PROJECT 5

Kankanamge Harsha

A. The dataset **vehicle mileage.sav**, presents the gasoline mileage (mpg) and weight (pounds) of 121 vehicles classified as car or non-car (includes sport utility vehicles, trucks and minivans). Use regression analysis to determine the effect on gasoline mileage of the weight of a vehicle and the type of vehicle.

1. Determine the proportion of vehicles in the two categories.

description

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	car	82	67.8	67.8	67.8
	non-car	39	32.2	32.2	100.0
	Total	121	100.0	100.0	

Proportion of the car is 0.678 and proportion of the non-car is 0.322.

2. Compute the new interaction variable (xz), where x is weight and z is car type.

WEIGHT	vehicle	MPG	cartype	XZ
2635	car	31	0	.00
3670	car	20	0	.00
3460	car	22	0	.00
3345	car	22	0	.00
3785	car	20	0	.00
3265	car	24	0	.00
3585	car	20	0	.00
2960	car	24	0	.00
3350	car	22	0	.00
3450	car	20	0	.00
3880	car	21	0	.00
3325	car	21	0	.00
3805	car	20	0	.00
4020	car	20	0	.00
4520	non-car	15	1	4520.00
4225	non-car	15	1	4225.00
2795	car	26	0	.00
3295	car	20	0	.00
3350	car	22	0	.00

3. Define a single multiple regression model that uses the data for both cars and non-cars and that defines straight line models for each group with possible differing intercepts and slopes. Use an interaction term.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.944[a]	.891	.888	1.386

a. Predictors: (Constant), XZ, X, Z

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1842.106	3	614.035	319.508	.000[b]
	Residual	224.852	117	1.922		
	Total	2066.959	120			

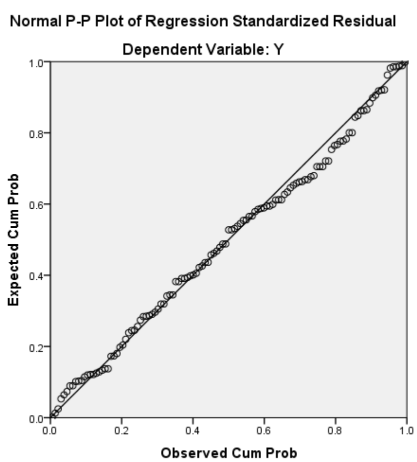
a. Dependent Variable: Y

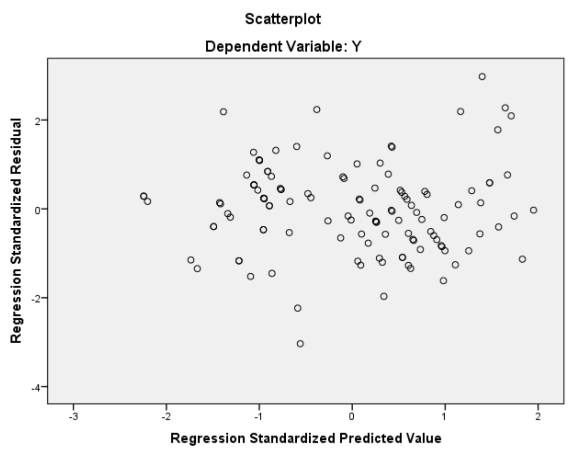
b. Predictors: (Constant), XZ, X, Z

Coefficients[a]

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
1	(Constant)	43.423	1.157		37.546	.000	41.133	45.714		
	X	-.006	.000	-.964	-17.369	.000	-.007	-.006	.302	3.314
	Z	-11.521	2.204	-1.303	-5.226	.000	-15.886	-7.155	.015	66.824
	XZ	.003	.001	1.258	4.609	.000	.002	.004	.012	80.126

a. Dependent Variable: Y





Most of the data points are on or very closer to the line. We can assume residual are normal. Residuals are randomly distributed (-2, +2) around the zero line. Therefore, variance of residuals is homogeneous.

Y=MPG

X= weight

Z= car type 0 → car

1 → non-car

Multiple regression model

Y=ß0+ß1weight +ß2Z +ß3XZ+Ɛ

X=weight

Z=Car type

Y^(estimated)=43.423-0.006X-11.521Z+0.003XZ

Straight line, model for the car (Z=0)

Y^(estimated)=43.423-0.006weight

Straight line, model for the non-car (Z=1)

Y^(estimated)= 43.423-0.006weight -11.521*1+0.003weight

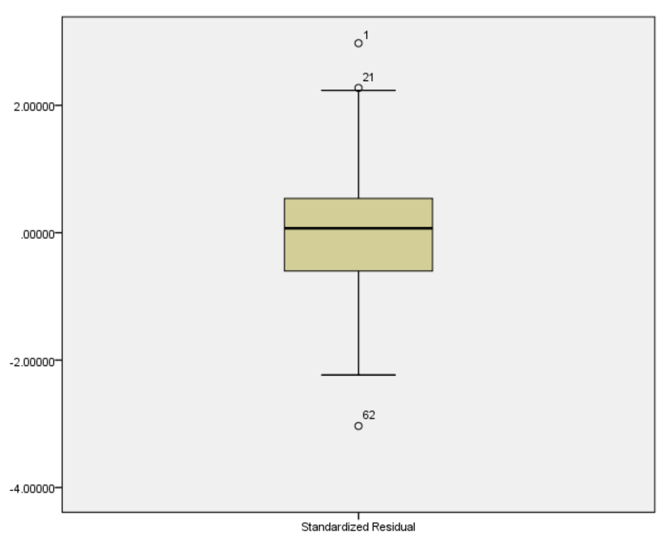
Y^(estimated)=31.902-0.003weight

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.062	121	.200*	.987	121	.290

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

a. Lilliefors Significance Correction



We can see there are outlier showing on standardized residual plot.

KS statistic is very low 0.062 and sig 0.200. Residuals are normal

Model without cartype

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.926[a]	.857	.856	1.577

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1771.176	1	1771.176	712.584	.000[b]
	Residual	295.783	119	2.486		
	Total	2066.959	120			

a. Dependent Variable: Y

b. Predictors: (Constant), X

Coefficients[a]

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
1	(Constant)	42.387	.799		53.031	.000	40.805	43.970		
	X	-.006	.000	-.926	-26.694	.000	-.006	-.006	1.000	1.000

a. Dependent Variable: Y

Y^(estimated)=42.387-0.006weight

But this model has lower R^2 (0.857, 0.891) higher S^2 (2.486,1.922). Therefore, Model with cartype is good

Best model is previous model, with Z1 and XZ

4. Define the intercept and slope for each straight line model in terms of the regression coefficients of the single regression model. Graph both fitted lines on the same chart.

MPG= mile per gallon

Straight line, model for the car (Z=0)

Y^(estimated)=43.423-0.006weight

Intercept of model for cars is 43.423

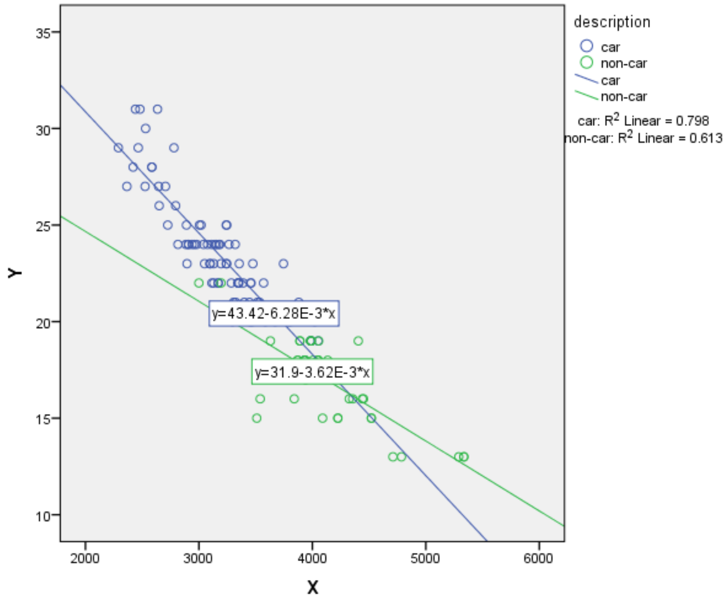
b1=-0.006(slope) mean that for every 1-pound increment of weight, average of MPG(Y) will be decreased by 0.006 for car.

Straight line, model for the non-car (Z=1)

Y^(estimated)=31.902-0.003weight

Intercept of model for non-car is 31.902

b1=-0.003(slope) mean that for every 1-pound increment of weight, average of MPG(Y) will be decreased by 0.003 for non-car



We can see that most of the data represent the car. Non-car has higher weight and low mpg. Car has low weight compare with non-car and higher number of miles per gasoline gallon.

5. For each of the three tests below, state the appropriate null hypothesis in terms of the regression coefficients of the model and use $\alpha = 0.05$.

a) Test for parallelism.

ß1=coefficient of weight

H0: ß1=0

H1: ß1≠0

Test statistic t =-17.369

p-value=0.000<0.05

Reject H0

Weight provide sufficient contribution to explain the variation of MPG of vehicle when cartype held constant.

b) Test for equal intercepts.

ß2=coefficient of cartype

H0: ß2=0

H1: ß2≠0

Test statistic t =-5.226

p-value=0.000<0.05

Reject H0

The distinction of the MPG between car and non-car is statically significant when weight held constant.

c) Test for coincidence.

ß3=coefficient of XZ

H0: ß3=0

H1: ß3≠0

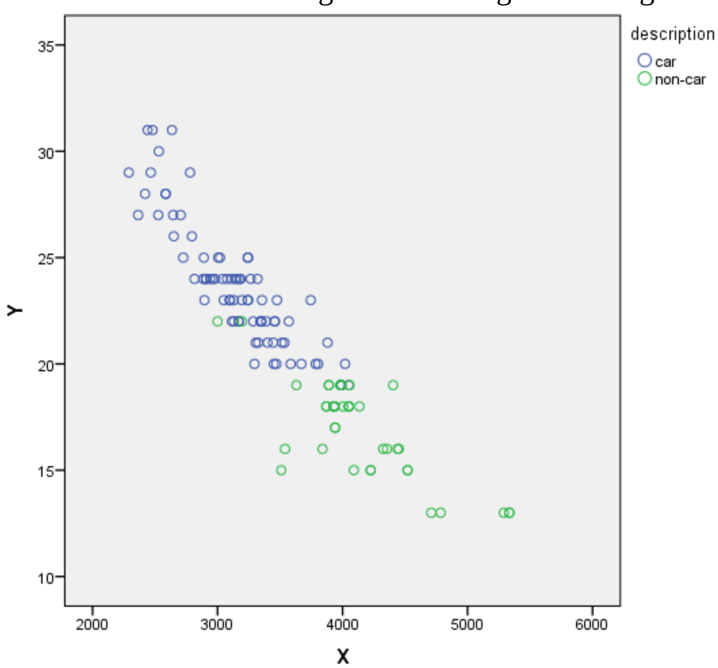
Test statistic t =4.609

p-value=0.000<0.05

Reject H0

There is a significant interaction between X and Z

6. Discuss the variation in gasoline mileage with weight between cars and non-cars.



We can see that most of the data represent the car. Non-car has higher weight and low mpg. Car has low weight compare with non-car and higher number of miles per gasoline gallon.

Every 1-pound increment of weight of car, MPG(Y) will be decreased by 0.006(mile per gallon)

Every 1-pound increment of weight of non-car, MPG(Y) will be decreased by 0.003(mile per gallon)

For given weight (X) in pound, MPG differences between car and non-car (MPG of car-noncar) is 11.521-0.003X. So car has 11.521-0.003X higher MPG than non-car if given weight(X)< 3840.333(11.521/0.003) pound. But if given weight > 3840.333 (11.521/0.003) pound non-car has higher MPG. If weight = 3840.333 (11.521/0.003) pound both car and non-car have same MPG.

B. In a certain locality, five residential houses that were sold recently were selected at random from each of three distinct neighborhoods (A, B, and C) in the city, and the selling price Y was compared to the property valuation X as determined by the local real estate assessor's office. In the data set **price_value_nbhd.sav**, selling price and property valuation are in thousands of dollars.

1. Estimate an appropriate regression model to these data assuming no interaction. How have the dummy variables been defined?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974[a]	.950	.936	10.0208

a. Predictors: (Constant), Z2, X, Z1

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20823.698	3	6941.233	69.125	.000[b]
	Residual	1104.579	11	100.416		
	Total	21928.277	14			

a. Dependent Variable: Y

b. Predictors: (Constant), Z2, X, Z1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	38.696	106.226		.364	.723	-195.106	272.499		
	X	.954	.324	.649	2.945	.013	.241	1.667	.094	10.596
	Z1	-31.303	20.554	-.386	-1.523	.156	-76.543	13.937	.071	14.024
	Z2	-3.337	12.958	-.041	-.258	.802	-31.858	25.184	.179	5.574

a. Dependent Variable: Y

Y= selling price

X= property valuation

Categorical variable has 3 groups neighborhood A B and C on define 2 dummy variable

Z1=1 for neighborhood A else 0

Z2=1 for neighborhood B else 0

Z1=0=Z2 for neighborhood C

	Z1	Z2
Neighborhood A	1	0
Neighborhood B	0	1
Neighborhood C	0	0

Regression model

Y^(selling price)= 0.954*property_valuation-31.303*Z1-3.337*Z2+38.696

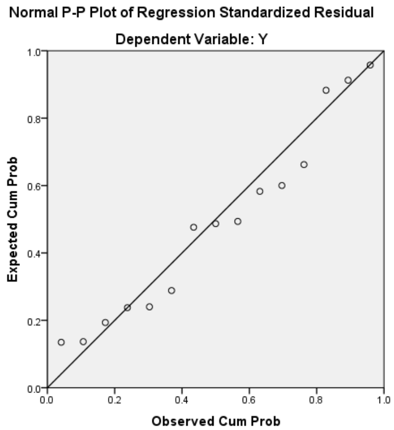
2. Interpret b_2 and b_3, the estimated regression coefficients of the two dummy variables.

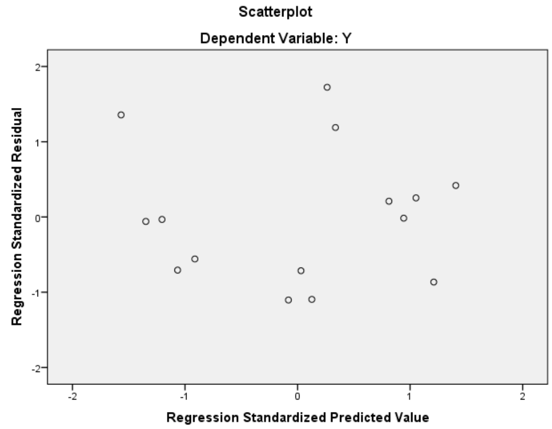
b2=-31.303 coefficient of Z1

Average selling price of the house in A neighborhood is $31303 less than the property in neighborhood C when property valuation held constant.

b3=-3.337 coefficient of Z2

Average selling price of the house in B neighborhood is $3337 less than the property in neighborhood C when property valuation held constant.





Most of the data points are on or very closer to the line. We can assume residual are normal. Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

3. Test whether the regression explained by the model is significant at the 0.05 level of significance.

H0: ß1=ß2= ß3=0

H1: At least one of the line model coefficient is non zero

Test statistic F = 69.125

p-value=0.000

p-value<0.05 Reject H0

property valuation and two dummy variables (Z1, Z2) appear to be explained the variation of selling price of house.

4. Test the hypothesis that α's = 0 at the 0.05 level of significance against the alternative α≠ 0 and interpret their significance.

ß1=coefficient of property valuation

H0: ß1=0

H1: ß1≠0

Test statistic t =2.945

p-value=0.013<0.05

Reject H0

Property valuation provide sufficient significant contribution to explain the variation of selling price of house when neighborhood held constant.

ß2=coefficient of Z1

H0: ß2=0

H1: ß2≠0

Test statistic t =-1.523

p-value=0.156>0.05

Do no reject H0

Distinction of Selling price of house between neighborhood A and C is not significant for fixed property valuation.

ß3=coefficient of Z3

H0: ß3=0

H1: ß3≠0

Test statistic t =-0.258

p-value=0.802>0.05

Do no reject H0

Distinction of Selling price of house between neighborhood B and C is not significant for fixed property valuation.

5. Evaluate the above model and revise it as necessary. Estimate the new regression model. Interpret the estimated regression coefficient of the new dummy variable. Compare the two models.

ß2 and ß3 are not significant but ß3 has higher p-value so we are going to make neighborhood B and C as one group.

Z3=1 for neighborhood A, 0 for neighborhood B and C (Now we consider B and C as one neighborhood)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974[a]	.949	.941	9.6231

a. Predictors: (Constant), Z3, X

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20817.039	2	10408.519	112.399	.000[b]
	Residual	1111.239	12	92.603		
	Total	21928.277	14			

a. Dependent Variable: Y

b. Predictors: (Constant), Z3, X

Coefficients[a]

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
1	(Constant)	14.458	47.291		.306	.765	-88.580	117.495		
	X	1.027	.152	.698	6.748	.000	.695	1.359	.395	2.535
	Z3	-26.512	8.391	-.327	-3.159	.008	-44.794	-8.229	.395	2.535

a. Dependent Variable: Y

Model with one dummy variable:

Z3=1 for neighborhood A, 0 for neighborhood B and C (Now we consider B and C as one neighborhood)

Y^(selling price)= 1.027*property_valuation-26.512*Z3+14.458

Average selling price of the house in A neighborhood is $26512 less than the house in neighborhood B and C when property valuation held constant.

H0: ß1=ß2=0

H1: At least one of the line model coefficient is non zero

Test statistic F = 112.399

p-value=0.000

p-value<0.05 Reject H0

property valuation and dummy variables (Z3) provide significant contribution to explained the variation of selling price of house.

ß1=coefficient of property valuation

H0: ß1=0

H1: ß1≠0

Test statistic t =6.748

p-value=0.000<0.05

Reject H0

Property valuation provide significant contribution to explain the variation of selling price of house when neighborhood held constant.

ß2=coefficient of Z3

H0: ß2=0

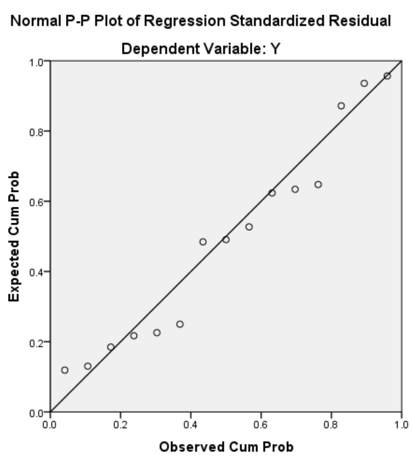
H1: ß2≠0

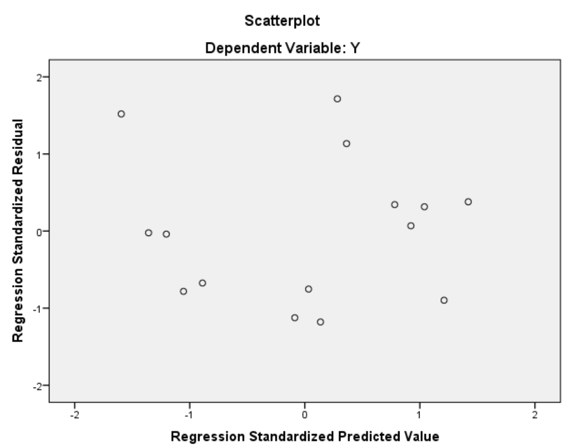
Test statistic t =-3.159

p-value=0.008<0.05

Reject H0

Distinction of Selling price of house between neighborhood A and BC-group is significant for fixed property valuation.





Most of the data points are on or very closer to the line. We can assume residual are normal. Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

Checking for model without dummy variables.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.952[a]	.907	.900	12.5135

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19892.639	1	19892.639	127.038	.000[b]
	Residual	2035.638	13	156.588		
	Total	21928.277	14			

a. Dependent Variable: Y

b. Predictors: (Constant), X

Coefficients[a]

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
1	(Constant)	-105.048	36.911		-2.846	.014	-184.790	-25.306		
	X	1.401	.124	.952	11.271	.000	1.133	1.670	1.000	1.000

a. Dependent Variable: Y

model without dummy variables:

Y^(selling price)= 1.401*property_valuation-105.048

H0: ß1=0

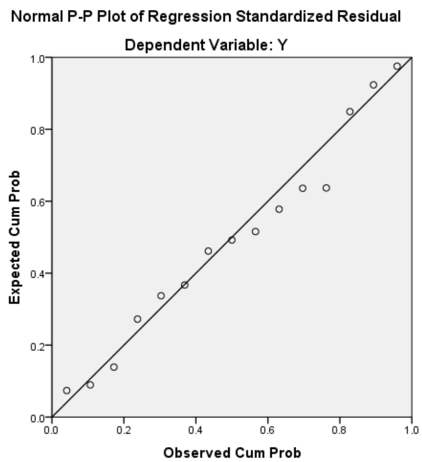
H1: ß1≠0

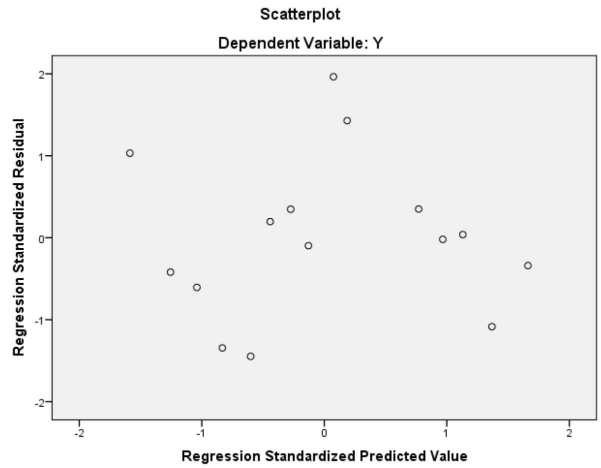
Test statistic F = 127.038

p-value=0.000

p-value<0.05 Reject H0

Property valuation provide sufficient significant contribution to explain the variation of selling price of house





Most of the data points are on or very closer to the line. We can assume residual are normal. Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

	Y^(selling price)= 0.954*property_valuation-31.303*Z1-3.337*Z2+38.696	**Y^(selling price)= 1.027*property_valuation-26.512*Z3+14.458**	**Y^(selling price)= 1.401*property_valuation-105.048**
	Model 1	**Model 2**	**Model 3**
variable	**X,Z1,Z2**	**X,Z3**	**X**
R^2	**0.95**	**0.949**	**0.907**
S^2	**100.416**	**92.603**	**156.588**
F	**69.125,sig=0.000**	**112.399, sig=0.000**	**127.038,sig=0.000**
t/sig for ßi	**2.945/0.013, -1.523/0.156, -0.258/0.802**	**6.748/0.000, -3.159/0.008**	**11.271/0.000**
VIF	**10.596,14.024,5.574**	**2.535,2.535**	**1.00**

Z3=1 for neighborhood A, 0 for neighborhood B and C (Now we consider B and C as one neighborhood)

After done the comparison best model is: Y^(selling price)= 1.027*property_valuation-26.512*Z3+14.458

R^2 almost same (0.95,0.949) for model1 and model2 which have highest R^2 so we ignore model3.Model2 has lower S^2(92.603,100.416) and VIF values. F values is higher in model2 and all Coefficient are significant in model2. Compare with other two models, model2 is the best model.

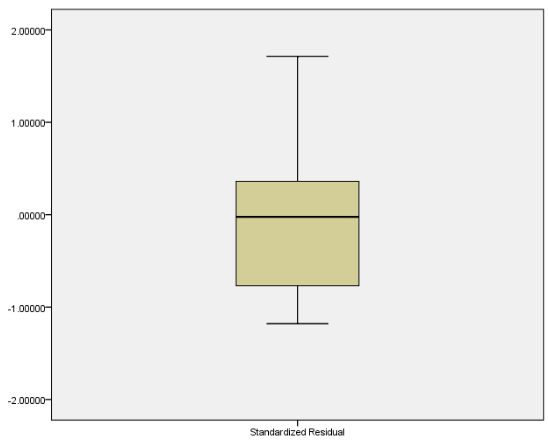
Normality test for the final model

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.167	15	.200*	.924	15	.222

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

a. Lilliefors Significance Correction



KS Test statistic is 0.167 P-value is 0.2. Residuals are normal. We can't see outlier on standardized residual plot.

6. From Step 1, write the equation of the straight line model of sale price and property valuation for each of the three neighborhoods. For a given property valuation X, which neighborhood has the highest mean sale price of houses?

<u>For neighborhood A Z1=1 Z2=0</u>

Y^(selling price)= 0.954*property_valuation-31.303*Z1-3.337*Z2+38.696

Y^(selling price)= 0.954*property_valuation-31.303*1-3.337*0+38.696

Y^(selling price)= 0.954*property_valuation+7.393

<u>For neighborhood B Z1=0 Z2=1</u>

Y^(selling price)= 0.954*property_valuation-31.303*0-3.337*1+38.696

Y^(selling price)= 0.954*property_valuation+35.359

<u>For neighborhood C Z1=0 Z2=0</u>

Y^(selling price)= 0.954*property_valuation+38.696

Neighborhood C has highest mean sale price of houses for a given property valuation.