PROJECT 6

Kankanamge Harsha

The data **sbpsmk.sav** shows systolic **blood pressure (mmHg), age (years), body mass index (kg/m_2),** and **smoking history** (smokers or non-smokers) of 32 white males over the age of 40.

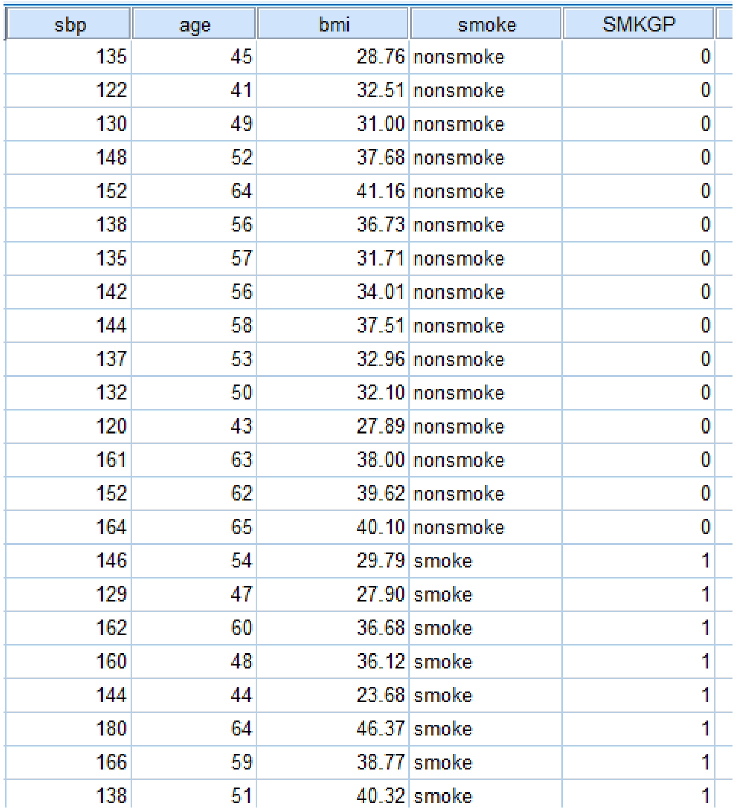
1)Determine the proportion of smokers and non-smokers in the data set.

smoke

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	nonsmoke	15	46.9	46.9	46.9
	smoke	17	53.1	53.1	100.0
	Total	32	100.0	100.0	

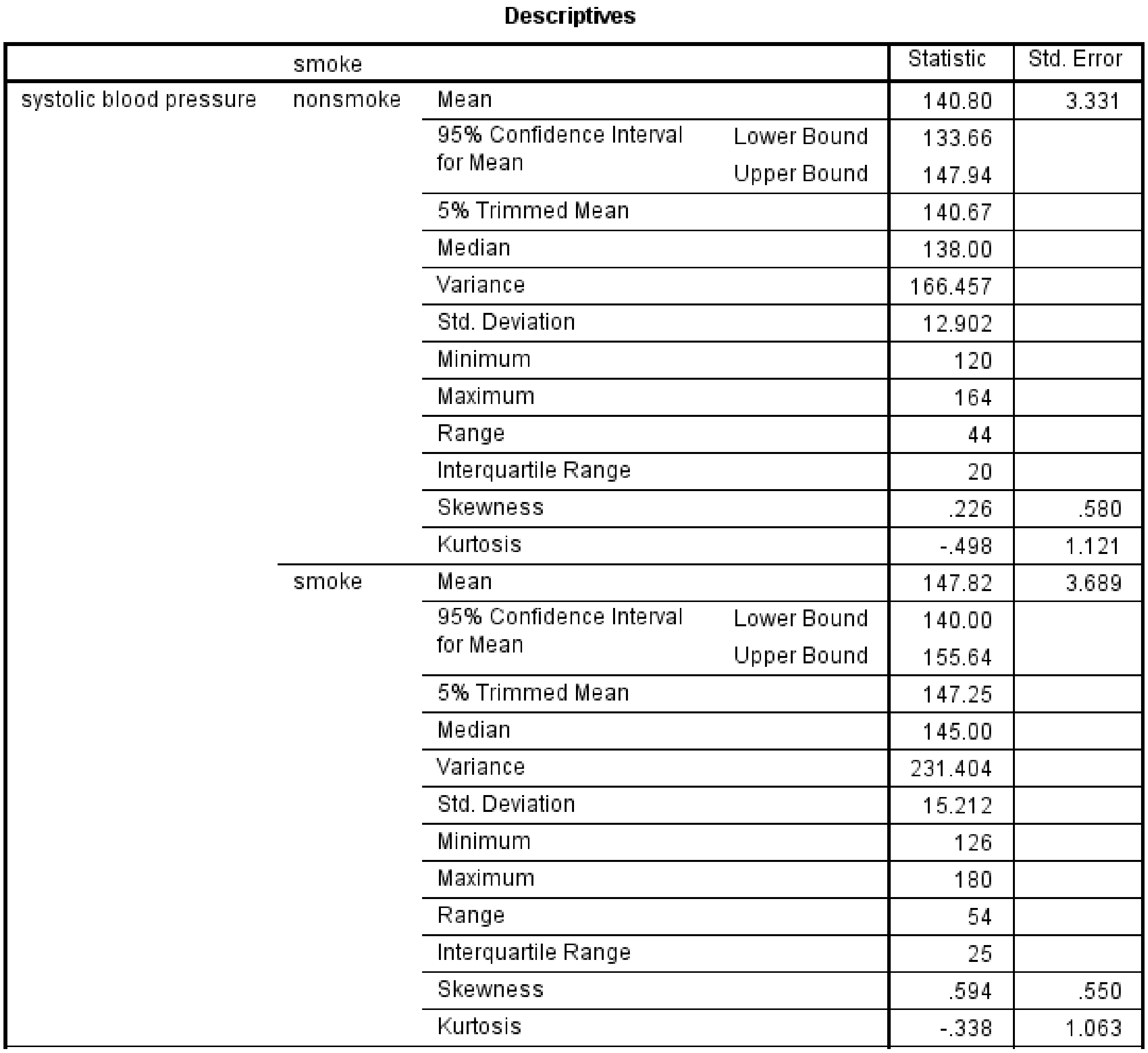
Sample represent nonsmokers proportion 0.469 and smokers proportion 0.531 on total sample proportion of 1.

2)Recode the 'SMOKE' variable into a numerical variable SMKGP using '0' for non- smokers and '1' for smokers.



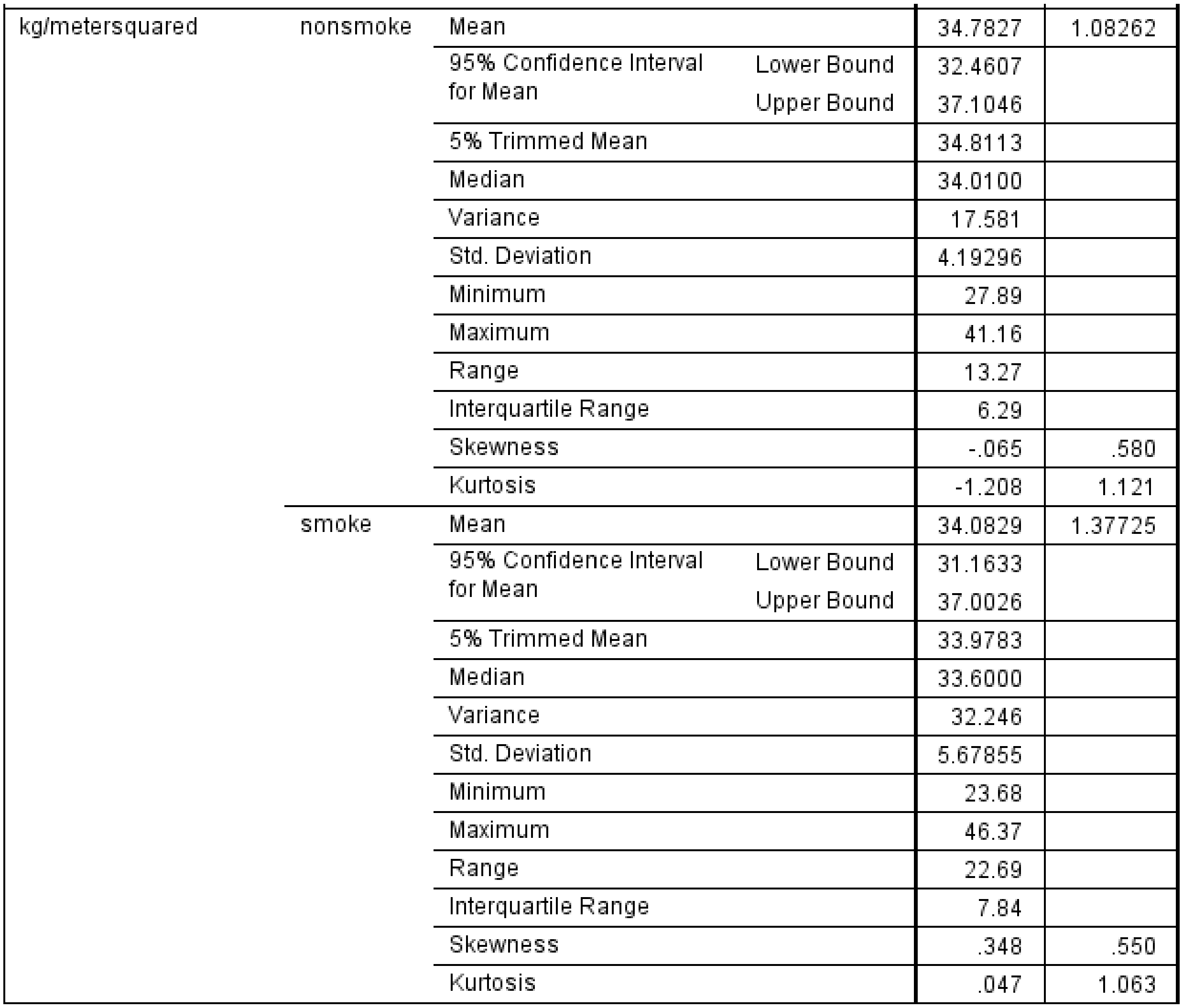
sbp	age	bmi	smoke	SMKGP
135	45	28.76	nonsmoke	0
122	41	32.51	nonsmoke	0
130	49	31.00	nonsmoke	0
148	52	37.68	nonsmoke	0
152	64	41.16	nonsmoke	0
138	56	36.73	nonsmoke	0
135	57	31.71	nonsmoke	0
142	56	34.01	nonsmoke	0
144	58	37.51	nonsmoke	0
137	53	32.96	nonsmoke	0
132	50	32.10	nonsmoke	0
120	43	27.89	nonsmoke	0
161	63	38.00	nonsmoke	0
152	62	39.62	nonsmoke	0
164	65	40.10	nonsmoke	0
146	54	29.79	smoke	1
129	47	27.90	smoke	1
162	60	36.68	smoke	1
160	48	36.12	smoke	1
144	44	23.68	smoke	1
180	64	46.37	smoke	1
166	59	38.77	smoke	1
138	51	40.32	smoke	1

3)Using ***Explore***, compare means and standard deviations of SBP, AGE, and BMI of smokers and non-smokers.

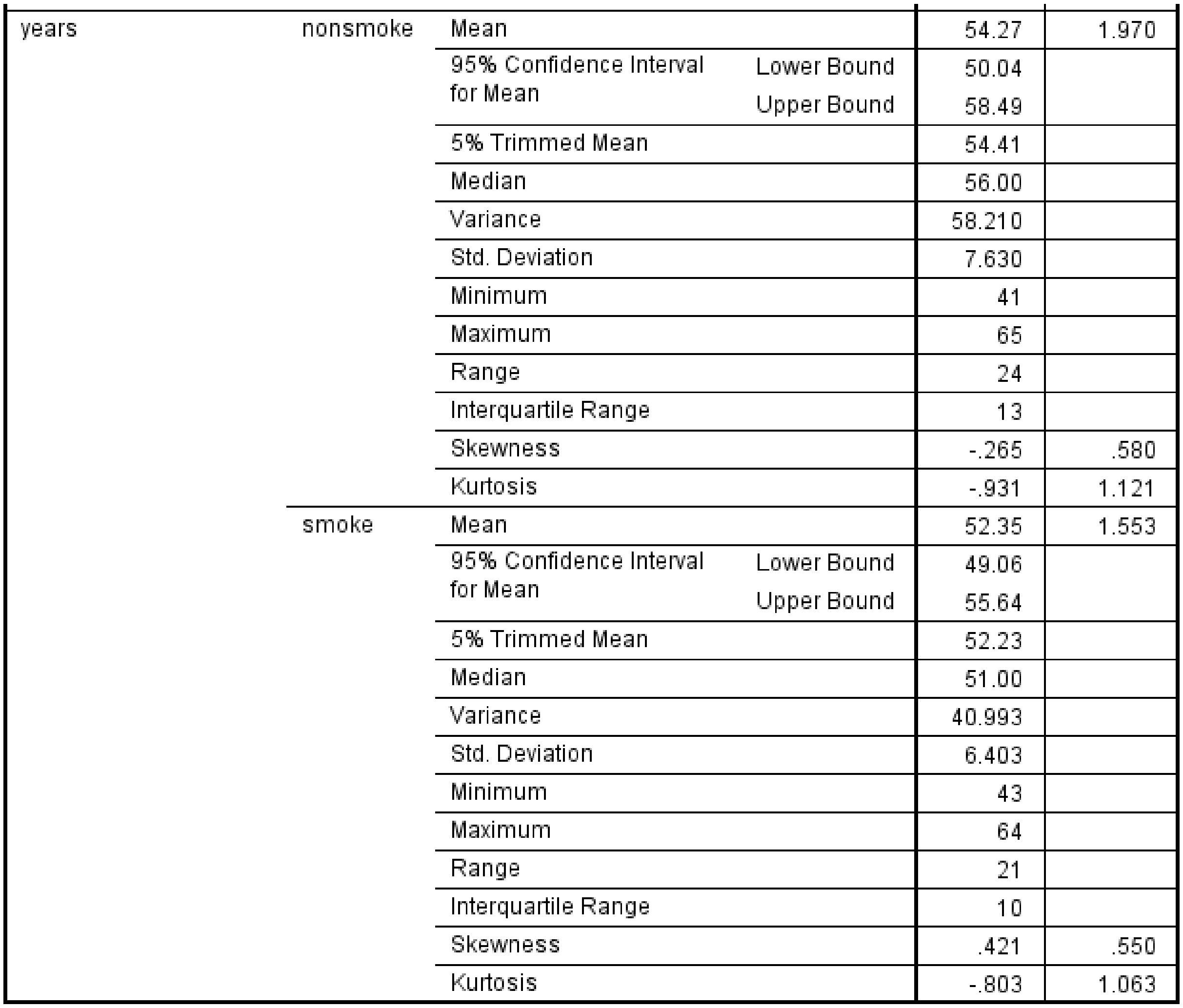


Descriptives

	smoke			Statistic	Std. Error
systolic blood pressure	nonsmoke	Mean		140.80	3.331
		95% Confidence Interval for Mean	Lower Bound	133.66	
			Upper Bound	147.94	
		5% Trimmed Mean		140.67	
		Median		138.00	
		Variance		166.457	
		Std. Deviation		12.902	
		Minimum		120	
		Maximum		164	
		Range		44	
		Interquartile Range		20	
		Skewness		.226	.580
		Kurtosis		-.498	1.121
	smoke	Mean		147.82	3.689
		95% Confidence Interval for Mean	Lower Bound	140.00	
			Upper Bound	155.64	
		5% Trimmed Mean		147.25	
		Median		145.00	
		Variance		231.404	
		Std. Deviation		15.212	
		Minimum		126	
		Maximum		180	
		Range		54	
		Interquartile Range		25	
		Skewness		.594	.550
		Kurtosis		-.338	1.063



kg/metersquared	nonsmoke	Mean		34.7827	1.08262
		95% Confidence Interval for Mean	Lower Bound	32.4607	
			Upper Bound	37.1046	
		5% Trimmed Mean		34.8113	
		Median		34.0100	
		Variance		17.581	
		Std. Deviation		4.19296	
		Minimum		27.89	
		Maximum		41.16	
		Range		13.27	
		Interquartile Range		6.29	
		Skewness		-.065	.580
		Kurtosis		-1.208	1.121
	smoke	Mean		34.0829	1.37725
		95% Confidence Interval for Mean	Lower Bound	31.1633	
			Upper Bound	37.0026	
		5% Trimmed Mean		33.9783	
		Median		33.6000	
		Variance		32.246	
		Std. Deviation		5.67855	
		Minimum		23.68	
		Maximum		46.37	
		Range		22.69	
		Interquartile Range		7.84	
		Skewness		.348	.550
		Kurtosis		.047	1.063



years	nonsmoke	Mean		54.27	1.970
		95% Confidence Interval for Mean	Lower Bound	50.04	
			Upper Bound	58.49	
		5% Trimmed Mean		54.41	
		Median		56.00	
		Variance		58.210	
		Std. Deviation		7.630	
		Minimum		41	
		Maximum		65	
		Range		24	
		Interquartile Range		13	
		Skewness		-.265	.580
		Kurtosis		-.931	1.121
	smoke	Mean		52.35	1.553
		95% Confidence Interval for Mean	Lower Bound	49.06	
			Upper Bound	55.64	
		5% Trimmed Mean		52.23	
		Median		51.00	
		Variance		40.993	
		Std. Deviation		6.403	
		Minimum		43	
		Maximum		64	
		Range		21	
		Interquartile Range		10	
		Skewness		.421	.550
		Kurtosis		-.803	1.063

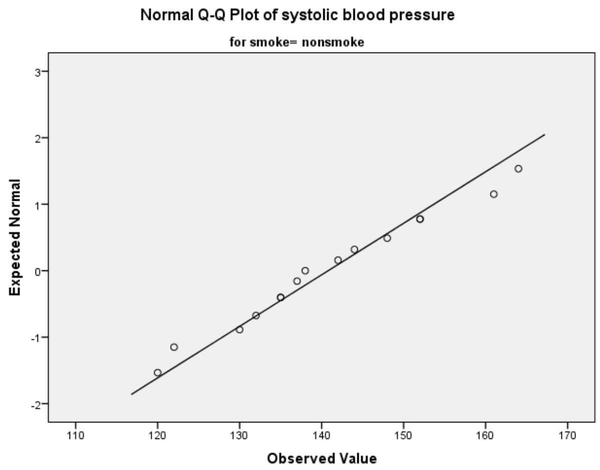
	Nonsmoke		smoke	
	Mean	Std	mean	Std
systolic blood pressure(sbp)	140.8	12.902	147.82	15.212
kg/metersquared(BMI)	34.7827	4.19296	34.0829	5.67855
years	54.27	7.63	52.35	6.403

WE can’t see significant different in mean values of BIM factor among nonsmokers and smokers in the sample. And also std(standard deviation) is little higher on smokers than nonsmokers. Which means in the sample smokers have little higher variation of BIM values than nonsmokers.

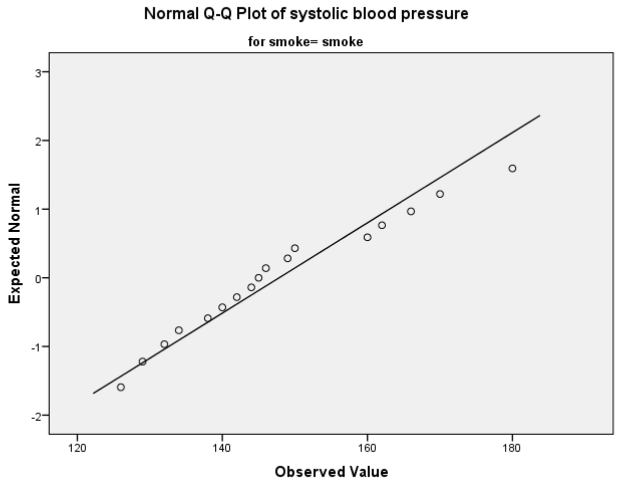
There is a significant different on mean and std of systolic blood pressure(sbp) among nonsmokers and smokers. Smokers have higher mean and std for systolic blood pressure(sbp) than nonsmokers that is showing on above comparison table. Which means smokers blood pressure is higher and its variations also higher than nonsmokers.

Nonsmoker have little higher mean and std for age than smokers. Which means in this sample nonsmokers are little older than smokers and age variation also little higher than smokers in this sample.

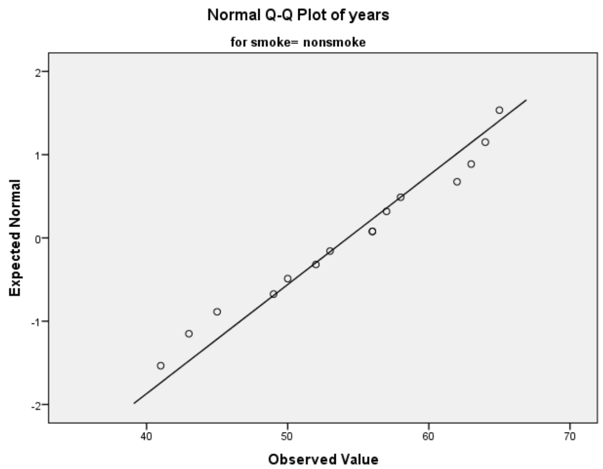
Even though nonsmokers have little higher age than smokers their systolic blood pressure(sbp) is lower compared with smokers on mean.


Normal Q-Q Plot of systolic blood pressure

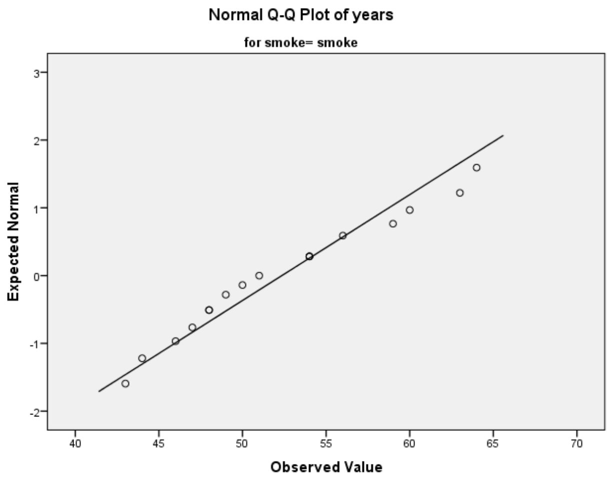
for smoke= nonsmoke


Normal Q-Q Plot of systolic blood pressure

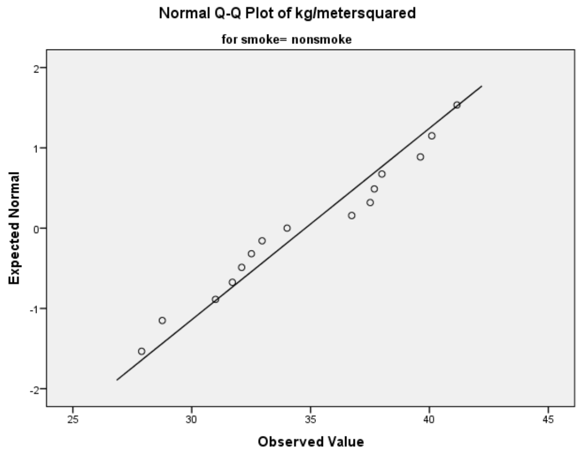
for smoke= smoke


Normal Q-Q Plot of years

for smoke= nonsmoke

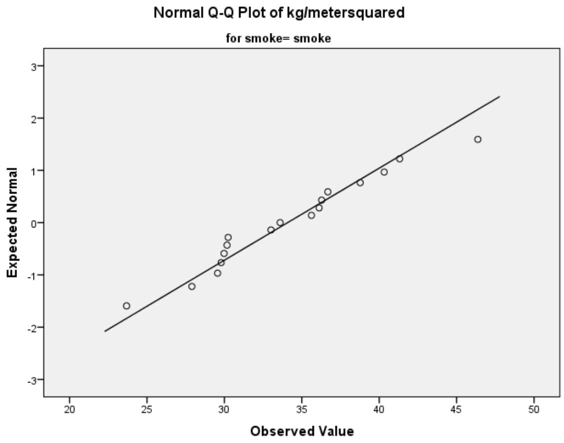

Normal Q-Q Plot of years

for smoke= smoke



Normal Q-Q Plot of kg/metersquared

for smoke= nonsmoke

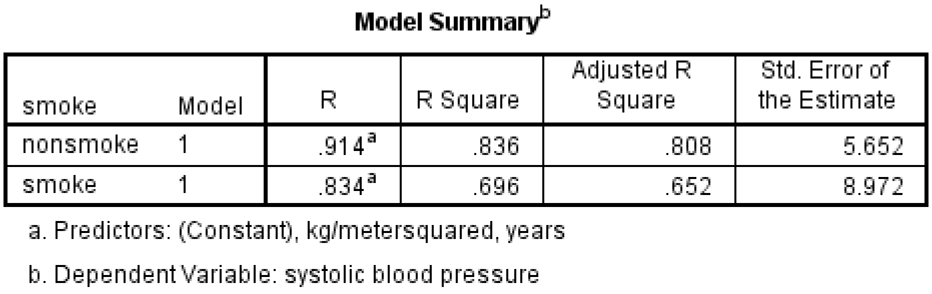


Normal Q-Q Plot of kg/metersquared

for smoke= smoke

All the graph shows that most of the data point fitting onto the line. Which mean data seems to be normal for smokers and nonsmokers among all categories.

4)Using ***Split-File***, compare the linear regression analysis of BMI and SBP separately for smokers and non-smokers. Tabulate relevant estimates from the regression analysis for smokers and non-smokers and summarize your conclusions.

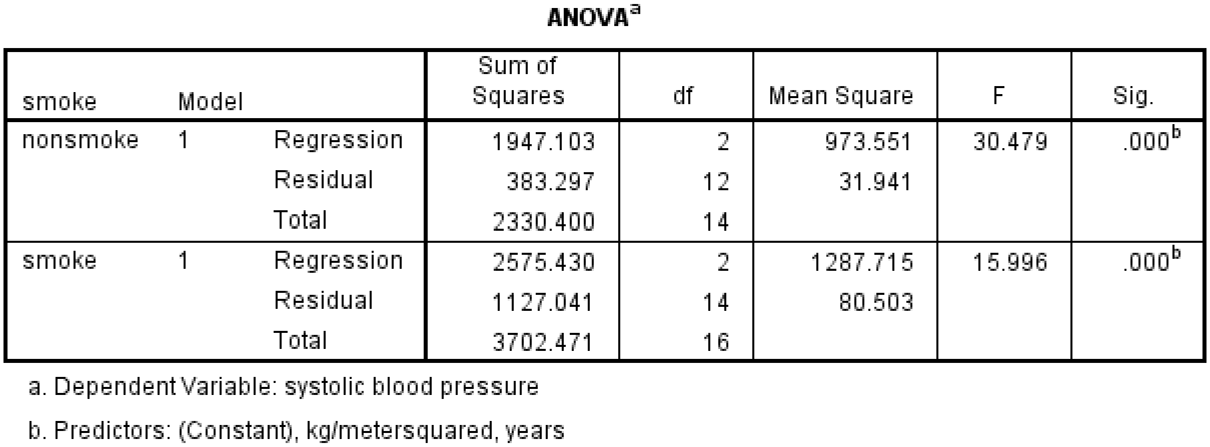


Model Summary[b]

smoke	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
nonsmoke	1	.914[a]	.836	.808	5.652
smoke	1	.834[a]	.696	.652	8.972

a. Predictors: (Constant), kg/metersquared, years

b. Dependent Variable: systolic blood pressure

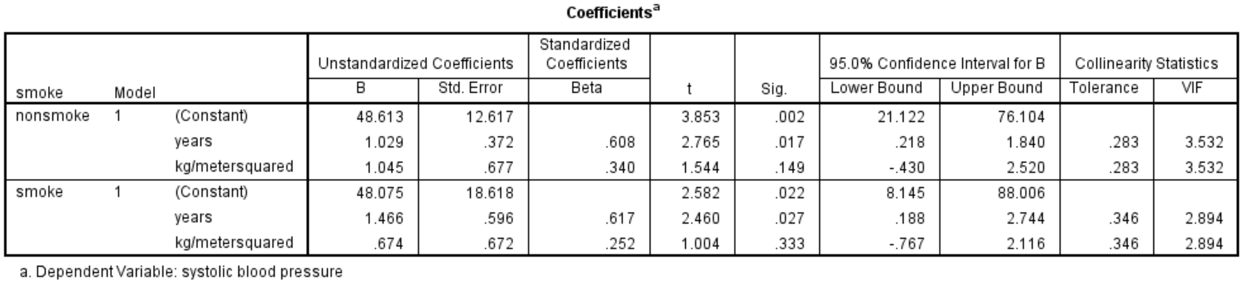


ANOVA[a]

smoke	Model		Sum of Squares	df	Mean Square	F	Sig.
nonsmoke	1	Regression	1947.103	2	973.551	30.479	.000[b]
		Residual	383.297	12	31.941		
		Total	2330.400	14			
smoke	1	Regression	2575.430	2	1287.715	15.996	.000[b]
		Residual	1127.041	14	80.503		
		Total	3702.471	16			

a. Dependent Variable: systolic blood pressure

b. Predictors: (Constant), kg/metersquared, years



Coefficients[a]

			Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
smoke	Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
nonsmoke	1	(Constant)	48.613	12.617		3.853	.002	21.122	76.104		
		years	1.029	.372	.608	2.765	.017	.218	1.840	.283	3.532
		kg/metersquared	1.045	.677	.340	1.544	.149	-.430	2.520	.283	3.532
smoke	1	(Constant)	48.075	18.618		2.582	.022	8.145	88.006		
		years	1.466	.596	.617	2.460	.027	.188	2.744	.346	2.894
		kg/metersquared	.674	.672	.252	1.004	.333	-.767	2.116	.346	2.894

a. Dependent Variable: systolic blood pressure

Systolic blood pressure(sbp) model for nonsmokers:

systolic blood pressure(sbp)^(estimated)= 1.045*kg/metersquared +1.029*years+48.613

Systolic blood pressure(sbp) model for smokers:

systolic blood pressure(sbp)^(estimated)= 0.674*kg/metersquared +1.466*years+48.075

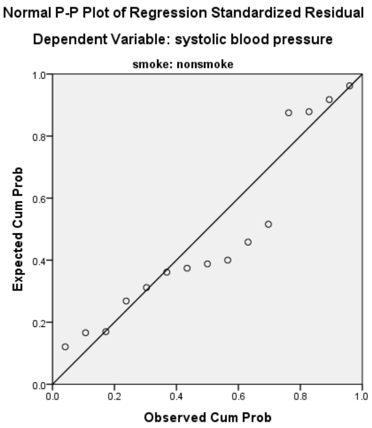
Coefficient of BMI is higher for nonsmoker (1.045) than smokers (0.674). Which mean every one value increment of BMI, sbp will increase more (0.371) for nonsmoker than smoker for fixed years. Coefficient of year is higher for smokers (1.466) than nonsmokers (1.029). Which mean every one-year increment of age, sbp will increase more (0.437) for smoker than nonsmoker for fixed BMI.

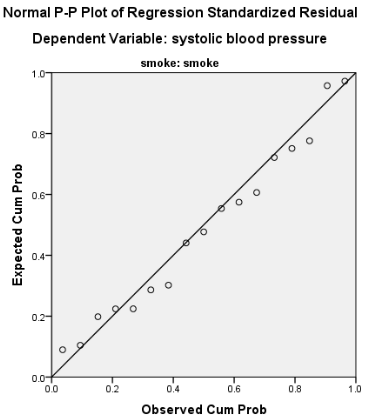
Nonsmokers: systolic blood pressure(sbp) will increase by 1.045 for every one value increase of kg/metersquared for fixed age.
Nonsmokers: systolic blood pressure(sbp) will increase by 1.029 for every one-year increase of age for fixed kg/metersquared.

Smokers: systolic blood pressure(sbp) will increase by 0.674 for every one value increase of kg/metersquared for fixed age.
Smokers: systolic blood pressure(sbp) will increase by 1.466 for every one-year increase of age for fixed kg/metersquared.

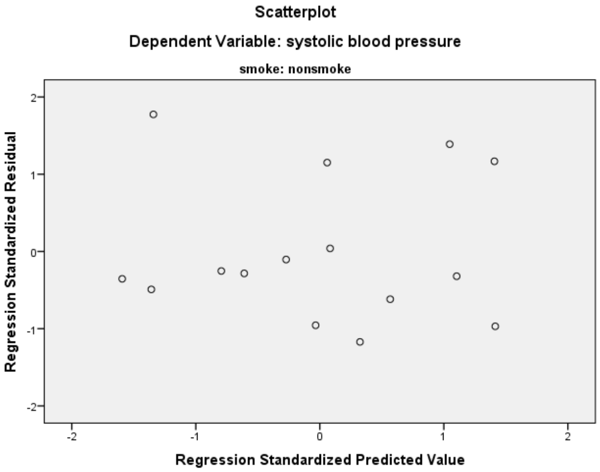
	Nonsmokers	Smokers
variables	kg/metersquared, years	kg/metersquared, years
R^2	0.836	0.696
S^2	31.941	80.503
F/sig	30.479/0.000	15.996/0.000
VIF	3.532/3.532	2.894/2.894
SSE	383.297	1127.041
t/sig for ßi	2.765/0.017, 1.544/0.149	2.460/0.027, 1.004/0.333
Residual analysis	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant

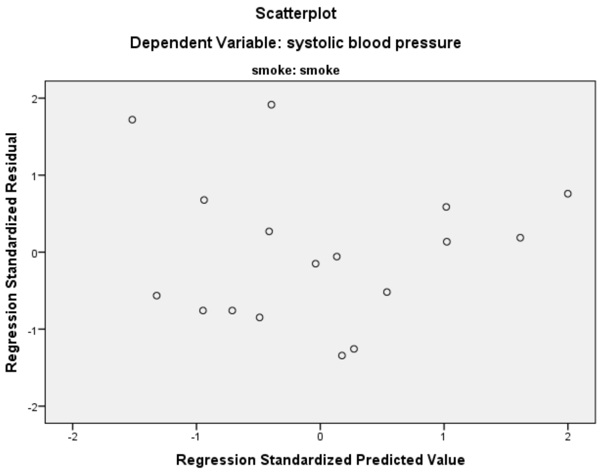
Regression model of nonsmokers has higher R^2 (0.836, 0.696) and lower S^2 (31.941, 80.503) compared with model for smokers. As well as F for nonsmokers is higher compare with smokers. VIF is higher and residual is lower of model for nonsmokers. t values, one (years) is border line significant other (kg/metersquared) is not for nonsmokers and smokers.
After comparison we can see model of sbp for nonsmokers has higher accuracy than model for smokers.





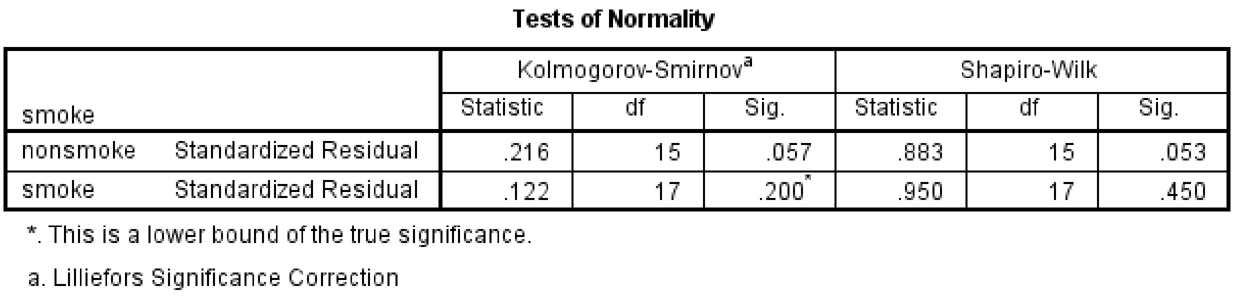
These normal p-p plots show that most points are fitting onto the line, which means the residuals are normally distributed for nonsmokers and smokers group.





The residual plots show that all points are within the band -2 to +2. As well as residuals don’t shows a pattern. Which means it is random. Therefore, variances of standardized residuals are homogeneous.

Test of normality for standardized residual:



Tests of Normality

smoke		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
nonsmoke	Standardized Residual	.216	15	.057	.883	15	.053
smoke	Standardized Residual	.122	17	.200*	.950	17	.450

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

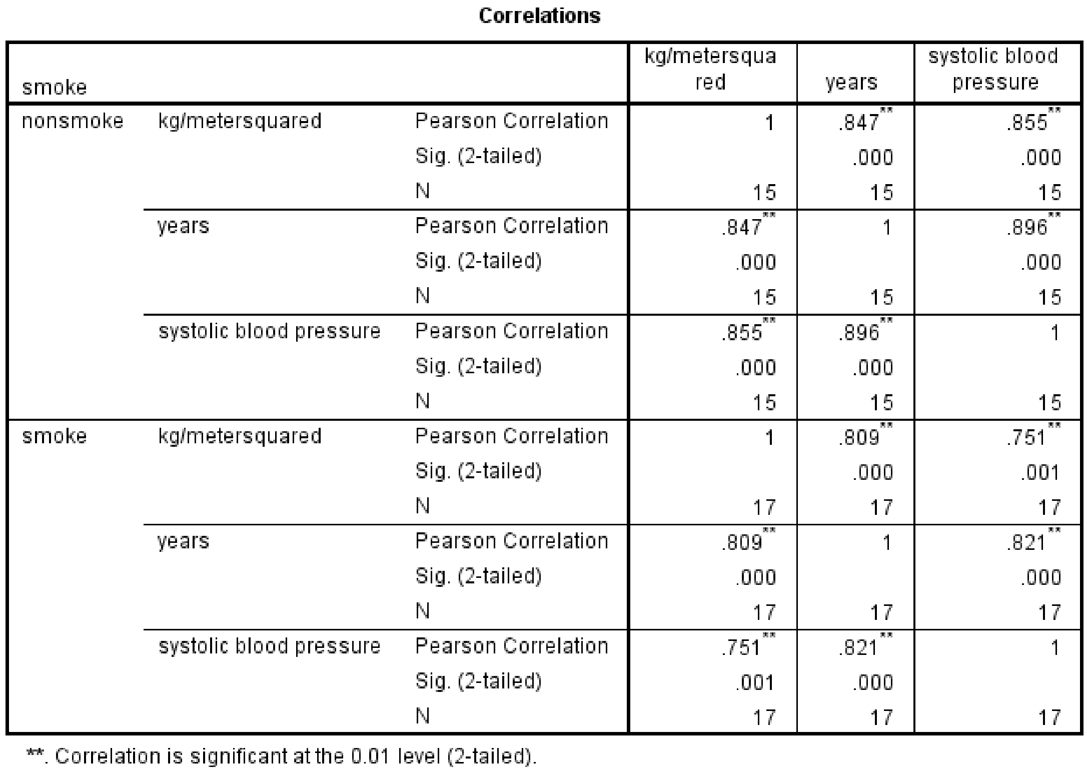
KS test statistics for nonsmokers: 0.216 and p-value = 0.057

KS test statistics for smokers: 0.122 and p-value=0.200

KS Test statistics are not significant. Standardized residuals are normal for nonsmoker and smokers.

5)Is AGE a confounder in the association of BMI and SBP for smokers and non-smokers? Explain. Use σ = 0.05

Which variable is a better predictor of SBP, AGE or BMI? Why? Confirm your answer



Correlations

smoke			kg/metersquared	years	systolic blood pressure
nonsmoke	kg/metersquared	Pearson Correlation	1	.847**	.855**
		Sig. (2-tailed)		.000	.000
		N	15	15	15
	years	Pearson Correlation	.847**	1	.896**
		Sig. (2-tailed)	.000		.000
		N	15	15	15
	systolic blood pressure	Pearson Correlation	.855**	.896**	1
		Sig. (2-tailed)	.000	.000	
		N	15	15	15
smoke	kg/metersquared	Pearson Correlation	1	.809**	.751**
		Sig. (2-tailed)		.000	.001
		N	17	17	17
	years	Pearson Correlation	.809**	1	.821**
		Sig. (2-tailed)	.000		.000
		N	17	17	17
	systolic blood pressure	Pearson Correlation	.751**	.821**	1
		Sig. (2-tailed)	.001	.000	
		N	17	17	17

**. Correlation is significant at the 0.01 level (2-tailed).

Nonsmoke:

Correlation:

Kg/metersquared, systolic blood pressure: 0.855 and p=value 0.000

Years, systolic blood pressure: 0.896 and p-value= 0.000

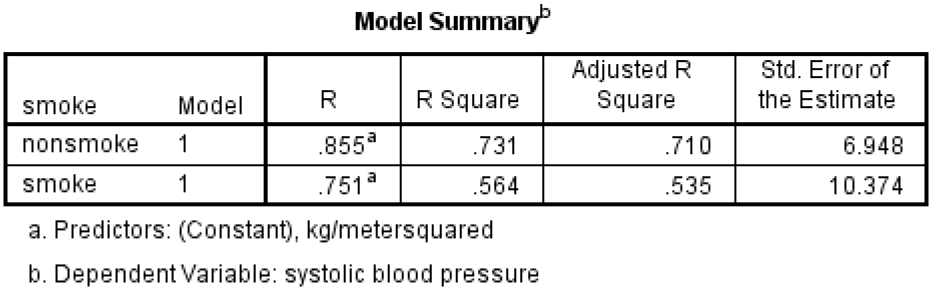
Kg/metersquared, years: 0.847 and p-value=0.000

smoke:

Correlation:
Kg/metersquared, systolic blood pressure: 0.751 and p=value 0.001
Years, systolic blood pressure: 0.821 and p-value= 0.000
Kg/metersquared, years: 0.809 and p-value=0.000

Correlation among each of the variable is high and significant. And highest correlation is in between Years and systolic blood pressure: 0.896 and p-value= 0.000. Therefore, there is a chance Age to be confounder.

KG/metersquared (without age):

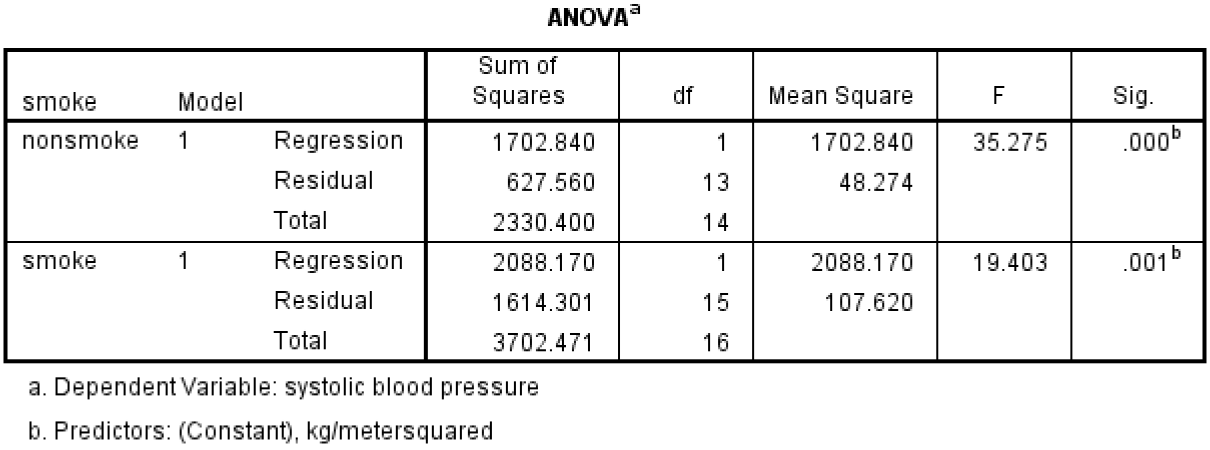


Model Summary[b]

smoke	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
nonsmoke	1	.855[a]	.731	.710	6.948
smoke	1	.751[a]	.564	.535	10.374

a. Predictors: (Constant), kg/metersquared

b. Dependent Variable: systolic blood pressure

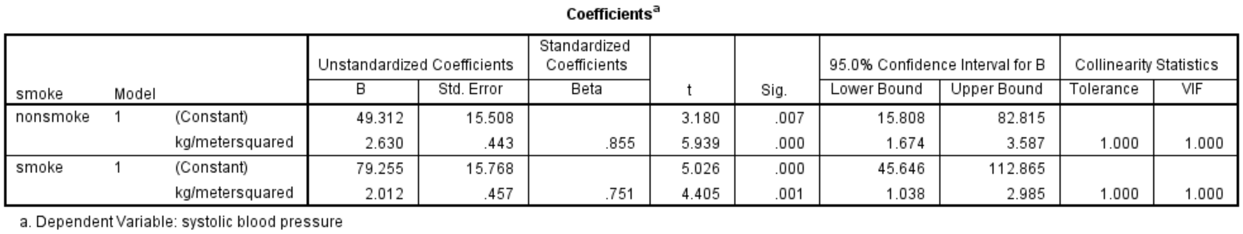


ANOVA[a]

smoke	Model		Sum of Squares	df	Mean Square	F	Sig.
nonsmoke	1	Regression	1702.840	1	1702.840	35.275	.000[b]
		Residual	627.560	13	48.274		
		Total	2330.400	14			
smoke	1	Regression	2088.170	1	2088.170	19.403	.001[b]
		Residual	1614.301	15	107.620		
		Total	3702.471	16			

a. Dependent Variable: systolic blood pressure

b. Predictors: (Constant), kg/metersquared



Coefficients[a]

			Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
smoke	Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
nonsmoke	1	(Constant)	49.312	15.508		3.180	.007	15.808	82.815		
		kg/metersquared	2.630	.443	.855	5.939	.000	1.674	3.587	1.000	1.000
smoke	1	(Constant)	79.255	15.768		5.026	.000	45.646	112.865		
		kg/metersquared	2.012	.457	.751	4.405	.001	1.038	2.985	1.000	1.000

a. Dependent Variable: systolic blood pressure

Systolic blood pressure(sbp) model for nonsmokers:

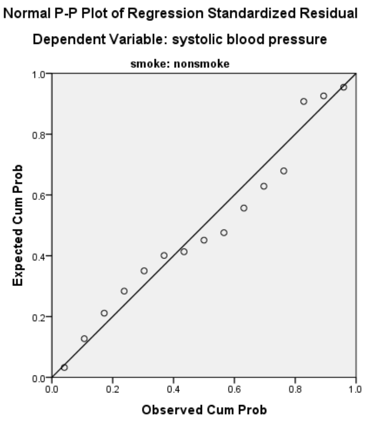
systolic blood pressure(sbp)^(estimated)= 2.630*kg/metersquared +49.312

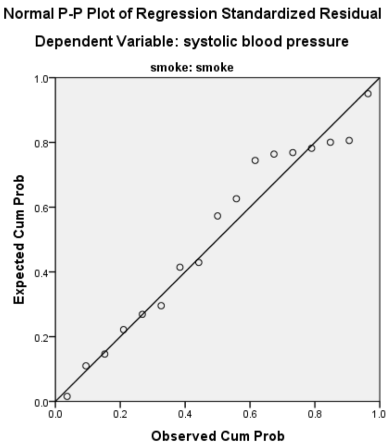
Systolic blood pressure(sbp) model for smokers:

systolic blood pressure(sbp)^(estimated)= 2.012*kg/metersquared +79.255

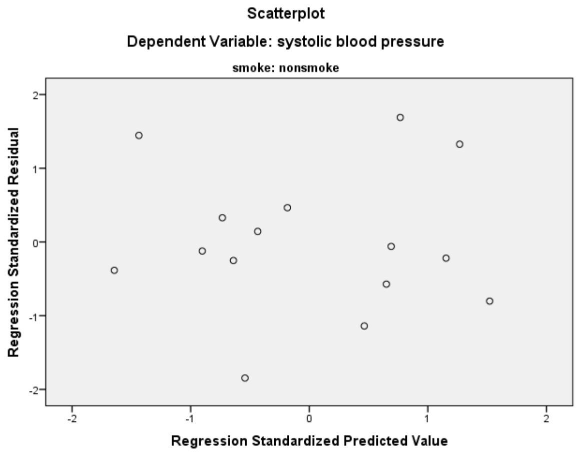
variable	nonsmoke	smoke
kg/metersquared years	systolic blood pressure(sbp)^(estimated)= 1.045*kg/metersquared +1.029*years+48.613	systolic blood pressure(sbp)^(estimated)= 0.674*kg/metersquared +1.466*years+48.075
kg/metersquared	systolic blood pressure(sbp)^(estimated)= 2.630*kg/metersquared +49.312	systolic blood pressure(sbp)^(estimated)= 2.012*kg/metersquared +79.255
coefficient increment for kg/metersquared when age dropped	1.585	1.338
Coefficient increment percentage	100*(2.63-1.045)/1.045 =151.67% (coefficient increment of kg/metersquared)>10%	100*(2.012-0.674)/0.674= 198.51% (coefficient increment of kg/metersquared)>10%

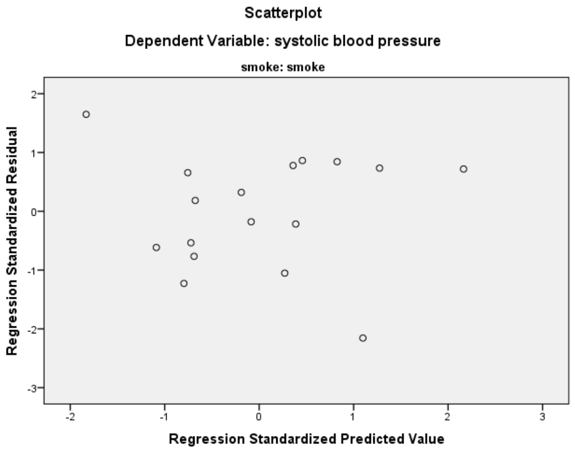
Value of the coefficient of 'kg/metersquared' is pretty high when age dropped from the model compared with when age present in the model. Therefore, when the age is present in the model effect of BMI on SBP is lower than when age dropped from the model. Which means AGE is a confounder in the association of BMI and SBP for smokers and non-smokers.





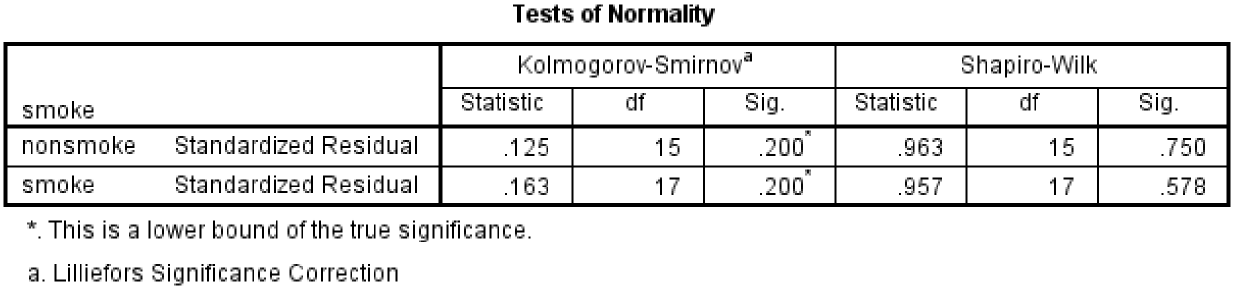
These normal p-p plots show that most points are fitting onto the line, which means the residuals are normally distributed for nonsmokers and smokers group.





The residual plots show that all points are within the band -2 to +2. As well as residuals don’t shows a pattern. Which means it is random. Therefore, variances of standardized residuals are homogeneous.

Test of normality for residuals:



Tests of Normality

smoke		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
nonsmoke	Standardized Residual	.125	15	.200*	.963	15	.750
smoke	Standardized Residual	.163	17	.200*	.957	17	.578

*. This is a lower bound of the true significance.

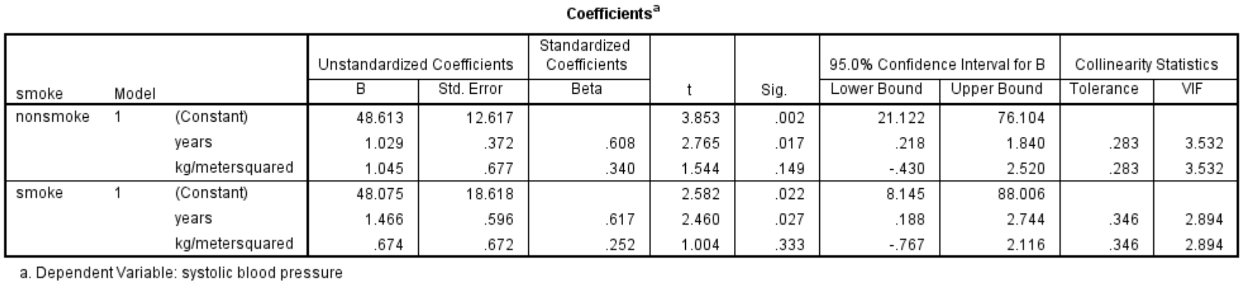
a. Lilliefors Significance Correction

KS test statistics for nonsmokers: 0.125 and p-value = 0.200

KS test statistics for smokers: 0.163 and p-value=0.200

KS Test statistics are not significant. Standardized residuals are normal for nonsmoker and smokers.

Which variable is a better predictor of SBP, AGE or BMI? Why? Confirm your answer



Coefficients[a]

smoke	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
			B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
nonsmoke	1	(Constant)	48.613	12.617		3.853	.002	21.122	76.104		
		years	1.029	.372	.608	2.765	.017	.218	1.840	.283	3.532
		kg/metersquared	1.045	.677	.340	1.544	.149	-.430	2.520	.283	3.532
smoke	1	(Constant)	48.075	18.618		2.582	.022	8.145	88.006		
		years	1.466	.596	.617	2.460	.027	.188	2.744	.346	2.894
		kg/metersquared	.674	.672	.252	1.004	.333	-.767	2.116	.346	2.894

a. Dependent Variable: systolic blood pressure

Nonsmoke:

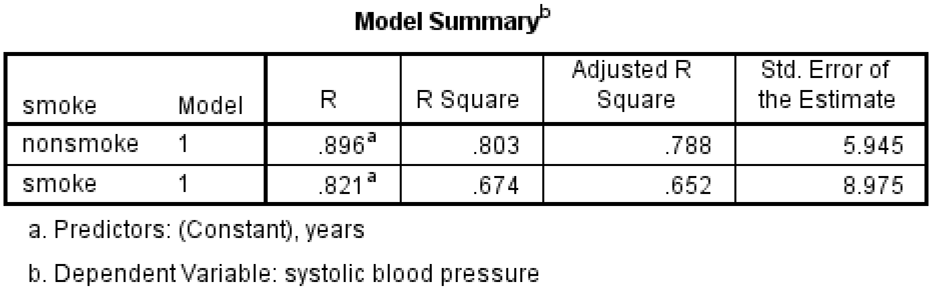
Years: t value = 2.765 p-value = 0.017
kg/metersquared: t value = 1.544 p-value = 0.149
α=0.05
years has higher t-value and lower p-value compare with kg/metersquared. Which means 'years' variable has higher significant association with sbp than kg/metersquared. Therefore, years is the better predictor for SBP.

smoke:

Years: t value = 2.46 p-value = 0.027
kg/metersquared: t value = 1.004 p-value = 0.333
α=0.05
Years has higher t-value and lower p-value compare with kg/metersquared. Which means 'years' variable has higher significant association with sbp than kg/metersquared. Therefore, years is the better predictor for SBP.

Variable 'age' is a better predictor of 'SBP'

Model with only age

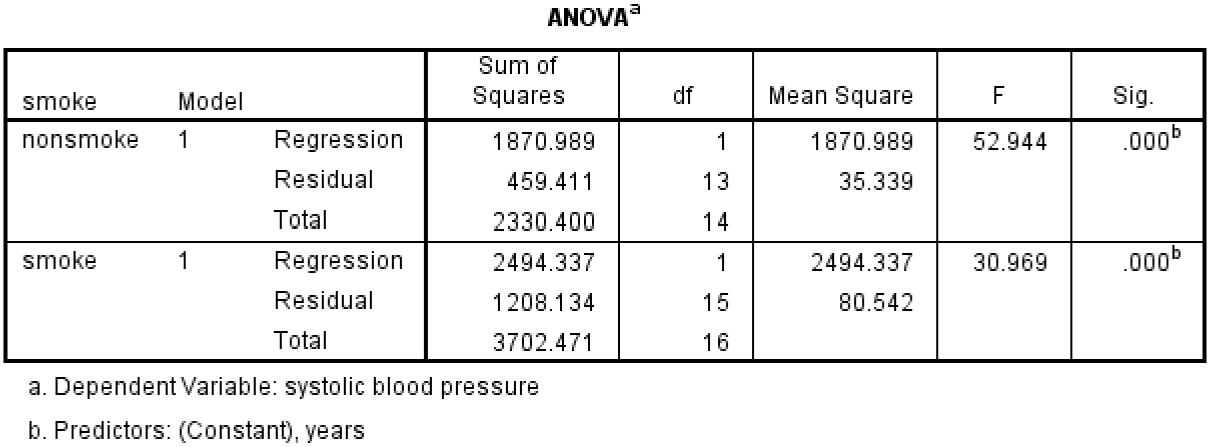


Model Summary[b]

smoke	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
nonsmoke	1	.896[a]	.803	.788	5.945
smoke	1	.821[a]	.674	.652	8.975

a. Predictors: (Constant), years

b. Dependent Variable: systolic blood pressure

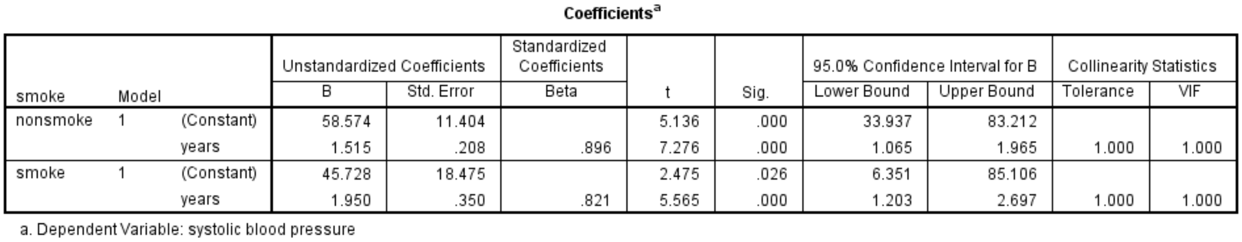


ANOVA[a]

smoke	Model		Sum of Squares	df	Mean Square	F	Sig.
nonsmoke	1	Regression	1870.989	1	1870.989	52.944	.000[b]
		Residual	459.411	13	35.339		
		Total	2330.400	14			
smoke	1	Regression	2494.337	1	2494.337	30.969	.000[b]
		Residual	1208.134	15	80.542		
		Total	3702.471	16			

a. Dependent Variable: systolic blood pressure

b. Predictors: (Constant), years



Coefficients[a]

			Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
smoke	Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
nonsmoke	1	(Constant)	58.574	11.404		5.136	.000	33.937	83.212		
		years	1.515	.208	.896	7.276	.000	1.065	1.965	1.000	1.000
smoke	1	(Constant)	45.728	18.475		2.475	.026	6.351	85.106		
		years	1.950	.350	.821	5.565	.000	1.203	2.697	1.000	1.000

a. Dependent Variable: systolic blood pressure

Nonsmoke:

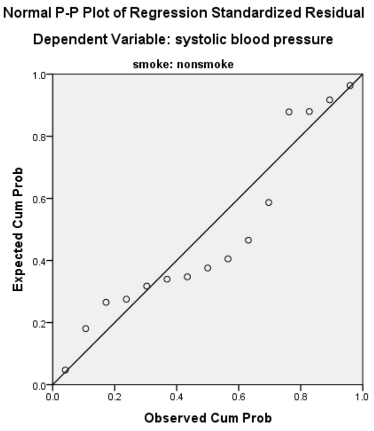
systolic blood pressure(sbp)^(estimated)= 1.515*years +58.574

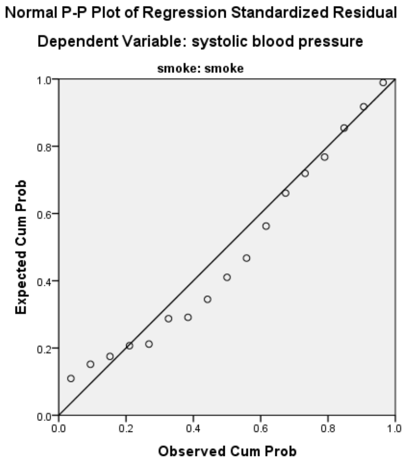
Smoke:

systolic blood pressure(sbp)^(estimated)= 1.950*years +45.728

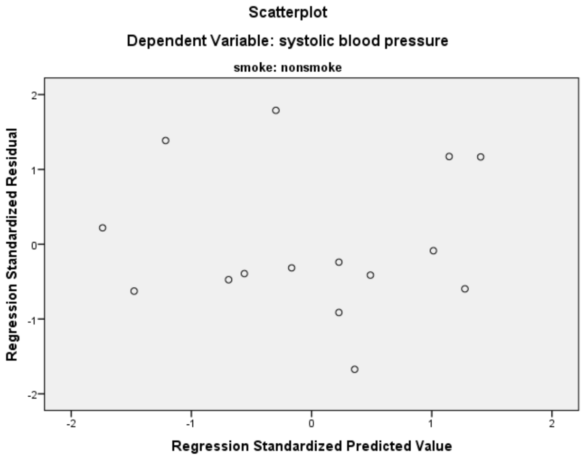
	Nonsmoke	smoke	Nonsmoke	Smokers
variable	kg/metersquared		years	
R^2	0.732	0.564	0.803	0.674
S^2	48.274	107.620	35.339	80.542
F/sig	35.275/0.00	19.403/0.001	52.944/0.000	30.969/0.000
t/sig for ßi	5.939/0.00	4.405/0.001	7.276/0.00	5.565/0.00
Residual analysis	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant	Residuals are normal (pp plot), Homogeneous (scatter plot for residual), KS statistics not significant

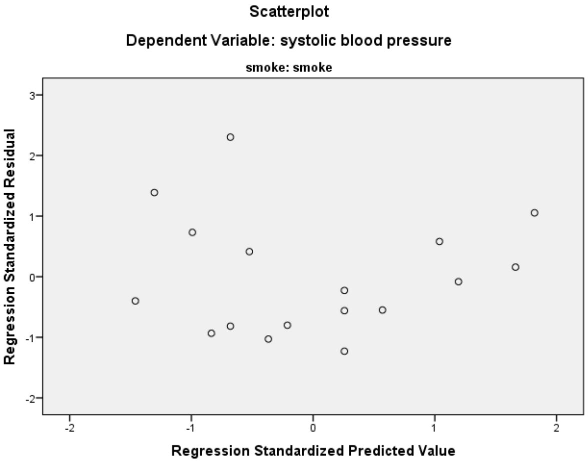
Here we can see that model with years has higher R^2 lower S^2 and good significant values coefficient. F value also high for model with years than model with kg/metersquared. Model with years is better than model with 'kg/metersquared'





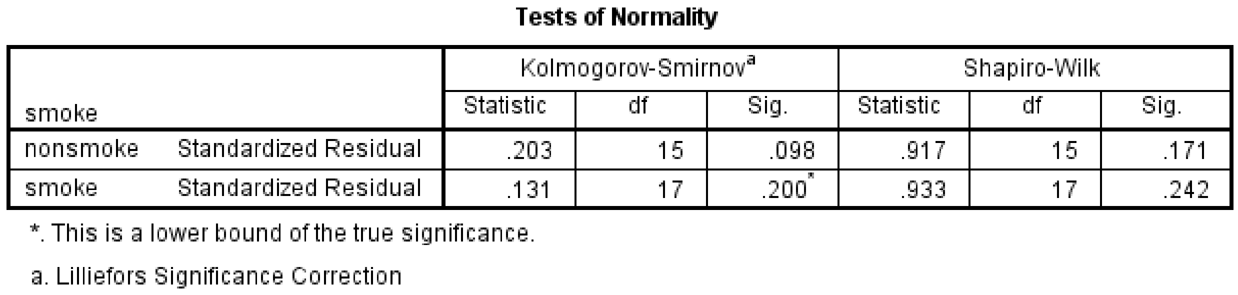
These normal p-p plots show that most points are fitting onto the line, which means the residuals are normally distributed for nonsmokers and smokers group.





The residual plots show that all points are within the band -2 to +2. As well as residuals don’t shows a pattern. Which means it is random. Therefore, variances of standardized residuals are homogeneous.

Normality test



Tests of Normality

smoke		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
nonsmoke	Standardized Residual	.203	15	.098	.917	15	.171
smoke	Standardized Residual	.131	17	.200*	.933	17	.242

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

KS test statistics for nonsmokers: 0.203 and p-value = 0.098

KS test statistics for smokers: 0.131 and p-value=0.200

KS test statistics are not significant. Residual are normal for nonsmoker and smokers.