CATEGORICAL DATA ANALYSIS PROJECT

DENNIS MULUMBI KYALO

05-13-2022

Contents

1. INTRODUCTION

This is an observational cohort study of infants born in the state of Florida in 1993. It is an exploratory epidemiological evaluation of a secondary database merging birth vital statistics (BVS), Florida's Death Data source supplied by the Florida Department of Health, Medicaid eligibility and enrollment data files provided by Florida's Agency of Health Care Administration; Women, Infants, and Children (WIC) Nutritional Supplement Program certification files supplied by the Florida WIC office; and the Florida Healthy Start prenatal risk screen score data file supplied by Florida Department of Health. It contains singletons only.

Table 1: Data summary

Name	data_tbl
Number of rows	2553
Number of columns	12
Column type frequency:	
numeric	12
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
smoking	0	1	1.54	0.50	1	1	2	2	2
drk	0	1	1.74	0.44	1	1	2	2	2
BI	0	1	2.03	0.81	1	1	2	3	3
edu	0	1	1.94	0.80	1	1	2	3	3
ms	0	1	1.48	0.50	1	1	1	2	2
mrace	0	1	2.10	0.91	1	1	2	3	3
bsex	0	1	1.50	0.50	1	1	1	2	2
kotel	0	1	1.55	0.50	1	1	2	2	2
med	0	1	1.46	0.50	1	1	1	2	2
momage	0	1	2.24	0.79	1	2	2	3	3
total	0	1	73.08	377.76	1	2	6	27	8728
vlbw	0	1	0.86	3.09	0	0	0	1	61

From the tables above, we can see that the dataset comprises of 2553 observations and 12 variables. Where 10 are explanatory variables - smoking, drk, BI, edu, ms, mrace, bsex, kotel, med, momage. Variable "Total" is the total number of infants, while variable "vlbw" is the total number of infants with vlbw and "ideath" is the total number of infants dead within 1st year of his/her life, within the combination of 10 explanatory variables. Your study's outcome is vlbw.

2. ANALYSIS

We shall be using this dataset to perform our analysis.

Table 3: Sample size and vlbw in each cohort

factor	level	Obs	total	vlbw Obs	vlbw Perc
BI	<=15 Months	787	40611	463	21.171
	>15 Months	876	68842	602	27.526
	First Birth	890	77110	1122	51.303
bsex	Female	1269	90352	1049	47.965
	Male	1284	96211	1138	52.035
drk	No	1886	183845	2111	96.525
	Yes	667	2718	76	3.475
edu	<HS	892	43400	622	28.441
	>HS	743	72313	647	29.584
	HS	918	70850	918	41.975
kotel	No	1139	23313	360	16.461
	Yes	1414	163250	1827	83.539
med	No	1179	104423	958	43.804
	Yes	1374	82140	1229	56.196
momage	<=20	564	24929	398	18.198
	>34	811	19641	271	12.391
	20-34	1178	141993	1518	69.410
mrace	Black	944	42900	999	45.679
	Other	416	3707	29	1.326
	White	1193	139956	1159	52.995
ms	No	1326	64323	1167	53.361
	Yes	1227	122240	1020	46.639
smoking	No	1380	158490	1769	80.887
	Yes	1173	28073	418	19.113

Table 3 above provides a concise summary of the sample size (total singletons) indicated by Obs, the total number of births, the total number of infants with "vlbw" denoted by vlbw in each cohort and as well as the vlbw percentage.

Table 4: Table with raw rates by explanatory variables

factor	level	Number of Births
BI	<=15 Months	40611
	>15 Months	68842
	First Birth	77110
bsex	Female	90352
	Male	96211
drk	No	183845
	Yes	2718
edu	<HS	43400
	>HS	72313
	HS	70850
kotel	No	23313
	Yes	163250
med	No	104423
	Yes	82140
momage	<=20	24929
	>34	19641
	20-34	141993
mrace	Black	42900
	Other	3707
	White	139956
ms	No	64323
	Yes	122240
smoking	No	158490
	Yes	28073

Table 4 above shows the raw rates by explanatory variables.

Next, we perform our analysis with the 10 explanatory variables and their corresponding two-factor interactions (vlbw and total). We create a logistic regression and apply (stepwise) model selection criterion by entering and removing p-values which are set as 0.05. We shall then use the Akaike Information Criterion (AIC) to choose the optimal model.

	Dependent variable:				
	cbind(vlbw, total)				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
smoking	-0.375^{***}	-0.361^{***}	-0.357^{***}	-0.355^{***}	-0.355^{***}
	(0.060)	(0.060)	(0.059)	(0.059)	(0.059)
drk	-0.588^{***}	-0.580^{***}	-0.582^{***}	-0.576^{***}	-0.577^{***}
	(0.124)	(0.124)	(0.124)	(0.124)	(0.124)
BI	-0.109^{***}	-0.112^{***}	-0.111^{***}	-0.110^{***}	-0.110^{***}
	(0.029)	(0.029)	(0.029)	(0.028)	(0.028)
edu	0.009	−0.008	−0.002	0.002	0.002
	(0.033)	(0.032)	(0.031)	(0.030)	(0.030)
ms	-0.383^{***}	-0.399^{***}	-0.387^{***}	-0.383^{***}	-0.383^{***}
	(0.055)	(0.055)	(0.052)	(0.051)	(0.051)
mrace	-0.450^{***}	-0.450^{***}	-0.447^{***}	-0.445^{***}	-0.445^{***}
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
bsex	−0.027	−0.027	−0.027	−0.027	
	(0.043)	(0.043)	(0.043)	(0.043)	
kotel	0.053	0.051	0.052		
	(0.061)	(0.061)	(0.061)		
med	0.037	0.036			
	(0.054)	(0.054)			
momage	-0.059^{**}				
	(0.030)				
Constant	-0.709^{**}	-0.832^{***}	-0.822^{***}	-0.764^{***}	-0.803^{***}
	(0.276)	(0.269)	(0.269)	(0.260)	(0.252)
Observations	2,553	2,553	2,553	2,553	2,553
Log Likelihood	−1,715.263	−1,717.249	−1,717.474	−1,717.843	−1,718.033
Akaike Inf. Crit.	3,452.527	3,454.498	3,452.949	3,451.686	3,450.065

Note: *p<0.1; **p<0.05; ***p<0.01

	Dependent variable:				
	cbind(vlbw, total)				
	Model 6	Model 7	Model 8	Model 9	Model 10
	(1)	(2)	(3)	(4)	(5)
smoking	−0.113**	−0.148**	−0.228***	−0.227***	−0.288***
	(0.057)	(0.057)	(0.056)	(0.056)	(0.055)
drk	−0.748***	−0.805***	−0.788***	−0.778***	
	(0.122)	(0.122)	(0.122)	(0.122)	
BI	−0.117***	−0.140***	−0.156***		
	(0.029)	(0.028)	(0.028)		
edu	−0.031	−0.210***			
	(0.030)	(0.028)			
ms	−0.724***				
	(0.047)				
Constant	−1.317***	−1.851***	−2.140***	−2.492***	−3.919***
	(0.251)	(0.249)	(0.246)	(0.238)	(0.101)
Observations	2,553	2,553	2,553	2,553	2,553
Log Likelihood	−1,874.582	−1,991.593	−2,019.625	−2,034.846	−2,051.591
Akaike Inf. Crit.	3,761.164	3,993.187	4,047.250	4,075.692	4,107.181
Note:	*p<0.1; **p<0.05; ***p<0.01				

From the Logistic regression models comparisons above, we can clearly see that model 5 had the least AIC of 3450.0065 while model 10 had the highest AIC of 4,107.181. Therefore, here we choose model 5.

	Dependent variable:
	cbind(vlbw, total)
	Model
smoking	−0.362***
	(0.059)
drk	−0.585***
	(0.124)
BI	−0.105***
	(0.029)
ms	−0.359***
	(0.050)
mrace	−0.445***
	(0.025)
momage	−0.054*
	(0.028)
Constant	−0.680***
	(0.259)
Observations	2,553
Log Likelihood	−1,716.265
Akaike Inf. Crit.	3,446.531
Note:	*p<0.1; **p<0.05; ***p<0.01

However, after conducting further analysis we find a model that is more optimal with a lower AIC of 3446.531. The table above shows the optimal model's variables, the estimates and standard errors.

We then proceed to get the adjusted main odds ratio of our best model.

Table 5: Adjusted Main Odds Ratios and CI

term	estimate	std.error	statistic	p.value	odd_ratio	lower_conf	upper_confint
(Intercept)	-0.6798920	0.2593723	-2.621298	0.0087596	0.5066717	0.3047506	0.8423815
smoking	-0.3620605	0.0589071	-6.146301	0.0000000	0.6962402	0.6203209	0.7814511
drk	-0.5852492	0.1239382	-4.722105	0.0000023	0.5569671	0.4368488	0.7101136
BI	-0.1053127	0.0285982	-3.682498	0.0002310	0.9000430	0.8509813	0.9519333
ms	-0.3587547	0.0502796	-7.135189	0.0000000	0.6985457	0.6329887	0.7708923
mrace	-0.4445555	0.0248933	-17.858452	0.0000000	0.6411092	0.6105798	0.6731651
momage	-0.0536011	0.0283328	-1.891841	0.0585122	0.9478101	0.8966109	1.0019330

Table 5 above shows the variables, estimates, standard error, statistic, p-value, odd_ratio, and the 95% confidence interval of the odds ratios. We can clearly see that the odds ratio confidence intervals contain 0 except for momage whose confidence interval includes 1. Therefore, this confirms an association among the variables.

Next, we go ahead and make a table with adjusted odds ratios for interactions between smoking and the sex of the baby, and interaction between smoking and drinking. Then create a 95% confidence interval for the odds ratios.

Interaction of smoking and bsex

	bsex absent	bsex present	Effect of bsex within the strata of smoking
	OR [95% CI]	OR [95% CI]	OR [95% CI]
smoking absent	1 [Reference]	0.93 [0.62, 1.38]	0.93 [0.62, 1.38]
smoking present	0.72 [0.51, 1]	0.68 [0.41, 1.14]	0.96 [0.79, 1.16]
Effect of smoking within the strata of bsex	0.72 [0.51, 1]	0.74 [0.64, 0.86]	
Multiplicative scale	1.03 [0.83, 1.28]		
RERI	0.04 [-0.21, 0.29]		
AP	0.06 [-0.34, 0.46]		

Based on the various odd-ratio 95% confidence intervals of interaction between smoking and sex of baby, we can clearly see that the confidence intervals contain 1, in which we strongly fail to reject the null hypothesis. Hence, we conclude that smoking and whether the sex of the baby is male or female are independent of each other.

Interaction of smoking and drk

	drk absent	drk present	Effect of drk within the strata of smoking
	OR [95% CI]	OR [95% CI]	OR [95% CI]
smoking absent	1 [Reference]	0.16 [0.08, 0.35]	0.16 [0.08, 0.35]
smoking present	0.16 [0.05, 0.53]	0.06 [0.02, 0.22]	0.37 [0.28, 0.48]
Effect of smoking within the strata of drk	0.16 [0.05, 0.53]	0.37 [0.2, 0.66]	
Multiplicative scale	2.25 [1.24, 4.08]		
RERI	0.73 [0.5, 0.97]		
AP	12.22 [-7.76, 32.21]		

On the other hand, the odd-ratio 95% confidence intervals between smoking and drinking do not contain 1; therefore, we reject the null hypothesis. Hence, we conclude that there is an association between smoking and whether drinking is present or absent.

3. CONCLUSION

From the analysis, we were able to find a suitable model that would help us predict whether or not an infant will have "vblw". The model's equation is as shown below:

$$\text{vblw}|\text{total} = -0.680 - 0.362(\text{Smoking}) - 0.585(\text{drk}) - 0.105(\text{BI}) - 0.359(\text{ms}) - 0.445(\text{mrace}) - 0.054(\text{momage})$$

We also found out that smoking and drinking during the pregnancy stage were highly associated and had a significant impact as to whether the child would be born with "vblw." In comparison, the interaction between smoking and the sex of the baby had no effect on the infant. In further research, we may investigate how the impact of the interaction between the mother's race and adequate prenatal care determines whether the child will be born with "vblw" or not.

4. REFERENCES

Wu, Samuel S., et al. "Risk factors for infant maltreatment: a population-based study." Child abuse & neglect 28.12 (2004): 1253-1264.

Agresti, Alan. Categorical data analysis. John Wiley & Sons, 2003.