
Problem Set Three

- I. The data give information about a high school honors class. We want to predict the likelihood of enrolling in an honors class **Honcomp**. Use the *hsb2.logistic* dataset.

Gender	Male	0
	Female	1
Race	White	0
	Asian	1
	AA/Black	2
	Hispanic	3
SES	Low	0
	Middle	1
	High	2
School	Public	0
	Private	1
Program	General	0
	Vocational	1
	Academic	2
Honors College Probability Plot	Not Accepted	0
	Accepted	1
Honors College Modeling	Not Accepted	1
	Accepted	0

1. Create a univariate frequency table of counts and percentages for *Honcomp* by Gender, Race, SES, School Type and Program.
2. Using chi-square analysis see if there are significant associations between the variable *Honcomp* and the variables *Gender*, *Race*, *SES*, *Schtyp* and *Prog*. Which variables are significantly associated with being enrolled in an honors course (Bivariate analysis).
3. Create a logistic regression model with *Gender*, *Race*, and *SES* to predict *Honcomp*.
4. Create a logistic regression model with *Gender*, *Race*, *SES*, and *School Type* to predict *Honcomp*.
5. Create a logistic regression model with *Gender*, *Race*, *SES*, *School Type* and *Program* to predict *Honcomp*.
6. In **ONE** table, for each model present the following—see example on page 4:
 1. Beta coefficients, standard error and odds ratio
 2. Negative log likelihood ratio, model chi-square, df and p-value, the Nagelkerke R^2 or Generalized R^2 , the whole-model p-value or Hosmer and Lemeshow Test p-value, and the classification accuracy (1 - misclassification rate).
7. Produce a table of predicted probabilities for *Gender*, *Race*, *SES*, *Schtyp* and *Prog* by the variable *Honcomp*.
8. Our interest is the association between *Gender*, *Race*, and *SES* and taking an honors course (*Honcomp*), but also how the relationship between *Gender*, *Race*, and *SES* changes as we account for other explanatory variables (*School Type* and *Program*).
 - A. How did the model change as you added the remaining two variables (*School Type* and *Program*)?
 - B. Interpret the significant predictors in the final model using odds ratios.
 - C. Discuss the final model's classification accuracy
9. Produce a ROC plot for the final model and interpret.

- II. The data collected were academic information on 316 students. The response variable is days absent during the school year (**Days_Absent**), from which we explore its relationship with math standardized tests score (**Math_Score**), language standardized tests score (**Language_Score**) and gender (**Gender: 0 = Male and 1 = Female**). Using the **Poisson** dataset answer the following questions:
1. Provide the mean, SD, min and max values for the variables **Math_Score** and **Language_Score**.
 2. Provide the mean, median, SD, min and max values for the variable **Days_Absent**.
 3. Create a histogram for the **Days_Absent**, **Math_Score** and **Language_Score** and comment on the distributions.
 4. Estimate a Poisson regression model for the number of days absent during the school year.
 5. Is the model over-dispersed?
 6. Produce a studentized deviance residual by predicted plot and comment on over-dispersion.
 7. In **ONE** table, present the following—see example on page 5:
 - A. Beta coefficients, SE and IRR, the chi-square/df value, along with the omnibus Test Results.
 - B. Interpret the incident rate ratios (IRR) for each predictor.
 8. Rerun the analysis using a negative binomial with a log-link function. In **ONE** table, present the following—see example on page 5:
 - A. Beta coefficients, SE and IRR, the chi-square/df value, along with the omnibus Test Results.
 - B. Interpret the incident rate ratios (IRR) for each predictor.
 9. Compare and contrast the IRR's from the Poisson and Negative Binomial models.

Table for Logistic Regression

	Model 1				Model 2				Model 3			
	B	Sig	SE	OR	B	Sig	SE	OR	B	Sig	SE	OR
Constant	1.20		0.82	3.34	1.35		0.90	3.85	1.36		0.92	3.89
Gender												
Male					Reference							
Female	0.73	*	0.36	2.08	0.74	*	0.36	2.10	0.73	*	0.37	2.08
Race												
White					Reference							
Asian	-1.47		0.78	0.22	-1.47		0.78	0.22	-1.58	*	0.80	0.20
AA/Black	-2.07	*	0.99	0.12	-2.05	*	0.99	0.12	-2.17	*	1.03	0.11
Hispanic	0.025		1.02		0.03		1.07	1.03	-0.40		1.08	0.96
SES												
Low					Reference							
Middle	0.83		0.48	2.29	0.83		0.48	2.29	0.48		0.51	1.62
High	1.43	**	0.29	4.19	1.43	**	0.29	4.19	1.18	**	0.41	3.27
School												
Public					Reference							
Private					-0.18		0.48	0.82	-0.46		0.50	0.63
Program												
General					Reference							
Vocational									1.24	**	0.49	3.47
Academic									1.32	**	0.52	3.75
-2ll			202.41				202.26				190.81	
Hosmer-Lemeshow			C ² (6)=2.82, p = 0.830				C ² (7)=6.16, p = 0.521				C ² (8)=8.27, p = 0.407	
Nagelkerke R²			0.196				0.197				0.267	
Classification Accuracy			74.5%				74.5%				77.0%	

* P < 0.05; **P < 0.01

Table for Count Models

	Poisson Model				Negative Binomial Model			
	B	Sig	SE	IRR	B	Sig	SE	IRR
Constant	2.68	**	0.07	14.69	2.71	**	0.20	10.04
Gender								
Female				Reference				
Male								
Math Score								
Language Score								
Pearson Chi-Square/DF						1.34		
Onibus Test						C ² (3)=26.73, p < 0.001		