Harsha Kankanamge **PROJECT 2**

The data set OBESITY.SAV represents data from a sample of 5857 subjects from the general population in the United States between 2009 and 2010. The goal is to identify risk factors that predict obesity.

Using the variables VIGRECEXR (vigorous recreational activities), HDL (high density lipoprotein), SEDMIN (minutes of sedentary activity per week), GENDER and OBESE (BMI status),

1)Conduct a descriptive analysis of the variables VIGRECEXR, HDL, SEDMIN, GENDER, and OBESE and summarize the results.

Statistics

		vigorous recreational activities	GENDER	bmi more than 35
N	Valid	5857	5857	5822
	Missing	0	0	35

vigorous recreational activities

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	yes	1078	18.4	18.4	18.4
	no	4779	81.6	81.6	100.0
	Total	5857	100.0	100.0	

In the sample 1078 people do “vigorous recreational activities” out of 5857 people. In general, 18.4% of the entire dataset represent that people who do “vigorous recreational activities”. 81.6% of the entire dataset represent that people who do not do “vigorous recreational activities”. As well as there are no missing values in this column.

GENDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	2820	48.1	48.1	48.1
	female	3037	51.9	51.9	100.0
	Total	5857	100.0	100.0	

In the sample 2820 people are males out of 5857 people. In general, 48.1% of the entire dataset represent males. 51.9% represent the females in the dataset. There are not missing values in this column.

bmi more than 35

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	4888	83.5	84.0	84.0
	yes	934	15.9	16.0	100.0
	Total	5822	99.4	100.0	
Missing	System	35	.6		
Total		5857	100.0		

In the sample 4888 people have BMI factor less than or equal 35 out of 5857 people. In general, 83.5% of the entire dataset represent that people who have BMI factor less than or equal 35, as well as 0.6% (35) of data missing in this column. 15.9% of people in the sample have “BMI more than 35”.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HDL cholesterol	5512	15	144	52.63	16.357
minutes of sedentary activity per week	5786	0	840	314.63	185.081
Valid N (listwise)	5445				

Mean and Std. deviation of variable “HDL cholesterol” are respectively 52.63 and 16.357. As well as it has maximum 144 and minimum 15.

Mean and Std. deviation of variable “minutes of sedentary activity per week” are respectively 314 minutes and 185.081. As well as it has maximum 840 minutes and minimum 0 minutes.

2. In CROSSTABS of VIGRECEXR (rows) and OBESE (columns), interpret row percentages

vigorous recreational activities * bmi more than 35 Crosstabulation

			bmi more than 35		
			no	yes	Total
vigorous recreational activities	yes	Count	981	95	1076
		% within vigorous recreational activities	91.2%	8.8%	100.0%
	no	Count	3907	839	4746
		% within vigorous recreational activities	82.3%	17.7%	100.0%
Total		Count	4888	934	5822
		% within vigorous recreational activities	84.0%	16.0%	100.0%

91.2% (981) of people have BMI less than or equal to 35 from people who do ‘vigorous recreational activities’. 82.3% (3907) of people have BMI less than or equal to 35 from people who do not do ‘vigorous recreational activities’. 17.7% (839) of people have BMI more than 35 from people who do not do ‘vigorous recreational activities’.

Chi-Square Test between VIGRECEXR and OBESE

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	50.995[a]	1	.000		
Continuity Correction[b]	50.340	1	.000		
Likelihood Ratio	57.432	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	50.986	1	.000		
N of Valid Cases	5822				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 172.62.

b. Computed only for a 2x2 table

H0: There is no association between the VIGRECEXR and OBESE.
H1: There is an association between the VIGRECEXR and OBESE.

Chi-Square Test statistic is 50.995 and p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between VIGRECEXR and OBESE.

VIGRECEXR should be in the model

3. In the logistic regression of OBESE on VIGRECEXR and SEDMIN, consider VIGRECEXR to be the risk factor and test

a) if SEDMIN (covariate) is a confounder in the association of VIGRECEXR with OBESE.

Chi-Square Test between VIGRECEXR and OBESE

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	50.995[a]	1	.000		
Continuity Correction[b]	50.340	1	.000		
Likelihood Ratio	57.432	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	50.986	1	.000		
N of Valid Cases	5822				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 172.62.

b. Computed only for a 2x2 table

H0: There is no association between VIGRECEXR and OBESE.
H1: There is an association between VIGRECEXR and OBESE.

Chi-Square Test statistic is 50.995 and p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between VIGRECEXR and OBESE.

VIGRECEXR should be in the model

T-test for SEDMIN and OBESE

Group Statistics

	bmi more than 35	N	Mean	Std. Deviation	Std. Error Mean
minutes of sedentary activity per week	no	4838	307.46	184.481	2.652
	yes	914	349.40	184.116	6.090

Average of SEDMIN for people who have obesity is 349.40 and average of SEDMIN for people who do not have obesity is 307.46. So people those who have 'bmi more than 35' have higher (349.40) average value for SEDMIN. Therefore, SEDMIN is suspected confounder

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
minutes of sedentary activity per week	Equal variances assumed	.230	.632	-6.306	5750	.000	-41.943	6.651	-54.983	-28.904
	Equal variances not assumed			-6.314	1283.469	.000	-41.943	6.643	-54.975	-28.912

F statistic = 0.230 and p-value = 0.632 >0.05 Do not reject H0 So we can assume the equal variances.

H0: There is no association between SEDMIN and OBESE ($\mu1=\mu2$).
H1: There is an association between SEDMIN and OBESE ($\mu1\neq\mu2$).

T statistic = -6.306 p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between SEDMIN and OBESE. Therefore, SEDMIN is suspected confounder

T-test for SEDMIN and VIGRECEXR

Group Statistics

	vigorous recreational activities	N	Mean	Std. Deviation	Std. Error Mean
minutes of sedentary activity per week	yes	1069	327.55	195.408	5.977
	no	4717	311.71	182.555	2.658

Average of SEDMIN for people who do the vigorous activities is 327.55 and average of SEDMIN for people who do not do the vigorous activities is 311.71. So people those who do the vigorous activities have higher (327.55) average value for SEDMIN. Therefore, SEDMIN is suspected confounder.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
minutes of sedentary activity per week	Equal variances assumed	10.776	.001	2.528	5784	.012	15.839	6.267	3.555	28.124
	Equal variances not assumed			2.422	1518.813	.016	15.839	6.541	3.009	28.670

F statistic = 10.776 and p-value = 0.001 < 0.05 Reject H0 So we can not assume the equal variances. So we will use un-pooled T test

H0: There is no association between SEDMIN and VIGRECEXR.
H1: There is an association between SEDMIN and VIGRECEXR.

T statistic = 2.422 p-value = 0.016

p-value < 0.05 Reject H0

So there is a significant association between SEDMIN and VIGRECEXR.

Therefore, SEDMIN is suspected confounder in the association of VIGRECEXR with OBESE.

Model 1: with only VIGRECEXR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR	.796	.114	48.811	1	.000	2.218	1.774	2.773
	Constant	-2.335	.107	472.107	1	.000	.097		

a. Variable(s) entered on step 1: VIGRECEXR.

Model 2: with only VIGRECEXR AND SEDMIN

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR	.815	.115	50.163	1	.000	2.260	1.803	2.832
	SEDMIN	.001	.000	43.006	1	.000	1.001	1.001	1.002
	Constant	-2.770	.129	461.748	1	.000	.063		

a. Variable(s) entered on step 1: VIGRECEXR, SEDMIN.

Model	constant	VIGRECEXR	SEDMIN	e^{β}
1	-2.335	0.796		2.218
2	-2.770	0.815	0.001	2.260

Coefficient of 'VIGRECEXR' changed from 0.796 to 0.815 =

(0.796 – 0.815)/0.815 = -0.023312883

Coefficient of 'VIGRECEXR' increase by about 2% when SEDMIN is added to the model.

Both 'VIGRECEXR' and 'SEDMIN' are significantly associated with OBESE. (p-values = 0.000<0.05)

However, 'SEDMIN' is not a major confounder in relationship between 'VIGRECEXR' and 'OBESE'. Since change of Coefficient of 'VIGRECEXR' is less than 10%.

Here, 2.218 is crude OR (Odds ratio) and 2.260 is 'SEDMIN' adjusted OR of 'VIGRECEXR'.

b) if SEDMIN (covariate)is an effect modifier (interaction between VIGRECEXR and SEDMIN) in the association of VIGRECEXR with OBESE. Illustrate graphically.

Categorical Variables Codings

		Frequency	Parameter coding (1)
vigorous recreational activities	yes	1067	.000
	no	4685	1.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	VIGRECEXR(1)	.496	.229	4.692	1	.030	1.642	1.048	2.571
	SEDMIN	.000	.001	.705	1	.401	1.000	.999	1.002
	SEDMIN by VIGRECEXR (1)	.001	.001	2.445	1	.118	1.001	1.000	1.002
	Constant	-2.489	.215	134.512	1	.000	.083		

a. Variable(s) entered on step 1: VIGRECEXR, SEDMIN, SEDMIN * VIGRECEXR .

$\ln[\frac{\hat{p}}{1-\hat{p}}]$= -2.489+0.496*VIGRECEXR(1)+0.000*SEDMIN+0.001* SEDMIN* VIGRECEXR(1)

For testing

H0: β3=0
H1: β3≠0

Wald test statistics = 2.445 p-value = 0.118
p-value 0.118 > 0.05 Do not reject H0
SEDMIN* VIGRECEXR has not significant contribution in the development of OBESE (bmi more than 35). So SEDMIN* VIGRECEXR should not be in the model.

Therefore, model is:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	VIGRECEXR(1)	.815	.115	50.163	1	.000	2.260	1.803	2.832
	SEDMIN	.001	.000	43.006	1	.000	1.001	1.001	1.002
	Constant	-2.770	.129	461.748	1	.000	.063		

a. Variable(s) entered on step 1: VIGRECEXR, SEDMIN.

$\ln[\frac{\hat{p}}{1-\hat{p}}]$= -2.770+0.815*VIGRECEXR+0.001*SEDMIN

Mathematically proving

VIGRECEXR=0 and SEDMIN=100

-2.770+0.815*0+0.001*100 = -2.67

VIGRECEXR=1 and SEDMIN=100

-2.770+0.815*1+0.001*100 = -1.855

Log odds ratio of VIGRECEXR (yes versus no) = -1.855 –(-2.67) = 0.815

VIGRECEXR=0 and SEDMIN=300

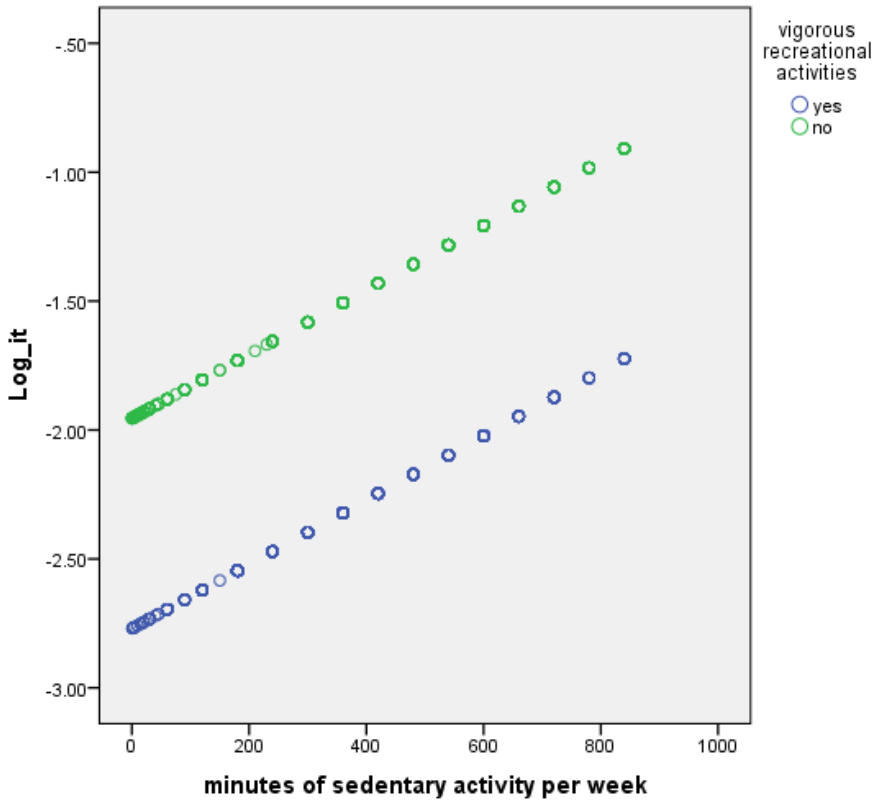
-2.770+0.815*0+0.001*300 = -2.47

VIGRECEXR=1 and SEDMIN=300

-2.770+0.815*1+0.001*300 = -1.655

Log odds ratio of VIGRECEXR (yes versus no) = -1.655 – (-2.47) = 0.815

This differences are same for all SEDMIN values. Which means the both lines are parallel. There is not interaction between VIGRECEXR and SEDMIN. So SEDMIN (covariate)is not an effect modifier in the association of VIGRECEXR with OBESE.



The graph clearly shows that there is not an interaction between VIGRECEXR and SEDMIN. Since both lines are parallel. Logit has minus values for both vigorous recreation activity yes or no for enter range of SEDMIN. But logit have little higher values for people who do not do vigorous recreation activity. So SEDMIN (covariate)is not an effect modifier in the association of VIGRECEXR with OBESE.

4. In the logistic regression of OBESE on VIGRECEXR and HDL, consider VIGRECEXR to be the risk factor and test

a) if HDL (covariate) is a confounder in the association of VIGRECEXR with OBESE.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HDL cholesterol	5512	15	144	52.63	16.357
Valid N (listwise)	5512				

Mean and Std. deviation of variable “HDL cholesterol” are respectively 52.63 and 16.357. As well as it has maximum 144 and minimum 15.

T-test for “HDL cholesterol” and Obese

Group Statistics

	bmi more than 35	N	Mean	Std. Deviation	Std. Error Mean
HDL cholesterol	no	4598	53.82	16.705	.246
	yes	881	46.78	12.965	.437

Average of HDL cholesterol for people who have obesity is 46.78 and average of HDL cholesterol for people who do not have obesity is 53.82. So people those who have obese have less (46.78) average value for HDL cholesterol. Therefore, HDL cholesterol is suspected confounder.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HDL cholesterol	Equal variances assumed	67.046	.000	11.839	5477	.000	7.037	.594	5.872	8.203
	Equal variances not assumed			14.033	1499.845	.000	7.037	.501	6.054	8.021

F statistic = 67.046 and p-value = 0.000 <0.05 Reject H0 So we can not assume the equal variances.

H0: There is no association between HDL and OBESE.
H1: There is an association between HDL and OBESE.

T statistic = 14.033 p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between HDL cholesterol and OBESE. Therefore, HDL cholesterol is suspected confounder

T-test for "HDL cholesterol" and "VIGRECEXR"

Group Statistics

	vigorous recreational activities	N	Mean	Std. Deviation	Std. Error Mean
HDL cholesterol	yes	1019	54.80	16.519	.517
	no	4493	52.14	16.282	.243

Average of HDL for people who do the vigorous activities is 54.80 and average of HDL for people who do not do the vigorous activities is 52.14. For people those who do the vigorous activities have higher (54.80) average value for HDL. Therefore, HDL is suspected confounder.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HDL cholesterol	Equal variances assumed	1.947	.163	4.698	5510	.000	2.661	.566	1.551	3.772
	Equal variances not assumed			4.655	1499.502	.000	2.661	.572	1.540	3.783

F statistic = 1.947 and p-value = 0.163 > 0.05 Do not reject H0 So we can assume the equal variances.

H0: There is no association between HDL and VIGRECEXR.
H1: There is an association between HDL and VIGRECEXR.

T statistic = 4.698 p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between HDL and VIGRECEXR.

Therefore, HDL is suspected confounder in the association of VIGRECEXR with OBESE

Model 1: with only VIGRECEXR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR(1)	.796	.114	48.811	1	.000	2.218	1.774	2.773
	Constant	-2.335	.107	472.107	1	.000	.097		

a. Variable(s) entered on step 1: VIGRECEXR.

Model 2: with HDL and VIGRECEXR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	HDL	-.031	.003	126.028	1	.000	.970	.965	.975
	VIGRECEXR(1)	.725	.118	37.969	1	.000	2.066	1.640	2.602
	Constant	-.740	.173	18.266	1	.000	.477		

a. Variable(s) entered on step 1: HDL, VIGRECEXR.

Coefficient of 'VIGRECEXR' changed from 0.796 to 0.725 =

(0.796 – 0.725)/0.725 = 0.09793

Coefficient of 'VIGRECEXR' decreased by about 10% when HDL is added to the model.

Both 'VIGRECEXR' and 'HDL' are significantly associated with OBESE. (p-values = 0.000<0.05)

However, 'HDL' is boundary line major confounder in relationship between 'VIGRECEXR' and 'OBESE'.

Therefore, conclude that confounding is present.

Here, 2.218 is crude OR (Odds ratio) and 2.066 is 'HDL' adjusted OR of 'VIGRECEXR'.

b) if HDL (covariate)is an effect modifier (interaction between VIGRECEXR and HDL) in the association of VIGRECEXR with OBESE. Illustrate graphically and mathematically.

Categorical Variables Codings

		Frequency	Parameter coding (1)
vigorous recreational activities	yes	1018	.000
	no	4461	1.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR(1)	-.212	.443	.229	1	.632	.809	.340	1.928
	HDL	-.048	.009	30.483	1	.000	.953	.937	.969
	HDL by VIGRECEXR(1)	.020	.009	4.604	1	.032	1.020	1.002	1.038
	Constant	.096	.419	.052	1	.820	1.100		

a. Variable(s) entered on step 1: VIGRECEXR, HDL, HDL * VIGRECEXR .

$\ln[\frac{\hat{p}}{1-\hat{p}}]$= 0.096 - 0.212*VIGRECEXR - 0.048*HDL + 0.020* HDL* VIGRECEXR

For testing

H0: β3=0
H1: β3≠0

Wald test statistics = 4.604 p-value = 0.032
p-value 0.032 < 0.05 Reject H0
HDL* VIGRECEXR has significant contribution in the development of OBESE (bmi more than 35).
So HDL * VIGRECEXR should be in the model.

So final model is:

$\ln[\frac{\hat{p}}{1-\hat{p}}]$= 0.096 - 0.212*VIGRECEXR - 0.048*HDL + 0.020* HDL* VIGRECEXR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR(1)	-.212	.443	.229	1	.632	.809	.340	1.928
	HDL	-.048	.009	30.483	1	.000	.953	.937	.969
	HDL by VIGRECEXR(1)	.020	.009	4.604	1	.032	1.020	1.002	1.038
	Constant	.096	.419	.052	1	.820	1.100		

a. Variable(s) entered on step 1: VIGRECEXR, HDL, HDL * VIGRECEXR .

Mathematically proving

VIGRECEXR=0 and HDL =40

0.096-0.048*40 = -1.824

VIGRECEXR=1 and HDL =40

0.096-0.212*1-0.048*40+0.020*1*40 = -1.236

Log odds ratio of VIGRECEXR (yes versus no) = -1.236-(-1.824) = 0.588

Ln[OR^] = β1^+β3^X = -0.212+0.02*40=0.588

VIGRECEXR=0 and HDL =70

0.096-0.048*70 = -3.264

VIGRECEXR=1 and HDL =70

0.096-0.212*1-0.048*70+0.020*1*70 = -2.076

Log odds ratio of VIGRECEXR (yes versus no) = -2.076-(-3.264) = 1.188

Ln[OR^] = β1^+β3^X = -0.212+0.02*70=1.188

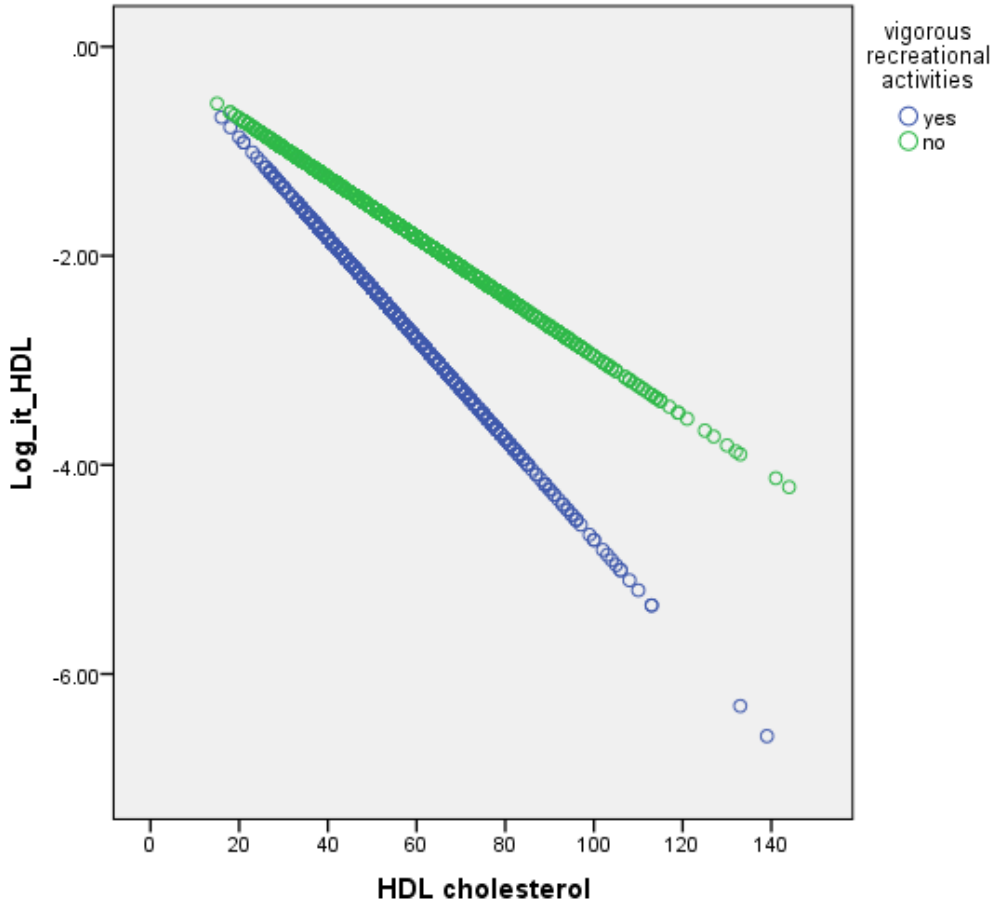
HDL =80

Ln[OR^] = β1^+β3^X = -0.212+0.02*80=1.388

HDL =90

Ln[OR^] = β1^+β3^X = -0.212+0.02*90=1.588

All Ln[OR^] are not same for different values of HDL. Which means the both lines are not parallel. There is an interaction between VIGRECEXR and HDL. So HDL (covariate)is an effect modifier in the association of VIGRECEXR with OBESE.



The graph clearly shows that there is an interaction between HDL and VIGRECEXR. Interaction is happening for a value of HDL closer to 20. For all other HDL values, Logit has higher values when 'VIGRECEXR = no'. When HDL values getting larger (Logit difference) Log odds ratio also getting larger. This graph clearly shows that when HDL values getting increase Logit values getting decrease for all the people those who do vigorous recreational activities and do not. But the slope is higher for the people those who do vigorous recreational activities.
Therefore, HDL (covariate)is an effect modifier (interaction between VIGRECEXR and HDL) in the association of VIGRECEXR with OBESE.

c) Find the probability of obesity for two subjects with and without vigorous recreational activity each having HDL = 55 and interpret the probabilities.

$\ln[\frac{p^\wedge}{1-p^\wedge}]$= 0.096 - 0.212*VIGRECEXR - 0.048*HDL + 0.020* HDL* VIGRECEXR

People those who do vigorous recreational activities. (VIGRECEXR=0)

$\ln[\frac{p^\wedge}{1-p^\wedge}]$= 0.096 - 0.212*0 - 0.048*55 + 0.020* 55* 0 =-2.544

p^ = 0.0728 is the probability of being obese for people those who do vigorous recreational activities when HDL = 55.

People those who do not do vigorous recreational activities. (VIGRECEXR=1)

$\ln[\frac{p^\wedge}{1-p^\wedge}]$= 0.096 - 0.212*1 - 0.048*55 + 0.020* 55* 1= -1.656

p^ = 0.1603 is the probability of being obese for people those who do not do vigorous recreational activities When HDL = 55.

d) Compare the log-odds of engaging and not engaging in vigorous recreational activity (i.e. logit difference) with HDL = 55, and the odds of the same.

log-odds of engaging and not engaging in vigorous recreational activity

Ln[OR^] = β1^+β3^X = - 0.212+0.02*55=0.888

VIGRECEXR=0 and HDL =55

0.096-0.048*55 = -2.544

VIGRECEXR=1 and HDL =55

0.096-0.212*1-0.048*55+0.020*1*55 = -1.656

Log odds ratio = -1.656-(-2.544) = 0.888

Person who does not do the vigorous recreational activity and HDL = 55 have Log odds of having obese 0.888 higher than when compare with person who does vigorous recreational activity.

Odds = $e^{0.88}$ = 2.43

The effect of VIGRECEXR on the odds of having obese increase with HDL.

In particular, person who does not do the vigorous recreational activity and HDL = 55 is estimated to have a risk being obese of 2.43 times that of person who have the same HDL and does the vigorous recreational activities.

5. In the logistic regression of OBESE on VIGRECEXR and GENDER, consider VIGRECEXR to be the risk factor and test
a) if GENDER (covariate) is a confounder in the association of VIGRECEXR with OBESE.

GENDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	2820	48.1	48.1	48.1
	female	3037	51.9	51.9	100.0
	Total	5857	100.0	100.0	

The sample represent 48.1% males which is 2820 out of 5857. As well as 51.9% females which is 3037 out of 5857.

Chi-Square Test between Gender and OBESE

GENDER * bmi more than 35 Crosstabulation

			bmi more than 35		
			no	yes	Total
GENDER	male	Count	2461	340	2801
		% within GENDER	87.9%	12.1%	100.0%
	female	Count	2427	594	3021
		% within GENDER	80.3%	19.7%	100.0%
Total		Count	4888	934	5822
		% within GENDER	84.0%	16.0%	100.0%

In the male group 87.9% (2461) do not have the “bmi more than 35” and 12.1% have the “bmi more than 35”. In the female group 80.3% (2427) do not have the “bmi more than 35” and 19.7% have the “bmi more than 35”.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	61.085[a]	1	.000		
Continuity Correction[b]	60.528	1	.000		
Likelihood Ratio	61.874	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	61.075	1	.000		
N of Valid Cases	5822				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 449.35.

b. Computed only for a 2x2 table

H0: There is no association between Gender and Obese.
H1: There is an association between Gender and Obese.

Chi-square statistic = 61.085 p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between Gender and Obese.

Therefore, gender is a suspected confounder.

Chi-Square Test between Gender and VIGRECEXR

GENDER * vigorous recreational activities Crosstabulation

			vigorous recreational activities		
			yes	no	Total
GENDER	male	Count	668	2152	2820
		% within GENDER	23.7%	76.3%	100.0%
	female	Count	410	2627	3037
		% within GENDER	13.5%	86.5%	100.0%
Total		Count	1078	4779	5857
		% within GENDER	18.4%	81.6%	100.0%

In the male group 23.7% (668) do vigorous recreational activities and 76.3% do not do vigorous recreational activities. In the female group 13.5% (410) do vigorous recreational activities and 86.5% do not do vigorous recreational activities.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	101.058[a]	1	.000		
Continuity Correction[b]	100.381	1	.000		
Likelihood Ratio	101.599	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	101.041	1	.000		
N of Valid Cases	5857				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 519.03.

b. Computed only for a 2x2 table

H0: There is no association between Gender and VIGRECEXR.
H1: There is an association between Gender and VIGRECEXR.

Chi-square statistic = 101.058 p-value = 0.000

p-value < 0.05 Reject H0

So there is a significant association between Gender and VIGRECEXR.

Therefore, gender is a suspected confounder.

Model 1: with only VIGRECEXR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR(1)	.796	.114	48.811	1	.000	2.218	1.774	2.773
	Constant	-2.335	.107	472.107	1	.000	.097		

a. Variable(s) entered on step 1: VIGRECEXR.

Model 2: with Gender and VIGRECEXR

Categorical Variables Codings

		Frequency	Parameter coding (1)
GENDER	male	2801	.000
	female	3021	1.000
vigorous recreational activities	yes	1076	.000
	no	4746	1.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VIGRECEXR(1)	.715	.115	38.714	1	.000	2.043	1.632	2.559
	GENDER(1)	.514	.074	47.765	1	.000	1.672	1.445	1.934
	Constant	-2.557	.114	506.664	1	.000	.078		

a. Variable(s) entered on step 1: VIGRECEXR, GENDER.

Coefficient of 'VIGRECEXR' changed from 0.796 to 0.715 =

(0.796 – 0.715)/0.715 = 0.1132

Coefficient of 'VIGRECEXR' decreased by about 11% when Gender is added to the model.

Both 'VIGRECEXR' and 'Gender' are significantly associated with OBESE. (p-values = 0.000<0.05)

Gender is a major confounder in relationship between 'VIGRECEXR' and 'OBESE'.

Therefore, conclude that confounding is present.

Here, 2.218 is crude OR (Odds ratio) and 2.043 is 'Gender' adjusted OR of 'VIGRECEXR'.

Categorical Variables Codings

GENDER			Frequency	Parameter coding (1)
male	vigorous recreational activities	yes	666	.000
		no	2135	1.000
female	vigorous recreational activities	yes	410	.000
		no	2611	1.000

Variables in the Equation

GENDER			B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
									Lower	Upper
male	Step 1[a]	VIGRECEXR(1)	.490	.153	10.331	1	.001	1.633	1.211	2.201
		Constant	-2.369	.139	292.457	1	.000	.094		
female	Step 1[a]	VIGRECEXR(1)	.974	.177	30.322	1	.000	2.649	1.873	3.746
		Constant	-2.281	.170	179.436	1	.000	.102		

a. Variable(s) entered on step 1: VIGRECEXR.

If confounder is a categorical variable, estimate odds ratios of outcome on risk factor by different categories of confounder.

Odds ratio = $e^{\beta 0 + \beta 1}/e^{\beta 0} = e^{\beta 1}$

For men

Odds ratio = $e^{0.49}$ = 1.632

In the male group person who does not do vigorous recreational activities has 1.632 times as likely to being obese compare with the person who does the vigorous recreational activities.

For Female

Odds ratio = $e^{0.974}$ = 2.648

In the female group person who does not do vigorous recreational activities has 2.648 times as likely to being obese compare with the person who does the vigorous recreational activities.

b) if GENDER (covariate)is an effect modifier (interaction between VIGRECEXR and GENDER) in the association of VIGRECEXR with OBESE. Illustrate graphically and mathematically.

Variables in the Equation

GENDER			B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
									Lower	Upper
male	Step 1[a]	VIGRECEXR(1)	.490	.153	10.331	1	.001	1.633	1.211	2.201
		Constant	-2.369	.139	292.457	1	.000	.094		
female	Step 1[a]	VIGRECEXR(1)	.974	.177	30.322	1	.000	2.649	1.873	3.746
		Constant	-2.281	.170	179.436	1	.000	.102		

a. Variable(s) entered on step 1: VIGRECEXR.

Mathematically proving

VIGRECEXR=0 and Gender = male

-2.369

VIGRECEXR=1 and Gender = male

-2.369 + 0.490 = -1.879

Log odds ratio = -1.879-(-2.369) = 0.49

VIGRECEXR=0 and Gender = female

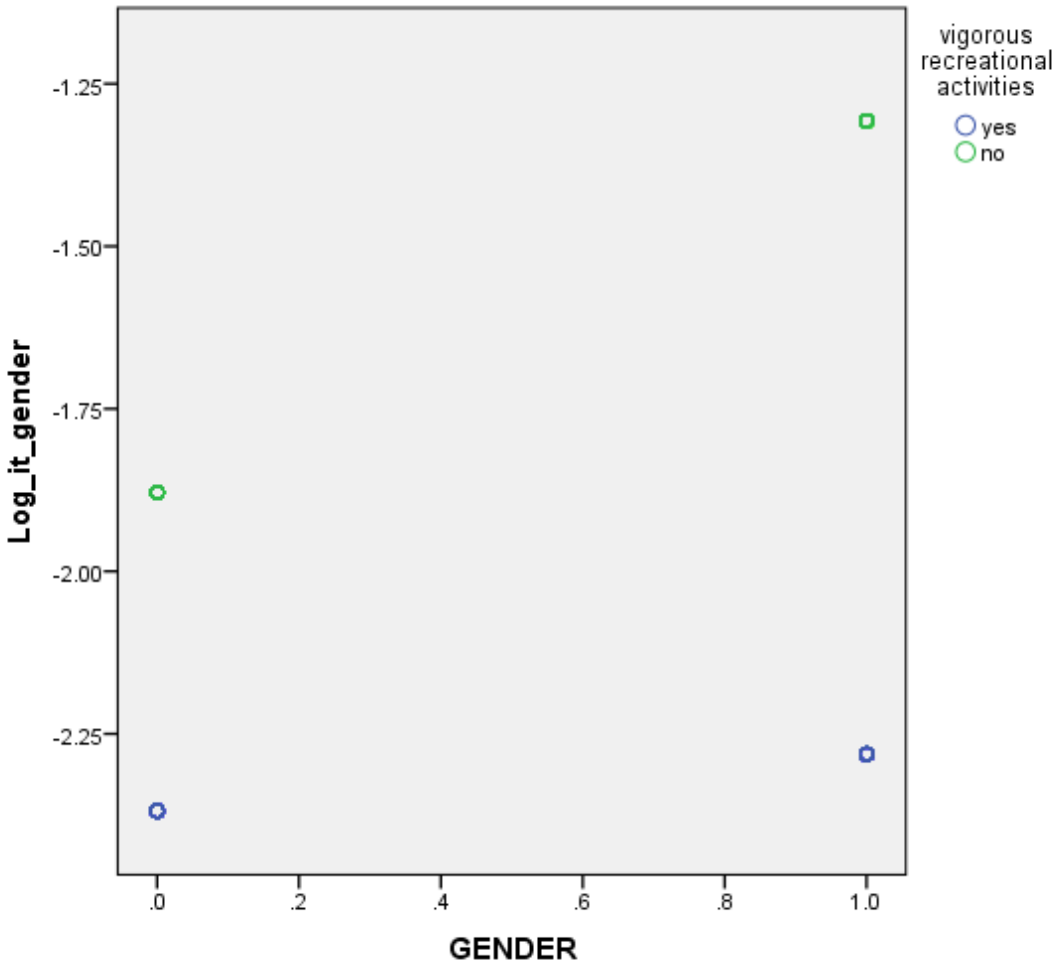
-2.281

VIGRECEXR=1 and Gender = female

-2.281 + 0.974 = -1.307

Log odds ratio = -1.307-(-2.281) = 0.974

Both Ln[OR^] are not same for male and female. Which means the both lines are not parallel. There is an interaction between VIGRECEXR and gender. So gender (covariate)is an effect modifier in the association of VIGRECEXR with OBESE.



The graph clearly shows that the two lines can't be parallel. Even though we can't see the interaction point on the graph, there should be an interaction somewhere. This graph clearly shows that females those who do not do vigorous recreational activities have considerably higher Logit value compare with male those who do not do the vigorous recreational activities. As well as the differences of Logit values for male and female of vigorous recreational activities do or not are clearly different. The slope is higher for the people those who do not do vigorous recreational activities.

Therefore, gender (covariate)is an effect modifier (interaction between VIGRECEXR and gender) in the association of VIGRECEXR with OBESE.