Case Study: SAS BRFSS 10 Categorical Table 2 and 3

This case study will continue your reintroduction to categorical variables and how you may use statistical procedures to investigate these types of data, display them in a "Table 2", and develop a results section describing your Table 2.

This case study will also introduce you to *adjusted* analysis of categorical variables and how you may use statistical procedures to investigate these types of data, display them in a "Table 3", and develop a results section describing your Table 3.

Continue using data from the Behavioral Health Needs Assessment Survey from 2010 to complete this case study.

The objective of this analysis is to investigate the association between **diabetes** and **BMI** after controlling for **exercise** and **sex**. The outcome variable is **diabetes** and the variable of interest (exposure) is **BMI**.

Conduct a **complete case, case control analysis** for this objective following these guidelines:

1. Use the raw variable categorization of BMI (_BMI4CAT)
2. Categorize gender (SEX) into a two-level variable (male=0, female=1) where male is category 1 of the raw variable and female is category 2
3. Categorize diabetes (DIABETE2) into a two-level variable (no=0, yes=1) where yes is category 1 of the raw variable and no is category 3
4. Categorize exercise (EXERANY2) into a two-level variable (no=0, yes=1) where yes is category 1 of the raw variable and no is category 2
5. For the complete case analysis, restrict your sample based on the following conditions:
    a. 18<=AGE<=99
    b. SEX: raw categories 1 and 2
    c. DIABETE2: raw categories 1 and 3
    d. EXERANY2: raw categories 1 and 2
    e. Education (EDUCA): raw categories 1-6
    f. _BMI4CAT: raw categories 1,2 and 3
    g. General health (GENHLTH): raw categories 1-5

1) (4 pts) Using PROC FREQ, show the simple frequency tables for sex, exercise, BMI, and diabetes.

The FREQ Procedure

sex_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	162430	39.26	162430	39.26
1	251318	60.74	413748	100.00

exerany2_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	111531	26.96	111531	26.96
1	302217	73.04	413748	100.00

_BMI4CAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	145352	35.13	145352	35.13
2	151781	36.68	297133	71.81
3	116615	28.19	413748	100.00

diabete2_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	360098	87.03	360098	87.03
1	53650	12.97	413748	100.00

2) (15 pts) Create your "Table 2" for this objective. You can use this table template:

Table 2. Characteristics of 413,748 BRFSS 2010 participants by presence of diabetes.

	Population		Diabetes - No		Diabetes - Yes		
Variable	N	%	n	%	n	%	p value *
	413,748	100%	360,098	87.0%	53,650	13.0%	
BMI							
Normal	145,352	35.3%	137,163	38.1%	8,189	15.3%	
Overweight	151,781	36.7%	134,539	37.4%	17,242	32.1%	
Obese	116,615	28.2%	88,396	24.6%	28,219	52.6%	<0.0001
Sex							
Male	162,430	39.3%	139,820	38.8%	22,610	42.1%	
Female	251,318	60.7%	220,278	61.2%	31,040	57.9%	<0.0001
Exercise							
No	111,531	27.0%	89,873	25.0%	21,658	40.3%	
Yes	302,217	73.0%	270,225	75.0%	31,992	59.6%	<0.0001

* p values based on Pearson chi-square test of association

3) (21 pts) Write the results section for this "Table 2".

Of the 451,075 BRFSS 2010 participants, 413,748 (91.7%) had complete data for the objective. The demographic characteristics of this population are compared in Table 1. Of the entire population, 60.7% were female, 73.0% exercised, and 13.0% had diabetes. There were proportionately more males than expected that had diabetes: 42.1% vs. 39.3% (p<0.0001). There were proportionately more non-

exercisers that had diabetes: 40.3% vs. 27.0% (p<0.0001). There were proportionately more people with an Obese BMI that had diabetes: 52.6% vs. 28.2% (p<0.0001).

4) (15 pts) Based on the information in your Table 2,
   a. Fill in Table 2B showing both the odds and probability of having diabetes:

TABLE2B. Odds and probabilities of having diabetes based on the characteristics of 413,748 BRFSS 2010 participants.

	Odds of Having Diabetes	Probability of Having Diabetes
Male	0.162	13.92%
Female	0.141	12.35%
Exerciser	0.118	10.59%
Non-exercisers	0.241	19.42%
Obese BMI	0.319	24.20%
Overweight BMI	0.128	11.36%
Normal BMI	0.060	5.63%

   b. Based on the odds of having diabetes, fill in Table 2C showing the odds ratio (OR). Also interpret both the sign and magnitude of each odds ratio in the space given.

(31,040 / 220,278) / (22,610 / 139,820) = 0.871
(31,992 / 270,225) / (21,658 / 89,873) = 0.491
(28,219 / 88,396) / (8,189 / 137,163) = 5.347

TABLE2C. Odds ratios of having diabetes based on the characteristics of 413,748 BRFSS 2010 participants.

	Odds Ratios (OR)	Interpretation of Odds Ratio (OR)
Female to male	0.871	Females are 0.871 times more likely to have diabetes compared to males.
Exerciser to non-exerciser	0.491	Exercisers are 0.491 times more likely to have diabetes than non-exercisers.
Obese BMI to normal BMI	5.347	People who have an Obese BMI are 5.347 times more likely to have diabetes than people who have a Normal BMI.

5) (5 pts) This study furthers our investigation by conducting a logistic regression explaining the presence of diabetes. Write the proposed model for this analysis.
DIABETES = f (SEX, EXERCISE, BMI)

SEX = 0/1 (male/female)
EXERCISE = 0/1 (does not exercise/does exercise)
BMI = 1/2/3 (Normal BMI/Overweight BMI/Obese BMI)

We can also write the model specification like this:
DIABETES = $\alpha + \beta_1$SEX + β_2EXERCISE + β_3BMI

6) (57 pts) Using SAS PROC LOGISTIC, run a logistic regression that estimates your model using a CLASS statement which specifies the reference category for each variable. As reference categories, use the following: DIABETES = 0; EXERCISE = 0; SEX = 0 _BMI4CAT = 1.
   a. Show the Analysis of Maximum Likelihood Estimates (MLEs) output table.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3337	0.0151	23910.6798	<.0001
sex_1	1	1	-0.0989	0.00977	102.4497	<.0001
exerany2_1	1	1	-0.5682	0.00994	3268.4451	<.0001
_BMI4CAT	2	1	0.7347	0.0141	2715.0696	<.0001
_BMI4CAT	3	1	1.5957	0.0134	14189.1763	<.0001

      i. Explain both the sign and magnitude of the coefficient "estimate" for **EXERCISE.**
      The magnitude (β) for EXERCISE is -0.5682. Since the sign is negative ($\beta < 0$), this means the variable is negatively related (as one variable increases, the other variable decreases, or vice versa) with our dependent variable (Diabetes). This tells us that it's more likely that exercisers will <u>not have</u> diabetes, and it's more likely that non-exercisers <u>will have</u> diabetes.

   b. Show the Odds Ratio Estimates output table.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex_1 1 vs 0	0.906	0.889	0.923
exerany2_1 1 vs 0	0.567	0.556	0.578
_BMI4CAT 2 vs 1	2.085	2.028	2.143
_BMI4CAT 3 vs 1	4.932	4.804	5.063

      i. Explain both the sign and magnitude of the odds ratio "point estimate" for **EXERCISE.**
      Exercisers are 0.567 times more likely to have diabetes than non-exercisers. If we include the 95% confidence interval, exercisers are

between 0.556 and 0.578 times more likely to have diabetes than non-diabetes.

c. Show how the odds ratio for **EXERCISE** is calculated using the coefficient estimates shown in the Analysis of MLEs table.
The odds ratio for EXERCISE can be calculated by taking the e^x of the coefficient estimate. The numbers are very similar, with a very slight difference.

Coefficient estimate for EXERCISE = $e^{-0.5682} = 0.5696$
Odds ratio for EXERCISE = 0.567

d. For each of the following goodness of fit tests, explain: (1) What question the test is attempting to answer; (2) what is the null hypothesis H_0; and (3) What the test statistic shows in this case supporting your answer with the relevant SAS output table.
   i. Log likelihood/AIC
   1. Is this model better than some other model?
   2. There is no H_0.
   3. Log likelihood/AIC can only be used to compare models estimated using the same data since there is no standard for determining whether any given value represents a good fitting model. If it's being compared with another model, smaller values of -2LogL or the AIC would indicate a better fitting model. In this case, since we don't have a model to compare to, Log likelihood/AIC is not very helpful.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	319213.28	296332.02
SC	319224.21	296386.68
-2 Log L	319211.28	296322.02

   ii. Global chi-squared
   1. Is this model better than no model? Specifically, if this model is better than one model with just an intercept. The global chi-squared test compares the log likelihood from an intercept-only model with that from an intercept plus covariates model.
   2. H_0 = all β on the current model equal 0.
   3. The Likelihood Ratio test has a chi-square value of 22889.2573, the Score test has a chi-square value of 23660.3049, and the Wald test has a chi-square value of 21059.4984. They all have p-values that are <0.0001. Therefore, since it is <0.05, we can reject H_0.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22889.2573	4	<.0001
Score	23660.3049	4	<.0001
Wald	21059.4984	4	<.0001

iii. Deviance chi-squared

1. Is there a better model than this one? Specifically, if the saturated model yields a better fit to the data than the unsaturated (current) model.

2. H_0 = the coefficients on the interaction terms are jointly equal to 0.

3. The Deviance has a value of 303.9816. The Pearson deviance has a value of 304.0900. Both the deviance and the Pearson deviance p-values are <0.0001. Therefore, since it is < 0.05, we can reject H_0.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	303.9816	7	43.4259	<.0001
Pearson	304.0900	7	43.4414	<.0001

Number of unique profiles: 12

iv. Hosmer and Lemeshow chi-squared

1. Is there a better model than this one?

2. H_0 = the current model fits the data well. A high p-value means the H_0 cannot be rejected and we conclude the model cannot be improved by adding additional terms, such as non- linearities and or interactions.

3. The HL chi-square statistic is 0.012 with a p-value of 0.994. Therefore, since the p-value is < 0.05, we can reject H_0 that the current model fits the data well.

Partition for the Hosmer and Lemeshow Test					
		diabete2_1 = 1		diabete2_1 = 0	
Group	Total	Observed	Expected	Observed	Expected
1	79687	3138	3776.22	76549	75910.78
2	33517	2132	1744.82	31385	31772.18
3	32148	2919	2667.95	29229	29480.05
4	57372	5388	5391.12	51984	51980.88
5	57035	5756	5859.34	51279	51175.66
6	37374	6098	5991.54	31276	31382.46
7	42152	8663	8304.02	33489	33847.98
8	32454	6915	6916.47	25539	25537.53
9	42009	12641	12998.51	29368	29010.49

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
267.4774	7	<.0001

e. Compare the *unadjusted* OR from Table 2C above and the *adjusted* OR from your logistic regression:

i. Fill in the table below with these odds ratios

	Unadjusted Odds Ratios (OR)	Adjusted Odds Ratios (OR)
Female to male	0.871	0.906
Exerciser to non-exerciser	0.491	0.567
Obese BMI to normal BMI	5.347	4.932

ii. In what sense are the ORs obtained from the logistic regression "adjusted"? Does this account for the differences between the unadjusted and adjusted ORs? Explain.

The "adjusted" OR considers our control variables (Sex, Exercise, BMI) into the analysis, while the "unadjusted" OR does not. Because of this, the unadjusted OR and the adjusted OR are slightly different (see table above).

7) (12 pts) Create your "Table 3" for this objective. You can use this table template:

Table 3. Logistic regression analysis comparing the adjusted odds ratio of diabetes in 116,615 obese BRFSS 2010 participants when compared to participants with normal BMI after controlling for exercise and sex.

	Diabetes - No		Diabetes - Yes		OR*	95% CI
Variable	n	%	n	%		
	360,098	87.0%	53,650	13.0%	-----	-----
BMI						
Normal	137,163	38.1%	8,189	15.3%	-----	-----
Overweight	134,539	37.4%	17,242	32.1%	2.085	2.028 - 2.143
Obese	88,396	24.6%	28,219	52.6%	4.804	4.804 - 5.063
Sex						
Male	139,820	38.8%	22,610	42.1%	-----	-----
Female	220,278	61.2%	31,040	57.9%	0.906	0.889 - 0.923
Exercise						
No	89,873	25.0%	21,658	40.3%	-----	-----
Yes	270,225	75.0%	31,992	59.6%	0.567	0.556 - 0.578

* 95% confidence intervals are for reported odds ratios.

8) (21 pts) Write the Table 3 results section/interpretation.

Adjusted odds ratios from the logistic regression are presented in Table 3 above. Those who reported exercising had roughly half the odds of reporting diabetes when compared to those not exercising after controlling for age and gender (OR=0.567; 95% CI = 0.556-0.578). Females were at slightly lower odds of reporting diabetes when compared to males after controlling for age and exercise (OR=0.906; 95% CI = 0.889-0.923). Those who had an Obese BMI had substantially greater odds (4.804 times) of reporting diabetes as compared to those who had a Normal BMI after controlling for BMI and Sex (OR=4.804; 95% CI = 4.804 - 5.063). Those who had an Overweight BMI had higher odds (2.085 times) of reporting diabetes as compared with those who had a Normal BMI after controlling for BMI and Sex (OR=2.085; 95% CI = 2.028 - 2.143).

Extra Credit (15 pts)

Estimate a fully saturated logistic model.

DIABETES = f (SEX, EXERCISE, BMI, SEX*EXERCISE, SEX*BMI, EXERCISE*BMI, SEX*EXERCISE*BMI)

1. Show the Maximum Likelihood Estimation (MLE) table for this regression.

Analysis of Maximum Likelihood Estimates								
Parameter				DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept				1	-2.1849	0.0340	4141.0873	<.0001
sex_1	1			1	-0.1729	0.0414	17.4496	<.0001
exerany2_1	1			1	-0.5043	0.0407	153.7992	<.0001
_BMI4CAT	2			1	0.5385	0.0403	178.2563	<.0001
_BMI4CAT	3			1	1.3486	0.0385	1225.4342	<.0001
exerany2_1*sex_1	1	1		1	-0.3322	0.0505	43.3447	<.0001
sex_1*_BMI4CAT	1	2		1	0.1926	0.0502	14.7147	0.0001
sex_1*_BMI4CAT	1	3		1	0.1628	0.0471	11.9659	0.0005
exerany2_1*_BMI4CAT	1	2		1	-0.0363	0.0482	0.5674	0.4513
exerany2_1*_BMI4CAT	1	3		1	0.0342	0.0466	0.5378	0.4634
exeran*sex_1*_BMI4CA	1	1	2	1	0.2327	0.0613	14.4374	0.0001
exeran*sex_1*_BMI4CA	1	1	3	1	0.2966	0.0581	26.0399	<.0001

2. Show the deviance output table for this regression. Does the estimated deviance statistic make sense? Explain.
   No, the estimated deviance statistic does not make sense because both Deviance and Pearson values are 0, 0 degrees of freedom, and no data for the Value/DF and p-values fields.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.0000	0	.	.
Pearson	0.0000	0	.	.

Number of unique profiles: 12

3. Identify which interactions should be removed.
   EXERCISE*BMI

4. Estimate the final model and show the MLE table.
   DIABETES = f (SEX, EXERCISE, BMI, SEX*EXERCISE, SEX*BMI, SEX*EXERCISE*BMI)

Analysis of Maximum Likelihood Estimates								
Parameter				DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept				1	-2.1871	0.0217	10180.4650	<.0001
sex_1	1			1	-0.1708	0.0321	28.3108	<.0001
exerany2_1	1			1	-0.5013	0.0157	1013.3911	<.0001
_BMI4CAT	2			1	0.5126	0.0221	540.1814	<.0001
_BMI4CAT	3			1	1.3707	0.0216	4011.4453	<.0001
exerany2_1*sex_1	1	1		1	-0.3352	0.0338	98.5841	<.0001
sex_1*_BMI4CAT	1	2		1	0.2185	0.0372	34.5927	<.0001
sex_1*_BMI4CAT	1	3		1	0.1408	0.0346	16.5066	<.0001
exeran*sex_1*_BMI4CA	1	1	2	1	0.1964	0.0378	26.9719	<.0001
exeran*sex_1*_BMI4CA	1	1	3	1	0.3307	0.0348	90.4689	<.0001

5. Show the final model deviance output table. Explain what the deviance statistic now shows.

   Now, the Deviance and Pearson values are available. The Deviance has a value of 4.1961. The Pearson deviance has a value of 4.2030. I removed the interaction between EXERCISE*BMI because it was not statistically significant (since its p-value was <0.05) in the MLE table. Since all the interactions are now statistically significant, the deviance statistic now shows.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	4.1961	2	2.0980	0.1227
Pearson	4.2030	2	2.1015	0.1223

Number of unique profiles: 12