Question 1

Full model for predicting uric acid levels using all other explanatory variables

$$\text{uric} = \beta_0 + \beta_1 * \text{dia} + \beta_2 * \text{hdl} + \beta_3 * \text{choles} + \beta_4 * \text{trig} + \beta_5 * \text{alco}$$

$$\text{uric} = 92.04641 + 1.42445 * \text{dia} + 4.59383 * \text{hdl} - 6.45949 * \text{choles} + 99.70139 * \text{trig} + 0.42497 * \text{alco}$$

Test if the variables hdl and choles can be (jointly) dropped together from the full model. Report an appropriate test value, p value and state your conclusion

Hypothesis

$H_0: \beta_2 = \beta_3 = 0$

$H_A: either\ \beta_2\ or\ \beta_3\ is\ not\ zero$

Test statistic for the partial F-test = 3.05

Since P-value is $0.0480 < 0.05$, we can conclude that (hdl, choles) are significant. Thus, in presence of other explanatory variables, they cannot be dropped together from the model.

Question 2

Find the best model(s) using adjusted R^2 criterion and stepwise selection method.

Selecting the best model using adjusted R^2 criterion:

The model with the highest adjusted $R^2 = 0.5207$ was selected. That is,

$$\text{uric} = \beta_0 + \beta_1\text{trig} + \beta_2\text{alco} + \beta_3\text{dia} + \beta_4\text{choles}$$

Selecting the best model using stepwise selection method:

$$\text{uric} = \beta_0 + \beta_1\text{trig} + \beta_2\text{alco} + \beta_3\text{dia} + \beta_4\text{choles}$$

Question 3

For one of the "best" models chosen above, check all assumptions (using all plots and tests discussed in class) and detect any outliers, influential points, and collinearity using appropriate diagnostic tools. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate plots, tests, statistics.

Chosen model is,

$$\text{uric} = \beta_0 + \beta_1\text{trig} + \beta_2\text{alco} + \beta_3\text{dia} + \beta_4\text{choles}$$

$$\text{uric} = 98.23003 + 98.58245\text{trig} + 0.43157\text{alco} + 1.43222\text{dia} - 6.19569\text{choles}$$

The p-value of ANOVA F test <0.001 suggests that the model is significant.

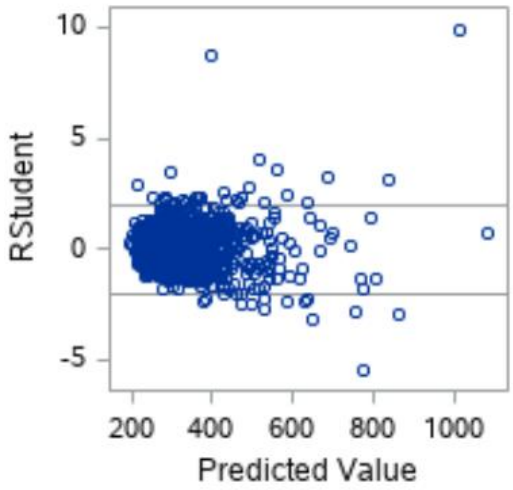
Checking the assumptions of simple linear regression model

Checking linearity of the model (Lack of fit test)

$H_0: E[uric] = \beta_0 + \beta_1 \text{trig} + \beta_2 \text{alco} + \beta_3 \text{dia} + \beta_4 \text{choles}$
$H_1: E[uric] \neq \beta_0 + \beta_1 \text{trig} + \beta_2 \text{alco} + \beta_3 \text{dia} + \beta_4 \text{choles}$

P-value = 0.8994 > 0.05 for the Lack of Fit test suggests the linearity holds for the model

Checking the Constant variance



The residual plot shows the increasing variance of residuals

(Other graphs are in output section)

Brown-Forsythe test and Breusch-Pagan test

H_0: Errors have constant variance
H_A: Errors does not have constant variance

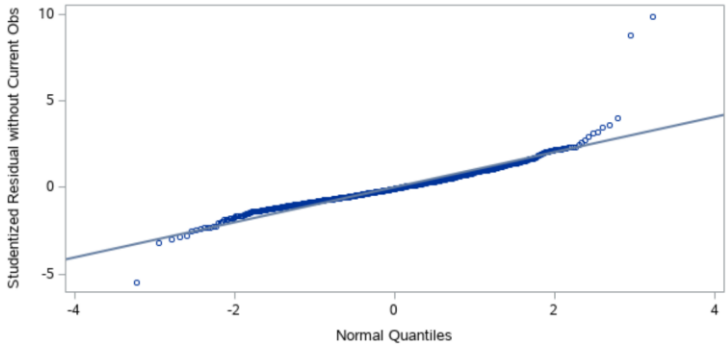
Variable	F value	P-value
Trig	58.80	<0.0001
Alco	20.87	<0.0001
Dia	15.36	<0.0001
Choles	6.63	0.0102

The Brown Forsythe test reject the constant variance assumption of the model as p values of all variables <0.05.

The Breusch-Pagan test reject the constant variance assumption of the model with p value <0.0001

Checking Normality

The Studentized Q-Q plot shows that residuals are not normally distributed as they are highly skewed to the right with heavy tails.



Shapiro-Wilk test

H_0: The errors are normally distributed
H_A: The errors are not normally distributed

The Shapiro Wilk test reject rejects the normality errors of the model with p value <0.001.

Detecting Outliers and Influential Observations:

To detect outliers and influential observations, DFFITS, DFBETAS, Cook's D and Hat matrix were computed. The following results were obtained.

Using Bonferroni method, observations 267, 477 and 483 are outliers.

Observations 22, 28, 31, 44, 46, 74, 85, 95, 103, 105, 124, 142, 149, 152, 160, 231, 233, 245, 246, 258, 303, 311, 383, 390, 397, 402, 406, 411, 421, 432, 440, 442, 449, 456, 477, 483, 490, 495, 499, 500, 506, 507, 508, 523, 525, 535, 544, 582, 583, 588, 592, 605, 625, 633, 634, 643, 661, 662, 724, 727, 736, 738, 818, 844, 851, 894, 900, 919, 924, 928, 944, 953, 969, 971, 981, 985 and 988 have high leverage.

Observations7,8,11,14,15,29,31,44,46,74,85,124,156,177,184,233,247,267,311,357,366,383,390,402,421,440,449,456,477,4853,490,499,500,508,523,534,535,600,621,646,662,696,720,736,803,823,88,897,900,924,944,953,969,981,985,988 have influence on their fitted values according to DFFITS criteria.

Observation 483 have Cook's D value that is considered to be influential. Hence this observation is influential on all the fitted values.

If we consider |DFBETAS|$>\frac{2}{\sqrt{998}}$,
observations
7,10,11,14,15,31,44,46,74,124,177,230,233,267,303,311,357,366,383,421,440,449,477,480,483,490,500,508,511,513,523,534,535,553,600,621,662,696,731,823,845,897,900,924,958,969,971,976,981,985,988 are influential on the effect of *trig*.

Observations
7,8,28,48,124,124145,227,233,244,258,267,311,366,390,441,449,456,477,483,490,499,523,582,588,605,621,633,720.736,771,818,851,888,944,953 is influential on the effect of *alco*.

Observations
15,29,52,63,103,114,1229,172,177,184,233,247,267,282,357,366,421,449,452,477,483,490,499,50

0,507,508,544,549,582,617,634,640,646,662,695,731,742,751,796,803,813,88,897,906,944,969,981 are influential on the effect of *dia*

Observations
5,8,14,23,29,34,37,43,50,55,71,74,106,11,124,145,156,165,177,103,233,247,267,274,287,323,326,383,402,421,449,277,483,490,499,523,544,562,646,673,687,762,782,796.803,823,851,872,892,895,907,924,939,948,969,976,981,985,988 are influential on the effect of *choles*

Collinearity diagnostic

To check for collinearity, VIF and Condition Index were used. All VIFs are smaller than 5 and all Condition Index are smaller than 30, so we can say there is no predictor variables which are linearly related.

Applying Transformations

In an attempt to remedy the situation, some transformations were tried and also used the Box-Cox method in identifying an appropriate transformation for the response variable *uric* based on other four variables *trig, alco, dia* and *choles.*

From the Box-Cox analysis, $\lambda = 0$, suggests that natural log transformation for *uric* is the best.

Below is the summary of the p-values from the various diagnostics tests conducted after regressing the transformed *uric* on other four variables.

Model	P-values			
	Shapiro Wilks	Breush-Pagan	Lack of fit	Model-signifiacance
uric	<0.0001	<0.0001	0.8994	<0.0001
Log uric	0.0016	<0.0001	0.9412	<0.0001
Inv sqrt uric	<0.0001	0.1553	0.9335	<0.0001

From the above table, we can observe that, at 5% significance level, both the inverse square root transformation and the log transformation fail the normality test, while the log transformation fails Breusch-Pagan test for homoscedasticity. Though Breusch-Pagan test fails for log transformation, we selected it as our final model because from the Box-Cox analysis suggests the same transformation and p value for Shapiro Wilks Higher in log uric than the inverse square root transformation.

Therefore the final linear regression model is,

$$\text{uric} = 4.99163 + 0.22011\text{trig} + 0.0010\text{alco} + 0.00541\text{dia} - 0.00927\text{choles}$$

The ANOVA F-test statistic shows that the model is significant.

Question4

Fitting a weighted least squares regression model

By fitting a weighted least squares regression model for uric vs trig, alco, dia, choles we obtained the following estimates for the regression parameters.

Variables	WLS	1st Iteration	2nd Iteration	3rd iteration
Intercept	77.79161	75.26328	75.07199	75.04400
Trig	86.59896	85.84075	85.76030	85.74604
alco	0.43564	0.44443	0.44709	0.44781
dia	1.68099	1.69389	1.69546	1.69580
choles	-3.85856	-3.48379	-3.46718	-3.46712

From the parameter estimates, we can see that the iterating process of estimating weights improve the estimates.

Question5

Iteratively Reweighted Least Squares approach to robust regression for dampening the influence of outlying cases.

IRLS was carried out for the model uric vs trig, alco, dia, choles. Using Bisquare weight function, the robust regression gave the following parameter estimates for three iterations. We used the mean absolute deviation (MAD) to estimate since it is robust to outliers.

Variables		1st Iteration	2nd Iteration	3rd iteration
Intercept	86.58116	85.75681	85.85647	85.96338
Trig	94.75207	92.55702	91.40628	90.82569
alco	0.37386	0.35124	0.34320	0.34024
dia	1.64759	1.68939	1.69944	1.70208
choles	-7.06153	-716299	-7.11892	-7.06758

From the parameter estimates, we can see that the robust regression improves the model.

Obs	Res1	wt1	Res2	wt2	Res3	wt3	Res4	wt4	Res5
1	-81.973	0.90522	-69.906	0.93120	-64.502	0.94087	-61.865	0.94557	-60.588
2	-25.119	0.99090	-15.492	0.99656	-12.337	0.99781	-11.022	0.99825	-10.447
3	-92.576	0.87994	-88.190	0.89165	-86.703	0.89447	-86.033	0.89611	-85.728
4	-95.347	0.87290	-94.836	0.87527	-93.509	0.87782	-92.570	0.88024	-92.045
5	132.925	0.76088	138.368	0.74445	140.686	0.73481	141.758	0.73120	142.243

Above is an output contained residuals and weights for 5 observations to compare the residuals and weights.

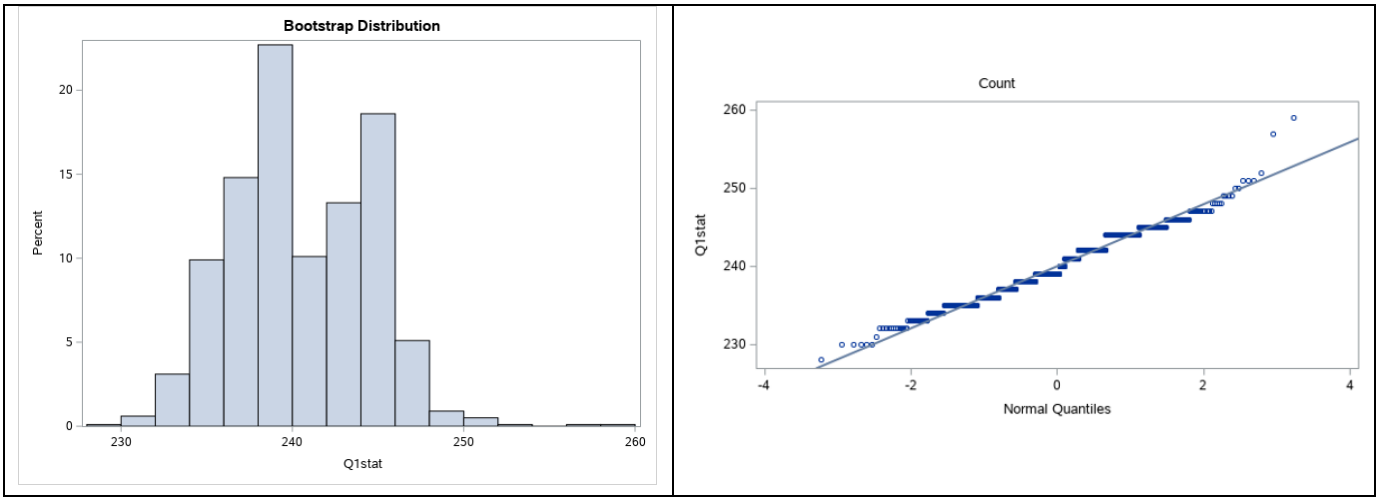
As number of iterations becomes higher the values of residuals and weights are getting closer and closer.

Bonus Question

Perform inference on the first quartile of uric acid ($\hat{\theta}$).

First quartile from the original sample $\hat{\theta} = 239$.

histogram and Q-Q plot of the bootstrap distribution of $\hat{\theta}$



Shapiro wilks test gives p-value <0.0001 suggests that Distribution is not normal.

Mean of $\hat{\theta}$ $= 239.986$

bias $= \widehat{B^*}$ =239.986 -239 =0.986

standard error of $\hat{\theta}$ $= \widehat{SE}^* = 3.9710922$

2.5th percentile of $\hat{\theta}$ $= \widehat{\theta^*_{\frac{\alpha}{2}}} = 233$

97.5th percentile of $\hat{\theta}$ $= \widehat{\theta^*_{1-\frac{\alpha}{2}}}$ =247

2.5th percentile of $\hat{\theta} - \theta$ = -6

97.5th percentile of $\hat{\theta} - \theta$ $= 8$

95% confidence interval for θ using three bootstrap methods

1. normal approximation-
   CI: $(\hat{\theta} - \widehat{B^*}) \mp Z_{1-\alpha/2}\widehat{SE}^* = (230.2306, 245.7973)$

2. basic bootstrap-
   CI: $(2\hat{\theta} - \widehat{\theta^*_{1-\frac{\alpha}{2}}}\ , 2\hat{\theta} - \widehat{\theta^*_{\frac{\alpha}{2}}}) = (231, 245)$

3. percentile bootstrap-
   CI: $(\widehat{\theta^*_{\frac{\alpha}{2}}}\ , \widehat{\theta^*_{1-\frac{\alpha}{2}}}) = (233\ ,\ 247\ )$

SAS OUTPUTS

Question 1

Full model for predicting uric acid levels using all other explanatory variables

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10118200	2023640	217.34	<.0001
Error	992	9236375	9310.86164		
Corrected Total	997	19354575			

Root MSE	96.49281	R-Square	0.5228
Dependent Mean	330.10120	Adj R-Sq	0.5204
Coeff Var	29.23128		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	92.04641	24.36508	3.78	0.0002
dia	1	1.42445	0.22632	6.29	<.0001
hdl	1	4.59383	7.72338	0.59	0.5521
choles	1	-6.45949	2.62444	-2.46	0.0140
trig	1	99.70139	4.16454	23.94	<.0001
alco	1	0.42497	0.04156	10.23	<.0001

The p-values of ANOVA F test show that the model is significant.

Test if the variables hdl and choles can be (jointly) dropped together from the full model.

Test 1 Results for Dependent Variable uric				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	28356	3.05	0.0480
Denominator	992	9310.86164		

Question 2

<u>Selecting the best model using adjusted R^2 criterion:</u>

Number in Model	Adjusted R-Square	R-Square	Variables in Model
4	0.5207	0.5226	dia choles trig alco
5	0.5204	0.5228	dia hdl choles trig alco
3	0.5184	0.5199	dia trig alco
4	0.5179	0.5199	dia hdl trig alco
3	0.5018	0.5033	choles trig alco
4	0.5017	0.5037	hdl choles trig alco
2	0.5012	0.5022	trig alco
3	0.5009	0.5024	hdl trig alco
4	0.4704	0.4725	dia hdl choles trig
3	0.4651	0.4667	dia choles trig
3	0.4635	0.4651	dia hdl trig
2	0.4599	0.4610	dia trig
3	0.4398	0.4415	hdl choles trig
2	0.4358	0.4370	hdl trig
2	0.4313	0.4325	choles trig
1	0.4290	0.4296	trig
4	0.2440	0.2471	dia hdl choles alco
3	0.2297	0.2320	dia hdl alco
3	0.1964	0.1988	hdl choles alco
3	0.1815	0.1839	dia choles alco
2	0.1692	0.1709	hdl alco
2	0.1688	0.1705	dia alco
3	0.1347	0.1373	dia hdl choles
2	0.1326	0.1343	choles alco
2	0.1263	0.1281	dia hdl
1	0.1076	0.1085	alco
2	0.1005	0.1023	dia choles
1	0.0926	0.0935	dia
2	0.0515	0.0534	hdl choles
1	0.0301	0.0310	hdl
1	0.0205	0.0215	choles

Selecting the best model using stepwise selection method:

Stepwise Selection: Step 1

Variable trig Entered: R-Square = 0.4296 and C(p) = 191.6626

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8315035	8315035	750.19	<.0001
Error	996	11039540	11084		
Corrected Total	997	19354575			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	195.54397	5.93639	12026383	1085.03	<.0001
trig	104.17280	3.80337	8315035	750.19	<.0001

Stepwise Selection: Step 2

Variable alco Entered: R-Square = 0.5022 and C(p) = 42.8678

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9719064	4859532	501.81	<.0001
Error	995	9635511	9683.93035		
Corrected Total	997	19354575			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	186.11152	5.60387	10681251	1102.99	<.0001
trig	100.15497	3.57070	7618879	786.75	<.0001
alco	0.48223	0.04005	1404029	144.99	<.0001

Stepwise Selection: Step 3

Variable dia Entered: R-Square = 0.5199 and C(p) = 8.0909

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10061489	3353830	358.73	<.0001
Error	994	9293086	9349.18098		
Corrected Total	997	19354575			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	73.23217	19.44747	132572	14.18	0.0002
dia	1.35686	0.22420	342425	36.63	<.0001
trig	96.07760	3.57254	6761821	723.25	<.0001
alco	0.44094	0.03994	1139633	121.90	<.0001

Stepwise Selection: Step 4

Variable choles Entered: R-Square = 0.5226 and C(p) = 4.3538

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10114906	2528727	271.77	<.0001
Error	993	9239669	9304.80238		
Corrected Total	997	19354575			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.23003	22.02861	185021	19.88	<.0001
dia	1.43222	0.22587	374119	40.21	<.0001
choles	-6.19569	2.58585	53417	5.74	0.0168
trig	98.58245	3.71421	6554998	704.47	<.0001
alco	0.43157	0.04003	1081310	116.21	<.0001

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	trig		1	0.4296	0.4296	191.663	750.19	<.0001
2	alco		2	0.0725	0.5022	42.8678	144.99	<.0001
3	dia		3	0.0177	0.5199	8.0909	36.63	<.0001
4	choles		4	0.0028	0.5226	4.3538	5.74	0.0168

Question 3

Chosen model

$$\text{uric} = \beta_0 + \beta_1 \text{trig} + \beta_2 \text{alco} + \beta_3 \text{dia} + \beta_4 \text{choles}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10114906	2528727	271.77	<.0001
Error	993	9239669	9304.80238		
Corrected Total	997	19354575			

Root MSE	96.46140	R-Square	0.5226
Dependent Mean	330.10120	Adj R-Sq	0.5207
Coeff Var	29.22177		

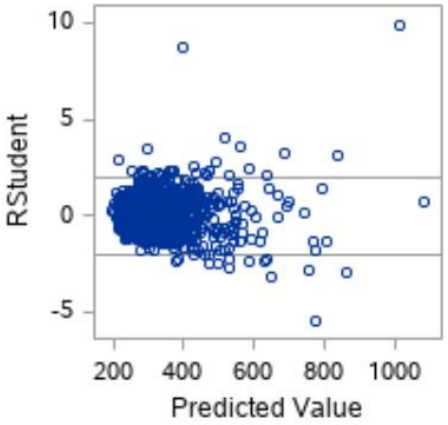
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	98.23003	22.02861	4.46	<.0001
trig	1	98.58245	3.71421	26.54	<.0001
alco	1	0.43157	0.04003	10.78	<.0001
dia	1	1.43222	0.22587	6.34	<.0001
choles	1	-6.19569	2.58585	-2.40	0.0168

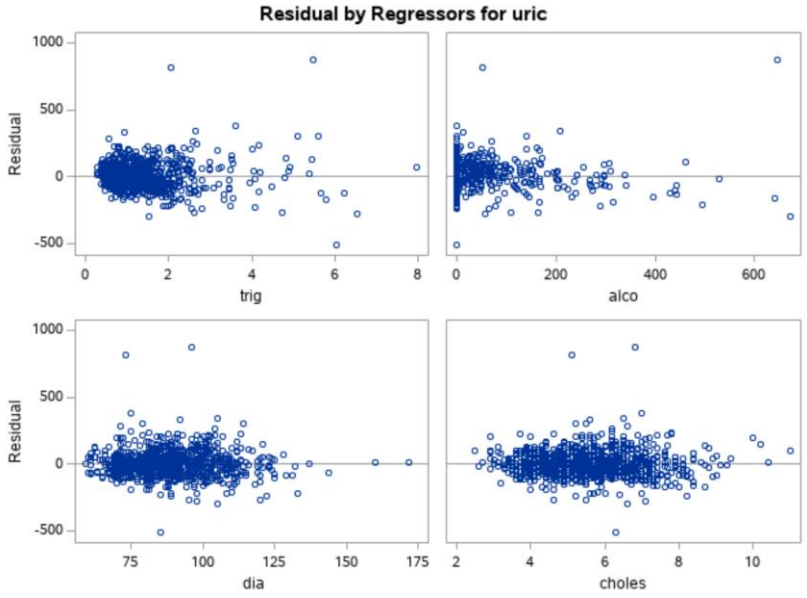
Checking the assumptions of simple linear regression model

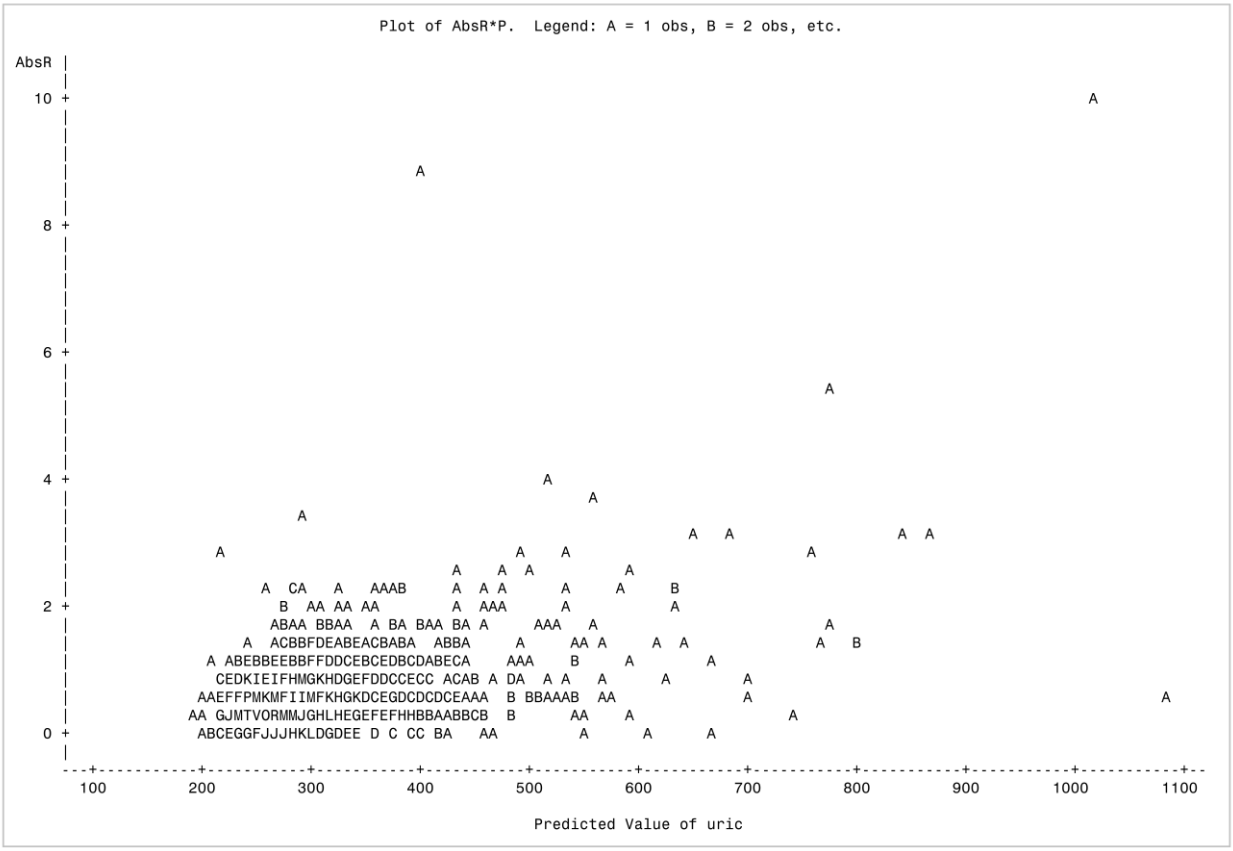
Checking linearity of the model (Lack of fit test)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10114906	2528727	271.77	<.0001
Error	993	9239669	9304.80238		
Lack of Fit	992	9214581	9288.89190	0.37	0.8994
Pure Error	1	25088	25088		
Corrected Total	997	19354575			

Checking the Constant variance







Brown-Forsythe test and Breusch-Pagan test

Heteroscedasticity Test					
Equation	**Test**	**Statistic**	**DF**	**Pr > ChiSq**	**Variables**
uric	White's Test	411.5	14	<.0001	Cross of all vars
	Breusch-Pagan	129.7	4	<.0001	trig, alco, dia, choles, 1

Trig

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
Group	1	29.1857	29.1857	58.80	<.0001
Error	996	494.4	0.4964		

Alco

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
Group	1	10.5983	10.5983	20.87	<.0001
Error	996	505.8	0.5078		

Dia

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
Group	1	7.9053	7.9053	15.36	<.0001
Error	996	512.7	0.5147		

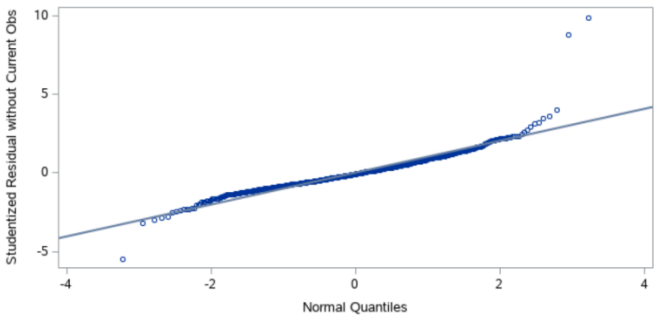
Choles

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
Group	1	3.4435	3.4435	6.63	0.0102
Error	996	517.0	0.5191		

Checking Normality



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.899635	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.065026	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.489844	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	9.503758	Pr > A-Sq	<0.0050

Detecting Outliers and Influential Observations:

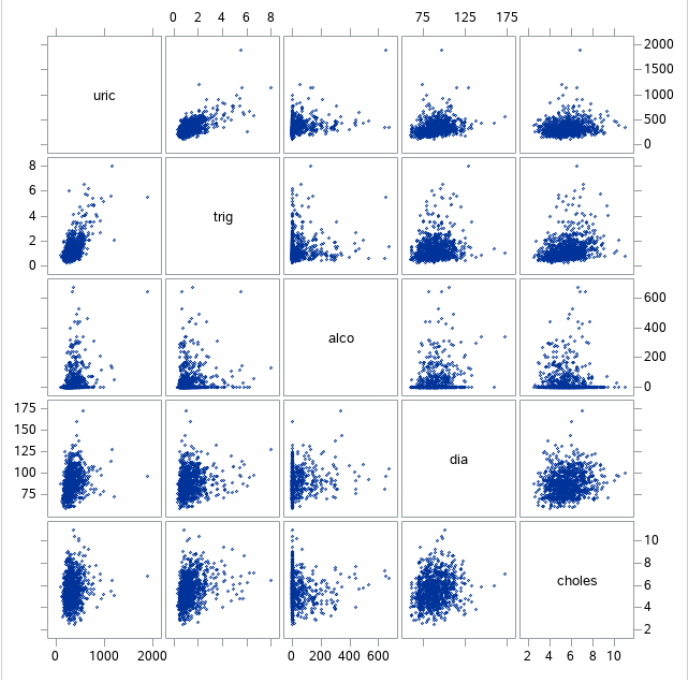
Bonferroni Method

Obs	RStudent
267	8.7634
477	-5.5039
483	9.8816

SAS output for only 5 observations

Obs	HatDiagonal	hilev	DFFITS	dfflag	CooksD	Fpercent	DFB_trig	b1flag	DFB_alco	b2flag	DFB_dia	b3flag	DFB_choles	b4flag
5	0.0080	0	0.1241	0	0.003	0.000156	0.0133	0	-0.0199	0	0.0318	0	0.0926	1
7	0.0084	0	0.3325	1	0.022	0.020253	0.1206	1	0.2345	1	0.0568	0	0.0423	0
8	0.0072	0	0.1937	1	0.007	0.001420	0.0096	0	0.1157	1	0.0352	0	0.1104	1
10	0.0050	0	0.0824	0	0.001	0.000020	0.0636	1	-0.0140	0	0.0076	0	0.0097	0
11	0.0087	0	-0.1577	1	0.005	0.000514	-0.1248	1	0.0254	0	0.0281	0	-0.0391	0

Collinearity diagnostic

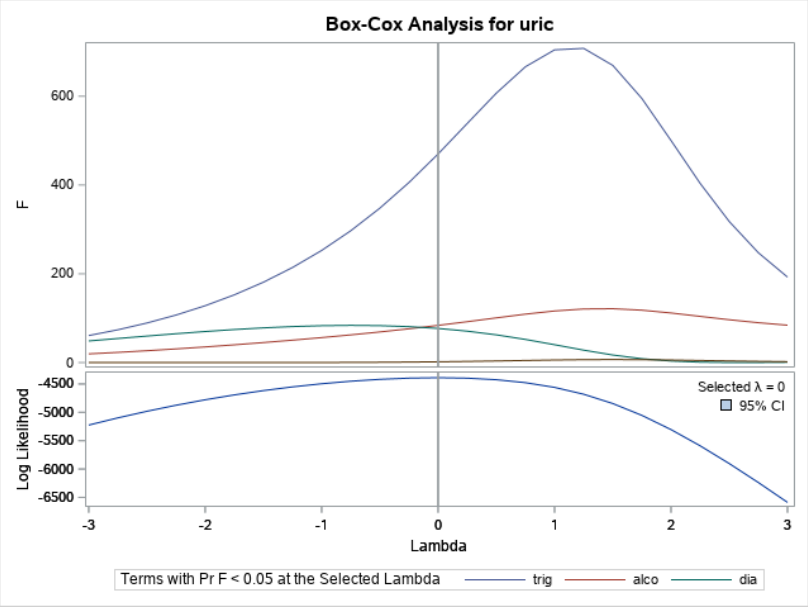


Pearson Correlation Coefficients, N = 998 Prob > |r| under H0: Rho=0					
	uric	trig	alco	dia	choles
uric	1.00000	0.65545 <.0001	0.32941 <.0001	0.30571 <.0001	0.14660 <.0001
trig	0.65545 <.0001	1.00000	0.09345 0.0031	0.20183 <.0001	0.30137 <.0001
alco	0.32941 <.0001	0.09345 0.0031	1.00000	0.18544 <.0001	-0.04235 0.1813
dia	0.30571 <.0001	0.20183 <.0001	0.18544 <.0001	1.00000	0.17675 <.0001
choles	0.14660 <.0001	0.30137 <.0001	-0.04235 0.1813	0.17675 <.0001	1.00000

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	98.23003	22.02861	4.46	<.0001	.	0
trig	1	98.58245	3.71421	26.54	<.0001	0.88027	1.13601
alco	1	0.43157	0.04003	10.78	<.0001	0.95316	1.04914
dia	1	1.43222	0.22587	6.34	<.0001	0.91321	1.09503
choles	1	-6.19569	2.58585	-2.40	0.0168	0.88663	1.12787

Collinearity Diagnostics							
Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	trig	alco	dia	choles
1	3.90256	1.00000	0.00120	0.01479	0.01229	0.00145	0.00266
2	0.82275	2.17792	0.00039544	0.00309	0.94699	0.00029376	0.00118
3	0.23163	4.10464	0.00809	0.93269	0.00111	0.00767	0.00848
4	0.03147	11.13666	0.04167	0.02712	0.02703	0.19314	0.89848
5	0.01160	18.34382	0.94864	0.02231	0.01258	0.79744	0.08920

Transformation

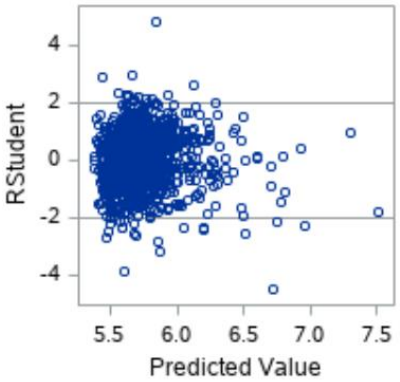


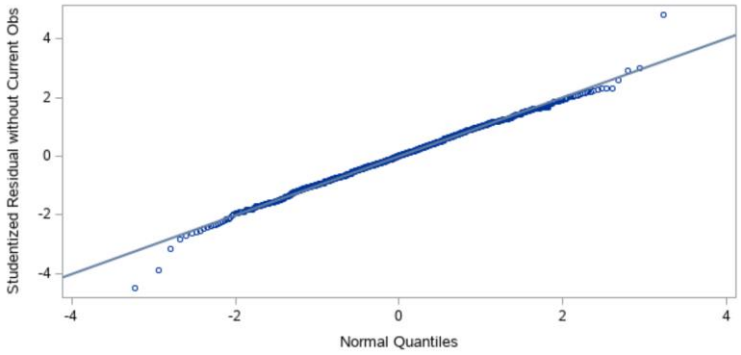
Log transformation

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.99163	0.06022	82.89	<.0001
trig	1	0.22011	0.01015	21.68	<.0001
alco	1	0.00100	0.00010944	9.15	<.0001
dia	1	0.00541	0.00061746	8.77	<.0001
choles	1	-0.00927	0.00707	-1.31	0.1901

Checking assumptions of log models

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	58.61986	14.65496	210.75	<.0001
Error	993	69.04953	0.06954		
Lack of Fit	992	68.80137	0.06936	0.28	0.9412
Pure Error	1	0.24816	0.24816		
Corrected Total	997	127.66939			



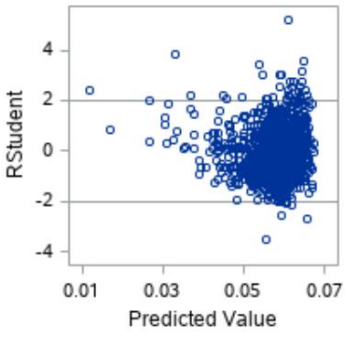


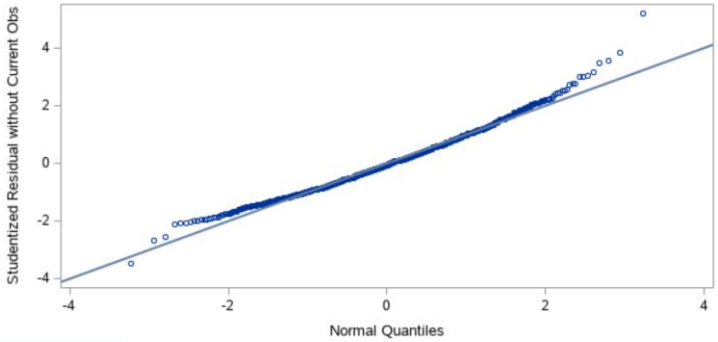
Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
loguric	White's Test	38.98	14	0.0004	Cross of all vars
	Breusch-Pagan	23.72	4	<.0001	trig, alco, dia, choles, 1

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994783	Pr < W	0.0016
Kolmogorov-Smirnov	D	0.013841	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.032337	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.317733	Pr > A-Sq	>0.2500

Inverse sqrt transformation

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.04095	0.01024	171.32	<.0001
Error	993	0.05935	0.00005976		
Lack of Fit	992	0.05914	0.00005962	0.30	0.9335
Pure Error	1	0.00020125	0.00020125		
Corrected Total	997	0.10030			





Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.984421	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.032528	Pr > D	0.0113
Cramer-von Mises	W-Sq	0.360824	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.535256	Pr > A-Sq	<0.0050

Question 4

Linear regression model

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	98.23003	22.02861	4.46	<.0001	55.00205	141.45801
trig	1	98.58245	3.71421	26.54	<.0001	91.29384	105.87106
alco	1	0.43157	0.04003	10.78	<.0001	0.35301	0.51013
dia	1	1.43222	0.22587	6.34	<.0001	0.98898	1.87546
choles	1	-6.19569	2.58585	-2.40	0.0168	-11.27005	-1.12134

Weighted least square

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	77.79161	17.32962	4.49	<.0001	43.78473	111.79850
trig	1	86.59896	5.12979	16.88	<.0001	76.53249	96.66542
alco	1	0.43564	0.05536	7.87	<.0001	0.32701	0.54428
dia	1	1.68099	0.18583	9.05	<.0001	1.31633	2.04564
choles	1	-3.85856	2.18272	-1.77	0.0774	-8.14184	0.42472

1st iteration

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	75.26328	17.26337	4.36	<.0001	41.38640	109.14017
trig	1	85.84075	5.15626	16.65	<.0001	75.72233	95.95917
alco	1	0.44443	0.05730	7.76	<.0001	0.33198	0.55688
dia	1	1.69389	0.18665	9.08	<.0001	1.32761	2.06017
choles	1	-3.48379	2.16356	-1.61	0.1077	-7.72947	0.76189

2nd iteration

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	75.07199	17.25729	4.35	<.0001	41.20704	108.93694
trig	1	85.76030	5.15525	16.64	<.0001	75.64386	95.87673
alco	1	0.44709	0.05785	7.73	<.0001	0.33357	0.56060
dia	1	1.69546	0.18677	9.08	<.0001	1.32894	2.06197
choles	1	-3.46718	2.16264	-1.60	0.1092	-7.71104	0.77669

3rd iteration

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	75.04400	17.25648	4.35	<.0001	41.18064	108.90735
trig	1	85.74604	5.15447	16.64	<.0001	75.63114	95.86094
alco	1	0.44781	0.05799	7.72	<.0001	0.33401	0.56161
dia	1	1.69580	0.18680	9.08	<.0001	1.32922	2.06237
choles	1	-3.46712	2.16262	-1.60	0.1092	-7.71095	0.77671

Question 5

Linear regression model

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	98.23003	22.02861	4.46	<.0001	55.00205	141.45801
trig	1	98.58245	3.71421	26.54	<.0001	91.29384	105.87106
alco	1	0.43157	0.04003	10.78	<.0001	0.35301	0.51013
dia	1	1.43222	0.22587	6.34	<.0001	0.98898	1.87546
choles	1	-6.19569	2.58585	-2.40	0.0168	-11.27005	-1.12134

0th iteration

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	86.58116	16.80564	5.15	<.0001
trig	1	94.75207	3.15925	29.99	<.0001
alco	1	0.37386	0.03377	11.07	<.0001
dia	1	1.64756	0.17291	9.53	<.0001
choles	1	-7.06153	1.99899	-3.53	0.0004

1st iteration

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	85.75681	16.80326	5.10	<.0001
trig	1	92.55702	3.16049	29.29	<.0001
alco	1	0.35124	0.03302	10.64	<.0001
dia	1	1.68939	0.17283	9.77	<.0001
choles	1	-7.16299	1.99801	-3.59	0.0004

2nd iteration

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	85.85647	16.77745	5.12	<.0001
trig	1	91.40628	3.16122	28.91	<.0001
alco	1	0.34320	0.03275	10.48	<.0001
dia	1	1.69944	0.17255	9.85	<.0001
choles	1	-7.11892	1.99520	-3.57	0.0004

3rd iteration

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	85.96338	16.77367	5.12	<.0001
trig	1	90.82569	3.16304	28.71	<.0001
alco	1	0.34024	0.03267	10.42	<.0001
dia	1	1.70208	0.17252	9.87	<.0001
choles	1	-7.06758	1.99487	-3.54	0.0004

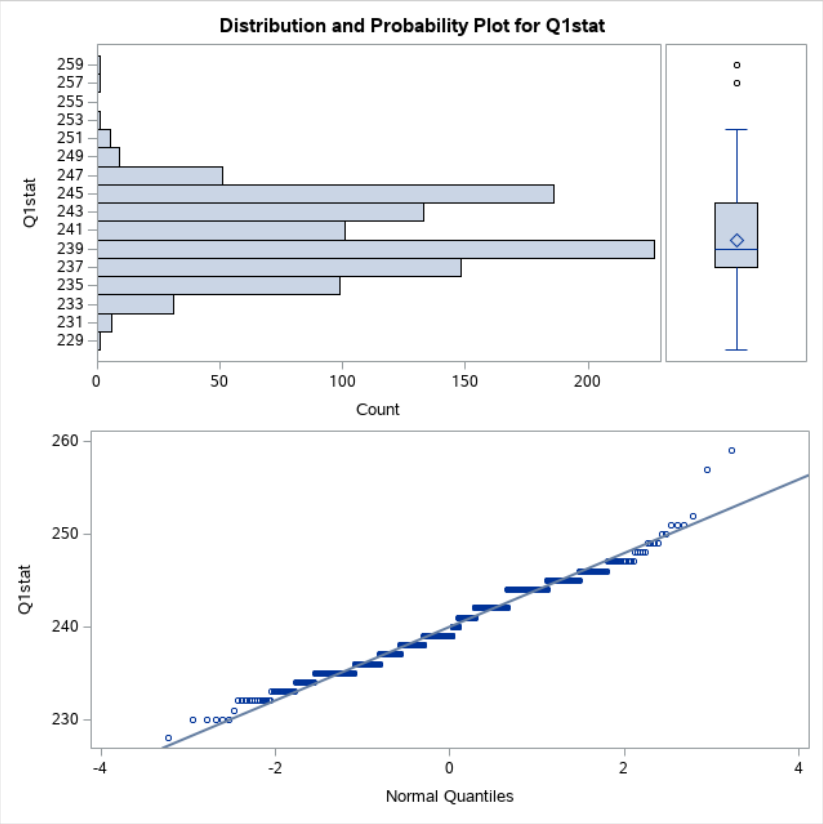
Printing 5 observations to compare the residuals and weights

Obs	Res1	wt1	Res2	wt2	Res3	wt3	Res4	wt4	Res5
1	-81.973	0.90522	-69.906	0.93120	-64.502	0.94087	-61.865	0.94557	-60.588
2	-25.119	0.99090	-15.492	0.99656	-12.337	0.99781	-11.022	0.99825	-10.447
3	-92.576	0.87994	-88.190	0.89165	-86.703	0.89447	-86.033	0.89611	-85.728
4	-95.347	0.87290	-94.836	0.87527	-93.509	0.87782	-92.570	0.88024	-92.045
5	132.925	0.76088	138.368	0.74445	140.686	0.73481	141.758	0.73120	142.243

QUESTION 6 (Bonus Question)

Basic Statistical Measures			
Location		Variability	
Mean	239.9860	Std Deviation	3.97109
Median	239.0000	Variance	15.76957
Mode	242.0000	Range	31.00000
		Interquartile Range	7.00000

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.978024	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.110047	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.432212	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	8.13021	Pr > A-Sq	<0.0050



The MEANS Procedure

Analysis Variable : Q1stat			
Mean	Std Dev	Lower Quartile	Upper Quartile
239.9860000	3.9710922	237.0000000	244.0000000

2.5th and 97.5th percentile of $\hat{\theta}$

Bootstrap Distribution

CI95_Lower	CI95_Upper
233	247

2.5th and 97.5th percentile of $\hat{\theta} - \theta$

CI95_Lower	CI95_Upper
-6	8

SAS CODES

```
filename cardio '/folders/myfolders/Project3/cardio.csv';

DATA cardio;
INFILE cardio DSD FIRSTOBS = 2 ;
INPUT uric dia hdl choles trig alco;
RUN;

PROC PRINT DATA=cardio (OBS=10);
RUN;

*.....;
```

/*Question 1*/

```
PROC REG Data=cardio;
MODEL uric = dia hdl choles trig alco; /*Fit a full model for
predicting uric acid level using other explanatory variables*/
TEST hdl=choles=0 /*Test if the variables hdl and choles can be
(jointly)dropped together from the full model */;
RUN;

*.....;
```

/*Question 2*/

```
/* Selection based on Adj R^2 */
PROC REG DATA = cardio;
MODEL uric = dia hdl choles trig alco / selection = adjrsq rsquare;
RUN;

/* Stepwise selection */
PROC REG DATA = cardio;
MODEL uric = dia hdl choles trig alco / selection = stepwise;
RUN;

*.....;
```

/*Question 3*/

```
/* Regression model for selected variables */
PROC REG Data=cardio;
MODEL uric = trig alco dia choles;
OUTPUT OUT=D RSTUDENT=R PREDICTED=P;
RUN;

/* Lack of fit test for linearity*/
```

```
PROC REG DATA=cardio;
MODEL uric = trig alco dia choles / lackfit ;
RUN;

/* Brown Forsythe Test for homogeneity*/
PROC MEANS DATA=cardio median; /* Get Median for variables*/
VAR trig alco dia choles;
RUN;

/* BF for variable trig*/
DATA D; SET D;
Group = (trig> 1.03);
RUN;

PROC GLM Data=D;
        class Group;
        model R=Group;
        means Group / hovtest=BF;
run;

/* BF for variable alco*/
DATA D; SET D;
Group = (alco> 0);
RUN;

PROC GLM Data=D;
        class Group;
        model R=Group;
        means Group / hovtest=BF;
run;

/* BF for variable dia*/
DATA D; SET D;
Group = (dia> 87);
RUN;

PROC GLM Data=D;
        class Group;
        model R=Group;
        means Group / hovtest=BF;
run;

/* BF for variable choles*/
DATA D; SET D;
Group = (choles> 5.5);
RUN;

PROC GLM Data=D;
```

```
        class Group;
        model R=Group;
        means Group / hovtest=BF;
run;

/* Breusch Pagan Test for homogeneity */
PROC MODEL DATA=D;
PARMS b0 b1 b2 b3 b4;
PSA=b0+b1*trig+b2*alco+b3*dia+b4*choles;
fit PSA /WHITE BREUSCH=(trig alco dia choles);
fit PSA /BREUSCH=(trig);
fit PSA /BREUSCH=(alco);
fit PSA /BREUSCH=(dia);
fit PSA /BREUSCH=(choles);
RUN;

 /* absolute residuals vs fitted values to check homogeneity
 assumption */

DATA D; SET D;
AbsR=ABS(R);

PROC PLOT DATA = D;
PLOT AbsR*P;
RUN;

/* Check normality of Studentized residuals */
PROC UNIVARIATE DATA=D NORMAL PLOT;
VAR R;
RUN;

/* Tests for outliers*/

/*calcutate hat matrix,Coock's distance, DFFITS and DFBETAS*/
PROC REG Data=cardio;
MODEL uric = trig alco dia choles / INFLUENCE R;
ods output outputstatistics=results;
RUN;

PROC PRINT Data=results (obs=5);
RUN;

/* Test for outliers using Bonferroni method */
DATA results; set results;
n=998; p= 5; alpha=0.05;
tvalue = tinv(1-alpha/(2*n), n-p-1);
if (abs(RStudent)) > tvalue then outlier=1;
else outlier=0;
```

```
RUN;

PROC PRINT data=results;
where outlier=1;
var RStudent;
RUN;

/* Test for outliers using R^2,hat matrix,Coock's distance,DFFITS and DFBETAS */
DATA results; SET results;

/* Checking if hii > 2*p/n */
if HatDiagonal > 2*(p/n) then hilev=1;
else hilev=0;

/* Checking if DFFITS > 2*sqrt(p/n) */
if (abs(DFFITS) > 2*sqrt(p/n)) then dfflag=1;
else dfflag=0;

/* Calculating percentile for each Cook's D value using F(p,n-p) */
Fpercent = 100*probf(CooksD, p, n-p);

/* Checking if each DFBETAS value > 1 */
if (abs(dfb_trig) > 1) then b1flag=1;
else b1flag=0;
if (abs(dfb_alco) > 1) then b2flag=1;
else b2flag=0;
if (abs(dfb_dia) > 1) then b3flag=1;
else b3flag=0;
if (abs(dfb_choles) > 1) then b4flag=1;
else b4flag=0;
RUN;

PROC PRINT DATA = results (obs=5);
where hilev=1 or dfflag=1 or Fpercent>20 or b1flag=1 or b2flag=1 or b3flag=1 or b4flag=1;
var HatDiagonal hilev DFFITS dfflag CooksD Fpercent dfb_trig b1flag dfb_alco b2flag dfb_dia
b3flag dfb_choles b4flag;
RUN;

PROC PRINT DATA = results;
where Fpercent>20;
var  Fpercent ;
RUN;

PROC PRINT DATA = results;
where dfflag=1;
var  DFFITS dfflag ;
RUN;
```

```
/* Collinearity diagnostics */
proc SGscatter data=cardio;
matrix uric trig alco dia choles;
run;

proc corr data=cardio plots=matrix;
var uric trig alco dia choles;
run;

PROC REG Data=cardio;
MODEL uric = trig alco dia choles / collin tol vif;
RUN;

/*Transformations*/

/* Find Box-Cox transformation power */
PROC TRANSREG DATA=cardio;
MODEL BoxCox(uric)=identity(trig alco dia choles);
RUN;

/* Tring log transformation */
DATA D; SET D;
loguric = log(uric);
RUN;

PROC REG Data=D;
MODEL loguric = trig alco dia choles/LACKFIT DWPROB;
OUTPUT OUT=E RSTUDENT=Rlog PREDICTED=Plog R=Res1;
RUN;

PROC MODEL DATA=D; /* Test for homogeneity assumption */
PARMS b0 b1 b2 b3 b4;
loguric = b0 + b1*trig + b2*alco + b3*dia + b4*choles;
fit loguric /WHITE BREUSCH=( trig alco dia choles);
RUN;

PROC UNIVARIATE DATA=E NORMAL PLOT; /* Checking normality of Studentized residuals
*/
VAR Rlog;
RUN;

/* Try INVSQRT transformation */
DATA D; SET D;
invsqrturic = 1/sqrt(uric);
RUN;
PROC REG Data=D;
MODEL invsqrturic = trig alco dia choles/LACKFIT DWPROB;
OUTPUT OUT=F RSTUDENT=Rsqinv PREDICTED=Psqinv;
```

```
RUN;

PROC MODEL DATA=D; /* Test for homogeneity assumption */
PARMS b0 b1 b2 b3 b4;
invsqrturic = b0 + b1*trig + b2*alco + b3*dia + b4*choles;
fit invsqrturic /WHITE BREUSCH=( trig alco dia choles);
RUN;

PROC UNIVARIATE DATA=F NORMAL PLOT; /* Checking normality of Studentized residuals
*/
VAR Rsqinv;
RUN;

*.....;
```

/*Qestion 4*/

```
PROC REG Data = cardio;
MODEL uric = trig alco dia choles /R clb;
output out=results r =residual;
RUN;

DATA Step2;
SET results;
absresid = abs(residual);
RUN;

PROC PRINT DATA=Step2(obs=5);
RUN;

PROC REG Data = Step2;
MODEL absresid = trig alco dia choles/p; /* option p requests fitted values */
output out = Step3 p =ehat;
RUN;

DATA STEP3;
SET Step3;
wt = 1/(ehat**2);
RUN;

PROC PRINT DATA=STEP3 (obs=5);
RUN;

PROC REG Data=Step3; /* weighted least squares regression */
MODEL uric = trig alco dia choles/R clb;
WEIGHT wt;
output out=iteration2 r =residual2;
RUN;
```

```
PROC PRINT DATA=iteration2 (obs=5);
RUN;

/* Reiterate the process - Iteratively reweighted least squares */

DATA iteration2;
SET iteration2;
absresid2 = abs(residual2);
RUN;

PROC PRINT DATA=iteration2 (obs=5);
RUN;

PROC REG Data = iteration2;
MODEL absresid2 = trig alco dia choles/p; /* option p requests fitted values */
output out = results2 p =ehat2;
RUN;
PROC PRINT DATA=results2 (obs=5);
RUN;

DATA results2;
SET results2;
wt2 = 1/(ehat2**2);
RUN;
PROC PRINT DATA=results2 (obs=5);
RUN;

PROC REG Data=results2; /* weighted least squares regression */
MODEL uric = trig alco dia choles/R clb;
WEIGHT wt2;
output out=iteration3 r =residual3;
RUN;
PROC PRINT DATA=iteration3 (obs=5);
RUN;

/* Reiterate the process - Iteratively reweighted least squares */

DATA iteration3;
SET iteration3;
absresid3 = abs(residual3);
RUN;

PROC PRINT DATA=iteration3 (obs=5);
RUN;

PROC REG Data = iteration3;
MODEL absresid3 = trig alco dia choles/p; /* option p requests fitted values */
output out = results3 p =ehat3;
```

```
RUN;
PROC PRINT DATA=results3 (obs=5);
RUN;

DATA results3;
SET results3;
wt3 = 1/(ehat3**2);
RUN;
PROC PRINT DATA=results3 (obs=5);
RUN;

PROC REG Data=results3; /* weighted least squares regression */
MODEL uric = trig alco dia choles/R clb;
WEIGHT wt3;
output out=iteration4 r =residual4;
RUN;
PROC PRINT DATA=iteration4(obs=5);
RUN;

/* Reiterate the process - Iteratively reweighted least squares */

DATA iteration4;
SET iteration4;
absresid4 = abs(residual4);
RUN;

PROC PRINT DATA=iteration4 (obs=5);
RUN;

PROC REG Data = iteration4;
MODEL absresid4 = trig alco dia choles/p; /* option p requests fitted values */
output out = results4 p =ehat4;
RUN;
PROC PRINT DATA=results4 (obs=5);
RUN;

DATA results4;
SET results4;
wt4 = 1/(ehat4**2);
RUN;
PROC PRINT DATA=results4 (obs=5);
RUN;

PROC REG Data=results4; /* weighted least squares regression */
MODEL uric = trig alco dia choles/R clb;
WEIGHT wt4;
output out=iteration5 r =residual5;
RUN;
```

```
PROC PRINT DATA=iteration5(obs=5);
RUN;

*.....;
```

/*Qestion 5*/

```
filename cardio '/folders/myfolders/Project3/cardio.csv';

DATA cardio;
INFILE cardio DSD FIRSTOBS = 2 ;
INPUT uric dia hdl choles trig alco;
RUN;

PROC PRINT DATA=cardio (OBS=10);
RUN;

PROC REG Data = cardio;
MODEL uric = trig alco dia choles /R clb;
output out=E r =Res1;
RUN;


PROC MEANS Data=E Median;
Var Res1;
RUN;

DATA E;
Set E;
AD1=abs(Res1+7.6743106);
RUN;

PROC MEANS Data=E Median;
Var AD1;
RUN;

DATA E;
Set E;
MAD1=53.5500285/0.6745;
u1=Res1/MAD1; /* Calculating scaled residuals */
If abs(u1) le 4.685 then wt1=(1-(u1/4.685)**2)**2; Else wt1=0; /* Using Bisquare weight
function */
RUN;

TITLE "Parameter Estimates from 1st Iteration";

PROC REG Data=E;
Model uric = trig alco dia choles / p;
Weight wt1;
```

```
Output Out=E2 R=Res2 P=P2;
RUN;

/* Iteratively Reweighted Least Squares - Second Iteration */
TITLE;
PROC MEANS Data=E2 Median;
Var Res2;
RUN;

DATA E2;
Set E2;
AD2=abs(Res2+5.1140754);
RUN;

PROC MEANS Data=E2 Median;
Var AD2;
RUN;

DATA E2;
Set E2;
MAD2=53.7849334/0.6745;
u2=Res2/MAD2; /* Calculating scaled residuals */
If abs(u2) le 4.685 then wt2=(1-(u2/4.685)**2)**2; Else wt2=0; /* Using Bisquare weight
function */
RUN;

TITLE "Parameter Estimates from 2nd Iteration";
PROC REG Data=E2;
Model uric = trig alco dia choles / P;
Weight wt2;
Output Out=E3 R=Res3 P=P3;
RUN;

/* Iteratively Reweighted Least Squares - Third Iteration */

TITLE;
PROC MEANS Data=E3 Median;
Var Res3;
RUN;

DATA E3;
Set E3;
AD3=abs(Res3+5.0397940);
RUN;

PROC MEANS Data=E3 Median;
Var AD3;
RUN;
```

```
DATA E3;
Set E3;
MAD3=53.6012022/0.6745;
u3=Res3/MAD3; /* Calculating scaled residuals */
If abs(u3) le 4.685 then wt3=(1-(u3/4.685)**2)**2; Else wt3=0; /* Using Bisquare weight
function */
RUN;

TITLE "Parameter Estimates from 3rd Iteration";
PROC REG Data=E3;
Model uric = trig alco dia choles / P;
Weight wt3;
Output Out=E4 R=Res4 P=P3;
RUN;


/* Iteratively Reweighted Least Squares - forth Iteration */

TITLE;
PROC MEANS Data=E4 Median;
Var Res4;
RUN;

DATA E4;
Set E4;
AD4=abs(Res4+5.1657584);
RUN;

PROC MEANS Data=E4 Median;
Var AD4;
RUN;

DATA E4;
Set E4;
MAD4=53.6156608/0.6745;
u4=Res4/MAD4; /* Calculating scaled residuals */
If abs(u4) le 4.685 then wt4=(1-(u4/4.685)**2)**2; Else wt4=0; /* Using Bisquare weight
function */
RUN;

TITLE "Parameter Estimates from 4rd Iteration";
PROC REG Data=E4;
Model uric = trig alco dia choles / P;
Weight wt4;
Output Out=E5 R=Res5 P=P4;
RUN;
```

```
proc print data=E5 (obs=5);
var res1 wt1 res2 wt2 res3 wt3 res4 wt4 res5;
run;
*.....;
```

BONUS QUESTION

```
filename sample '/folders/myfolders/Project3/cardio.csv';

DATA sample;
INFILE sample DSD FIRSTOBS = 2 ;
INPUT uric dia hdl choles trig alco;
RUN;

Proc means data=sample q1;
var uric; /*1st quartile 239.0000000*/
run;

%let NumSamples = 1000;      /* number of bootstrap resamples */
/* 2. Generate many bootstrap samples */
proc surveyselect data=sample NOPRINT seed=98638
    out=BootSSFreq(rename=(Replicate=SampleID))
    method=urs            /* resample with replacement */
    samprate=1            /* each bootstrap sample has N observations */
    /* OUTHITS              option to suppress the frequency var */
    reps=&NumSamples;      /* generate NumSamples bootstrap resamples */
run;

/* 3. Compute the statistic for each bootstrap sample */
proc means data=BootSSFreq noprint;
  by SampleID;
  freq NumberHits;
  var uric;
  output out=OutStats Q1=Q1stat;  /* approx sampling distribution */
run;

proc print data=OutStats (obs=5);
run;

/*Visualize the bootstrap distribution*/
title "Bootstrap Distribution";
proc sgplot data=OutStats;
  histogram Q1stat;
run;

proc univariate data=OutStats normal plot;
var Q1stat;
run;
```

```
Proc means data=OutStats mean std q1 q3;
var Q1stat;
run;

/*mean=239.9860000, std = 3.9710922
*/

/*CI*/
proc univariate data=OutStats noprint;
  var Q1stat;
  output out=Pctl pctlpre =CI95_
       pctlpts =2.5  97.5      /* compute 95% bootstrap confidence interval */
       pctlname=Lower Upper;
run;

proc print data=Pctl noobs; run;

data OutStats;
set outstats;
dif=Q1stat-239;
run;

proc print data=outstats (obs=5);
run;

proc univariate data=OutStats noprint;
  var dif;
  output out=Pct2 pctlpre =CI95_
       pctlpts =2.5  97.5      /* compute 95% bootstrap confidence interval */
       pctlname=Lower Upper;
run;

proc print data=Pct2 noobs; run;
```