Case Study: SAS BRFSS 10 Categorical Table 3

This case study will build on last week's introduction to adjusted analyses with categorical variables and dive into the area of model building. There are many ways in which analysts come to a final model and we will explore some of them in this case study. The case study will end with a more developed "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, age, education, general health**, and **sex**. The outcome variable is **diabetes,** and the variable of interest (exposure) is **BMI**.

Conduct a **complete case 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. Categorize general health (GENHLTH) into a three-level variable: (raw category 1 = 2; raw categories 2 and 3 = 1; raw categories 4 and 5 = 0)
6. Categorize education (EDUCA) into a three-level variable: (raw categories 1 and 2 = 1), raw categories 3 and 4 = 2), (raw categories 5 and 6 = 3)
7. Categorize age (AGE) into a three-level variable: (18<=age<=34 = 1, 35<=age<=54 = 2, 55<=age = 3)
8. 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) (15 pts) Expand Table 2 to include the 3 new categorized variables age, education, and general health. Use the template below. Double check that the values in the pre-populated cells match your SAS output. Note: use the following classifications for Education and General Health:

If EDUCA 1 or 2 = "Elementary"	**If GENHLTH 1 = "Excellent"**
If EDUCA 3 or 4 = "HS"	**If GENHLTH 2 or 3 = "Good"**
If EDUCA 5 or 6 = "College"	**IF GENHLTH 4 or 5 = "Poor"**

The SAS code for defining the set of complete cases has not changed from homework #2 but has been amended to include the categorization of the additional variables:

```
/* subset data and create variables */
data work.tmp; set work.cdbrfs10;
       where (18<=age<=99) and (sex in(1,2)) and (diabete2 in(1,3)) and
       (exerany2 in(1,2)) and (educa in (1,2,3,4,5,6))
       and (_bmi4cat in(1,2,3)) and (genhlth in(1,2,3,4,5)) ;

       if 18<=age<=34            then agecat = 1;
       if 35<=age<=54            then agecat = 2;
       if 55<=age                then agecat = 3;

/* have you ever been told by a doctor that you have diabetes?  1 = yes, 3 =
no */
       if diabete2=1       then diabetes = 1;
       if diabete2=3       then diabetes = 0;

/* did you participate in any physical activity during the past month? 1 =
yes, 2 = no */
       if exerany2=1       then exercise = 1;
       if exerany2=2       then exercise = 0;

/* sex: 1 = male, 2 = female */
       if sex = 2 then sex2 = 1;
       if sex = 1 then sex2 = 0;

/* education: 1 = no school, 2 = elemen, 3 = some hs, 4 = HS grad, 5 = some
college, 6 = college grad+, 9 = refused */
       if educa in (1,2) then educat = 1;
       if educa in (3,4) then educat = 2;
       if educa in (5,6) then educat = 3;

       /* bmi: 1 = ok 2 = overweight 3 = obese */
       if _bmi4cat = 1 then bmi = "normal    ";
       if _bmi4cat = 2 then bmi = "overweight";
       if _bmi4cat = 3 then bmi = "obese";

/* genhlth: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor, 7 =
dk, 9 = refused */
       if genhlth = 1 then healthcat = 2;
       if genhlth in(2,3) then healthcat = 1;
       if genhlth in(4,5) then healthcat = 0;

       drop sex;

run;

data work.tmp; set work.tmp; rename sex2 = sex; run;
```

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	100.0%	360,098	87.0%	53,650	13.0%	
BMI							
Normal	145,352	35.1%	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	24.5%	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	40.4%	
Yes	302,217	73.0%	270,225	75.0%	31,992	59.6%	<0.0001
Age							
18 to 34	44,159	10.7%	43,260	12.0%	899	1.7%	
35 to 54	133,743	32.3%	124,039	34.4%	9,704	18.1%	
55+	235,846	57.0%	192,799	53.5%	43,047	80.2%	<0.0001
Education							
Elementary	12,867	3.1%	9,549	2.7%	3,318	6.2%	
HS	149,121	36.0%	125,304	34.8%	23,817	44.4%	
College	251,760	60.8%	225,245	62.6%	26,515	49.4%	<0.0001
General Health							
Poor	82,162	19.9%	57,115	15.9%	25,047	46.7%	
Good	258,008	62.4%	231,036	64.2%	26,972	50.3%	
Excellent	73,578	17.8%	71,947	20.0%	1,631	3.0%	<0.0001

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

2) (20 pts) Amend your Homework #2 results section for this expanded Table 2.

The demographic characteristics of the BRFSS 2010 population are compared in Table 2 with respect to the presence of diabetes. Overall, 13% of the entire population had a diagnosis of diabetes. There were proportionately fewer females than expected diagnosed with diabetes (57.9% vs. 60.7%; $p<0.0001$) and proportionately fewer exercisers than expected diagnosed with diabetes (59.6% vs. 73.0%; $p<0.0001$). With respect to BMI, there were proportionately fewer normal and overweight BMI diagnosed with diabetes than expected (15.3% vs. 35.1% and 32.1% vs. 36.7%; $p<0.0001$). But there were there proportionately more obese BMI diagnosed with diabetes than expected (52.6% vs. 28.2%; $p<0.0001$). With respect to additional control variables, there were proportionately more aged 55+ than expected diagnosed with diabetes (80.2% vs. 57.0%; $p<0.0001$), proportionately fewer college educated than expected diagnosed with diabetes (49.4% vs. 60.8%; $p<0.0001$), and proportionately fewer people with excellent general health than expected diagnosed with diabetes (3.0% vs. 17.8%; $p<0.0001$).

3) **(15 pts) Based on the data in the expanded Table 2, which are the top 3 categories in terms of the unadjusted probability of having diabetes? Use the template below.**

As shown in the table below, the top 3 categories are poor health (probability = 30.5%), an elementary (or less) education (25.8%) and an obese BMI (24.2%). Remember the odds are calculated as the ratio of the "yes" to the "no" in Table 2 above. The probability is found as odds / (1+odds).

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	13.9%
Female	0.141	12.4%
Exerciser	0.118	10.6%
Non-exercisers	0.241	19.4%
Normal BMI	0.060	5.6%
Overweight BMI	0.128	11.4%
Obese BMI	0.319	24.2%
Age 18 to 34	0.021	2.0%
Age 35 to 54	0.078	7.3%
Age 55+	0.223	18.3%
Education: Elem	0.347	25.8%
Education: HS	0.190	16.0%
Education: College	0.118	10.5%
Poor health	0.439	30.5%
Good health	0.117	10.5%
Excellent health	0.023	2.2%

4) **(5pts) Write the proposed model for this objective.**

Like what we did for homework #2, but now with more control variables,

Diabetes = f(BMI, exercise, sex, age, education, general health)

where the "f" is shorthand for "a function of". We can be a bit more specific by stating:

Diabetes(yes/no) = f(_BMI4CAT (normal, overweight, obese), EXERCISE (yes/no), SEX(male/female), AGE (18-34, 35-54, 55+), EDUCATION (elementary, high school, college), HEALTH (poor, good, excellent))

So, our **outcome** variable is diabetes, our **exposure** variable is BMI, and our **control** variables are sex, exercise, age, education, and general health.

5) **(15 pts) Estimate a base logistic regression model. <u>Do not</u> use a CLASS statement specifying reference values for each variable. Use the raw continuous variable AGE…not your categorized version. (*Note – we have a continuous variable AGE from which we created a categorized version. We can use one or the other in our model. Here we are using the continuous version.*) Use the DESCENDING option to order your dependent variable diabetes from highest to lowest (1 to 0).**

```
proc logistic data=work.tmp descending;
      model diabetes = AGE var2 .... varN; run;
```

Based on the variable names shown above in question 1, the SAS code is:

```
/* base logistic model */
proc logistic data=work.tmp descending;
      model diabetes = exercise sex healthcat age educat _bmi4cat / lackfit;
      run;
```

a. **Interpret the sign on the coefficient (β) of AGE. Show the relevant SAS output.**

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6489	0.0402	13390.7938	<.0001
exercise	1	-0.1148	0.0108	111.9599	<.0001
sex	1	-0.1833	0.0102	320.7196	<.0001
healthcat	1	-1.1068	0.00938	13930.6419	<.0001
AGE	1	0.0367	0.000359	10447.3192	<.0001
educat	1	-0.0789	0.00878	80.8301	<.0001
_BMI4CAT	1	0.8380	0.00695	14559.9362	<.0001

Because we used the DESCENDING option, a positive coefficient means the variable is **positively** related to the diagnosis of diabetes. Since the sign on AGE is > 0, **<u>older</u> people are more likely to have diabetes than <u>younger</u> people**.

b. **Interpret the OR value for AGE (remember AGE is a continuous variable). Show the relevant SAS output.**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
exercise	0.892	0.873	0.911
sex	0.833	0.816	0.849
healthcat	0.331	0.325	0.337
AGE	1.037	1.037	1.038
educat	0.924	0.908	0.940
_BMI4CAT	2.312	2.281	2.343

Subtracting 1 from the OR and multiplying by 100 gives us the % change in the odds from a 1-unit change in the variable. So, **a one-year increase** in AGE is associated with a (1.037 -1)*100 or **3.7% increase in the odds** of having diabetes, after controlling for exercise, health, education and BMI.

c. Interpret the OR value for BMI.

For a multicategory variable like _BMI4CAT, the same interpretation applies: moving from one category of BMI to next "higher" category is associated with a (2.312 – 1)*100 or 131% increase in the odds of having diabetes, after controlling for exercise, health, education and age.

Note – since we did not use a CLASS statement, SAS did not know what reference category to use. So, what we get is an average effect across the categories. That is, whether the BMI change is from "normal" to "overweight" or from "overweight" to "obese", the estimated effect is the same: 131% increase in the odds. When we use a CLASS statement later, we will see that the association in fact differs across the BMI categories.

6) (10 pts) Investigate whether multicollinearity exists among the variables using PROC REG (switch to using the categorized version of AGE). Which two variables have the highest VIF values and what does that mean? Is multicollinearity an issue with these variables? Show the relevant SAS output.

Using PROC REG with the VIF option:

```
/* multicolinearity test */
proc reg data=work.tmp;
      model diabetes = agecat exercise sex _bmi4cat educat healthcat / vif;
run;
```

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-0.01091	0.00353	-3.09	0.0020	0
agecat		1	0.07020	0.00073247	95.84	<.0001	1.03162
exercise		1	-0.02343	0.00116	-20.20	<.0001	1.10224
sex		1	-0.00959	0.00101	-9.49	<.0001	1.01312
_BMI4CAT	COMPUTED BODY MASS INDEX CATEGORIES	1	0.07096	0.00063557	111.65	<.0001	1.05613
educat		1	-0.01698	0.00092178	-18.43	<.0001	1.08260
healthcat		1	-0.10481	0.00086334	-121.39	<.0001	1.16609

The variance inflation factor (VIF) measures how collinear the variables are in the model (i.e., how correlated they are). Multicollinearity among variables can lead to coefficient (β) standard errors being too large which makes the variables statistically nonsignificant (t-values too low). That is the variance is "inflated". Essentially multicollinearity means we have redundant information. Large VIF's greater than 10 (but should start to get concerned when they are around 5) indicates the presence multicollinearity. In this case all VIF are < 2 so we do not have an issue. The general health and exercise variables have the highest VIF (1.166 and 1.102).

7) **(15 pts) Investigate whether any of the control variables are true confounding variables with respect to the exposure variable _BMI4CAT using PROC LOGISTIC. <u>Do not</u> use a class statement specifying reference values for each variable. <u>Continue using the categorized version of AGE for this question and all subsequent questions</u>. Compute the % change in the OR for _BMI4CAT when each of the variables are removed from the model. Use the following template. What is your conclusion? Is there a relationship between BMI and the presence of diabetes? Explain.**

Confounding is when a third variable clouds the estimated relationship between two other variables. In our case, we are interested in the relationship between BMI and the presence of diabetes. Suppose we estimate the simple logistic model, diabetes = f(BMI), and find a statistically significant coefficient (β) on BMI. It is possible that there is another variable that is correlated with BMI that has the true association with diabetes and not BMI. Including this variable in the model could render the β on BMI statistically nonsignificant while the new variable's β is statistically significant. This would be a dead giveaway of a strong "confounding effect." But confounding can exist even if both variables are statistically significant when both are in the model.

To test whether any of our control variables are confounders we can drop each variable one at a time and see what happens to the OR on BMI. If the % change in BMI's OR is large, then we might want to investigate any partial confounding effect.

Using our base PROC LOGISTIC model using the categorized version of AGE, here is the first run where EXERCISE has been removed:

```
/* confounding test - remove each variable except exposure...measure % change
in the OR of the exposure variable */
proc logistic data=work.tmp descending;
*      model diabetes = exercise sex healthcat agecat educat _bmi4cat;
       model diabetes =          sex healthcat agecat educat _bmi4cat;
run;
```

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.861	0.844	0.878
healthcat	0.319	0.313	0.325
agecat	2.788	2.732	2.846
educat	0.869	0.854	0.884
_BMI4CAT	2.173	2.144	2.202

The results of the full test are shown below.

	Exposure (_BMI4CAT)	
Variable	**OR**	**% Chg OR**
_BMI4CAT	2.159	
Remove exercise	2.173	0.65%
Remove sex	2.166	0.32%
Remove general health	2.325	7.69%
Remove age	2.039	-5.56%
Remove education	2.159	0.00%

With all variables in the model, the OR for our BMI variable is 2.159. Removal of GENERAL HEALTH and AGE have the greatest effect on the OR of BMI, but the effect is under the 10% threshold value. So, we may have a small amount of confounding and, to be safe, we should keep general health and age in the model.

8) (20 pts) Re-estimate your logistic model, now using the CLASS statement showing the reference categories for each variable. Use the following reference category for each variable:

Sex – male
Exercise – no exercise
BMI – normal
Age – 18 to 34
Education – elementary
General Health – poor

```
/* final model */
proc logistic data=work.tmp;
       class diabetes (ref='0') exercise (ref='0') sex (ref='0')
        _bmi4cat (ref='1') healthcat (ref='0') agecat (ref='1') educat
       (ref='1') / param=ref;
       model diabetes = exercise sex healthcat agecat educat _bmi4cat /
       lackfit;
run;
```

a. Show the SAS output table which presents the model coefficients (β's).

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.1812	0.0432	5434.8199	<.0001
exercise	1	1	-0.1409	0.0109	168.2041	<.0001
sex	1	1	-0.1656	0.0103	259.2367	<.0001
healthcat	1	1	-1.0751	0.0109	9708.3419	<.0001
healthcat	2	1	-2.3554	0.0270	7607.6319	<.0001
agecat	2	1	1.1250	0.0359	982.6820	<.0001
agecat	3	1	2.1277	0.0347	3761.6806	<.0001
educat	2	1	-0.1716	0.0235	53.5508	<.0001
educat	3	1	-0.2912	0.0235	152.9884	<.0001
_BMI4CAT	2	1	0.6714	0.0146	2125.4736	<.0001
_BMI4CAT	3	1	1.5035	0.0140	11556.3011	<.0001

i. Since the estimated β's for BMI are statistically different from zero, is this proof that the included control variables are not confounding the relationship between BMI and diabetes? Explain.

Not necessarily. A confounder is a variable that is correlated with both the outcome and exposure variable. A control variable that is correlated with the exposure variable (BMI) can enter with a statistically significant coefficient. And its correlation with BMI can be to a degree that does not render the β on BMI statistically insignificant. But its presence can affect the size of β on BMI. The extent to which the β on BMI is being "confounded" is best determined as we did in Q #7. This drop-variable approach we used suggests a minimal level of confounding.

b. Show the SAS output table which presents the adjusted ORs.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
exercise 1 vs 0	0.869	0.850	0.887
sex 1 vs 0	0.847	0.830	0.865
healthcat 1 vs 0	0.341	0.334	0.349
healthcat 2 vs 0	0.095	0.090	0.100
agecat 2 vs 1	3.080	2.871	3.305
agecat 3 vs 1	8.395	7.844	8.986
educat 2 vs 1	0.842	0.804	0.882
educat 3 vs 1	0.747	0.714	0.783
_BMI4CAT 2 vs 1	1.957	1.902	2.014
_BMI4CAT 3 vs 1	4.497	4.376	4.622

i. What are the top 3 categories in terms of the adjusted ORs?

From the output above we see, ranked from largest to smallest, the top 3 adjusted ORs are: age 55+ vs. age 18-34 (8.395), "obese" BMI vs. "normal" BMI (4.497) and age 34-54 vs. age 18-34 (3.080)

ii. Do these top 3 categories match what you find if you calculate the unadjusted OR using Table 2? What accounts for the difference if there is any? Explain. Use the following table template to support your answer:

We can create a table which shows the unadjusted OR from Table 2 above next to the adjusted OR from the logistic regression as follows.

	Odds Ratios (OR)	Adjusted Odds Ratios
Female to male	0.871	0.847
Exerciser to non-exerciser	0.491	0.869
Overweight to normal BMI	2.147	1.957
Obese to normal BMI	5.347	4.497
Age 35-54 to 18-24	3.765	3.080
Age 55+ to 18-24	10.744	8.395
HS to elementary education	0.547	0.842
College to elementary education	0.339	0.747
Good to poor health	0.266	0.341
Excellent to poor health	0.052	0.095

So, the answer is, yes, the logistic regression yields the same top 3 categories but the **magnitude** of the adjusted OR differs from the unadjusted. This is because the ORs obtained from the logistic regression control for other variables that are

associated with the presence of diabetes. The ORs obtained from the contingency table (Table 2) do not, so are "unadjusted".

c. So, what is your overall conclusion as to the relationship between BMI and diabetes? <u>Does your conclusion differ from what you found in Homework 2</u>? Even with the inclusion of additional control variables? Explain. Be specific in your answer using comparative analysis of estimated effects.

Our HW #2 findings still stand: there is a statistically significant relationship between BMI and the diagnosis of diabetes, especially for obese BMI.

In Homework #2 we found an adjusted OR for obese BMI vs normal of 4.932 (95% CI 4.804 – 5.063). After adding AGE, Education and Health as additional control variables, the adjusted OR for obese BMI vs normal is 8.8% lower but still sizable at 4.497 (95% CI 4.376 – 4.622).

Likewise, for overweight BMI, the effect of controlling for additional variables resulted in a smaller estimated relationship but still statistically significant. In Homework #2, the estimated OR was 2.085 (95% CI 2.028 – 2.143). After adding AGE, Education and Health as additional control variables, the adjusted OR for overweight is 6.1% lower at 1.957 (95% CI 1.902 – 2.014).

So, we can conclude that:

(1) BMI is associated with the presence of diabetes, after controlling for other potentially confounding and modifying variables.
(2) The higher the BMI, the greater the likelihood of diabetes, all else constant.

A surprising finding is that the highest adjusted OR is associated with being 55 years old or older…such individuals stand a 740% greater chance of having diabetes, all else constant. Of course, this is all based on this simple analysis. There may be many more control variables we should try.

9) (15 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, sex, age, education, and general health

	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%	1.957	1.902 - 2.014
Obese	88,396	24.5%	28,219	52.6%	4.497	4.376 - 4.622
Sex						
Male	139,820	38.8%	22,610	42.1%	-----	-----
Female	220,278	61.2%	31,040	57.9%	0.847	0.830 - 0.865
Exercise						
No	89,873	25.0%	21,658	40.4%	-----	-----
Yes	270,225	75.0%	31,992	59.6%	0.869	0.850 - 0.887
Age						
18 to 34	43,260	12.0%	899	1.7%	-----	-----
35 to 54	124,039	34.4%	9,704	18.1%	3.080	2.871 - 3.305
55+	192,799	53.5%	43,047	80.2%	8.395	7.844 - 8.986
Education						
Elementary	9,549	2.7%	3,318	6.2%	-----	-----
HS	125,304	34.8%	23,817	44.4%	0.842	0.804 - 0.882
College	225,245	62.6%	26,515	49.4%	0.747	0.714 - 0.783
General Health						
Poor	57,115	15.9%	25,047	46.7%	-----	-----
Good	231,036	64.2%	26,972	50.3%	0.341	0.334 - 0.349
Excellent	71,947	20.0%	1,631	3.0%	0.095	0.090 - 0.100

* 95% confidence intervals are for reported odds ratios.

10) (20 pts) Write the Table 3 results section/interpretation.

Table 3 presents the adjusted odds ratios for the study on the association between BMI and the diagnosis of diabetes in the BRFSS 2010 population. Those who reported exercising had 13% fewer odds of being diagnosed with diabetes than those who reported they did not exercise, after controlling for sex, BMI, age, education, and general health (OR = 0.869; 95% CI = 0.850-0.887). Females had lower odds (15%) relative to males of being diagnosed with diabetes, after controlling for exercise, BMI, age education and general health (OR = 0.847; 95% CI = 0.830-0.865). For the variable of primary interest BMI, those with a BMI rated as "obese" had substantially greater odds (350%) of being diagnosed with diabetes as compared to those with a BMI rates as "normal", after

controlling for sex, exercise, age, education, and general health (OR = 4.497; 95% CI = 4.376-4.622). With respect to additional control variables, being aged 55 years or more is associated with 740% greater odds of having diabetes, the largest OR among all the included variable categories, after controlling for sex, exercise, BMI, education, and general health (OR = 8.395; CI = 7.844-8.986). Having attended some college or better is associated with having 25% fewer odds of having diabetes after controlling for sex, exercise, BMI, age, and general health (OR = 0.747; 95% CI = 0.714-0.783). And having "excellent" general health is associated with having 91% fewer odds of having diabetes after controlling for sex, exercise, BMI, age, and education (OR = 0.095; 95% CI = 0.090-0.100).

Extra Credit (15 pts)

Test for an interaction between SEX and the categorized AGE variable.

1. **What is your conclusion? Does a significant interaction exist?**
2. **Show the MLE table and Type 3 Analysis of Effects ("Joint test") output table to support your conclusion.**
3. **Compute and interpret the OR for your interaction: SEX at AGECAT = 1, AGECAT = 2 and AGECAT = 3.**

We can test for an interaction using the following syntax:

```
proc logistic data=work.tmp descending;
       class agecat (ref='1') / param=ref;
       model diabetes = exercise sex healthcat agecat educat _bmi4cat
       sex*agecat / lackfit;
       oddsratio sex / at(agecat='1' '2' '3');
run;
```

We find that the β on both interactions are statistically significant.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.1286	0.0637	4204.2007	<.0001
exercise		1	-0.1401	0.0108	166.7355	<.0001
sex		1	0.2480	0.0715	12.0252	0.0005
healthcat		1	-1.1149	0.00937	14149.9576	<.0001
agecat	2	1	1.2873	0.0600	460.4739	<.0001
agecat	3	1	2.3968	0.0583	1692.6702	<.0001
educat		1	-0.1314	0.00877	224.6545	<.0001
_BMI4CAT		1	0.7687	0.00679	12805.7603	<.0001
sex*agecat	2	1	-0.2654	0.0749	12.5558	0.0004
sex*agecat	3	1	-0.4568	0.0725	39.7282	<.0001

And that in the aggregate (across all AGE categories) the interaction is statistically significant:

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
exercise	1	166.7355	<.0001
sex	1	12.0252	0.0005
healthcat	1	14149.9576	<.0001
agecat	2	4724.7995	<.0001
educat	1	224.6545	<.0001
_BMI4CAT	1	12805.7603	<.0001
sex*agecat	2	91.2413	<.0001

The ODDSRATIO statement produces the estimated OR for SEX at both AGE categories:

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
sex at agecat=1	1.281	1.114	1.474
sex at agecat=2	0.983	0.941	1.027
sex at agecat=3	0.812	0.793	0.830

The odds of a female having diabetes are 28.1% higher than for a male in the 18-34 age group, 1.7% lower than for a male in the 35-54 age group and are 18.8% lower than for a male in the 55+ age group.

Since the OR varies with AGE, there is a clear interaction between SEX and AGE. AGE is an effect modifier of the association between SEX and DIABETES.

By the way, how would we hand calculate the OR in this case? For simplicity, suppose we estimate using:

```
proc logistic data=work.tmp descending;
       class agecat (ref='1') /
       param=ref;
       model diabetes = agecat sex
       sex*agecat;
run;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.0251	0.0569	4997.9013	<.0001
agecat	2	1	1.5068	0.0592	647.0794	<.0001
agecat	3	1	2.6521	0.0575	2124.8228	<.0001
sex		1	0.2432	0.0706	11.8501	0.0006
sex*agecat	2	1	-0.2937	0.0738	15.8342	<.0001
sex*agecat	3	1	-0.4534	0.0715	40.2437	<.0001

The model we are estimating can be represented as follows since we are using the class agecat (ref='1') / param=ref statement:

$$D = \alpha + \beta_1 AGECAT2 + \beta_2 AGECAT3 + \beta_3 SEX + \beta_4 SEX*AGECAT2 + \beta_5 SEX*AGECAT3$$

Since we have 6 profiles, we can find the odds for each profile as follows:

Profile	AGECAT2	AGECAT3	SEX	SEX*AGECAT2	SEX*AGECAT3
A	1	0	1	1	0
B	1	0	0	0	0
C	0	1	1	0	1
D	0	1	0	0	0
E	0	0	1	0	0
F	0	0	0	0	0

Profile	Log Odds	Odds
A	$\alpha + \beta_1(1) + \beta_2(0) + \beta_3(1) + \beta_4(1) + \beta_5(0)$	$e^{\alpha + \beta1 + \beta3 + \beta4}$
B	$\alpha + \beta_1(1) + \beta_2(0) + \beta_3(0) + \beta_4(0) + \beta_5(0)$	$e^{\alpha + \beta1}$
C	$\alpha + \beta_1(0) + \beta_2(1) + \beta_3(1) + \beta_4(0) + \beta_5(1)$	$e^{\alpha + \beta2 + \beta3 + \beta5}$
D	$\alpha + \beta_1(0) + \beta_2(1) + \beta_3(0) + \beta_4(0) + \beta_5(0)$	$e^{\alpha + \beta2}$
E	$\alpha + \beta_1(0) + \beta_2(0) + \beta_3(1) + \beta_4(0) + \beta_5(0)$	$e^{\alpha + \beta3}$
F	$\alpha + \beta_1(0) + \beta_2(0) + \beta_3(0) + \beta_4(0) + \beta_5(0)$	e^{α}

So, the odd ratio for a female vs a male:

At AGECAT1 $= E/F = e^{\alpha + \beta3} / e^{\alpha} = e^{\beta3}$ $= 1.275$

At AGECAT2 $= A/B = e^{\alpha + \beta1 + \beta3 + \beta4} / e^{\alpha + \beta1} = e^{\beta3 + \beta4}$ $= 0.951$

At AGECAT3 $= C/D = e^{\alpha + \beta2 + \beta3 + \beta5} / e^{\alpha + \beta2} = e^{\beta3 + \beta5}$ $= 0.810$

Which is verified by using the ODDSRATIO statement:

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
sex at agecat=1	1.275	1.110	1.465
sex at agecat=2	0.951	0.912	0.991
sex at agecat=3	0.810	0.793	0.828