Harsha Kankanamge **ACM 660** **PROJECT 3**

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.

Consider the variable OBESE as the outcome variable and the variables AGE, GENDER, TOTCHOL, HDL, SYSBP, DIASBP, SEDMIN, WLKBIK, and RECACT in the data set as potential covariates.

1)Conduct univariate analysis of all the selected variables.

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. In the data set majority represent females.

recact

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	inactive	3237	55.3	55.3	55.3
	low	1542	26.3	26.3	81.6
	moderate	379	6.5	6.5	88.1
	high	698	11.9	11.9	100.0
	Total	5856	100.0	100.0	
Missing	System	1	.0		
Total		5857	100.0		

In the sample 55.3% represent inactive group, 26.3% represent low group, 6.5% represent moderate group and 11.9% represent high group. There is a one missing value. In the data set majority of people in the inactive group.

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”. In the dataset majority of people do not have obesity.

walk or bicycle

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	yes	1524	26.0	26.0	26.0
	no	4333	74.0	74.0	100.0
	Total	5857	100.0	100.0	

In the sample 4333 people out of 5857 people do not “walk or bicycle”. It is 74% percent of the entire valid data. 1524 people do the “walk or bicycle” out of 5857. It is 26% of entire valid data. In the dataset majority of people do not do ‘walk or bicycle’.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	5857	20	80	49.42	17.791
HDL cholesterol	5512	15	144	52.63	16.357
minutes of sedentary activity per week	5786	0	840	314.63	185.081
systolic blood pressure	5361	90	220	124.15	18.560
diastolic blood pressure	5330	40	134	70.40	11.815
total cholesterol	5512	90	383	195.50	41.131
Valid N (listwise)	4943				

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.

Mean and Std. deviation of variable “Age” are respectively 49.42 and 17.791. As well as it has maximum 80 and minimum 20.

Mean and Std. deviation of variable “systolic blood pressure” are respectively 124.15 and 18.56. As well as it has maximum 220 and minimum 90.

Mean and Std. deviation of variable “diastolic blood pressure” are respectively 70.40 and 11.815. As well as it has maximum 134 and minimum 40.

Mean and Std. deviation of variable “Total cholesterol” are respectively 195.50 and 41.131As well as it has maximum 383 and minimum 90.

2.)Conduct bivariate analysis of each covariate with the outcome.

Group Statistics

	bmi more than 35	N	Mean	Std. Deviation	Std. Error Mean
AGE	no	4888	49.47	18.135	.259
	yes	934	49.39	16.037	.525
total cholesterol	no	4598	195.94	41.755	.616
	yes	881	193.79	37.681	1.269
HDL cholesterol	no	4598	53.82	16.705	.246
	yes	881	46.78	12.965	.437
systolic blood pressure	no	4492	123.84	18.751	.280
	yes	843	125.75	17.580	.606
diastolic blood pressure	no	4471	70.10	11.716	.175
	yes	833	71.89	12.235	.424
minutes of sedentary activity per week	no	4838	307.46	184.481	2.652
	yes	914	349.40	184.116	6.090

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									99% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AGE	Equal variances assumed	31.997	.000	.125	5820	.901	.079	.636	-1.560	1.719
	Equal variances not assumed			.136	1428.404	.892	.079	.585	-1.430	1.589
total cholesterol	Equal variances assumed	11.489	.001	1.422	5477	.155	2.151	1.513	-1.747	6.048
	Equal variances not assumed			1.524	1328.728	.128	2.151	1.411	-1.489	5.790
HDL cholesterol	Equal variances assumed	67.046	.000	11.839	5477	.000	7.037	.594	5.506	8.569
	Equal variances not assumed			14.033	1499.845	.000	7.037	.501	5.744	8.331
systolic blood pressure	Equal variances assumed	2.980	.084	-2.741	5333	.006	-1.910	.697	-3.707	-.114
	Equal variances not assumed			-2.864	1229.402	.004	-1.910	.667	-3.631	-.190
diastolic blood pressure	Equal variances assumed	3.726	.054	-4.035	5302	.000	-1.797	.445	-2.944	-.649
	Equal variances not assumed			-3.917	1134.407	.000	-1.797	.459	-2.980	-.613
minutes of sedentary activity per week	Equal variances assumed	.230	.632	-6.306	5750	.000	-41.943	6.651	-59.082	-24.805
	Equal variances not assumed			-6.314	1283.469	.000	-41.943	6.643	-59.079	-24.808

Mean age for people who have obese 49.39 and mean age for people who do not have obese 49.47. Difference is very small. Not significant. Which is confirm by independent sample t test. T-statistic 0.136 and p-value = 0.892>0.01. which means mean age is not significantly different in the two groups (obese group and not obese group). Age should not be in the model.

Mean 'Total cholesterol' for people who have obese 193.79 and mean 'Total cholesterol' for people who do not have obese 195.94. Differences are very small (2.15). Not significant. Which is confirm by independent sample t test. T-statistic 1.524 and p-value = 0.128 >0.01. which means mean 'Total cholesterol' is not significantly different in the two groups (obese group and not obese group). Variable 'Total cholesterol' should not be in the model.

Mean 'HDL cholesterol' for people who have obese 46.78 and mean 'HDL cholesterol' for people who do not have obese 53.82. Differences are not very small (7.04). significant. Which is confirm by independent sample t test. T-statistic 14.033 and p-value = 0.000 <0.01. which means mean 'HDL cholesterol' is significantly different in the two groups (obese group and not obese group). Should be in the model.

Mean 'systolic blood pressure' for people who have obese 125.75 and mean 'systolic blood pressure' for people who do not have obese 123.84. Differences are not very small (-1.9). significant. Which is confirm by independent sample t test. T-statistic -2.741 and p-value = 0.006<0.01. which means mean 'systolic blood pressure' is significantly different in the two groups (obese group and not obese group).

Mean 'diastolic blood pressure' for people who have obese 71.89 and mean 'diastolic blood pressure' for people who do not have obese 70.10. Differences are not very small (-1.8). significant. Which is confirm by independent sample t test. T-statistic -4.035 and p-value = 0.000<0.01. which means mean 'diastolic blood pressure' is significantly different in the two groups (obese group and not obese group).

Mean 'minutes of sedentary activity per week' for people who have obese 349.40 and mean 'minutes of sedentary activity per week' for people who do not have obese 307.46. Differences are not very small (-41.9). significant. Which is confirm by independent sample t test. T-statistic -6.306 and p-value = 0.000<0.01. which means mean 'minutes of sedentary activity per week' is significantly different in the two groups (obese group and not obese group).

Crosstab

			bmi more than 35		
			no	yes	Total
recact	1	Count	2590	620	3210
		% within recact	80.7%	19.3%	100.0%
	2	Count	1317	219	1536
		% within recact	85.7%	14.3%	100.0%
	3	Count	342	37	379
		% within recact	90.2%	9.8%	100.0%
	4	Count	638	58	696
		% within recact	91.7%	8.3%	100.0%
Total		Count	4887	934	5821
		% within recact	84.0%	16.0%	100.0%

People who are in the inactive group 19.3% have “bmi more than 35”.
People who are in the low group 14.3% have “bmi more than 35”.
People who are in the moderate group 9.8% have “bmi more than 35”.
People who are in the high group 8.3% have “bmi more than 35”. Looking at this we can say that people who in the high group have less risk to being obese than people in other groups.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	70.948[a]	3	.000
Likelihood Ratio	76.697	3	.000
Linear-by-Linear Association	68.762	1	.000
N of Valid Cases	5821		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.81.

Variable Recact has significant association with obese.

Crosstab

			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%

People who are in the male group 12.1% have “bmi more than 35”.
People who are in the female group 19.7% have “bmi more than 35”. Females have higher risk of being obese compare with males.

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

Variable gender has significant association with obese

Crosstab

			bmi more than 35		
			no	yes	Total
walk or bicycle	yes	Count	1335	184	1519
		% within walk or bicycle	87.9%	12.1%	100.0%
	no	Count	3553	750	4303
		% within walk or bicycle	82.6%	17.4%	100.0%
Total		Count	4888	934	5822
		% within walk or bicycle	84.0%	16.0%	100.0%

In the group people who are doing ‘work or bicycle’ have 12.1% obese people. In the group people who are not doing ‘work or bicycle’ have 17.4% obese people. So people who are not doing ‘walk or bicycle’ have higher risk to be obese than people who do the walk or bicycle.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	23.560[a]	1	.000		
Continuity Correction[b]	23.167	1	.000		
Likelihood Ratio	24.745	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	23.556	1	.000		
N of Valid Cases	5822				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 243.69.

b. Computed only for a 2x2 table

Variable ‘work or bicycle’ has significant association with obese

3. Check crosstabs of all categorical variables (all cells should have expected frequencies greater than one and no more than 20% of cells should have frequencies less than five).

Gender and walk or bicycle

Crosstab

			walk or bicycle		
			yes	no	Total
GENDER	male	Count	793	2027	2820
		% within GENDER	28.1%	71.9%	100.0%
	female	Count	731	2306	3037
		% within GENDER	24.1%	75.9%	100.0%
Total		Count	1524	4333	5857
		% within GENDER	26.0%	74.0%	100.0%

71.9% males do not do 'walk bicycle' as well as 75.9% female do not do 'walk or bicycle'. Higher proportion of male do the 'walk or bicycle' than females. "Walk or bicycle" proportion significantly different in male and female group. Which is confirm by Chi-square test.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	12.464[a]	1	.000		
Continuity Correction[b]	12.255	1	.000		
Likelihood Ratio	12.459	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	12.462	1	.000		
N of Valid Cases	5857				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 733.77.

b. Computed only for a 2x2 table

Recact and Walk or bicycle

Crosstab

			walk or bicycle		
			yes	no	Total
recact	1	Count	766	2471	3237
		% within recact	23.7%	76.3%	100.0%
	2	Count	390	1152	1542
		% within recact	25.3%	74.7%	100.0%
	3	Count	135	244	379
		% within recact	35.6%	64.4%	100.0%
	4	Count	233	465	698
		% within recact	33.4%	66.6%	100.0%
Total		Count	1524	4332	5856
		% within recact	26.0%	74.0%	100.0%

23.7% of people who are in the inactive group, do the walk or bicycle.
25.3% of people who are in the low group do the walk or bicycle.
35.6% of people who are in the moderate group do the walk or bicycle.
33.4% of people who are in the high group do the walk or bicycle. Looking at this we can say that people who in the high group and moderate group are doing walk or bicycle more than other two groups. So differences of proportion of walk bicycle or not are significantly different in all four groups. Which is showing in the Chi-square test.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	47.548[a]	3	.000
Likelihood Ratio	45.541	3	.000
Linear-by-Linear Association	39.904	1	.000
N of Valid Cases	5856		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 98.63.

Gender and Recact

recact * GENDER Crosstabulation

			GENDER		Total
			male	female	
recact	1	Count	1441	1796	3237
		% within recact	44.5%	55.5%	100.0%
	2	Count	711	831	1542
		% within recact	46.1%	53.9%	100.0%
	3	Count	260	119	379
		% within recact	68.6%	31.4%	100.0%
	4	Count	407	291	698
		% within recact	58.3%	41.7%	100.0%
Total		Count	2819	3037	5856
		% within recact	48.1%	51.9%	100.0%

55.5% of people who are in the inactive group, are females.
53.9% of people who are in the low group are females.
68.6% of people who are in the moderate group are males.
58.3% of people who are in the high group are males. Looking at this we can say that most of the people who are in the high group and moderate group are males and most of the people in other two group are females. So proportion of males and females are significantly different in all four groups. Which is showing in the Chi-square test.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	112.045[a]	3	.000
Likelihood Ratio	113.348	3	.000
Linear-by-Linear Association	75.170	1	.000
N of Valid Cases	5856		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 182.45.

4. Find a best model using FORWARD LR selection. Write the equation of the estimated logistic probabilities and also the estimated logit transformation of the final model. Interpret odds ratios of all coefficients in the final model.

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
recact	inactive	2667	.000	.000	.000
	low	1325	1.000	.000	.000
	moderate	332	.000	1.000	.000
	high	594	.000	.000	1.000
GENDER	male	2428	.000		
	female	2490	1.000		
walk or bicycle	yes	1280	.000		
	no	3638	1.000		

Variables in the Equation

								99% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	HDL	-.030	.003	109.466	1	.000	.970	.963	.977
	Constant	-.171	.144	1.415	1	.234	.843		
Step 2[b]	GENDER(1)	.959	.087	121.364	1	.000	2.609	2.085	3.265
	HDL	-.041	.003	166.675	1	.000	.960	.952	.968
	Constant	-.173	.149	1.348	1	.246	.841		
Step 3[c]	GENDER(1)	.983	.088	125.801	1	.000	2.672	2.132	3.348
	HDL	-.041	.003	166.960	1	.000	.960	.952	.968
	SEDMIN	.001	.000	33.100	1	.000	1.001	1.001	1.002
	Constant	-.583	.166	12.361	1	.000	.558		
Step 4[d]	GENDER(1)	.913	.089	105.890	1	.000	2.492	1.983	3.132
	HDL	-.039	.003	148.667	1	.000	.962	.954	.970
	SEDMIN	.001	.000	36.556	1	.000	1.001	1.001	1.002
	recact			40.341	3	.000			
	recact(1)	-.395	.097	16.417	1	.000	.674	.524	.866
	recact(2)	-.708	.208	11.627	1	.001	.493	.289	.841
	recact(3)	-.758	.160	22.487	1	.000	.469	.311	.707
	Constant	-.466	.167	7.771	1	.005	.627		
Step 5[e]	GENDER(1)	.976	.090	117.324	1	.000	2.655	2.105	3.349
	HDL	-.038	.003	146.164	1	.000	.962	.954	.970
	DBP	.017	.003	24.788	1	.000	1.018	1.008	1.027
	SEDMIN	.001	.000	36.022	1	.000	1.001	1.001	1.002
	recact			40.990	3	.000			
	recact(1)	-.407	.098	17.341	1	.000	.665	.517	.856
	recact(2)	-.711	.208	11.709	1	.001	.491	.287	.839
	recact(3)	-.757	.160	22.369	1	.000	.469	.311	.709
	Constant	-1.742	.308	32.062	1	.000	.175		
Step 6[f]	GENDER(1)	.971	.090	115.728	1	.000	2.640	2.093	3.331
	HDL	-.038	.003	144.507	1	.000	.963	.955	.970
	DBP	.018	.003	25.579	1	.000	1.018	1.009	1.027
	SEDMIN	.001	.000	32.560	1	.000	1.001	1.001	1.002
	recact			38.833	3	.000			
	recact(1)	-.406	.098	17.215	1	.000	.666	.518	.857
	recact(2)	-.681	.208	10.704	1	.001	.506	.296	.865
	recact(3)	-.732	.160	20.860	1	.000	.481	.318	.727
	WLKBIK	.300	.101	8.846	1	.003	1.350	1.041	1.751
	Constant	-1.986	.319	38.719	1	.000	.137		

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

b. Variable(s) entered on step 2: GENDER.

c. Variable(s) entered on step 3: SEDMIN.

d. Variable(s) entered on step 4: recact.

e. Variable(s) entered on step 5: DBP.

f. Variable(s) entered on step 6: WLKBIK.

In the final model we have six variables. As well as all six variables in final model are significant. P-value <0.01

Estimated logit transformation

$\ln[\frac{p^\wedge}{1-p^\wedge}]$= -1.986 +0.971*Gender - 0.038*HDL + 0.001*SEDMIN+0.018*DBP+0.3*WLKBIK – 0.406*recact(low) – 0.681*recact(moderate) -0.732*recact(high)

logistic probabilities

$$P^\wedge = \frac{1}{1+e^{-(-1.986 +0.971*\text{Gender} - 0.038*\text{HDL} + 0.001*\text{SEDMIN}+0.018*\text{DBP}+0.3*\text{WLKBIK} - 0.406*\text{recact(low)} - 0.681*\text{recact(moderate)} -0.732*\text{recact(high)})}}$$

Coefficient of Gender = 0.971

OR^=exp(0.971) = 2.64 odds ratio adjusted for gender. The odds of developing obese are 2.64 times higher among females as compared to males while others variable remain constant.

Coefficient of HDL = -0.038

OR^=exp(-0.038) = 0.963 odds ratio adjusted for HDL. The odds of developing obese are 0.963 times lower for each one unite increment of HDL while others variable remain constant.

Coefficient of SEDMIN = 0.001

OR^=exp(0.001) = 1.001 odds ratio adjusted for SEDMIN. The odds of developing obese are 1.001 times higher for each one unite increment of SEDMIN while others variable remain constant.

Coefficient of DBP = 0.018

OR^=exp(0.018) = 1.018 odds ratio adjusted for DBP. The odds of developing obese are 1.018 times higher for each one unite increment of DBP while others variable remain constant.

Coefficient of WLKBIK = 0.3

OR^=exp(0.3) = 1.349 odds ratio adjusted for WLKBIK. The odds of developing obese are 1.349 times higher among people who do not do walk or bicycle as compare with those who do walk or bicycle while others variable remain constant.

Recact

OR^(low)=0.666 The odds of Obese in the low group are 33.3% less likely than in the inactive group while others variable remains constant.

OR^(moderate)=0.506 The odds of Obese in the moderate group are 49.3% less likely than in the inactive group while others variable remains constant.
OR^(high)=0.48 The odds of Obese in the high group are 52% less likely than in the inactive group while others variable remains constant.

5. Check for any confounding among selected covariates from the final FORWARD LR model.

variable	β crude	β Adjust
Gender	**0.572**	**0.971**
HDL	**-0.031**	**-0.038**
SEDMIN	**0.001**	**0.001**
DBP	**0.013**	**0.018**
WLKBIK	**0.426**	**0.3**
Recact(1,2,3)	**-0.364/-0.794/-0.968**	**-0.406/-0.681/-0.732**

Gender = 0.572-0.971/0.971 = -0.4109 = 41%

WLKBIK= 0.426-0.3/0.3 = 0.42 = 42%

Coefficient of 'gender and walk or bicycle' increase by about 41% when other variables are added to the model. So other variables are confounding Gender and WLKBIK. Coefficient of other variables have not been changed in larger scale. Confounding is present for gender and wlkbik.

6. Check interaction of each pair of the selected covariates from the final FORWARD LR model. No graphs necessary.

Gender and HDL

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	HDL	-.046	.005	72.532	1	.000	.955	.942	.969
	GENDER(1)	.710	.301	5.567	1	.018	2.034	.937	4.415
	GENDER(1) by HDL	.005	.006	.712	1	.399	1.005	.989	1.022
	Constant	.061	.233	.068	1	.795	1.063		

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

H0: $\beta 3=0$
H1: $\beta 3\neq 0$
Wald test statistics = 0.712 p-value = 0.399
p-value 0.399 > 0.01 Do not reject H0 Not significant

Gender*HDL should not be in the model

Gender and SEDMIN

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	GENDER(1)	.403	.152	7.017	1	.008	1.497	1.011	2.216
	GENDER(1) by SEDMIN	.001	.000	1.899	1	.168	1.001	1.000	1.002
	SEDMIN	.001	.000	8.888	1	.003	1.001	1.000	1.002
	Constant	-2.289	.119	371.297	1	.000	.101		

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

H0: β3=0
H1: β3≠0
Wald test statistics = 1.899 p-value = 0.168
p-value 0.168 > 0.01 Do not reject H0 not significant

Gender*SEDMIN should not be in the model

Gender and DBP

H0: β3=0
H1: β3≠0
Wald test statistics = 0.24 p-value = 0.877>0.01 Reject H0 Gender*DBP should not be in the model.

Gender and WLKBIK

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	GENDER(1)	.837	.165	25.822	1	.000	2.310	1.511	3.532
	GENDER(1) by WLKBIK	-.352	.184	3.653	1	.056	.703	.437	1.130
	WLKBIK	.616	.146	17.678	1	.000	1.851	1.269	2.698
	Constant	-2.449	.131	347.662	1	.000	.086		

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

H0: β3=0
H1: β3≠0
Wald test statistics = 3.653 p-value = 0.056>0.01 But borderline significant Gender*WLKBIK Could be in the model.

Gender and Recact

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	GENDER(1)	.685	.152	20.355	1	.000	1.984	1.342	2.935
	GENDER(1) by recact	-.110	.085	1.677	1	.195	.896	.720	1.115
	recact	-.259	.061	18.202	1	.000	.772	.660	.903
	Constant	-1.523	.116	171.504	1	.000	.218		

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

H0: β3=0
H1: β3≠0

Wald test statistics = 1.677 p-value = 0.195>0.01 Reject H0 Gender*Recact should not be in the model. Not significant.

HDL and SEDMIN

HDL*SEDMIN is not significant. Should not be in the model. P-value 0.925>0.01

HDL and DBP

HDL*DBP is not significant. Should not be in the model. P-value 0.333>0.01

HDL and WLKBIK

HDL*WLKBIK is not significant. Should not be in the model. P-value 0.621>0.01

HDL and Recact

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	HDL	-.013	.006	5.183	1	.023	.987	.973	1.002
	HDL by recact	-.011	.003	11.278	1	.001	.989	.980	.997
	recact	.216	.163	1.748	1	.186	1.241	.815	1.889
	Constant	-.429	.277	2.398	1	.121	.651		

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

H0: $\beta 3=0$
H1: $\beta 3\neq 0$
Wald test statistics = 11.278 p-value = 0.001<0.01 Do not reject H0 HDL*Recact should be in the model. HDL*Recact is significant.

SEDMIN and DBP

SEDMIN*DBP is not significant. Should not be in the model. P-value 0.696>0.01

SEDMIN and WLKBIK

SEDMIN*WLKBIK is not significant. Should not be in the model. P-value 0.847>0.01

SEDMIN and Recact

SEDMIN*Recact is not significant. Should not be in the model. P-value 0.194>0.01

DBP and WLKBIK

DBP*WLKBIK is not significant. Should not be in the model. P-value 0.185>0.01

DBP and Recact

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	DBP	-.006	.007	.980	1	.322	.994	.977	1.010
	DBP by recact	.013	.004	11.864	1	.001	1.013	1.003	1.023
	recact	-1.313	.286	21.099	1	.000	.269	.129	.562
	Constant	-.605	.481	1.581	1	.209	.546		

a. Variable(s) entered on step 1: DBP, DBP * recact , recact.

H0: $\beta 3=0$
H1: $\beta 3\neq 0$
Wald test statistics = 11.864 p-value = 0.001<0.01 Do not reject H0 DBP*Recact should be in the model. DBP*Recact is significant.

WLKBIK and Recact

WLKBIK*Recact is not significant. Should not be in the model. P-value 0.219>0.01

7. Check for linear relationship between each pair of the selected continuous covariates and their respective logits i.e. is the logit linear in the continuous covariates?

HDL

Statistics

		minutes of sedentary activity per week	HDL cholesterol	diastolic blood pressure
N	Valid	5786	5512	5330
	Missing	71	345	527
Minimum		0	15	40
Maximum		840	144	134
Percentiles	25	180.00	41.00	62.00
	50	300.00	50.00	70.00
	75	480.00	62.00	78.00

HDLGP	Midpoint
15-41	28
42-50	46
51-62	56.5
63- 144	103.5

HDLGP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1472	25.1	26.7	26.7
	1	1347	23.0	24.4	51.1
	2	1407	24.0	25.5	76.7
	3	1286	22.0	23.3	100.0
	Total	5512	94.1	100.0	
Missing	System	345	5.9		
Total		5857	100.0		

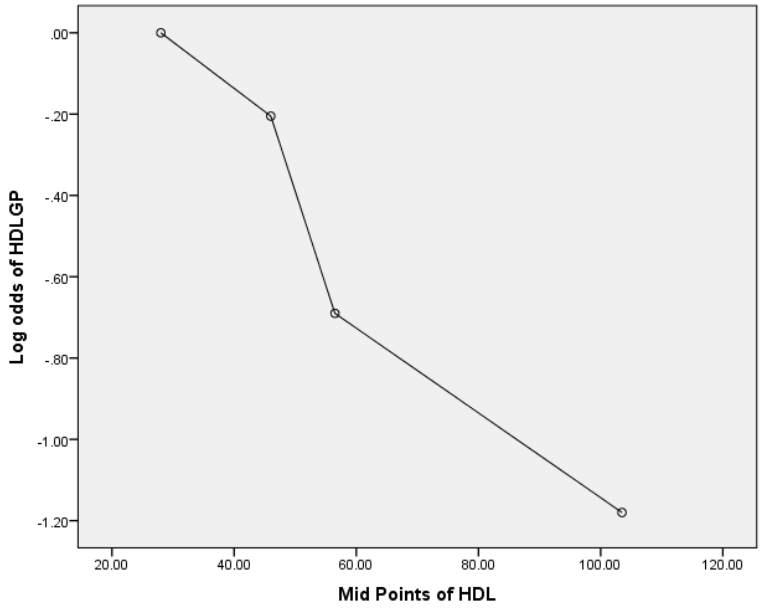
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	DBP	.010	.003	9.910	1	.002	1.011	1.002	1.019
	SEDMIN	.001	.000	28.452	1	.000	1.001	1.001	1.002
	HDLGP			105.125	3	.000			
	HDLGP(1)	-.205	.100	4.230	1	.040	.815	.630	1.053
	HDLGP(2)	-.694	.108	40.969	1	.000	.500	.378	.661
	HDLGP(3)	-1.177	.127	85.206	1	.000	.308	.222	.428
	Constant	-2.361	.258	83.556	1	.000	.094		

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

H0: $\beta i = 0$
H1: $\beta i \neq 0$ i = 1,2,3

Wald statistic are 9.91, 28.452, 105.125 all the p-value (0.00) < 0.01 Reject H0
All the variables are significant and should be in the model. In HDLGP group 42-50 is borderline significant compare with 15-41 group (p-value =0.04).



There is no departure from the linear sight. Keep decreasing. Logodds keep changing linearly.

SEDMIN

SEDMINGP	Midpoint
0-180	90
181-300	240.5
301-480	390.5
481- 840	660.5

SEDMINGP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1974	33.7	34.1	34.1
	1	1426	24.3	24.6	58.8
	2	1539	26.3	26.6	85.4
	3	847	14.5	14.6	100.0
	Total	5786	98.8	100.0	
Missing	System	71	1.2		
Total		5857	100.0		

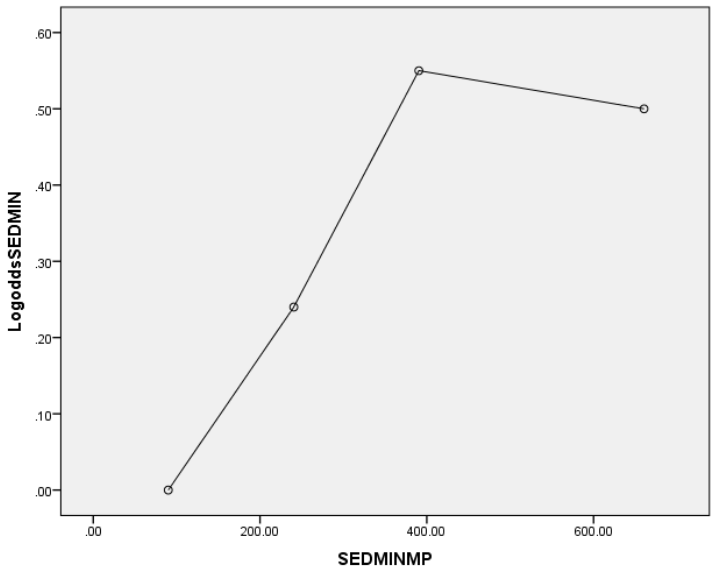
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	DBP	.011	.003	10.928	1	.001	1.011	1.002	1.020
	HDL	-.030	.003	105.541	1	.000	.971	.964	.978
	SEDMINGP			33.080	3	.000			
	SEDMINGP(1)	.241	.111	4.733	1	.030	1.272	.957	1.691
	SEDMINGP(2)	.549	.103	28.250	1	.000	1.731	1.327	2.258
	SEDMINGP(3)	.504	.124	16.644	1	.000	1.656	1.204	2.277
	Constant	-1.278	.294	18.939	1	.000	.279		

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

H0: $\beta i = 0$
H1: $\beta i \neq 0$ i = 1,2,3

Wald statistic are 10.928, 105.541, 33.08 all the p-value (0.00) < 0.01 Reject H0
All the variables are significant and should be in the model. In the SEDMINGP group 181-300 is borderline significant when compare with 0-180 group (p-value =0.03).



Difficult to tell departure from the linear sight or just random variation.

DBP

HDLGP	Midpoint
40-62	51
63-70	66.5
71-78	74.5
79- 134	106.5

DBPGP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1419	24.2	26.6	26.6
	1	1371	23.4	25.7	52.3
	2	1313	22.4	24.6	77.0
	3	1227	20.9	23.0	100.0
	Total	5330	91.0	100.0	
Missing	System	527	9.0		
Total		5857	100.0		

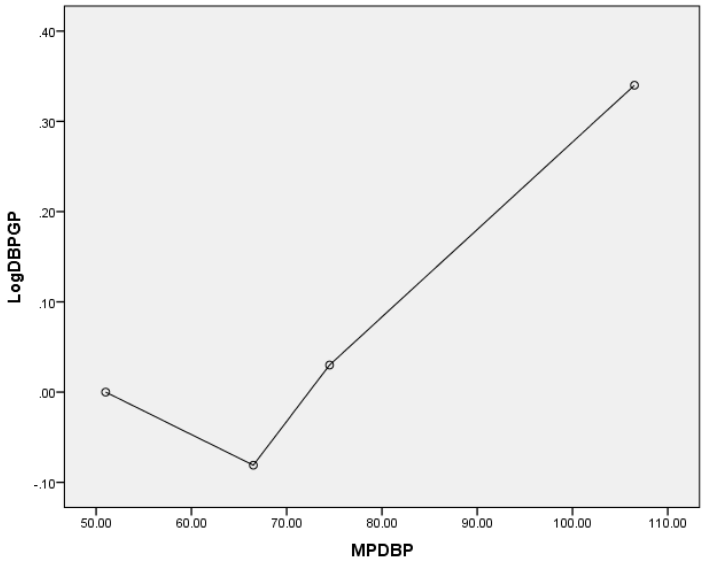
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	HDL	-.029	.003	104.084	1	.000	.971	.964	.978
	SEDMIN	.001	.000	28.504	1	.000	1.001	1.001	1.002
	DBPGP			17.429	3	.001			
	DBPGP(1)	-.081	.115	.500	1	.480	.922	.685	1.240
	DBPGP(2)	.032	.114	.079	1	.778	1.033	.769	1.387
	DBPGP(3)	.344	.110	9.838	1	.002	1.411	1.064	1.873
	Constant	-.652	.176	13.696	1	.000	.521		

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

H0: $\beta i = 0$
H1: $\beta i \neq 0$ i = 1,2,3

Wald statistic are 104.084, 28.504, 17.429 all the p-value (0.00) < 0.01 Reject H0
All the variables are significant and should be in the model. In the DBPGP groups 63-70 and 71-78 are not significant compare with 40-62 group (p-value =0.03).



Difficult to tell departure from the linear sight or just random variation.

8. Using the best model, include an overall assessment of model fit (Likelihood ratio tests, Wald tests, Hosmer Lemeshow test, Classification tables) and use of diagnostic graphs of leverage, standardized residuals, Cooks distance, deviances, and Dfbetas in the final model.
INTERPRET ALL TABLES AND GRAPHS.

Test for the significant of the model

H0: $\beta1=\beta2=\beta3=\beta4=\beta5=\beta6=\beta7=\beta8$

H1: At least one coefficient is different from zero.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	99% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	HDL	-.030	.003	109.466	1	.000	.970	.963	.977
	Constant	-.171	.144	1.415	1	.234	.843		
Step 2[b]	GENDER(1)	.959	.087	121.364	1	.000	2.609	2.085	3.265
	HDL	-.041	.003	166.675	1	.000	.960	.952	.968
	Constant	-.173	.149	1.348	1	.246	.841		
Step 3[c]	GENDER(1)	.983	.088	125.801	1	.000	2.672	2.132	3.348
	HDL	-.041	.003	166.960	1	.000	.960	.952	.968
	SEDMIN	.001	.000	33.100	1	.000	1.001	1.001	1.002
	Constant	-.583	.166	12.361	1	.000	.558		
Step 4[d]	recact			40.341	3	.000			
	recact(1)	-.395	.097	16.417	1	.000	.674	.524	.866
	recact(2)	-.708	.208	11.627	1	.001	.493	.289	.841
	recact(3)	-.758	.160	22.487	1	.000	.469	.311	.707
	GENDER(1)	.913	.089	105.890	1	.000	2.492	1.983	3.132
	HDL	-.039	.003	148.667	1	.000	.962	.954	.970
	SEDMIN	.001	.000	36.556	1	.000	1.001	1.001	1.002
	Constant	-.466	.167	7.771	1	.005	.627		
Step 5[e]	recact			40.990	3	.000			
	recact(1)	-.407	.098	17.341	1	.000	.665	.517	.856
	recact(2)	-.711	.208	11.709	1	.001	.491	.287	.839
	recact(3)	-.757	.160	22.369	1	.000	.469	.311	.709
	GENDER(1)	.976	.090	117.324	1	.000	2.655	2.105	3.349
	HDL	-.038	.003	146.164	1	.000	.962	.954	.970
	DBP	.017	.003	24.788	1	.000	1.018	1.008	1.027
	SEDMIN	.001	.000	36.022	1	.000	1.001	1.001	1.002
	Constant	-1.742	.308	32.062	1	.000	.175		
Step 6[f]	WLKBIK(1)	.300	.101	8.846	1	.003	1.350	1.041	1.751
	recact			38.833	3	.000			
	recact(1)	-.406	.098	17.215	1	.000	.666	.518	.857
	recact(2)	-.681	.208	10.704	1	.001	.506	.296	.865
	recact(3)	-.732	.160	20.860	1	.000	.481	.318	.727
	GENDER(1)	.971	.090	115.728	1	.000	2.640	2.093	3.331
	HDL	-.038	.003	144.507	1	.000	.963	.955	.970
	DBP	.018	.003	25.579	1	.000	1.018	1.009	1.027
	SEDMIN	.001	.000	32.560	1	.000	1.001	1.001	1.002
	Constant	-1.986	.319	38.719	1	.000	.137		

a. Variable(s) entered on step 1: HDL.
b. Variable(s) entered on step 2: GENDER.
c. Variable(s) entered on step 3: SEDMIN.
d. Variable(s) entered on step 4: recact.
e. Variable(s) entered on step 5: DBP.
f. Variable(s) entered on step 6: WLKBIK.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	124.627	1	.000
	Block	124.627	1	.000
	Model	124.627	1	.000
Step 2	Step	127.368	1	.000
	Block	251.995	2	.000
	Model	251.995	2	.000
Step 3	Step	32.550	1	.000
	Block	284.546	3	.000
	Model	284.546	3	.000
Step 4	Step	42.958	3	.000
	Block	327.503	6	.000
	Model	327.503	6	.000
Step 5	Step	24.876	1	.000
	Block	352.379	7	.000
	Model	352.379	7	.000
Step 6	Step	9.170	1	.002
	Block	361.549	8	.000
	Model	361.549	8	.000

LR test for overall sig of the 6 coefficients for the independent variables in the model is based on change -2LL = 4261.435 – 3899.887 = 361.55 Chi-square p-value 0.000<0.01 reject H0

At least one or and perhaps all coefficients are not equal to zero.

Overall model is good

Wald test

H0: $\beta i=0$
H1: $\beta i \neq 0$ i=1,2,3,4,5,6

Wald statistics = 8.846,38.833,115.728,144.507,25.579,32.560
All p-value 0.00<0.01 Reject H0

All the variable in the model are significant. So all the variables should be in the model.

Hosmer Lemeshow test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.088	8	.080
2	7.405	8	.494
3	8.526	8	.384
4	3.806	8	.874
5	4.566	8	.803
6	15.211	8	.055

H0: Model is good

H1: Model is not good

Chi-square test statistic = 15.211 and p-value = 0.055 > 0.01 Do not reject H0 (Boarder line significant)

The Hosmer & Lemeshow test of the goodness of fit suggests the model is a good fit to the data as p=0.055 (>.01)

Therefore, overall model is good fit.

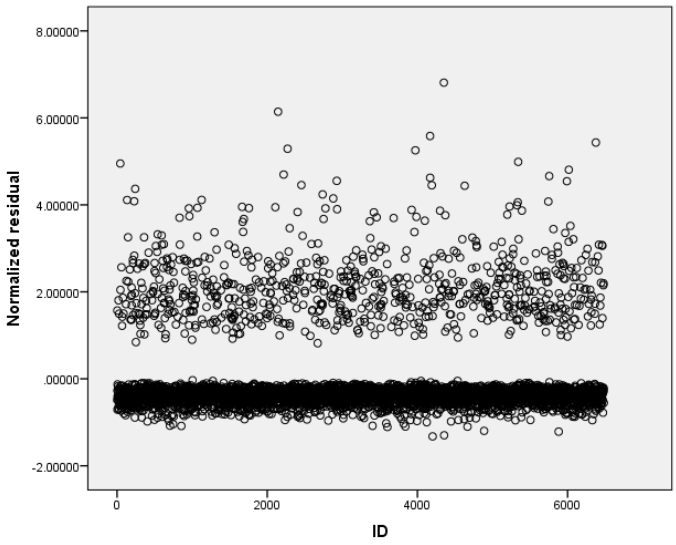
Classification Table[a]

Observed			Predicted		
			bmi more than 35		Percentage Correct
			no	yes	
Step 1	bmi more than 35	no	4150	0	100.0
		yes	768	0	.0
	Overall Percentage				84.4
Step 2	bmi more than 35	no	4150	0	100.0
		yes	768	0	.0
	Overall Percentage				84.4
Step 3	bmi more than 35	no	4143	7	99.8
		yes	766	2	.3
	Overall Percentage				84.3
Step 4	bmi more than 35	no	4142	8	99.8
		yes	763	5	.7
	Overall Percentage				84.3
Step 5	bmi more than 35	no	4134	16	99.6
		yes	761	7	.9
	Overall Percentage				84.2
Step 6	bmi more than 35	no	4138	12	99.7
		yes	758	10	1.3
	Overall Percentage				84.3

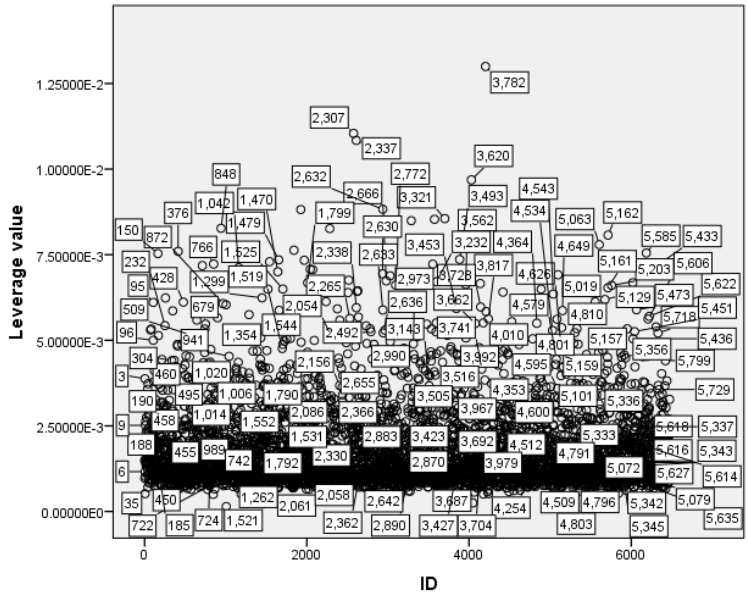
a. The cut value is .500

12 people who do not have obese wrongly classified as they do have obese in our prediction.
758 people who do have obese wrongly classified as they do not have obese in our predictions.

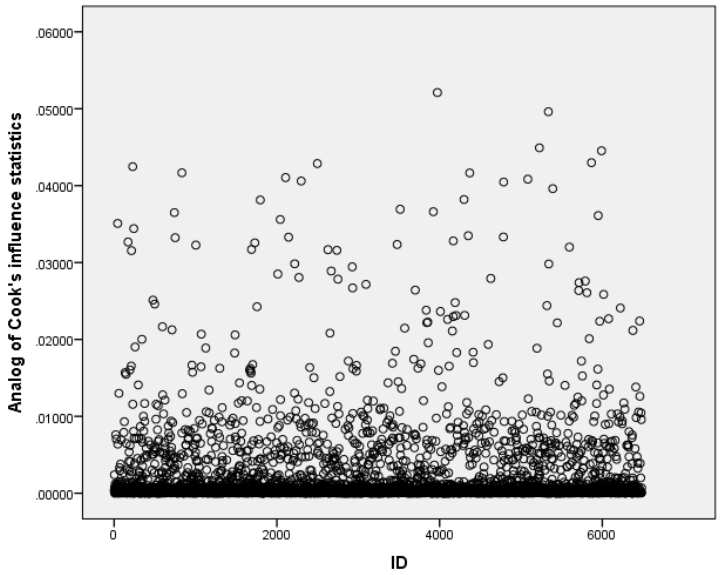
Overall accuracy of our prediction is 84.3%. When we see overall accuracy from step 1 it is keep improving when new variable added to the model.



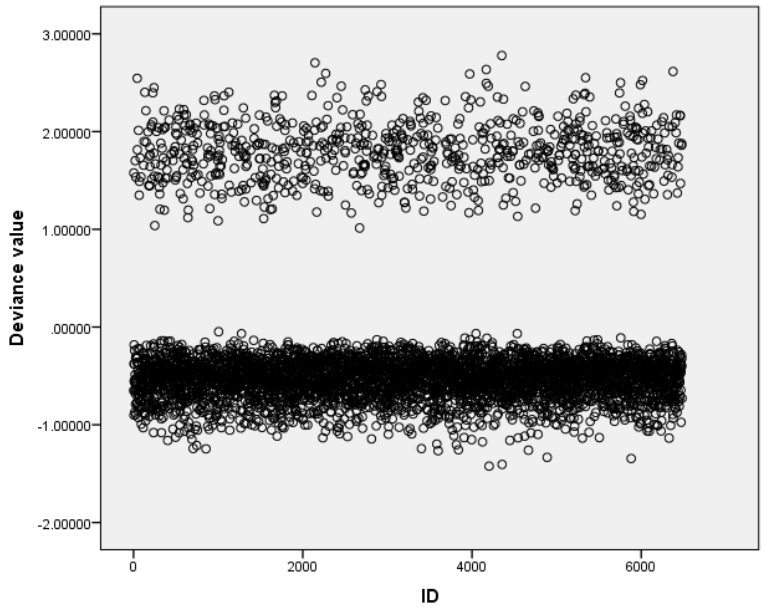
Residual data represent difference between the observed and predicted probability of the obese. Standardized residuals should be approximately normally distributed with mean zero and variance 1. We can see here thick black line of data points stay closer to zero Which means most of the data that we predicted are correct. There are considerable amount of data stay outside of that rage too.



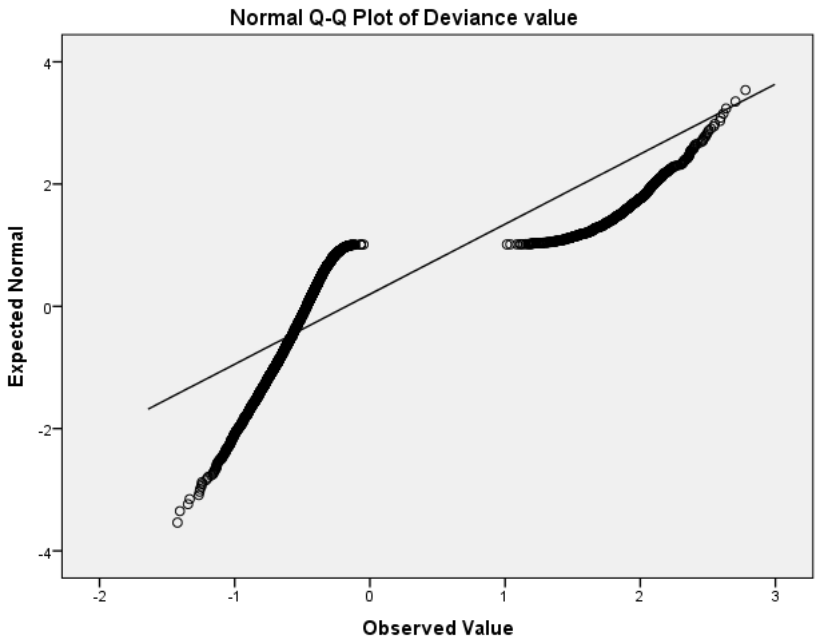
Leverage values are bounded by 0 and 1. Used for detecting observations having a large impact on predicted values. On the graph we can see most of the points are very closer to zero which is showing by black thick line of data points. Some of them way off from the zero (data points such as 3782, 2307,2337) but still stay inside the bounded regent. In general, we do not see any observations having a large impact on predicted values unnecessarily.



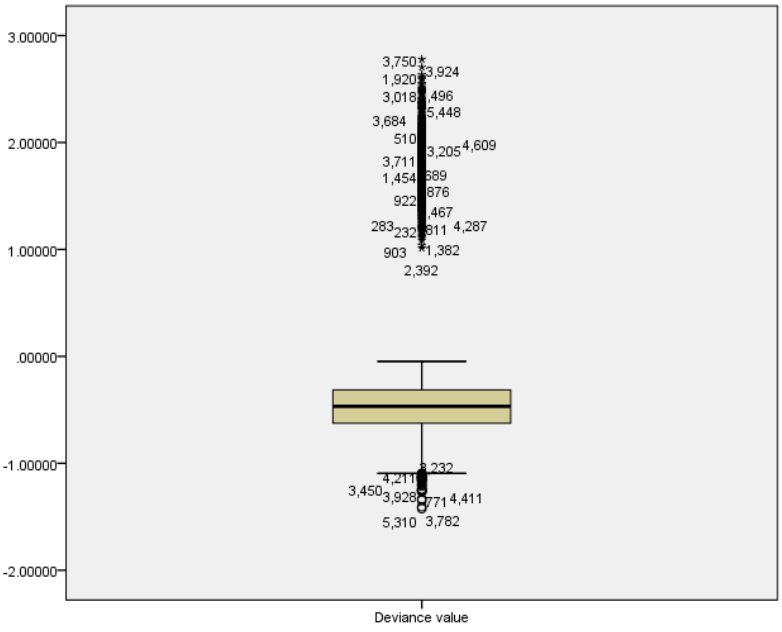
COOK's DISTANCE is a measure of overall influence of a case on the model and values greater than one may be cause for concern. In this graph we can see that black thick line of data very closer to zero which is good. There are no data value greater than one. Most of the values stay very closer to zero.



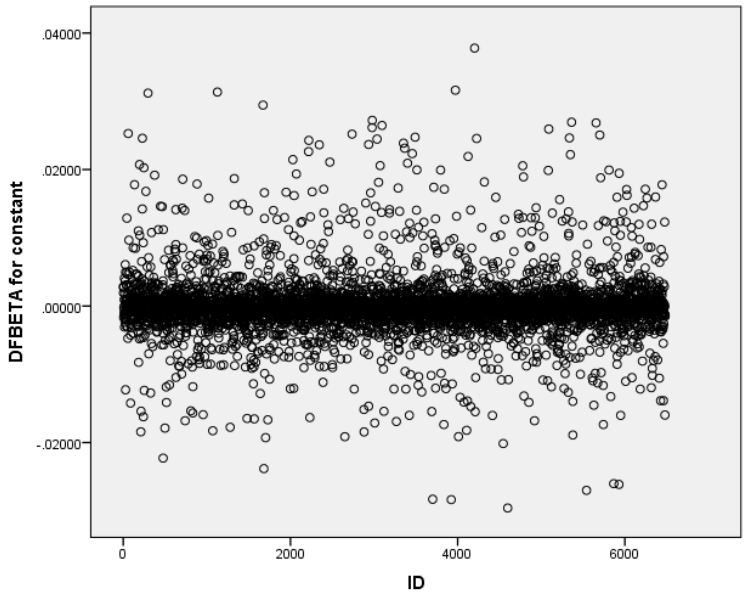
Deviance is the lack of fit or deviance from the observed data. Deviances and -2log likelihoods are conceptually identical. So we can see overall model good or not. Deviance in logistic regression plays the same role as SSE (Residual sum of squares) in linear regression. Therefore, when all points getting closer to zero the model is best. In here most of data points stay closer to zero. Which is good. Our model is good.



Most of the data point are not on the line or closer to the line. So deviance value are not normally distributed.



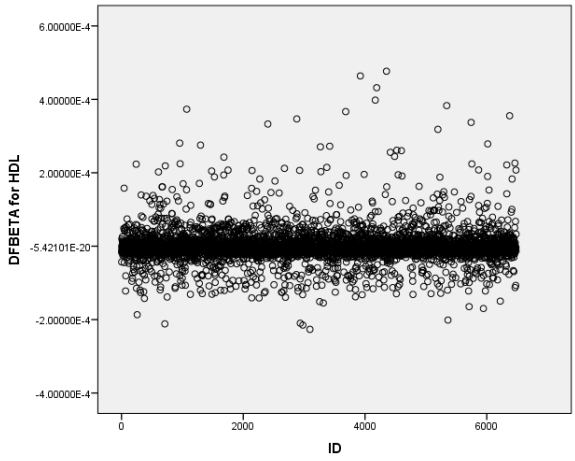
There are considerable amount of data outside the mean area. We consider those values as outlies.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

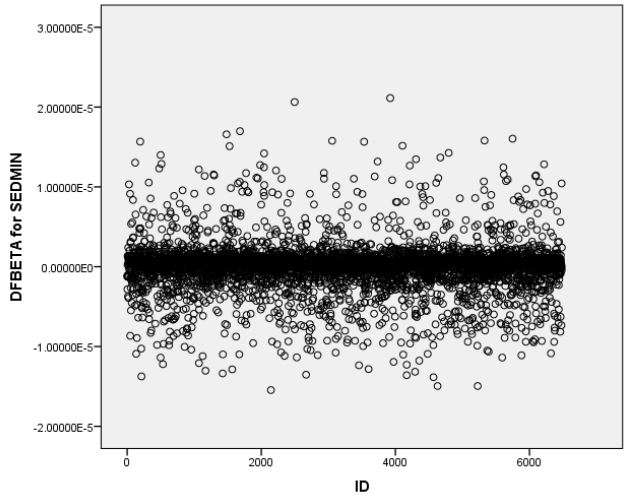
Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on constant.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

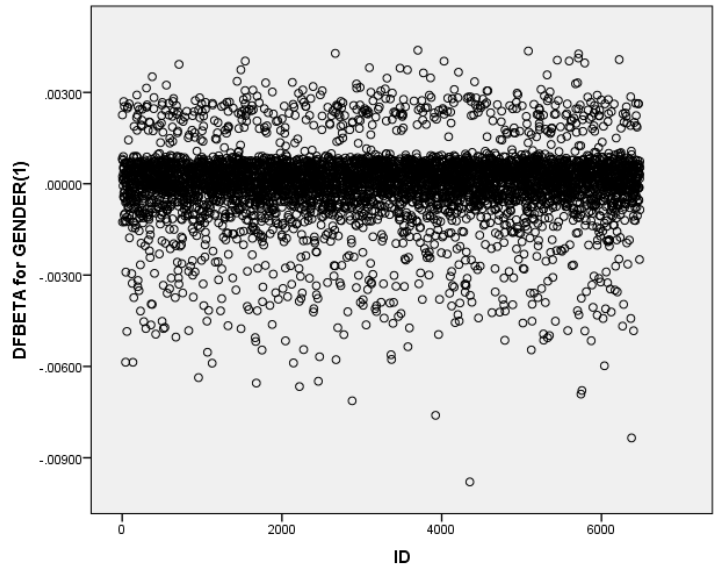
Most of the data points stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of HDL.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

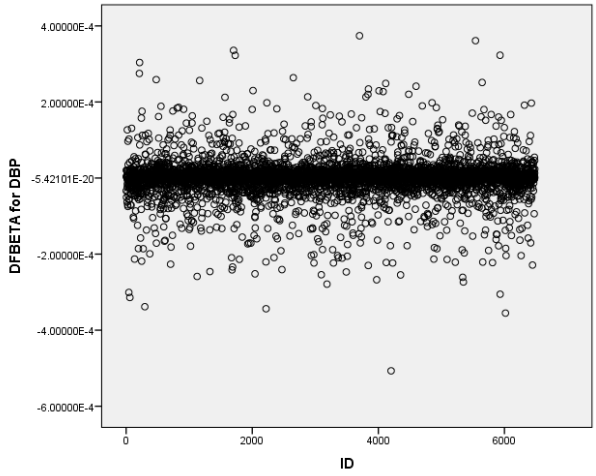
Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of SEDMIN.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

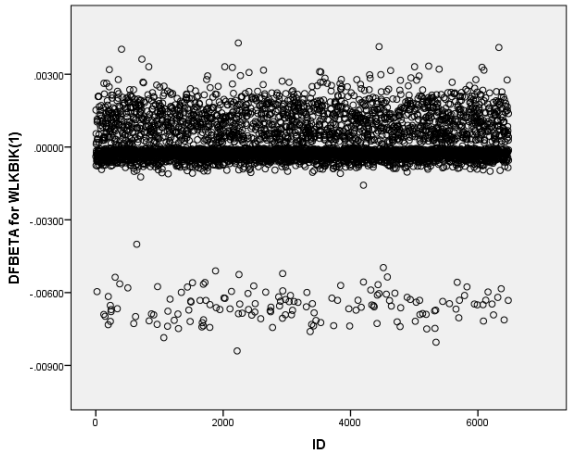
Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of gender. But when compare with other Dfbeta (B1i) this difference is considerably high.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

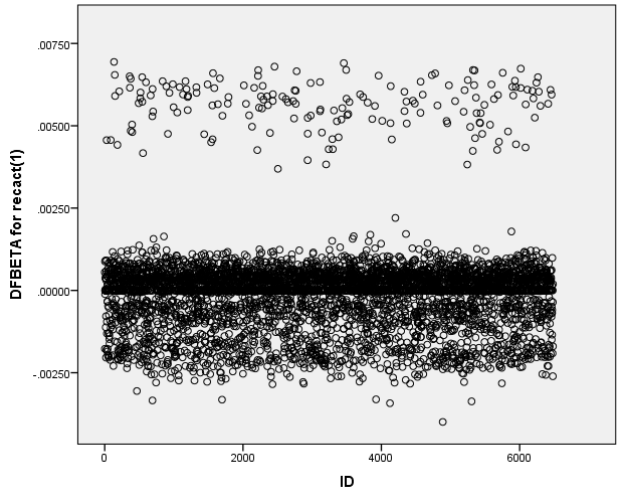
Most of the data points stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of DBP.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

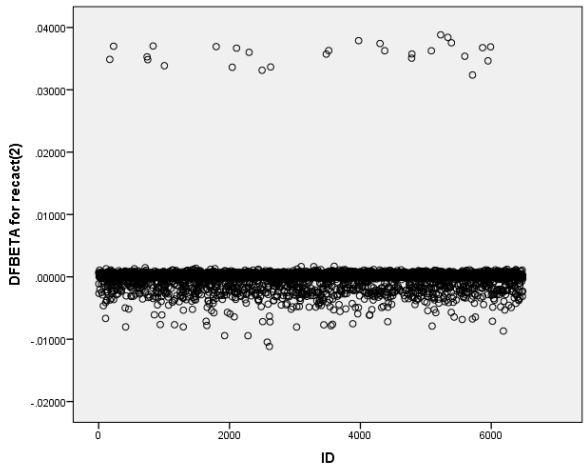
Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of WLKBIK. But when compare with other Dfbeta (B1i) this difference is considerably high.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

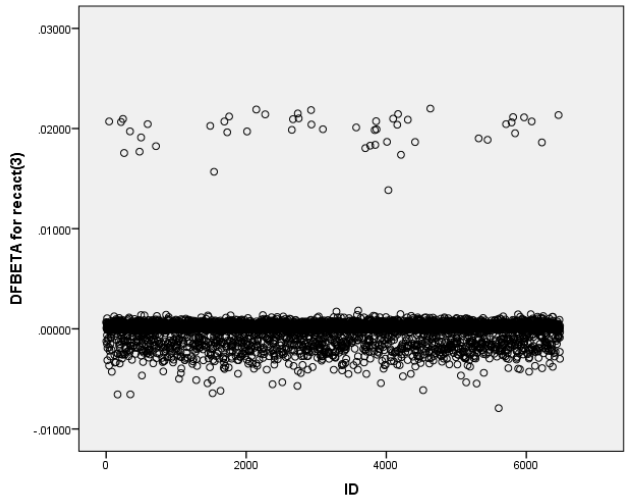
Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of recact(low). But when compare with other Dfbeta (B1i) this difference is considerably high.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of recact(Moderate). But some of data points way off from zero.



Dfbeta (B1i) = B1 - B1i

Where B1 is the value of the coefficient when all cases are included and B1i is the value of the coefficient when the ith case is excluded.

Most of the data points zero or stay very closer to zero. Which mean when most of the case excluding from the model, it does not make big difference on coefficient of recact(high). But some of data points way off from zero.

9. Would you suggest any changes to the final model? Explain.

There were suspected confounding. So interaction between HDL and Recact as well as interaction between DBP and Recact were significant. So we can add those interactions to the model and see how model performs. We can look for other variable which might be significant that we did not look while doing this analysis such as weight.