Q1) we can see through correlation matrix that Cancer Volume, Weight, and Penetration seems to have correlation values away from 0. It hints that there is a strong linear relationship between the PSA level and the three variables. While the correlation between PSA and Age and Benign doesn’t seem to be that strong. Similarly Scatterplot also points to that result.

Q2) $Y^{\wedge} \approx 17.60290\ X0 + 2.46112X1 + 0.00470X2 - 0.31140X3 + 1.14931X4 + 2.52247X5$ is the required equation as per the Multiple linear regression model mentioned below where Cancer Volume is X1, Weight is X2, Age is X3, Hyperplasia is X4 and Penetration is X5 while Y predictor value is PSA.

Lack of Fit:-

H0:- there is no lack of fit in the simple linear regression model

HA:- there is lack of fit in the simple linear regression model

Since p value is less than .05 and is actually .001 we reject H0 that there is no lack of fit in the simple linear regression model

Residual vs. Predicted Values:-

we can see that all the data points in the residual vs. fitted scatter plot form a fan-shaped pattern. Hence errors are not independent.

Absolute Residuals vs. Fitted Values Plot :

There is no independence of error as we can see that data points begin to scatter everywhere in the plot, and data points form an upward slope fan-shaped pattern. Hence, that homogeneity assumption is not satisfied.

Brown-Forsythe Test

since p-value $\approx 0.0029 \le 0.05$, we must reject H0 meaning we conclude that the given data doesn’t meet the assumption of the homogeneity of variance.

Breush-Pagan Test

Result of Breusch-Pagan test was do not reject Ho because the pvalue=0.009 is small, so Variance Homogeneity doesnot hold.

Similarly Shapiro Wilks value for both error and PSA itself is less than 0.05 and we can clearly see that from QQ plot that data is not normal.

Q3)

Non-normality can be affected by outliers as well as We can see that there is an outlier at obs 94 and 96 and we cannot delete it as it is influential which is proven by below diagram.

I have used log transformation to remedy the situation.

After using log transformation, QQ plot is perfectly normal, Shapiro Wilks Value is above 0.05 and Brown -Forsythe shows homogeneity of variance.

Q4)

We can see from Output of the Type I SS and Type III SS of the Transformed Full Model that Age , Weight and Capsular are not significant as their p-value is 0.1125, 0.0899, 0.0124, and 0.2880 respectively. Hence after removing these variables one by one we arrive at the conclusion that the final model should contain benign and cancer variables.

Q4 Second part) Now we can see if these variables can be dropped from the model or not.

$SSR(X2,X3,X5\ |X1, X4) = SSE(X1, X4\ ) - SSE(\text{All})$

$SSR(X2,X3,X5\ |X1, X4) \approx 64.80167 - 63.42966$

$SSR(X2,X3,X5\ |X1, X4) \approx 1.37201$

Similarly,

$MSR(X2,X3\ |X1,X4, X5) \approx 1.37201\ /3 \approx 0.45734$

$MSE(X1, X2,X3,X4, X5)$ = 63.42966 /91 ≈ 0.69703 (because 97-5-1=91)

F * = 0.45734 /0.69703

= 0.65612

Finding F value $P(F_{3,89})$=.5812

Since $F_{3,89}$<F* we can conclude that $H0$: $\beta2 = \beta3 = \beta5 = 0$.

Hence these variables can be dropped.

Q5)

Here we have $R = \sqrt{R^2} \approx \sqrt{0.4928} \approx 0.7019$.

And $SSE(X1,X4)$ = 64.80167

$SSE(X4)$=124.60324

So $R^2_{Y(1|4)}$= $(SSE(X4) - SSE(X1, X4))/ SSE(X4)$= 0.47994

SS$E(X1)$= 72.60508

So $R^2_{Y(4|1)}$= $(SSE(X1) - SSE(X1, X4))/ SSE(X4)$= 0.10748

If we square them they will be . .69278 and .32784

we do have R^2 as .479 as per SAS output. Thus we have found that the coefficient of simple determination, R^2, is approximately 0.4799, which matches the value of our largest coefficient of partial determination, $R_{Y1|4}^2 \approx 0.4799$.

Hence verified.

Q6)

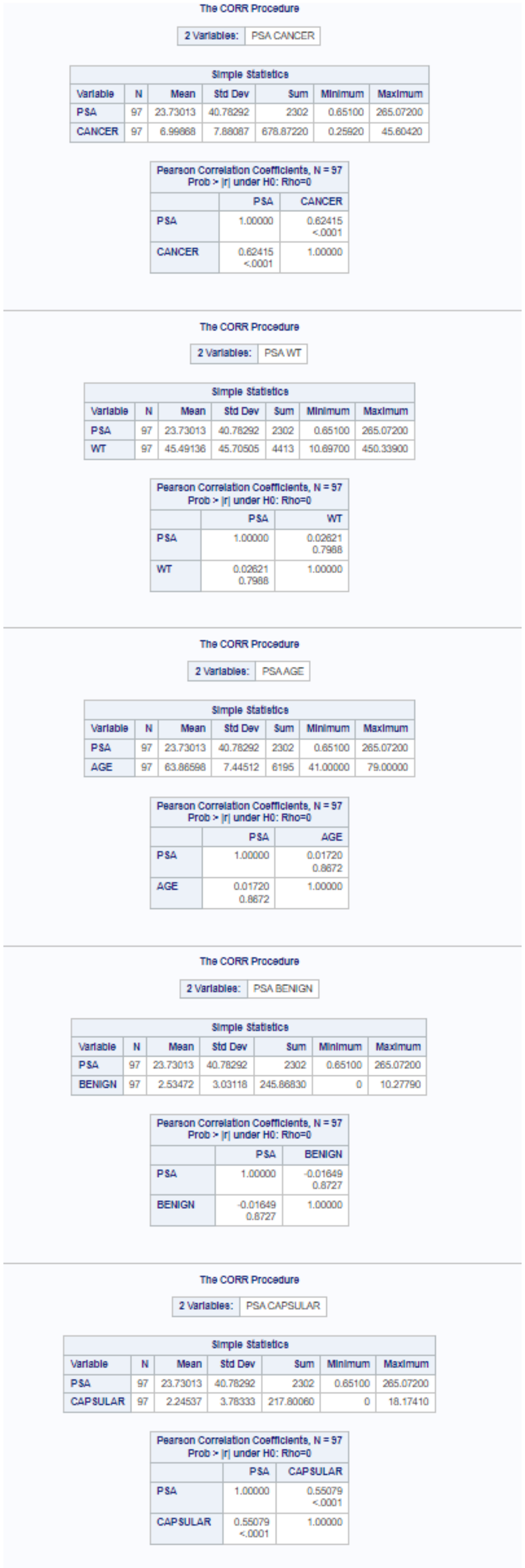
95% interval estimates for the mean response , 95% interval estimates for the mean response for a future observation and 95% interval estimates for the mean response for a future observation has been created. each of the 3 plots, working Hoelting is the largest interval and can take the largest number of possible values. PI is the next largest interval estimate of possible predictions, and CI is the smallest and takes the smallest number of possible values.

Q7) we have Test statistic at 657.0637 while critical value is at 8.06259. Hence we reject H0 meaning that the regression coefficients, $\beta0$, $\beta1$, and $\beta2$, are all not equal to 0.

Simultaneous CI has been attached to output below.

Q8) We know that Bonferroni CI is $b0 \pm t\ (1- \alpha/ 2)*n-p\ S*$sq rt$(((X'X)^{-1})$

Outputs for Q1)



The CORR Procedure

2 Variables:	PSA CANCER

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PSA	97	23.73013	40.78292	2302	0.65100	265.07200
CANCER	97	6.99868	7.88087	678.87220	0.25920	45.60420

Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	PSA	CANCER
PSA	1.00000	0.62415 <.0001
CANCER	0.62415 <.0001	1.00000

The CORR Procedure

2 Variables:	PSA WT

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PSA	97	23.73013	40.78292	2302	0.65100	265.07200
WT	97	45.49136	45.70505	4413	10.69700	450.33900

Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	PSA	WT
PSA	1.00000	0.02621 0.7988
WT	0.02621 0.7988	1.00000

The CORR Procedure

2 Variables:	PSA AGE

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PSA	97	23.73013	40.78292	2302	0.65100	265.07200
AGE	97	63.86598	7.44512	6195	41.00000	79.00000

Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	PSA	AGE
PSA	1.00000	0.01720 0.8672
AGE	0.01720 0.8672	1.00000

The CORR Procedure

2 Variables:	PSA BENIGN

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PSA	97	23.73013	40.78292	2302	0.65100	265.07200
BENIGN	97	2.53472	3.03118	245.86830	0	10.27790

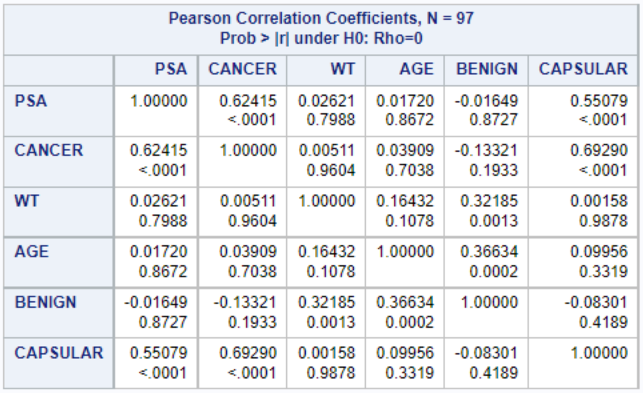
Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	PSA	BENIGN
PSA	1.00000	-0.01649 0.8727
BENIGN	-0.01649 0.8727	1.00000

The CORR Procedure

2 Variables:	PSA CAPSULAR

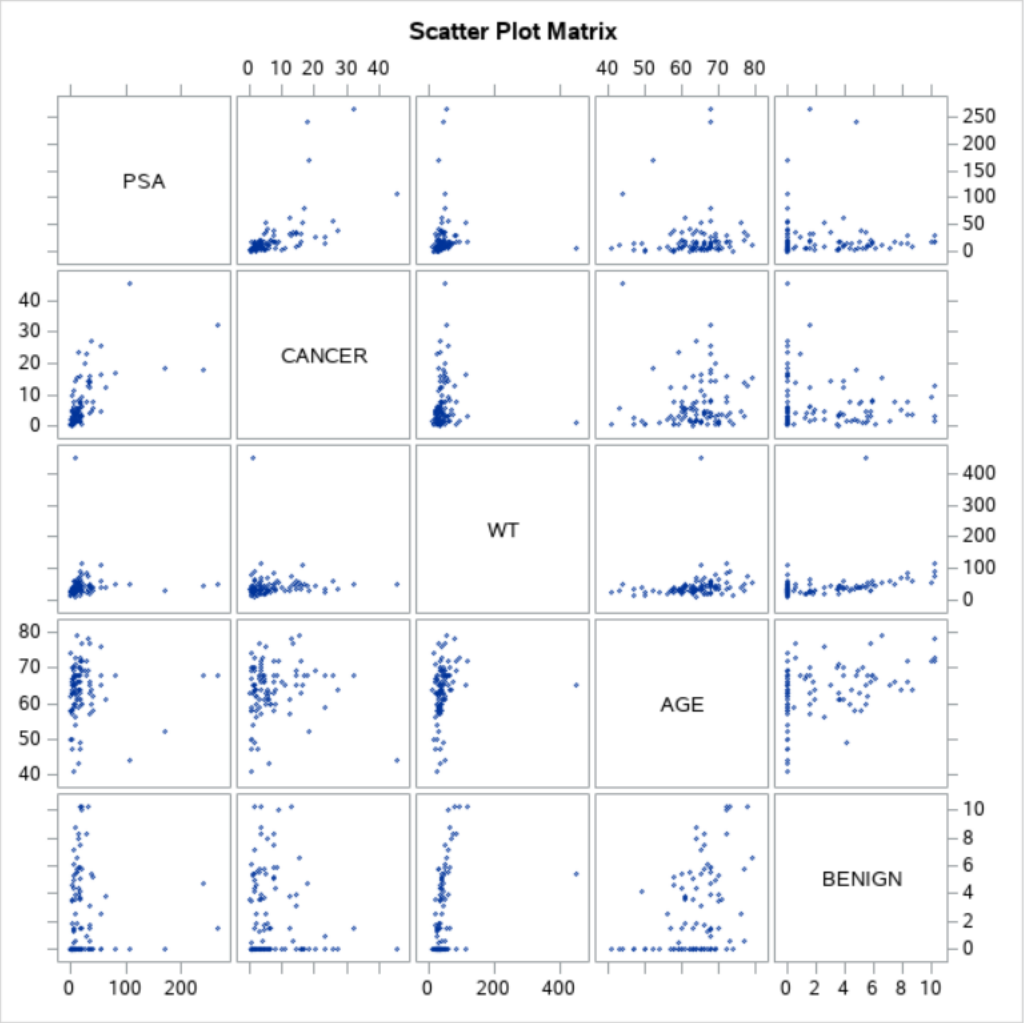
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PSA	97	23.73013	40.78292	2302	0.65100	265.07200
CAPSULAR	97	2.24537	3.78333	217.80060	0	18.17410

Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	PSA	CAPSULAR
PSA	1.00000	0.55079 <.0001
CAPSULAR	0.55079 <.0001	1.00000

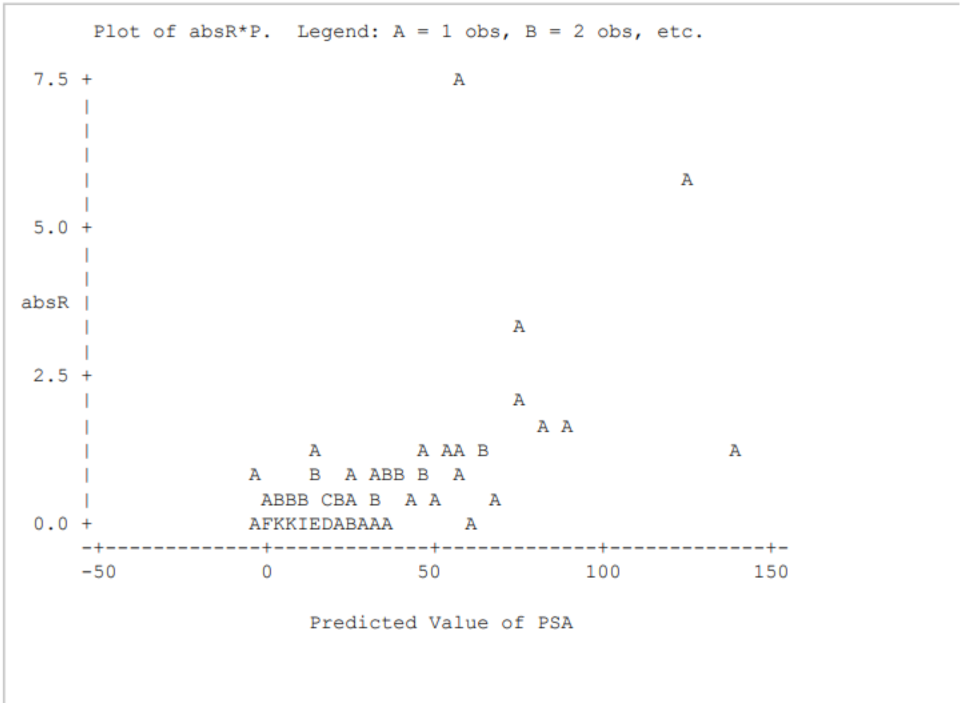


Pearson Correlation Coefficients, N = 97
Prob > |r| under H0: Rho=0

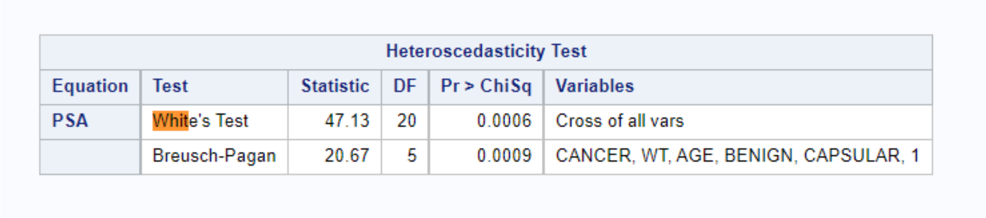
	PSA	CANCER	WT	AGE	BENIGN	CAPSULAR
PSA	1.00000	0.62415 <.0001	0.02621 0.7988	0.01720 0.8672	-0.01649 0.8727	0.55079 <.0001
CANCER	0.62415 <.0001	1.00000	0.00511 0.9604	0.03909 0.7038	-0.13321 0.1933	0.69290 <.0001
WT	0.02621 0.7988	0.00511 0.9604	1.00000	0.16432 0.1078	0.32185 0.0013	0.00158 0.9878
AGE	0.01720 0.8672	0.03909 0.7038	0.16432 0.1078	1.00000	0.36634 0.0002	0.09956 0.3319
BENIGN	-0.01649 0.8727	-0.13321 0.1933	0.32185 0.0013	0.36634 0.0002	1.00000	-0.08301 0.4189
CAPSULAR	0.55079 <.0001	0.69290 <.0001	0.00158 0.9878	0.09956 0.3319	-0.08301 0.4189	1.00000



Output for Q2

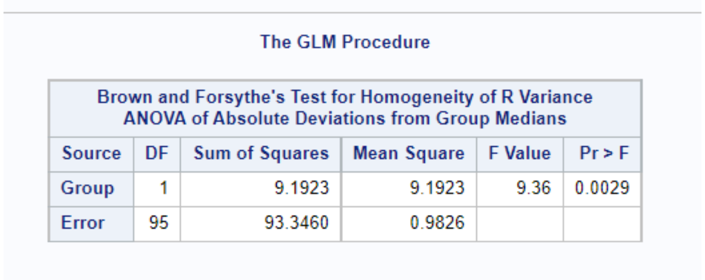


```
      Plot of absR*P.  Legend: A = 1 obs, B = 2 obs, etc.

 7.5 +                              A
     |
     |
     |
     |                                                 A
     |
 5.0 +
     |
     |
absR |
     |                                   A
     |
 2.5 +
     |                                   A
     |                                     A A
     |                  A        A AA B                    A
     |             A    B  A ABB B  A
     |              ABBB CBA B  A A    A
 0.0 +             AFKKIEDABAAA      A
     -+-------------+-------------+-------------+-------------+-
     -50            0            50            100           150

                        Predicted Value of PSA
```



Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
PSA	White's Test	47.13	20	0.0006	Cross of all vars
	Breusch-Pagan	20.67	5	0.0009	CANCER, WT, AGE, BENIGN, CAPSULAR, 1

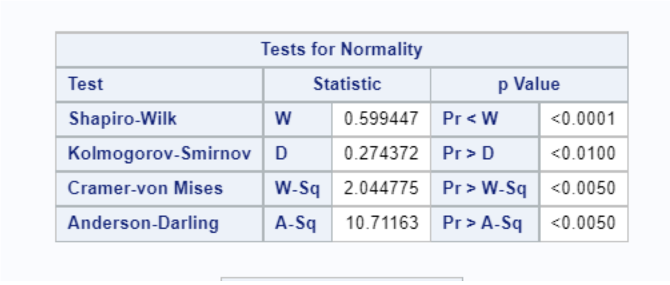
Brow forsythe



The GLM Procedure

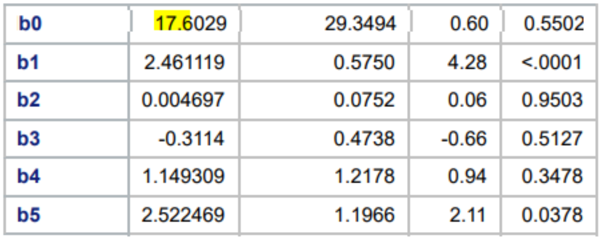
Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	1	9.1923	9.1923	9.36	0.0029
Error	95	93.3460	0.9826		

Shapiro Wilk value disproving normality



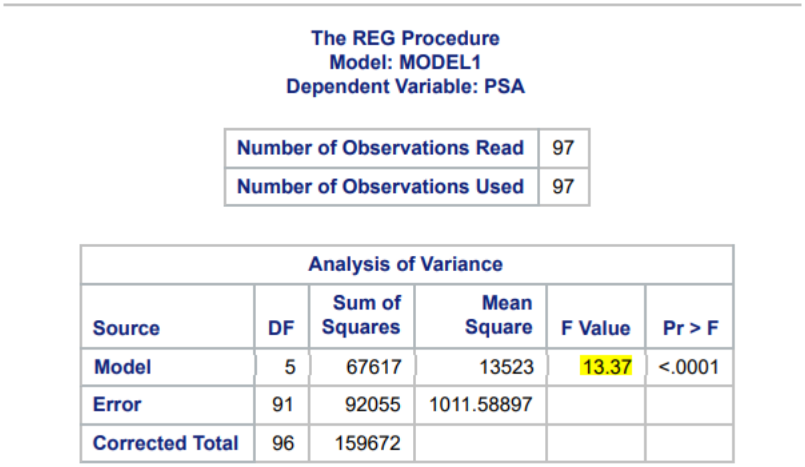
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.599447	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.274372	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.044775	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	10.71163	Pr > A-Sq	<0.0050

Intercept



b0	17.6029	29.3494	0.60	0.5502
b1	2.461119	0.5750	4.28	<.0001
b2	0.004697	0.0752	0.06	0.9503
b3	-0.3114	0.4738	-0.66	0.5127
b4	1.149309	1.2178	0.94	0.3478
b5	2.522469	1.1966	2.11	0.0378

Lack of fit

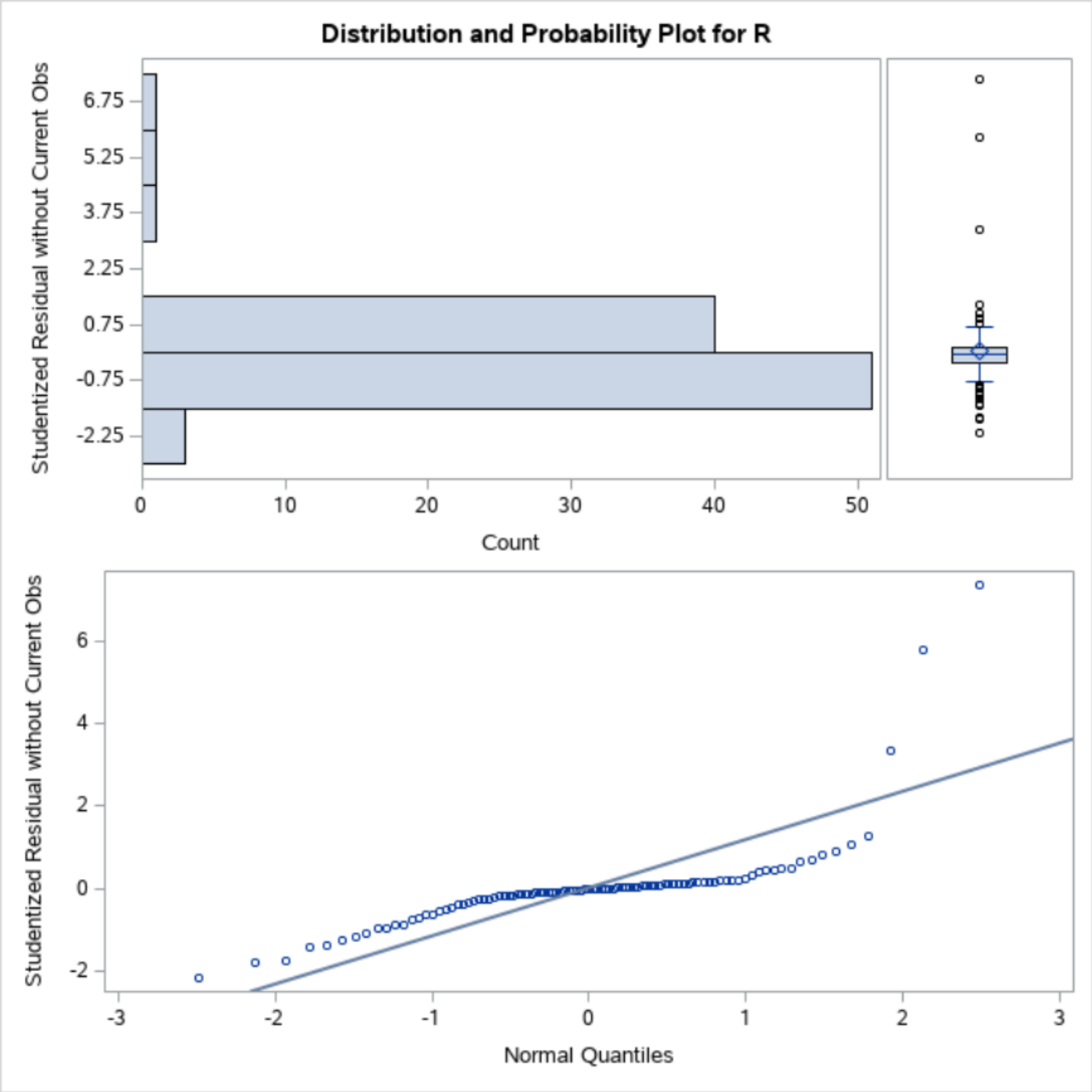


The REG Procedure
Model: MODEL1
Dependent Variable: PSA

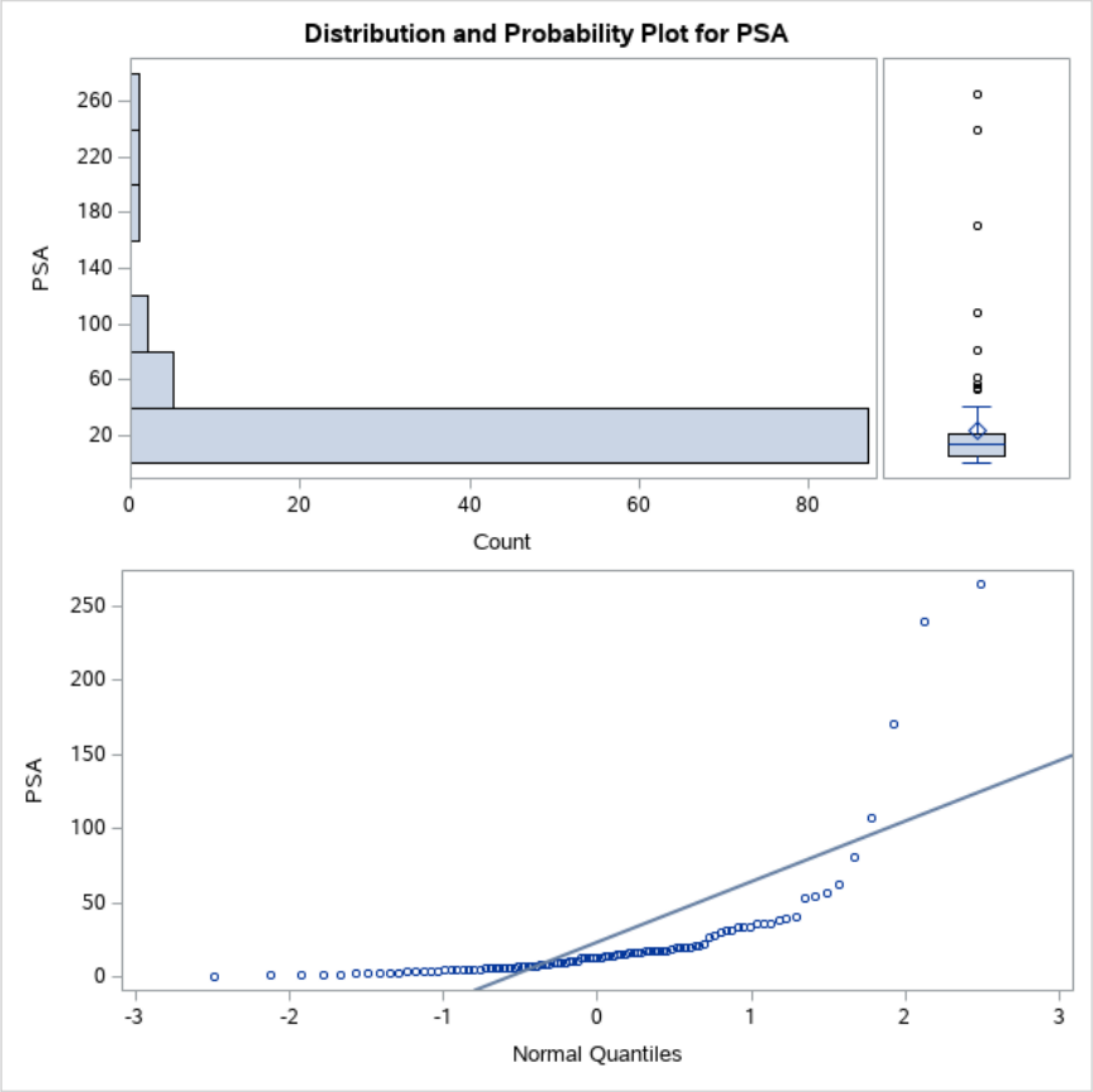
Number of Observations Read	97
Number of Observations Used	97

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	67617	13523	13.37	<.0001
Error	91	92055	1011.58897		
Corrected Total	96	159672			

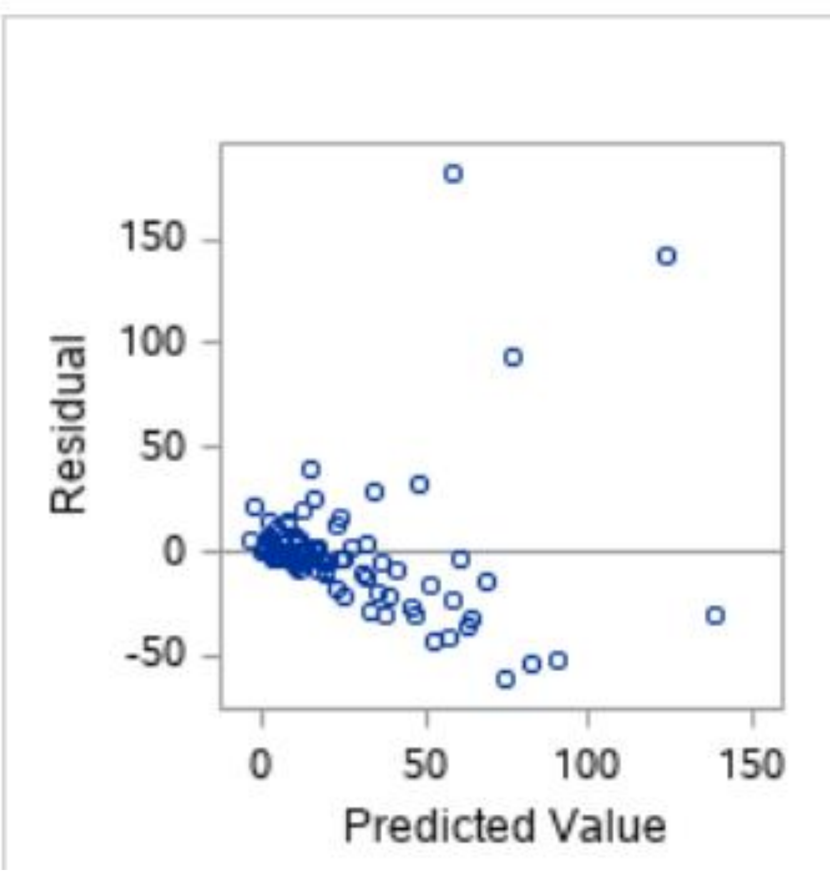
Normality of errors



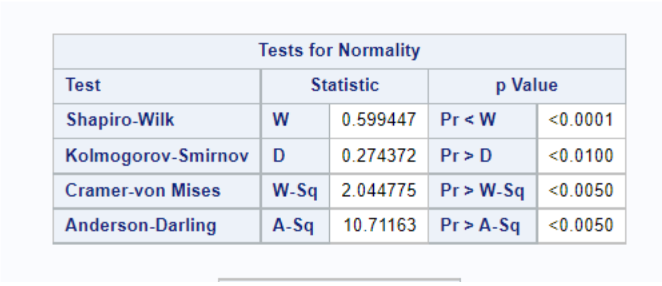
Distribution and Probability Plot for R



Distribution and Probability Plot for PSA

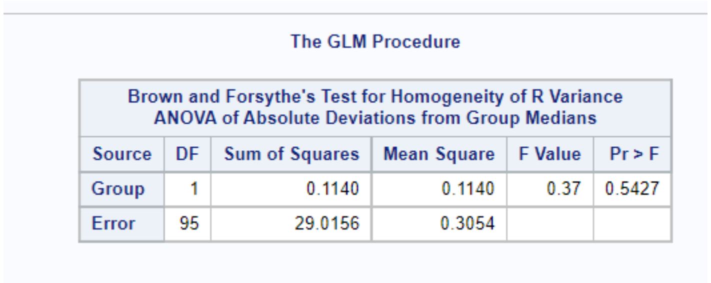


Shapiro wilk of value



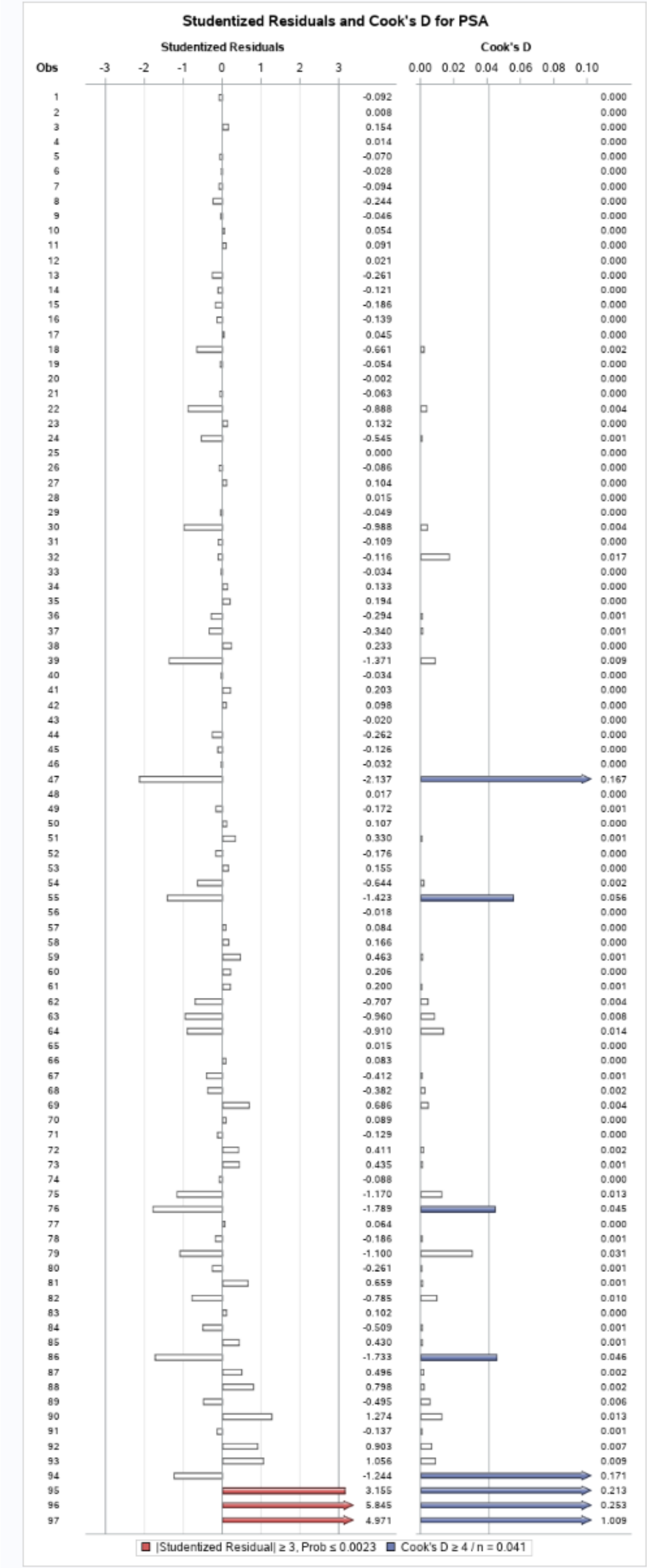
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.599447	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.274372	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.044775	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	10.71163	Pr > A-Sq	<0.0050

Output for q 3

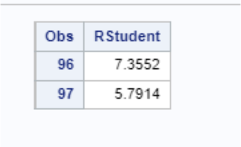


The GLM Procedure

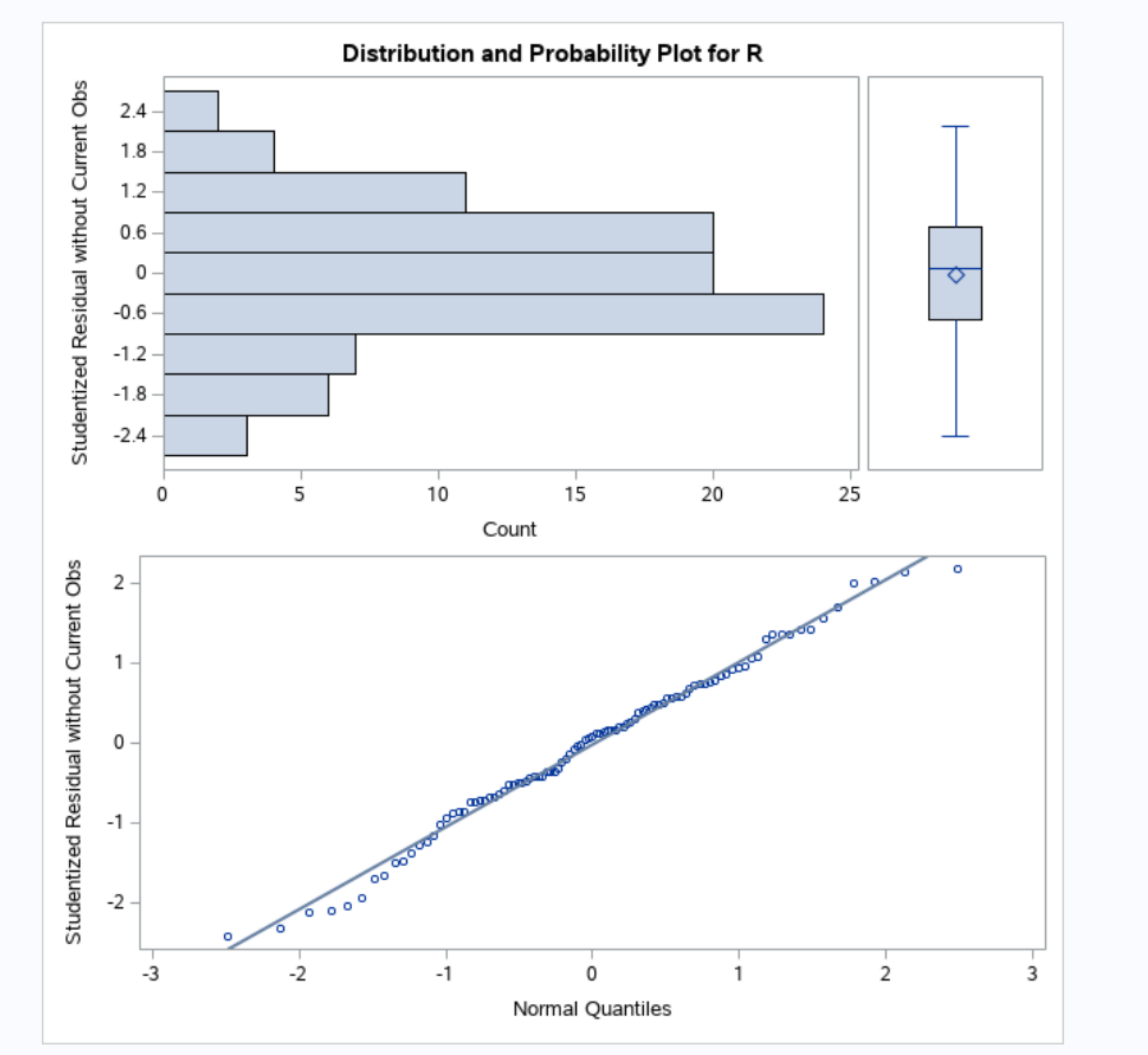
Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	1	0.1140	0.1140	0.37	0.5427
Error	95	29.0156	0.3054		

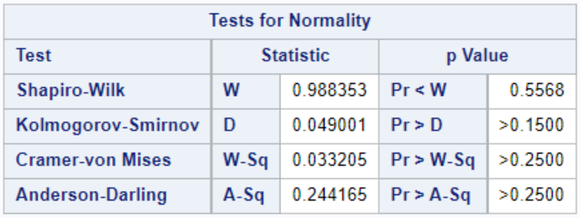


outlier



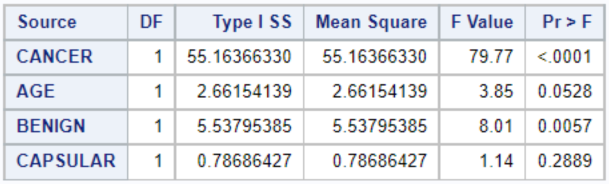
Obs	RStudent
96	7.3552
97	5.7914





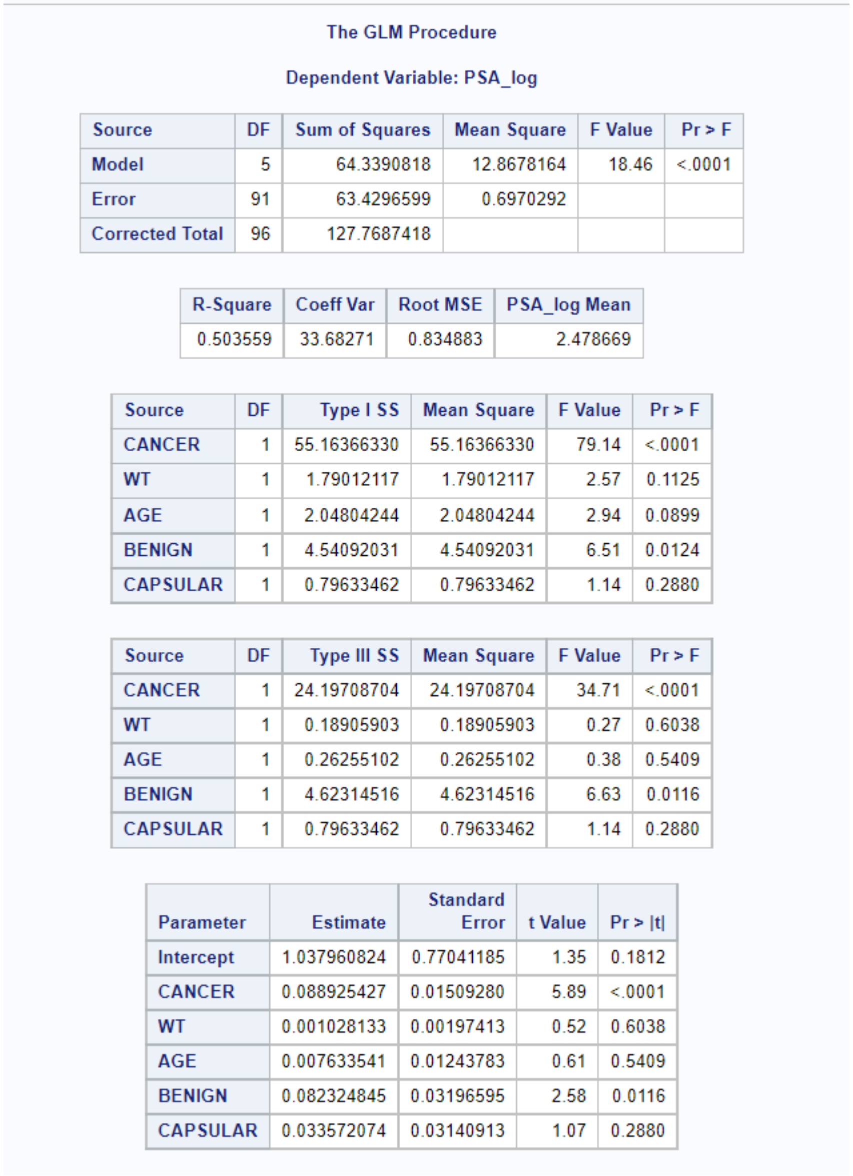
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.988353	Pr < W	0.5568
Kolmogorov-Smirnov	D	0.049001	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.033205	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.244165	Pr > A-Sq	>0.2500

Output for q 4)



Source	DF	Type I SS	Mean Square	F Value	Pr > F
CANCER	1	55.16366330	55.16366330	79.77	<.0001
AGE	1	2.66154139	2.66154139	3.85	0.0528
BENIGN	1	5.53795385	5.53795385	8.01	0.0057
CAPSULAR	1	0.78686427	0.78686427	1.14	0.2889

All variables



The GLM Procedure

Dependent Variable: PSA_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	64.3390818	12.8678164	18.46	<.0001
Error	91	63.4296599	0.6970292		
Corrected Total	96	127.7687418			

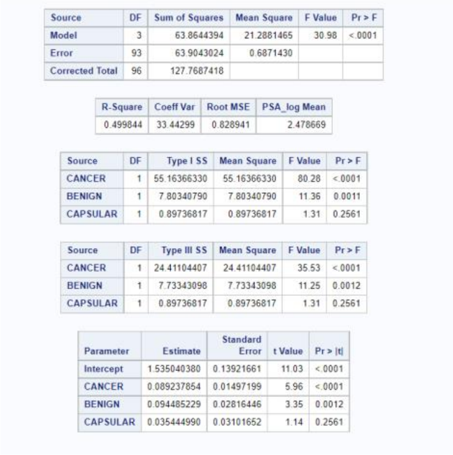
R-Square	Coeff Var	Root MSE	PSA_log Mean
0.503559	33.68271	0.834883	2.478669

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CANCER	1	55.16366330	55.16366330	79.14	<.0001
WT	1	1.79012117	1.79012117	2.57	0.1125
AGE	1	2.04804244	2.04804244	2.94	0.0899
BENIGN	1	4.54092031	4.54092031	6.51	0.0124
CAPSULAR	1	0.79633462	0.79633462	1.14	0.2880

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CANCER	1	24.19708704	24.19708704	34.71	<.0001
WT	1	0.18905903	0.18905903	0.27	0.6038
AGE	1	0.26255102	0.26255102	0.38	0.5409
BENIGN	1	4.62314516	4.62314516	6.63	0.0116
CAPSULAR	1	0.79633462	0.79633462	1.14	0.2880

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.037960824	0.77041185	1.35	0.1812
CANCER	0.088925427	0.01509280	5.89	<.0001
WT	0.001028133	0.00197413	0.52	0.6038
AGE	0.007633541	0.01243783	0.61	0.5409
BENIGN	0.082324845	0.03196595	2.58	0.0116
CAPSULAR	0.033572074	0.03140913	1.07	0.2880

Age ommited



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63.8644394	21.2881465	30.98	<.0001
Error	93	63.9043024	0.6871430		
Corrected Total	96	127.7687418			

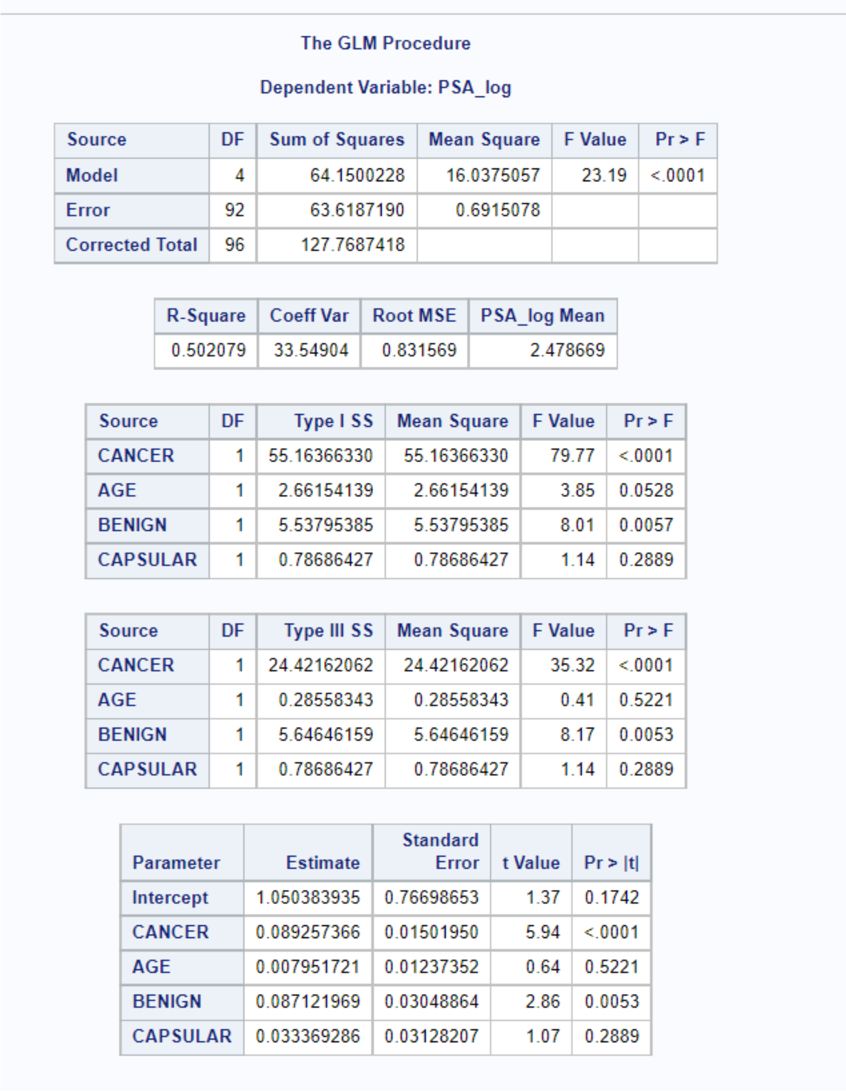
R-Square	Coeff Var	Root MSE	PSA_log Mean
0.499844	33.44299	0.828941	2.478669

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CANCER	1	55.16366330	55.16366330	80.28	<.0001
BENIGN	1	7.80340790	7.80340790	11.36	0.0011
CAPSULAR	1	0.89736817	0.89736817	1.31	0.2561

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CANCER	1	24.41104407	24.41104407	35.53	<.0001
BENIGN	1	7.73343098	7.73343098	11.25	0.0012
CAPSULAR	1	0.89736817	0.89736817	1.31	0.2561

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.535040380	0.13921661	11.03	<.0001
CANCER	0.089237854	0.01497199	5.96	<.0001
BENIGN	0.094485229	0.02816446	3.35	0.0012
CAPSULAR	0.035444990	0.03101652	1.14	0.2561

Weight omitted



The GLM Procedure

Dependent Variable: PSA_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	64.1500228	16.0375057	23.19	<.0001
Error	92	63.6187190	0.6915078		
Corrected Total	96	127.7687418			

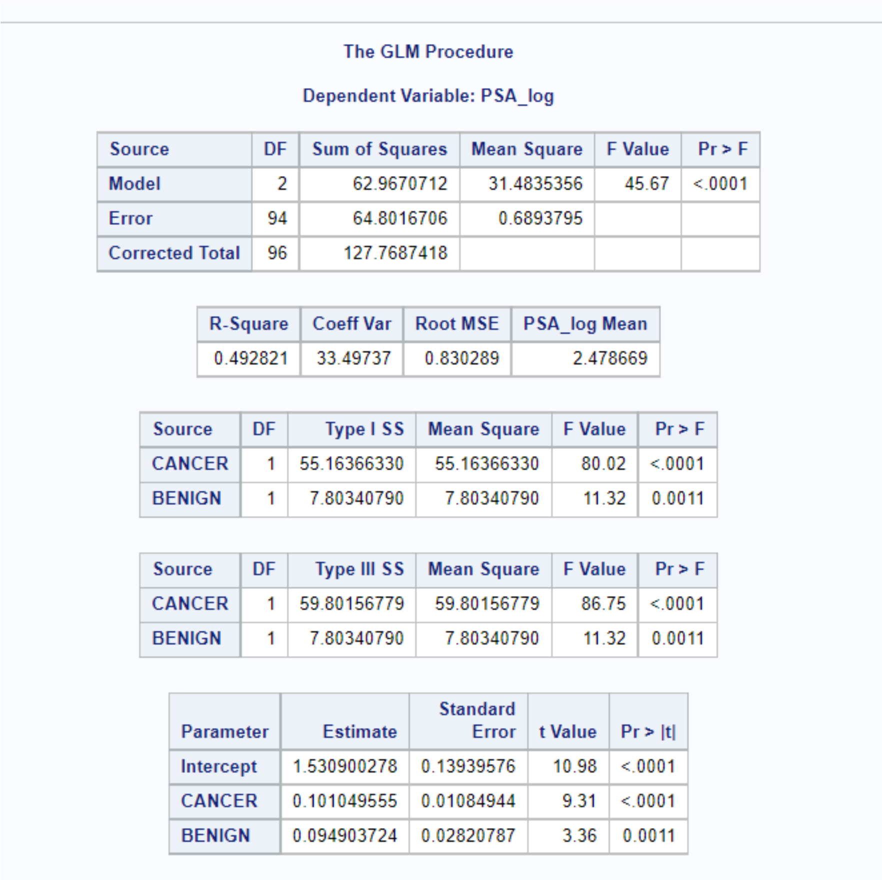
R-Square	Coeff Var	Root MSE	PSA_log Mean
0.502079	33.54904	0.831569	2.478669

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CANCER	1	55.16366330	55.16366330	79.77	<.0001
AGE	1	2.66154139	2.66154139	3.85	0.0528
BENIGN	1	5.53795385	5.53795385	8.01	0.0057
CAPSULAR	1	0.78686427	0.78686427	1.14	0.2889

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CANCER	1	24.42162062	24.42162062	35.32	<.0001
AGE	1	0.28558343	0.28558343	0.41	0.5221
BENIGN	1	5.64646159	5.64646159	8.17	0.0053
CAPSULAR	1	0.78686427	0.78686427	1.14	0.2889

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.050383935	0.76698653	1.37	0.1742
CANCER	0.089257366	0.01501950	5.94	<.0001
AGE	0.007951721	0.01237352	0.64	0.5221
BENIGN	0.087121969	0.03048864	2.86	0.0053
CAPSULAR	0.033369286	0.03128207	1.07	0.2889

Final model



The GLM Procedure

Dependent Variable: PSA_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	62.9670712	31.4835356	45.67	<.0001
Error	94	64.8016706	0.6893795		
Corrected Total	96	127.7687418			

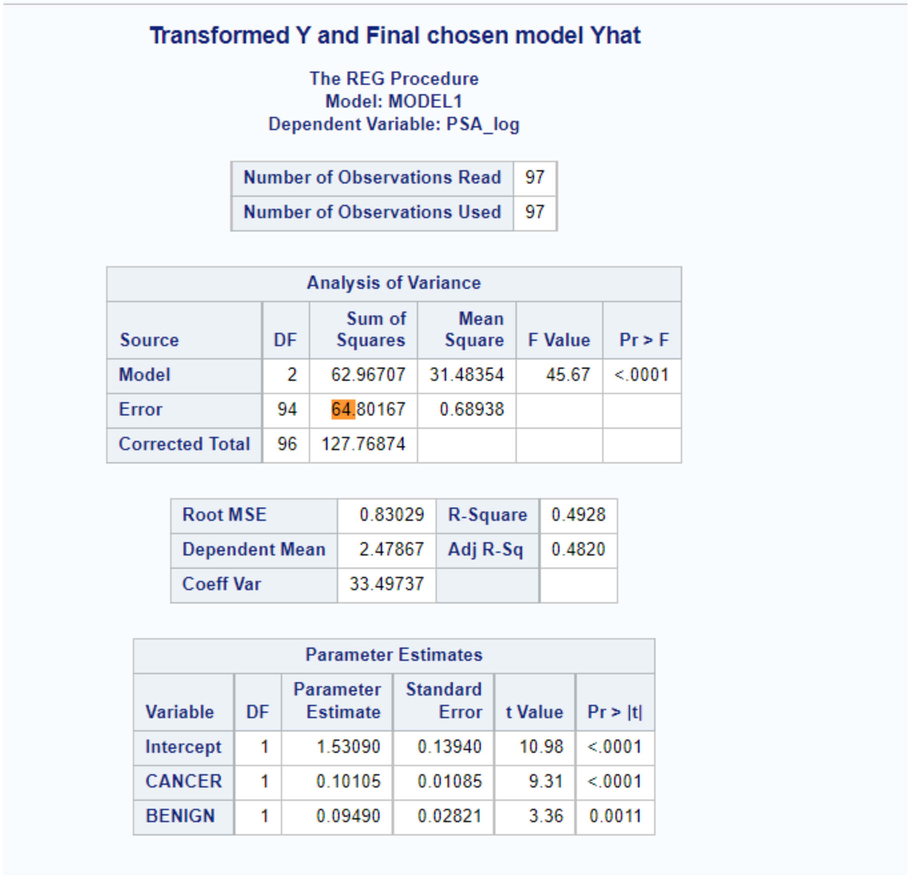
R-Square	Coeff Var	Root MSE	PSA_log Mean
0.492821	33.49737	0.830289	2.478669

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CANCER	1	55.16366330	55.16366330	80.02	<.0001
BENIGN	1	7.80340790	7.80340790	11.32	0.0011

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CANCER	1	59.80156779	59.80156779	86.75	<.0001
BENIGN	1	7.80340790	7.80340790	11.32	0.0011

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.530900278	0.13939576	10.98	<.0001
CANCER	0.101049555	0.01084944	9.31	<.0001
BENIGN	0.094903724	0.02820787	3.36	0.0011

Output for Q5



Transformed Y and Final chosen model Yhat

The REG Procedure
Model: MODEL1
Dependent Variable: PSA_log

Number of Observations Read	97
Number of Observations Used	97

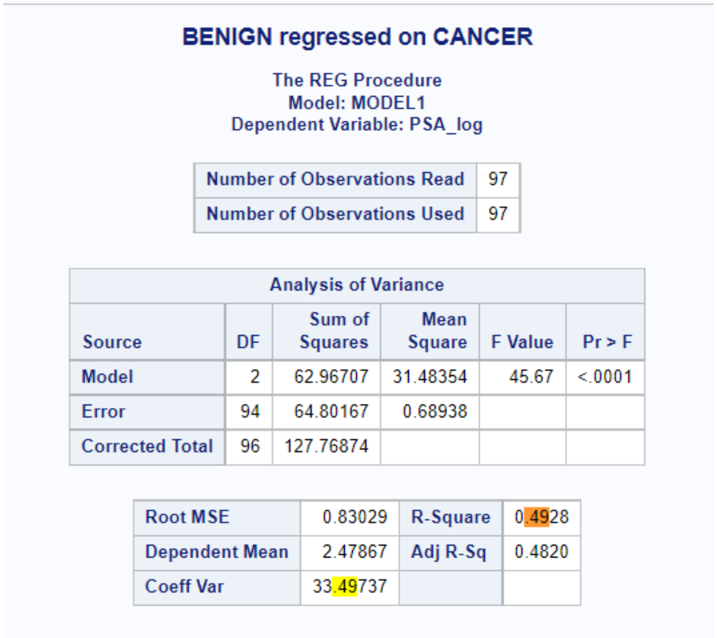
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	62.96707	31.48354	45.67	<.0001
Error	94	64.80167	0.68938		
Corrected Total	96	127.76874			

Root MSE	0.83029	R-Square	0.4928
Dependent Mean	2.47867	Adj R-Sq	0.4820
Coeff Var	33.49737		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.53090	0.13940	10.98	<.0001
CANCER	1	0.10105	0.01085	9.31	<.0001
BENIGN	1	0.09490	0.02821	3.36	0.0011



BENIGN regressed on CANCER

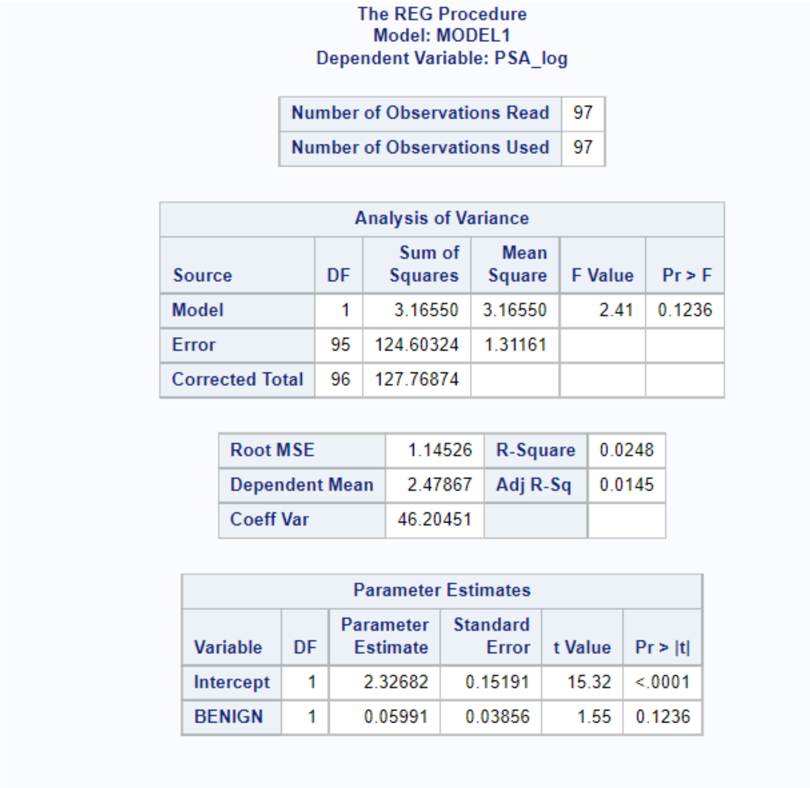
The REG Procedure
Model: MODEL1
Dependent Variable: PSA_log

Number of Observations Read	97
Number of Observations Used	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	62.96707	31.48354	45.67	<.0001
Error	94	64.80167	0.68938		
Corrected Total	96	127.76874			

Root MSE	0.83029	R-Square	0.4928
Dependent Mean	2.47867	Adj R-Sq	0.4820
Coeff Var	33.49737		



The REG Procedure
Model: MODEL1
Dependent Variable: PSA_log

Number of Observations Read	97
Number of Observations Used	97

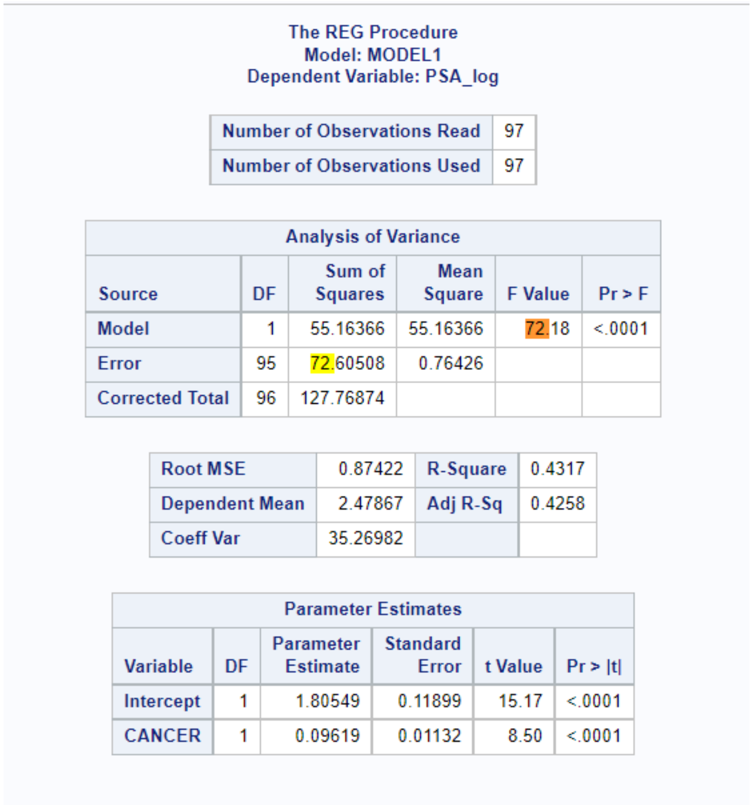
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.16550	3.16550	2.41	0.1236
Error	95	124.60324	1.31161		
Corrected Total	96	127.76874			

Root MSE	1.14526	R-Square	0.0248
Dependent Mean	2.47867	Adj R-Sq	0.0145
Coeff Var	46.20451		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.32682	0.15191	15.32	<.0001
BENIGN	1	0.05991	0.03856	1.55	0.1236



The REG Procedure
Model: MODEL1
Dependent Variable: PSA_log

Number of Observations Read	97
Number of Observations Used	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	55.16366	55.16366	72.18	<.0001
Error	95	72.60508	0.76426		
Corrected Total	96	127.76874			

Root MSE	0.87422	R-Square	0.4317
Dependent Mean	2.47867	Adj R-Sq	0.4258
Coeff Var	35.26982		

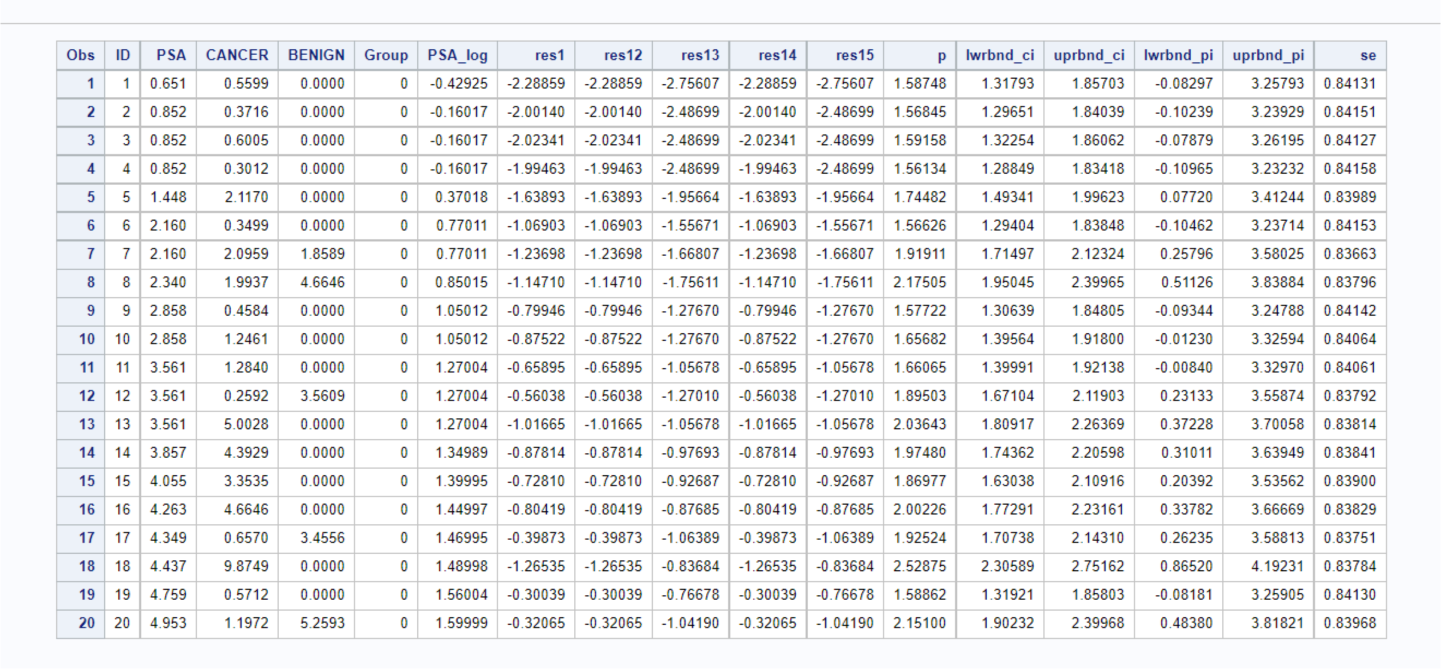
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.80549	0.11899	15.17	<.0001
CANCER	1	0.09619	0.01132	8.50	<.0001

Output for Q6

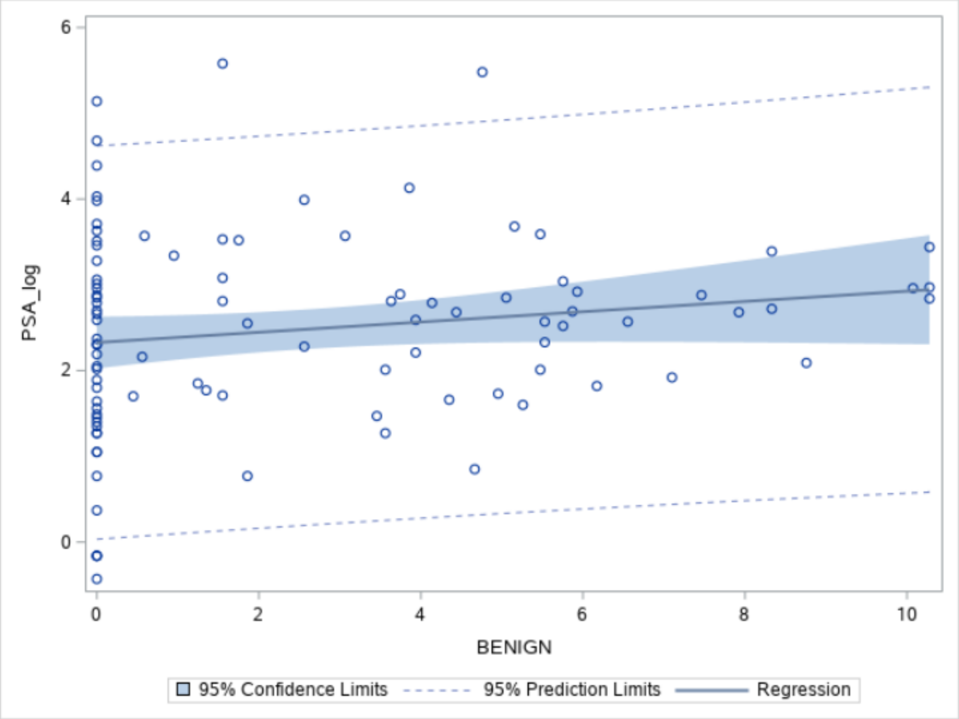
95 % SI

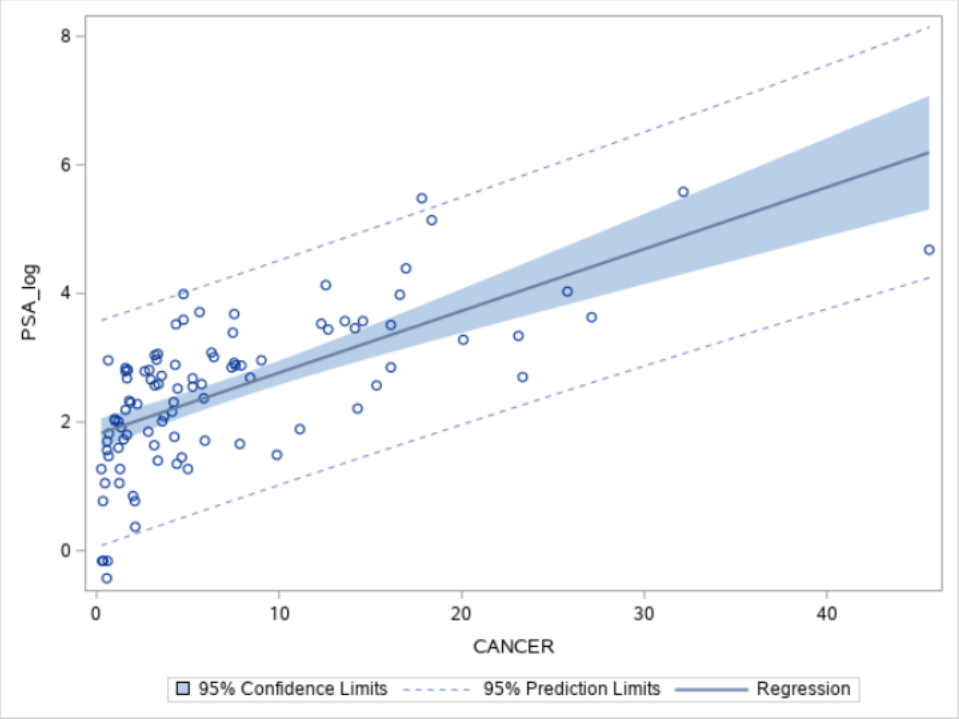
CO

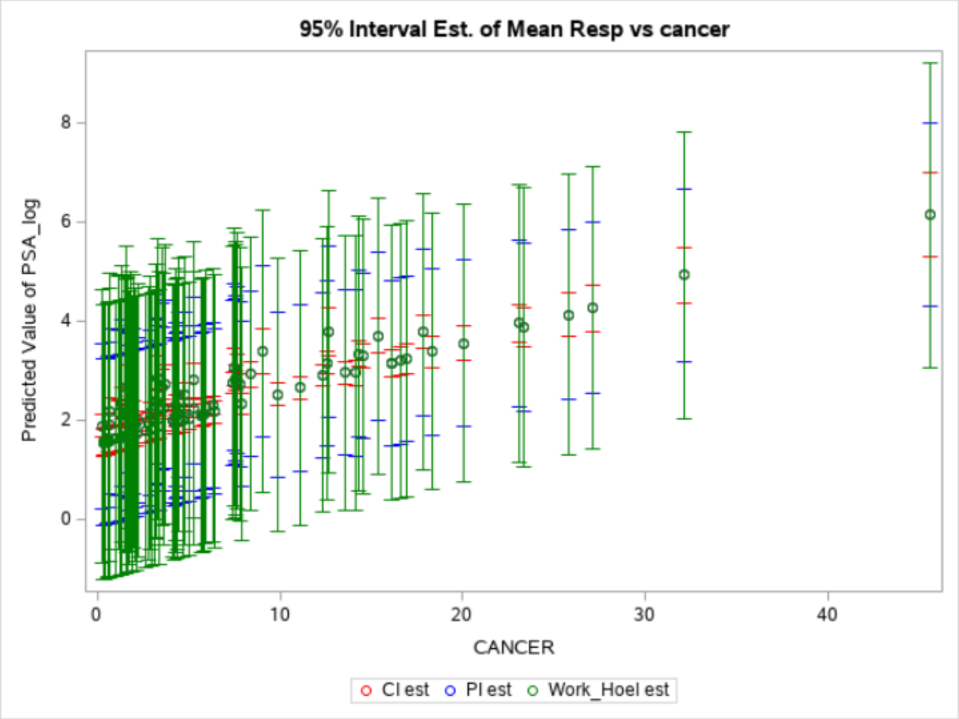


Obs	ID	PSA	CANCER	BENIGN	Group	PSA_log	res1	res12	res13	res14	res15	p	lwrbnd_ci	uprbnd_ci	lwrbnd_pi	uprbnd_pi	se
1	1	0.651	0.5599	0.0000	0	-0.42925	-2.28859	-2.28859	-2.75607	-2.28859	-2.75607	1.58748	1.31793	1.85703	-0.08297	3.25793	0.84131
2	2	0.852	0.3716	0.0000	0	-0.16017	-2.00140	-2.00140	-2.48699	-2.00140	-2.48699	1.56845	1.29651	1.84039	-0.10239	3.23929	0.84151
3	3	0.852	0.6005	0.0000	0	-0.16017	-2.02341	-2.02341	-2.48699	-2.02341	-2.48699	1.59158	1.32254	1.86062	-0.07879	3.26195	0.84127
4	4	0.852	0.3012	0.0000	0	-0.16017	-1.99463	-1.99463	-2.48699	-1.99463	-2.48699	1.56134	1.28849	1.83418	-0.10965	3.23232	0.84158
5	5	1.448	2.1170	0.0000	0	0.37018	-1.63893	-1.63893	-1.95664	-1.63893	-1.95664	1.74482	1.49341	1.99623	0.07720	3.41244	0.83989
6	6	2.160	0.3499	0.0000	0	0.77011	-1.06903	-1.06903	-1.55671	-1.06903	-1.55671	1.56626	1.29404	1.83848	-0.10462	3.23714	0.84153
7	7	2.160	2.0959	1.8589	0	0.77011	-1.23698	-1.23698	-1.66807	-1.23698	-1.66807	1.91911	1.71497	2.12324	0.25796	3.58025	0.83663
8	8	2.340	1.9937	4.6646	0	0.85015	-1.14710	-1.14710	-1.75611	-1.14710	-1.75611	2.17505	1.95045	2.39965	0.51126	3.83884	0.83796
9	9	2.858	0.4584	0.0000	0	1.05012	-0.79946	-0.79946	-1.27670	-0.79946	-1.27670	1.57722	1.30639	1.84805	-0.09344	3.24788	0.84142
10	10	2.858	1.2461	0.0000	0	1.05012	-0.87522	-0.87522	-1.27670	-0.87522	-1.27670	1.65682	1.39564	1.91800	-0.01230	3.32594	0.84064
11	11	3.561	1.2840	0.0000	0	1.27004	-0.65895	-0.65895	-1.05678	-0.65895	-1.05678	1.66065	1.39991	1.92138	-0.00840	3.32970	0.84061
12	12	3.561	0.2592	3.5609	0	1.27004	-0.56038	-0.56038	-1.27010	-0.56038	-1.27010	1.89503	1.67104	2.11903	0.23133	3.55874	0.83792
13	13	3.561	5.0028	0.0000	0	1.27004	-1.01665	-1.01665	-1.05678	-1.01665	-1.05678	2.03643	1.80917	2.26369	0.37228	3.70058	0.83814
14	14	3.857	4.3929	0.0000	0	1.34989	-0.87814	-0.87814	-0.97693	-0.87814	-0.97693	1.97480	1.74362	2.20598	0.31011	3.63949	0.83841
15	15	4.055	3.3535	0.0000	0	1.39995	-0.72810	-0.72810	-0.92687	-0.72810	-0.92687	1.86977	1.63038	2.10916	0.20392	3.53562	0.83900
16	16	4.263	4.6646	0.0000	0	1.44997	-0.80419	-0.80419	-0.87685	-0.80419	-0.87685	2.00226	1.77291	2.23161	0.33782	3.66669	0.83829
17	17	4.349	0.6570	3.4556	0	1.46995	-0.39873	-0.39873	-1.06389	-0.39873	-1.06389	1.92524	1.70738	2.14310	0.26235	3.58813	0.83751
18	18	4.437	9.8749	0.0000	0	1.48998	-1.26535	-1.26535	-0.83684	-1.26535	-0.83684	2.52875	2.30589	2.75162	0.86520	4.19231	0.83784
19	19	4.759	0.5712	0.0000	0	1.56004	-0.30039	-0.30039	-0.76678	-0.30039	-0.76678	1.58862	1.31921	1.85803	-0.08181	3.25905	0.84130
20	20	4.953	1.1972	5.2593	0	1.59999	-0.32065	-0.32065	-1.04190	-0.32065	-1.04190	2.15100	1.90232	2.39968	0.48380	3.81821	0.83968

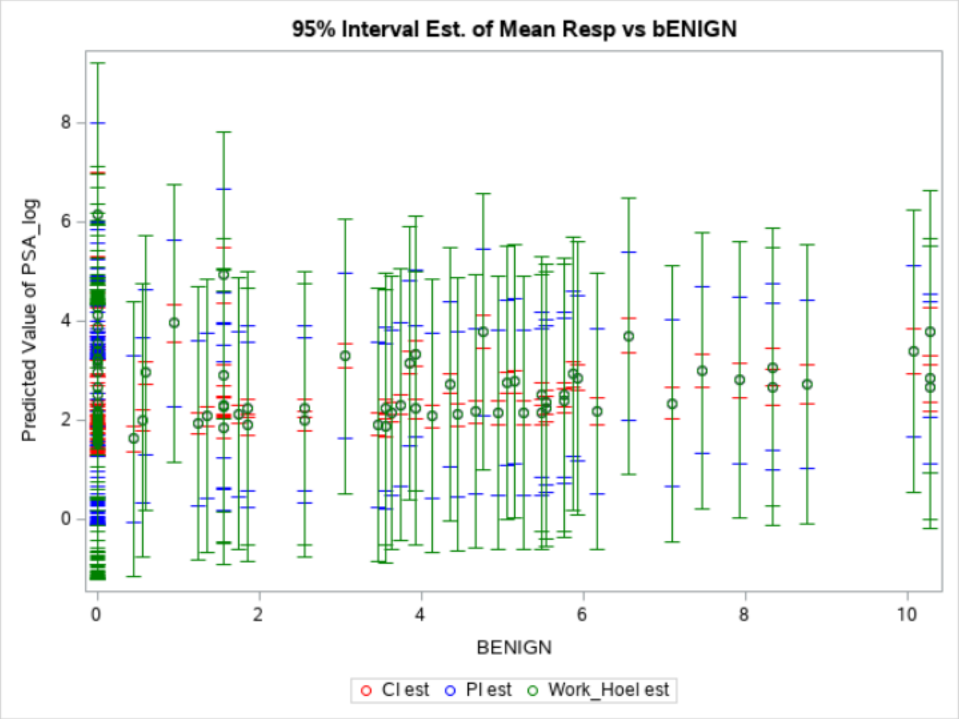
Confidence Band


6
4
2
0
PSA_log
0
2
4
6
8
10
BENIGN
□ 95% Confidence Limits
95% Prediction Limits
Regression


8
6
4
2
0
PSA_log
0
10
20
30
40
CANCER
□ 95% Confidence Limits
95% Prediction Limits
Regression

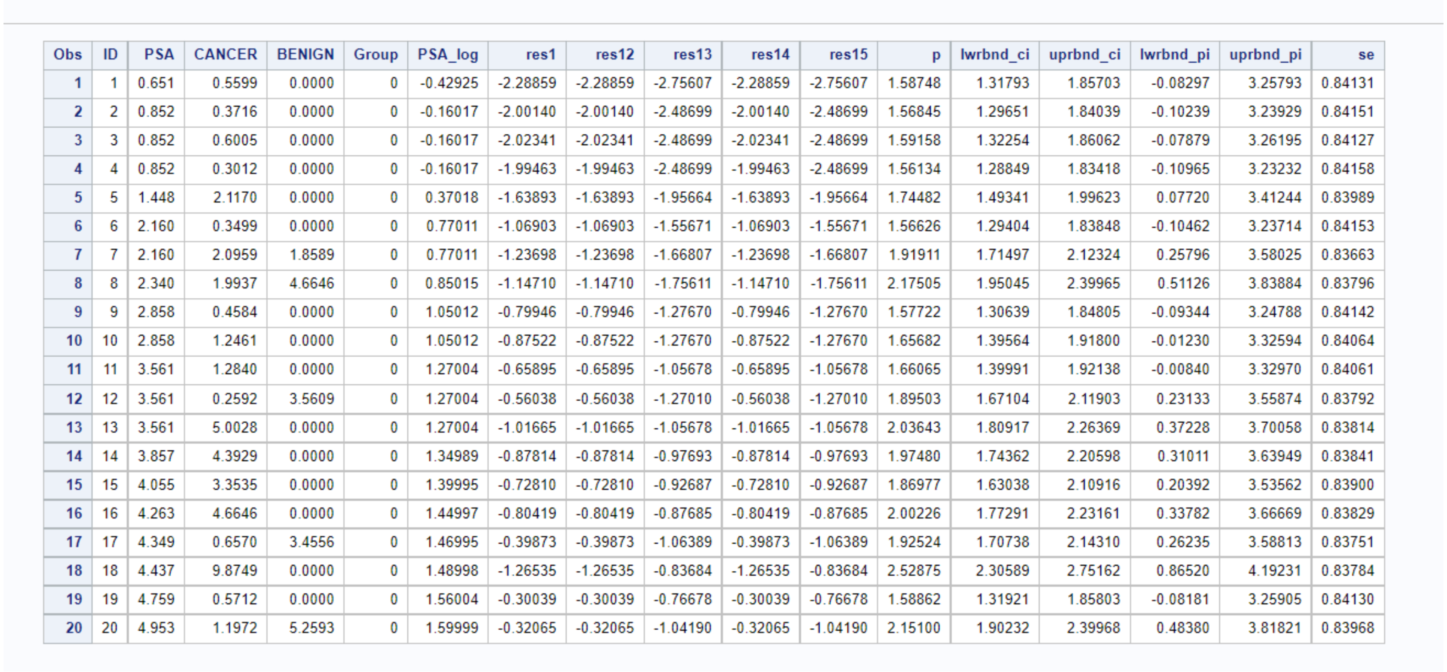


95% Interval Est. of Mean Resp vs cancer



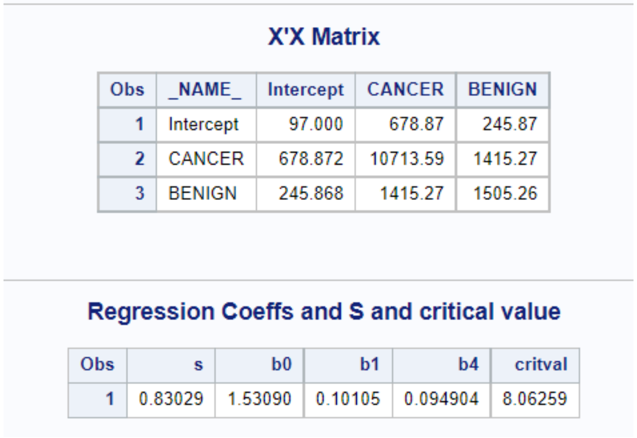
95% Interval Est. of Mean Resp vs bENIGN

Prediction interval



Obs	ID	PSA	CANCER	BENIGN	Group	PSA_log	res1	res12	res13	res14	res15	p	lwrbnd_ci	uprbnd_ci	lwrbnd_pi	uprbnd_pi	se
1	1	0.651	0.5599	0.0000	0	-0.42925	-2.28859	-2.28859	-2.75607	-2.28859	-2.75607	1.58748	1.31793	1.85703	-0.08297	3.25793	0.84131
2	2	0.852	0.3716	0.0000	0	-0.16017	-2.00140	-2.00140	-2.48699	-2.00140	-2.48699	1.56845	1.29651	1.84039	-0.10239	3.23929	0.84151
3	3	0.852	0.6005	0.0000	0	-0.16017	-2.02341	-2.02341	-2.48699	-2.02341	-2.48699	1.59158	1.32254	1.86062	-0.07879	3.26195	0.84127
4	4	0.852	0.3012	0.0000	0	-0.16017	-1.99463	-1.99463	-2.48699	-1.99463	-2.48699	1.56134	1.28849	1.83418	-0.10965	3.23232	0.84158
5	5	1.448	2.1170	0.0000	0	0.37018	-1.63893	-1.63893	-1.95664	-1.63893	-1.95664	1.74482	1.49341	1.99623	0.07720	3.41244	0.83989
6	6	2.160	0.3499	0.0000	0	0.77011	-1.06903	-1.06903	-1.55671	-1.06903	-1.55671	1.56626	1.29404	1.83848	-0.10462	3.23714	0.84153
7	7	2.160	2.0959	1.8589	0	0.77011	-1.23698	-1.23698	-1.66807	-1.23698	-1.66807	1.91911	1.71497	2.12324	0.25796	3.58025	0.83663
8	8	2.340	1.9937	4.6646	0	0.85015	-1.14710	-1.14710	-1.75611	-1.14710	-1.75611	2.17505	1.95045	2.39965	0.51126	3.83884	0.83796
9	9	2.858	0.4584	0.0000	0	1.05012	-0.79946	-0.79946	-1.27670	-0.79946	-1.27670	1.57722	1.30639	1.84805	-0.09344	3.24788	0.84142
10	10	2.858	1.2461	0.0000	0	1.05012	-0.87522	-0.87522	-1.27670	-0.87522	-1.27670	1.65682	1.39564	1.91800	-0.01230	3.32594	0.84064
11	11	3.561	1.2840	0.0000	0	1.27004	-0.65895	-0.65895	-1.05678	-0.65895	-1.05678	1.66065	1.39991	1.92138	-0.00840	3.32970	0.84061
12	12	3.561	0.2592	3.5609	0	1.27004	-0.56038	-0.56038	-1.27010	-0.56038	-1.27010	1.89503	1.67104	2.11903	0.23133	3.55874	0.83792
13	13	3.561	5.0028	0.0000	0	1.27004	-1.01665	-1.01665	-1.05678	-1.01665	-1.05678	2.03643	1.80917	2.26369	0.37228	3.70058	0.83814
14	14	3.857	4.3929	0.0000	0	1.34989	-0.87814	-0.87814	-0.97693	-0.87814	-0.97693	1.97480	1.74362	2.20598	0.31011	3.63949	0.83841
15	15	4.055	3.3535	0.0000	0	1.39995	-0.72810	-0.72810	-0.92687	-0.72810	-0.92687	1.86977	1.63038	2.10916	0.20392	3.53562	0.83900
16	16	4.263	4.6646	0.0000	0	1.44997	-0.80419	-0.80419	-0.87685	-0.80419	-0.87685	2.00226	1.77291	2.23161	0.33782	3.66669	0.83829
17	17	4.349	0.6570	3.4556	0	1.46995	-0.39873	-0.39873	-1.06389	-0.39873	-1.06389	1.92524	1.70738	2.14310	0.26235	3.58813	0.83751
18	18	4.437	9.8749	0.0000	0	1.48998	-1.26535	-1.26535	-0.83684	-1.26535	-0.83684	2.52875	2.30589	2.75162	0.86520	4.19231	0.83784
19	19	4.759	0.5712	0.0000	0	1.56004	-0.30039	-0.30039	-0.76678	-0.30039	-0.76678	1.58862	1.31921	1.85803	-0.08181	3.25905	0.84130
20	20	4.953	1.1972	5.2593	0	1.59999	-0.32065	-0.32065	-1.04190	-0.32065	-1.04190	2.15100	1.90232	2.39968	0.48380	3.81821	0.83968

Q7 output



X'X Matrix

Obs	_NAME_	Intercept	CANCER	BENIGN
1	Intercept	97.000	678.87	245.87
2	CANCER	678.872	10713.59	1415.27
3	BENIGN	245.868	1415.27	1505.26

Regression Coeffs and S and critical value

Obs	s	b0	b1	b4	critval
1	0.83029	1.53090	0.10105	0.094904	8.06259

Codes:-

```
DATA D ; INFILE "/home/u59316208/Prostate.dat";

INPUT ID PSA CANCER WT AGE BENIGN SEMINAL CAPSULAR GLEASON;

RUN;


/* DROPPING QUALTITATIVE VARIABLES */

data C; set D;

drop SEMINAL GLEASON;

run;

/* scatterplot matrix */

PROC CORR DATA=C PLOTS=MATRIX;
```

```
VAR PSA CANCER WT AGE BENIGN  CAPSULAR  ;
RUN;

/* FINDING CORRELATION OF VARIABLES WITH PREDICTABLE VARIABLES */
proc corr data=C;
var PSA CANCER;
run;

proc corr data=C;
var PSA WT;
run;

proc corr data=C;
var PSA AGE;
run;

proc corr data=C;
var PSA BENIGN;
run;

proc corr data=C;
var PSA CAPSULAR;
run;




/*Question 2*/
PROC REG Data=C;
MODEL PSA = CANCER WT AGE BENIGN CAPSULAR/ lackfit xpx i ;
OUTPUT OUT=A RSTUDENT=R PREDICTED=P; /*predicted (yhat) and residuals*/
RUN;



/*Breusch Pagan Test for Normality*/
PROC MODEL DATA=C;
```

```
PARMS b0 b1 b2 b3 b4 b5;
PSA = b0 + b1*CANCER + b2*WT + b3*AGE + b4*BENIGN+b5*CAPSULAR;
fit PSA /WHITE BREUSCH=(CANCER WT AGE BENIGN CAPSULAR);
fit PSA /BREUSCH=(CANCER);
fit PSA /BREUSCH=(WT);
fit PSA /BREUSCH=(AGE);
fit PSA /BREUSCH=(BENIGN);
fit PSA /BREUSCH=(CAPSULAR);
RUN;

/*Brown-Forsythe Test*/
DATA C;set C;
Group = (PSA > 13.3); /* median r = -0.0333717 from above */
RUN;

Proc print data=X(obs=15);
run;

PROC GLM Data=A;
class Group;
model R=Group;
means Group / hovtest=BF; /*Brown-Forsythe Test*/
run;

/*checks for normality of errors*/
proc univariate data=A normal plot;
var r;
run;

proc univariate data=C normal plot;
var PSA;
run;
PROC plot Data=C;
PLOT PSA*(CANCER WT AGE BENIGN CAPSULAR);
RUN;
```

```
/*Find residual and predicted values with plots*/
PROC PLOT Data=A HPERCENT=50 VPERCENT=50; /* Residual plot of each variables */
plot R*(CANCER WT AGE BENIGN CAPSULAR);
RUN;


DATA A; SET A;
absR = abs(R); /* save absolute value of residuals */
RUN;

PROC PLOT DATA = A HPERCENT=50 VPERCENT=50;
PLOT R*P; /* residuals vs fitted values */
RUN;

PROC PLOT DATA = A HPERCENT=50 VPERCENT=50;
PLOT absR*P; /* absolute residuals vs fitted values to check homogeneity assumption */
RUN;

/*Question 3*/
/*checking for outliers*/
PROC REG DATA=C;
MODEL PSA = CANCER WT AGE BENIGN CAPSULAR / INFLUENCE R;
ods output outputstatistics=results;
RUN;

PROC PRINT Data=results (obs=10);
RUN;

DATA results; set results; /* Test for outliers using Bonferroni method 95% C-level, 4 variables*/
tvalue = tinv(0.999742268, 91);/*.999528302=1-.05/(2*97) and df=n-p-1=97-5-1=91*/
if (abs(RStudent)) > tvalue then outlier=1;
else outlier=0;
RUN;
```

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

/*notice outliers at Obs 96 97*/

proc print data=results(obs=10);
run;

/*remedy Non-Normality to make it follow normality*/
/*before*/
proc univariate data=A plot normal;
var r;
run;
/*remedy Non-Normality to make it follow normality*/
/*before*/
proc univariate data=A plot normal;
var r;
run;

/*now*/
DATA C1; SET C;
PSA_log = LOG(PSA);
RUN;

PROC REG Data=C1 noprint;
MODEL PSA_log = CANCER WT AGE BENIGN CAPSULAR;
OUTPUT OUT=Q RSTUDENT=R PREDICTED=P; /*predicted (yhat) and residuals*/
RUN;
proc univariate data=q plot normal;
var r;
run;
proc print data=C(obs=20);
run;
```

```
proc print data=C1(obs=20);
run;

proc univariate data=C1;
var PSA_log;
run;

DATA X1; SETC1;
Group = (PSA_log > 2.5); /* median r = 2.5 from above */
RUN;

PROC GLM Data=X1;
class Group;
model R=Group;
means Group / hovtest=BF; /*Brown-Forsythe Test*/
run;




/*Question 4*/
PROC GLM DATA=C1; /* both Type-I and Type-III SS along with partial F tests */
MODEL PSA_log = CANCER WT AGE BENIGN CAPSULAR;
RUN;

PROC GLM DATA=C1; /* both Type-I and Type-III SS along with partial F tests */
MODEL PSA_log = CANCER  AGE BENIGN CAPSULAR;
RUN;

PROC GLM DATA=C1; /* both Type-I and Type-III SS along with partial F tests */
MODEL PSA_log = CANCER   BENIGN CAPSULAR;
RUN;

PROC GLM DATA=C1; /* both Type-I and Type-III SS along with partial F tests */
```

```
MODEL PSA_log = CANCER   BENIGN ;
RUN;

/*Question 5*/
/* gives partial correlation coefficient for all combos*/
proc corr data=C1;
var PSA_log CANCER;
partial WT AGE BENIGN CAPSULAR ;
run;
proc corr data=C1;
var PSA_log WT;
partial CANCER  AGE BENIGN CAPSULAR;
run;
proc corr data=C1;
var PSA_log AGE;
partial CANCER WT  BENIGN CAPSULAR;
run;
proc reg data=C1;
model PSA_log= CANCER WT AGE BENIGN CAPSULAR;
run;
/* second part of q 5 */



proc reg data=C1;
model PSA_log =CANCER ;
output out=c1 residual=res1;
run;

proc reg data=C1;
model PSA_log =BENIGN ;
output out=c1 residual=res1;
run;

proc reg data=C1;
```

```
model BENIGN=CANCER;

output out=c2 residual=res2;

run;


proc print data=c1(obs=10);

Title "Y regressed on CANCER AND BENIGN";

run;


proc print data=c2(obs=10);

Title "BENIGN regressed on CANCER";

run;


data ress; merge c1 c2;

run;

proc corr data=ress;

var res1 res2;

run;



/*next part*/

proc reg data=C1;

model PSA_LOG=CANCER BENIGN;

output out=d1 predicted=p;

run;

proc print data=d1(obs=10);

Title "Transformed Y and Final chosen model Yhat ";

run;

proc corr data=d1;

var p PSA_LOG;

run;

/* other way to do it*/

PROC CORR DATA=C1;

VAR PSA_log CANCER;

PARTIAL BENIGN;

RUN; /* Partial correlation coeff for horsepower */
```

```
PROC CORR DATA=C1;
VAR PSA_log BENIGN;
PARTIAL CANCER;
RUN; /* Partial correlation coeff for weight */

/* Alt explanantion of partial determination for var = weight */
PROC REG DATA=C1 NOPRINT;
MODEL PSA_log = CANCER;
OUTPUT OUT=R23 RESIDUAL=Res23;
RUN;
PROC REG DATA=C1 NOPRINT;
MODEL BENIGN = CANCER;
OUTPUT OUT=R1 RESIDUAL=Res1;
RUN;
DATA q4; MERGE R1 R23;
OUTPUT;
PROC CORR DATA=q4;
VAR Res1 Res23;
RUN;
PROC REG DATA=q4;
MODEL PSA_log = CANCER BENIGN;
OUTPUT OUT=q4 PREDICTED=P;
RUN;
PROC CORR DATA=q4;
VAR PSA_log P;
RUN;



/*Question 6*/
data C2; set C1;
drop WT AGE CAPSULAR;
run;
proc print data=C2(obs=20);
run;
```

```
/*95% interval estimates*/
proc reg data=C2 ALPHA=0.05;
model PSA_log=CANCER BENIGN/clm cli;
output out=xx predicted=p stdi=se lclm=lwrbnd_ci uclm=uprbnd_ci lcl=lwrbnd_pi ucl=uprbnd_pi;
run;
proc print data=xx(obs=20);
run;
/*add working hoelting confidence bands*/
data xx; set xx;
WHLwr=p-(sqrt(finv(0.95,3,97-3)*4)*se);
WHUpr=p+(sqrt(finv(0.95,3,97-3)*4)*se);
run;
proc print data=xx(obs=10);
Title "Table of Lower and Upper Bounds for CI, PI, and Working Hoelting Conf.Bands";
run;
proc sgplot data=xx;
Title "95% Interval Est. of Mean Resp vs cancer";
scatter x=CANCER y=p/ yerrorlower=lwrbnd_ci yerrorupper=uprbnd_ci markerattrs= (COLOR=red) legendlabel="CI est";
scatter x=CANCER y=p/ yerrorlower=lwrbnd_pi yerrorupper=uprbnd_pi markerattrs= (COLOR=blue) legendlabel="PI est";
scatter x=CANCER y=p/ yerrorlower=WHLwr yerrorupper=WHUpr markerattrs= (COLOR=green) legendlabel="Work_Hoel est";
run;
proc sgplot data=xx;
Title "95% Interval Est. of Mean Resp vs bENIGN";
scatter x=BENIGN y=p/ yerrorlower=lwrbnd_ci yerrorupper=uprbnd_ci markerattrs= (COLOR=red) legendlabel="CI est";
scatter x=BENIGN y=p/ yerrorlower=lwrbnd_pi yerrorupper=uprbnd_pi markerattrs= (COLOR=blue) legendlabel="PI est";
scatter x=BENIGN y=p/ yerrorlower=WHLwr yerrorupper=WHUpr markerattrs= (COLOR=green) legendlabel="Work_Hoel est";
run;

PROC CAPABILITY DATA=C2;
INTERVALS BENIGN/METHOD=2;
RUN;

PROC SGPLOT DATA=C2;
REG X=BENIGN Y=PSA_LOG/CLM CLI;
RUN;
```

```
/*CONFIDENCE BAND*/

PROC SGPLOT DATA=C2;

REG X=CANCER Y=PSA_LOG/CLM CLI;

RUN;


/*q 7*/

PROC REG DATA=C2;

MODEL PSA_LOG=CANCER BENIGN/XPX;

RUN;


/*Question 7*/

DATA C1; SET C;

PSA_log = LOG(PSA);

RUN;


PROC REG DATA=C1 OUTEST=est1;

MODEL PSA_log=CANCER BENIGN;

RUN;


/*reg coeff and s*/

data est1; set est1;

s = _rmse_; /* root MSE = estimated standard deviation */

b0 = intercept; /* estimated intercept b0 */

b1 = CANCER; /* estimated slope b1 */

b4 = BENIGN; /* estimated slope b4 */

KEEP s b0 b1  b4;

RUN;


/*reg coeff and s*/

proc print data=est1;

Title "Regression Coeffs and S";

run;


proc reg data=C1 OUTSSCP= est3 ;
```

```
MODEL PSA_log= CANCER BENIGN/XPX;

run;

data est3; set est3;

drop PSA_log;

run;


/*deleting unneeded rows and columns*/

data est3; set est3;

if _NAME_ = 'PSA_log' then delete;

if _TYPE_ ='N' then delete;

drop _TYPE_;

run;

proc print data=est3;

Title "X'X Matrix";

run;

data est1; set est1;

critval=finv(0.95,5,97-5)*5*(0.83488)**2;

run;

proc print data=est1;

Title "Regression Coeffs and S and critical value";

run;

proc iml;

XX={97 678.87 245.87 678.87 10713.59 1415.7 245.868 1415.27 1505.26};

b={1.53090, 0.10105, 0.094904};

/*test if slopes equal to zero*/

teststat=b`XX*b;

print teststat;

run;


PROC REG DATA=C1;

MODEL PSA_log= CANCER BENIGN / XPX ;

RUN;


PROC REG DATA=C1;

MODEL PSA_log= CANCER BENIGN / I ;
```

```
RUN;
```