Michael Shoemate

Professor Choudhary

Advanced Statistical Methods Project III

**Question 1**

I applied the best subset selection procedure to select a subset of models with the best performance. Since the number of predictors is still relatively small, I opted to run the best subset procedure to avoid local minima that may be encountered when conducting stepwise methods.

There is not enough information provided about the Gleason Score variable to determine if said variable is ordinal or categorical. However, model performance increases marginally when dummy variables are used in model selection compared to treating GS as an ordinal variable. Therefore, Gleason Score is modeled as a categorical variable. Further analysis uses said Gleason Score modeling assumption. Plots 1.a and 1.b are attached in the SAS output, that visualize the best scoring model for each variable count. The metrics demonstrate a preference toward a predictor count of three. There is a lesser preference toward a predictor count of five (MSE). Regardless of the metric used, the same models are selected at each step. The best performances for each metric at each variable count is available in SAS appendix 1.b.

Analysis is conducted on the model PSAL = CV SV GS\_8 first. Three observations are flagged as influencers (1.d). Observation 94 has a significant hat diagonal. Observation 96 has significant DFBetas scores for the SV and GS\_8 predictors. Observation 97 has an F-percent over 20%, which is computed from Cook’s D score. The flags indicate the observations have a significant effect on the model. The variance inflation factors do not indicate an issue with collinearity, since all VIFs are well below ten (1.e). There are three outliers, determined by having an absolute RStudentized residual greater than t(1-α/(2p), n) ≈ 2.545.

**Regression Assumption checks for PSAL = CV SV GS\_8**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe CV | 11.96 | 0.0008 | No |
|  | Brown-Forsythe SV | 21.28 | <.0001 | No |
|  | Brown-Forsythe GS\_8 | 21.31 | <.0001 | No |
| Homoscedasticity | Breusch-Pagan | 24.122 | <.0001 | No |
| Normality | Shapiro-Wilke (1.g) | .6372 | <.0001 | No |
|  | QQPlot (1.g) |  |  | No |
| Linearity | Residual Plot (1.g) |  |  | No |
| Independence | Residual Plot (1.g) |  |  | Yes |

The F-statistic for model significance is 25.94 with a p-value < .0001. Therefore we conclude that the model is significant (appendix 1.h). The R² value of .4556 indicates ~46% of variability is explained by the model.

I attempted to remedy violations of regression assumptions and address the previously flagged observations. First off, applying BoxCox corrects the violation of normality. I could not find a suitable transformation for CV that helps pass homoscedasticity assumptions without sacrificing model significance. After the chosen BoxCox transformation the outliers and flagged observations are reduced to one point observation each, which happen to be the same observation. Removing the outlier has a negligible effect on the model (plots in appendix 1.i). After applying the BoxCox transformation, the normality and linearity assumptions pass. Violations of homoscedasticity are reduced, but still fail. The residual plots appear to be much more reasonably distributed.

**Regression Assumption checks for BoxCox(PSAL) = CV SV GS\_8 for λ = .25**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe CV | 5.14 | 0.0257 | No |
|  | Brown-Forsythe SV | 7.10 | .0091 | No |
|  | Brown-Forsythe GS\_8 | 7.76 | .0064 | No |
| Homoscedasticity | Breusch-Pagan | 15.40 | .001 | No |
| Normality | Shapiro-Wilke (1.j) | .9892 | .6224 | Yes |
|  | QQPlot (1.j) |  |  | Yes |
| Linearity | Residual Plot (1.j) |  |  | Yes |
| Independence | Residual Plot (1.j) |  |  | Yes |

The variance inflation factors remain insignificant after transformation, with values less than ten (1.k), so collinearity is not a problem. The F-statistic for model significance is 43.40 with a p-value < .0001 (1.l), so the model is considered significant. The R² value of .5833 indicates ~58% of variability is explained by the model. Notice that the R² value increased by 12% after applying BoxCox transformation.  
  
Analysis is also conducted on the 5 variable model PSAL = CV AGE BPH SV GS\_8 that was previously mentioned. By adding AGE and BPH, the MSE drops to its lowest, and there is a second bend on the other metrics. The only additional flagged observation compared to the untransformed three-variable model is 95, where the hat-diagonal diagnostic is significant (1.m).

**Regression Assumption checks for PSAL = CV AGE BPH SV GS\_8**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **Test** | **Statistic** | **P-Value** | **Accept** |
| Homoscedasticity | Brown-Forsythe CV | 11.61 | 0.0110 | No |
|  | Brown-Forsythe AGE | 2.52 | .1155 | Yes |
|  | Brown-Forsythe BPH | .42 | .5172 | Yes |
|  | Brown-Forsythe SV | 26 | <.0001 | No |
|  | Brown-Forsythe GS\_8 | 21.23 | <.0001 | No |
| Homoscedasticity | Breusch-Pagan | 24.7 | <.0001 | No |
| Normality | Shapiro-Wilke (1.n) | .6384 | <.0001 | No |
|  | QQPlot (1.n) |  |  | No |
| Linearity | Residual Plot (1.n) |  |  | No |
| Independence | Residual Plot (1.n) |  |  | Yes |

The newly introduced variables AGE and BPH pass the Brown-Forsythe tests for homoscedasticity. Variance inflation factors remain insignificant for all predictors (1.o). The F-statistic for model significance is 16.08 with a p-value < .0001 (1.p). The F-statistic dropped considerably from the previous models, but the new model is overall still significant. The R² value of .469 indicates ~47% of variability is explained by the model. This is a relatively small increase in R² value given that two new variables were added. The negligible increase in explained variance is related to the low significance, judging by the t-values, of the newly added AGE and BPH regression coefficients. Further transforming PSAL with BoxCox yields an R² value of .57, which is worse than the previous model (1.p).

Based on the best subset selection results, insignificance of AGE and BPH, and the prior analysis on the three-predictor model CV SV GS\_8, the best overall model is shown to be:

BoxCox(PSAL) = CV SV GS\_8

**Question 2**

The scatter plot between GPA and ACT suggests a weak correlation (2.a). For a strong correlation I would be expect a more noticeable line on the scatter plot. The computed correlation is .2695, which supports the visual analysis. The two variables also have a weak positive correlation on the plot. I collected 1000 samples from the bootstrap distribution of the point estimate. The histogram shows the computed correlations for each sample dataset (2.b). The data is clearly not skewed, judging by the attached visual (2.b).

The mean of the computed correlations is the bootstrap point estimate, with a value of .2636. The standard error of the point estimate is the deviance of the set of correlations from the bootstrap-sampled datasets, with a value of .10381. The difference between the constant correlation from above and the bootstrap estimate is the bias, with a value of -.0011. Conducting a T-Test with α=.05 of H₀: ρ’=.2695 provides a p-value of .73, which further justifies that the bias is negligible.

We may construct confidence intervals because there is no bias or skew. Normal, basic and percentile confidence intervals are provided below.

|  |  |  |
| --- | --- | --- |
|  | **Lower** | **Upper** |
| **Normal** | .0671 | .4741 |
| **Basic** | .0627 | .4684 |
| **Percentile** | .0705 | .4762 |

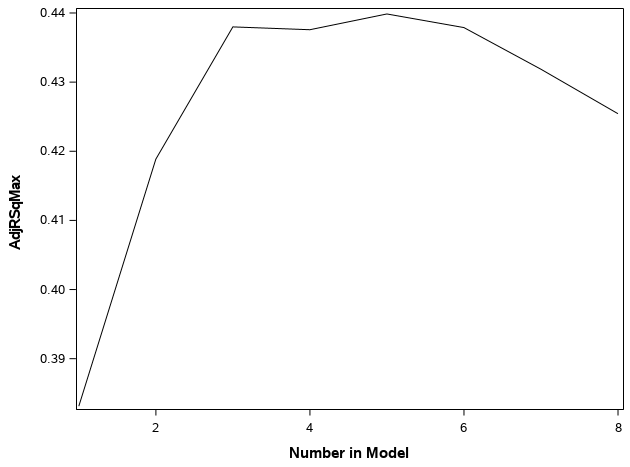
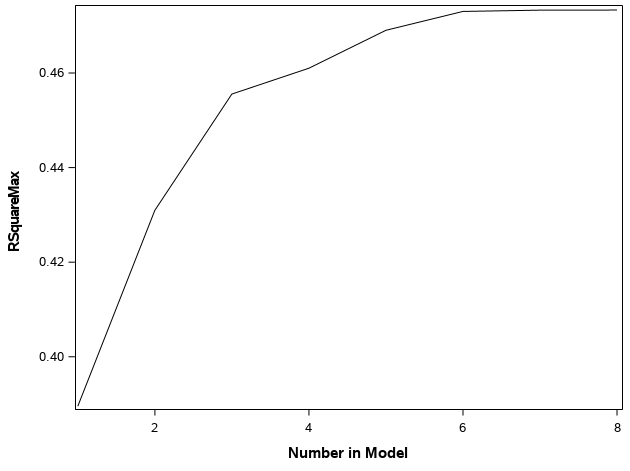
The normal and basic intervals are more conservative, while the percentile bootstrap is slightly tighter. Since there is no skew or bias, we can expect the estimates to be very similar, as they are.

Advanced Statistical Methods I Project III – SAS Output

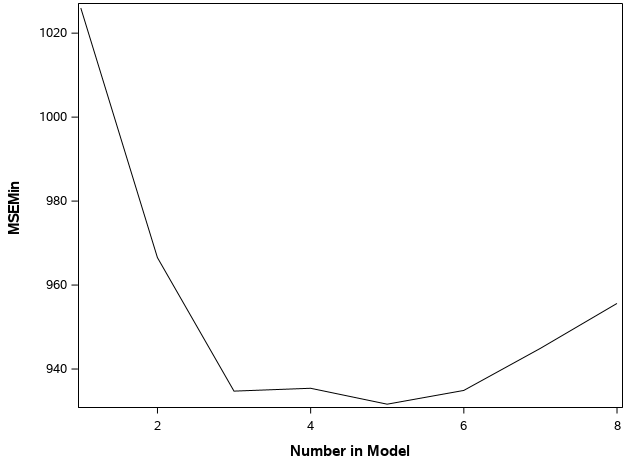
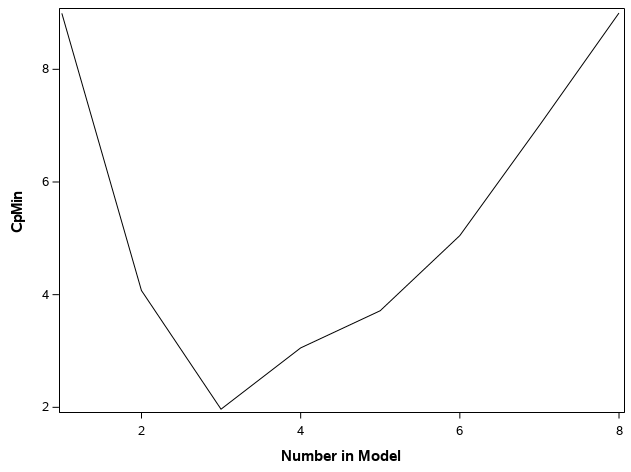
I omitted the PROC GLM-generated output for each of the Brown-Forsythe tests, due to the large number of times (11) I repeated it. The Breusch-Pagan tests do not have typical SAS output, they are just numbers printed via PROC IML.

1.a. Performance visualized by each metric:

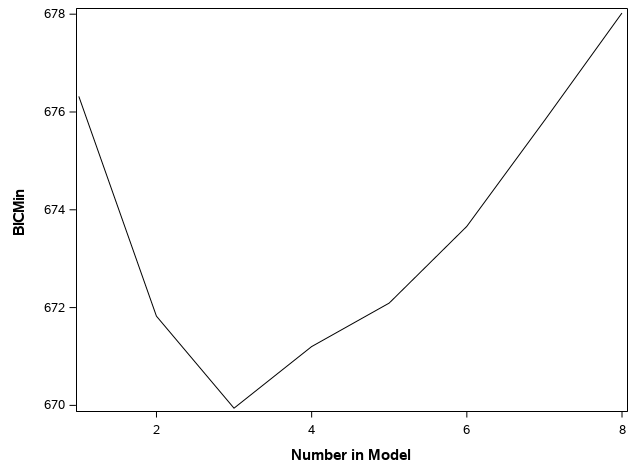
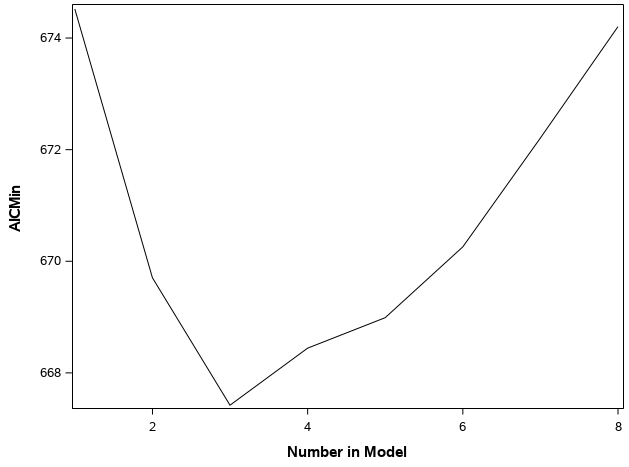
Maximum RSquare Maximum Adjusted RSquare



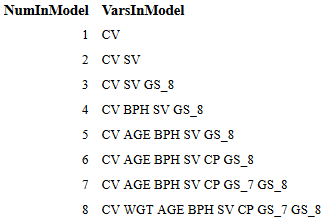
Minimum Mallow’s Cₚ Minimum MSE



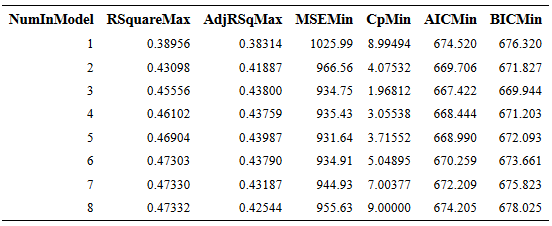
Minimum Akaike IC Minimum Bayesian IC



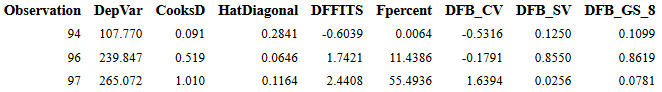
1.b The best models at each variable count:



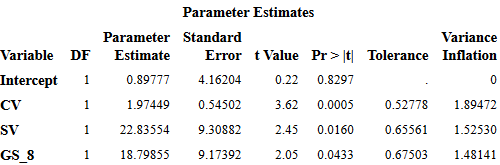
1.c The best performances for each metric, for each variable count:



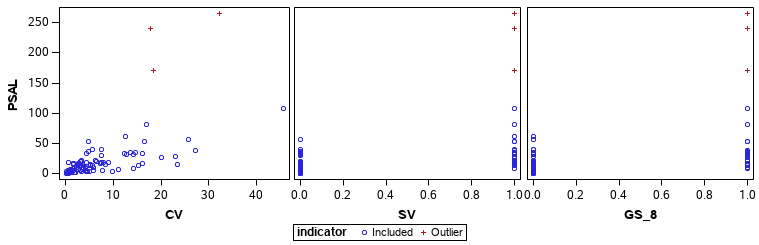
1.d Diagnostics for CV SV GS\_8:



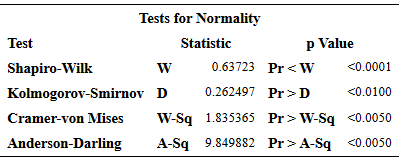
1.e Collinearity checks via VIF/TOL:

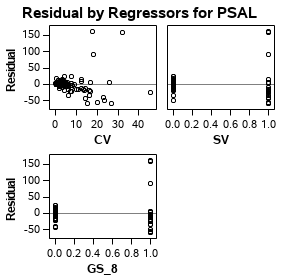
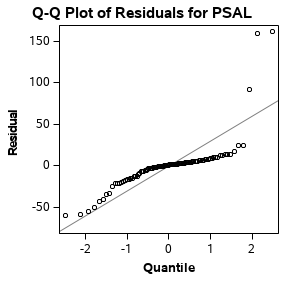


1.f Outliers for PSAL = CV SV GS\_8:

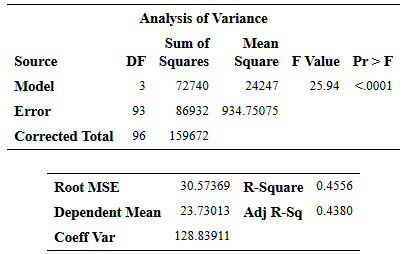


1.g Shapiro-Wilke, QQPlot and Residual plots for PSAL = CV SV GS\_8:

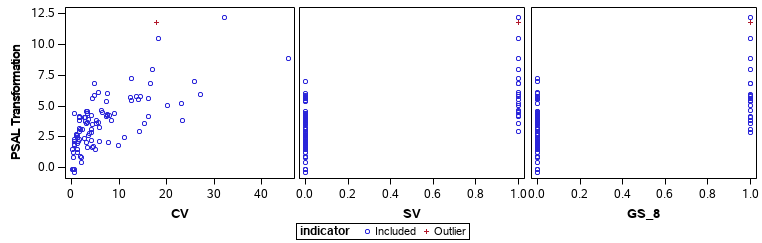




1.h Regression results for PSA = CV SV GS\_8:

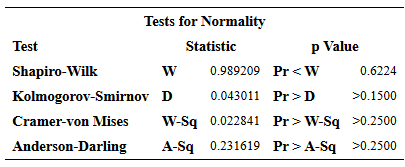


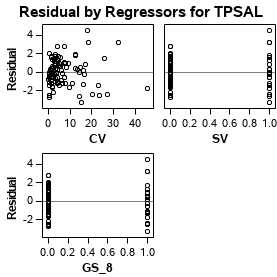
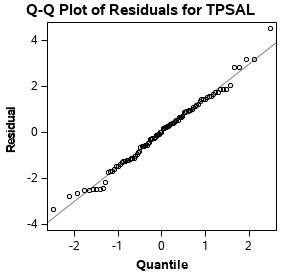
1.i Outliers and diagnostics after transformation. Observation 94 is an outlier:



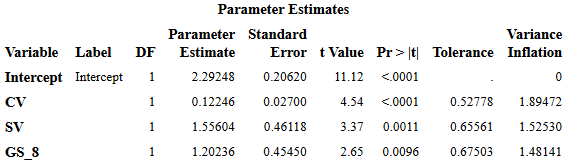


1.j Normality checks, QQPlot and Residual plots for BoxCox(PSAL) = CV SV GS\_8:

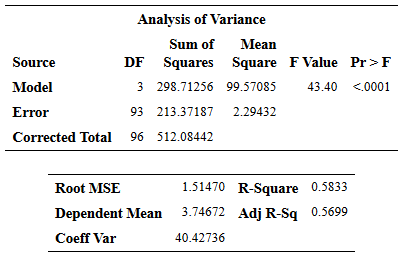




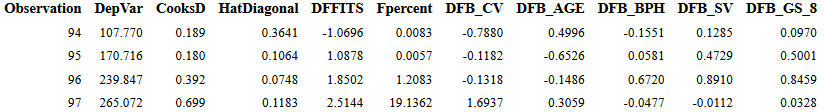
1.k Collinearity checks via VIF/TOL:



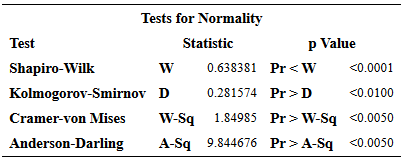
1.l Regression results for BoxCox(PSA) = CV SV GS\_8:

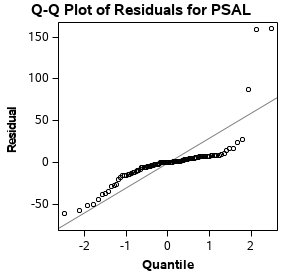


1.m Flagged observations for PSAL = CV AGE BPH SV GS\_8:

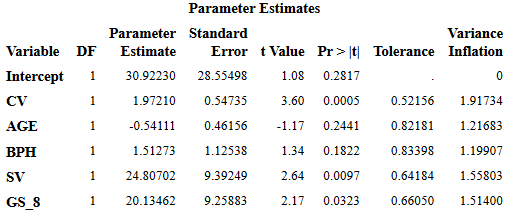


1.n Normality checks, QQPlot and Residual plots for PSAL = CV AGE BPH SV GS\_8:

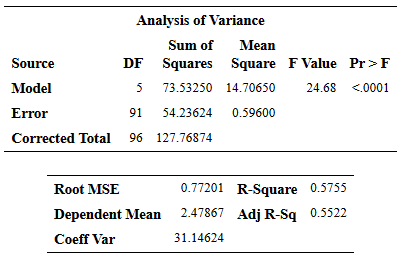
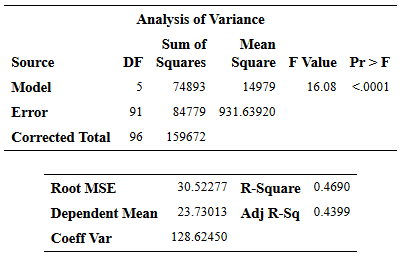




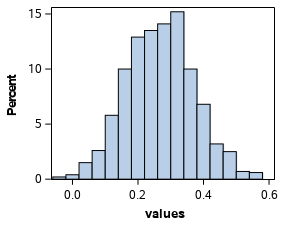
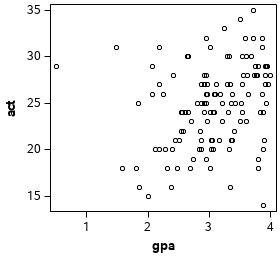
1.o Collinearity checks via VIF/TOL:



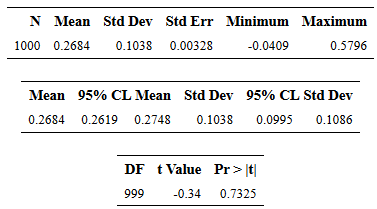
1.p Regression diagnostics (untransformed and transformed with CV AGE BPH SV GS\_8):



2.a & 2.b Scatter plot of GPA vs ACT, histogram of bootstrapped correlations:



2.c TTest for zero bias of bootstrap correlation estimate:



Michael Shoemate

Professor Choudhary

Advanced Statistical Methods I Project III – 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. Problem 1 is on page 16. The call ‘evaluate\_model’ performs collinearity, diagnostics, all the regression assumption checks, and the regression.

%LET seed = 160830;

DATA prostate;

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

INPUT ID PSAL CV WGT AGE BPH SV CP GS;

DATA gpa;

INFILE '/folders/myfolders/gpa.csv' DLM=',' FIRSTOBS=2;

INPUT gpa act;

/\* prints the best set of variables for each predictor count \*/

%MACRO find\_best\_variables(metric, sign);

PROC SQL;

CREATE TABLE models\_&metric AS

SELECT d1.NumInModel, d1.VarsInModel FROM SubsetSummary AS d1

LEFT OUTER JOIN SubsetSummary AS d2

ON (d1.NumInModel = d2.NumInModel and d1.&metric &sign d2.&metric)

WHERE d2.NumInModel IS NULL GROUP BY d1.NumInModel;

TITLE "Models by &metric";

PROC PRINT DATA=models\_&metric;

RUN;

%MEND find\_best\_variables;

%MACRO model\_search(datatable, predictors, predicted);

ODS EXCLUDE ALL;

PROC RSQUARE DATA=&datatable;

MODEL &predicted=&predictors / CP ADJRSQ SSE MSE AIC BIC;

ODS OUTPUT SubsetSelSummary=SubsetSummary;

RUN;

ODS EXCLUDE NONE;

PROC SQL;

CREATE TABLE SubsetSummaryStats AS

SELECT NumInModel, MAX(RSquare) AS RSquareMax, MAX(AdjRSq) AS AdjRSqMax,

MIN(MSE) AS MSEMin, MIN(Cp) AS CpMin,

MIN(AIC) AS AICMin, MIN(BIC) AS BICMin

FROM SubsetSummary GROUP BY NumInModel;

PROC PRINT DATA=SubsetSummaryStats NOOBS;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=RSquareMax;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=AdjRSqMax;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=CpMin;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=MSEMin;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=AICMin;

RUN;

PROC SGPLOT DATA=SubsetSummaryStats;

SERIES X=NumInModel Y=BICMin;

RUN;

%find\_best\_variables(Cp, GT);

/\* %find\_best\_variables(RSquare, LT); \*/

/\* %find\_best\_variables(AdjRsq, LT); \*/

/\* %find\_best\_variables(MSE, GT); \*/

/\* %find\_best\_variables(AIC, LT); \*/

/\* %find\_best\_variables(BIC, LT); \*/

%MEND model\_search;

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

PROC SQL;

DROP TABLE results;

ODS EXCLUDE ALL;

PROC REG DATA=&datatable;

MODEL &predicted=&predictors / INFLUENCE R;

ODS OUTPUT outputstatistics=results;

RUN;

ODS EXCLUDE NONE;

/\* Just to count the number of rows \*/

PROC IML;

USE &datatable;

READ ALL VAR{&predicted};

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

%LET p = %SYSEVALF(%LENGTH(&predictors) + 1);

DATA results; set results;

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

dfflag = ABS(DFFITS) > 1;

/\* calculate percentile for each Cook's D value using F(p, n-p) dist\*/

Fpercent = 100\*probf(CooksD, &p, &n - &p);

/\* 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;

%LET dfbetas\_flags = ;

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

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

%LET dfbetas\_flags = &dfbetas\_flags DFB\_&var;

%END;

TITLE2 '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 &dfbetas\_flags;

RUN;

TITLE2;

%MEND test\_influential;

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

PROC REG DATA=&datatable;

ODS SELECT WHERE=(\_name\_ = 'CollinDiag' | \_name\_ = 'ParameterEstimates');

MODEL &predicted = &predictors / VIF TOL COLLIN;

/\* multiple correlation coefficients - not used \*/

/\* %DO i=1 %TO %SYSFUNC(COUNTW(&predictors, ' ')); \*/

/\* %LET label = %SCAN(&predictors, &i, ' '); \*/

/\* \*/

/\* %LET otherParams = %SYSFUNC(TRANWRD(&predictors, &label, %STR( ))); \*/

/\* PROC REG DATA=&datatable; \*/

/\* ODS SELECT WHERE=(\_name\_ = 'CollinDiag' | \_name\_ = 'ParameterEstimates'); \*/

/\* MODEL &label = &otherParams / VIF TOL COLLIN; \*/

/\* RUN; \*/

/\* %END; \*/

%MEND test\_collinearity;

%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=8in HEIGHT=3in;

PROC SGSCATTER DATA=&datatable;

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

RUN;

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

TITLE2;

%MEND outliers;

%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);

TITLE "Fit for &predicted vs &predictors";

ODS GRAPHICS ON;

PROC REG DATA=&datatable;

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

MODEL &predicted=&predictors / CLB;

RUN;

%MEND fit;

%MACRO evaluate\_model(dataset, predictors, predicted);

TITLE "Outliers: &predicted = &predictors";

%outliers(&dataset, &predictors, &predicted, &tvalue);

TITLE "Diagnostics: &predicted = &predictors";

%test\_influential(&dataset, &predictors, &predicted);

TITLE "Collinearity: &predicted = &predictors";

%test\_collinearity(&dataset, &predictors, &predicted);

TITLE "Homoscedasticity: &predicted = &predictors";

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

TITLE "Normality: &predicted = &predictors";

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

TITLE "Linearity/independence: &predicted = &predictors";

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

TITLE "Evaluate Model: &predicted = &predictors";

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

%MEND evaluate\_model;

%MACRO problem\_1();

/\* Check performance when GS is treated as an ordinal variable \*/

/\* %LET predictors=CV WGT AGE BPH SV CP GS; \*/

/\* %model\_search(prostate, &predictors, PSAL); \*/

PROC SQL;

ALTER TABLE prostate DROP ID;

ALTER TABLE prostate ADD GS\_7 NUMERIC, GS\_8 NUMERIC;

UPDATE prostate SET GS\_7= CASE WHEN GS=7 THEN 1 ELSE 0 END;

UPDATE prostate SET GS\_8= CASE WHEN GS=8 THEN 1 ELSE 0 END;

ALTER TABLE prostate DROP GS;

%LET predictors=CV WGT AGE BPH SV CP GS\_7 GS\_8;

/\* Check performance when GS is treated as an categorical variable \*/

%model\_search(prostate, &predictors, PSAL);

/\* Determine the T-value to be used throughout the problem first \*/

%LET alpha = .05;

PROC IML;

USE prostate;

READ ALL VARS{PSAL};

CALL SYMPUTX('tvalue', TINV(1 - &alpha/(2 \* 4), NROW(PSAL)));

PRINT 'T-Value for absolute RStudent cutoff';

PRINT &tvalue;

/\* MODEL ONE \*/

%evaluate\_model(prostate, CV SV GS\_8, PSAL);

/\* transforms on CV do not improve performance \*/

/\* PROC SQL; \*/

/\* ALTER TABLE prostate ADD CV\_log NUMERIC; \*/

/\* UPDATE prostate SET CV\_log=log(CV); \*/

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

PROC TRANSREG DATA=prostate;

MODEL BoxCox(PSAL)=identity(CV SV GS\_8);

OUTPUT OUT=transdata;

RUN;

%evaluate\_model(transdata, CV SV GS\_8, TPSAL);

/\* MODEL TWO \*/

%evaluate\_model(prostate, CV AGE BPH SV GS\_8, PSAL);

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

PROC TRANSREG DATA=prostate;

MODEL BoxCox(PSAL)=identity(CV AGE BPH SV GS\_8);

OUTPUT OUT=transdata;

RUN;

%evaluate\_model(transdata, CV AGE BPH SV GS\_8, TPSAL);

%MEND problem\_1;

%MACRO problem\_2();

TITLE "scatter plot of GPA vs ACT";

PROC SGPLOT DATA=gpa;

SCATTER Y=act X=gpa;

PROC CORR DATA=gpa OUTP=correlation\_constant NOPRINT;

VAR gpa act;

RUN;

PROC IML;

USE correlation\_constant;

READ ALL VAR {gpa} WHERE(\_name\_='act');

CALL SYMPUTX('correlation\_constant', gpa);

%LET numSamples = 1000;

PROC SURVEYSELECT DATA=gpa SEED=&seed NOPRINT

OUT=bootstrapSample(rename=(Replicate=SampleID))

METHOD=urs /\* resampling with replacement \*/

SAMPRATE=1 /\* each bootstrap resample has N observations \*/

REPS=&numSamples;

PROC CORR DATA=bootstrapSample OUTP=correlations\_with\_fluff NOPRINT;

VAR gpa act;

BY SampleId;

FREQ NumberHits;

PROC SQL;

CREATE TABLE correlations AS SELECT gpa AS values FROM correlations\_with\_fluff WHERE \_name\_ EQ 'act';

TITLE "histogram of (nonparametric) bootstrap distribution of the point estimate";

PROC SGPLOT DATA=correlations; /\* histogram of bootstrap distribution \*/

HISTOGRAM values;

RUN;

PROC MEANS DATA=correlations NOPRINT;

VAR values;

OUTPUT OUT=correlation\_stats MEAN=mean STD=std;

RUN;

TITLE "constant POINT ESTIMATE of ρ";

PROC IML;

PRINT(&correlation\_constant);

RUN;

TITLE "bootstrap POINT ESTIMATE and STANDARD ERROR of ρ";

PROC PRINT DATA=correlation\_stats NOOBS;

VAR mean std;

RUN;

TITLE "Bootstrap bias";

PROC IML;

USE correlation\_stats;

READ ALL VAR {mean};

bias = mean - &correlation\_constant;

PRINT bias;

CALL SYMPUTX('bias', bias);

PROC TTEST DATA=correlations H0=&correlation\_constant PLOTS(SHOWH0) SIDES=2 ALPHA=0.05;

VAR values;

RUN;

TITLE;

DATA biases;

SET correlations;

bias = values - &correlation\_constant;

PROC UNIVARIATE DATA=correlations NOPRINT;

VAR values;

OUTPUT OUT=correlation\_confidences PCTLPRE=CI95\_

PCTLPTS=2.5 97.5 /\* compute 2.5th & 97.5th percentiles of sampling distribution of median \*/

PCTLNAME=Lower Upper;

RUN;

/\* normal approximation \*/

DATA normal;

SET correlation\_stats;

z1 = QUANTILE('NORMAL', .975);

z2 = QUANTILE('NORMAL', .025);

lower\_normal = (&correlation\_constant - &bias) - z1 \* std;

upper\_normal = (&correlation\_constant - &bias) - z2 \* std;

RUN;

TITLE "Normal 95% confidence interval for correlation";

PROC PRINT DATA=normal NOOBS;

VAR lower\_normal upper\_normal;

/\* basic bootstrap \*/

DATA basic;

SET correlation\_confidences;

lower\_basic = 2 \* &correlation\_constant - CI95\_Upper;

upper\_basic = 2 \* &correlation\_constant - CI95\_Lower;

RUN;

TITLE "Basic 95% confidence interval for correlation";

PROC PRINT DATA=basic NOOBS;

VAR lower\_basic upper\_basic;

/\* percentile bootstrap \*/

DATA percentile;

SET correlation\_confidences;

lower\_percentile = CI95\_Lower;

upper\_percentile = CI95\_Upper;

RUN;

TITLE "Percentile 95% confidence interval for bias";

PROC PRINT DATA=percentile NOOBS;

VAR lower\_percentile upper\_percentile;

RUN;

TITLE;

%MEND problem\_2;

%problem\_1();

%problem\_2();