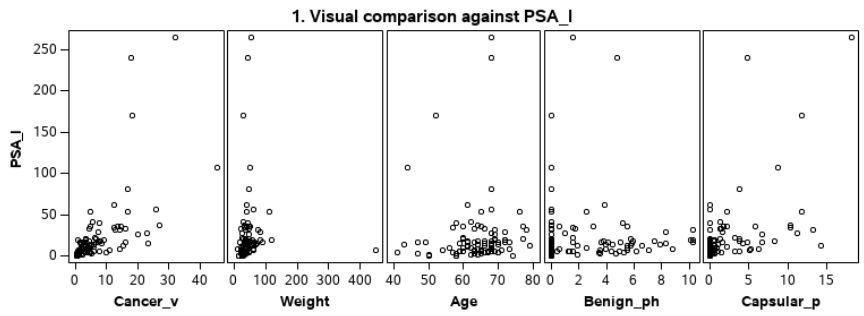
Michael Shoemate

Professor Choudhary

Advanced Statistical Methods I Project II

I am assuming a significance level of α = .05 throughout the project. Extensive SAS output for statistical tests like Brown-Forsythe, Shapiro-Wilke, and Breusch Pagan have been omitted, their results are tabularized throughout the report. In all questions, there is no repetition in the predictors, so the lack of fit test is not used. I computed the Breusch-Pagan statistic via the Lagrange Multiplier method: N\*R², where R² is from the regression E[ε²|X].

**Question 1**



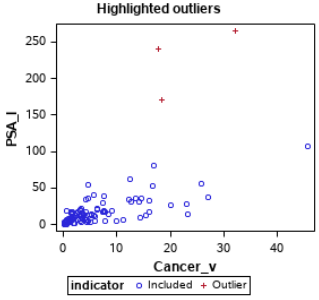
**Correlations**



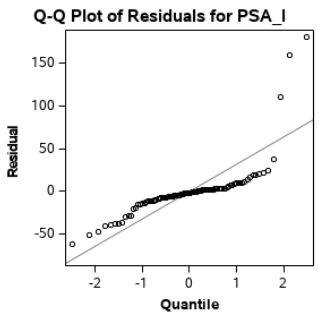
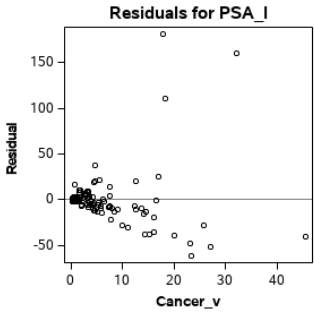
Judging from the plots and maximum correlation, Cancer\_v is the best predictor for PSA\_l.

**Question 1.a**

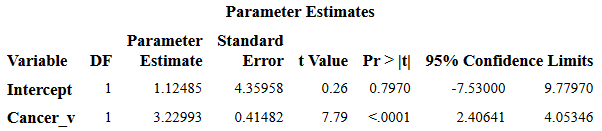
In this problem, I am considering any point with an absolute studentized residual greater than 2. This level seems to be the default for SAS and Minitab and is provided in the example SAS notes. In this problem the Bonferroni intervals have unreasonably many outlier points.



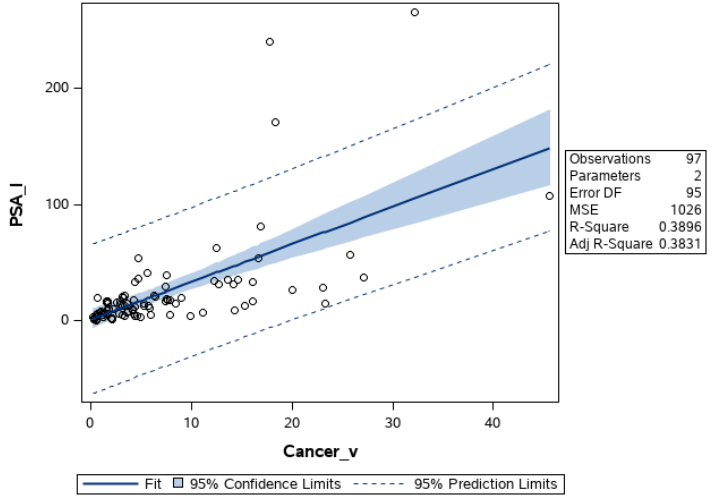
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | F 11.32 | .0011 | No |
| Homoscedasticity | Breusch-Pagan | X²(1) 18.22 | <.0001 | No |
| Normality | Shapiro-Wilke | F .626 | <.0001 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |
|  |  |  |  |  |

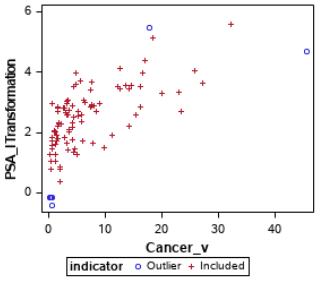
I fit the regression model even though all assumptions were violated. The F-statistic for model significance is 60.63 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .3896 indicates ~39% of variability is explained by the model.



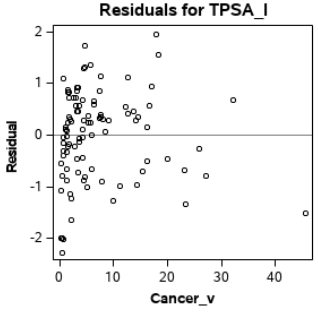
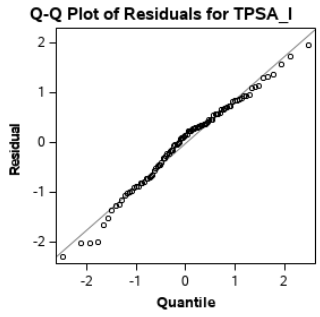
**Fit Plot for PSA\_l**



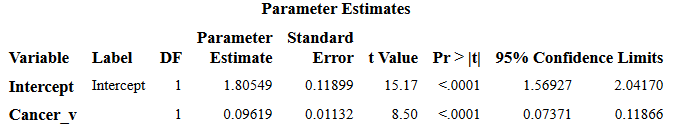
To attempt to address the violations with regression assumptions, I applied a log transformation on the dependent variable, PSA\_l and reapplied the above analysis. The transformed variable is TPSA\_l. The transformed space, with outliers recomputed is shown below. Take note that there are more outliers in the projected space than in the original space.



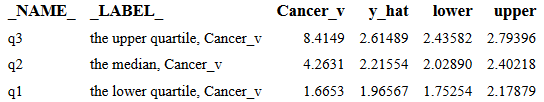
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | .04 | .8448 | Yes |
| Homoscedasticity | Breusch-Pagan | X²(2) .2174 | .6451 | Yes |
| Normality | Shapiro-Wilke | .9828 | .2360 | Yes |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

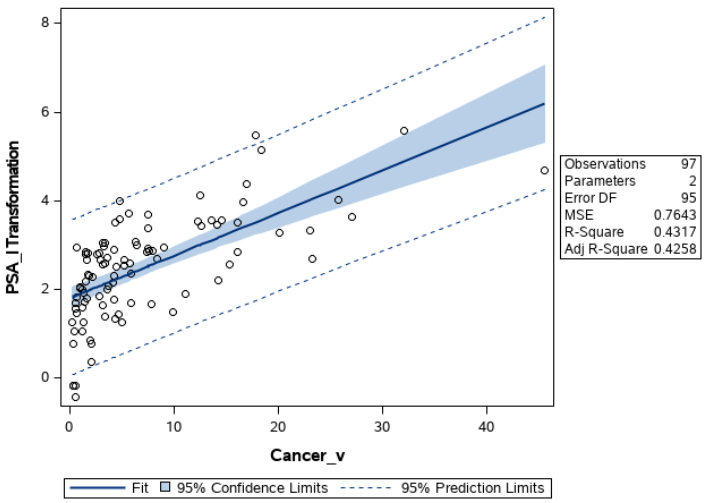


The revised F-statistic for model significance is 72.18 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the revised model is significant. The R² value of .4317 indicates ~43% of variability is explained by the model, which is an improvement over the original model.



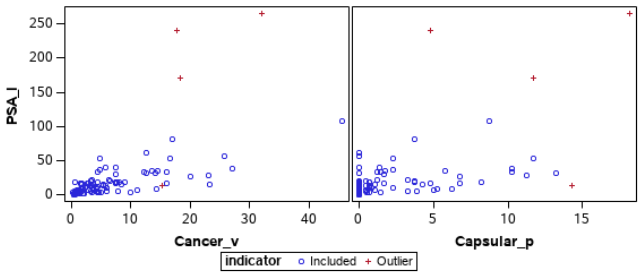
**Quartile Prediction Intervals at α = .05**



**Fit Plot for PSA\_l Under Box-Cox Transformation**

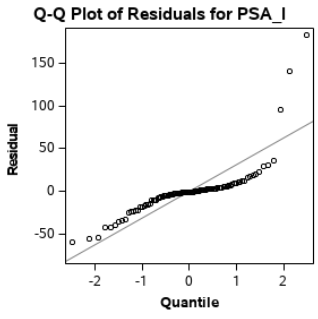
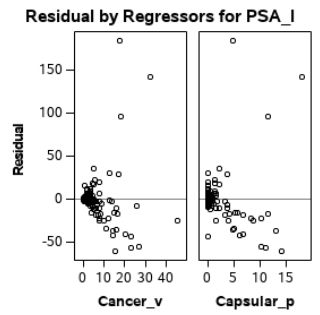
**Question 1.b**

Judging from the scatter plots and correlations, Capsular\_p is the second-best variable for regression. One thing to consider is the correlation between Capsular\_p and Cancer\_v, which is .69. If there is a variable that has significantly less correlation with the other predictors and significant correlation with the dependent variable, it may produce a better R² than Capsular\_p. Regardless, we will continue with PSA\_l vs Cancer\_v, Capsular\_p.

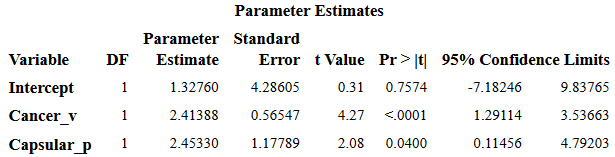
I made two scatterplots with the different axes for X. An observation/outlier may appear to be within the data on one axis but is extreme when viewed from a different axis.

I ran Brown-Forsythe for both Cancer\_v and Capsular\_p:

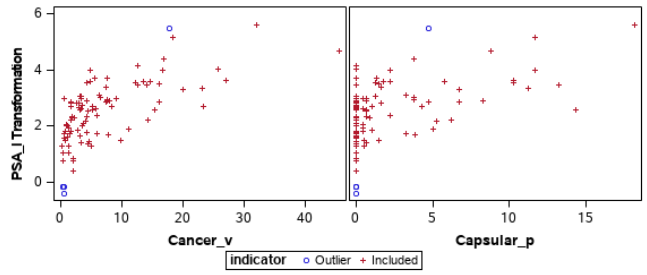
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe CV | 13.69 | .0004 | No |
|  | Brown-Forsythe CP | 11.5 | .0010 | No |
| Homoscedasticity | Breusch-Pagan | 19.77 | <.0001 | No |
| Normality | QQPlot |  |  | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | No |

Just like in part A, I fit the regression model even though all assumptions were violated. The F-statistic for model significance is 33.55 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .4165 indicates ~42% of variability is explained by the model. Take note the large variability the intercept estimate.

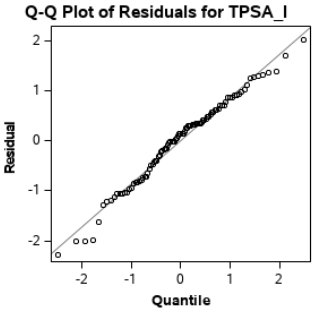
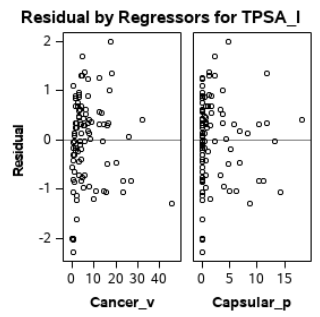


Now in the multivariate setting, I found that a Box-Cox transform on the dependent variable is sufficient to address most of the assumptions. Again outliers are shown amongst the scatterplots in the projected space.

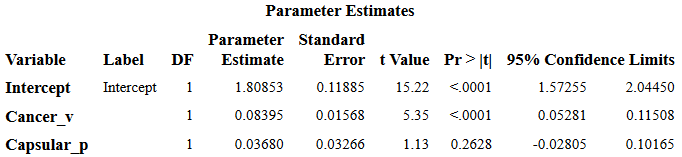


I ran Brown-Forsythe for both Cancer\_v and Capsular\_p:

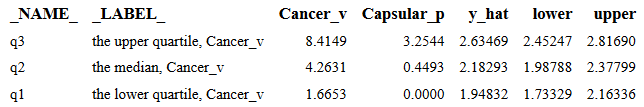
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe CV | .03 | .8728 | Yes |
|  | Brown-Forsythe CP | .15 | .6987 | Yes |
| Homoscedasticity | Breusch-Pagan | .47 | .7870 | Yes |
| Normality | QQPlot |  |  | Yes |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

I still fit the regression in spite of nonlinearity in the residuals. The F-statistic for model significance is 36.83 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .4393 indicates ~44% of variability is explained by the model. The transformation only marginally improved the R² value. Notice the parameter estimate of Capsular\_p is not significant, since H₀: β₂ = 0 is accepted with a p-value = .76. It can equivalently be said that β₂ is not significant because the confidence interval contains zero. We can conclude Capsular\_p does not have a significant effect on the model.

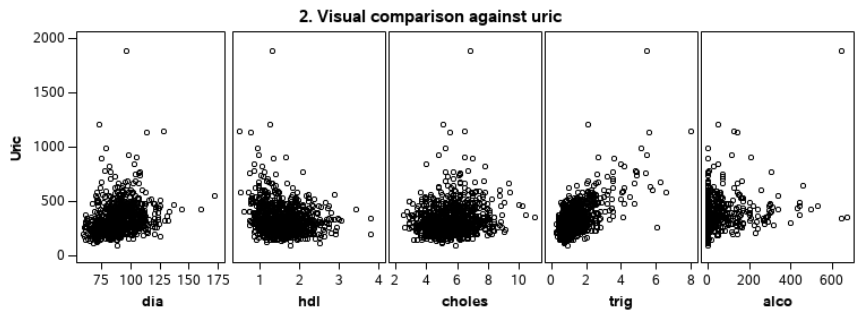


**Quartile Mean Intervals at α = .05**



The model from 1.a is nested within the model from 1.b. We can also say SS(extra) the reduction of error when Capsular\_p is added to the transformed model is 97469 – 93170 = 4299. A partial F-test in this case is equivalent to conducting a T test on H₀: β₂ = 0, shown above.

**Question 2**



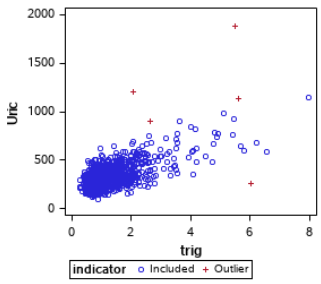
**Correlations**



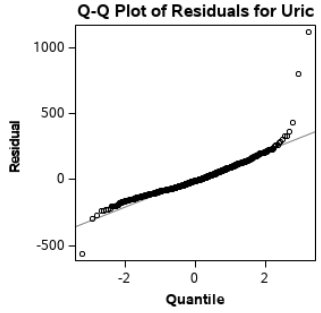
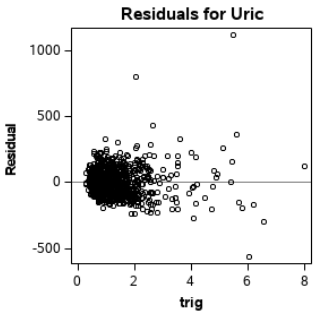
Judging from the plots and predictor with maximum absolute correlation, trig is the best predictor for Uric.

**Question 2.a**

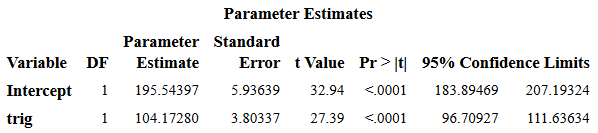
On the basis of question 2.b, I am considering the 5 points with the greatest absolute studentized residuals as outliers.



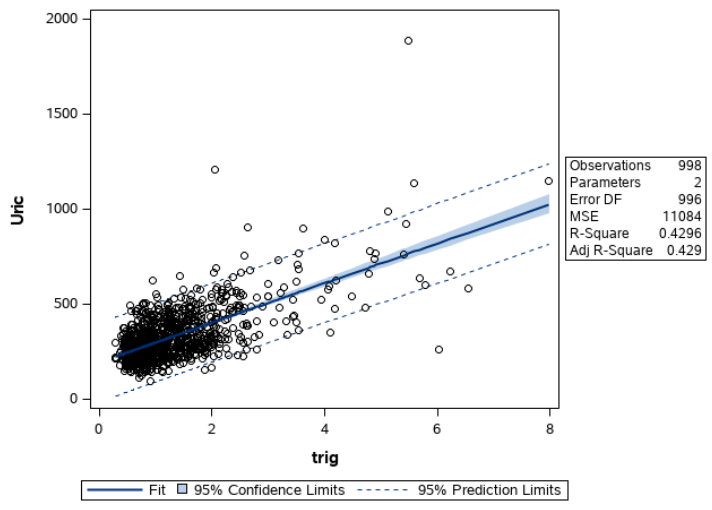
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 51.47 | <.0001 | No |
| Homoscedasticity | Breusch-Pagan | 76.504 | <.0001 | No |
| Normality | Shapiro-Wilke | .9011 | <.0001 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |
|  |  |  |  |  |

I fit the regression model even though all assumptions were violated. The F-statistic for model significance is 750.19 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .4296 indicates ~43% of variability is explained by the model.

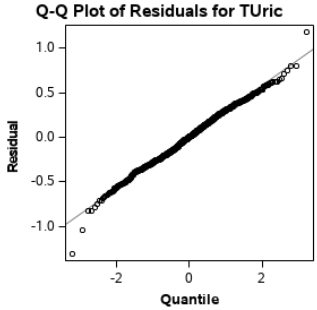
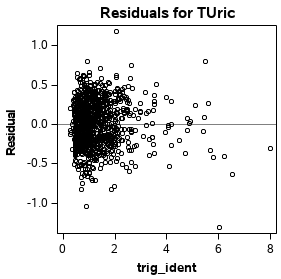


**Fit Plot for Uric**



I looked for a suitable transformation to satisfy the assumptions of a linear model but could not find one. However, I have documented the process. Initially to address normality I applied a Box-Cox transformation on Y. The selected λ is 0, so Uric is transformed to log(Uric) = TUric.

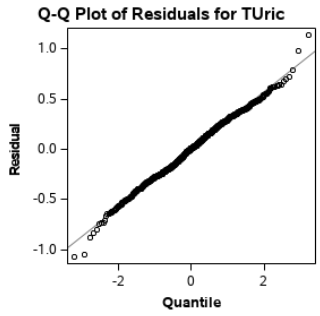
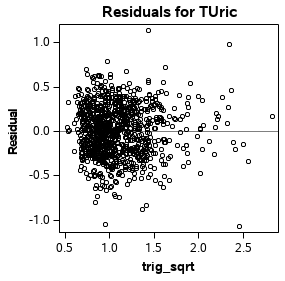
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 5.34 | .0211 | No |
| Homoscedasticity | Breusch-Pagan | 18.16 | <.0001 | No |
| Normality | Shapiro-Wilke | .995 | .0035 | No |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

The QQPlot looks much better than before, but Shapiro-Wilke still doesn’t pass. We will continue with this transformation on Uric and attempt to address homoscedasticity and linearity by checking some transforms on trig.

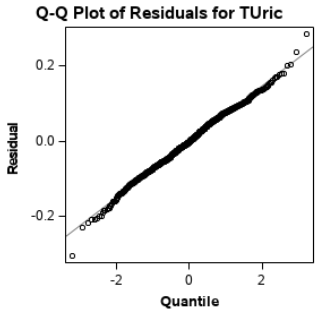
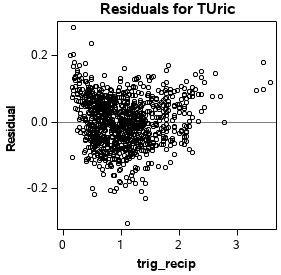
**BoxCox(Uric) = sqrt(trig**) **where λ = 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 5.06 | .0247 | No |
| Homoscedasticity | Breusch-Pagan | 8.75 | .0031 | No |
| Normality | Shapiro-Wilke | .996 | .0152 | No |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

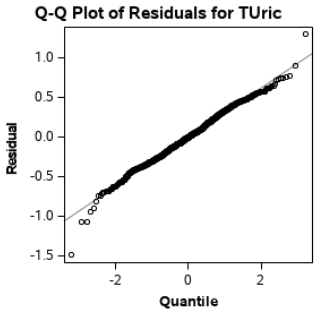
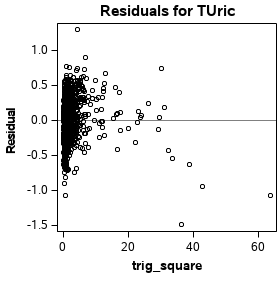
**BoxCox(Uric) = 1/trig where λ = -.25**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 1.32 | .2503 | Yes |
| Homoscedasticity | Breusch-Pagan | .96 | .3269 | Yes |
| Normality | Shapiro-Wilke | .996 | .0181 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |

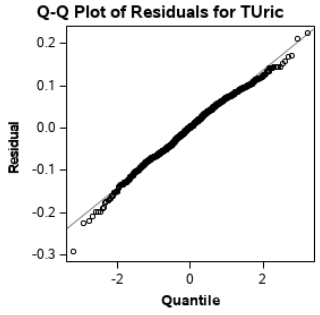
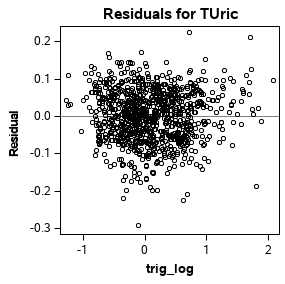
**BoxCox(Uric) = trig² where λ = 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 7.14 | .0077 | No |
| Homoscedasticity | Breusch-Pagan | 87.61 | <.0001 | No |
| Normality | Shapiro-Wilke | .994 | .0004 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | No |

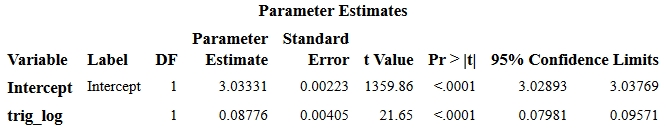
**BoxCox(Uric) = log(trig) where λ = -.25**

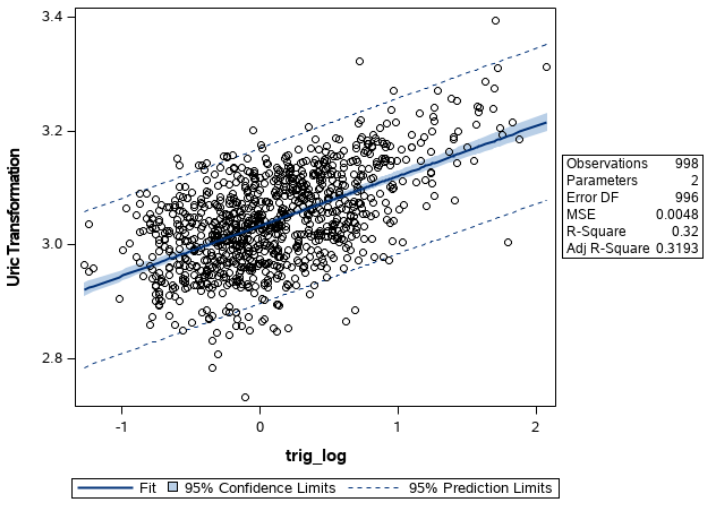
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | .15 | .6959 | Yes |
| Homoscedasticity | Breusch-Pagan | .279 | .5971 | Yes |
| Normality | Shapiro-Wilke | .994 | .0004 | No |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

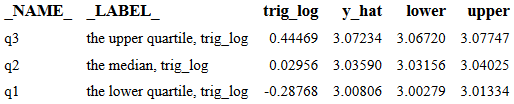
Even though it doesn’t pass Shapiro-Wilke for normality, the best transformation among these is the log, since it passes homoscedasticity checks and the residual plot looks unstructured.

I fit the regression model on BoxCox(Uric) = log(trig) even though some assumptions were still violated. The F-statistic for model significance is 468.73 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .32 indicates ~32% of variability is explained by the model. Take note that after the transformations to attempt to meet the assumptions of the model, the R² value dropped.





**Quartile Mean Intervals at α = .05**



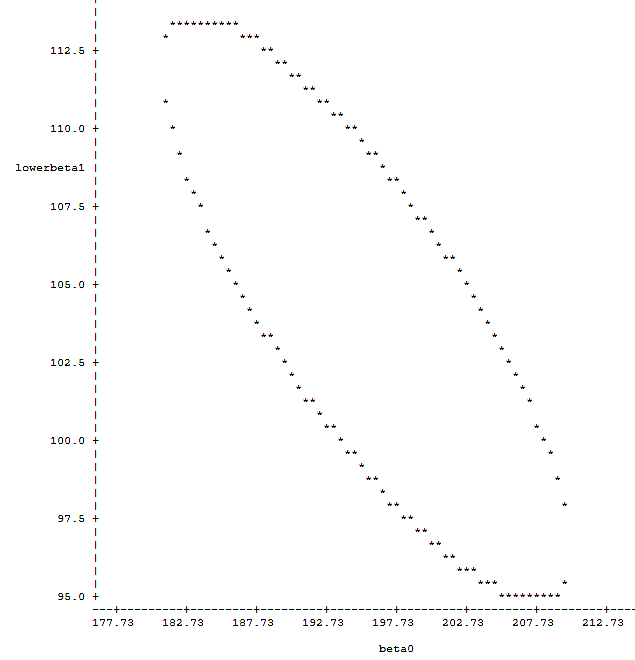
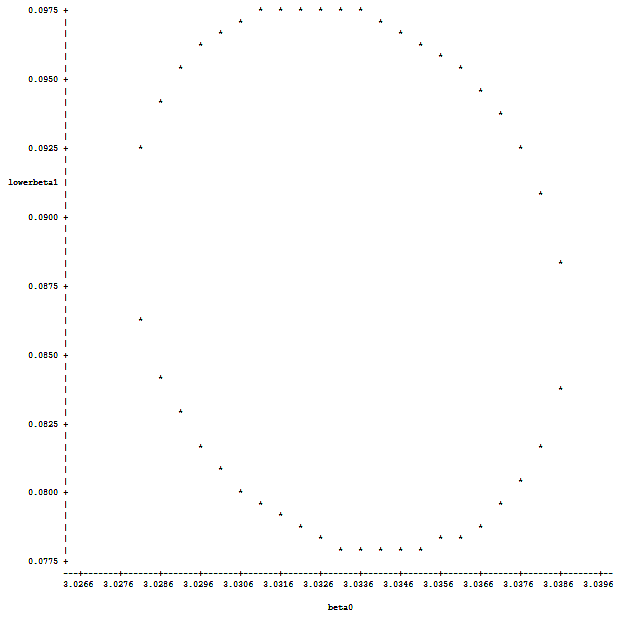
I computed a 95% confidence ellipse for the parameters β₀ and β₁ of the regression equation Uric=trig, as well as BoxCox(Uric)=log(trig). Both tests are staged as follows:

H₀: β = (100, 100)

H₁: β ≠ (100, 100)

Plot 1. Before transformation: β₀ axis ∈ (177, 212), β₁ axis ∈ (95, 112)

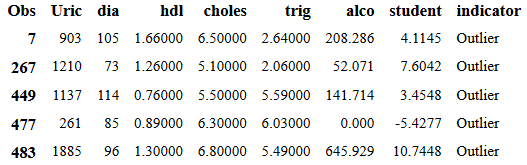
Plot 2. After transformation: β₀ axis ∈ (3.0266, 3.0396), β₁ axis ∈ (.0775, .0975)

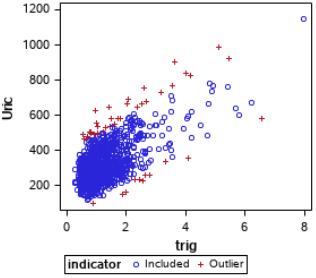
The point (100, 100) is clearly outside the 95% confidence ellipse of both, therefore we reject H₀ for both tests and conclude β ≠ (100, 100).

**Question 2.b**

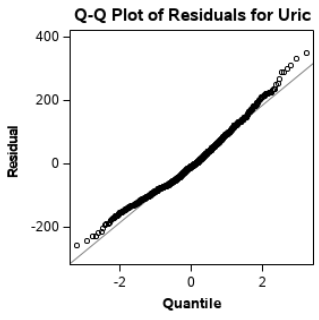
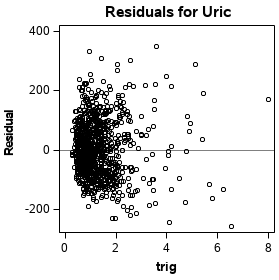
I removed the five points with the greatest absolute residuals:



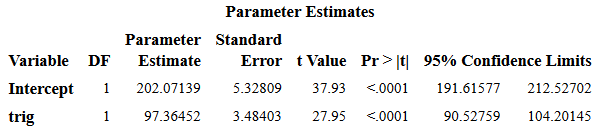
The following is a scatterplot with the outliers removed. All remaining observations with a studentized residual greater than 2 are marked with a cross.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 59.76 | <.0001 | No |
| Homoscedasticity | Breusch-Pagan | 103.65 | <.0001 | No |
| Normality | Shapiro-Wilke | .9842 | <.0001 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |

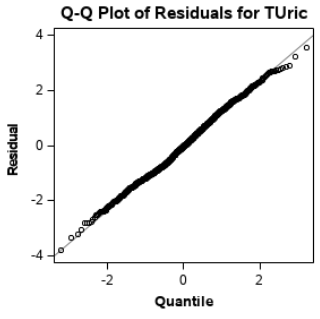
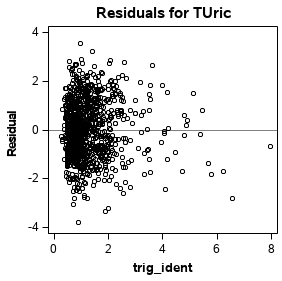
 

I fit the regression model on Filtered Uric = trig even though some assumptions were still violated. The F-statistic for model significance is 780.97 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .44 indicates ~44% of variability is explained by the model. Compare this to the fit with the five outliers, the variability explained has increased by 1%.



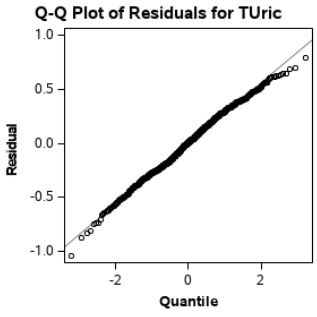
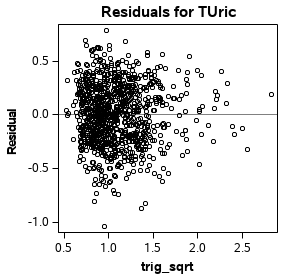
**Filtered BoxCox(Uric) = trig where λ = 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 11.31 | .0008 | No |
| Homoscedasticity | Breusch-Pagan | 8.736 | .0031 | No |
| Normality | Shapiro-Wilke | .996 | .0169 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |

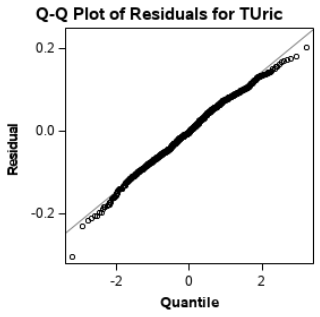
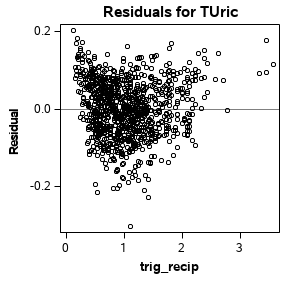
**Filtered BoxCox(Uric) = sqrt(trig) where λ = 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 3.07 | .0799 | Yes |
| Homoscedasticity | Breusch-Pagan | .07 | .7911 | Yes |
| Normality | Shapiro-Wilke | .9958 | .0079 | No |
| Linearity | Residual Plot |  |  | No |
| Independence | Residual Plot |  |  | Yes |

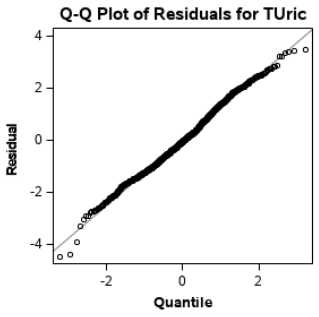
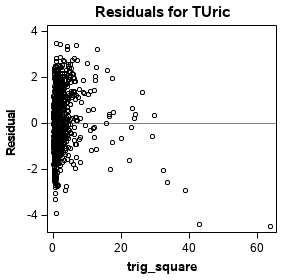
**Filtered BoxCox(Uric) = 1/trig where λ = -.25**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | .6 | .44 | Yes |
| Homoscedasticity | Breusch-Pagan | .0387 | .844 | Yes |
| Normality | Shapiro-Wilke | .9944 | .001 | No |
| Linearity | Residual Plot |  |  | No |
| Independent | Residual Plot |  |  | No |
|  |  |  |  |  |

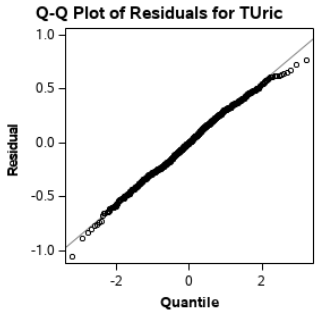
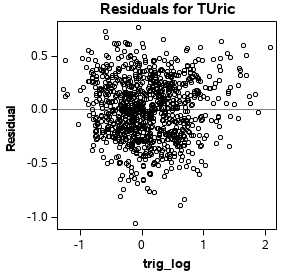
**Filtered BoxCox(Uric) = trig² where λ = .25**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 13.88 | .0002 | No |
| Homoscedasticity | Breusch-Pagan | 87.10 | <.0001 | No |
| Normality | Shapiro-Wilke | .9947 | .0014 | No |
| Linearity | Residual Plot |  |  | No |
| Independent | Residual Plot |  |  | No |

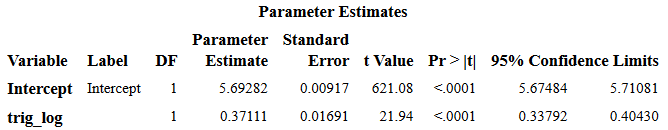
 

**Filtered BoxCox(Uric) = log(trig) where λ = 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe | 2.57 | .1091 | Yes |
| Homoscedasticity | Breusch-Pagan | 1.544 | .214 | Yes |
| Normality | Shapiro-Wilke | .9959 | .00104 | No |
| Linearity | Residual Plot |  |  | Yes |
| Independence | Residual Plot |  |  | Yes |

The best transformation among these is still BoxCox(Uric) = log(trig). Since λ = 0, this is log(Uric) = log(trig). The homoscedasticity tests are better for 1/trig, but the residual plot looks worse. The F-statistic for model significance is 481.45 with a p-value < .0001. Therefore H₀: β₁ = 0 is rejected, and we conclude that the model is significant. The R² value of .327 indicates ~33% of variability is explained by the model. The difference between this model and the transformed fit with the five outliers is negligible, I would prefer not removing the outliers since they don’t have a large effect.



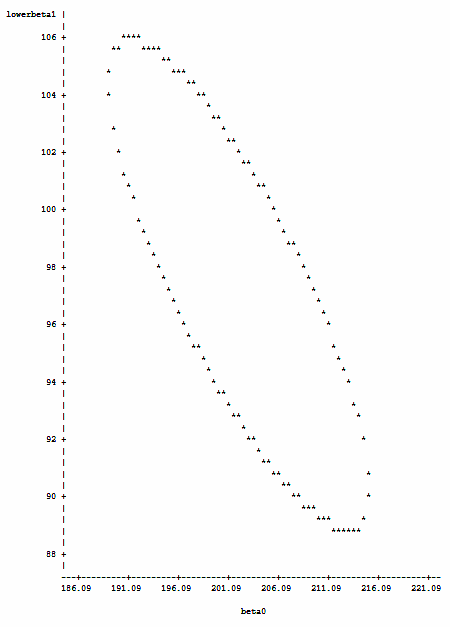
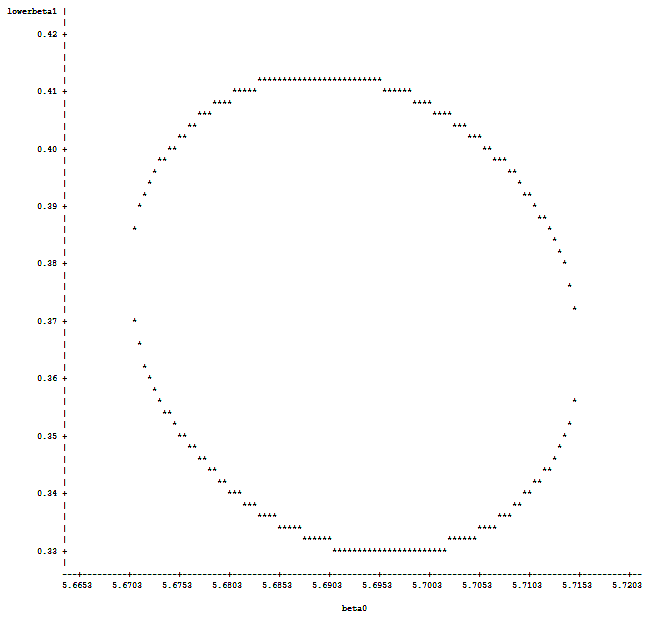
After transforming, I computed a 95% confidence ellipse for the parameters β₀ and β₁ of the regression equation Uric=trig, as well as BoxCox(Uric)=log(trig). Both tests are staged as follows:

H₀: β = (100, 100)

H₁: β ≠ (100, 100)

Plot 1. Before transformation: β₀ axis ∈ (186, 221), β₁ axis ∈ (88, 106)

Plot 2. After transformation: β₀ axis ∈ (5.663, 5.7202), β₁ axis ∈ (.33, .42)

The point (100, 100) is clearly outside the 95% confidence ellipse of both, therefore we reject H₀ for both tests and conclude β ≠ (100, 100).

Michael Shoemate

Professor Choudhary

Advanced Statistical Methods I Project II – SAS Code

I modularized the code into macros: outliers, diagnostics, assumption\_homoscedasticity, assumption\_normality, assumption\_linear\_indep, fit, test\_transformations and ellipse. Each problem also has its own macro at the end.

%LET seed=160830;

DATA prostate;

INFILE '/folders/myfolders/Prostate.dat';

INPUT ID PSA\_l Cancer\_v Weight Age Benign\_ph Seminal\_vi Capsular\_p Gleason\_s;

DATA cardio\_all;

INFILE '/folders/myfolders/Cardiodata.csv' DSD FIRSTOBS=2;

INPUT age bmi waisthip smok choles trig hdl ldl sys dia Uric sex alco apoa;

PROC SQL;

CREATE TABLE cardio AS SELECT uric, dia, hdl, choles, trig, alco FROM cardio\_all;

%MACRO outliers(datatable, predictors, predicted, cutoff);

PROC SQL;

ALTER TABLE &datatable DROP student, residuals, indicator;

ODS EXCLUDE ALL;

PROC REG DATA=&datatable

PLOTS(label)=RStudentByLeverage;

/\* ODS SELECT WHERE=(\_name\_ ? 'RStudentByLeverage'); \*/

MODEL &predicted=&predictors;

OUTPUT OUT=&datatable STUDENT=student;

RUN;

ODS EXCLUDE NONE;

/\* add an indicator variable for outliers \*/

PROC SQL;

ALTER TABLE &datatable ADD indicator VARCHAR(10);

UPDATE &datatable SET indicator = CASE WHEN ABS(student) > &cutoff THEN 'Outlier' ELSE 'Included' END;

RUN;

TITLE2 "Highlighted outliers";

ODS GRAPHICS ON / WIDTH=3in HEIGHT=3in;

PROC SGSCATTER DATA=&datatable;

COMPARE Y=&predicted X=(&predictors) / GROUP=indicator;

RUN;

TITLE2;

%MEND outliers;

%MACRO diagnostics(datatable, predictors, predicted, n);

ODS EXCLUDE ALL;

PROC REG DATA=&datatable;

MODEL &predicted=&predictors / INFLUENCE R;

ODS OUTPUT outputstatistics=results;

RUN;

ODS EXCLUDE NONE;

DATA results; set results;

%LET p = %LENGTH(&predictors);

hilev = HatDiagonal > 2\*(&p/&n);

dfflag = ABS(DFFITS) > 1;

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

/\* check DFB for each parameter \*/

%DO i=1 %TO %SYSFUNC(COUNTW(&predictors));

%LET var = %SCAN(&predictors, &i);

&var.\_flag = ABS(DFB\_&var > 1);

IF &var.\_flag THEN flag = 1;

%END;

RUN;

TITLE2 'Diagnostics: flagged observations';

PROC PRINT DATA=results NOOBS;

WHERE hilev=1 OR dfflag=1 OR Fpercent>20 OR flag=1;

VAR Observation DepVar CooksD HatDiagonal DFFITS Fpercent;

RUN;

TITLE2;

%MEND diagnostics;

%MACRO assumption\_homoscedasticity(datatable, predictors, predicted);

PROC SQL;

ALTER TABLE &datatable DROP rstudent, residual;

ODS EXCLUDE ALL;

PROC REG DATA=&datatable;

MODEL &predicted=&predictors;

/\* ODS SELECT WHERE=(\_name\_ = 'ANOVA'); \*/

OUTPUT OUT=&datatable RSTUDENT=rstudent RESIDUAL=residual;

RUN;

PROC SQL;

ALTER TABLE &datatable ADD residual\_square NUMERIC;

UPDATE &datatable SET residual\_square=residual\*residual;

PROC REG DATA=&datatable OUTEST=outest;

MODEL residual\_square=&predictors / RSQUARE;

RUN;

ODS EXCLUDE NONE;

PROC IML;

USE &datatable;

READ ALL VAR{&predicted};

CALL SYMPUTX("n", NROW(&predicted));

RUN;

TITLE2 "Homoscedasticity of &predicted Breusch-Pagan via Lagrange Multipliers";

PROC IML;

USE outest;

READ ALL VAR {\_RSQ\_};

bp\_test = &n \* \_RSQ\_;

PRINT bp\_test;

bp\_p\_value = 1 - CDF("chisquare", bp\_test, %SYSFUNC(COUNTW(&predictors)));

PRINT bp\_p\_value;

QUIT;

%DO j=1 %TO %SYSFUNC(COUNTW(&predictors));

%LET variable = %SCAN(&predictors, &j);

PROC IML;

USE &datatable;

READ ALL VAR {&variable};

CALL SYMPUTX("median", MEDIAN(&variable));

QUIT;

DATA &datatable; SET &datatable;

group = &variable > &median;

RUN;

TITLE2 "Homoscedasticity of &predicted vs &variable Brown-Forsythe";

PROC GLM DATA=&datatable;

ODS SELECT WHERE=(\_name\_ ? 'HOVFTest');

CLASS group; MODEL rstudent=group; MEANS group / HOVTEST=BF;

RUN;

%END;

PROC SQL;

ALTER TABLE &datatable DROP rstudent, residual, residual\_square, group;

TITLE2;

%MEND assumption\_homoscedasticity;

%MACRO assumption\_normality(datatable, predictors, predicted);

TITLE2 "Normality of &predicted";

ODS EXCLUDE ALL;

PROC REG DATA=&datatable;

MODEL &predicted=&predictors;

OUTPUT OUT=&datatable RESIDUAL=residual;

RUN;

ODS EXCLUDE NONE;

PROC UNIVARIATE DATA=&datatable NORMAL;

ODS SELECT WHERE=(\_name\_ = 'TestsForNormality');

VAR residual;

RUN;

ODS GRAPHICS ON / WIDTH=3in HEIGHT=3in;

PROC REG DATA=&datatable

PLOTS(label)=QQPlot;

ODS SELECT WHERE=(\_name\_ ? 'QQPlot');

MODEL &predicted=&predictors;

RUN;

PROC SQL;

ALTER TABLE &datatable DROP residual;

ODS GRAPHICS OFF;

TITLE2;

%MEND assumption\_normality;

%MACRO assumption\_linear\_indep(datatable, predictors, predicted);

PROC SQL;

ALTER TABLE &datatable DROP residual;

RUN;

ODS GRAPHICS ON / WIDTH=3in HEIGHT=3in;

PROC REG DATA=&datatable;

MODEL &predicted=&predictors;

OUTPUT OUT=&datatable RESIDUAL=residual;

ODS SELECT WHERE=(\_name\_ = 'ResidualPlot');

RUN;

TITLE2;

%MEND assumption\_linear\_indep;

%MACRO fit(datatable, predictors, predicted);

ODS EXCLUDE ALL;

%DO j=1 %TO %SYSFUNC(COUNTW(&predictors));

%LET var = %SCAN(&predictors, &j);

PROC UNIVARIATE DATA=&datatable;

VAR &var;

OUTPUT OUT=quantiles\_&j P25=q1 P50=q2 P75=q3;

RUN;

%END;

ODS EXCLUDE NONE;

DATA quantiles\_h;

SET %DO j=1 %TO %SYSFUNC(COUNTW(&predictors));quantiles\_&j %END;;

PROC TRANSPOSE DATA=quantiles\_h OUT=quantiles;

/\* give names to columns \*/

DATA quantiles;

SET quantiles;

%DO j=1 %TO %SYSFUNC(COUNTW(&predictors));

%LET var = %SCAN(&predictors, &j);

&var = COL&j;

%END;

/\* remove unnamed columns and vertical stack quantiles and data \*/

DATA regdata;

SET &datatable quantiles (DROP=COL1 -- COL%SYSFUNC(COUNTW(&predictors)));;

ODS GRAPHICS / RESET=all;

PROC REG DATA=regdata OUTEST=est ALPHA=.05;

ODS SELECT WHERE=(\_name\_ = 'ANOVA' | \_name\_ = 'ParameterEstimates' | \_name\_ = 'FitStatistics' | \_name\_ = 'FitPlot');

MODEL &predicted=&predictors / CLB;

OUTPUT OUT=outdata PREDICTED=y\_hat LCLM=lower UCLM=upper;

RUN;

TITLE2 "Intervals for mean at quantile points";

PROC PRINT DATA=outdata NOOBS;

WHERE &predicted IS NULL;

VAR \_name\_ \_label\_ &predictors y\_hat lower upper;

RUN;

PROC SQL;

DROP TABLE outdata;

TITLE;

%MEND fit;

%MACRO test\_transformations(datatable, predictors, predicted);

%LET labels=ident sqrt recip square log;

PROC SQL;

ALTER TABLE &datatable ADD

&predictors.\_ident NUMERIC,

&predictors.\_sqrt NUMERIC,

&predictors.\_recip NUMERIC,

&predictors.\_square NUMERIC,

&predictors.\_log NUMERIC;

UPDATE &datatable SET

&predictors.\_ident=&predictors,

&predictors.\_sqrt=sqrt(&predictors),

&predictors.\_recip=1/&predictors,

&predictors.\_square=&predictors \* &predictors,

&predictors.\_log=log(&predictors);

%DO i=1 %TO %SYSFUNC(COUNTW(&labels));

%LET label = %SCAN(&labels, &i);

PROC SQL;

DROP TABLE transdata;

TITLE "BoxCox(y)=&label(x) transformed regression assumption checks";

PROC TRANSREG DATA=&datatable;

MODEL BoxCox(&predicted)=identity(&predictors.\_&label);

OUTPUT OUT=transdata;

RUN;

%assumption\_normality(transdata, &predictors.\_&label, T&predicted);

%assumption\_homoscedasticity(transdata, &predictors.\_&label, T&predicted);

%assumption\_linear\_indep(transdata, &predictors.\_&label, T&predicted);

RUN;

%END;

RUN;

%MEND test\_transformations;

%MACRO ellipse(datatable, predictors, predicted, stepsize);

PROC SQL;

DROP TABLE xdata, est;

ODS EXCLUDE ALL;

PROC MEANS DATA=&datatable;

VAR &predictors;

OUTPUT OUT=xdata N=n MEAN=xbar CSS=Sxx;

RUN;

PROC REG DATA=&datatable OUTEST=est;

MODEL &predicted=&predictors;

RUN;

ODS EXCLUDE NONE;

DATA est; SET est;

s = \_rmse\_; /\* root MSE = estimated standard deviation \*/

b0 = intercept; /\* estimated intercept b0 \*/

b1 = &predictors; /\* estimated slope b1 \*/

KEEP s b0 b1;

RUN;

DATA ellipse;

MERGE xdata est;

sb1 = s/SQRT(Sxx); /\* standard deviation of b1 \*/

sb0 = s\*SQRT(1/n+xbar\*\*2/Sxx); /\* standard deviation of b0 \*/

F95=finv(0.95,2,n-2)\*2\*s\*\*2; /\* 95% upper bound for the quadratic form \*/

DO beta0=b0-3\*sb0 BY &stepsize TO b0+3\*sb0; /\* for a fixed value of beta0, solve for beta1 using the quadratic form of ellipse \*/

D = (n\*xbar\*(beta0-b0))\*\*2 - (n\*xbar\*\*2+Sxx)\*(n\*(beta0-b0)\*\*2-F95); /\*discriminant \*/

if D < 0 then do; upperbeta1 = .; lowerbeta1 = .; end; /\* discard beta0 values with D<0 \*/

else do; upperbeta1 = b1+(n\*xbar\*(b0-beta0)+sqrt(D))/(n\*xbar\*\*2+Sxx);

lowerbeta1 = b1+(n\*xbar\*(b0-beta0)-sqrt(D))/(n\*xbar\*\*2+Sxx); end;

OUTPUT;

END;

RUN;

PROC PLOT DATA=ellipse; TITLE 'Confidence region';

PLOT (lowerbeta1 upperbeta1)\*beta0 = '\*' / overlay;

RUN;

TITLE;

%MEND ellipse;

%MACRO problem1();

%LET datatable = prostate;

%LET predicted = PSA\_l;

TITLE "1. Visual comparison against &predicted";

ODS GRAPHICS ON / WIDTH=8in HEIGHT=3in;

PROC SGSCATTER DATA=&datatable;

COMPARE Y=&predicted X=(Cancer\_v Weight Age Benign\_ph Capsular\_p);

RUN;

ODS GRAPHICS / RESET=all;

TITLE "1. &datatable correlation matrix";

ODS GRAPHICS ON;

ODS SELECT WHERE=(\_label\_ ? 'Pearson');

PROC CORR DATA=&datatable;

VAR PSA\_l Cancer\_v Weight Age Benign\_ph Seminal\_vi Capsular\_p Gleason\_s;

RUN;

ODS GRAPHICS OFF;

/\* correlation matrix and plots indicate that Cancer\_v is the best predictor \*/

%LET predictors=Cancer\_v;

/\* Rule of thumb is 2, could also use tinv for bonferroni, but would be far more restrictive \*/

TITLE "1.a outliers";

%outliers(&datatable, &predictors, &predicted, 2);

TITLE "1.a regression diagnostics";

%diagnostics(&datatable, &predictors, &predicted, n=97);

TITLE "1.a test assumptions";

%assumption\_homoscedasticity(&datatable, &predictors, &predicted);

%assumption\_normality(&datatable, &predictors, &predicted);

%assumption\_linear\_indep(&datatable, &predictors, &predicted);

TITLE "1.a regression fit &predicted vs &predictors";

%fit(&datatable, &predictors, &predicted);

PROC TRANSREG DATA=&datatable;

MODEL log(&predicted)=identity(&predictors);

OUTPUT OUT=transdata RESIDUALS;

RUN;

TITLE "1.a transformed outliers";

%outliers(transdata, &predictors, T&predicted, 2);

%assumption\_normality(transdata, &predictors, T&predicted);

%assumption\_homoscedasticity(transdata, &predictors, T&predicted);

TITLE "1.a transformed regression assumption checks";

%assumption\_linear\_indep(transdata, &predictors, T&predicted);

TITLE "1.a regression fit &predicted vs &predictors";

%fit(transdata, &predictors, T&predicted);

/\* PART B \*/

%LET predictors = Cancer\_v Capsular\_p;

TITLE "1.b outliers";

%outliers(&datatable, &predictors, &predicted, 2);

TITLE "1.b test assumptions";

%assumption\_homoscedasticity(&datatable, &predictors, &predicted);

%assumption\_normality(&datatable, &predictors, &predicted);

%assumption\_linear\_indep(&datatable, &predictors, &predicted);

TITLE "1.b regression fit &predicted vs &predictors";

%fit(&datatable, &predictors, &predicted);

PROC TRANSREG DATA=&datatable;

MODEL log(&predicted)=identity(&predictors);

OUTPUT OUT=transdata RESIDUALS;

RUN;

TITLE "1.a transformed outliers";

%outliers(transdata, &predictors, T&predicted, 2);

TITLE "1.b transformed regression assumption checks";

%assumption\_homoscedasticity(transdata, &predictors, T&predicted);

%assumption\_normality(transdata, &predictors, T&predicted);

%assumption\_linear\_indep(transdata, &predictors, T&predicted);

TITLE "1.b regression fit T&predicted vs &predictors";

%fit(transdata, &predictors, T&predicted);

%MEND problem1;

%MACRO problem2();

%LET basetable = cardio;

%LET predicted = uric;

TITLE "2. Visual comparison against &predicted";

ODS GRAPHICS ON / WIDTH=8in HEIGHT=3in;

PROC SGSCATTER DATA=&basetable;

COMPARE Y=&predicted X=(dia hdl choles trig alco);

RUN;

ODS GRAPHICS / RESET=all;

TITLE "2. &basetable correlation matrix";

ODS GRAPHICS ON;

ODS SELECT WHERE=(\_label\_ ? 'Pearson');

PROC CORR DATA=&basetable;

VAR uric dia hdl choles trig alco;

RUN;

ODS GRAPHICS OFF;

/\* correlation matrix and plots indicate that trig is the best predictor \*/

%LET predictor=trig;

%MACRO analysis(part);

TITLE "&part test assumptions";

%assumption\_homoscedasticity(&basetable, &predictor, &predicted);

%assumption\_normality(&basetable, &predictor, &predicted);

%assumption\_linear\_indep(&basetable, &predictor, &predicted);

TITLE "&part regression fit &predicted vs &predictor";

%fit(&basetable, &predictor, &predicted);

%ellipse(&basetable, &predictor, &predicted, 0.5);

%test\_transformations(&basetable, &predictor, &predicted);

%fit(transdata, &predictor.\_log, T&predicted);

TITLE "&part regression fit &predicted vs &predictor";

%ellipse(transdata, &predictor.\_log, T&predicted, 0.0005);

%MEND analysis;

TITLE "2.a outliers";

%outliers(&basetable, &predictor, &predicted, 3.4);

%analysis('2.a');

TITLE '2.b Observations removed';

PROC PRINT DATA=&basetable;

WHERE indicator='Outlier';

PROC SQL;

DELETE FROM &basetable WHERE indicator='Outlier';

TITLE "2.b outliers";

%outliers(&basetable, &predictor, &predicted, 2);

%analysis('2.b');

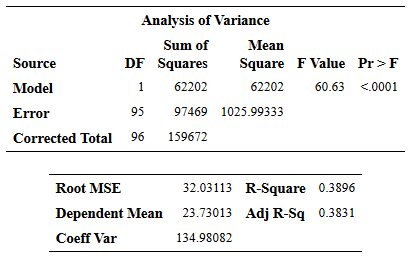
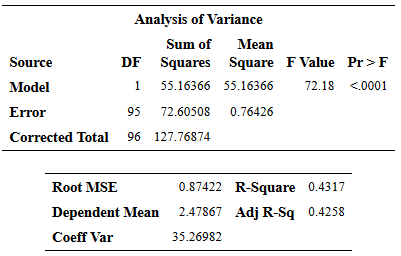
%MEND problem2;

%problem1();

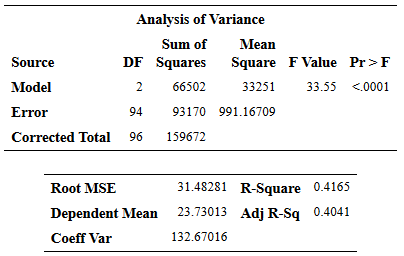
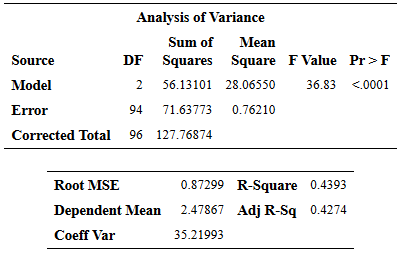
%problem2();

Advanced Statistical Methods I Project II – Supplementary SAS Output

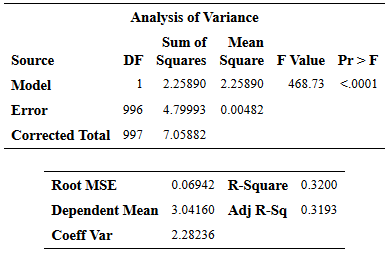
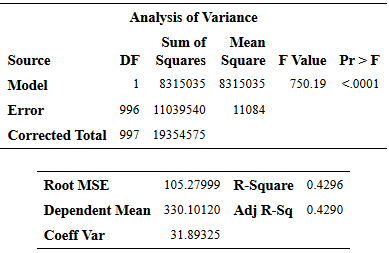
1.a. Regression diagnostics (untransformed and transformed):

1.b. Regression diagnostics (untransformed and transformed):

2.a. Regression diagnostics (untransformed and transformed):



2.b. Regression diagnostics (untransformed and transformed):

