Niranjan Poudel
STAT 5100
Spring 2019

Homework 4

Exercise 1.

a) The sequence plots for all three predictors and the response, and the order in data and response are provided below.

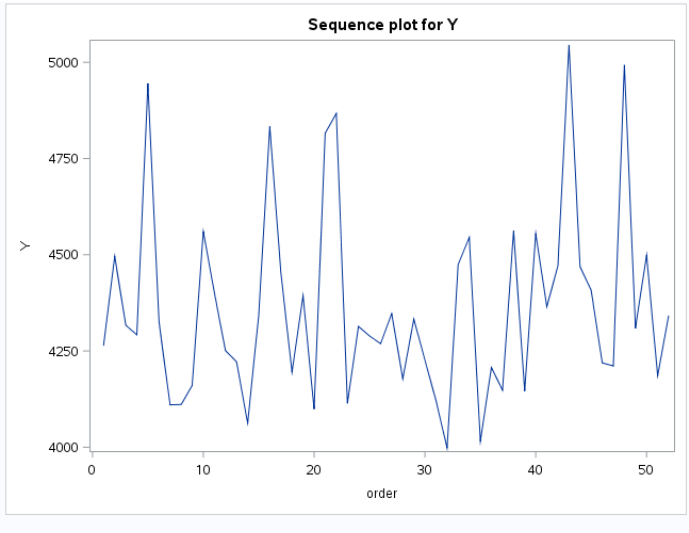


Fig 1: sequence plot of order vs Y

- The sequence plot has no definite order, so it's hard to explain the pattern of Y.

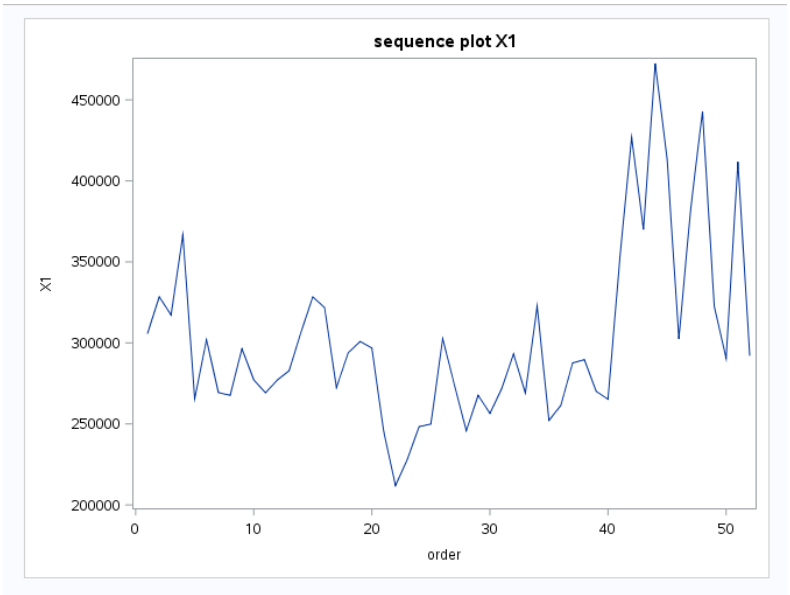
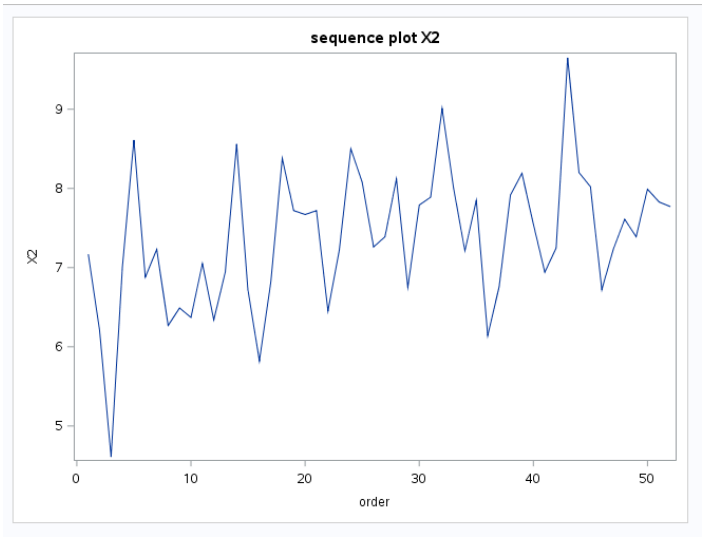


Fig 2: Sequence plot of X1 vs Y

- Above sequence plot does not show any definite order and distribution.

Niranjan Poudel
STAT 5100
Spring 2019

Fig 3: Sequence plot of X2 vs Y



- Above sequence plot shows increasing trend in average of the X2 variable, still there is no much information.

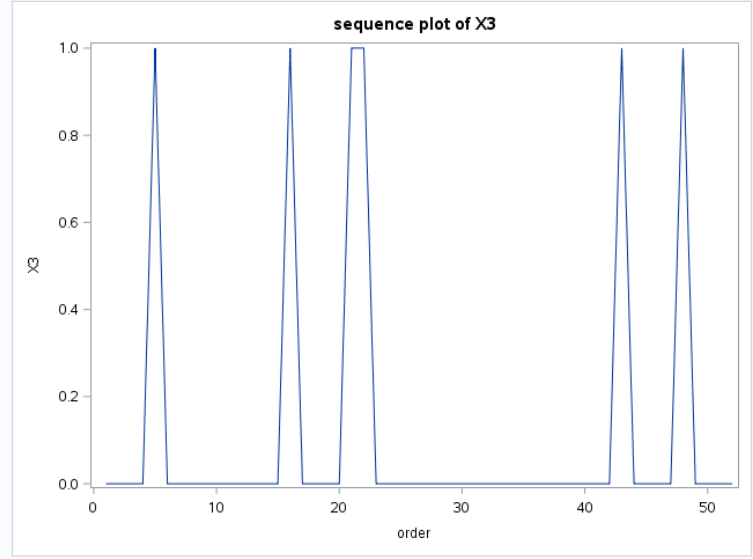


Fig 4: Sequence plot of X3 vs Y

- Above sequence plot does not show any definite order and distribution but shows the data fluctuating from 0 to 1 for X3.

All of the sequence plots above does not have definite order so it is hard to explain any trends.

Niranjan Poudel
STAT 5100
Spring 2019

b) The scatterplot matrix and correlation matrix are provided below.

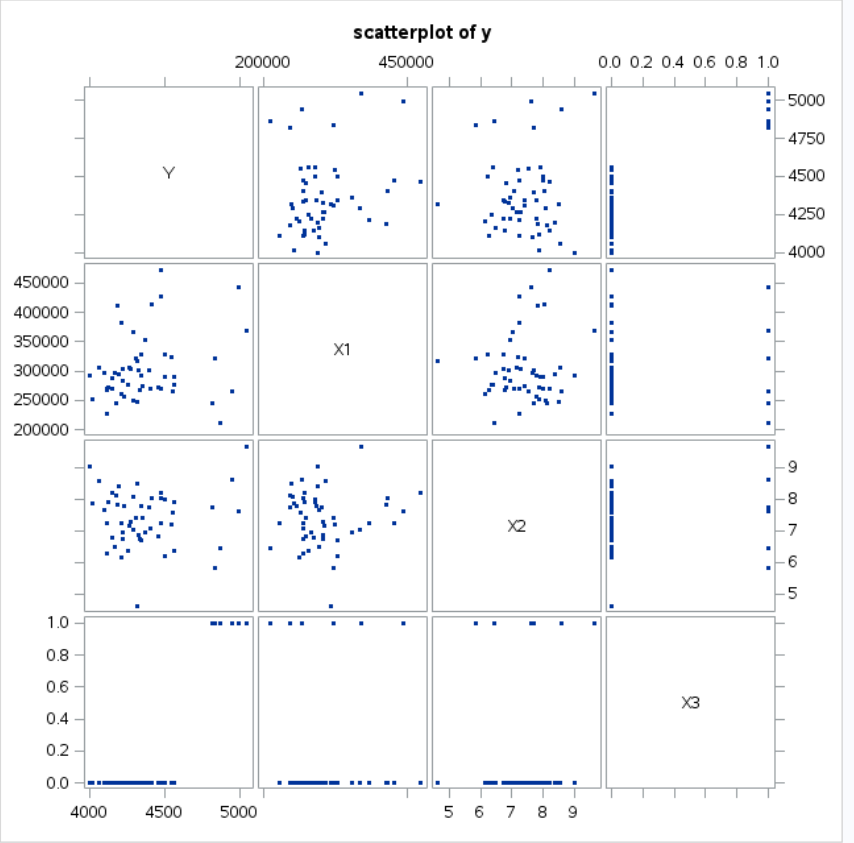


Fig 5: scatterplot matrix

- The scatter plot of X1 vs Y and X2 vs Y shows scattered plots where it seems a linear relationship does not fit well. Also, the scatterplot of X3 vs Y shows that there are only two values of X3, 0 and 1. Value of Y seems to be correlated more with X3.

Pearson Correlation Coefficients, N = 52 Prob > |r| under H0: Rho=0				
	Y	**X1**	**X2**	**X3**
Y	1.00000	0.20766 0.1396	0.06003 0.6725	0.81058 <.0001
X1	0.20766 0.1396	1.00000	0.08490 0.5496	0.04566 0.7479
X2	0.06003 0.6725	0.08490 0.5496	1.00000	0.11337 0.4236
X3	0.81058 <.0001	0.04566 0.7479	0.11337 0.4236	1.00000

Fig 6: correlation matrix

- From above correlation matrix, the p-value of correlation of Y with X1 is 0.1396, X2 is 0.6725 and with X3 is <0.0001. So the correlation with X3 is significant than others while the correlation is 0.81058 between X3 and Y. p-value of correlation between X1

and X2 is 0.5496 between X2 and X3 is 0.4236 and between X1 and X3 is 0.7479 which are greater than 0.05 which suggest no significant correlation between them.

Exercise 2.

a) The regression model output is as follow;

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4149.88721	195.56541	21.22	<.0001
X1	1	0.00078708	0.00036455	2.16	0.0359
X2	1	-13.16602	23.09173	-0.57	0.5712
X3	1	623.55448	62.64095	9.95	<.0001

- From above SAS output we can estimate the regression function as;

$$Y = 4149.887 + 0.00078708\ X1 - 13.167\ X2 + 623.554\ X3$$

Also, we get from above, β_0=4149.887 and p-value < 0.0001, which implies $\beta_0 \neq 0$

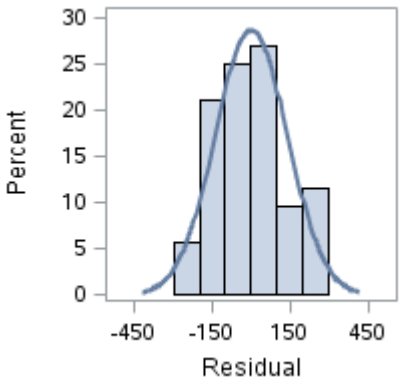
B_1 = 0.000787 and p-value =0.0359 < 0.05 which implies $\beta_1 \neq 0$

B_2 = -13.16602 and p-value = 0.5712 > 0.05 we do not have enough evidence to not support β_2 =0.

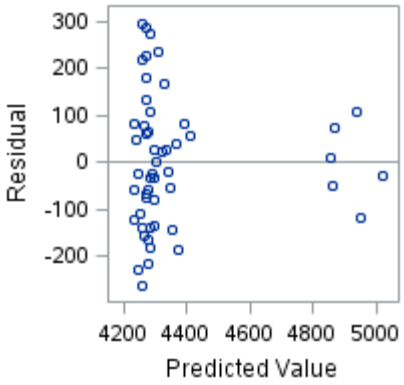
β_3 = 623.55448 and p-value < 0.0001 which suggest $\beta_3 \neq 0$.

From above model we can suggest there is enough evidence to support that X1 and X3 affects the Y value whereas there is not enough evidence to say X3 affect the Y value.

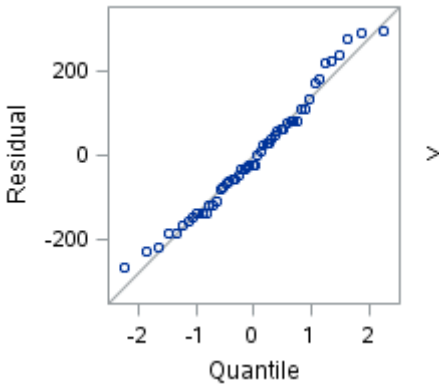
b) The graphical and numerical checks are as follow;



Histogram plot

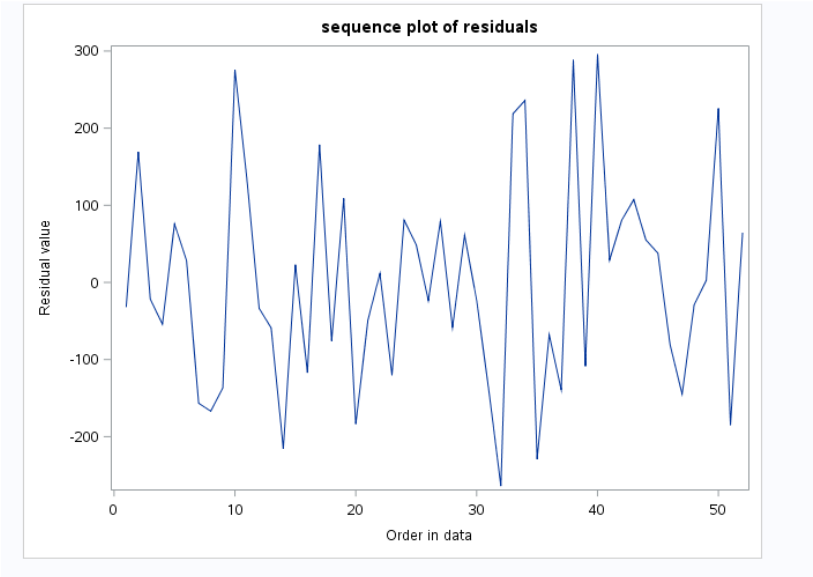


residual vs fitted plot



normal probability

Niranjan Poudel
STAT 5100
Spring 2019



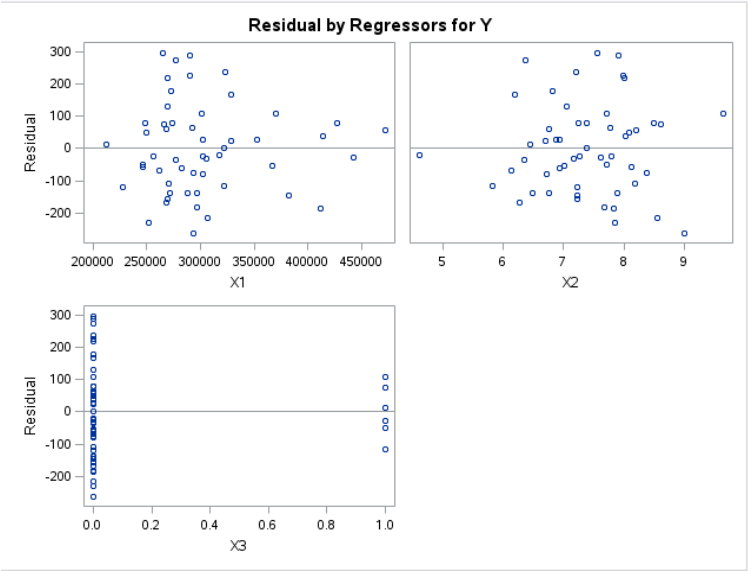


Fig: Residual plots for each predictor.

- The graphical plots suggest that the residuals are not totally normally distributed and have constant variance. The data does not have the fix order to be defined from sequence plot.

Niranjan Poudel
STAT 5100
Spring 2019

Brown-Forsythe test of constant variance is as follow.

P-value for Brown-Forsythe test of constant variance in Residual vs. Predicted Value

Obs	t_BF	BF_pvalue
1	2.66204	0.010418

- Bf_pvalue = 0.010418 < 0.05 which does not show enough evidence for constant variance.

Correlation test of normality is as follow;

Pearson Correlation Coefficients, N = 52
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid Residual	1.00000	0.99087 <.0001
expectNorm	0.99087 <.0001	1.00000

- From table B.6 the critical value of correlation coefficient is between 0.977 and 0.980, from above table correlation coefficient is 0.99087 > 0.980 and p-value < 0.0001 which suggest that the error terms are normal.

Exercise 3.

Bonferroni and Scheffe intervals for the predictions of shipments with given characters are as follow;

Simultaneous 95% interval of individual predivtion
At four X-profilels, usding scheffe and Bonferroni

Obs	X1	X2	X3	Yhat	S_lower	S_upper	B_lower	B_upper
53	230000	7.5	0	4232.17	3760.37	4703.97	3849.91	4614.43
54	250000	7.3	0	4250.55	3782.68	4718.41	3871.48	4629.61
55	280000	7.1	0	4276.79	3811.89	4741.69	3900.12	4653.46
56	340000	6.9	0	4326.65	3859.19	4794.10	3947.91	4705.38

- From above table we can see that Bonferroni predictions are tighter than Scheffe predictions so, Bonferroni is the efficient prediction interval.

Exercise 4.

a. From the subset F-test model we have following estimates

Niranjan Poudel
STAT 5100
Spring 2019

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	4149.88721	195.56541	21.22	<.0001	989877440	.
X1	1	0.00078708	0.00036455	2.16	0.0359	136366	0.04312
X3	1	623.55448	62.64095	9.95	<.0001	2033565	0.67208
X2	1	-13.16602	23.09173	-0.57	0.5712	6674.58809	0.00673

$SSR(X_1)$ = 136366.
SSR(X3|X1) = 2033565
SSR (X2|X 1, X3) = 6674.58809

b. Sub set F-test of H_0: $\beta_2 = 0$ is as follow;

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	4149.88721	195.56541	21.22	<.0001	989877440	.
X1	1	0.00078708	0.00036455	2.16	0.0359	136366	0.04312
X3	1	623.55448	62.64095	9.95	<.0001	2033565	0.67208
X2	1	-13.16602	23.09173	-0.57	0.5712	6674.58809	0.00673

subsest f-test , automatically

The REG Procedure
Model: MODEL1

Test subsetcheck Results for Dependent Variable Y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	6674.58809	0.33	0.5712
Denominator	48	20532		

- From above tables P-value = 0.5712 > 0.05 so we do not have enough evidence to reject the null hypothesis and so it suggests that: $\beta_2 = 0$. Furthermore, we have SSE (X1, X2, X3) = 985530, SSE (X2|X1, X3) m =6,674.
- F* = (6674/1) / (985530/48) = 0.32491 and also from above table F value =0.33.

c. Again after fitting two models with X1 and X2 as multiple predictors we observe the following results;

First we obtain the model for X1 followed by X2 as following.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	3995.47867	337.76602	11.83	<.0001	989877440	.
X1	1	0.00091916	0.00063120	1.46	0.1517	136366	0.04312
X2	1	12.12052	39.76555	0.30	0.7618	5725.92181	0.00189

- From above table we have the following values

SSR(X1) = 136366 And SSR (X2 | X1) = 5725.92181

Again, we have the following results as we model X2 preceding X1;

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	3995.47867	337.76602	11.83	<.0001	989877440	.
X2	1	12.12052	39.76555	0.30	0.7618	11395	0.00360
X1	1	0.00091916	0.00063120	1.46	0.1517	130697	0.04148

SSR(X2) = 11395 and SSR(X1|X2) = 130697.

Now, SSR(X1) + SSR(X2|X1) = 136366 + 5725.922 = 142091.922

And, SSR(X2) + SSR(X1|X2) = 11395 + 130697 = 142092.

- Therefore, SSR(X1) + SSR (X2|X1 = SSR(X2) + SSR (X1|X2 = 142092.

Yes, these two must be always equal because both of the value represents the amount of variation in Y when both X1 and X2 are in the model, which should be same from both the process.

Niranjan Poudel
STAT 5100
Spring 2019

Exercise 5.

The test for multicollinearity is done following are the result for the test;

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	X1	X2	X3
1	3.13677	1.00000	0.00102	0.00308	0.00133	0.02004
2	0.83411	1.93923	0.00032068	0.00091374	0.00035401	0.97010
3	0.02283	11.72211	0.03448	0.88968	0.16309	0.00033831
4	0.00629	22.32696	0.96418	0.10632	0.83523	0.00953

From the above table we can see that the condition index for number 3 and 4 are 11.72 and 22.32696 and we can see predictors variability more than 50% for only one predictor variable, in both the cases so this does not show multicollinearity.

Again test for multicollinearity using Variance inflation factor we have the following results.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4149.88721	195.56541	21.22	<.0001	0
X1	1	0.00078708	0.00036455	2.16	0.0359	1.00860
X2	1	-13.16602	23.09173	-0.57	0.5712	1.01960
X3	1	623.55448	62.64095	9.95	<.0001	1.01436

Here we have V IF1 = 1.0086, V IF2 = 1.0196, V IF3 = 1.0144. There is no VIF greater than 10 and the average looks to be around 1. So, we can say that there is no multicollinearity.

Niranjan Poudel
STAT 5100
Spring 2019

Appendix

```
/* Exercise 1 */

/* Input data */
data grocery;
   input Y X1 X2 X3 @@; cards;
   4264   305657   7.17   0      4496   328476   6.20   0
   4317   317164   4.61   0      4292   366745   7.02   0
   4945   265518   8.61   1      4325   301995   6.88   0
   4110   269334   7.23   0      4111   267631   6.27   0
   4161   296350   6.49   0      4560   277223   6.37   0
   4401   269189   7.05   0      4251   277133   6.34   0
   4222   282892   6.94   0      4063   306639   8.56   0
   4343   328405   6.71   0      4833   321773   5.82   1
   4453   272319   6.82   0      4195   293880   8.38   0
   4394   300867   7.72   0      4099   296872   7.67   0
   4816   245674   7.72   1      4867   211944   6.45   1
   4114   227996   7.22   0      4314   248328   8.50   0
   4289   249894   8.08   0      4269   302660   7.26   0
   4347   273848   7.39   0      4178   245743   8.12   0
   4333   267673   6.75   0      4226   256506   7.79   0
   4121   271854   7.89   0      3998   293225   9.01   0
   4475   269121   8.01   0      4545   322812   7.21   0
   4016   252225   7.85   0      4207   261365   6.14   0
   4148   287645   6.76   0      4562   289666   7.92   0
   4146   270051   8.19   0      4555   265239   7.55   0
   4365   352466   6.94   0      4471   426908   7.25   0
   5045   369989   9.65   1      4469   472476   8.20   0
   4408   414102   8.02   0      4219   302507   6.72   0
   4211   382686   7.23   0      4993   442782   7.61   1
   4309   322303   7.39   0      4499   290455   7.99   0
   4186   411750   7.83   0      4342   292087   7.77   0
;
run;

/* Look at sequence plots */
data grocery; set grocery;
  order = _n_;
proc sgplot data=grocery;
   series x=order y=Y / lineattrs=(pattern=solid) ;
```

```
  title1 'Sequence plot for Y';
run;
/* repeat sgplot for y = (other variables) */
data grocery; set grocery;
       order = _n_;
proc sgplot data=grocery;
   series x=order y=X1 / lineattrs=(pattern=solid) ;
  title1 'sequence plot X1';
run;

data grocery; set grocery;
       order = _n_;
proc sgplot data=grocery;
   series x=order y=X2/ lineattrs=(pattern=solid) ;
  title1 'sequence plot X2';
run;

data grocery; set grocery;
       order = _n_;
proc sgplot data=grocery;
   series x=order y=X3 / lineattrs=(pattern=solid) ;
  title1 'sequence plot of X3';
run;

/* Scatterplot matrix and correlation matrix*/
proc sgscatter data= grocery;
       matrix Y X1 X2 X3/ markerattrs=(symbol=CIRCLEFILLED size= 3pt);
title1 'scatterplot of y';
run;

proc corr data=grocery;
       var Y X1 X2 X3;
title1 'Correlation matrix';
run;

/* regression model*/
proc reg data= grocery;
       model Y= X1 X2 X3;
       output out = out1 residual=resid predicted=pred;
       title1 'regression of Y';
       title2  'full model';
run;

data temp; set out1;
       order = _n_;
```

Niranjan Poudel
STAT 5100
Spring 2019

```
proc sgplot data=temp;
       series x=order y=resid / lineattrs=(pattern=solid);
       xaxis label = 'Order in data';
       yaxis label= 'Residual value';
       title1 'sequence plot of residuals';
run;


%macro resid_num_diag(dataset,datavar,label='requested
variable',predvar=' ',predlabel='predicted variable'); title; data
shortfourplotdataset; set &dataset;  label &datavar = &label;  if
&datavar ne .; run; proc means data=shortfourplotdataset noprint;  var
&datavar; output out=shortfourplotoutset N=nval mean=meanval; data
shortfourplotoutset; set shortfourplotoutset;  xn=nval; CALL
SYMPUT('nval',xn);   xmean=meanval; CALL SYMPUT('meanval',xmean);
%global nvalue; %let nvalue=&nval; %global meanvalue; %let
meanvalue=&meanval; run; %if &predvar ne ' '  %then %do;     data
shortfourplotdataset; set shortfourplotdataset;       label &predvar =
&predlabel;     proc sort data=shortfourplotdataset
out=shortfourplottemp;        by descending &predvar;     data
shortfourplottemp; set shortfourplottemp;       shortfourplotorder =
_n_;       shortfourplotgroup = 1-(shortfourplotorder <
ceil(&nvalue/2));     proc means data=shortfourplottemp median
noprint;       by shortfourplotgroup;       var &datavar;       output
out=shortfourplotouttemp median=medresid;     run;    data
shortfourplottempnew; merge shortfourplottemp shortfourplotouttemp; by
shortfourplotgroup;       d = abs(&datavar-medresid);     run;
       run;     proc ttest data=shortfourplottempnew plots=none;
class shortfourplotgroup;       var d;       ods output
TTests=shortfourplotBFtemp; title1 '(Ignore this nuisance output)';
       run;   run;     data shortfourplotBFtemp2; set
shortfourplotBFtemp;       if method = 'Pooled';       t_BF =
abs(tValue);       BF_pvalue = probt;       keep t_BF BF_pvalue;
proc print data=shortfourplotBFTemp2;       title1 'P-value for Brown-
Forsythe test of constant variance';       title2 'in ' &label ' vs. '
&predlabel;     run; %end; proc sort data=shortfourplotdataset
out=shortfourplottemp;     by &datavar; data shortfourplottemp; set
shortfourplottemp;   n=&nvalue;   expectNorm = probit((_n_-
.375)/(n+.25)); proc corr data=shortfourplottemp;   var &datavar
expectNorm;   title1 'Output for correlation test of normality of '
&label;   title2 '(Check text Table B.6 for threshold)'; run; title;
quit; %mend resid_num_diag;


%resid_num_diag(dataset=out1, datavar=resid,
```

```
 label='Residual', predvar=pred,
 predlabel='Predicted Value');
run;


/* Exercise 3 */

data dummy; input X1 X2 X3 check; cards;
  230000 7.5 0  1
  250000 7.3 0  1
  280000 7.1 0  1
  340000 6.9 0  1
;
data temp; set grocery dummy;
proc reg data= temp noprint;
       model Y= X1 X2 X3;
       output out=out2 predicted=Yhat stdi=seYhatnew;
data out2; set out2;
       alpha= 0.05;
       p=4;
       n= 52;
       g=4;
       S = sqrt(g*finv(1-alpha,g,n-p));
       t = tinv(1-alpha/(2*g),n-p);
       S_upper = Yhat + S*seYhatnew;
       S_lower = Yhat - S*seYhatnew;
       B_upper = Yhat + t*seYhatnew;
       B_lower = Yhat - t*seYhatnew;
 proc print data= out2;
       where check= 1;
       var  X1 X2 X3 Yhat S_lower S_upper B_lower B_upper;
       title1 'Simultaneous 95% interval of individual predivtion';
       title2 'At four X-profilels, usding scheffe and Bonferroni';
run;

/*question 4*/

proc reg data = grocery;
       model Y = X1 X3 X2 / ss1 pcorr1;
       subsetcheck: test X3= X2= 0;
       title1 'subsest f-test , automatically';
run;

proc reg data = grocery;
       model Y = X1 X3 X2 / ss1 pcorr1;
```

```
       subsetcheck: test X2= 0;
       title1 'subsest f-test , automatically';
run;

proc reg data = grocery;
       model Y= X1 X2 / ss1 pcorr1;
       title1 'regression using X1 and X2';
run;

proc reg data =grocery;
       model Y = X2 X1 / ss1 pcorr1;
       title1 'regression using X2 and X1';
run;

proc reg data= grocery;
       model Y = X1 X2 X3 / vif collin;
       title1 'test for multicollinearity';
run;
```