STAT 6337
Advanced Statistical Methods I (Fall 2022)
Project 3

This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions if something is not clear.

Sign on this page below and attach with your project. You project will not be graded without it.

This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* **of UT Dallas** https://policy.utdallas.edu/utdsp5003.

YOUR NAME ______________________

DATE ______________________

YOUR SIGNATURE ______________________

`From the scatterplot matrix, X2, X4, X5, X8, X9, X10 are more linearly related to Y compared to the other predictor variables. ALso, both of the categorical variables seem linearly releated as well. So, X2, X4, X5, X6, X7, X8, X9, X10

seems like good predictors of Y .

Scatter plot is in the output section.

Lack of Fit Test:

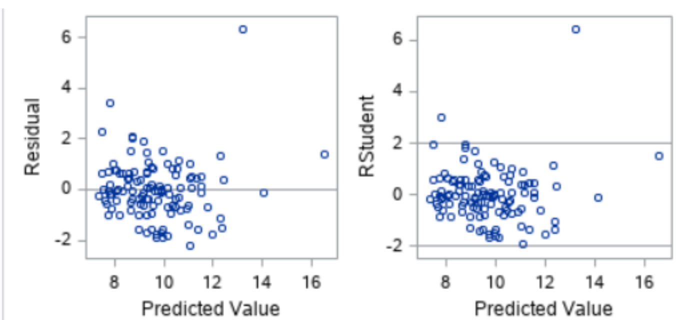
Lack of fit test cannot be performed because as can see that the degree of SSPE is 0 there are no replications in this data.

Number of Observations Used	113

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	256.67902	25.66790	17.16	<.0001
Error	102	152.53136	1.49541		
Lack of Fit	102	152.53136	1.49541	.	.
Pure Error	0	0	.		
Corrected Total	112	409.21038			

Independence of Errors:

We can see outliers in the residual plots where some points are greater than 2 .



Checking for homogeneity of variance:

The P-value > 0.05, indicates that errors follow constant variance assumption and the output of Breush pagan test shows that all variables are > .05.

Brown Forsythe Tests for homogeneity

The homogeneity of residuals hold for all the nine variables as all their p values are greater than 5.

Based on Shapiro wilks and QQ plot we can see that there is violation of normality as Shapiro wilks > .05 .

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.864616	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.104197	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.292998	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.867247	Pr > A-Sq	<0.0050

QQ plot is in the output section.

As we can see that there is violation of normality hence we apply the box cox transformation after finding box cox power.

After applying box cox transformation we can see that the data is normal.

After doing stepwise selection these variables namely risk of infection, daily census, chest xray ratio, age, number of beds, number of nurses and culturing seems to be good predictors of Y.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.97591	1.45210	0.67	0.5030	-1.90333	3.85515
RISK	1	0.28447	0.11022	2.58	0.0112	0.06592	0.50301
CENSUS	1	0.02013	0.00363	5.54	<.0001	0.01293	0.02733
XRAY	1	0.01620	0.00642	2.53	0.0130	0.00348	0.02892
AGE	1	0.09153	0.02568	3.56	0.0006	0.04061	0.14245
BEDS	1	-0.00899	0.00301	-2.98	0.0036	-0.01497	-0.00301
NURSES	1	-0.00546	0.00196	-2.79	0.0063	-0.00935	-0.00158
CULTURING	1	0.03359	0.01350	2.49	0.0144	0.00683	0.06036

I have performed box cox transformation on the length of stay . and got length of stay or Y = β0 + β1*age + β2*risk + β3*culturing + β4*chestxray +β5*beds + β8*census + β9*nurses Firstly residuals from the OLS model is computed , then I regress the residuals on the independent variables, and find the weight, as per Huber's formula [w=1.345/abs(u) for abs(u) > 1.345 and w=1 elsewhere], with u computed with MAD (Median Absolute Deviation). The details are taken care in code and I have run 3 iterations after regression of residuals on independent variables. Here are the parameter estimates for this question:-

Iteration 1:-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.91863	1.45755	0.63	0.5299	-1.97142	3.80867
RISK	1	0.28868	0.11066	2.61	0.0104	0.06926	0.50810
CENSUS	1	0.02021	0.00365	5.54	<.0001	0.01298	0.02745
XRAY	1	0.01640	0.00643	2.55	0.0122	0.00365	0.02916
AGE	1	0.09233	0.02577	3.58	0.0005	0.04123	0.14343
BEDS	1	-0.00897	0.00302	-2.96	0.0038	-0.01496	-0.00297
NURSES	1	-0.00560	0.00197	-2.84	0.0054	-0.00951	-0.00169
CULTURING	1	0.03297	0.01356	2.43	0.0167	0.00609	0.05986

Iteration 2:-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.97591	1.45210	0.67	0.5030	-1.90333	3.85515
RISK	1	0.28447	0.11022	2.58	0.0112	0.06592	0.50301
CENSUS	1	0.02013	0.00363	5.54	<.0001	0.01293	0.02733
XRAY	1	0.01620	0.00642	2.53	0.0130	0.00348	0.02892
AGE	1	0.09153	0.02568	3.56	0.0006	0.04061	0.14245
BEDS	1	-0.00899	0.00301	-2.98	0.0036	-0.01497	-0.00301
NURSES	1	-0.00546	0.00196	-2.79	0.0063	-0.00935	-0.00158
CULTURING	1	0.03359	0.01350	2.49	0.0144	0.00683	0.06036

Iteration 3:-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.98673	1.44746	0.68	0.4969	-1.88332	3.85678
RISK	1	0.28342	0.10972	2.58	0.0112	0.06587	0.50098
CENSUS	1	0.02013	0.00362	5.57	<.0001	0.01296	0.02730
XRAY	1	0.01623	0.00639	2.54	0.0125	0.00356	0.02889
AGE	1	0.09127	0.02560	3.56	0.0005	0.04050	0.14203
BEDS	1	-0.00900	0.00300	-3.00	0.0034	-0.01495	-0.00305
NURSES	1	-0.00544	0.00195	-2.79	0.0063	-0.00931	-0.00157
CULTURING	1	0.03374	0.01343	2.51	0.0135	0.00711	0.06038

Iteration 4 :-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.98743	1.44557	0.68	0.4961	-1.87887	3.85373
RISK	1	0.28309	0.10951	2.58	0.0111	0.06594	0.50023
CENSUS	1	0.02013	0.00361	5.58	<.0001	0.01297	0.02728
XRAY	1	0.01625	0.00638	2.55	0.0123	0.00361	0.02889
AGE	1	0.09122	0.02557	3.57	0.0005	0.04052	0.14192
BEDS	1	-0.00900	0.00299	-3.01	0.0033	-0.01494	-0.00307
NURSES	1	-0.00544	0.00195	-2.79	0.0063	-0.00930	-0.00157
CULTURING	1	0.03378	0.01341	2.52	0.0133	0.00719	0.06036

As we can see, the change in magnitude becomes insignificant after 3rd iteration.

Q 2:-

As per the question I have generated1000 bootstrap samples using random X sampling. For each bootstrap sample, and then I regress Y on X using OLS method and obtain the residuals. After that I estimated the standard deviation function by regressing the absolute residuals on X and then use the fitted standard deviation function to obtain weights, and finally I use WLS to regress Y and X and obtain the bootstrap estimated regression b ∗ 1.

Fitting Standard deviation:-

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	23.04683	23.04683	5.89	0.0415
Error	8	31.32500	3.91562		
Corrected Total	9	54.37183			

Root MSE	1.97879	R-Square	0.4239
Dependent Mean	3.12691	Adj R-Sq	0.3519
Coeff Var	63.28278		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.42370	5.62034	-1.85	0.1008
age	1	0.56227	0.23176	2.43	0.0415

Results of WLS of b0 and b1

The REG Procedure
Model: MODEL1
Dependent Variable: dbp

Number of Observations Read	10
Number of Observations Used	10

Weight: wt

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.96041	13.96041	9.56	0.0149
Error	8	11.68538	1.46067		
Corrected Total	9	25.64579			

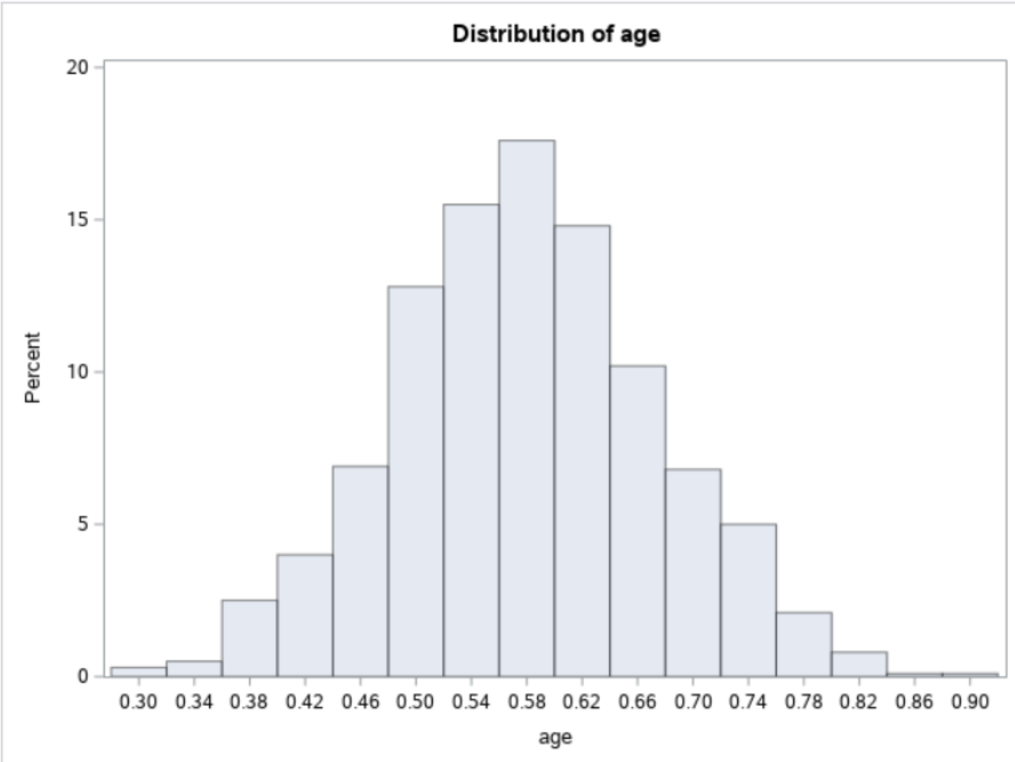
Root MSE	1.20858	R-Square	0.5444
Dependent Mean	66.31340	Adj R-Sq	0.4874
Coeff Var	1.82253		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	41.23538	8.14334	5.06	0.0010
age	1	1.17804	0.38106	3.09	0.0149

Bootstrap histogram:-

The UNIVARIATE Procedure



Value of b1*;

I have found out the bootstrap estimated regression, $b1*$, is approximately 0.58003. Here is the output to justify that :-

Values of s, b0, and b1

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	Diastolic_BP
1	MODEL1	PARMS	Diastolic_BP	8.14575	56.1569	0.58003	-1

Bootstrap sample

Obs	SampleID	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	age	dbp
1	1	MODEL1	PARMS	dbp	7.50147	49.8348	0.76591	-1
2	2	MODEL1	PARMS	dbp	7.65142	58.5624	0.47349	-1
3	3	MODEL1	PARMS	dbp	7.01931	49.4862	0.73008	-1
4	4	MODEL1	PARMS	dbp	7.36390	55.0548	0.62670	-1
5	5	MODEL1	PARMS	dbp	9.82496	51.3620	0.72519	-1
6	6	MODEL1	PARMS	dbp	7.89100	57.3685	0.54307	-1
7	7	MODEL1	PARMS	dbp	7.18217	59.7710	0.42970	-1
8	8	MODEL1	PARMS	dbp	6.95926	59.5471	0.43689	-1
9	9	MODEL1	PARMS	dbp	6.98917	52.0152	0.76367	-1
10	10	MODEL1	PARMS	dbp	7.91466	52.6806	0.66404	-1

Ashish Mani Acharya AXA190076 STAT 6337 Project 3

I tried all the method as mentioned in our class as you can see from my code. Unfortunately I could only get CI estimations from percentile bootsrap and basic bootstrap CI
here are the results:-

Basic bootstrap and percentile bootstrap respectively:-

Values of s, b0, and b1

Obs	r2	bias	p_lb	p_ub
1	.	.	0.3963	0.4077

Values of s, b0, and b1

CI95_Lower	CI95_Upper
0.30476	0.87520

Bonus Question:-

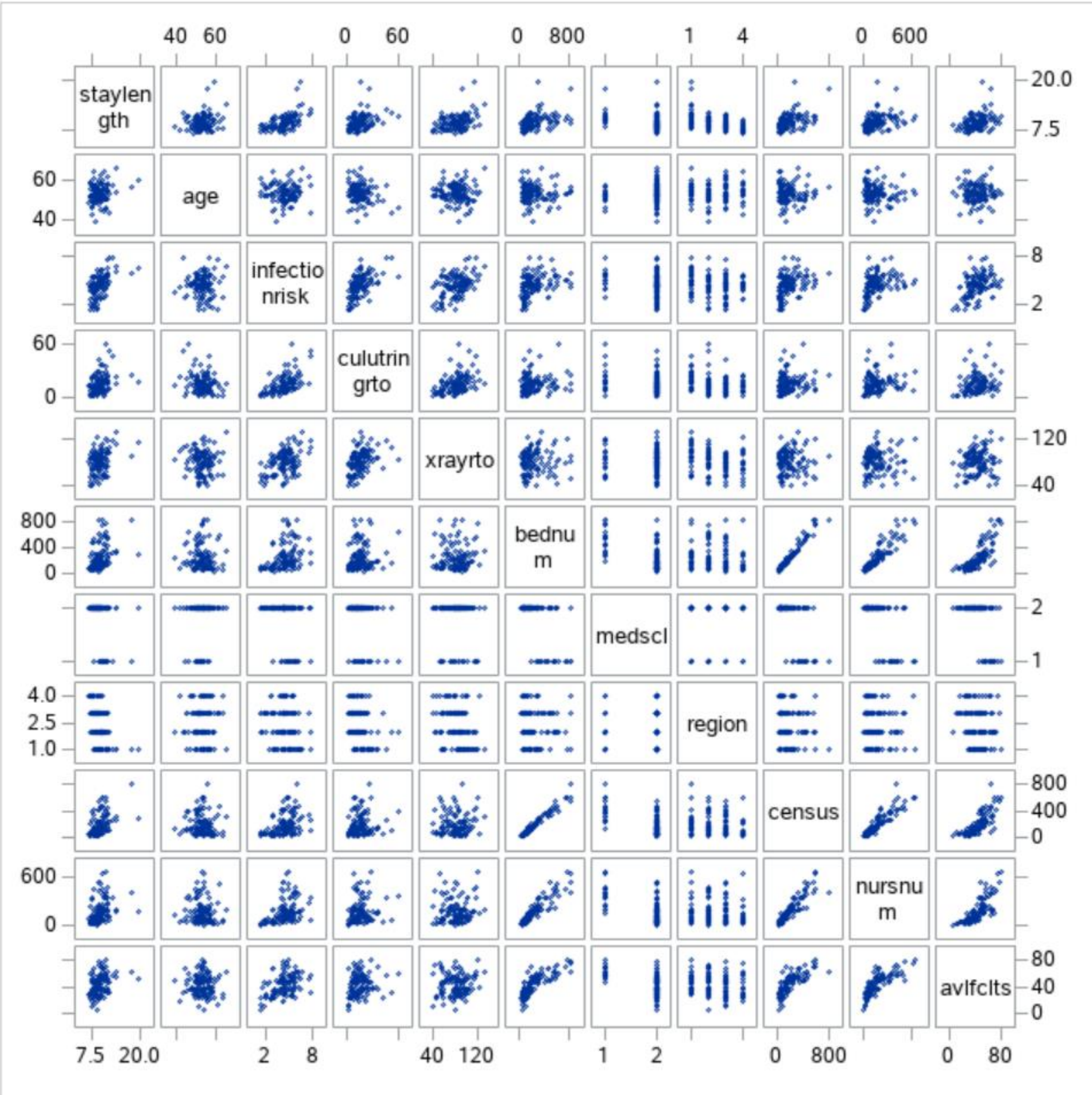
Heart dataset contains 15 independent variables; among them the ones significant at α=0.1 level (as per the simple logistic model) are

1. male (categorical, denotes the sex. male = 0 for females)
2. age
3. education (categorical)
4. CigsPerDay
5. BPMeds (categorical)
6. PrevalentStroke (categorical)
7. PrevalentHyp (categorical)
8. Diabetes (categorical)
9. TotChol
10. sysBP
11. diaBP
12. BMI
13. glucose

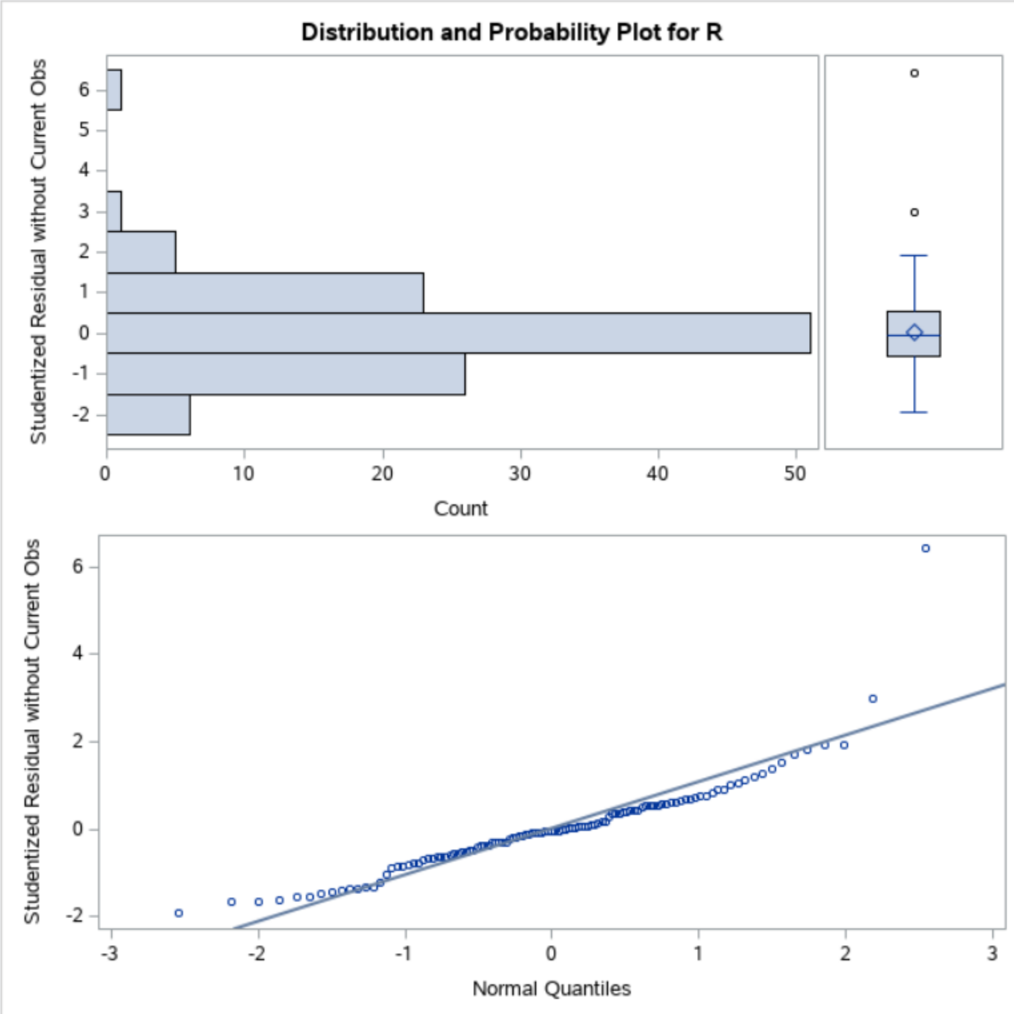
We can conclude that the model with these 13 variables jointly significant and stepwise forward selection is run on this model. The details are in the output section.

Ashish Mani Acharya *AXA190076* *STAT 6337* *Project 3*

Output for Q 1:-



Before transformation QQ plot



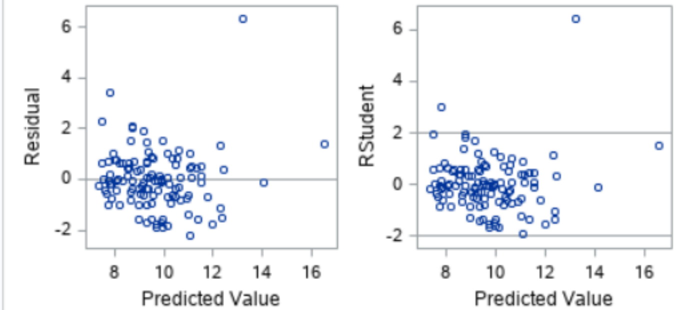
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.864616	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.104197	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.292998	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.867247	Pr > A-Sq	<0.0050

Lack of Fit test

Number of Observations Used	113

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	256.67902	25.66790	17.16	<.0001
Error	102	152.53136	1.49541		
Lack of Fit	102	152.53136	1.49541	.	.
Pure Error	0	0	.		
Corrected Total	112	409.21038			

Residual plot:-



Bresusch pagan test:-

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.30	1	0.5847	FACILITIES, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.00	1	0.9915	NURSES, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.65	1	0.4185	CENSUS, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	1.32	1	0.2505	REGION, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.00	1	0.9769	MEDSCHOOL, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.30	1	0.5822	BEDS, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	1.79	1	0.1805	XRAY, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	0.11	1	0.7398	CULTURING, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	1.59	1	0.2070	RISK, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	1.04	1	0.3078	AGE, 1

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
STAYLENGTH	Breusch-Pagan	8.76	10	0.5546	AGE, RISK, CULTURING, XRAY, BEDS, MEDSCHOOL, REGION, CENSUS, NURSES, FACILITIES, 1

Brown Forsythe test:-

BROWN FORSYTHE TEST (XRAY)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group4	1	0.1181	0.1181	0.19	0.6635
Error	111	69.4683	0.6258		

BROWN FORSYTHE TEST (XRAY)

The GLM Procedure

Level of Group4	N	R	
		Mean	Std Dev
0	57	0.08546222	0.82248359
1	56	-0.05700994	1.19680156

BROWN FORSYTHE TEST (BEDS)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group5	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (BEDS)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group5	1	1.0783	1.0783	1.76	0.1874
Error	111	67.5264	0.6083		

BROWN FORSYTHE TEST (BEDS)

The GLM Procedure

Level of Group5	N	R	
		Mean	Std Dev
0	57	-0.08035269	0.88147519
1	56	0.13499029	1.22213714

BROWN FORSYTHE TEST (MEDSCHOOL)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group6	1	0

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (MEDSCHOOL)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group6	0	0	.	.	.
Error	112	69.8063	0.6233		

BROWN FORSYTHE TEST (MEDSCHOOL)

The GLM Procedure

Level of Group6	N	R	
		Mean	Std Dev
0	113	0.01465655	1.06486767

BROWN FORSYTHE TEST (REGION)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group7	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (REGION)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group7	1	0.4757	0.4757	0.76	0.3845
Error	111	69.2637	0.6240		

BROWN FORSYTHE TEST (REGION)

The GLM Procedure

Level of Group7	N	R	
		Mean	Std Dev
0	60	-0.00833518	1.19973699
1	53	0.04111130	0.89887622

The MEANS Procedure

Variable	Median
AGE	53.2000000
RISK	4.4000000
CULTURING	14.1000000
XRAY	82.3000000
BEDS	186.0000000
MEDSCHOOL	2.0000000
REGION	2.0000000
CENSUS	143.0000000
NURSES	132.0000000
FACILITIES	42.9000000

BROWN FORSYTHE TEST (AGE)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group1	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (AGE)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group1	1	0.4119	0.4119	0.67	0.4156
Error	111	68.4685	0.6168		

BROWN FORSYTHE TEST (AGE)

The GLM Procedure

Level of Group1	N	R	
		Mean	Std Dev
0	60	-0.08749268	0.91429487
1	53	0.13872382	1.21169139

BROWN FORSYTHE TEST (RISK)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group2	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (RISK)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group2	1	0.6257	0.6257	1.00	0.3186
Error	111	69.2084	0.6235		

BROWN FORSYTHE TEST (RISK)

The GLM Procedure

Level of Group2	N	R	
		Mean	Std Dev
0	58	0.01374361	0.88572269
1	55	0.01603019	1.23445512

BROWN FORSYTHE TEST (CULTURING)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group3	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (CULTURING)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group3	1	0.00547	0.00547	0.01	0.9259
Error	111	69.9117	0.6298		

BROWN FORSYTHE TEST (CULTURING)

The GLM Procedure

Level of Group3	N	R	
		Mean	Std Dev
0	57	-0.02193633	0.94224816
1	56	0.05230644	1.18423874

BROWN FORSYTHE TEST (CENSUS)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group8	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (CENSUS)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group8	1	1.2014	1.2014	1.95	0.1656
Error	111	68.4751	0.6169		

BROWN FORSYTHE TEST (CENSUS)

The GLM Procedure

Level of Group8	N	R	
		Mean	Std Dev
0	57	-0.05987168	0.87167166
1	56	0.08894413	1.23562340

(2) BIDn ForsYthe Test (NURSES)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group9	2	0 1

Number of Observations Read	113
Number of Observations Used	113

(2) BIDn ForsYthe Test (NURSES)

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group9	1	0.4218	0.4218	0.68	0.4125
Error	111	69.1781	0.6232		

(2) BIDn ForsYthe Test (NURSES)

The GLM Procedure

Level of Group9	N	R	
		Mean	Std Dev
0	57	-0.05317100	0.90646666
1	56	0.08499887	1.20955823

BROWN FORSYTHE TEST (FACILITIES)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Group10	2	0 1

Number of Observations Read	113
Number of Observations Used	113

BROWN FORSYTHE TEST (FACILITIES)

The GLM Procedure

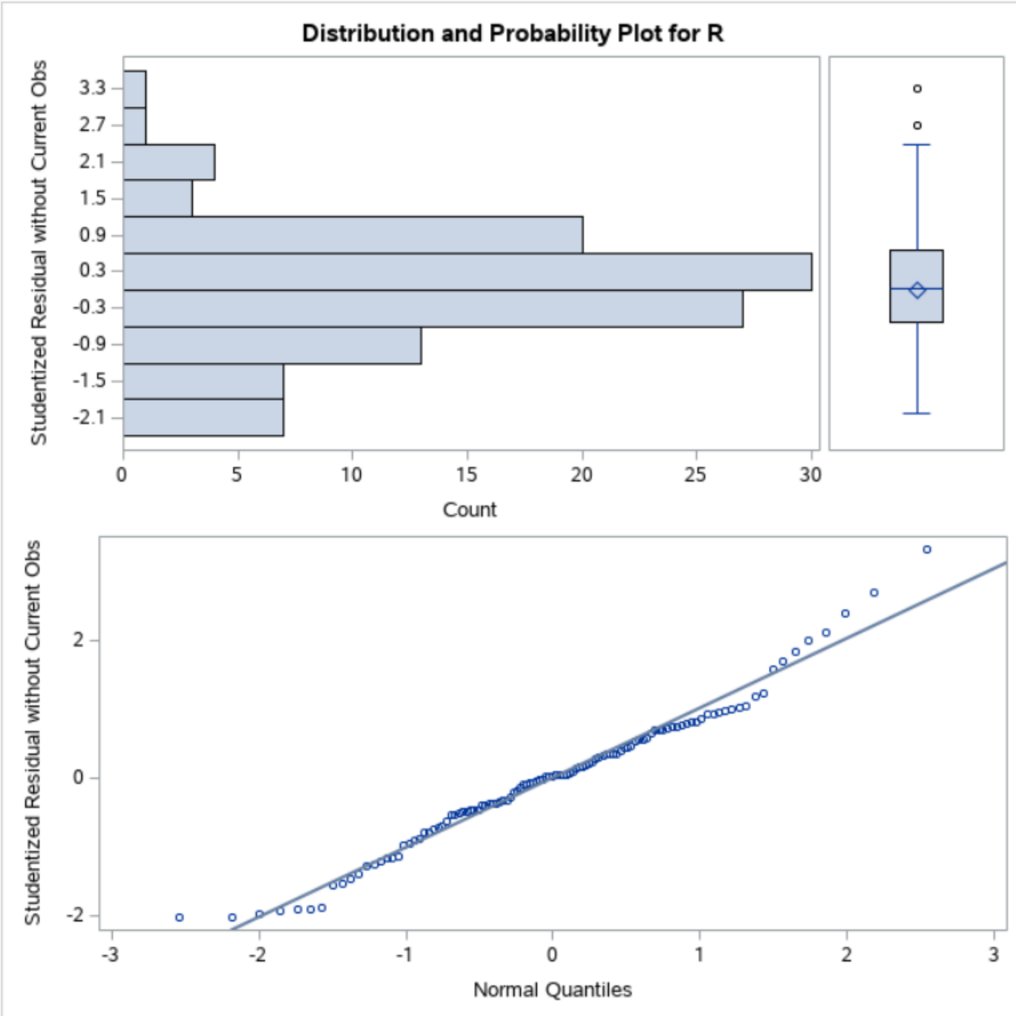
Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group10	1	0.6083	0.6083	0.97	0.3266
Error	111	69.5470	0.6265		

BROWN FORSYTHE TEST (FACILITIES)

The GLM Procedure

Level of Group10	N	R	
		Mean	Std Dev
0	57	-0.01332796	0.90455141
1	56	0.04354434	1.21427192

-1.90769	14	3.32298	43



Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
STAYLENGTH	113	9.64832	1.91146	1090	6.70000	19.56000
AGE	113	53.23186	4.46161	6015	38.80000	65.90000
RISK	113	4.35487	1.34091	492.10000	1.30000	7.80000
CULTURING	113	15.79292	10.23471	1785	1.60000	60.50000
XRAY	113	81.62832	19.36383	9224	39.60000	133.50000
BEDS	113	252.16814	192.84269	28495	29.00000	835.00000
MEDSCHOOL	113	1.84956	0.35910	209.00000	1.00000	2.00000
REGION	113	2.36283	1.00944	267.00000	1.00000	4.00000
CENSUS	113	191.37168	153.75956	21625	20.00000	791.00000
NURSES	113	173.24779	139.26539	19577	14.00000	656.00000
FACILITIES	113	43.15929	15.20086	4877	5.70000	80.00000

Results of Stepwise calculation:-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.97591	1.45210	0.67	0.5030	-1.90333	3.85515
RISK	1	0.28447	0.11022	2.58	0.0112	0.06592	0.50301
CENSUS	1	0.02013	0.00363	5.54	<.0001	0.01293	0.02733
XRAY	1	0.01620	0.00642	2.53	0.0130	0.00348	0.02892
AGE	1	0.09153	0.02568	3.56	0.0006	0.04061	0.14245
BEDS	1	-0.00899	0.00301	-2.98	0.0036	-0.01497	-0.00301
NURSES	1	-0.00546	0.00196	-2.79	0.0063	-0.00935	-0.00158
CULTURING	1	0.03359	0.01350	2.49	0.0144	0.00683	0.06036

Medians for IWLS procedures:-

Stepwise Selection

The UNIVARIATE Procedure
Variable: residual3 (Residual)

Moments			
N	113	Sum Weights	113
Mean	0.07238228	Sum Observations	8.17919818
Std Deviation	1.27799712	Variance	1.63327663
Skewness	1.54189456	Kurtosis	6.8034814
Uncorrected SS	183.519012	Corrected SS	182.926983
Coeff Variation	1765.62141	Std Error Mean	0.12022386

Basic Statistical Measures			
Location		Variability	
Mean	0.07238	Std Deviation	1.27800
Median	-0.06523	Variance	1.63328
Mode	.	Range	9.35588
		Interquartile Range	1.67119

Stepwise Selection

The UNIVARIATE Procedure
Variable: abs_Res3_median

Moments			
N	113	Sum Weights	113
Mean	0.94745611	Sum Observations	107.062541
Std Deviation	0.85620587	Variance	0.73308849
Skewness	3.56118276	Kurtosis	22.2412887
Uncorrected SS	183.542969	Corrected SS	82.1059108
Coeff Variation	90.3689214	Std Error Mean	0.08054507

Basic Statistical Measures			
Location		Variability	
Mean	0.947456	Std Deviation	0.85621
Median	0.853649	Variance	0.73309
Mode	.	Range	7.02541
		Interquartile Range	0.80589

Stepwise Selection

The UNIVARIATE Procedure
Variable: abs_Res2_median

Moments			
N	113	Sum Weights	113
Mean	0.94586304	Sum Observations	106.882523
Std Deviation	0.86573804	Variance	0.74950236
Skewness	3.56142672	Kurtosis	22.1970005
Uncorrected SS	185.040493	Corrected SS	83.9442644
Coeff Variation	91.5289011	Std Error Mean	0.08144178

Basic Statistical Measures			
Location		Variability	
Mean	0.945863	Std Deviation	0.86574
Median	0.820663	Variance	0.74950
Mode	.	Range	7.09655
		Interquartile Range	0.76453

Stepwise Selection

The UNIVARIATE Procedure
Variable: abs_Res1_median

Moments			
N	113	Sum Weights	113
Mean	0.94584309	Sum Observations	106.88027
Std Deviation	0.86304221	Variance	0.74484186
Skewness	3.53209255	Kurtosis	21.927057
Uncorrected SS	184.514253	Corrected SS	83.4222878
Coeff Variation	91.2458118	Std Error Mean	0.08118818

Basic Statistical Measures			
Location		Variability	
Mean	0.945843	Std Deviation	0.86304
Median	0.839637	Variance	0.74484
Mode	.	Range	7.05974
		Interquartile Range	0.78435

Stepwise Selection

The UNIVARIATE Procedure
Variable: abs_Res0_median

Moments			
N	113	Sum Weights	113
Mean	0.94907174	Sum Observations	107.245107
Std Deviation	0.85034257	Variance	0.72308249
Skewness	3.21267173	Kurtosis	18.8854307
Uncorrected SS	182.768539	Corrected SS	80.9852387
Coeff Variation	89.5972911	Std Error Mean	0.0799935

Basic Statistical Measures			
Location		Variability	
Mean	0.949072	Std Deviation	0.85034
Median	0.851565	Variance	0.72308
Mode	.	Range	6.76150
		Interquartile Range	0.81256

Stepwise Selection

The UNIVARIATE Procedure
Variable: residual1 (Residual)

Moments			
N	113	Sum Weights	113
Mean	0.06128768	Sum Observations	6.92550832
Std Deviation	1.2768921	Variance	1.63045344
Skewness	1.52021449	Kurtosis	6.68724627
Uncorrected SS	183.035234	Corrected SS	182.610785
Coeff Variation	2083.43996	Std Error Mean	0.1201199

Basic Statistical Measures			
Location		Variability	
Mean	0.06129	Std Deviation	1.27689
Median	-0.06850	Variance	1.63045
Mode	.	Range	9.31355
		Interquartile Range	1.67929

Stepwise Selection

The UNIVARIATE Procedure
Variable: residual2 (Residual)

Moments			
N	113	Sum Weights	113
Mean	0.0698938	Sum Observations	7.89799921
Std Deviation	1.27783508	Variance	1.63286248
Skewness	1.5389978	Kurtosis	6.78956565
Uncorrected SS	183.432619	Corrected SS	182.880598
Coeff Variation	1828.25244	Std Error Mean	0.12020861

Basic Statistical Measures			
Location		Variability	
Mean	0.06989	Std Deviation	1.27784
Median	-0.06836	Variance	1.63286
Mode	.	Range	9.34943
		Interquartile Range	1.67141

Stepwise Selection

The UNIVARIATE Procedure
Variable: residual0 (Residual)

Moments			
N	113	Sum Weights	113
Mean	0	Sum Observations	0
Std Deviation	1.27262568	Variance	1.61957611
Skewness	1.33923159	Kurtosis	5.63163326
Uncorrected SS	181.392524	Corrected SS	181.392524
Coeff Variation	.	Std Error Mean	0.11971855

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	1.27263
Median	-0.11035	Variance	1.61958
Mode	.	Range	9.18728
		Interquartile Range	1.70037

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-6.5	Pr >= |M|	0.2589
Signed Rank	S	-215.5	Pr >= |S|	0.5394

Parameters during IWLS calculations:-

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.98673	1.44746	0.68	0.4969	-1.88332	3.85678
RISK	1	0.28342	0.10972	2.58	0.0112	0.06587	0.50098
CENSUS	1	0.02013	0.00362	5.57	<.0001	0.01296	0.02730
XRAY	1	0.01623	0.00639	2.54	0.0125	0.00356	0.02889
AGE	1	0.09127	0.02560	3.56	0.0005	0.04050	0.14203
BEDS	1	-0.00900	0.00300	-3.00	0.0034	-0.01495	-0.00305
NURSES	1	-0.00544	0.00195	-2.79	0.0063	-0.00931	-0.00157
CULTURING	1	0.03374	0.01343	2.51	0.0135	0.00711	0.06038

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.97591	1.45210	0.67	0.5030	-1.90333	3.85515
RISK	1	0.28447	0.11022	2.58	0.0112	0.06592	0.50301
CENSUS	1	0.02013	0.00363	5.54	<.0001	0.01293	0.02733
XRAY	1	0.01620	0.00642	2.53	0.0130	0.00348	0.02892
AGE	1	0.09153	0.02568	3.56	0.0006	0.04061	0.14245
BEDS	1	-0.00899	0.00301	-2.98	0.0036	-0.01497	-0.00301
NURSES	1	-0.00546	0.00196	-2.79	0.0063	-0.00935	-0.00158
CULTURING	1	0.03359	0.01350	2.49	0.0144	0.00683	0.06036

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.91863	1.45755	0.63	0.5299	-1.97142	3.80867
RISK	1	0.28868	0.11066	2.61	0.0104	0.06926	0.50810
CENSUS	1	0.02021	0.00365	5.54	<.0001	0.01298	0.02745
XRAY	1	0.01640	0.00643	2.55	0.0122	0.00365	0.02916
AGE	1	0.09233	0.02577	3.58	0.0005	0.04123	0.14343
BEDS	1	-0.00897	0.00302	-2.96	0.0038	-0.01496	-0.00297
NURSES	1	-0.00560	0.00197	-2.84	0.0054	-0.00951	-0.00169
CULTURING	1	0.03297	0.01356	2.43	0.0167	0.00609	0.05986

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.98743	1.44557	0.68	0.4961	-1.87887	3.85373
RISK	1	0.28309	0.10951	2.58	0.0111	0.06594	0.50023
CENSUS	1	0.02013	0.00361	5.58	<.0001	0.01297	0.02728
XRAY	1	0.01625	0.00638	2.55	0.0123	0.00361	0.02889
AGE	1	0.09122	0.02557	3.57	0.0005	0.04052	0.14192
BEDS	1	-0.00900	0.00299	-3.01	0.0033	-0.01494	-0.00307
NURSES	1	-0.00544	0.00195	-2.79	0.0063	-0.00930	-0.00157
CULTURING	1	0.03378	0.01341	2.52	0.0133	0.00719	0.06036

Output for Q 3:-

Simple logistic regression for 15 variables:-

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.0792	0.2754	340.1573	<.0001
sysBP	1	0.0246	0.00194	162.0188	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.4758	0.3187	197.1868	<.0001
diaBP	1	0.0326	0.00366	79.4236	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0391	0.2502	147.5679	<.0001
TotChol	1	0.00550	0.00101	29.8338	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1819	0.1078	120.3148	<.0001
Diabetes	0	1	-0.5784	0.1078	28.8144	<.0001

PrevalentHyp	1	114.6518	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6043	0.0469	1170.0348	<.0001
PrevalentHyp	0	1	-0.5022	0.0469	114.6518	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1060	0.2259	23.9777	<.0001
PrevalentStroke	0	1	-0.6205	0.2259	7.5473	0.0060

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.2270	0.1034	140.7184	<.0001
BPMeds	0	1	-0.5339	0.1034	26.6404	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8270	0.0593	948.7787	<.0001
CigsPerDay	1	0.0115	0.00367	9.8977	0.0017

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0738	0.2927	50.1933	<.0001
HeartRate	1	0.00471	0.00379	1.5392	0.2147

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.5700	0.1440	318.6051	<.0001
glucose	1	0.0102	0.00161	40.1968	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7098	0.0463	1363.2869	<.0001
male	0	1	-0.2552	0.0463	30.3750	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8012	0.0545	1091.7449	<.0001
education	1	1	0.3557	0.0714	24.8376	<.0001
education	2	1	-0.2009	0.0855	5.5239	0.0188
education	3	1	-0.1560	0.1029	2.3014	0.1293

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7161	0.0460	1389.6044	<.0001
CurrentSmoker	0	1	-0.0534	0.0460	1.3436	0.2464

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7098	0.0463	1363.2869	<.0001
male	0	1	-0.2552	0.0463	30.3750	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0862	0.2855	116.8226	<.0001
BMI	1	0.0525	0.0107	24.1851	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7098	0.0463	1363.2869	<.0001
male	0	1	-0.2552	0.0463	30.3750	<.0001

full model with 13 variables . they look jointly significant.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	367.9099	15	<.0001
Score	371.4981	15	<.0001
Wald	310.1566	15	<.0001

SAS codes:-

```
/* Q1*/

filename Senic "/home/u59316208/6337/Senic.dat";

DATA A;
INFILE Senic;
INPUT ID STAYLENGTH AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES;
RUN;

/* Scatter plots*/
PROC SGSCATTER DATA=A;
```

```
MATRIX STAYLENGTH AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES;

RUN;




/*Correlation coefficients*/

PROC CORR DATA = A;


VAR STAYLENGTH AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES;

RUN;




/* Regression Diagnostic:*/


/* linearity test*/

PROC REG DATA = A;

MODEL STAYLENGTH = AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES/ LACKFIT;


OUTPUT OUT=B RSTUDENT=R PREDICTED=P;


RUN;


/* normality test*/

PROC UNIVARIATE DATA=B NORMAL PLOT;

VAR R;

RUN;


PROC MODEL DATA = A;


/*breush pagan test*/

PARMS b0 b1 b2 b3 b4 b5 b6 b7 b8 b9 b10;
```

```
STAYLENGTH = b0 + b1*AGE + b2*RISK + b3*CULTURING +b4*XRAY +b5*BEDS +b6*MEDSCHOOL +b7*REGION
+b8*CENSUS +b9*NURSES +b10*FACILITIES ;

FIT STAYLENGTH /BREUSCH=(AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES);

FIT STAYLENGTH /BREUSCH=(AGE);

FIT STAYLENGTH /BREUSCH=(RISK);

FIT STAYLENGTH /BREUSCH=(CULTURING);

FIT STAYLENGTH /BREUSCH=(XRAY);

FIT STAYLENGTH /BREUSCH=(BEDS);

FIT STAYLENGTH /BREUSCH=(MEDSCHOOL);

FIT STAYLENGTH /BREUSCH=(REGION);

FIT STAYLENGTH /BREUSCH=(CENSUS);

FIT STAYLENGTH /BREUSCH=(NURSES);

FIT STAYLENGTH /BREUSCH=(FACILITIES);

RUN;


/* Get medians for BROWN FORSYTHE TEST */

PROC MEANS DATA = A MEDIAN;

VAR AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES;

RUN;


DATA B; SET B;

Group1 = (AGE > 53.2000000);

Group2 = (RISK > 4.4000000);

Group3 = (CULTURING > 14.1000000);

Group4 = (XRAY > 82.3000000);

Group5 = (BEDS > 186.0000000);

Group6 = (MEDSCHOOL > 2.0000000);

Group7 = (REGION > 2.0000000);

Group8 = (CENSUS > 143.0000000);

Group9 = (NURSES > 132.0000000);
```

```
Group10 = (FACILITIES > 42.9000000);

RUN;


/* BROWN FORSYTHE TESTs for homogeneity*/

TITLE "BROWN FORSYTHE TEST   (AGE)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group1;

MODEL R = Group1 / NOUNI;

MEANS Group1 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (RISK)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group2;

MODEL R = Group2 / NOUNI;

MEANS Group2 / HOVTEST = BF;

RUN;

TITLE "BROWN FORSYTHE TEST   (CULTURING)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group3;

MODEL R = Group3 / NOUNI;

MEANS Group3 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (XRAY)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group4;

MODEL R = Group4 / NOUNI;

MEANS Group4 / HOVTEST = BF;

RUN;
```

```
TITLE "BROWN FORSYTHE TEST   (BEDS)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group5;

MODEL R = Group5 / NOUNI;

MEANS Group5 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (MEDSCHOOL)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group6;

MODEL R = Group6 / NOUNI;

MEANS Group6 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (REGION)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group7;

MODEL R = Group7 / NOUNI;

MEANS Group7 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (CENSUS)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group8;

MODEL R = Group8 / NOUNI;

MEANS Group8 / HOVTEST = BF;

RUN;


TITLE "(2) BIDn ForsYthe Test (NURSES)";
```

```
PROC GLM DATA = B PLOTS = NONE;

CLASS Group9;

MODEL R = Group9 / NOUNI;

MEANS Group9 / HOVTEST = BF;

RUN;


TITLE "BROWN FORSYTHE TEST   (FACILITIES)";

PROC GLM DATA = B PLOTS = NONE;

CLASS Group10;

MODEL R = Group10 / NOUNI;

MEANS Group10 / HOVTEST = BF;

RUN;


PROC TRANSREG DATA=A; /* Finding Box-Cox transformation power */

MODEL BoxCox(STAYLENGTH)=identity(AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES
FACILITIES);

RUN;


DATA A; SET A;

YTRANS = (STAYLENGTH**-0.75-1)/-0.75;

RUN;


PROC REG Data = A NOPRINT;

MODEL YTRANS = AGE RISK CULTURING XRAY BEDS MEDSCHOOL REGION CENSUS NURSES FACILITIES;

OUTPUT OUT=F RSTUDENT=R;

RUN;


PROC UNIVARIATE DATA=F NORMAL PLOT;

VAR R;
```

```
RUN;

PROC REG DATA = A;
MODEL YTRANS = AGE RISK CULTURING XRAY BEDS CENSUS NURSES FACILITIES/ selection = stepwise; /* stepwise selection */
TITLE "Stepwise Selection";
RUN;

PROC REG DATA = A NOPRINT;
MODEL STAYLENGTH = RISK CENSUS XRAY AGE BEDS NURSES CULTURING /R CLB;
output out = iteration0 r = residual0;
RUN; /* fitting the OLS model */

DATA iteration0; SET iteration0;
abs_res0 = abs(residual0);
RUN;

PROC UNIVARIATE DATA = iteration0;
VAR residual0;
RUN; /* median = -0.11035 for MAD computation */
DATA iteration0; SET iteration0;
abs_Res0_median = abs(residual0 + 0.11035);
RUN;

PROC UNIVARIATE DATA = iteration0;
VAR abs_Res0_median;
RUN; /* median = 0.851565, MAD = 0.071617/0.6745 */
DATA iteration0; SET iteration0;
```

```
u = residual0/0.851565*0.6745; w = 1;

IF abs(u) > 1.345 THEN w = 1.345/abs(u); /* Huber wt */

RUN;


PROC REG Data= iteration0; /* IWLS with Iteration0 wt */

MODEL STAYLENGTH = RISK CENSUS XRAY AGE BEDS NURSES CULTURING /R CLB;

WEIGHT w;

output out = iteration1 r = residual1;

RUN; /* IWLS Iteration 1 complete */

DATA iteration1; SET iteration1;

abs_res1 = abs(residual1);

RUN;


PROC UNIVARIATE DATA = iteration1;

VAR residual1;

RUN; /* median = -0.06850         for MAD computation */

DATA iteration1; SET iteration1;

abs_Res1_median = abs(residual1 + 0.06850       );

RUN;


PROC UNIVARIATE DATA = iteration1;

VAR abs_Res1_median;

RUN; /* median = 0.839637       , MAD = 0.072674/0.6745 */

DATA iteration1; SET iteration1;

u1 = residual1/0.839637             *0.6745; w1 = 1;

IF abs(u1) > 1.345 THEN w1 = 1.345/abs(u1); /* Huber wt */

RUN;


PROC REG Data= iteration1; /* IWLS with Iteration1 wt */

MODEL STAYLENGTH = RISK CENSUS XRAY AGE BEDS NURSES CULTURING/R CLB;
```

```
WEIGHT w1;
output out = iteration2 r = residual2;
RUN; /* IWLS Iteration 2 complete */

DATA iteration2; SET iteration2;
abs_res2 = abs(residual2);
RUN;

PROC UNIVARIATE DATA = iteration2;
VAR residual2;
RUN; /* median = -0.06836         for MAD computation */
DATA iteration2; SET iteration2;
abs_Res2_median = abs(residual2 + 0.06836       );
RUN;

PROC UNIVARIATE DATA = iteration2;
VAR abs_Res2_median;
RUN; /* median = 0.820663, MAD = 0.072447/0.6745 */

DATA iteration2; SET iteration2;
u2 = residual2/0.820663*0.6745; w2 = 1;
IF abs(u2) > 1.345 THEN w2 = 1.345/abs(u2); /* Huber wt */
RUN;

PROC REG Data= iteration2; /* IWLS with Iteration2 wt */
MODEL STAYLENGTH = RISK CENSUS XRAY AGE BEDS NURSES CULTURING /R CLB;
WEIGHT w2;
output out = iteration3 r = residual3;
RUN; /* IWLS Iteration 3 complete */
```

```
DATA iteration3; SET iteration3;
abs_res3 = abs(residual3);
RUN;

PROC UNIVARIATE DATA = iteration3;
VAR residual3;
RUN; /* median =          -0.06523 for MAD computation */

DATA iteration3; SET iteration3;
abs_Res3_median = abs(residual3 +       0.06523);
RUN;

PROC UNIVARIATE DATA = iteration3;
VAR abs_Res3_median;
RUN; /* median = 0.812700, MAD = 0.072404/0.6745 */

DATA iteration3; SET iteration3;
u3 = residual2/0.812700*0.6745; w3 = 1;
IF abs(u3) > 1.345 THEN w3 = 1.345/abs(u3); /* Huber wt */
RUN;

PROC REG Data= iteration3; /* IWLS with Iteration3 wt */
MODEL STAYLENGTH = RISK CENSUS XRAY AGE BEDS NURSES CULTURING /R CLB;
WEIGHT w3;
output out = iteration4 r = residual4;
RUN; /* IWLS Iteration 4 complete *
/*Q2*/
```

```
data BP1;
	infile '/home/u59603516/CH11TA011.dat';
	input Age1 $ Diastolic_BP;
run;

title "Contents of the Data, BP1";
proc contents data = BP1;

data BP;
	set BP1;
	Age = input(Age1, comma9.);
	drop Age1;
run;

title "Contents of the data, senic_original";
proc contents data = BP;

%let N = 1000;
proc surveyselect data = BP noprint
	out = BootFreq(rename = (replicate = SampleID)) noprint
	seed = 67138
	method = urs samprate = 1
	outhits reps = &N;
run;

proc reg data = BootFreq noprint;
by SampleID;
freq NumberHits;
model Diastolic_BP = Age/R clb;
output out = results0 r = residual0;
```

```
run;

data results0;
set results0;
absresid0 = abs(residual0);
run;

title "Standard Deviation Function";
proc reg data = results0 noprint;
model absresid0 = Age / p; /* option p requests fitted values */
output out = results1 p = yhat;
run;

data results1;
set results1;
wt = 1/(yhat**2);
run;

proc reg data = results1 noprint; /* to obtain the weighted least squares regression */
model Diastolic_BP = Age / R clb;
weight wt;
run;

*/ WLS to regress Y and X and obtain the bootstrap estimayed regression b*1;
title "ANOVA Table and the Values of the parameters beta0 and beta1";
proc reg data = BP outest = est;
model Diastolic_BP = Age;
run;

title "Values of s, b0, and b1";
```

```
proc print data = est;
run;

data est; set est;
s = _rmse_;
b0 = intercept;
b1 = Age;
keep s b0 b1;
run;

*/ To make a histogram of bootstrap distribution of b1;
proc reg data = BootFreq outest = bootstrap noprint;
by SampleID;
freq NumberHits;
model Diastolic_BP = Age/R clb;
run;

proc univariate data = bootstrap;
	var Age;
	histogram;
run;

*/to check for Normality of the Bootstrap Distribution;
proc univariate data = bootstrap normal plot;
/* Check the normality*/
var Age;
run;

*/Calculate the Normal approximation CI;
ods output FitStatistics = t0;
```

```
proc reg data = BP;

        model Diastolic_BP = Age;

run;

quit;


*store the estimated r-square;

data _null_;

        set t0;

        if label2 = "R-square" then

        call symput('r2bar', cvalue2);

run;


%let rep = 500;

proc surveyselect data = BP out=bootsample

        seed = 1347 method = urs

        samprate = 1 outhits rep = &rep;

run;

ods listing close;


* character type to numeric type;

data t1;

        set t0;

        r2 = cvalue2 + 0;

run;


* creating CI, normal distribution theory method;

* Z- distribution;

%let alphalev = .05;

ods listing;

proc sql;
```

```
        select r2bar as r2,
                mean(r2) - r2bar as bias,
                std(r2) as std_err,
                r2bar - 1.96*(1-&alphalev/2)*std(r2) as lb,
                r2bar + 1.96*(1-&alphalev/2)*std(r2) as hb
        from t1;
quit;

*/Studentized Bootstrap CI;
* creating CI, normal distribution theory method;
* using the t-distribution;
%let alphalev = .05;
ods listing;
proc sql;
        select  r2bar as r2,
                        mean(r2) - r2bar as bias,
                        std(r2) as std_err,
                        r2bar - tinv(1-&alphalev/2, &rep-1)*std(r2) as lb,
                        r2bar - tinv(&alphalev/2, &rep-1)*std(r2) as hb
        from t1;
quit;

*/Calculating the basic bootstrap CI;
%let alphalev = .05;
%let alpha1 = %sysevalf(1 - &alphalev/2);
%put &alpha1;
proc sql;
        select sum(r2<=r2bar)/count(r2) into :z0bar
        from t1;
quit;
```

```
data _null_;
    z0 = probit(&z0bar);
    zalpha = probit(&alpha1);
    p1 = put(probnorm(2*z0 - zalpha)*100, 3.0);
    p2 = put(probnorm(2*z0 + zalpha)*100, 3.0);
    output;
    call symput('a1', p1);
    call symput('a2', p2);
run;

*creating CI, bias-corrected method;
proc univariate data = t1 alpha = .05;
    var r2;
    output out = pmethod mean = r2hat pctlpts = 5 95 pctlpre = p pctlname = _lb _ub ;
run;

data t2;
    set pmethod;
    bias = r2hat - r2bar;
    r2 = r2bar;
run;

ods listing;

proc print data = t2;
    var r2 bias p_lb p_ub;
run;

*/Calculate the percentile Bootstrap CI;
```

```
*/ Using approx sampling distribution to make statistical inferences.;

proc univariate data = bootstrap noprint;

    var Age;

    output out = Pctl pctlpre = CI95_

        pctlpts = 2.5 97.5      /* 95% bootstrap CI */

        pctlname = Lower Upper;

run;


proc print data = Pctl noobs; run;



/*bonus question */

DATA heart;

INFILE "/home/u59316208/6337/heart (1).csv" DLM=',' MISSOVER DSD FIRSTOBS=1;

INPUT male age education CurrentSmoker CigsPerDay BPMeds PrevalentStroke PrevalentHyp Diabetes

TotChol sysBP diaBP BMI HeartRate glucose TenyearCHD;

RUN;

/* Code for data input ends here */

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS male; /* category var */

MODEL TenyearCHD = male;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = age;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS education; /* category var */

MODEL TenyearCHD = education;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */
```

```
CLASS CurrentSmoker; /* category var */

MODEL TenyearCHD = CurrentSmoker;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = CigsPerDay;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS BPMeds; /* category var */

MODEL TenyearCHD = BPMeds;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS PrevalentStroke; /* category var */

MODEL TenyearCHD = PrevalentStroke;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS PrevalentHyp; /* category var */

MODEL TenyearCHD = PrevalentHyp;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS Diabetes; /* category var */

MODEL TenyearCHD = Diabetes;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = TotChol;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = sysBP;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = diaBP;
```

```
RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = BMI;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = HeartRate;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

MODEL TenyearCHD = glucose;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS male education BPMeds PrevalentStroke PrevalentHyp Diabetes; /* category var */

MODEL TenyearCHD = male age education CigsPerDay BPMeds PrevalentStroke PrevalentHyp Diabetes

TotChol sysBP diaBP BMI glucose ;

RUN;

PROC Logistic Data = heart descending; /* model P(Y=1) */

CLASS male education BPMeds PrevalentStroke PrevalentHyp Diabetes; /* category var */

MODEL TenyearCHD = male age education CigsPerDay BPMeds PrevalentStroke PrevalentHyp Diabetes

TotChol sysBP diaBP BMI glucose / SELECTION = F SLENTRY=0.1 SLSTAY=0.1; /* stepwise forward selection*/

RUN;
```