962	,000
	Race4	−8,811	3,198	−,282	−2,755	,008

[a] Dependent variable: strengthscore

The above table is in the output. It shows that race 1, 3, 4 are significant predictors of physical strength compared to race 2. The results can be interpreted as follows.

The underneath regression equation is used:

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

a = intercept
b_1 = regression coefficient for age
b_2 = hispanics
b_3 = asians
b_4 = white

If an individual is black (race 2), then x_2, x_3, and x_4 will turn into 0, and the regression equation becomes

$$y = a + b_1x_1$$

If hispanic, $y = a + b_1x_1 + b_2x_2$

If asian, $y = a + b_1x_1 + b_3x_3$

If white, $y = a + b_1x_1 + b_4x_4.$

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

$$y = 97.270 + 0.20{*}25 - 8.811{*}1 = 93.459,$$
$$(* = \text{sign of multiplication}).$$

Obviously, all of the races are negative predictors of physical strength, but the blacks scored highest and the asians lowest. All of these results are adjusted for age.

If we return to the data file page, we will observe that SPSS has added a new variable entitled “PRE_1”. It represents the individual strengthscores as predicted by the recoded linear model. They are pretty similar to the measured values.

We can now with the help of the Scoring Wizard and the exported XML (eXtended Markup Language) file entitled “exportcategoriesaspredictor” try and predict strength scores of future patients with known race and age.

race	age
1,00	40,00
2,00	70,00
3,00	54,00
4,00	45,00
1,00	36,00
2,00	46,00
3,00	50,00
4,00	36,00

First, recode the stepping variable race into 4 categorical variables.

race	age	race1	race3	race4
1,00	40,00	1,00	,00	,00
2,00	70,00	,00	,00	,00
3,00	54,00	,00	1,00	,00
4,00	45,00	,00	,00	1,00
1,00	36,00	1,00	,00	,00
2,00	46,00	,00	,00	,00
3,00	50,00	,00	1,00	,00
4,00	36,00	,00	,00	1,00

Then Command:

click Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the exportcategoriesaspredictor.xml file….click Select….in Scoring Wizard click Next….click Finish.

race	age	race1	race3	race4	predicted strength score
1,00	40,00	1,00	,00	,00	71,81
2,00	70,00	,00	,00	,00	83,30
3,00	54,00	,00	1,00	,00	60,83
4,00	45,00	,00	,00	1,00	79,48
1,00	36,00	1,00	,00	,00	72,60
2,00	46,00	,00	,00	,00	88,09
3,00	50,00	,00	1,00	,00	61,62
4,00	36,00	,00	,00	1,00	81,28

The above data file now gives predicted strength scores of the 8 future patients as computed with help of the XML file.

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 have to 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….OK.

Example 2

Particularly, if Gaussian distributions in the outcome are uncertain, automatic non-parametric testing is an adequate and very convenient modern alternative. Three parallel-groups were treated with different sleeping pills. Both hours of sleep and side effect score were assessed.

Group	efficacy	gender	comorbidity	side effect score
0	6,00	,00	1,00	45,00
0	7,10	,00	1,00	35,00
0	8,10	,00	,00	34,00
0	7,50	,00	,00	29,00
0	6,40	,00	1,00	48,00
0	7,90	1,00	1,00	23,00
0	6,80	1,00	1,00	56,00
0	6,60	1,00	,00	54,00
0	7,30	1,00	,00	33,00
0	5,60	,00	,00	75,00

Only the first ten patients are shown. The entire data file is in extras.springer.com and is entitled “categoriesaspredictor2”. Automatic nonparametric tests is available in SPSS 18 and up. Start by opening the above data file.

Command:

Analyze….Nonparametric Tests….Independent Samples….click Objective….mark Automatically compare distributions across groups….click Fields….in Test fields: enter "hours of sleep" and "side effect score"….in Groups: enter "group"….click Settings….Choose Tests….mark "Automatically choose the tests based on the data"….click Run.

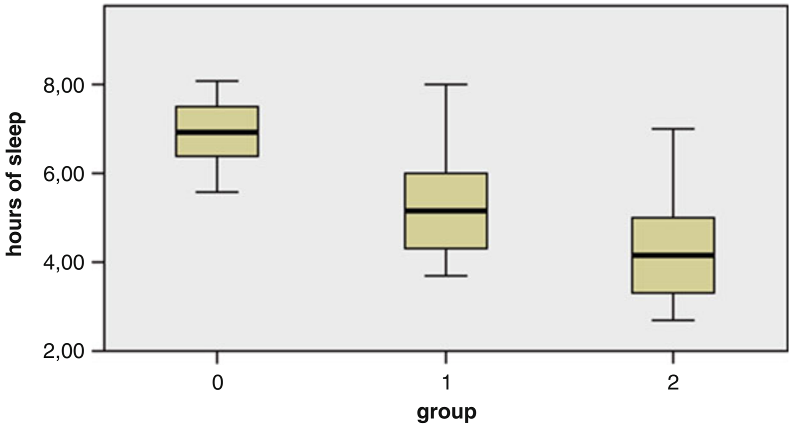
In the interactive output sheets the underneath table is given. Both the distribution of hours of sleep and side effect score are significantly different across the three categories of treatment. The traditional assessment of these data would have been a multivariate analysis of variance (MANOVA) with treatment-category as predictor and both hours of sleep and side effect score as outcome. However, normal distributions are uncertain in this example, and the correlation between the two outcome measures may not be zero, reducing the sensitivity of MANOVA. A nice thing about the automatic nonparametric tests is that, like discriminant analysis (Machine learning in medicine part one, Chap. 17, Discriminant analysis for supervised data, pp 215–224, Springer Heidelberg Germany, 2013, from the same authors), they assume orthogonality of the two outcomes, which means that the correlation level between the two does not have to be taken into account. By double-clicking the table you will obtain an interactive set of views of various details of the analysis, entitled the Model Viewer.

Hypothesis test summary

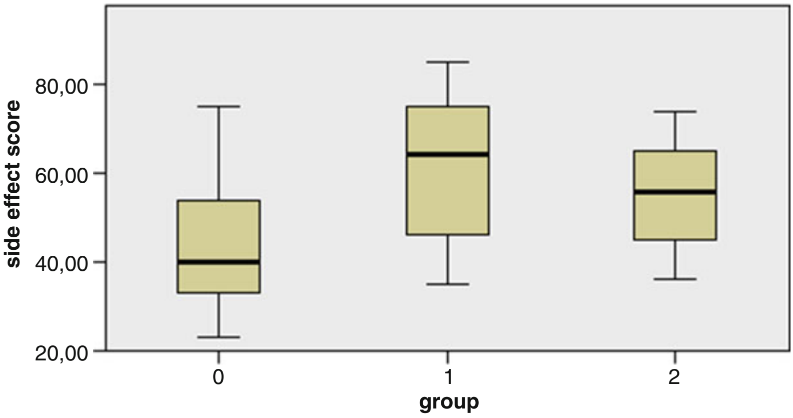
	Null hypothesis	Test	Sig.	Decision
1	The distribution of hours of sleep is the same across categories of group.	Independent-Samples Kruskal-Wallis Test	,001	Reject the null hypothesis.
2	The distribution of side effect score is the same across categories of group.	Independent-Samples Kruskal-Wallis Test	,036	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is,05

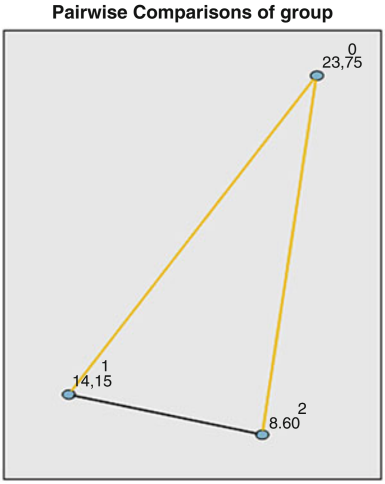
One view provides the box and whiskers graphs (medians, quartiles, and ranges) of hours of sleep of the three treatment groups. Group 0 seems to perform better than the other two, but we don't know where the significant differences are.



Also the box and whiskers graph of side effect scores is given. Some groups again seem to perform better than the other. However, we cannot see whether 0 vs 1, 1 vs 2, and/or 0 vs 2 are significantly different.



In the view space at the bottom of the auxiliary view (right half of the Model Viewer) several additional options are given. When clicking Pairwise Comparisons, a distance network is displayed with yellow lines corresponding to statistically significant differences, and black ones to insignificant ones. Obviously, the differences in hours of sleep of group 1 vs (versus) 0 and group 2 vs 0 are statistically significant, and 1 vs 2 is not. Group 0 had significantly more hours of sleep than the other two groups with p = 0.044 and 0.0001.

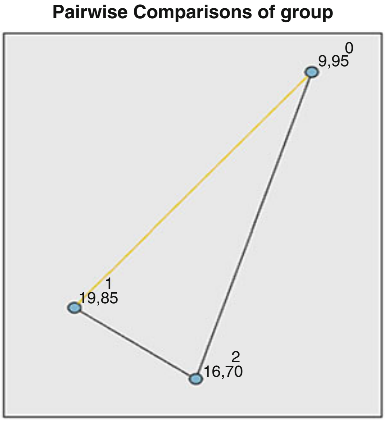


Each node shows the sample average rank of group

Sample1-	Sample2	Test statistic	Std. Error	Std. Test statistic	Sig.	Adj.Sig.
2-	1	5,550	3,936	1,410	,158	,475
2-	0	15,150	3,936	3,849	,000	,000
1-	0	9,600	3,936	2,439	,015	,044

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same
Asymptotic significances (2-sided tests) are displayed. The significance level is ,05

As shown below, the difference in side effect score of group 1 vs 0 is also statistically significant, and 1 vs 0, and 1 vs 2 are not. Group 0 has a significantly better side effect score than the 1 with p=0.035, but group 0 vs 2 and 1 vs 2 are not significantly different.


Pairwise Comparisons of group

Each node shows the sample average rank of group

Sample1-	Sample2	Test statistic	Std. Error	Std. Test statistic	Sig.	Adj.Sig.
0-	2	−6,750	3,931	−1,717	,086	,258
0-	1	−9,900	3,931	−2,518	,012	,035
2-	1	3,150	3,931	,801	,423	1,000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same
Asymptotic significances (2-sided tests) are displayed. The significance level is ,05

Conclusion

Predictor variables with a categorical rather than linear character should be recoded into categorical variables before analysis in a regression model. An example is given. Particularly if the Gaussian distributions in the outcome are uncertain, automatic non-parametric testing is an adequate and very convenient alternative.

Note

More background theoretical and mathematical information of categories as predictor is given in SPSS for starters part two, Chap. 5, Categorical data, pp 21–24, and Statistics applied to clinical studies 5th edition, Chap. 21, Races as a categorical variable, pp 244–252, both from the same authors and edited by Springer Heidelberg Germany 2012.