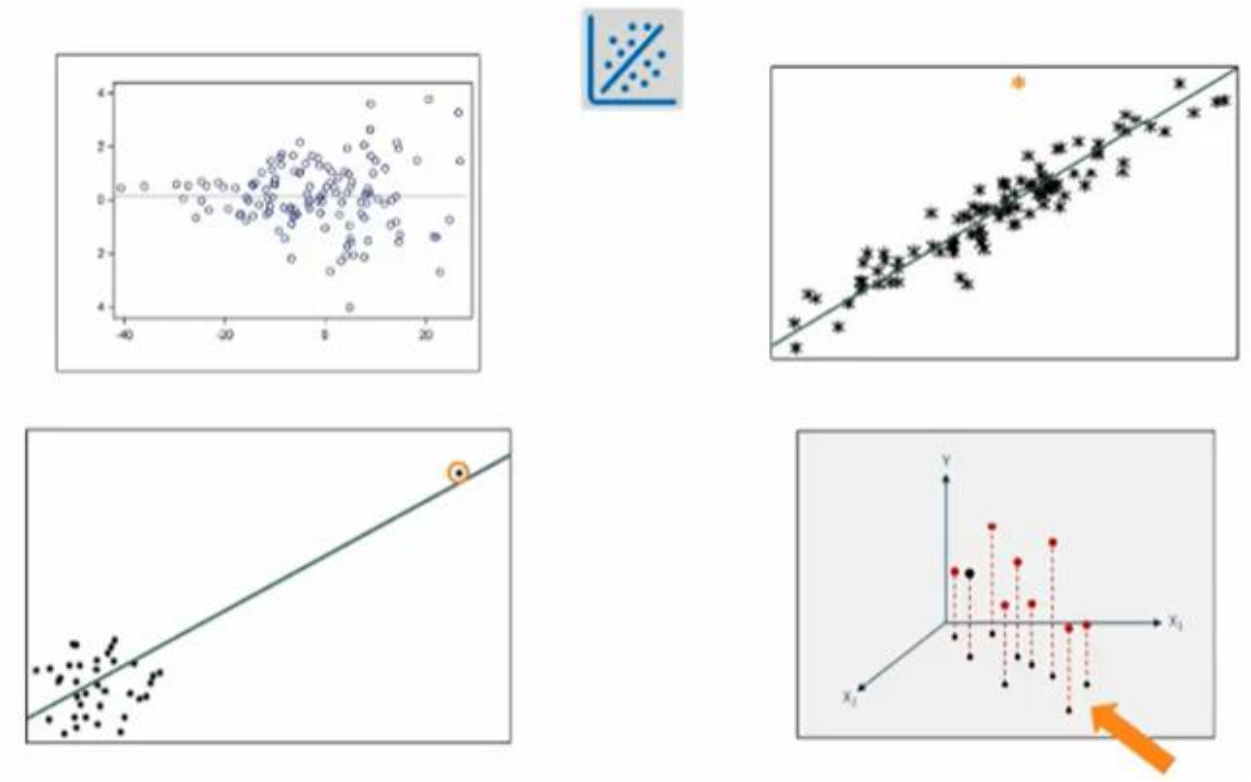
SBA Statistical Business Analyst with SAS

SBA2 Regression Modeling Fundamentals

SBA202 Model Building for Scoring and Predictions

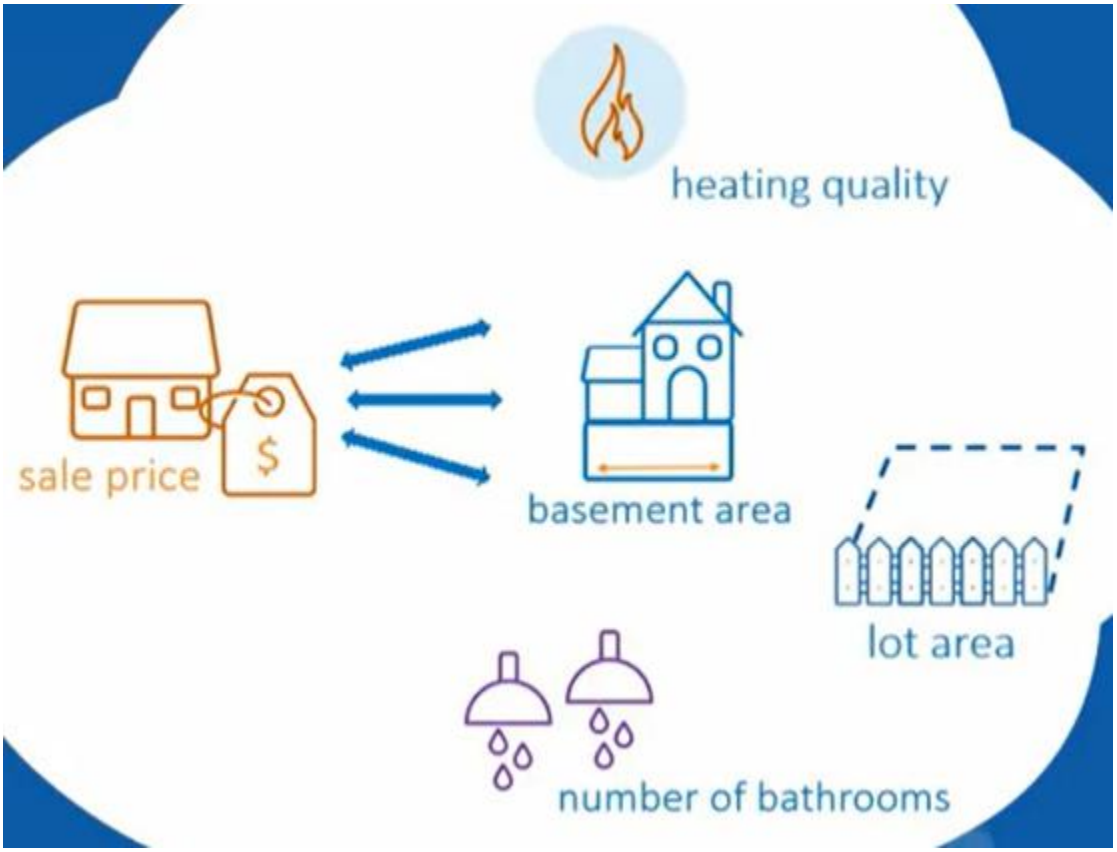
Overview

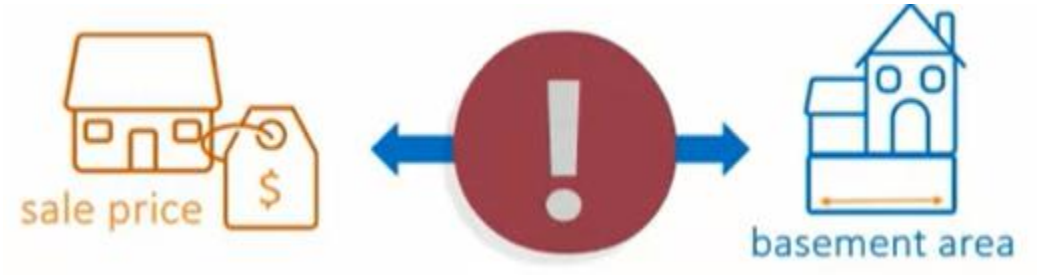

1
linear relationship
3
constant variance
2
normally distributed
4
independent



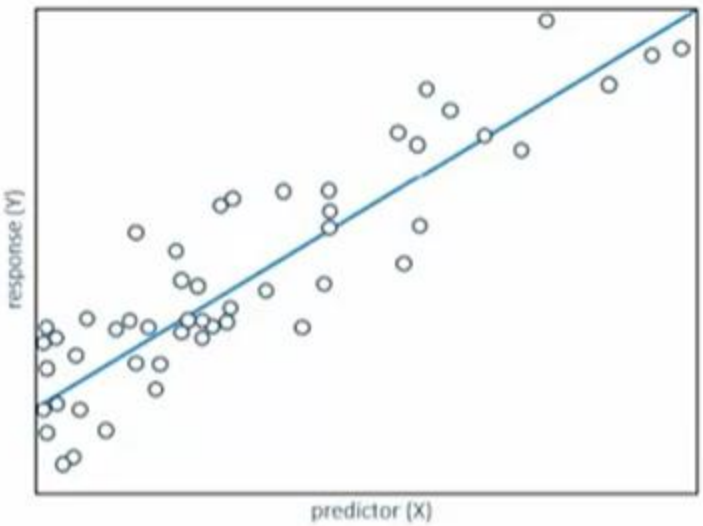
Examining Residuals

Scenario

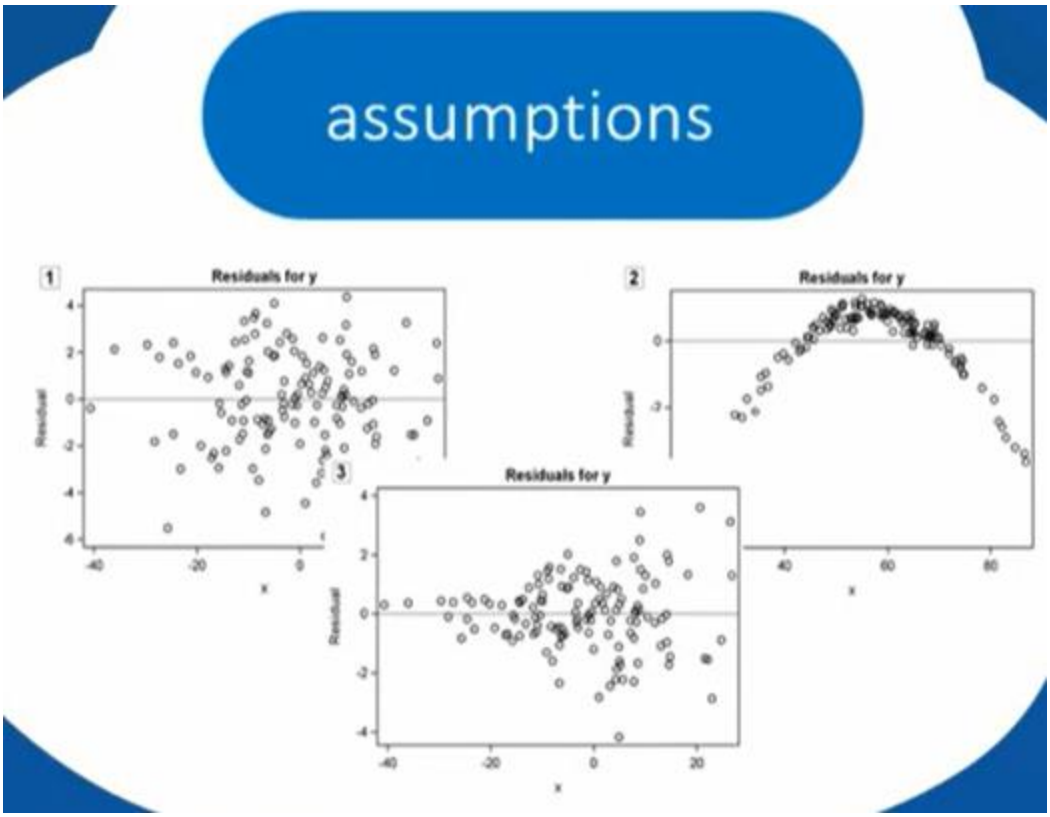




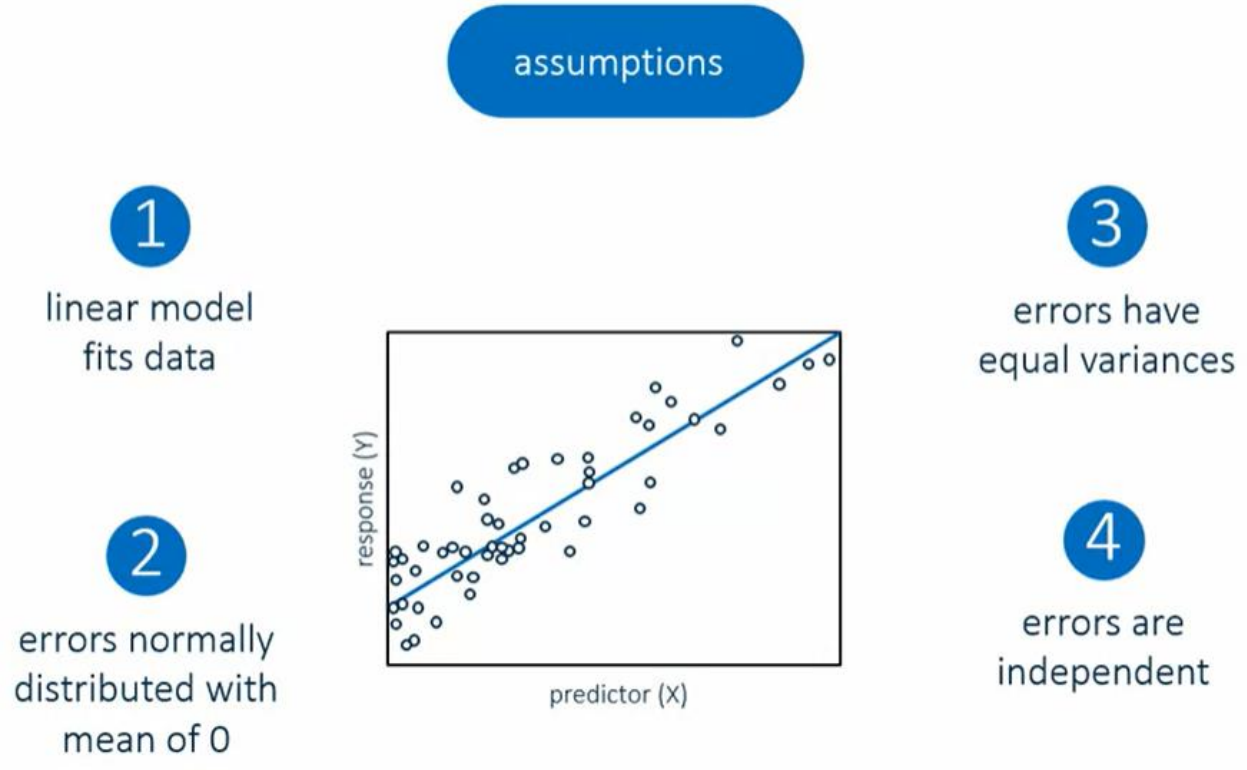
$\neq 0$



p-value

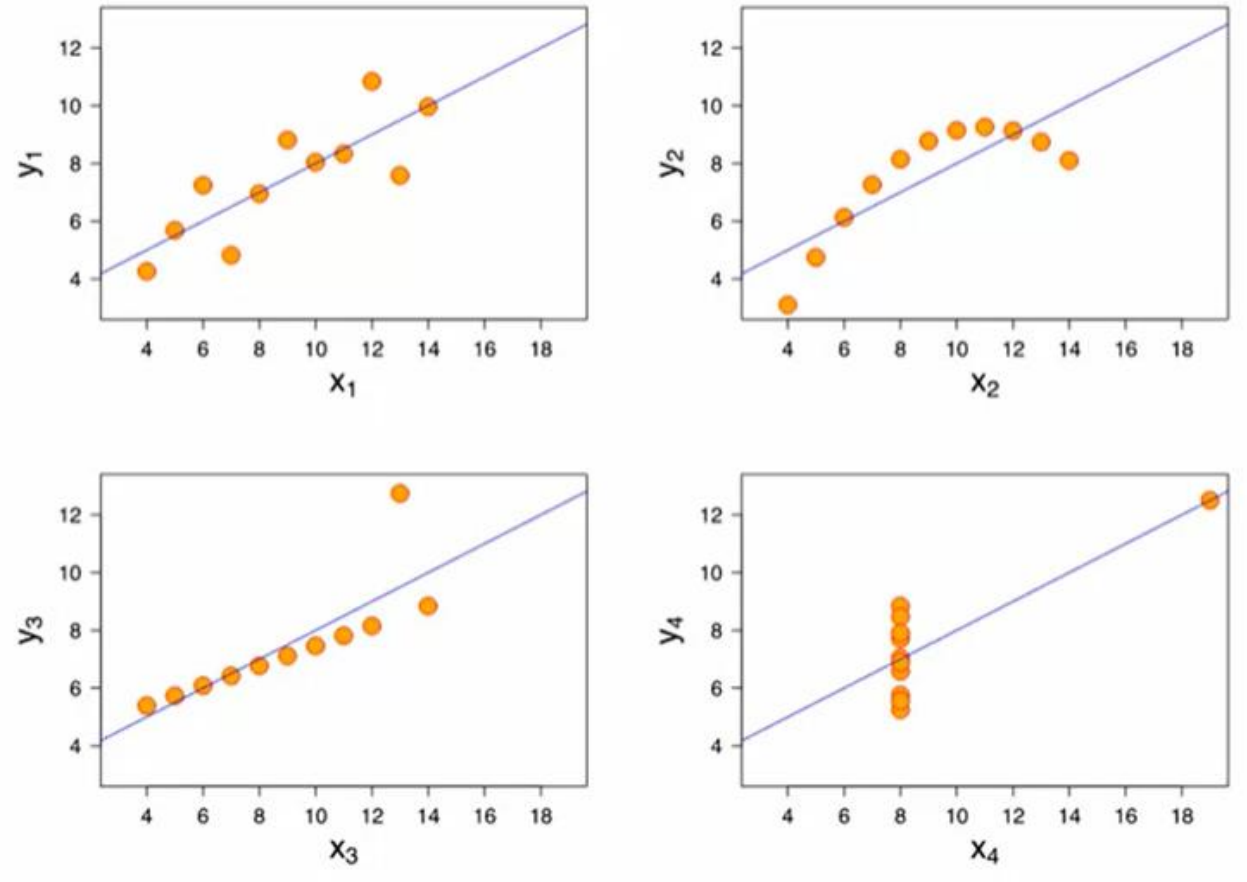

assumptions
1
Residuals for y
Residual
x
2
Residuals for y
Residual
x
3
Residuals for y
Residual
x

Assumptions for Regression

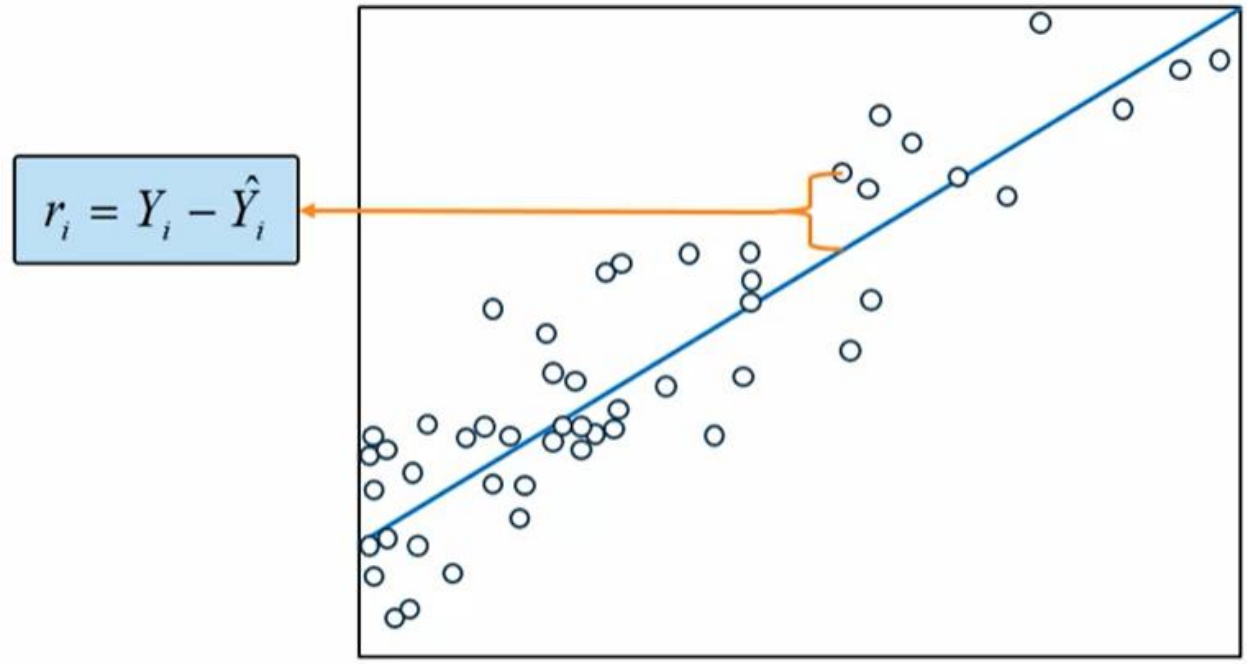


Model: $Y = 3.0 + 0.5X$

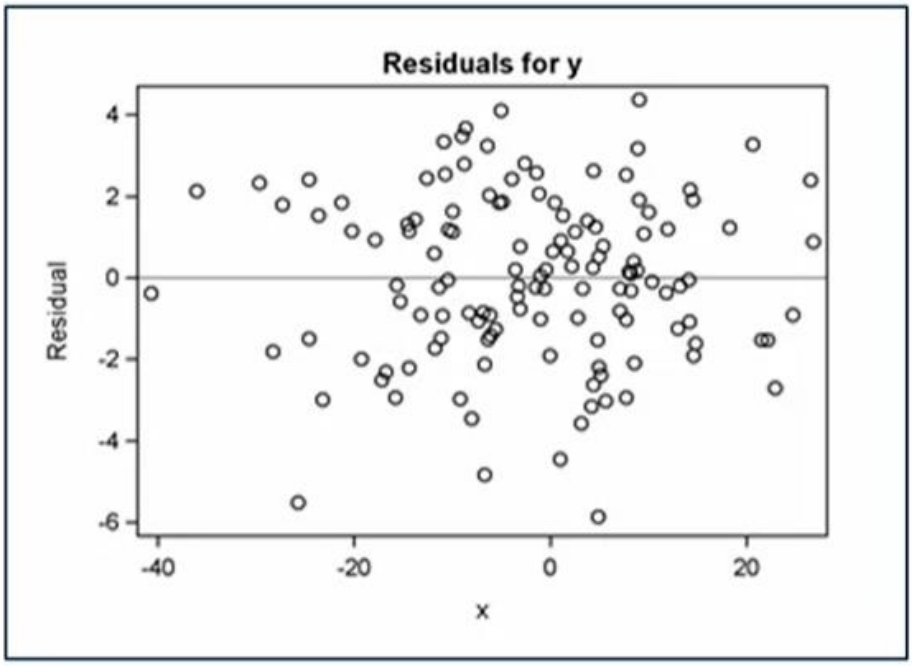
R^2=.67



Verifying Assumptions Using Residual Plots

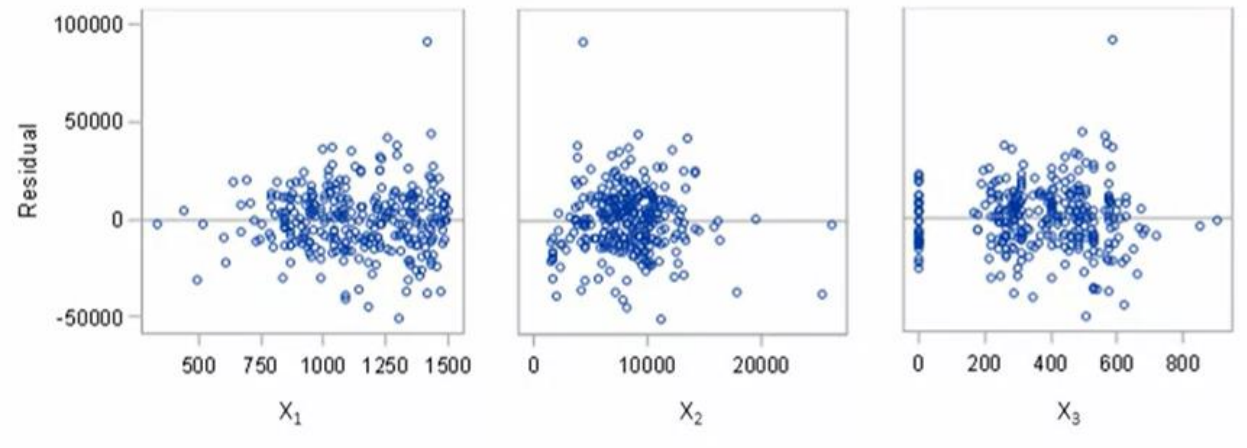


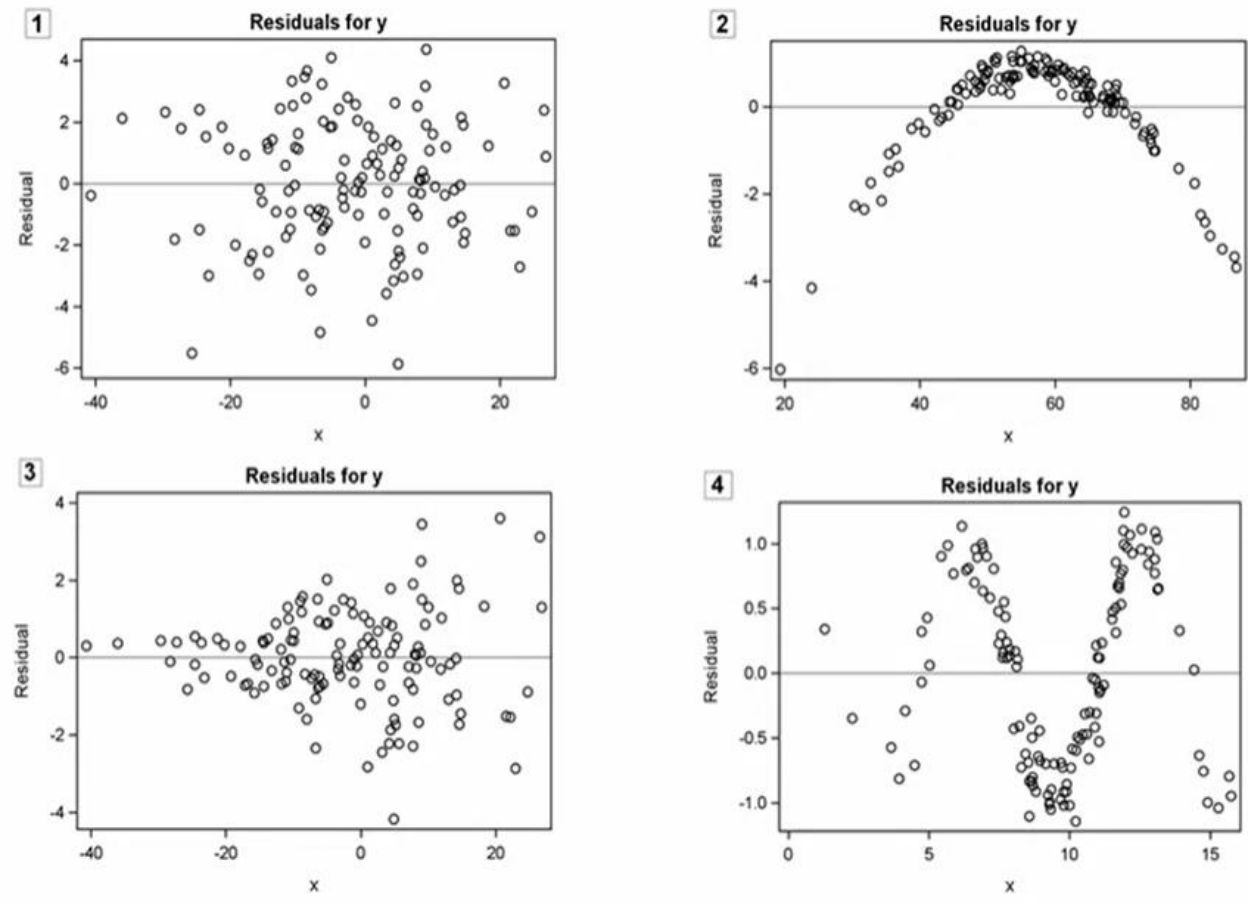
residuals versus predicted values

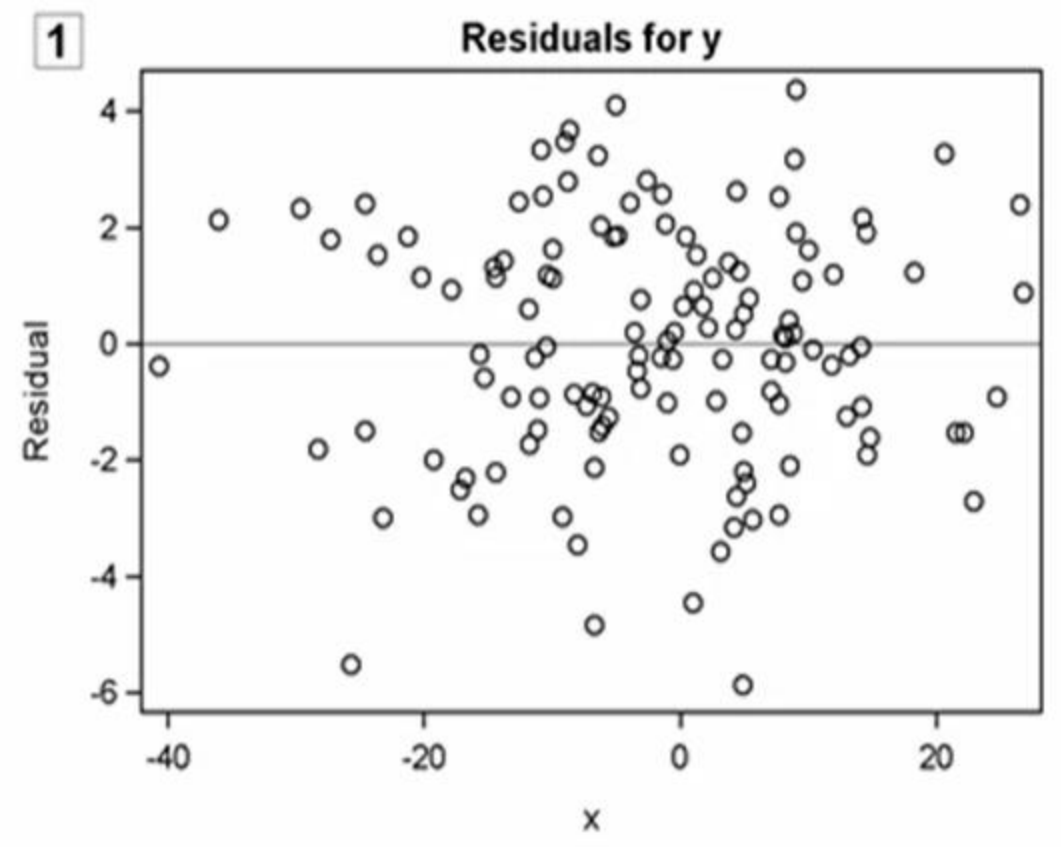


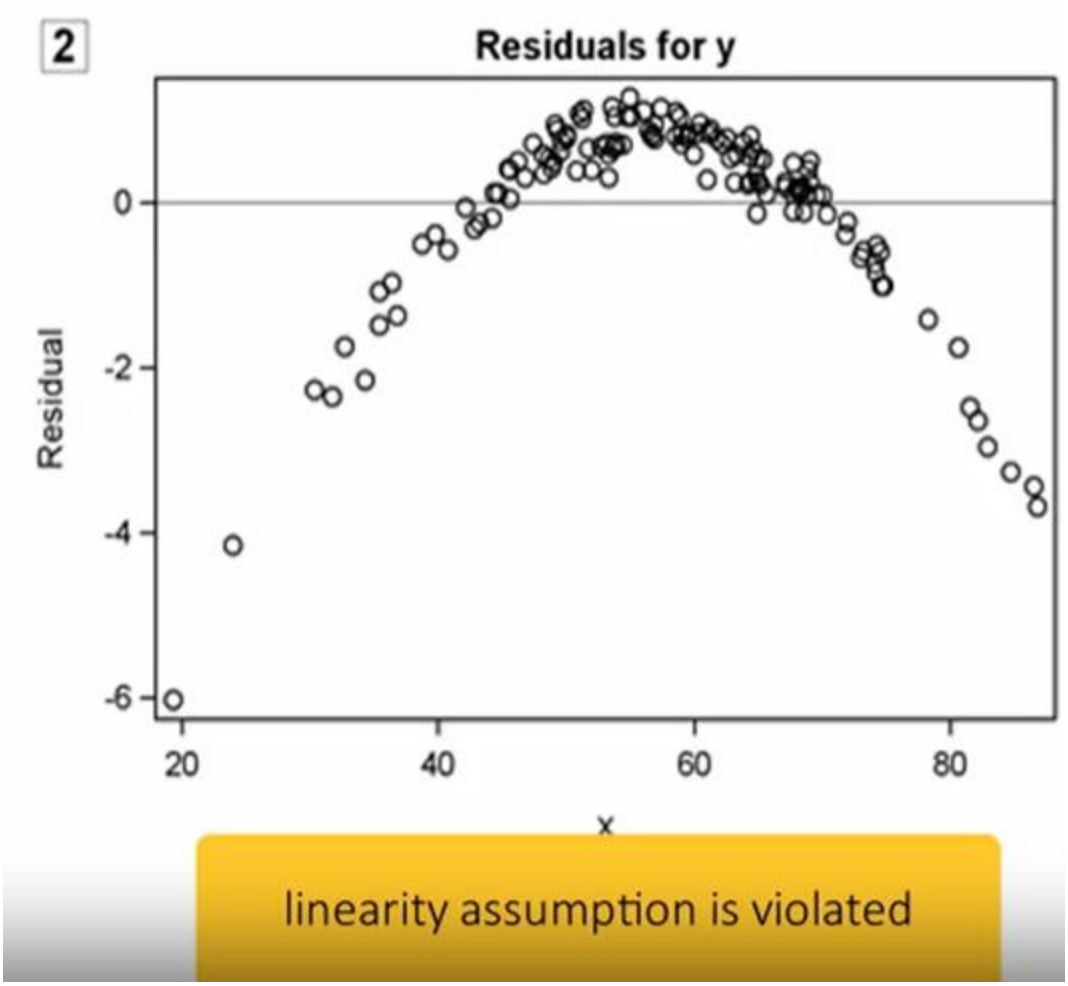
check for violations of equal variances, linearity, independence

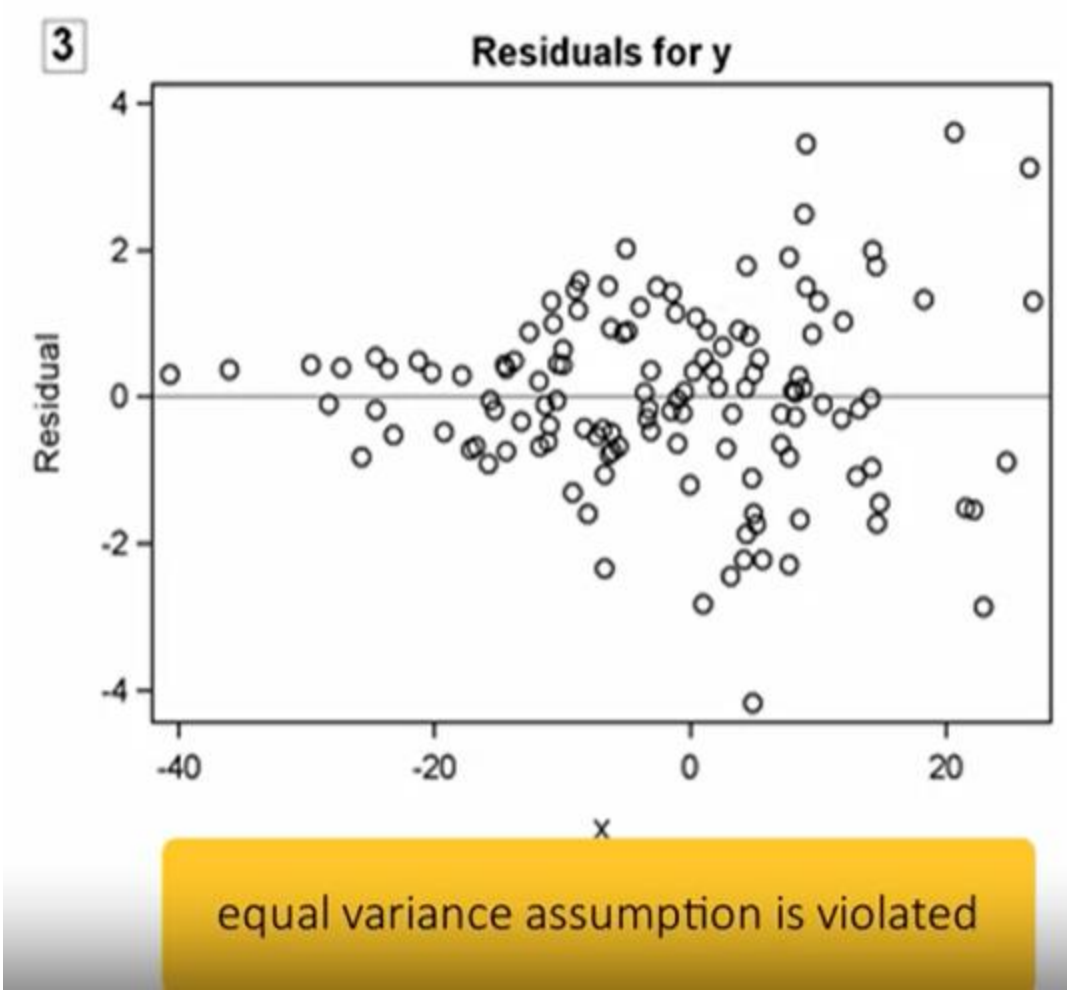
residuals versus independent variables

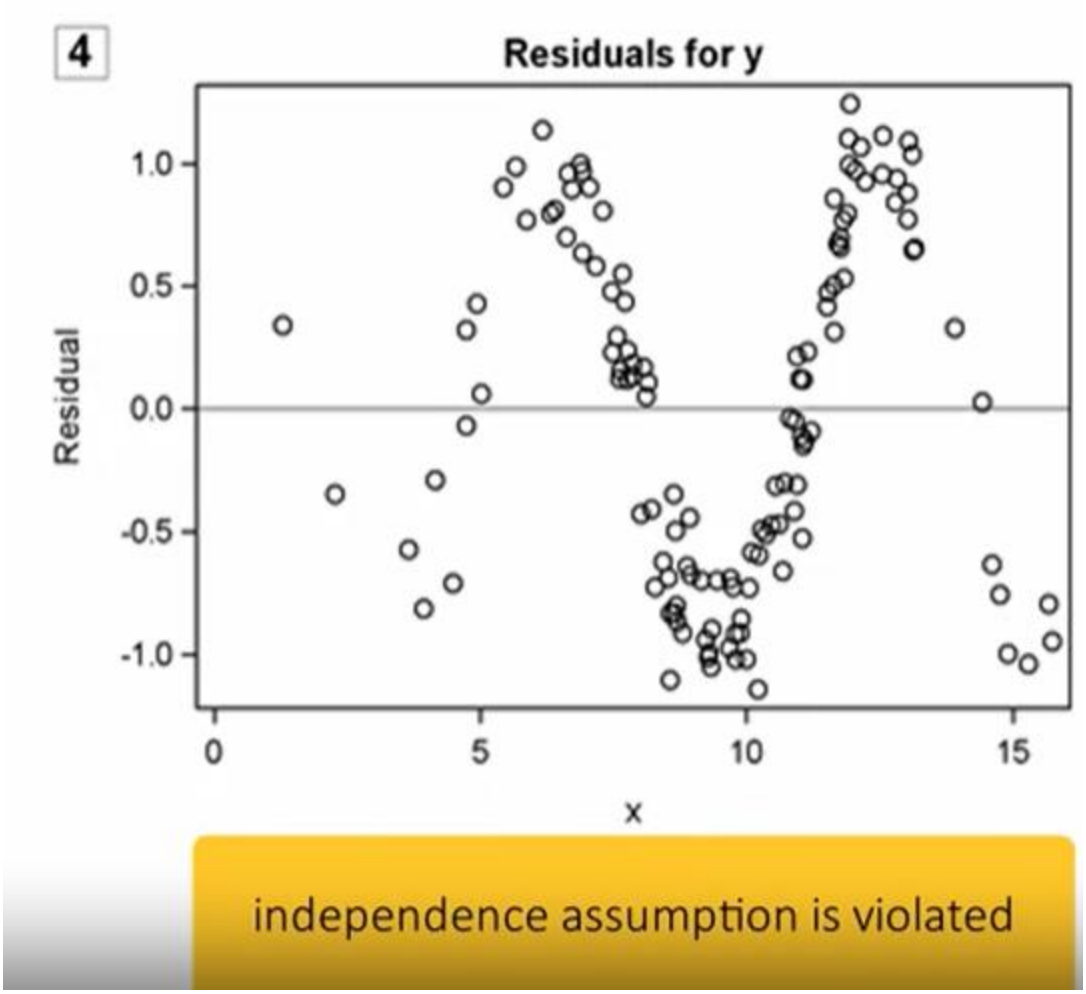


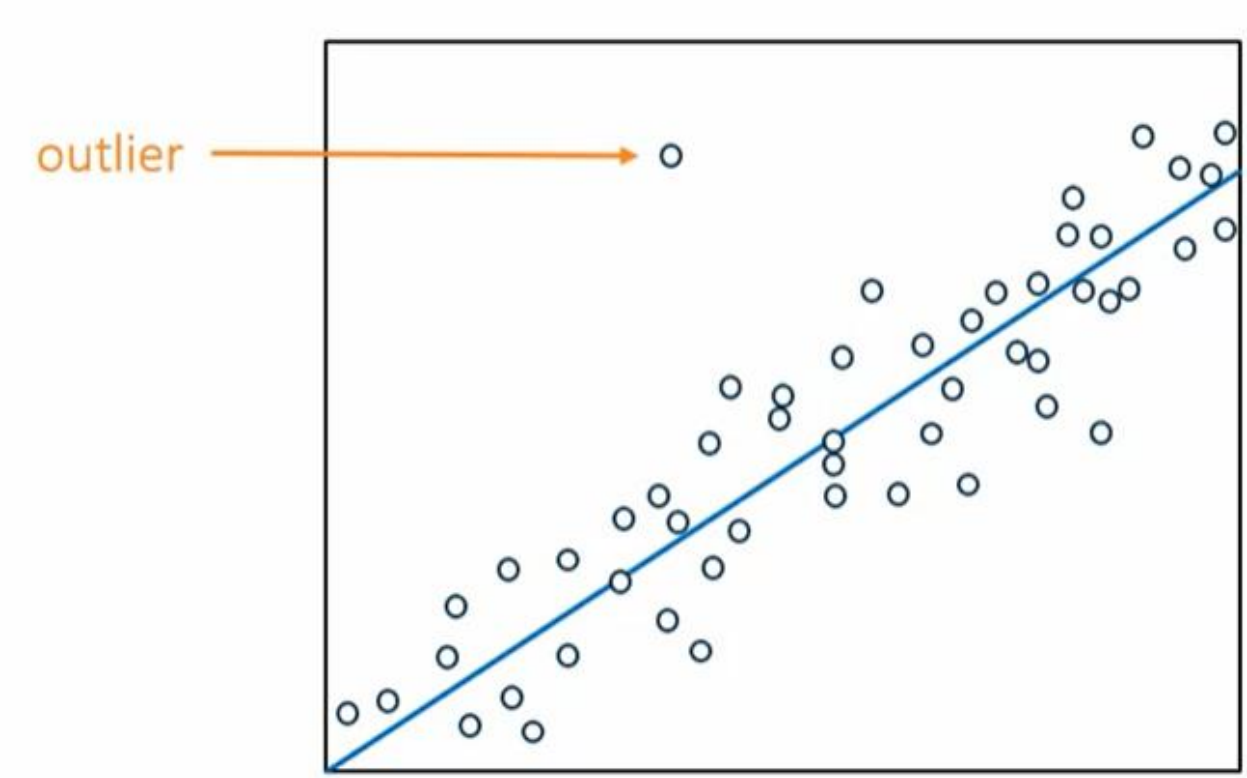



1
Residuals for y
Residual
4
2
0
-2
-4
-6
-40
-20
0
20
x

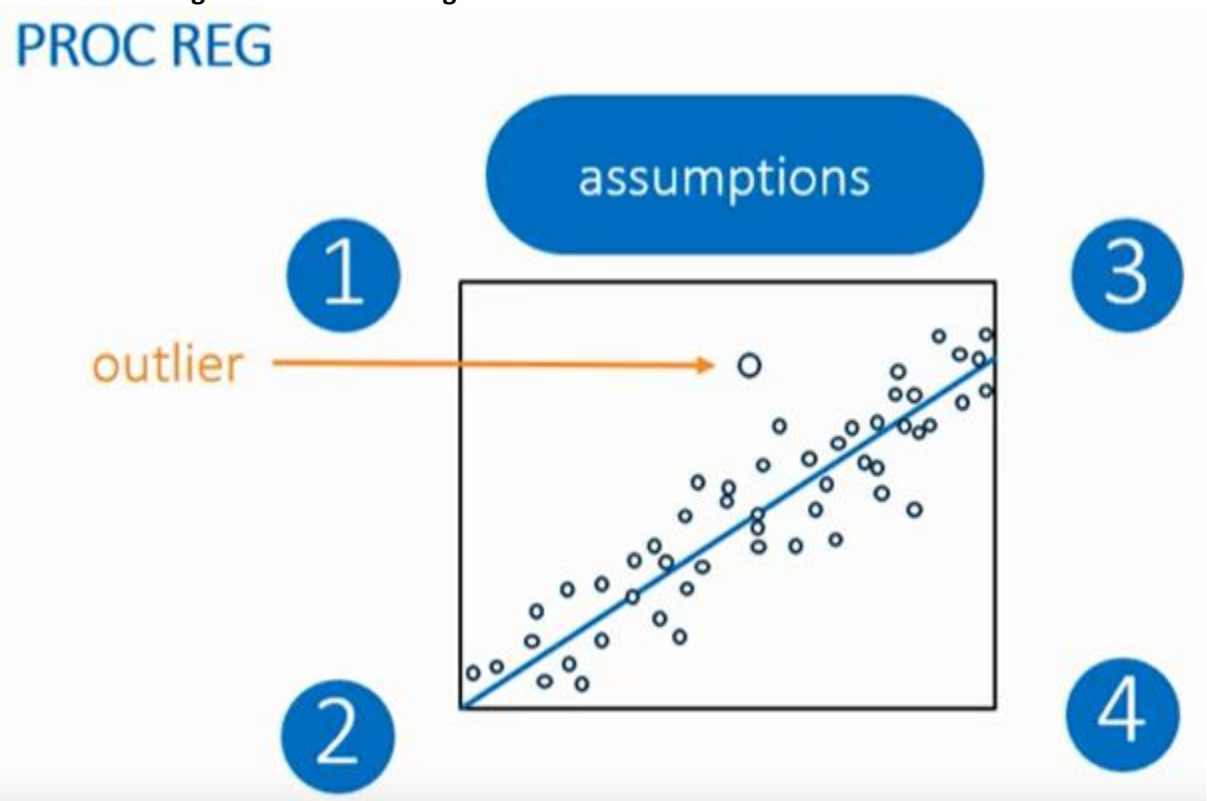

2
Residuals for y
Residual
0
-2
-4
-6
20
40
60
80
x
linearity assumption is violated

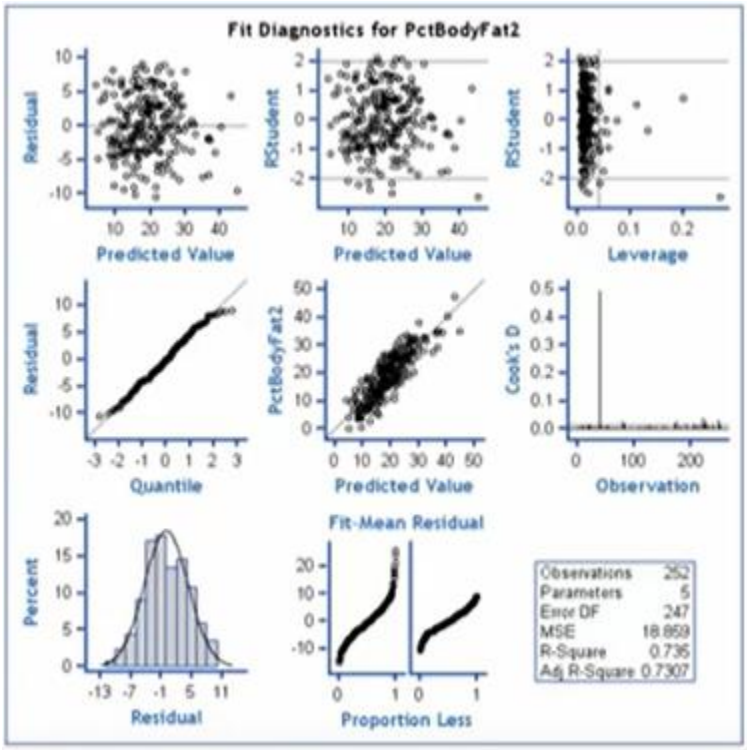

3
Residuals for y
Residual
4
2
0
-2
-4
-40
-20
0
20
x
equal variance assumption is violated


4
Residuals for y
Residual
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
x
independence assumption is violated


outlier

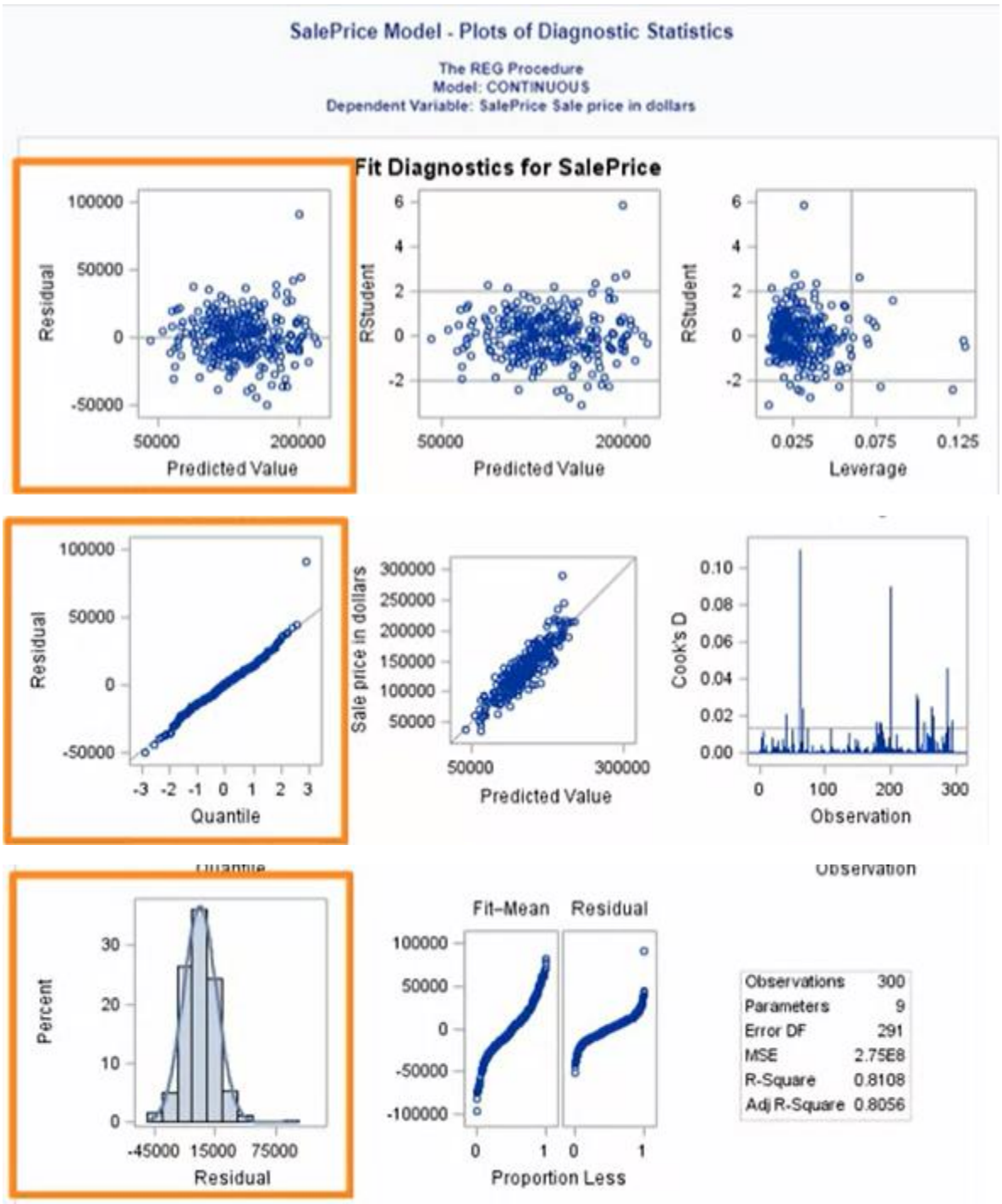
Demo Examining Residual Plots Using PROC REG

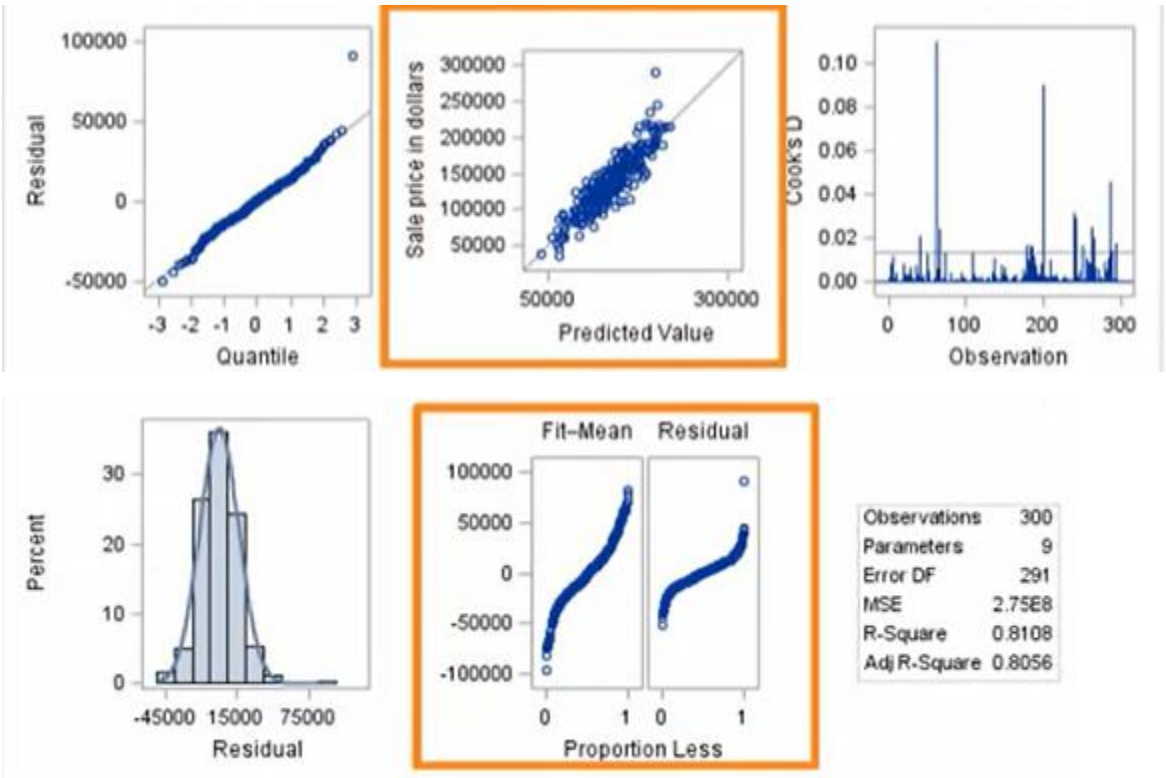




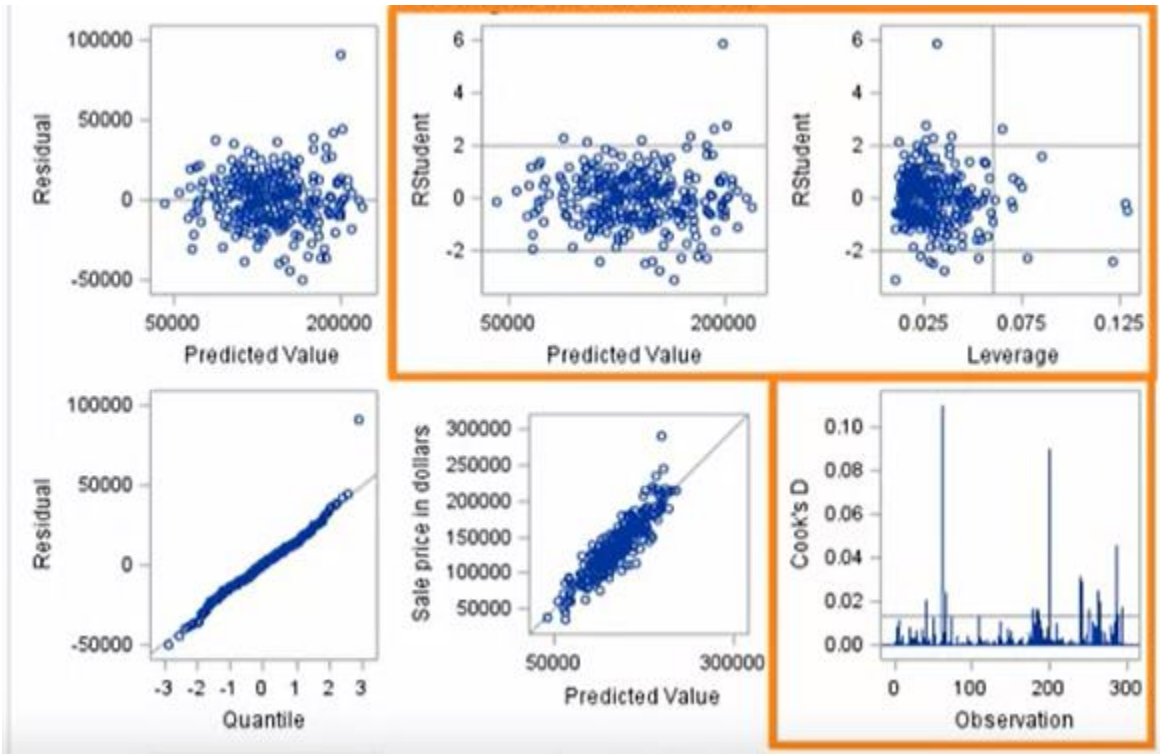
```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d01.sas*/  /*Part A*/
ods graphics on;
proc reg data=STAT1.ameshousing3;
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

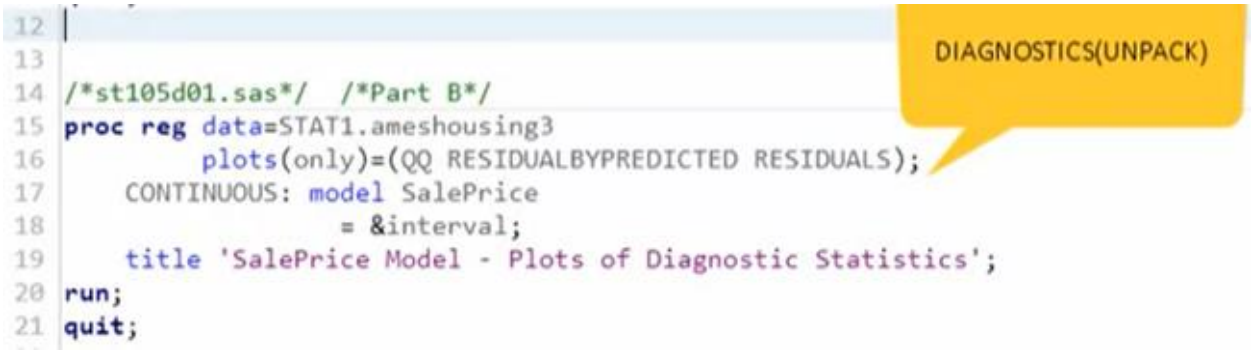
PROC REG DATA=*SAS-data-set <options>*;
MODEL *dependents = <regressors> < / options>*;
RUN;





These three charts below could be used to detect outliers





/*st105s01.sas*/

ods graphics / imagemap=on;

proc reg data=STAT1.BodyFat2

plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);

FORWARD: model PctBodyFat2

= Abdomen Weight Wrist Forearm;

id Case;

title 'FORWARD Model - Plots of Diagnostic Statistics';

run;

quit;

FORWARD Model - Plots of Diagnostic Statistics

The REG Procedure
Model: FORWARD
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

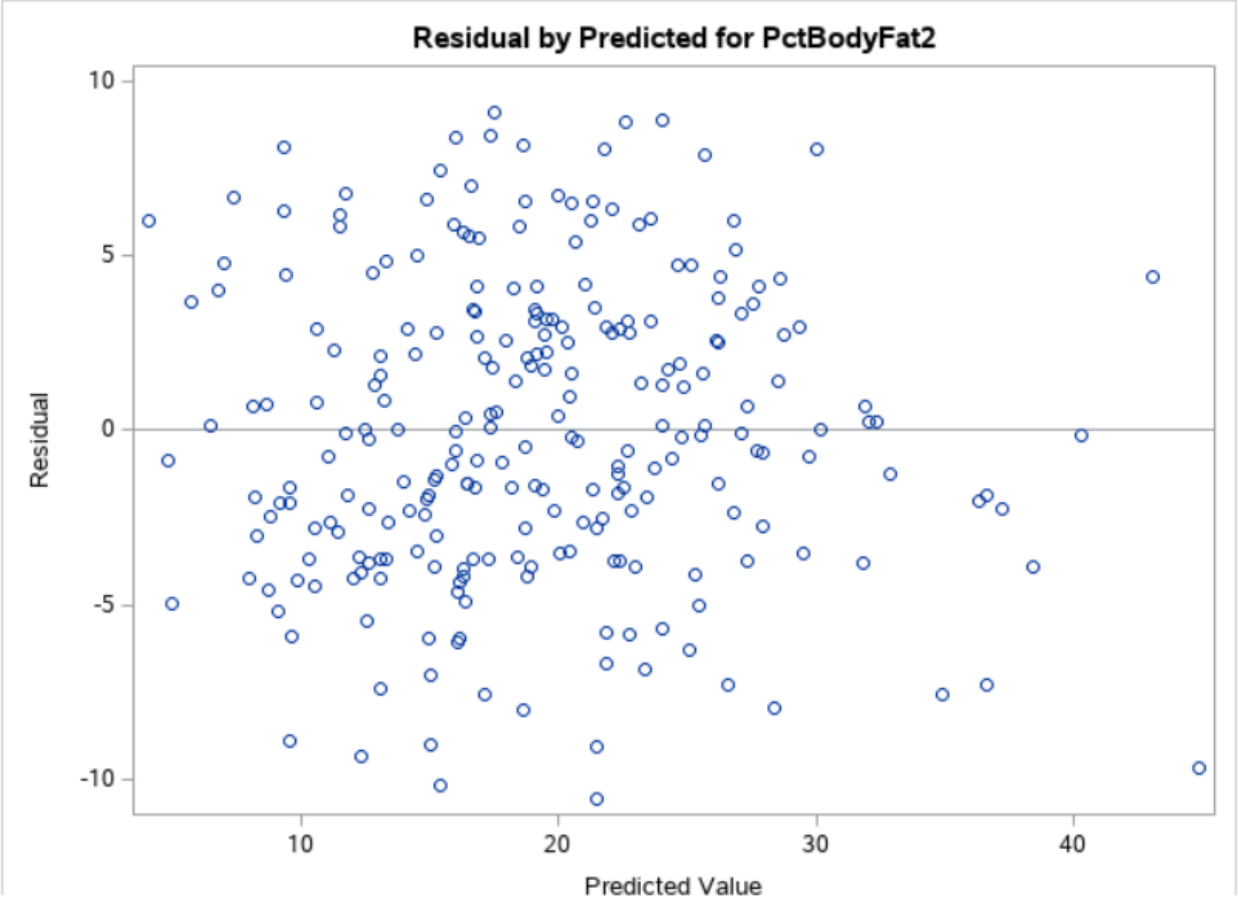
Root MSE	4.34272	R-Square	0.7350
Dependent Mean	19.15079	Adj R-Sq	0.7307
Coeff Var	22.67647		

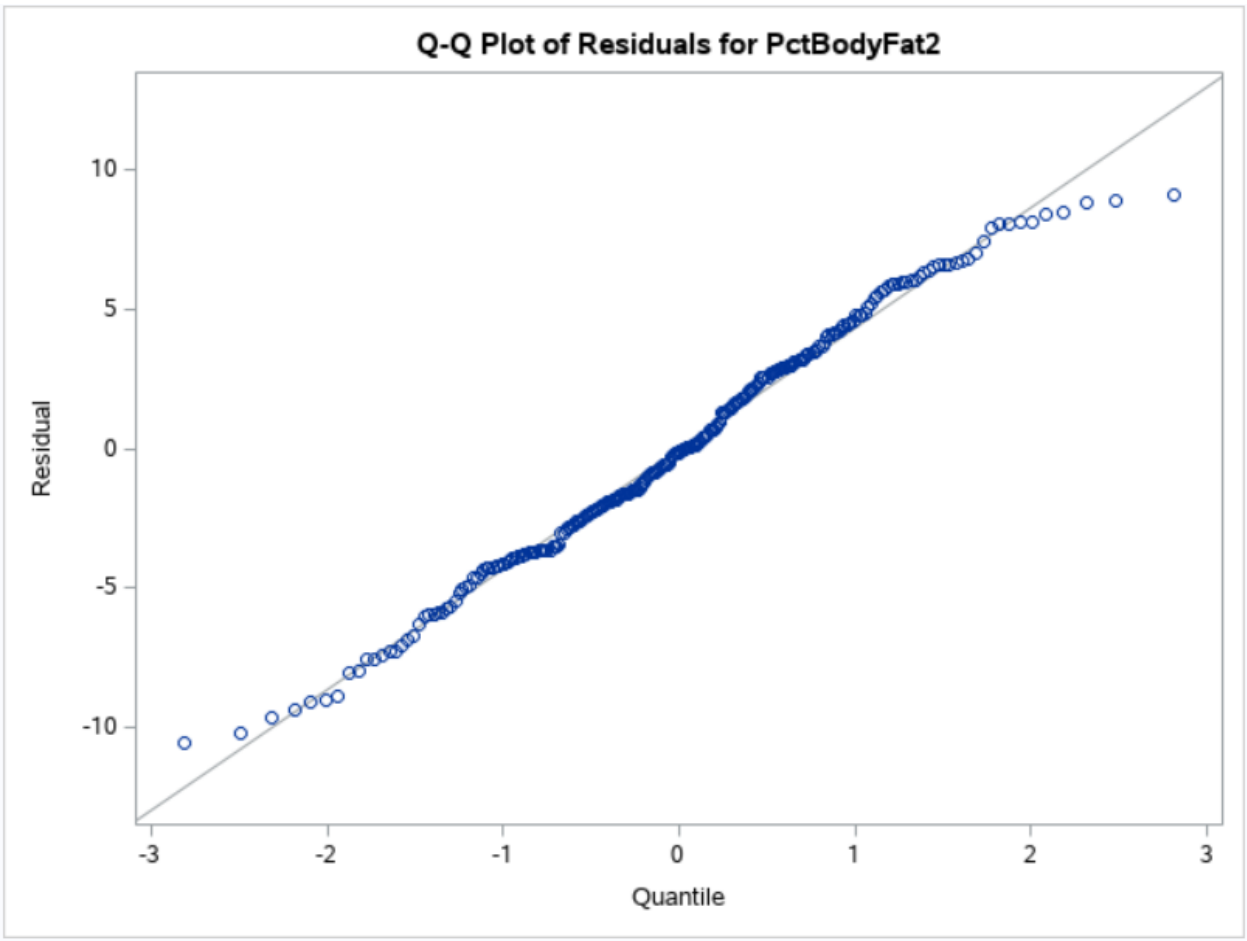
Parameter Estimates

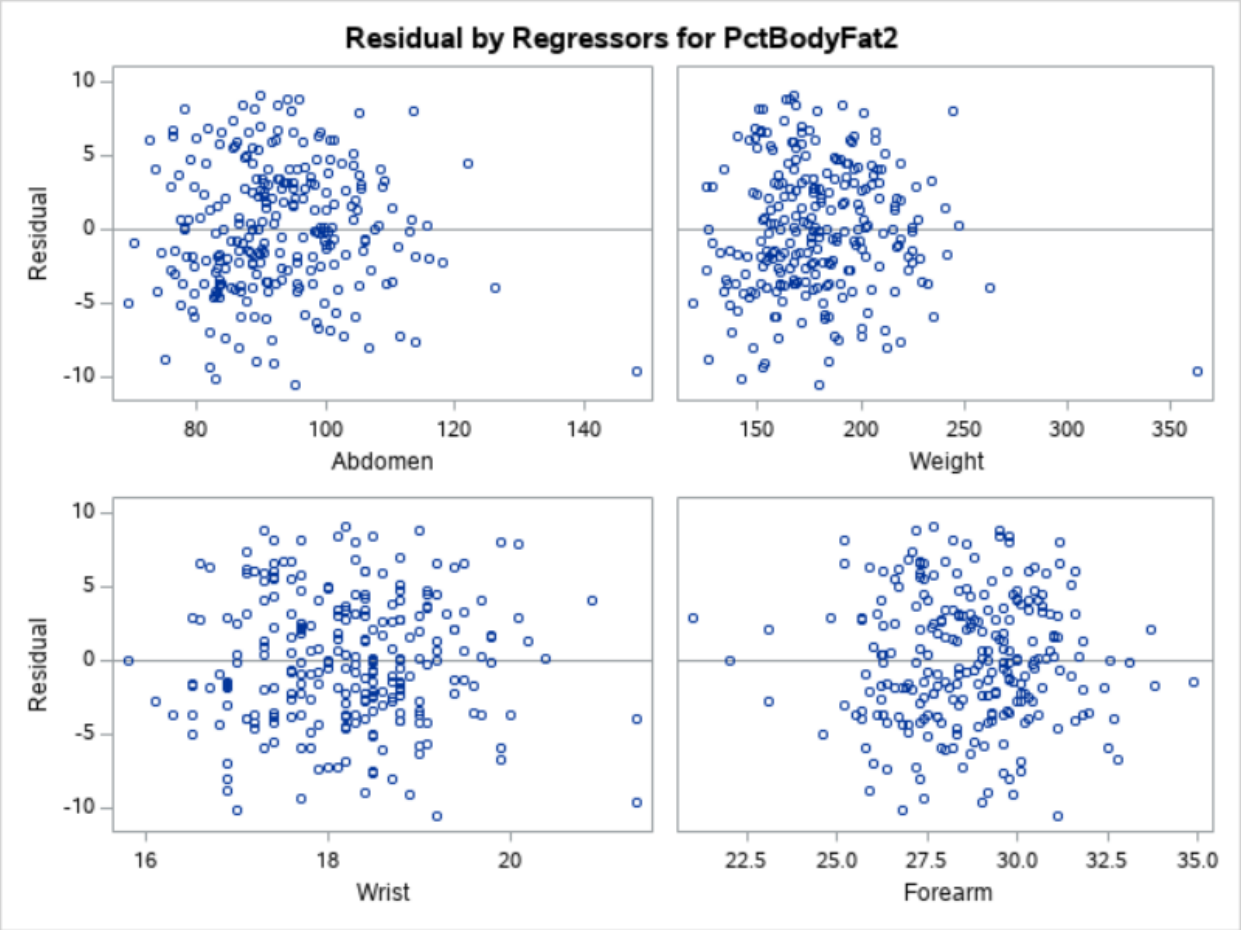
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-34.85407	7.24500	-4.81	<.0001
Abdomen	1	0.99575	0.05607	17.76	<.0001
Weight	1	-0.13563	0.02475	-5.48	<.0001
Wrist	1	-1.50556	0.44267	-3.40	0.0008
Forearm	1	0.47293	0.18166	2.60	0.0098

FORWARD Model - Plots of Diagnostic Statistics

The REG Procedure
Model: FORWARD
Dependent Variable: PctBodyFat2




Q-Q Plot of Residuals for PctBodyFat2
Residual
10
5
0
-5
-10
-3
-2
-1
0
1
2
3
Quantile


Residual by Regressors for PctBodyFat2
Residual
10
5
0
-5
-10
80
100
120
140
Abdomen
150
200
250
300
350
Weight
16
18
20
Wrist
22.5
25.0
27.5
30.0
32.5
35.0
Forearm

Practice: Using PROC REG to Examine Residuals

Question 1

Run a regression on **PctBodyFat2** in the **stat1.bodyfat2** data set to examine residuals.

1. Use PROC REG to run a regression model of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**.
2. Create plots of the residuals by the four regressors and by the predicted values, and a normal Q-Q plot.
3. Submit the code and view the results.

Do the residual plots indicate any problems with the constant variance assumption?

It doesn't appear that the data violate the assumption of constant variance. Also, the residuals show nice random scatter and indicate no problem with model specification. Solution code:

```
/*st105s01.sas*/

ods graphics / imagemap=on;

proc reg data=STAT1.BodyFat2
         plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
   FORWARD: model PctBodyFat2 = Abdomen Weight Wrist Forearm;
   id Case;
   title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

Question 2

Are there any outliers indicated by the evidence in any of the residual plots?

There are a few outliers for **Wrist** and **Forearm**, and one clear outlier in both **Abdomen** and **Weight**.

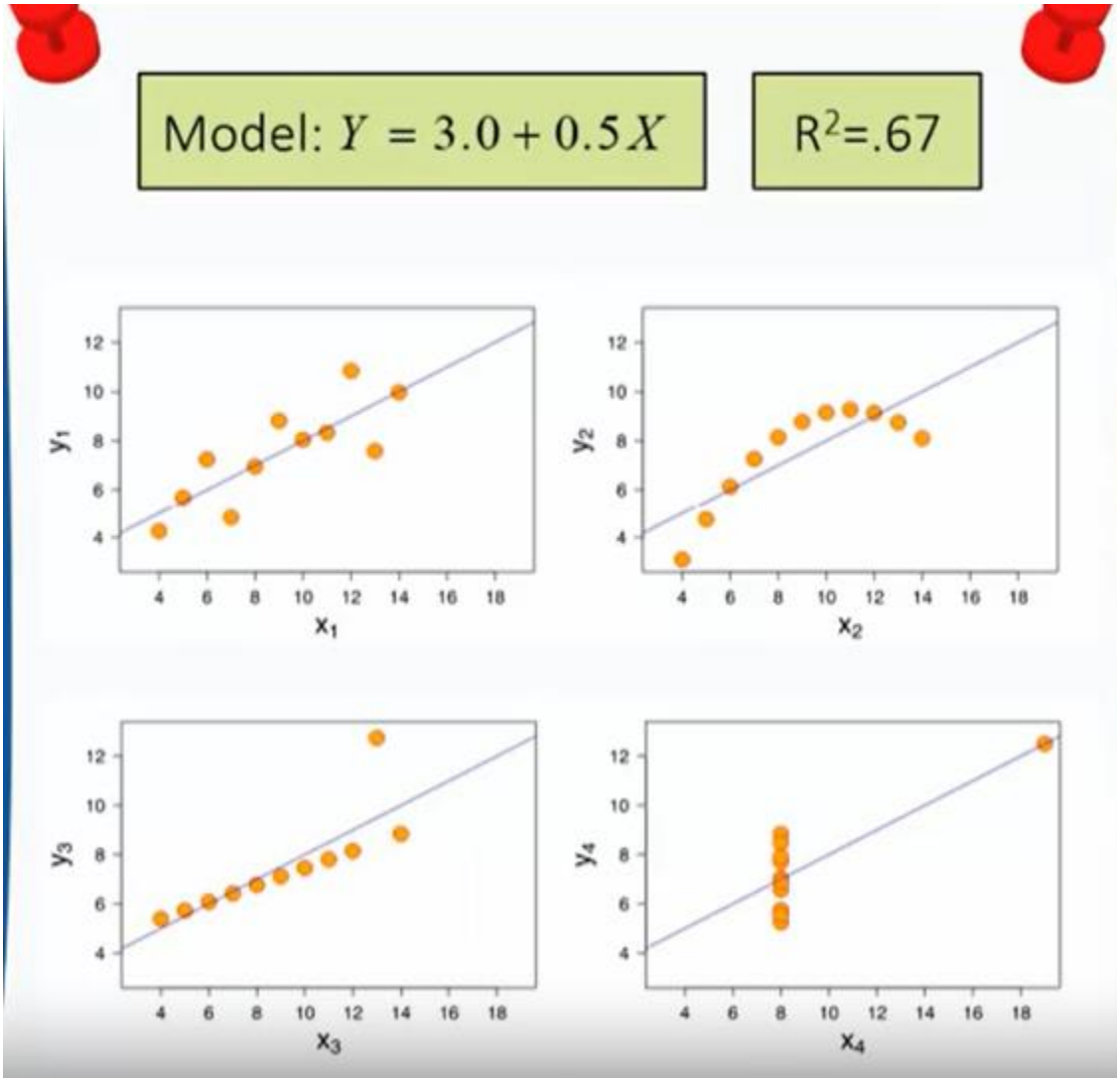
Question 3

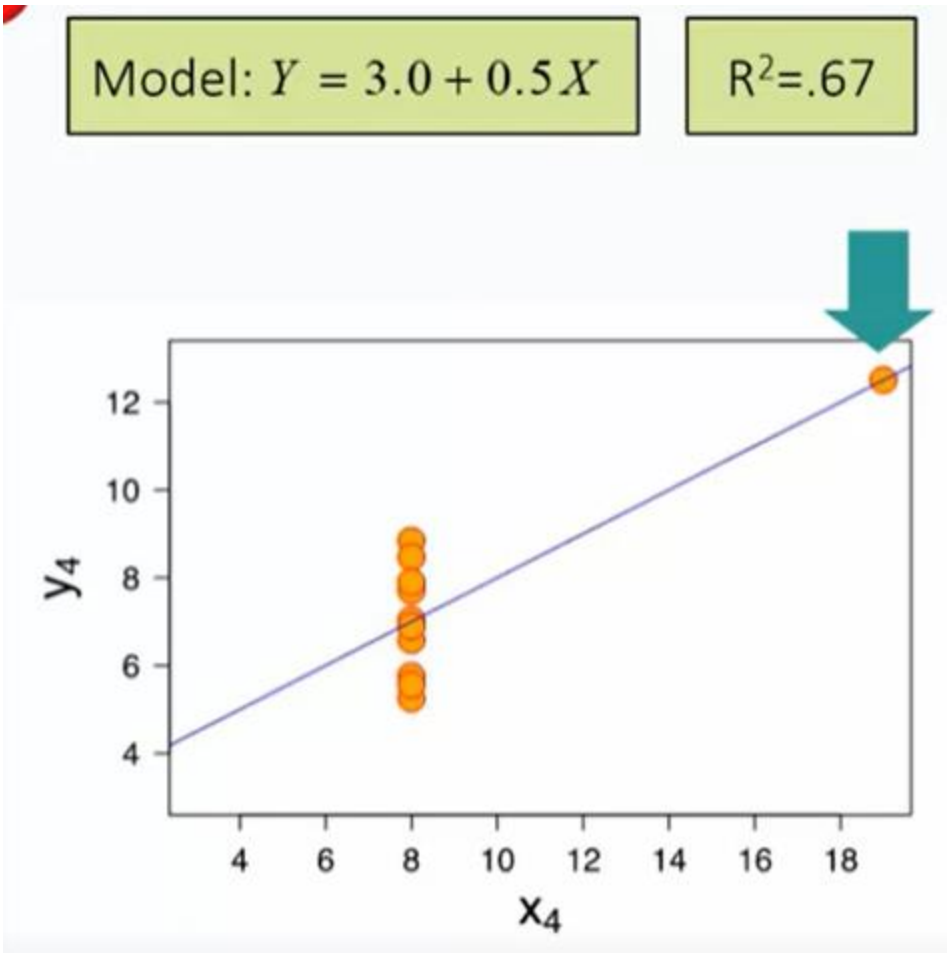
Does the Q-Q plot indicate any problems with the normality assumption?

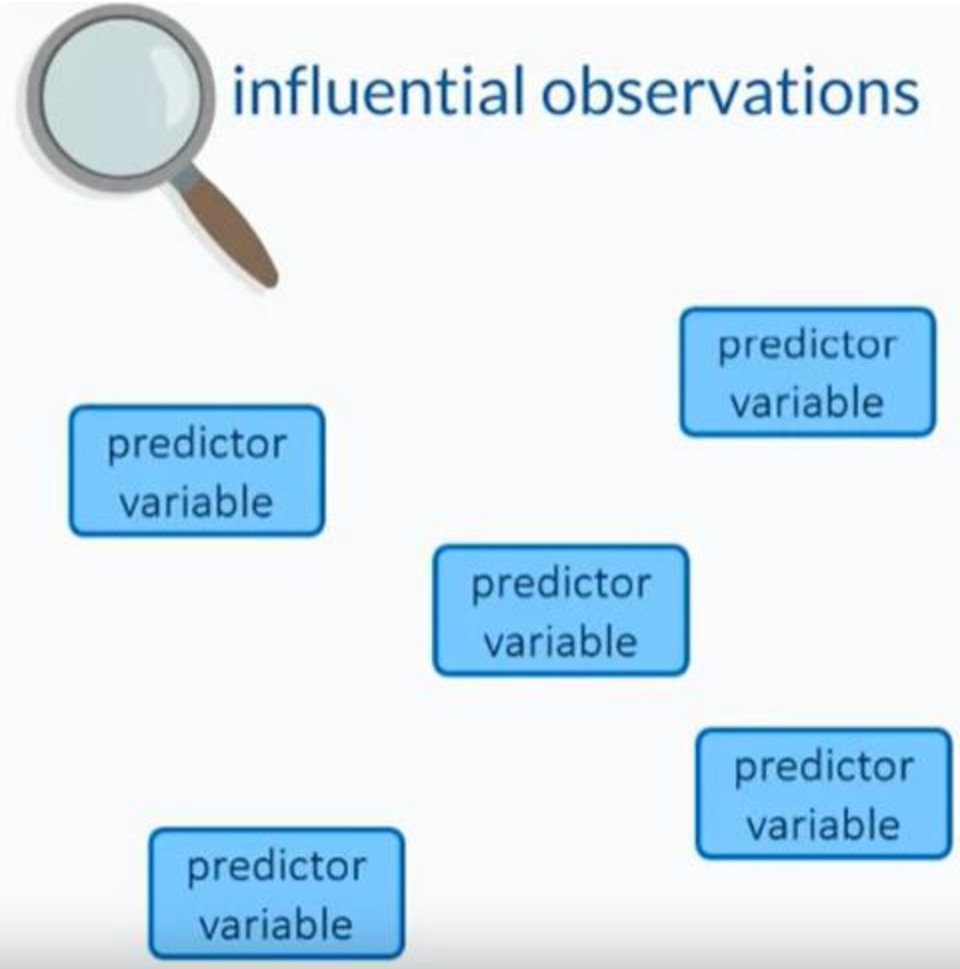
The normality assumption seems to be met.

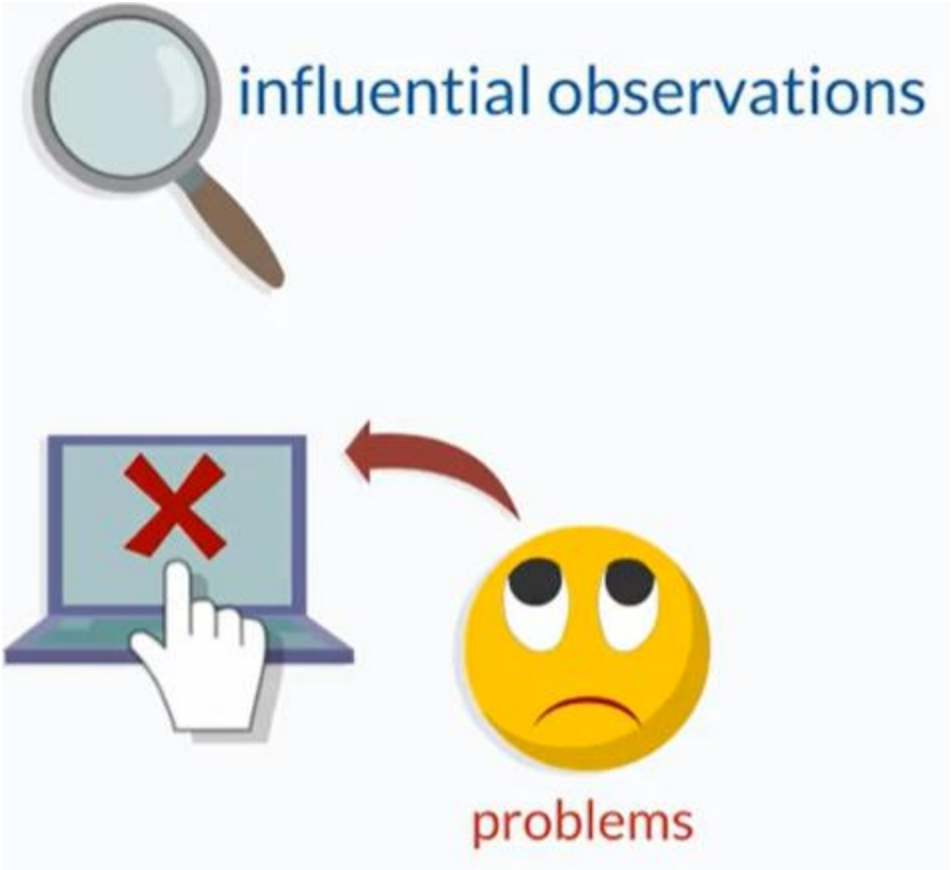
Influential Observations

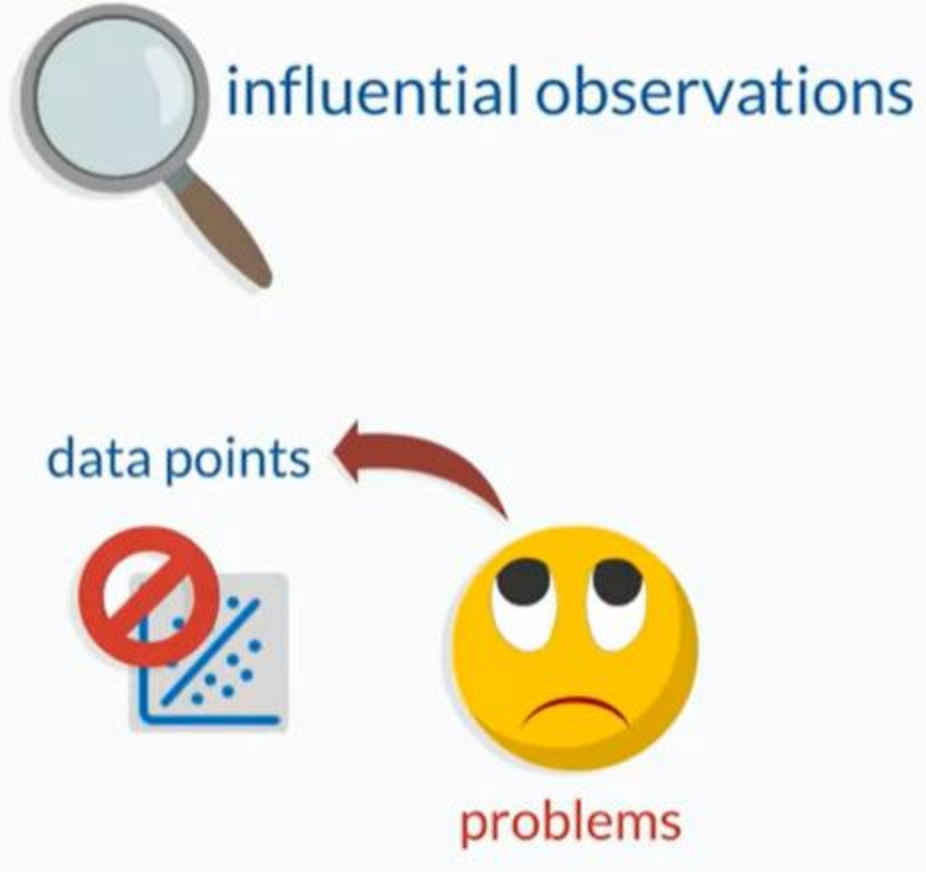
Scenario



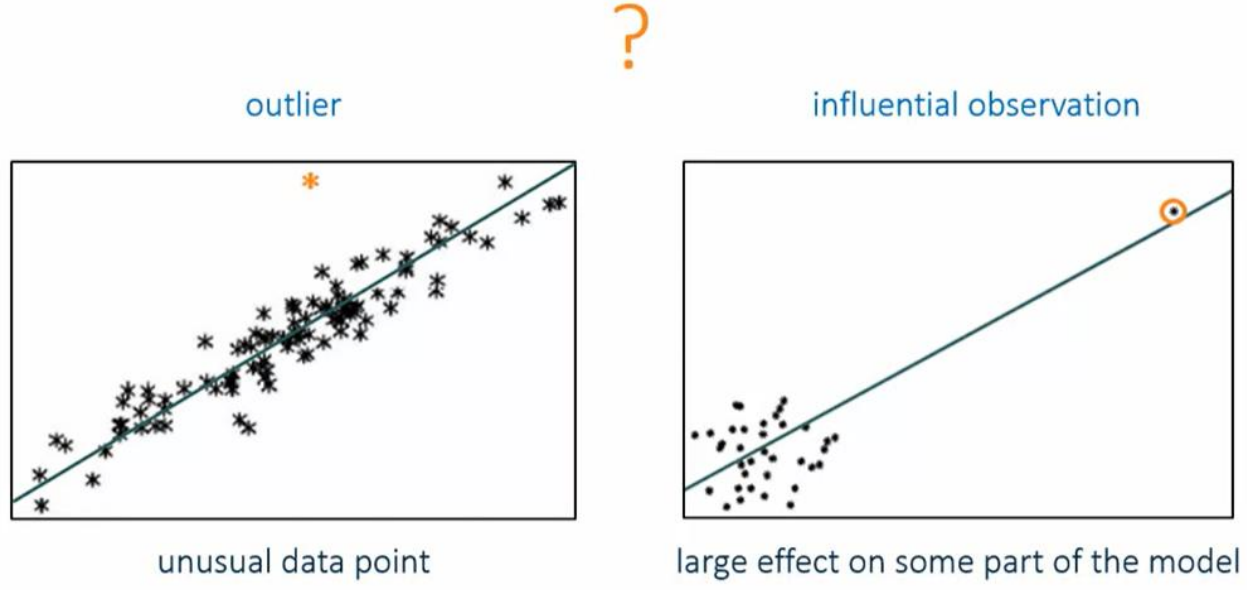

Model: $Y = 3.0 + 0.5X$
R^2=.67
12
10
8
6
4
y_4
4
6
8
10
12
14
16
18
x_4

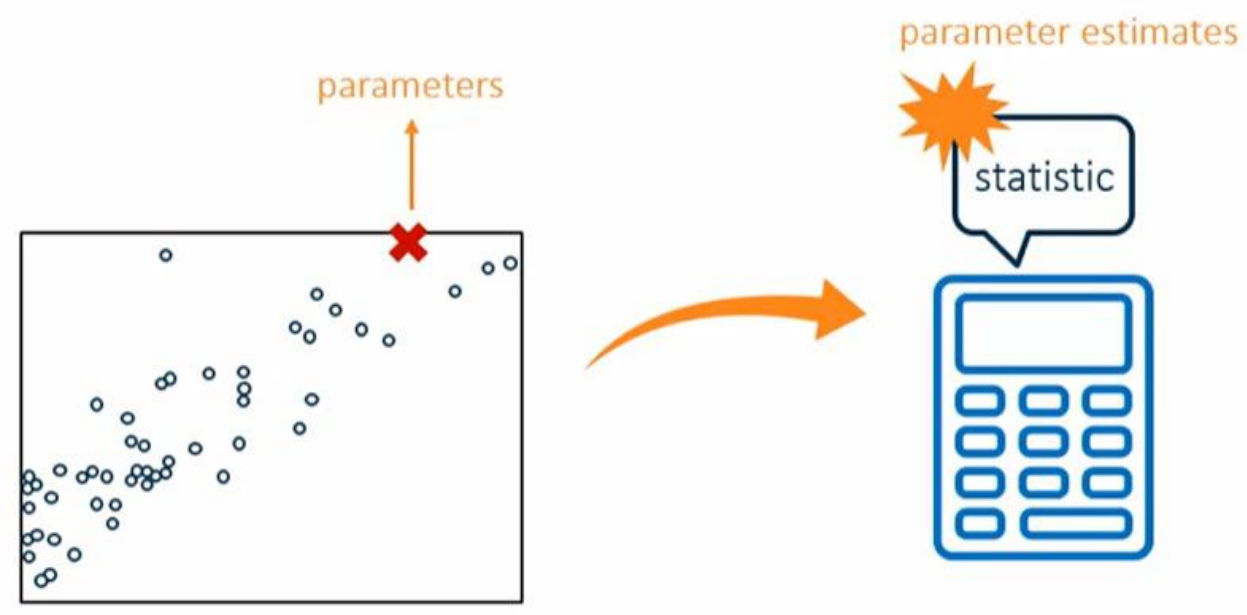

influential observations
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable

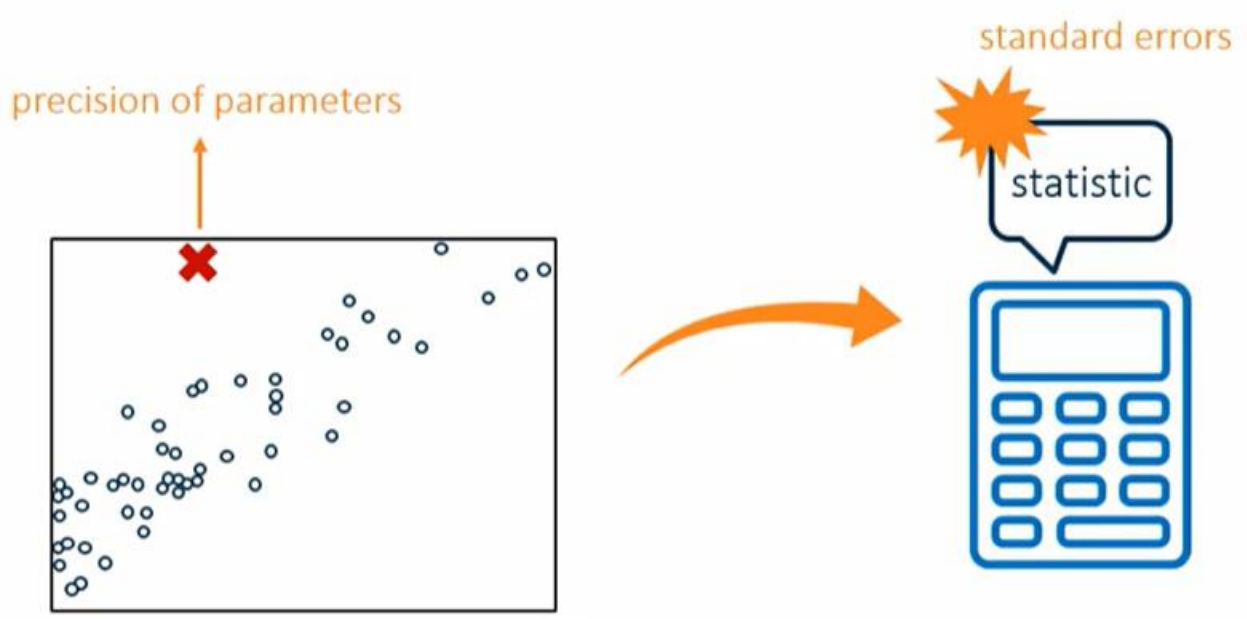

influential observations
problems

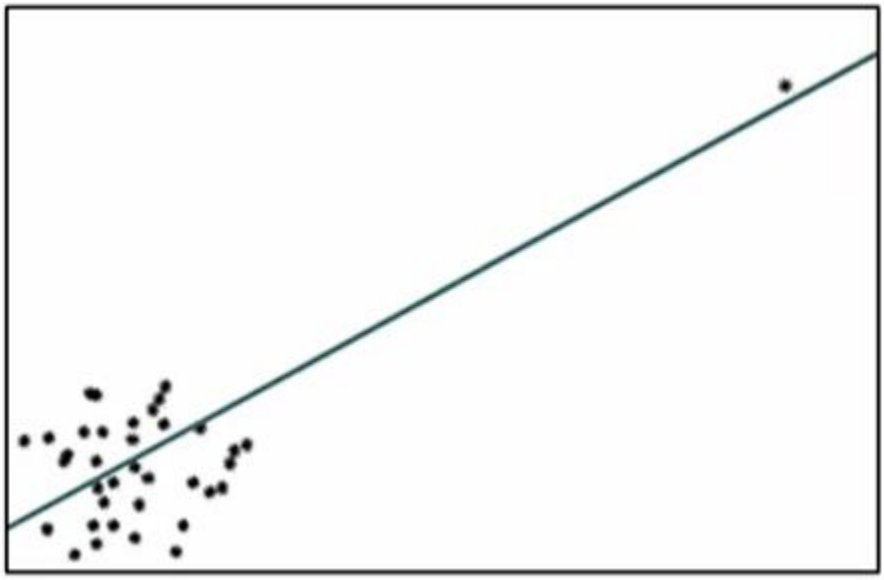

influential observations
data points
problems

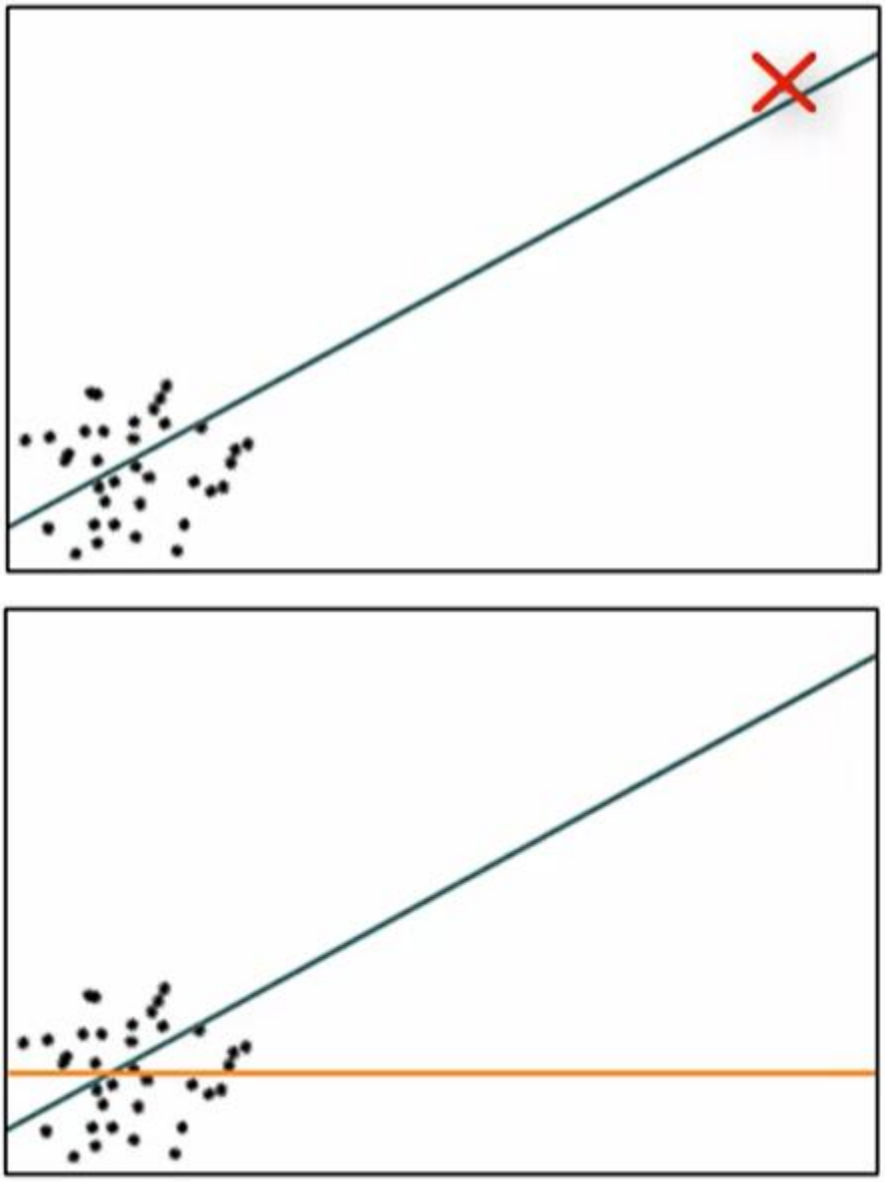
Identifying Influential Observations





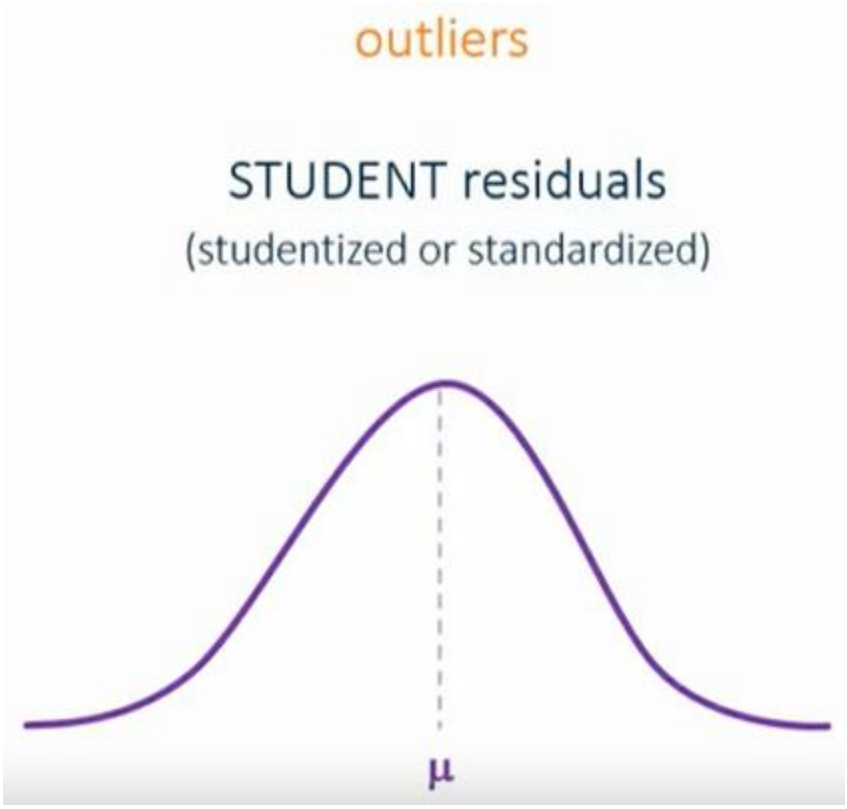

standard errors
precision of parameters
statistic

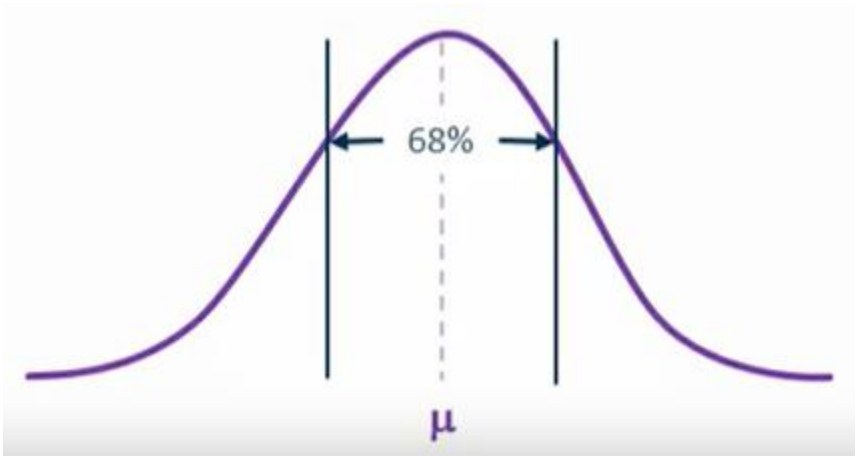


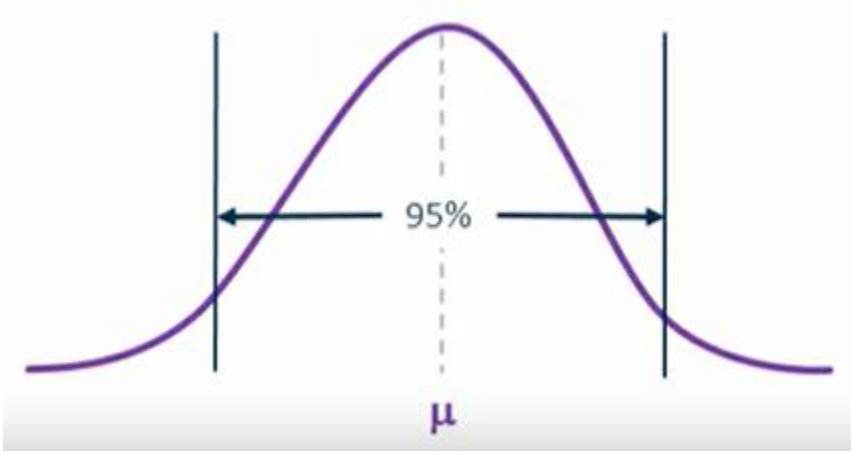


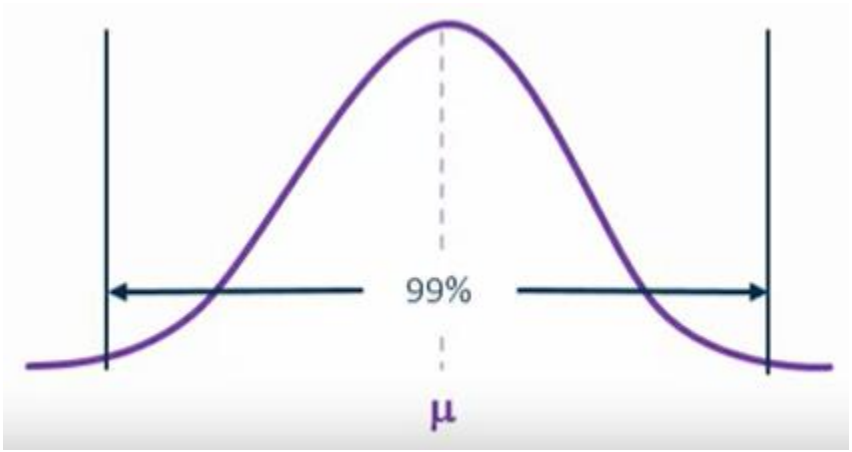
outliers	influential observations
STUDENT	Cook's *D*
	RSTUDENT
	DFFITS
	DFBETAS

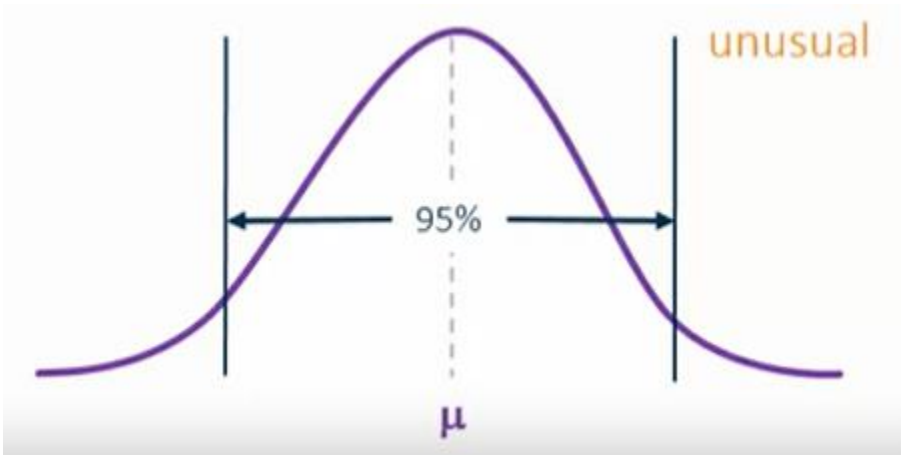
Checking for Outliers with STUDENT Residuals

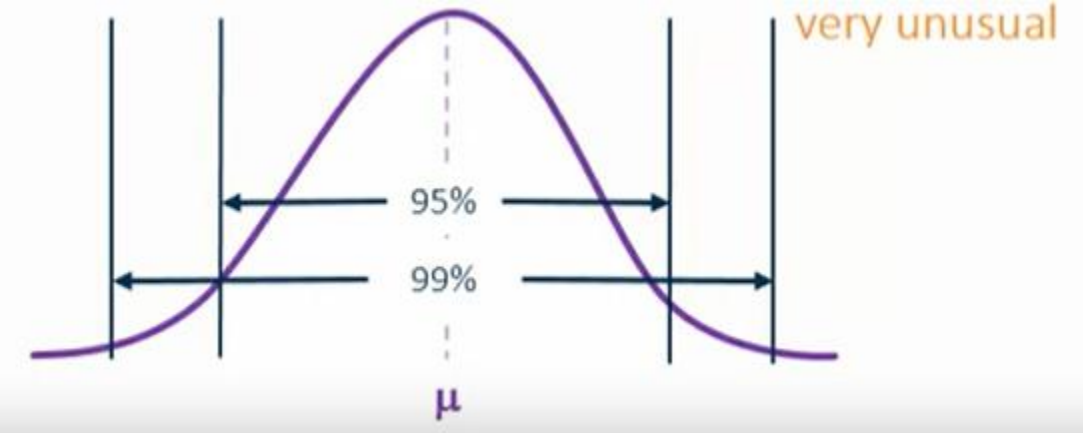





95%
μ


99%
μ


unusual
95%
μ

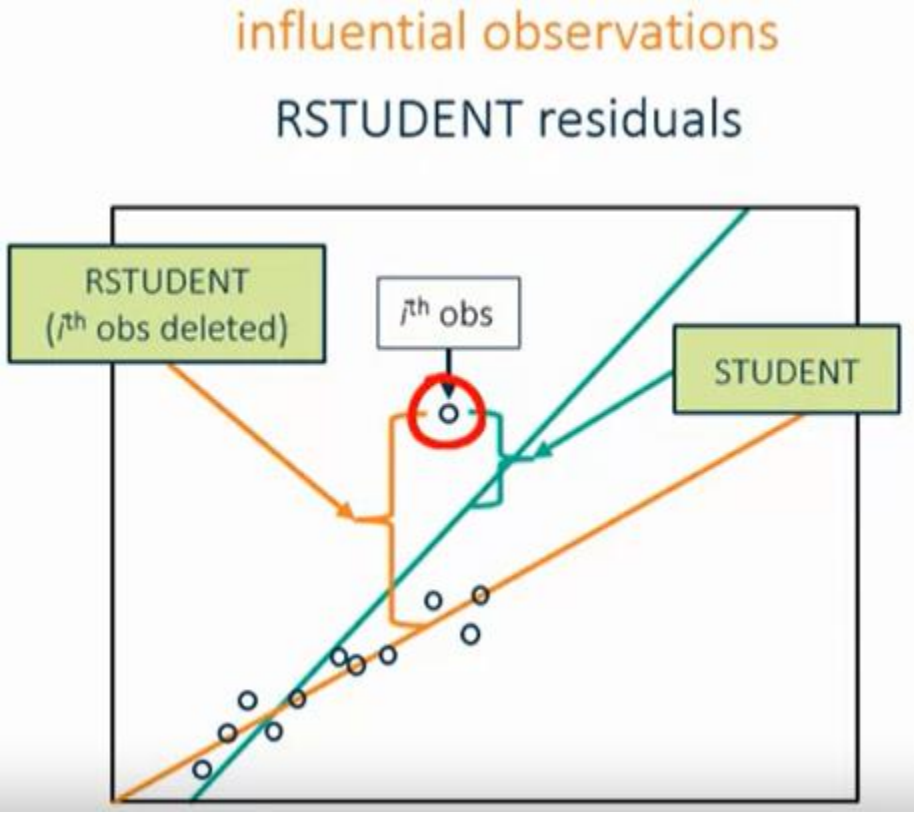


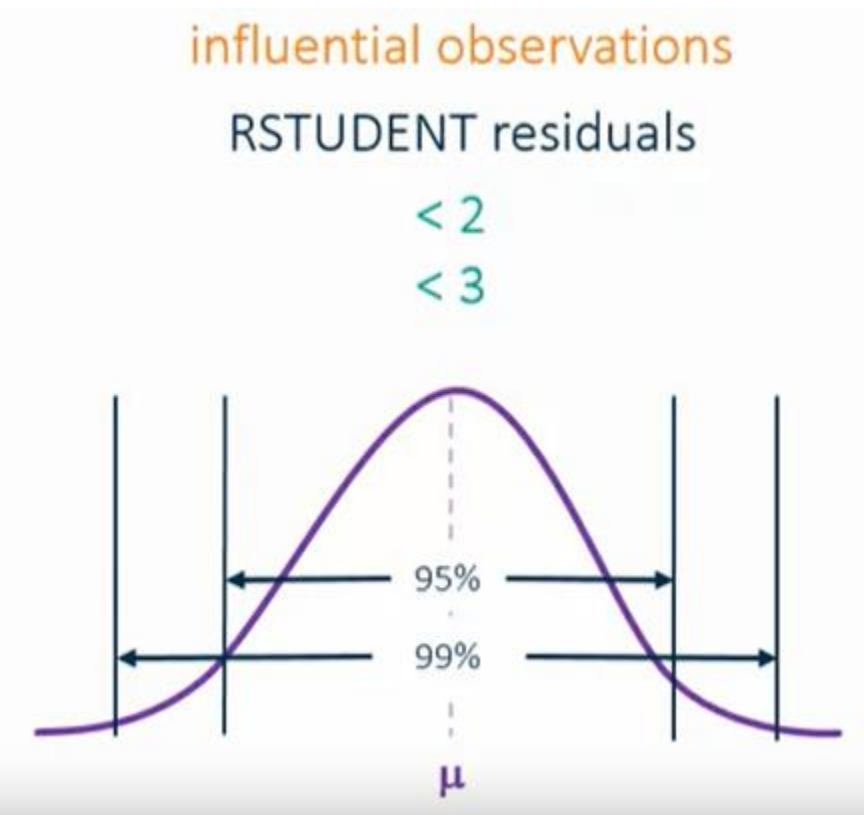
outliers

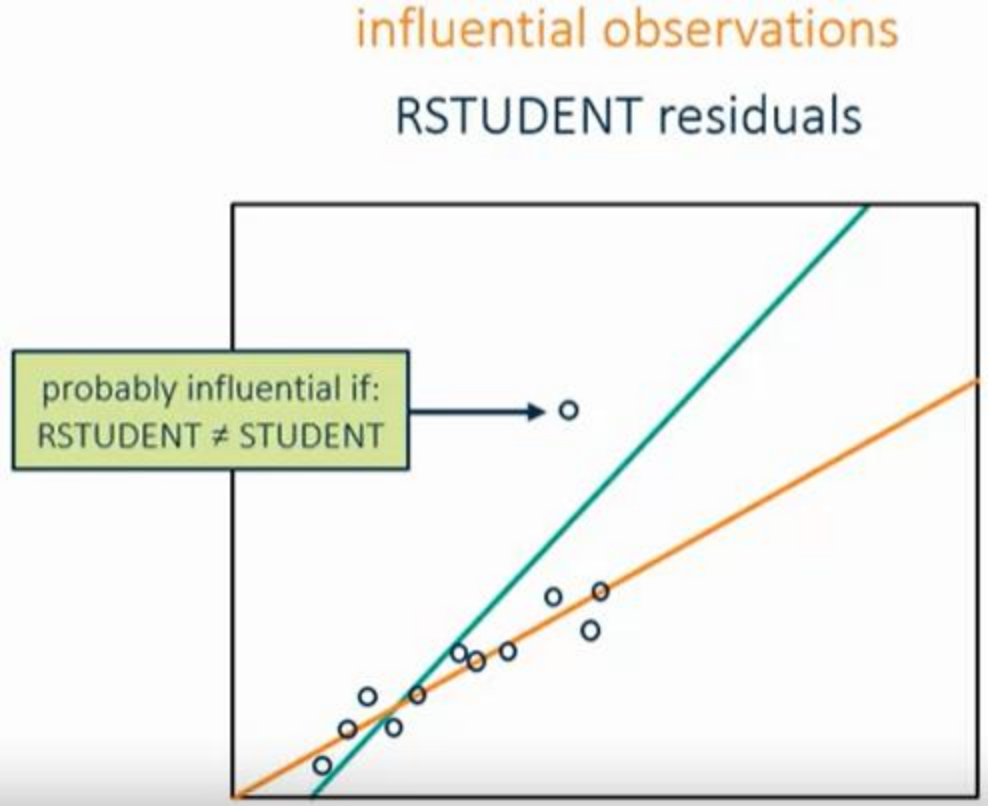
STUDENT residuals

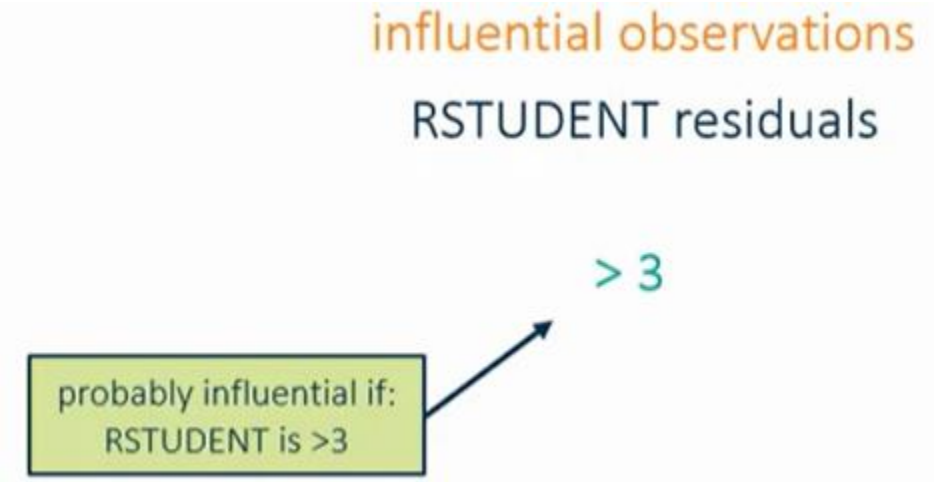
< 2	occur by chance
2 - 3	infrequent
> 3	rare – investigate!

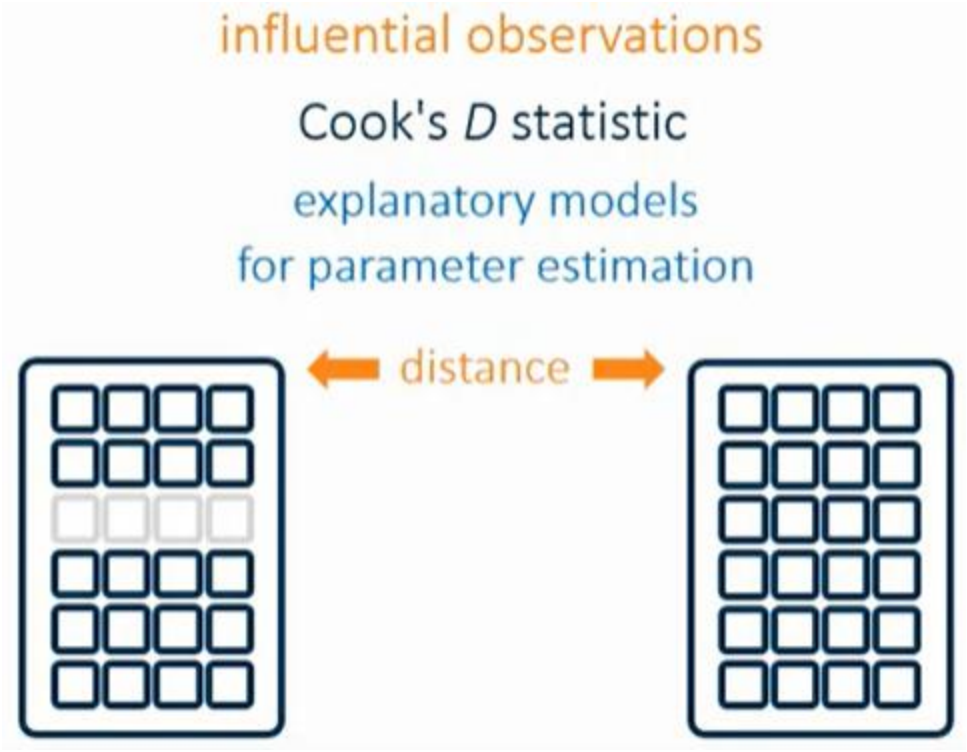
Checking for Influential Observations

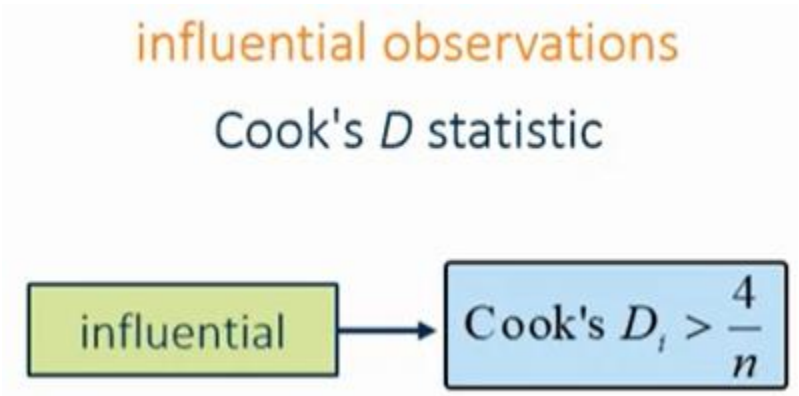


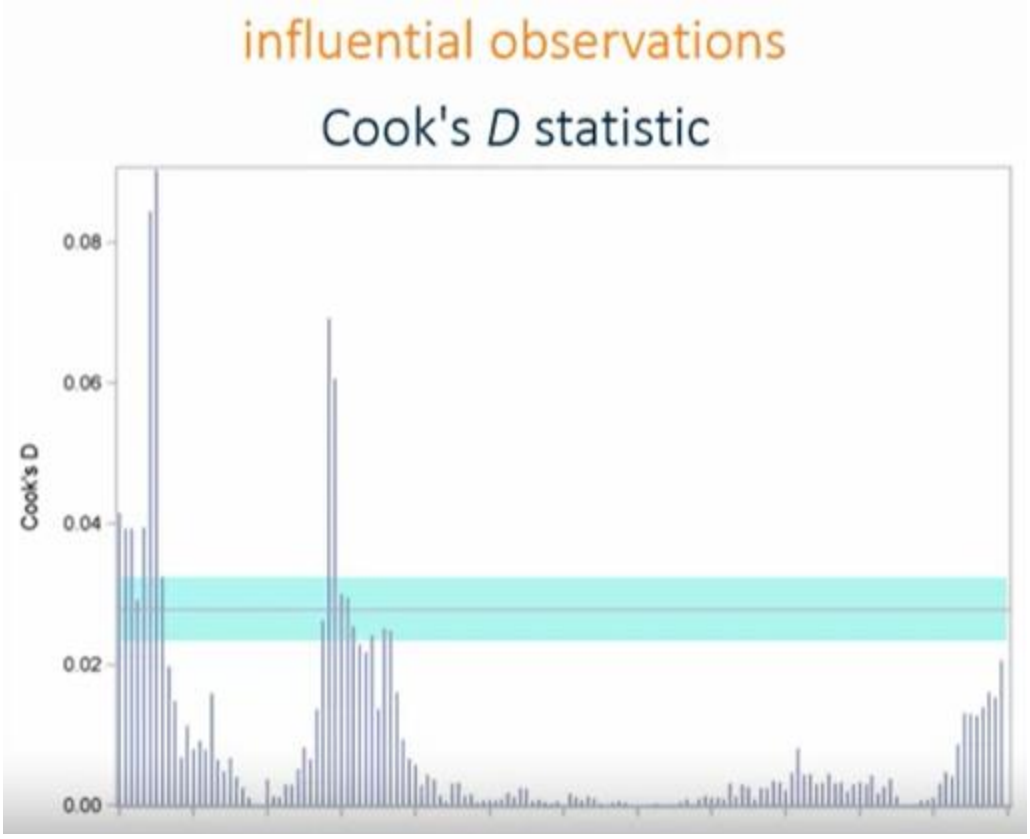


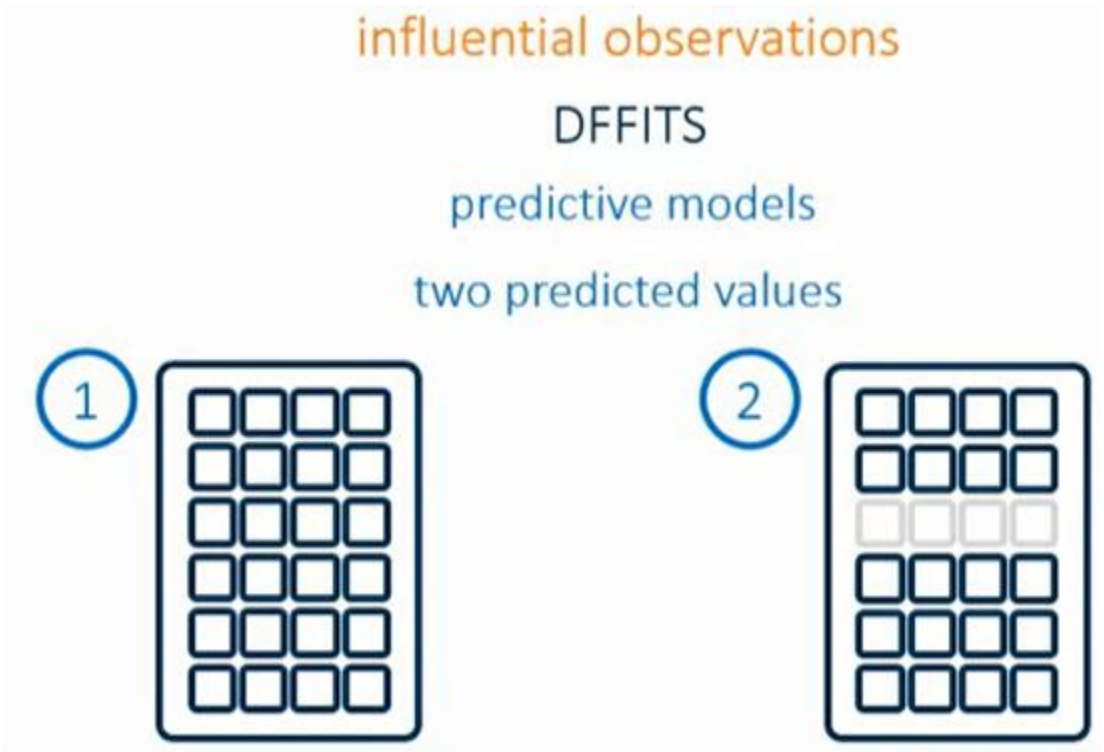

influential observations
RSTUDENT residuals
probably influential if:
RSTUDENT ≠ STUDENT

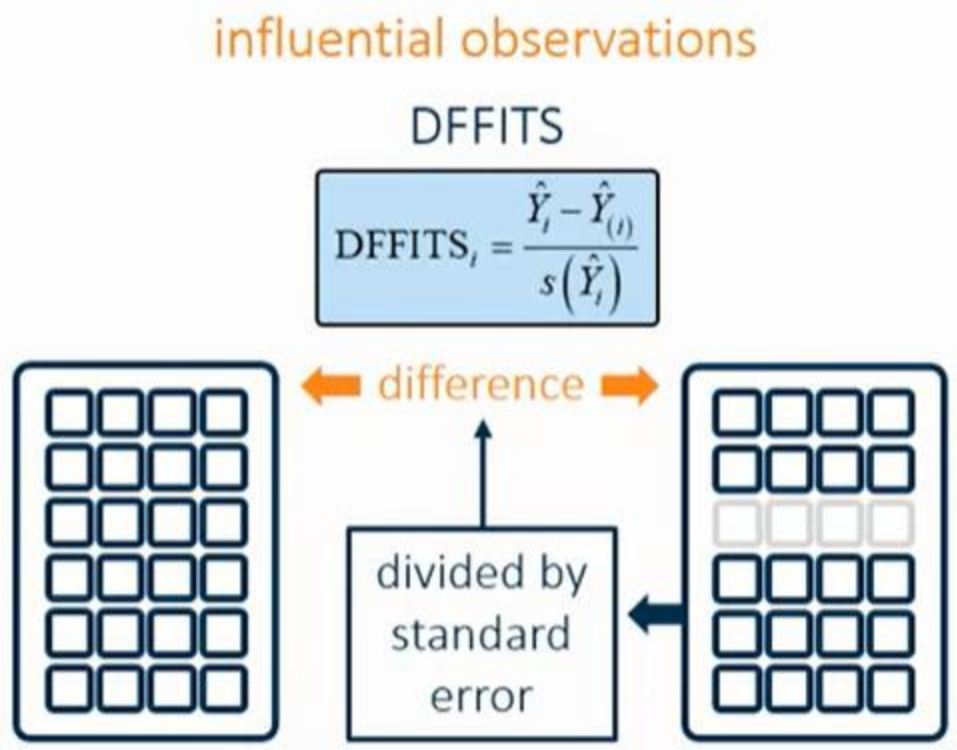

influential observations
RSTUDENT residuals
> 3
probably influential if:
RSTUDENT is >3

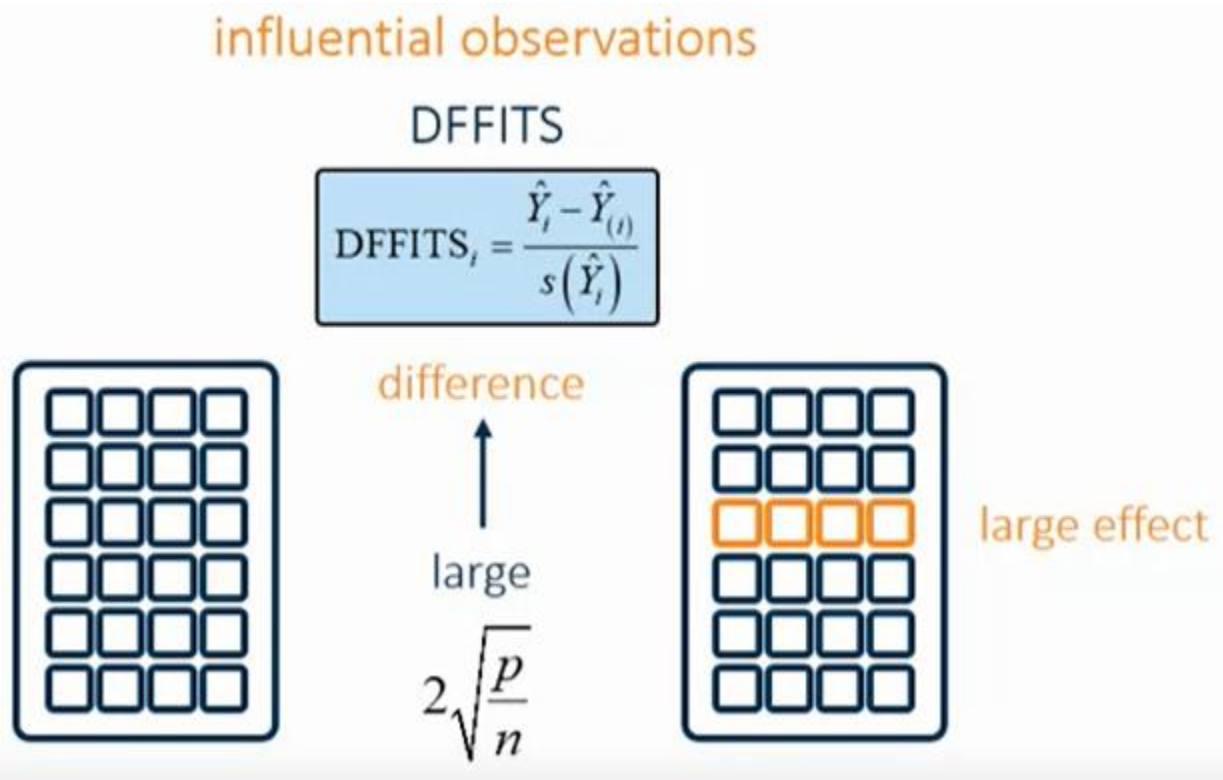

influential observations
Cook's D statistic
explanatory models
for parameter estimation
distance


influential observations
Cook's D statistic
influential
Cook's $D_i > \frac{4}{n}$

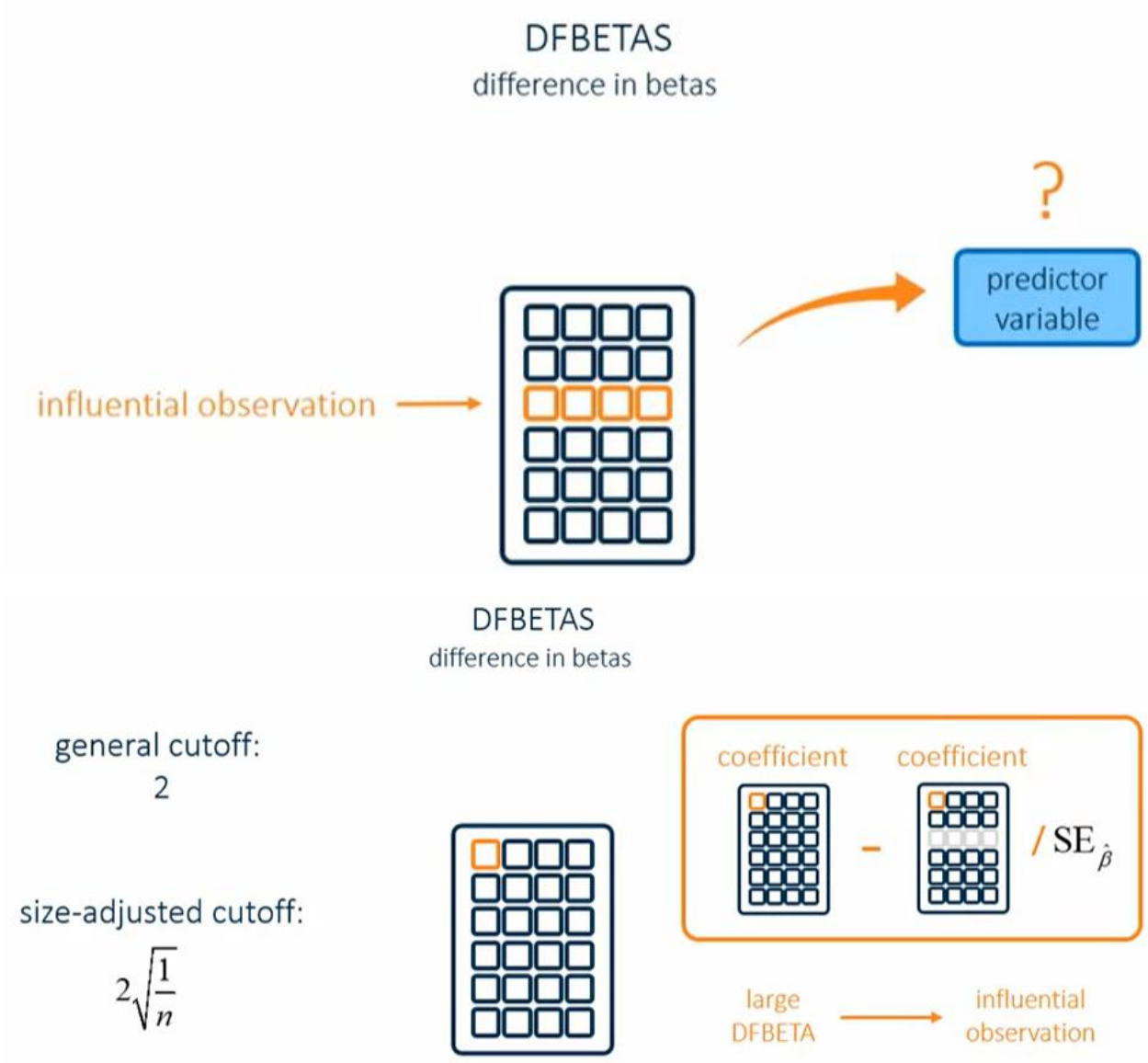

influential observations
Cook's D statistic
Cook's D
0.08
0.06
0.04
0.02
0.00


influential observations
DFFITS
predictive models
two predicted values
1
2

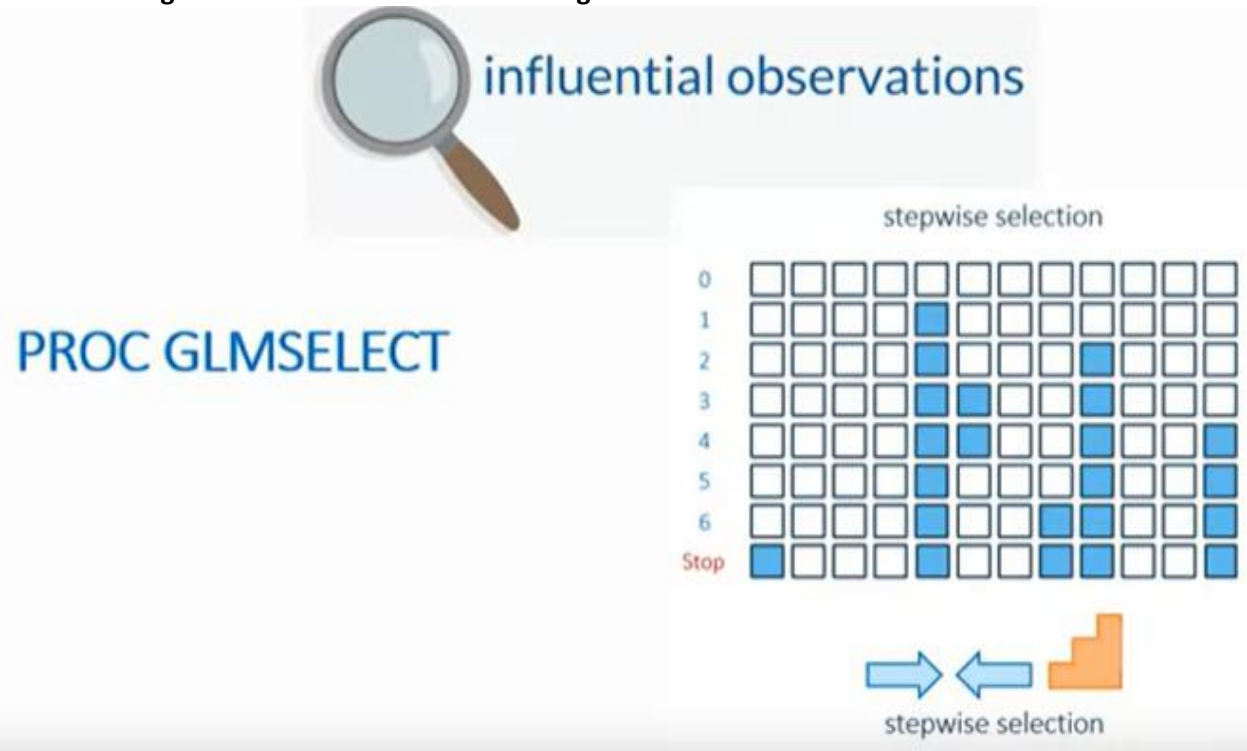

influential observations
DFFITS
$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$
difference
divided by standard error

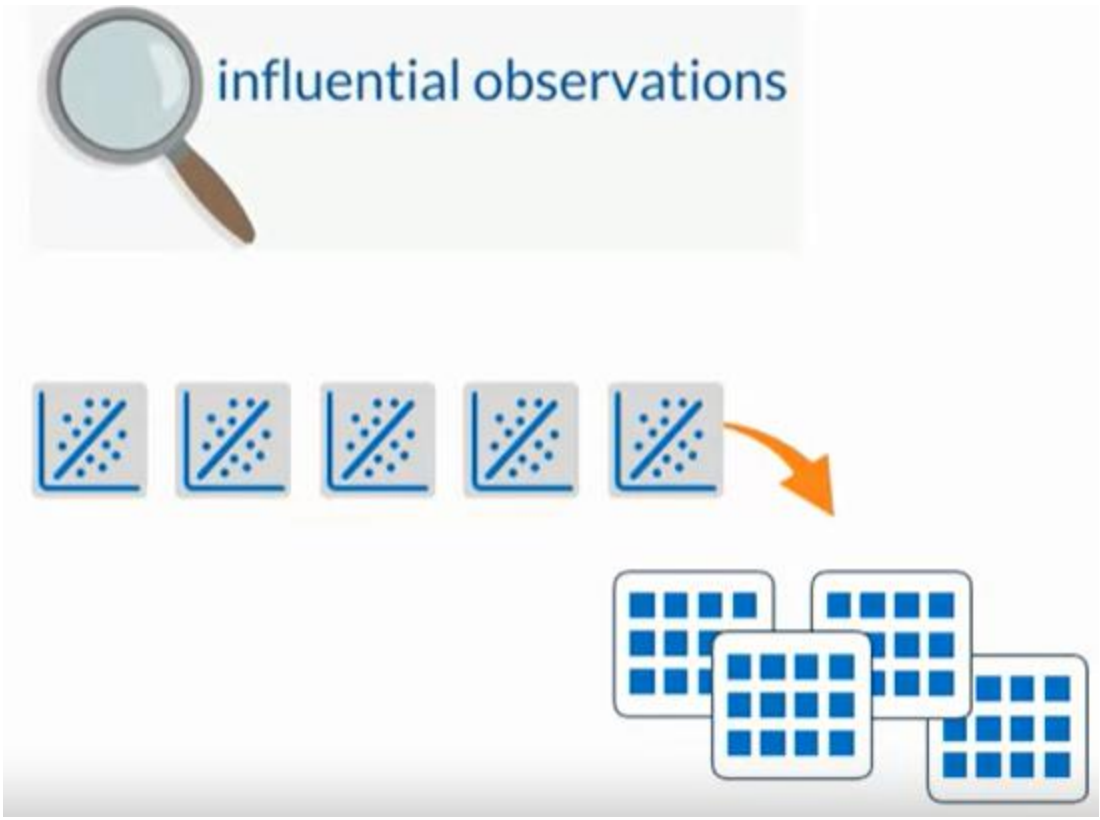

influential observations
DFFITS
$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$
difference
large
$2\sqrt{\frac{p}{n}}$
large effect

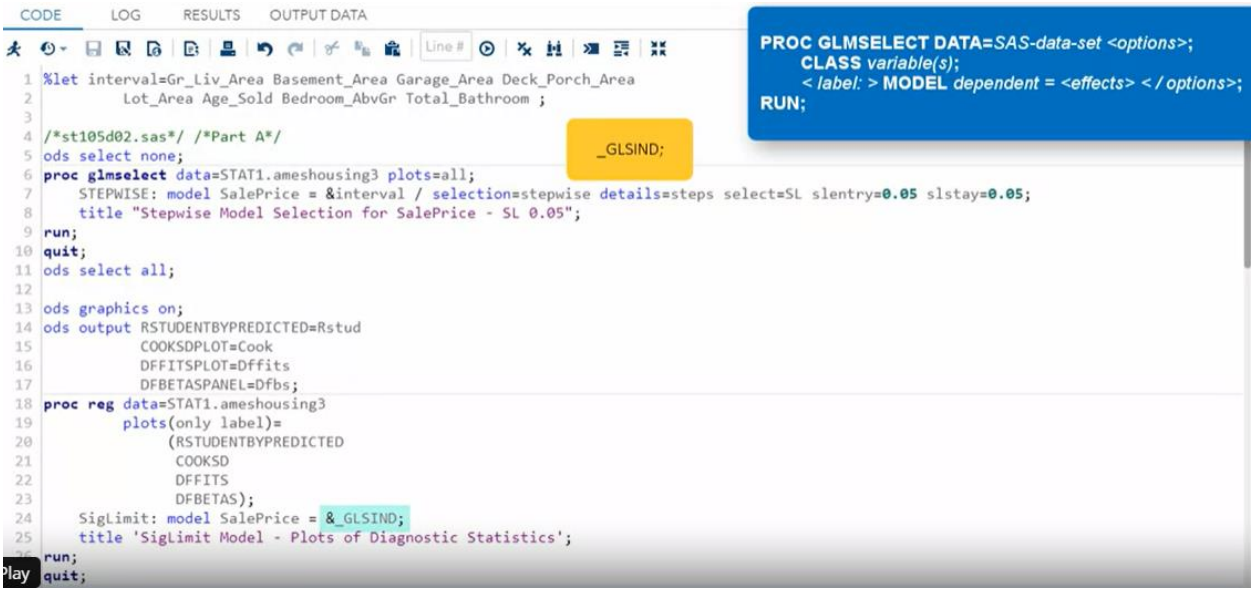
Detecting Influential Observations with DFBETAS



Demo Looking for Influential Observations Using PROC GLMSELECT and PROC REG




influential observations


CODE
LOG
RESULTS
OUTPUT DATA
PROC GLMSELECT DATA=SAS-data-set <options>;
CLASS variable(s);
< label: > MODEL dependent = <effects> < / options>;
RUN;
_GLSIND;
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;
/*st105d02.sas*/ /*Part A*/
ods select none;
proc glmselect data=STAT1.ameshousing3 plots=all;
STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slentry=0.05 slstay=0.05;
title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
quit;
ods select all;
ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
COOKSDPLOT=Cook
DFFITSPLOT=Dffits
DFBETASPANEL=Dfbs;
proc reg data=STAT1.ameshousing3
plots(only label)=
(RSTUDENTBYPREDICTED
COOKSD
DFFITS
DFBETAS);
SigLimit: model SalePrice = &_GLSIND;
title 'SigLimit Model - Plots of Diagnostic Statistics';
run;
quit;

```
63         %let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
64                  Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;
65
66         /*st105d02.sas*/ /*Part A*/
67         ods select none;
68         proc glmselect data=STAT1.ameshousing3 plots=all;
69         STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL
69       ! slentry=0.05 slstay=0.05;
70         title "Stepwise Model Selection for SalePrice - SL 0.05";
71         run;

NOTE: There were 300 observations read from the data set STAT1.AMESHOUSING3.
```

%put &_GLSIND;

SigLimit Model - Plots of Diagnostic Statistics

Model: SigLimit
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

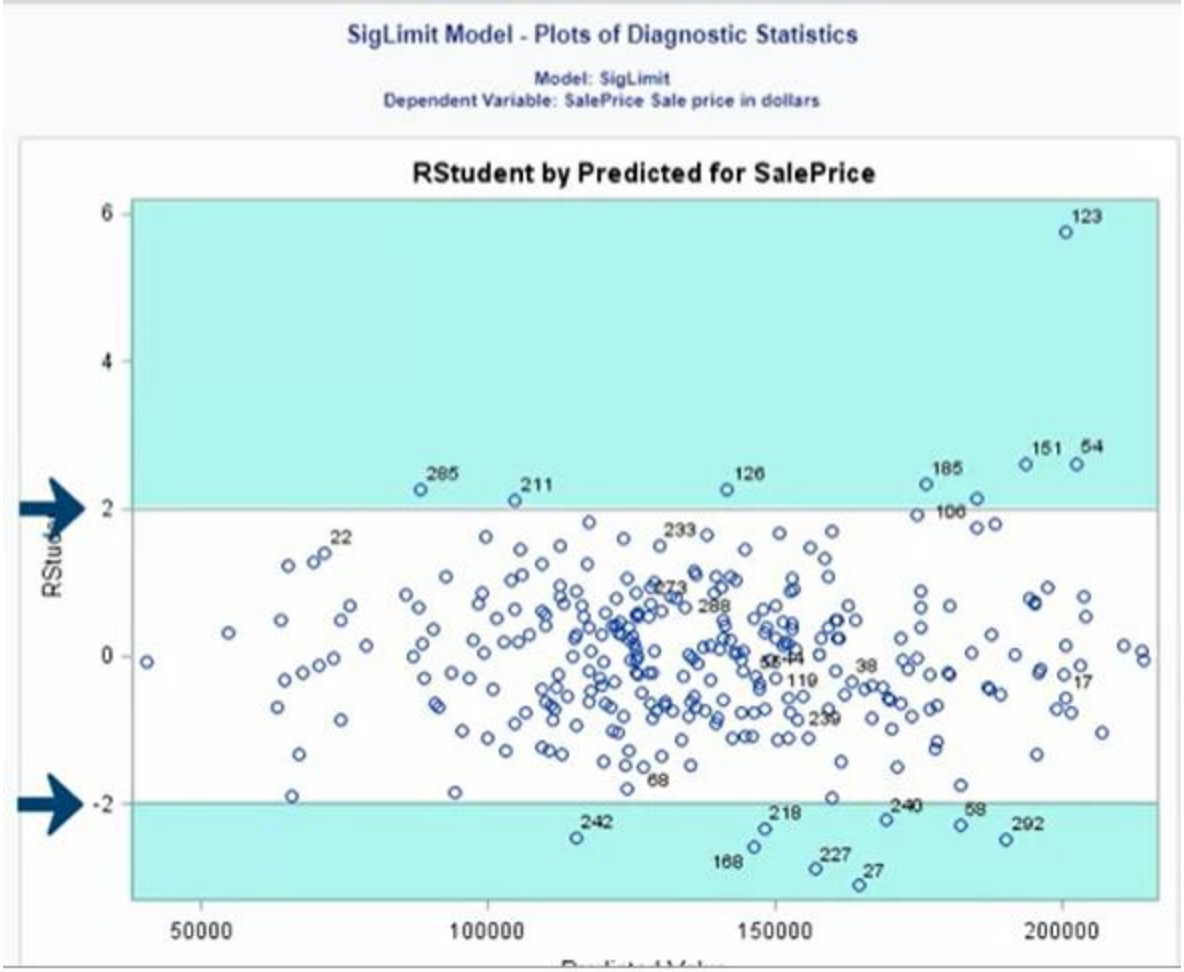
Analysis of Variance

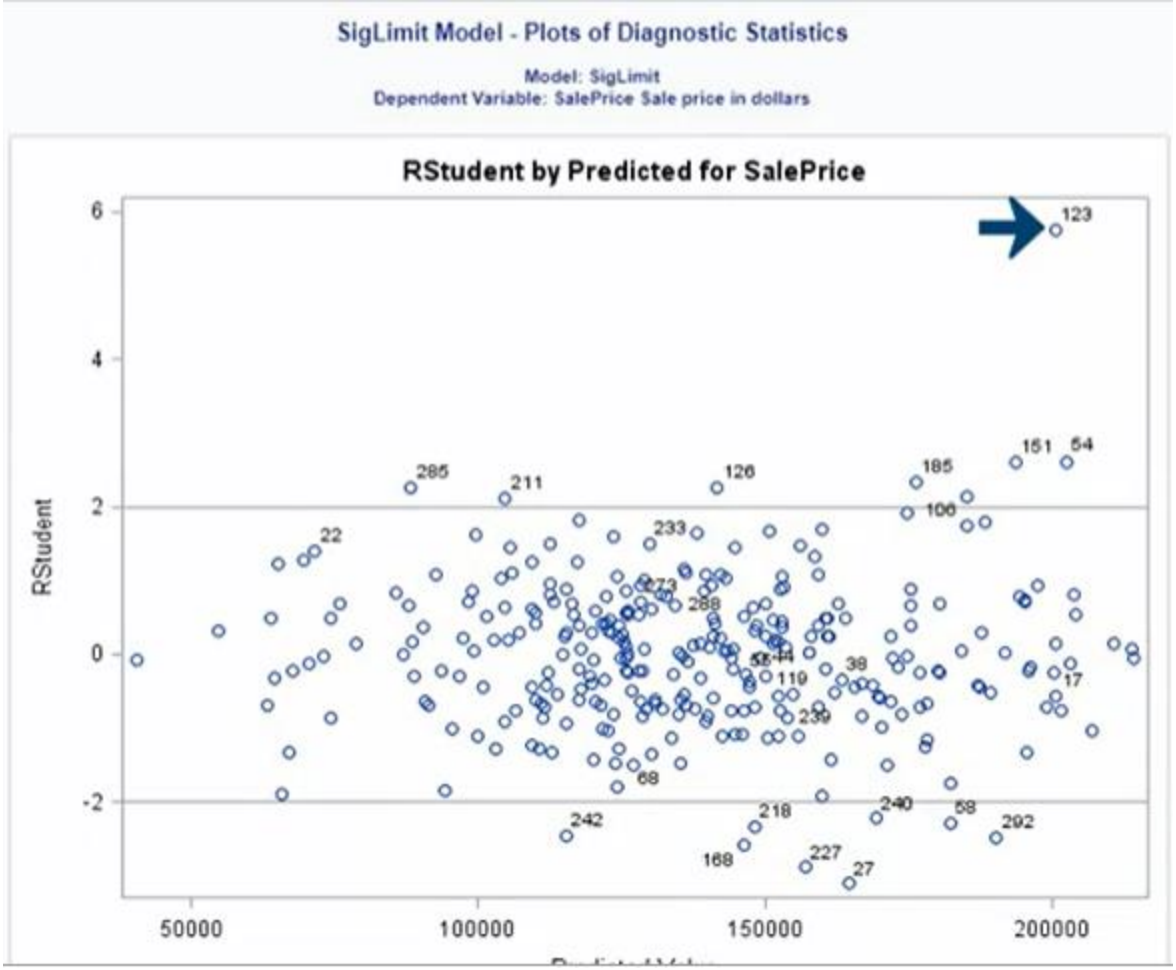
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	3.424508E11	48921543221	176.86	<.0001
Error	292	80772716963	276618894		
Corrected Total	299	4.232235E11			

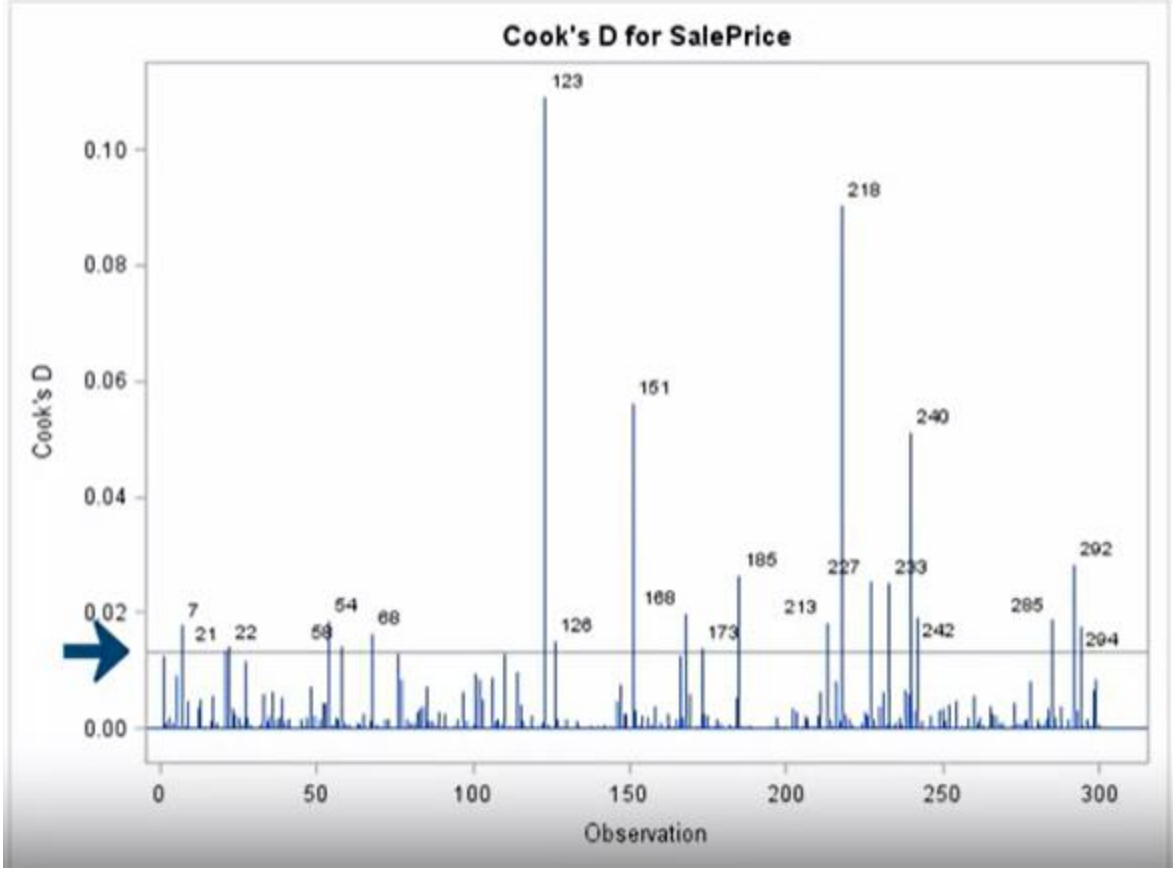
Root MSE	16632	R-Square	0.8091
Dependent Mean	137525	Adj R-Sq	0.8046
Coeff Var	12.09371		

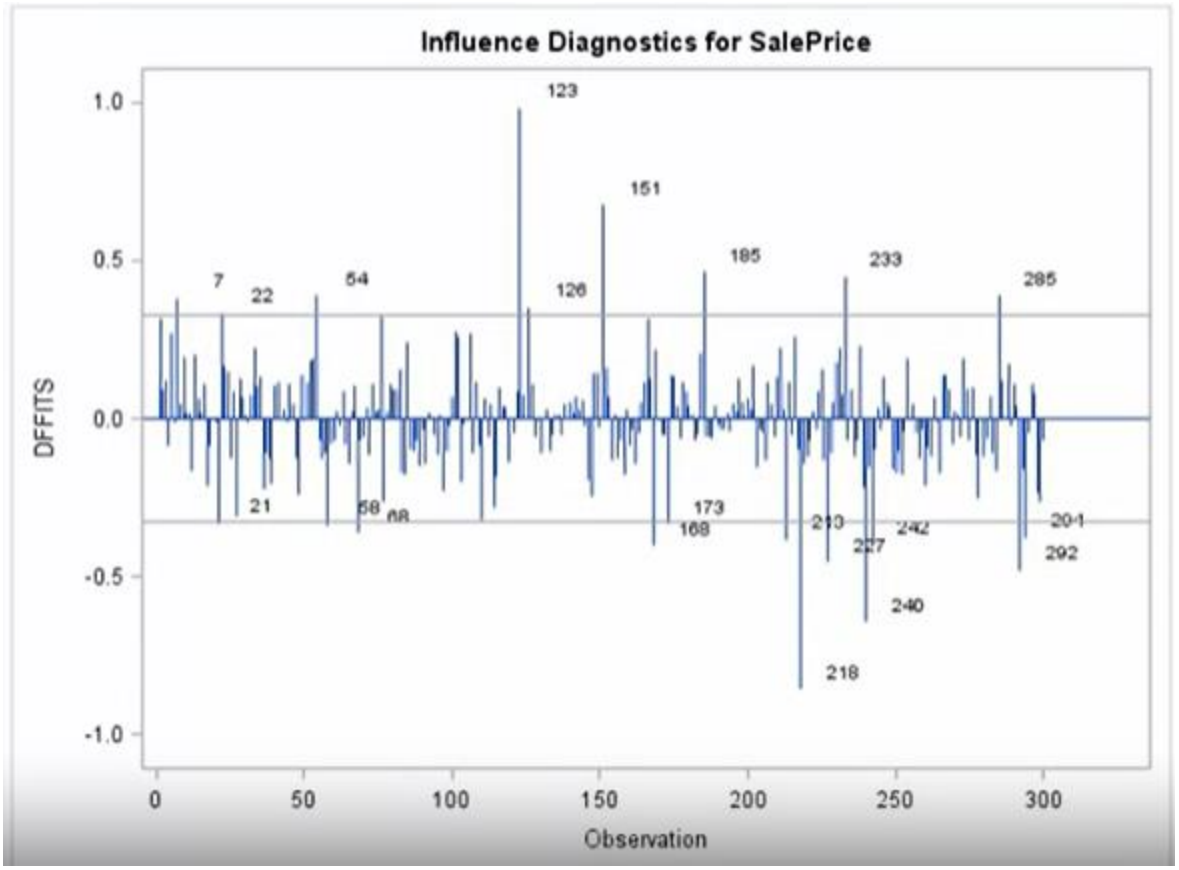
Parameter Estimates

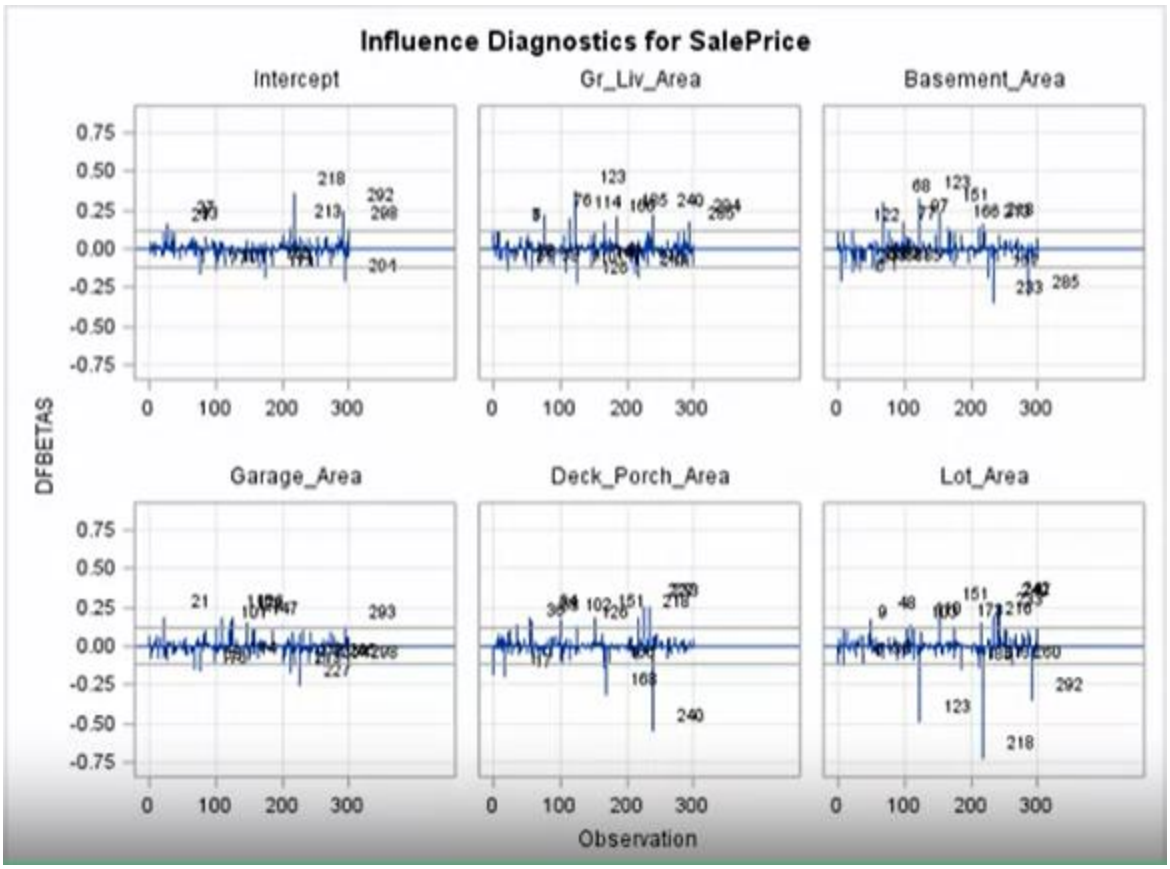
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	47463	5880.67404	8.07	<.0001
Gr_Liv_Area	Above grade (ground) living area square feet	1	65.30372	5.43667	12.01	<.0001
Basement_Area	Basement area in square feet	1	29.84908	3.34540	8.92	<.0001
Garage_Area	Size of garage in square feet	1	36.30961	6.45241	5.63	<.0001
Deck_Porch_Area	Total area of decks and porches in square feet	1	32.05255	7.96766	4.02	<.0001
Lot_Area	Lot size in square feet	1	0.70813	0.31751	2.23	0.0265
Age_Sold	Age of house when sold, in years	1	-447.19868	41.01931	-10.90	<.0001
Bedroom_AbvGr	Bedrooms above grade	1	-5042.76650	1687.92817	-2.99	0.0031

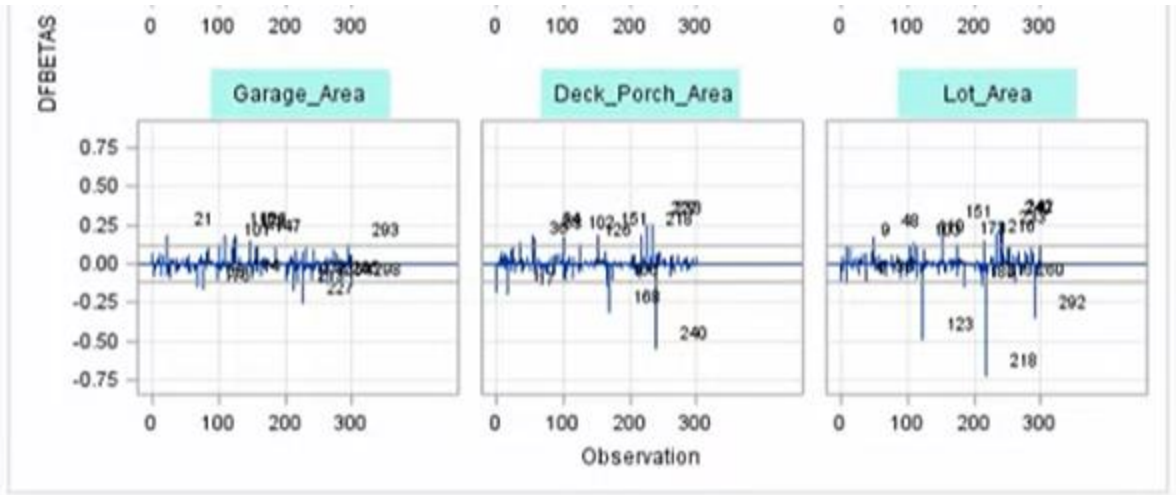

SigLimit Model - Plots of Diagnostic Statistics
Model: SigLimit
Dependent Variable: SalePrice Sale price in dollars
RStudent by Predicted for SalePrice
RStudent
6
4
2
0
-2
50000
100000
150000
200000
123
285
211
126
185
151
54
106
22
233
273
288
38
119
17
239
68
242
218
240
58
292
168
227
27


SigLimit Model - Plots of Diagnostic Statistics
Model: SigLimit
Dependent Variable: SalePrice Sale price in dollars
RStudent by Predicted for SalePrice
RStudent
6
4
2
0
-2
50000
100000
150000
200000
123
285
211
126
185
151
54
106
22
233
273
268
38
119
17
239
68
242
218
240
58
292
168
227
27


Cook's D for SalePrice
Cook's D
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
300
Observation
123
218
151
240
292
185
227
233
168
54
7
213
285
68
126
21
22
58
173
242
294


Influence Diagnostics for SalePrice
DFFITS
Observation
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
250
300
123
151
185
233
285
7
22
54
126
21
58
68
173
168
210
227
242
240
218
204
292


Influence Diagnostics for SalePrice
Intercept
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
DFBETAS
Observation
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
0
100
200
300



%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area

Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d02.sas*/ /*Part A*/

ods select none;

proc glmselect data=STAT1.ameshousing3 plots=all;

STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slentry=0.05 slstay=0.05;

title "Stepwise Model Selection for SalePrice - SL 0.05";

run;

quit;

ods select all;

ods graphics on;

```
ods output RSTUDENTBYPREDICTED=Rstud
          COOKSDPLOT=Cook
          DFFITSPLOT=Dffits
          DFBETASPANEL=Dfbs;
proc reg data=STAT1.ameshousing3
        plots(only label)=
            (RSTUDENTBYPREDICTED
             COOKSD
             DFFITS
             DFBETAS);
   SigLimit: model SalePrice = &_GLSIND;
   title 'SigLimit Model - Plots of Diagnostic Statistics';
run;
quit;
```

Demo Examining the Influential Observations Using PROC PRINT

exceed cutoffs

```
/*st105d02.sas*/  /*Part B*/
title;
proc print data=Rstud;
run;

proc print data=Cook;
run;

proc print data=Dffits;
run;

proc print data=Dfbs;
run;
```

Obs	Model	Dependent	RStudent	PredictedValue	outLevLabel	Observation
1	SigLimit	SalePrice	1.73092	185283.46	.	1
2	SigLimit	SalePrice	0.67964	180284.34	.	2
3	SigLimit	SalePrice	0.63948	104541.46	.	3
4	SigLimit	SalePrice	-0.58261	169597.56	.	4
5	SigLimit	SalePrice	1.32153	158490.53	.	5
6	SigLimit	SalePrice	-0.05738	125932.40	.	6
7	SigLimit	SalePrice	1.92473	174736.56	.	7
8	SigLimit	SalePrice	0.36232	153008.35	.	8
9	SigLimit	SalePrice	1.47568	156222.61	.	9
10	SigLimit	SalePrice	0.10202	140446.83	.	10
11	SigLimit	SalePrice	0.15571	125421.52	.	11
12	SigLimit	SalePrice	-1.12570	150512.33	.	12
13	SigLimit	SalePrice	1.65684	150725.00	.	13
14	SigLimit	SalePrice	0.54343	126013.31	.	14
15	SigLimit	SalePrice	0.15339	151460.95	.	15
16	SigLimit	SalePrice	0.55394	125741.85	.	16
17	SigLimit	SalePrice	-0.56164	200736.13	17	17
18	SigLimit	SalePrice	-0.64482	90640.01	.	18
19	SigLimit	SalePrice	-0.06749	120117.08	.	19
20	SigLimit	SalePrice	-0.05783	124845.09	.	20
21	SigLimit	SalePrice	-1.47822	123941.30	.	21
22	SigLimit	SalePrice	1.40038	71388.82	22	22
23	SigLimit	SalePrice	1.24923	117326.83	.	23

Obs	Model	Dependent	Observation	DFFITS	DFFITSOUT
1	SigLimit	SalePrice	1	0.31861	.
2	SigLimit	SalePrice	2	0.09029	.
3	SigLimit	SalePrice	3	0.12177	.
4	SigLimit	SalePrice	4	-0.08573	.
5	SigLimit	SalePrice	5	0.26928	.
6	SigLimit	SalePrice	6	-0.01301	.
7	SigLimit	SalePrice	7	.	0.37928
8	SigLimit	SalePrice	8	0.04386	.
9	SigLimit	SalePrice	9	0.19752	.
10	SigLimit	SalePrice	10	0.01636	.
11	SigLimit	SalePrice	11	0.01720	.
12	SigLimit	SalePrice	12	-0.16660	.
13	SigLimit	SalePrice	13	0.20071	.
14	SigLimit	SalePrice	14	0.06417	.
15	SigLimit	SalePrice	15	0.01742	.
16	SigLimit	SalePrice	16	0.10651	.
17	SigLimit	SalePrice	17	-0.21458	.
18	SigLimit	SalePrice	18	-0.08653	.
19	SigLimit	SalePrice	19	-0.00777	.
20	SigLimit	SalePrice	20	-0.01146	.
21	SigLimit	SalePrice	21	.	-0.33145

Intercept

▸ Table of Contents

_DFBETASOUT1	_DFBETAS2	_DFBETASOUT2	_DFBETAS3	_DFBETASOUT3	_DFBETAS4	_DFBETASOUT4	_DFBETAS5	_DFBETASOUT5	_DFBETAS6	_DFBETASOUT6
.	0.10763	.	.	0.11744	0.07141	.	.	-0.18180	.	-0.12067
.	0.02128	.	0.03237	.	0.00993	.	0.00943	.	-0.04517	.
.	-0.04403	.	0.01625	.	-0.08221	.	-0.01416	.	0.03778	.
.	0.00052	.	-0.01769	.	0.00315	.	-0.01452	.	0.05082	.
.	.	0.12008	.	-0.21199	-0.03160	.	0.04749	.	0.01234	.
.	0.00672	.	0.00198	.	0.00249	.	-0.01051	.	-0.00022	.
.	.	0.11635	0.10354	.	0.01939	.	0.08053	.	.	-0.13126
.	-0.00500	.	0.02086	.	0.01649	.	-0.02605	.	0.00021	.
.	-0.07039	.	0.02326	.	0.03276	.	0.06244	.	.	0.12030
.	-0.00134	.	0.00327	.	-0.00608	.	0.01138	.	0.00328	.

```
48 data Dfbs01;
49     set Dfbs (obs=300);
50 run;
51
52 data Dfbs02;
53     set Dfbs (firstobs=301);
54 run;
55
56 data Dfbs2;
57     update Dfbs01 Dfbs02;
58     by Observation;
59 run;
60
```


```
62 data influential;
63 /*  Merge datasets from above.*/
64     merge Rstud
65           Cook
66           Dffits
67           Dfbs2;
68     by observation;
69
70 /*  Flag observations that have exceeded at least one cutpoint;*/
71     if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
72     array dfbetas{*} _dfbetasout: ;
73     do i=2 to dim(dfbetas);
74          if dfbetas{i} then flag=1;
75     end;
```

```
77 /*  Set to missing values of influence statistics for those*/
78 /*  that have not exceeded cutpoints;*/
79     if ABS(Rstudent)<=3 then RStudent=.;
80     if Cooksdlabel eq ' ' then CooksD=.;
81
82 /*  Subset only observations that have been flagged.*/
83     if flag=1;
84     drop i flag;
85 run;
```

```
87 title;
88 proc print data=influential;
89     id observation;
90     var Rstudent CooksD Dffitsout _dfbetasout:;
91 run;
```

Observation	RStudent	CooksD	DFFITSOUT	_DFBETASOUT1	_DFBETASOUT2	_DFBETASOUT3	_DFBETASOUT4	_DFBETASOUT5	_DFBETASOUT6	_DFBETASOUT7	_DFBETASOUT8
1	.	.	.								
5	.	.	.	.	0.12008	-0.21199	.	.	.	-0.15614	.
7	.	0.01782	0.37928	.	0.11635	.	.	.	-0.13126	.	-0.25391
9	.	.	.	.	.	.	.	.	0.12030	.	.
17	.	.	.	.	.	.	.	-0.19555	.	.	.
21	.	0.01368	-0.33145	0.12113	-0.13573	-0.14595	0.18720	.	.	-0.13114	.
22	.	0.01406	0.33593	.	.	0.12409	.	.	.	0.25385	.
23	.	.	.	.	-0.11834	.	.	.	.	.	.
27	-3.10785	.	.	0.16637	.	-0.12996	.	.	.	.	.
33	.	.	.	0.13368	.	-0.13475	.	.	.	.	.

/*st105d02.sas*/ /*Part B*/

title;

proc print data=Rstud;

run;

proc print data=Cook;

run;

proc print data=Dffits;

run;

proc print data=Dfbs;

run;

data Dfbs01;

set Dfbs (obs=300);

run;

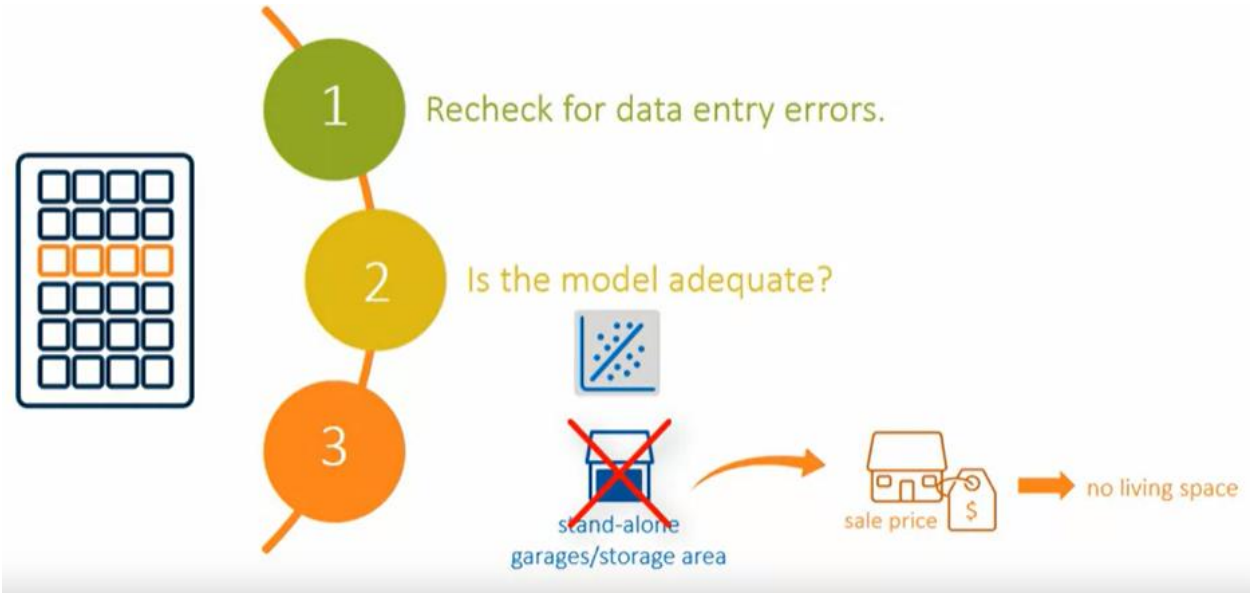
data Dfbs02;

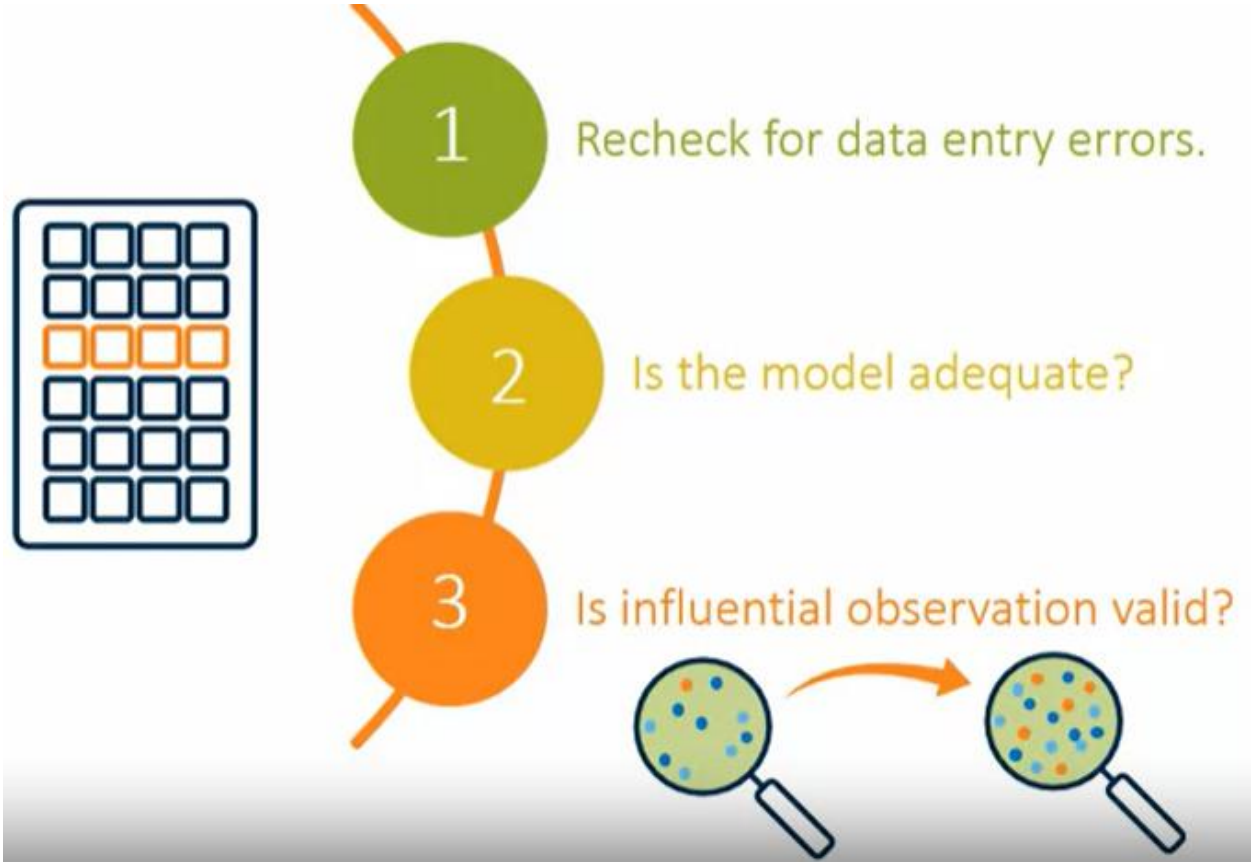
set Dfbs (firstobs=301);

```
run;

data Dfbs2;
        update Dfbs01 Dfbs02;
        by Observation;
run;



data influential;
/*  Merge datasets from above.*/
   merge Rstud
       Cook
       Dffits
                   Dfbs2;
   by observation;

/*  Flag observations that have exceeded at least one cutpoint;*/
   if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
   array dfbetas{*} _dfbetasout: ;
   do i=2 to dim(dfbetas);
      if dfbetas{i} then flag=1;
   end;

/*  Set to missing values of influence statistics for those*/
/*  that have not exceeded cutpoints;*/
   if ABS(Rstudent)<=3 then RStudent=.;
   if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
```

```
   if flag=1;
   drop i flag;
run;

title;
proc print data=influential;
   id observation;
   var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

Handling Influential Observations

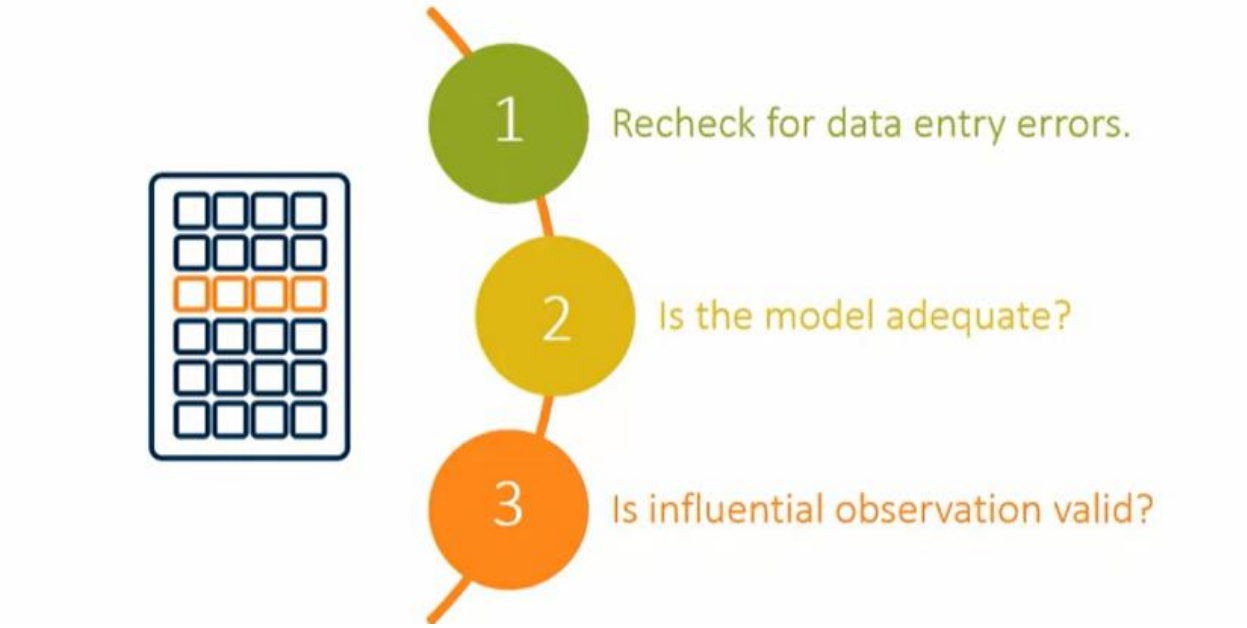




you should not exclude data, but if you do:

description of observations

limitation of conclusions



```
/*st105s02.sas*/  /*Part A*/
ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
        COOKSDPLOT=Cook
        DFFITSPLOT=Dffits
        DFBETASPANEL=Dfbs;
proc reg data=STAT1.BodyFat2
       plots(only label)=
          (RSTUDENTBYPREDICTED
           COOKSD
           DFFITS
           DFBETAS);
   FORWARD: model PctBodyFat2
            = Abdomen Weight Wrist Forearm;
   id Case;
   title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

FORWARD Model - Plots of Diagnostic Statistics

The REG Procedure
Model: FORWARD
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

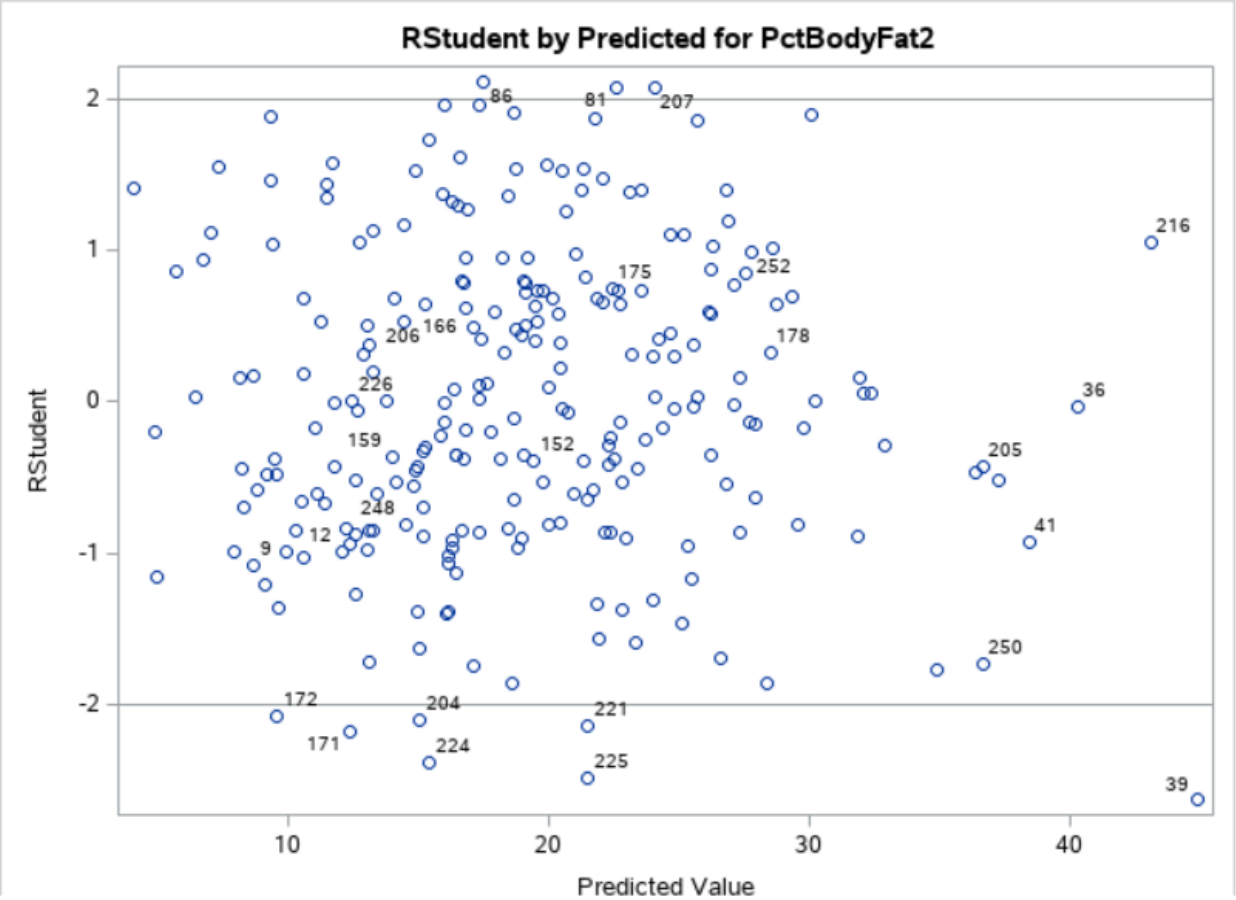
Root MSE	4.34272	R-Square	0.7350
Dependent Mean	19.15079	Adj R-Sq	0.7307
Coeff Var	22.67647		

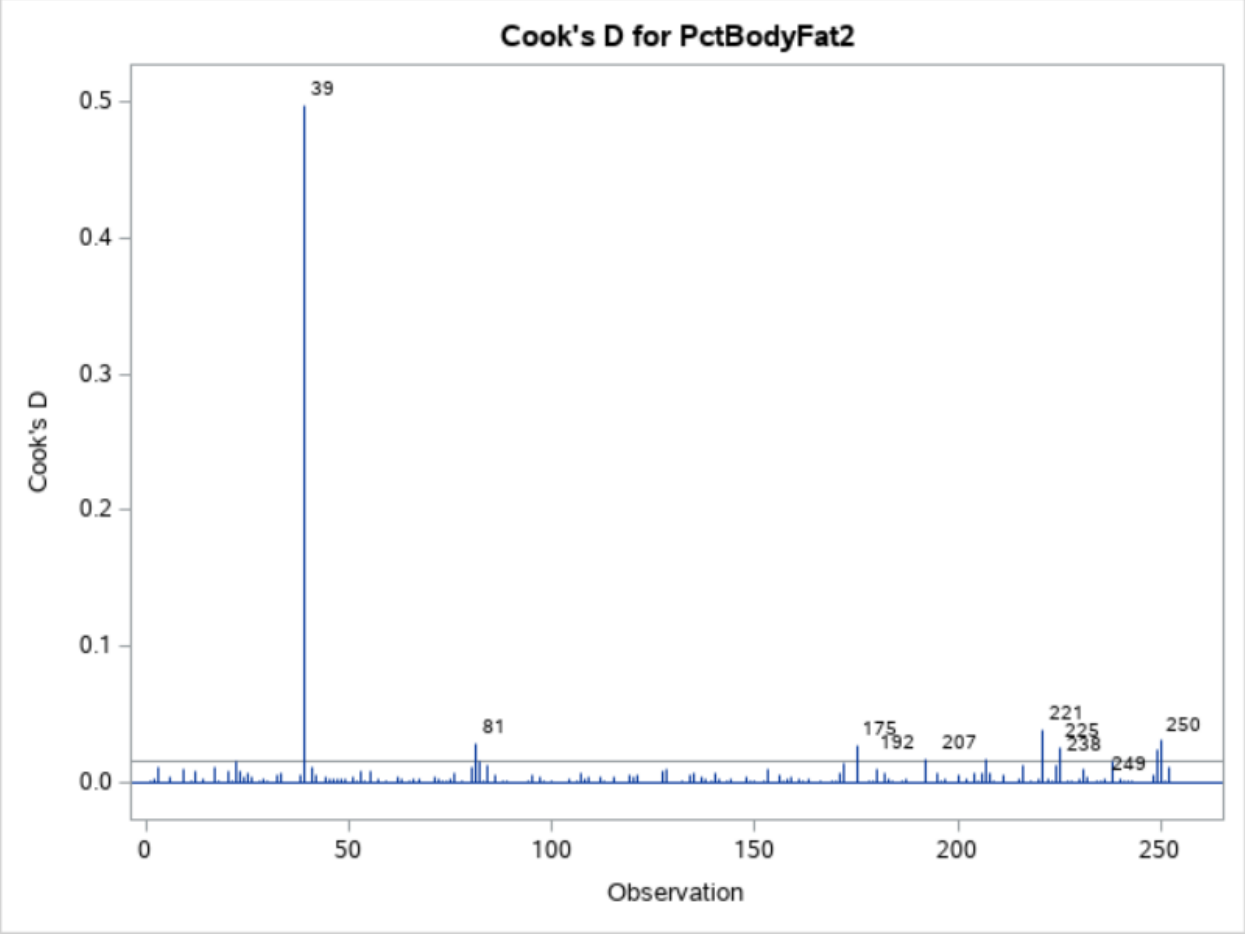
Parameter Estimates

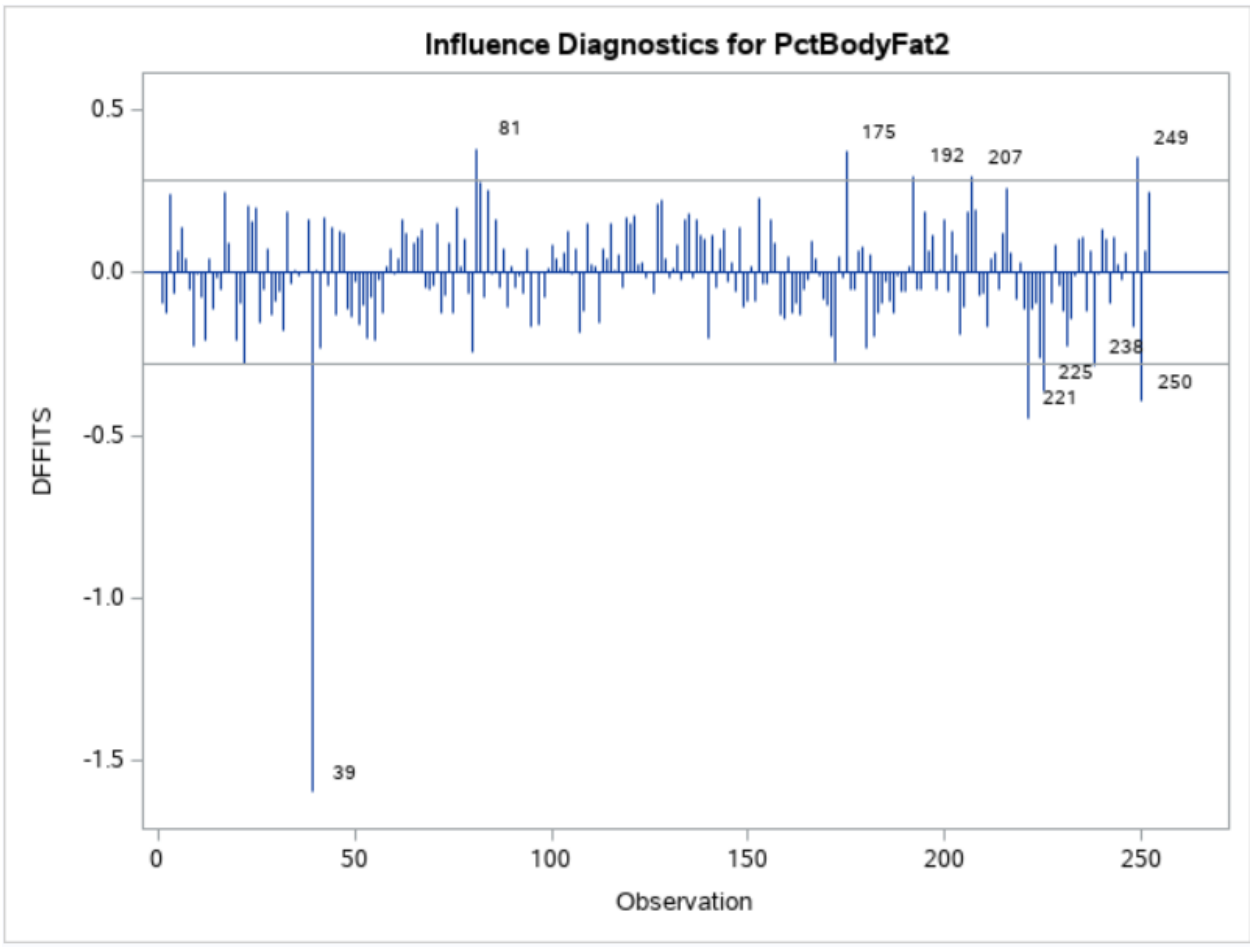
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-34.85407	7.24500	-4.81	<.0001
Abdomen	1	0.99575	0.05607	17.76	<.0001
Weight	1	-0.13563	0.02475	-5.48	<.0001
Wrist	1	-1.50556	0.44267	-3.40	0.0008
Forearm	1	0.47293	0.18166	2.60	0.0098

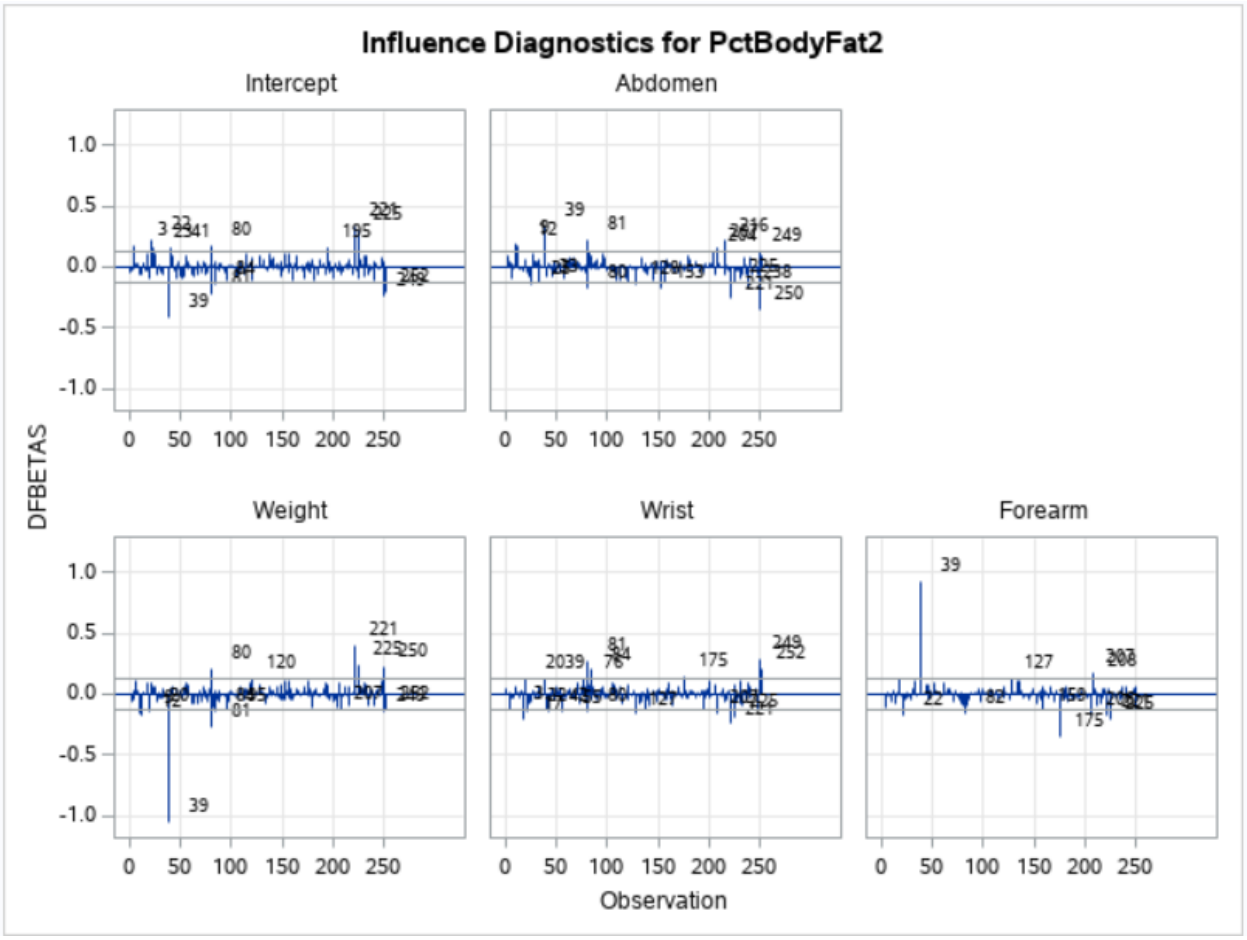
FORWARD Model - Plots of Diagnostic Statistics

The REG Procedure
Model: FORWARD
Dependent Variable: PctBodyFat2




Cook's D for PctBodyFat2
Cook's D
0.5
0.4
0.3
0.2
0.1
0.0
39
81
175
192
207
221
225
238
249
250
0
50
100
150
200
250
Observation


Influence Diagnostics for PctBodyFat2
DFFITS
0.5
0.0
-0.5
-1.0
-1.5
81
175
192
207
249
238
225
221
250
39
0
50
100
150
200
250
Observation




```
/*st105s02.sas*/  /*Part B*/
data influential;
/*  Merge datasets from above.*/
   merge Rstud
         Cook
         Dffits
                  Dfbs;
   by observation;

/*  Flag observations that have exceeded at least one cutpoint;*/
   if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
   array dfbetas{*} _dfbetasout: ;
   do i=2 to dim(dfbetas);
```

```
    if dfbetas{i} then flag=1;
  end;

/*  Set to missing values of influence statistics for those*/
/*  who have not exceeded cutpoints;*/
  if ABS(Rstudent)<=3 then RStudent=.;
  if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
  if flag=1;
  drop i flag;
run;

proc print data=influential;
  id observation ID1;
  var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

Observation	id1	RStudent	CooksD	DFFITSOUT	_DFBETASOUT1	_DFBETASOUT2	_DFBETASOUT3	_DFBETASOUT4	_DFBETASOUT5
3	3	.	.	.	0.17943	.	.	-0.12815	.
9	9	.	.	.	.	0.18911	-0.15600	.	.
12	12	.	.	.	.	0.18169	-0.18076	.	.
17	17	.	.	.	.	.	.	-0.20902	.
20	20	.	.	.	.	.	-0.13786	0.13273	.
22	22	.	.	.	0.22887	.	.	-0.14080	-0.16797
25	25	.	.	.	.	-0.14080	.	.	.
33	33	.	.	.	.	-0.12765	.	.	.
39	39	.	0.49632	-1.59408	-0.41792	0.33576	-1.05761	0.13217	0.93125
42	42	.	.	.	.	.	.	-0.13688	.
55	55	.	.	.	.	.	.	-0.14907	.
76	76	.	.	.	.	.	.	0.13108	.
80	80	.	.	.	0.17122	-0.17507	0.20391	-0.14744	.

Practice: Using PROC REG to Generate Potential Outliers

Question 1

Generate statistics for potential outliers in the **stat1.bodyfat2** data set. Write this data to an output data set, and print your results.

1. Use PROC REG to run a regression model of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**. Create plots to identify potential influential observations that are based on the suggested cutoff values.
2. Submit the code.

What do you notice in the results?

In the RStudent by Predicted for PctBodyFat2 scatter plot, only a modest number of observations are further than two standard error units from the mean of 0.

- In the Cook's D for PctBodyFat2 plot, there are 10 labeled outliers, but observation 39 is clearly the most extreme.
- In the Influence Diagnostics for PctBodyFat2 plot, the same observations are shown to be influential by the DFFITS statistic.
- In the panel plot, DFBETAS are particularly high for observation 39 on the parameters for **Weight** and **Forearm** circumference.

```
/*st105s02.sas*/  /*Part A*/

ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=STAT1.BodyFat2
         plots(only label)=
              (RSTUDENTBYPREDICTED
               COOKSD
               DFFITS
               DFBETAS);
   FORWARD: model PctBodyFat2
                  = Abdomen Weight Wrist Forearm;
   id Case;
   title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

Question 2

1. Write the residuals output to a data set named **influential**, subset the data to select only observations that are potentially influential outliers, and print your results.
2. Submit the code and view the results.

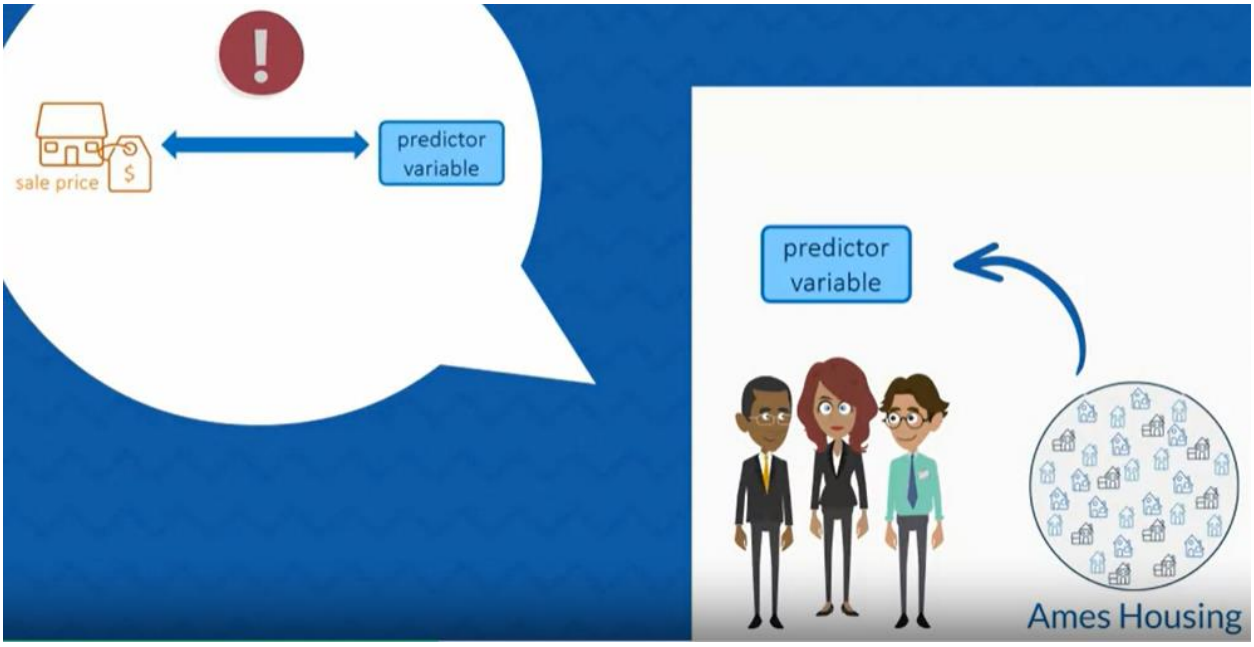
```
/*st105s02.sas*/  /*Part B*/

data influential;
/*  Merge datasets from above.*/
    merge Rstud
          Cook
          Dffits
          Dfbs;
    by observation;
```

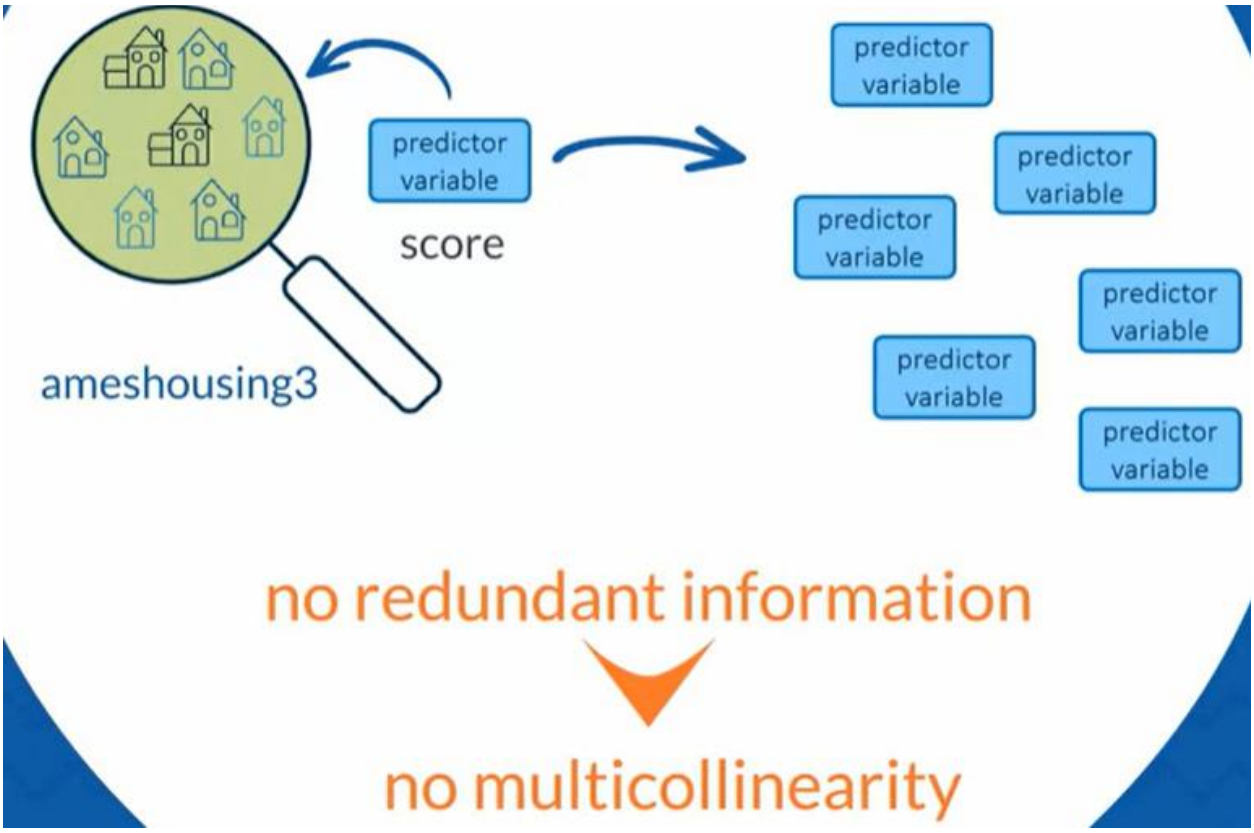
```
/*  Flag observations that have exceeded at least one cutpoint;*/
    if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
    array dfbetas{*} _dfbetasout: ;
    do i=2 to dim(dfbetas);
        if dfbetas{i} then flag=1;
    end;

/*  Set to missing values of influence statistics for those*/
/*  who have not exceeded cutpoints;*/
    if ABS(Rstudent)<=3 then RStudent=.;
    if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
    if flag=1;
    drop i flag;
run;

proc print data=influential;
    id observation ID1;
    var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

The same observations appear in the PROC PRINT report as in the plots. Examine the values of observation 39 to see what is causing problems. You might find it interesting.

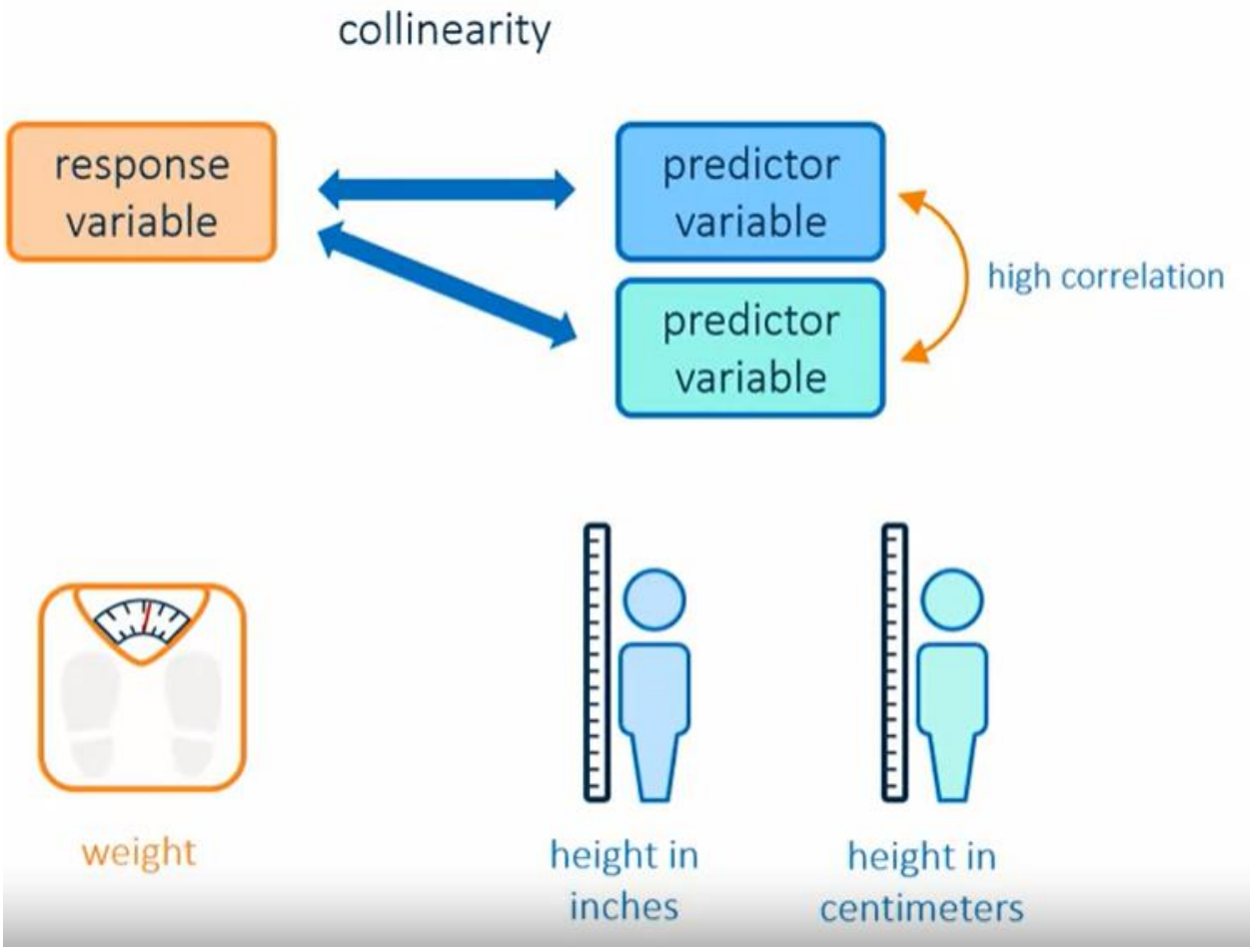
Collinearity

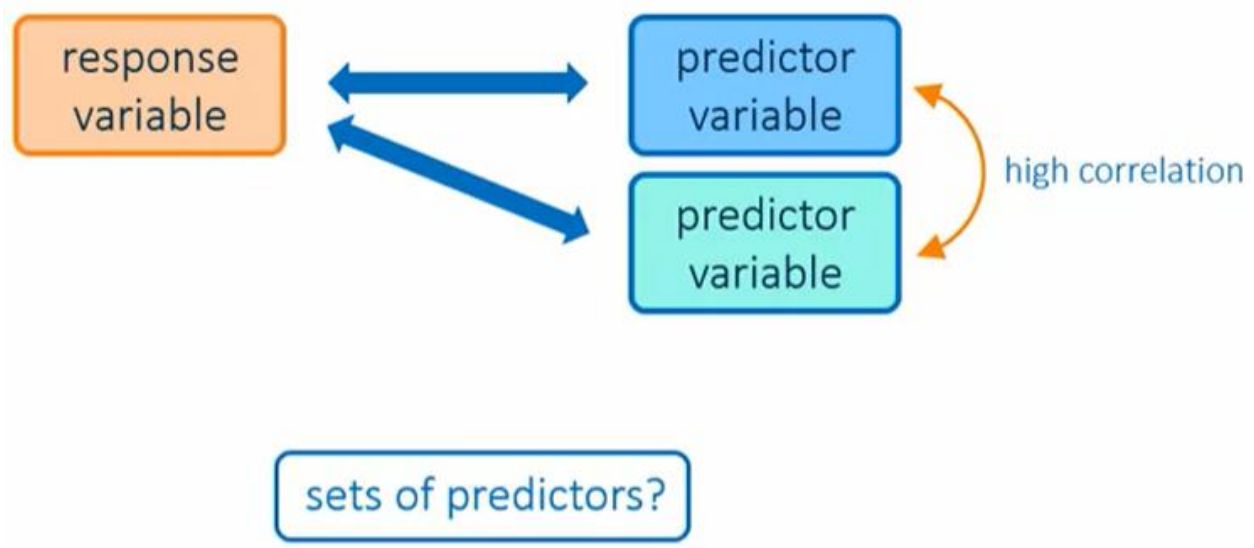
Scenario

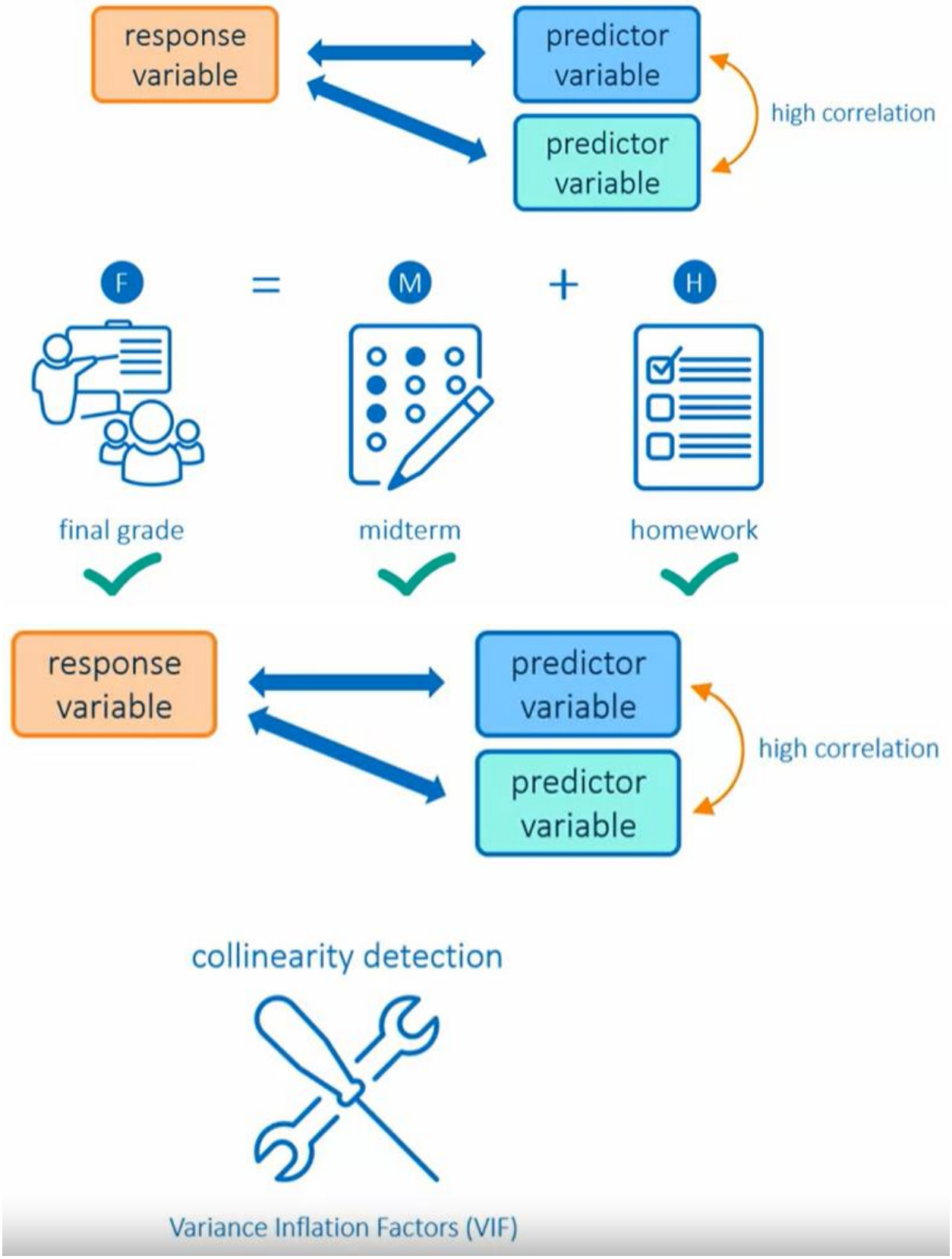


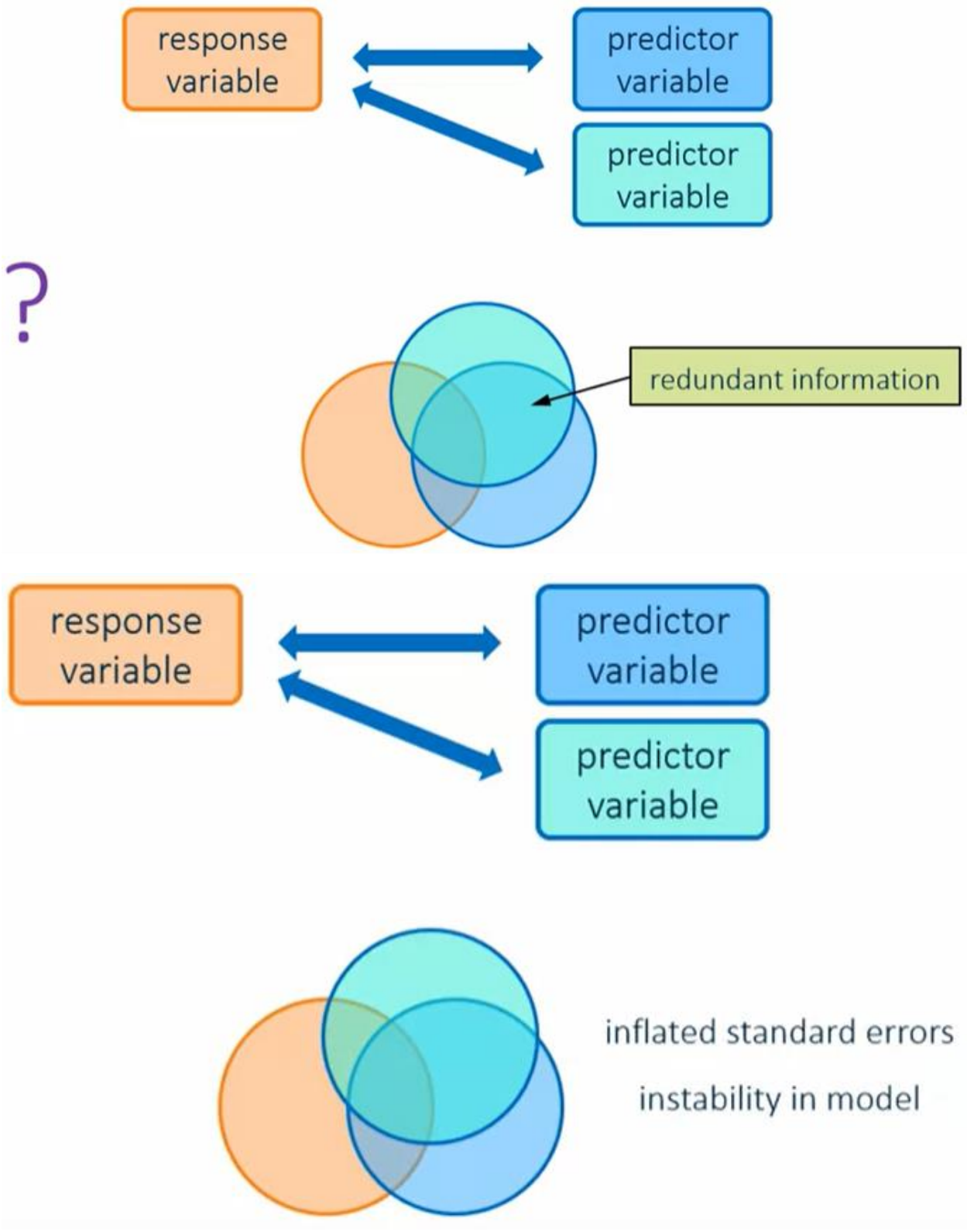


Exploring Collinearity

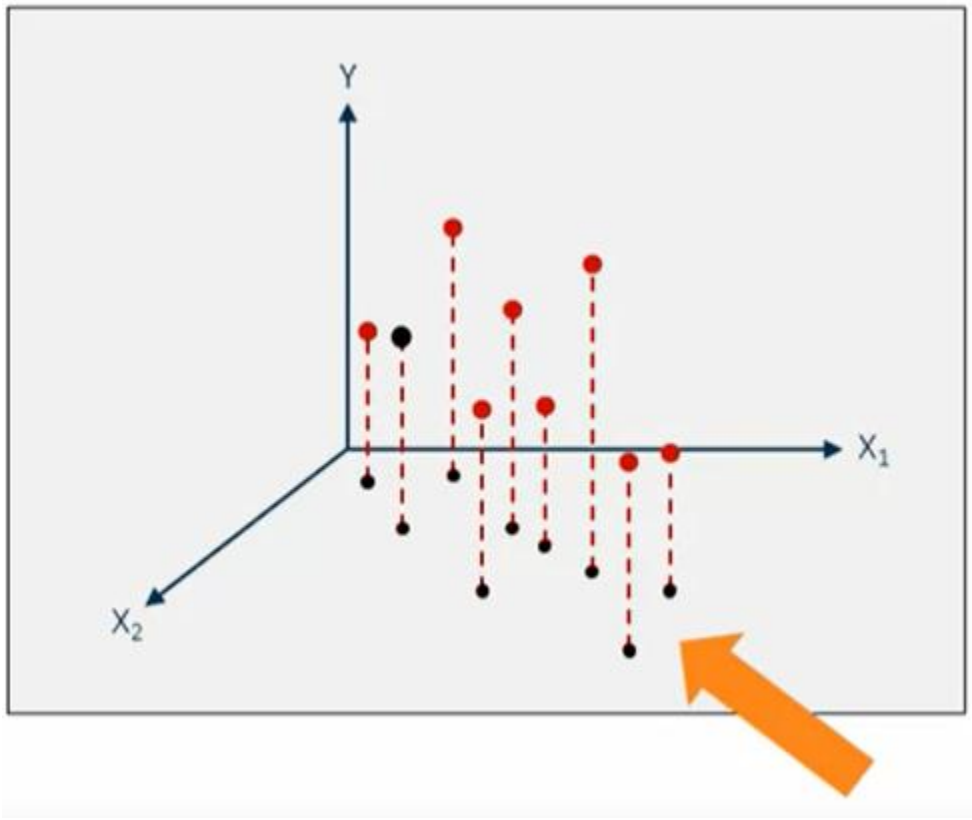


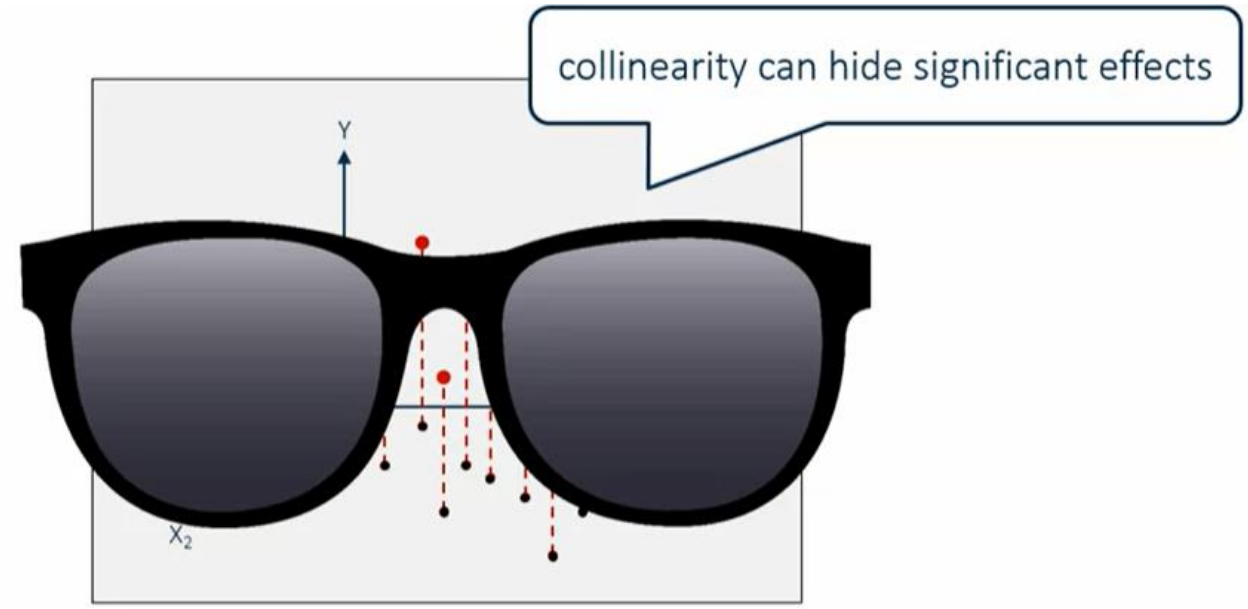


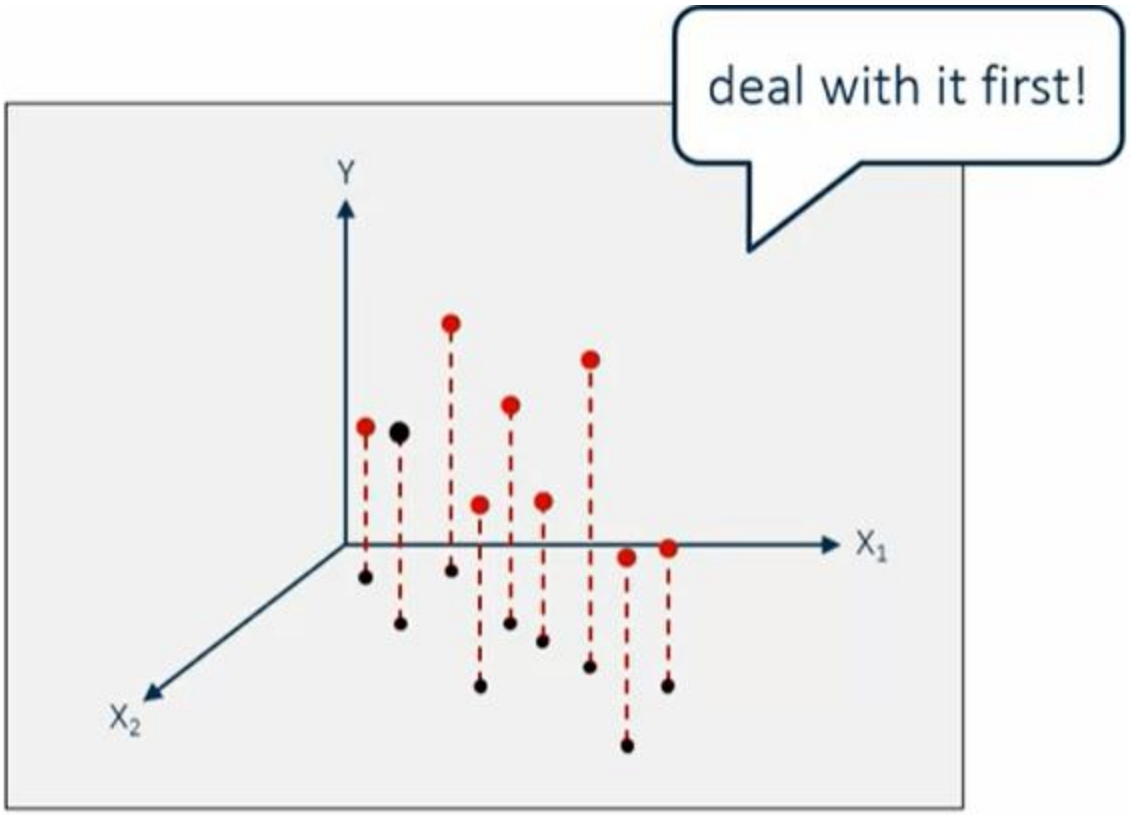

response variable
predictor variable
predictor variable
high correlation
F
=
M
+
H
final grade
midterm
homework
response variable
predictor variable
predictor variable
high correlation
collinearity detection
Variance Inflation Factors (VIF)

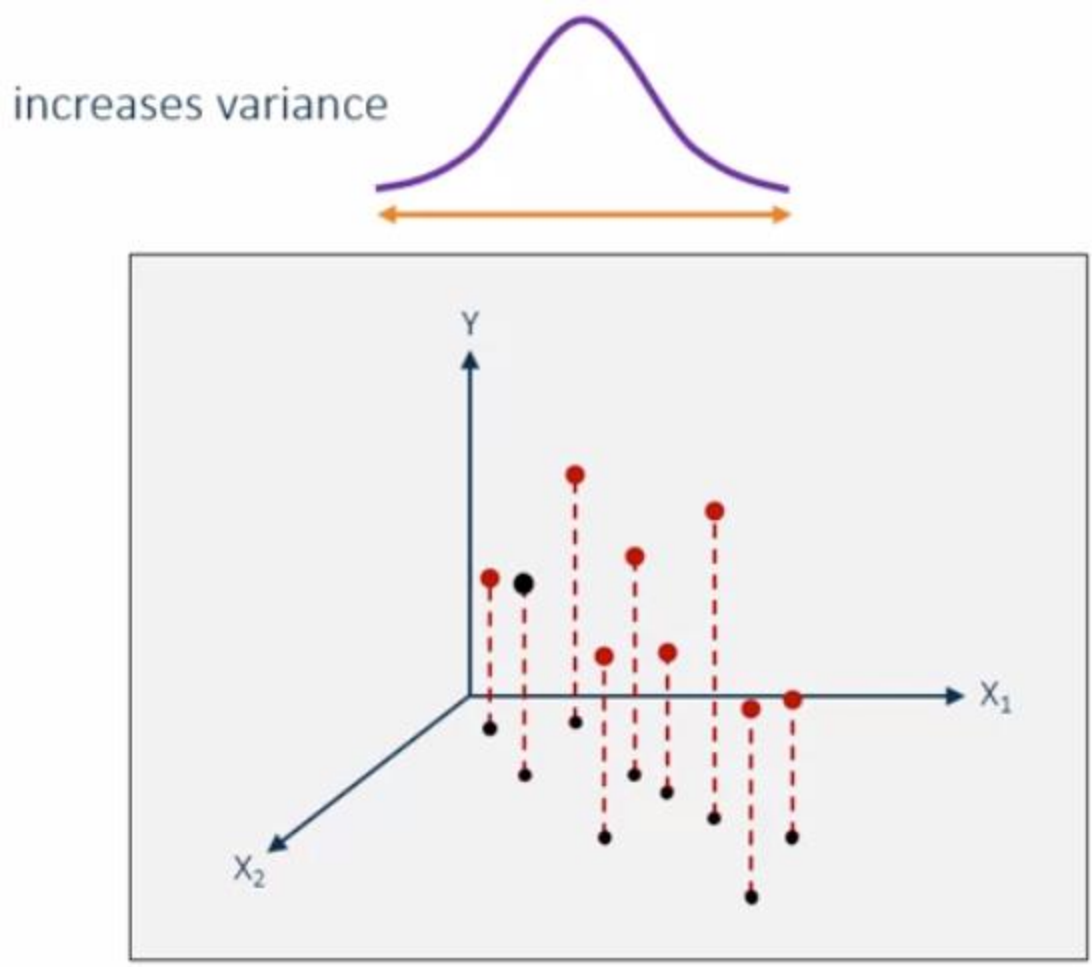

response variable
predictor variable
predictor variable
?
redundant information
response variable
predictor variable
predictor variable
inflated standard errors
instability in model

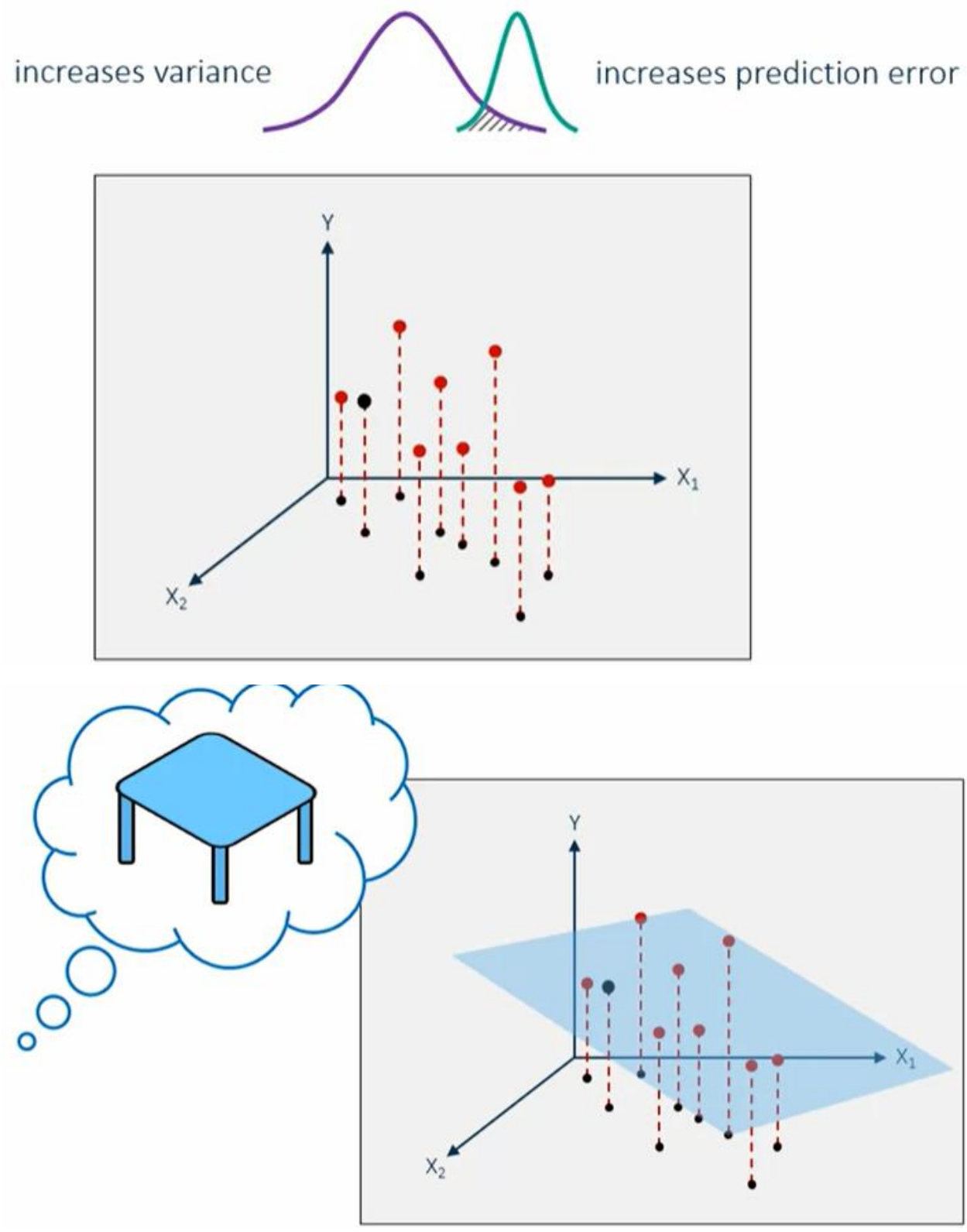
Visualizing Collinearity

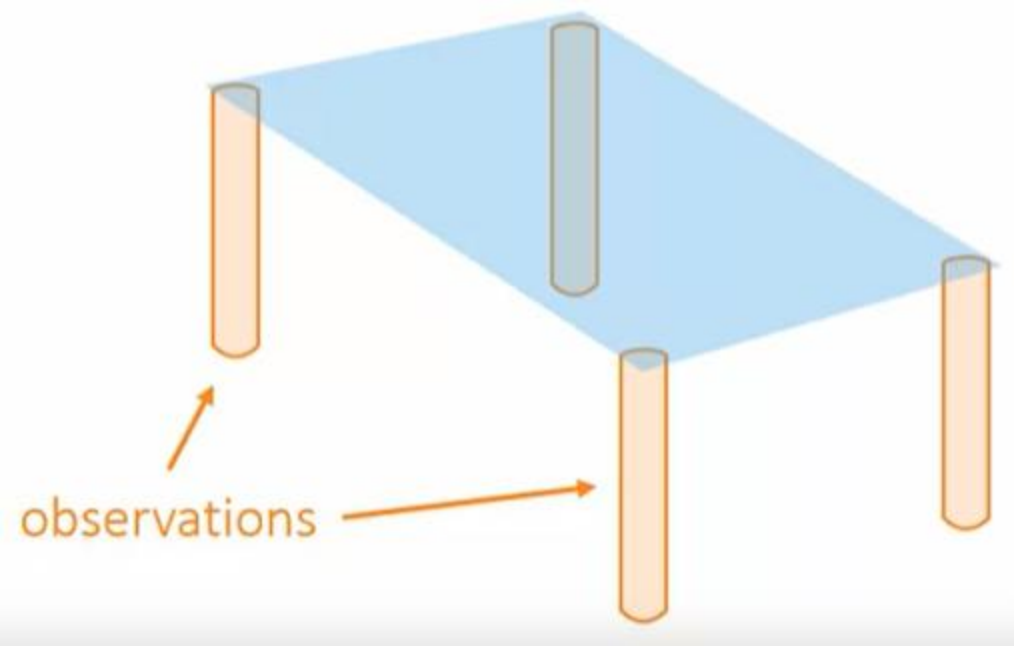


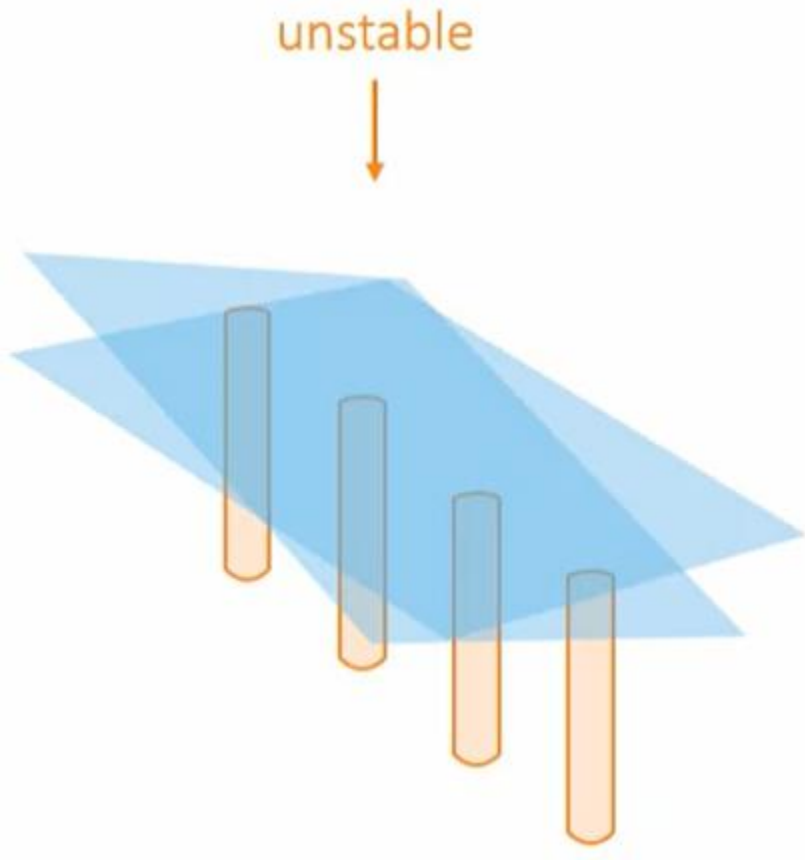


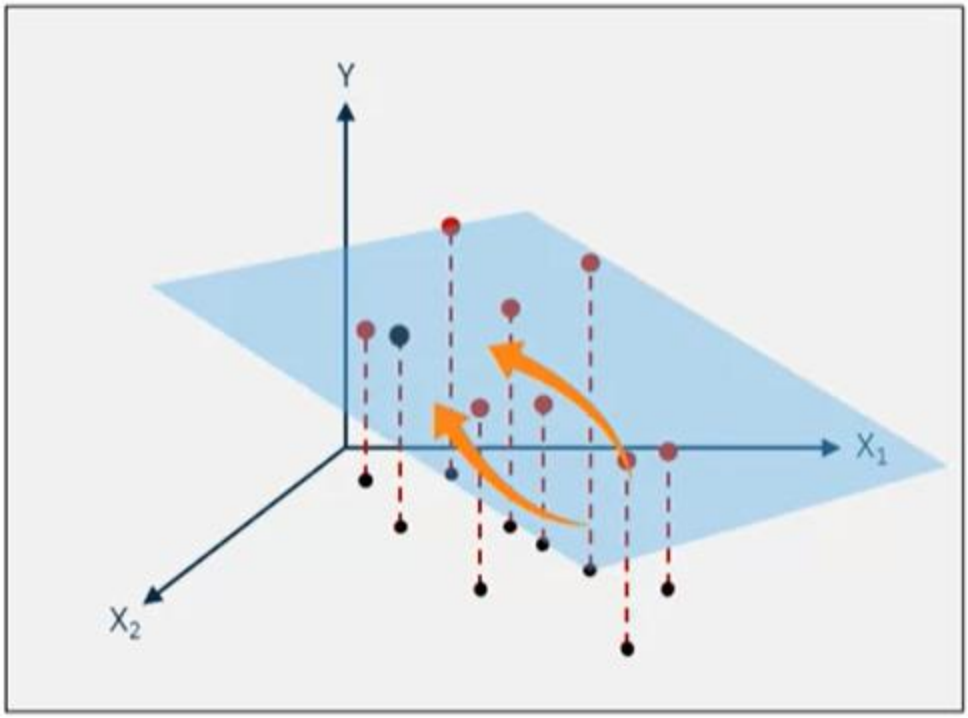

deal with it first!
Y
X1
X2

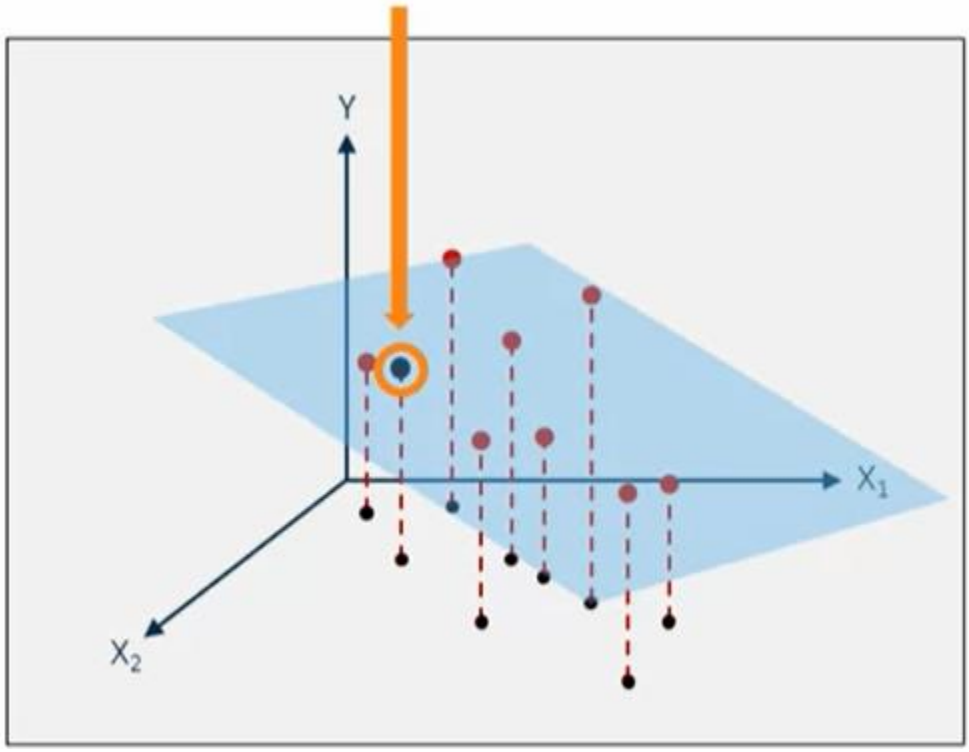

increases variance
Y
X1
X2

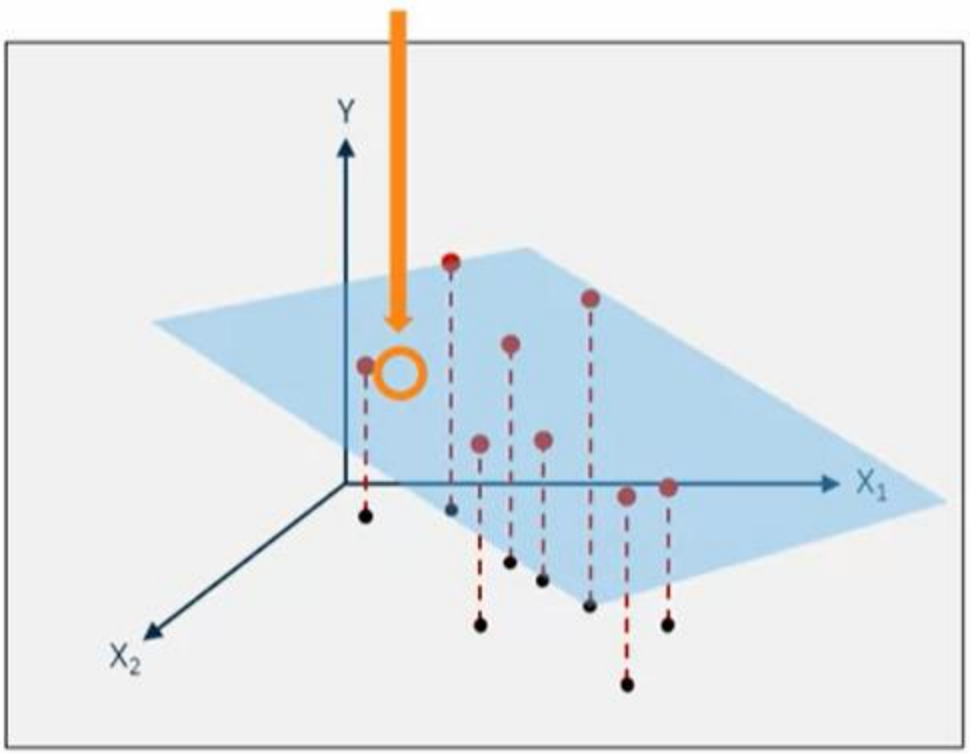

increases variance
increases prediction error
Y
X1
X2
Y
X1
X2

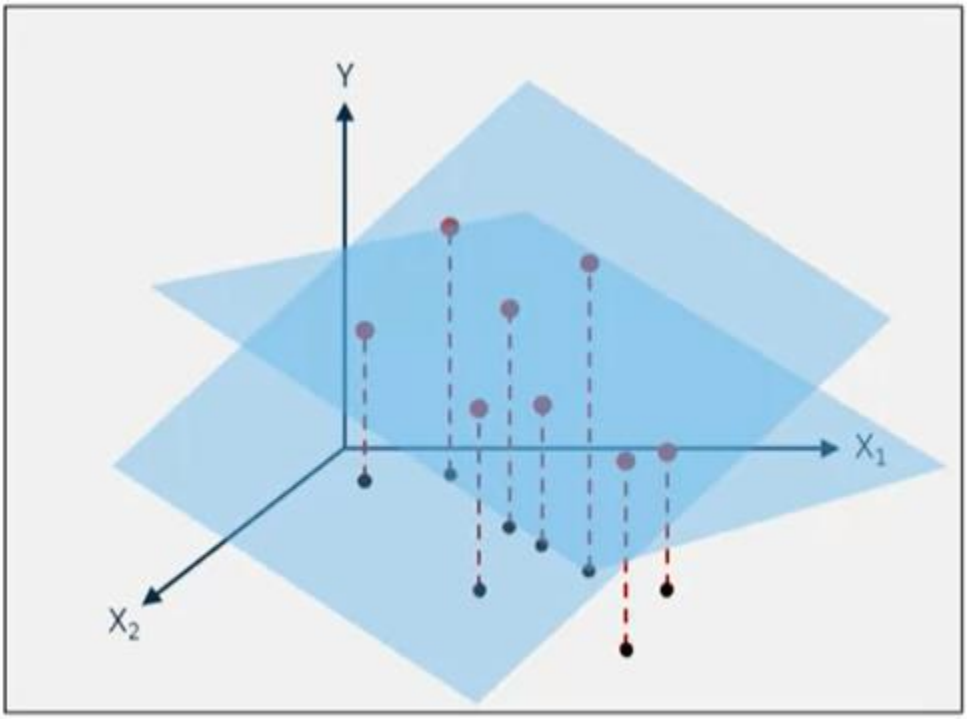

observations

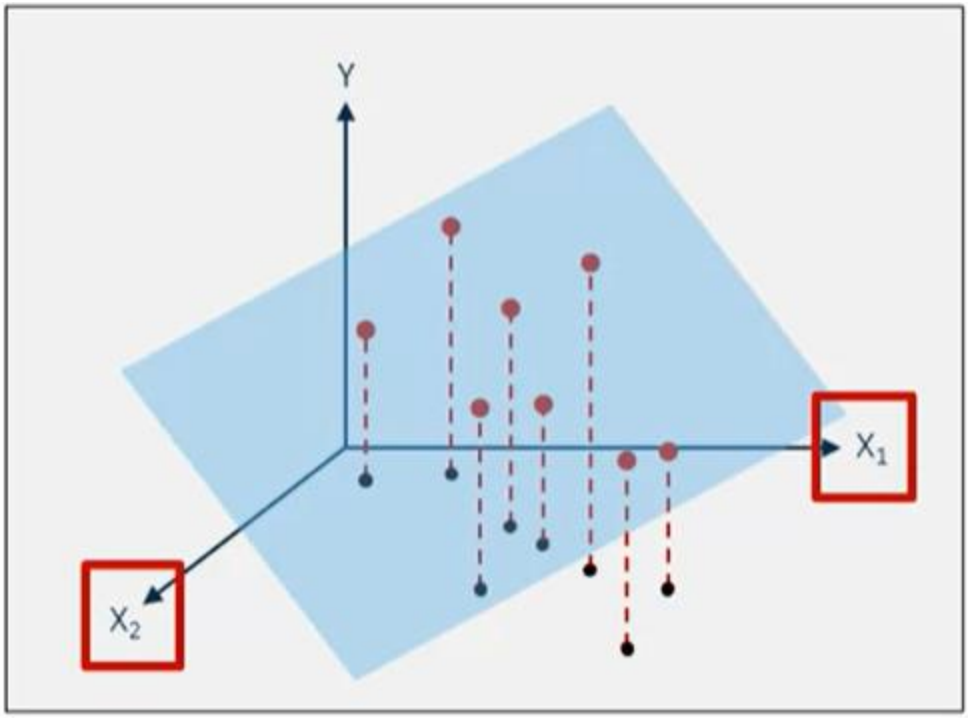

unstable


Y
X1
X2

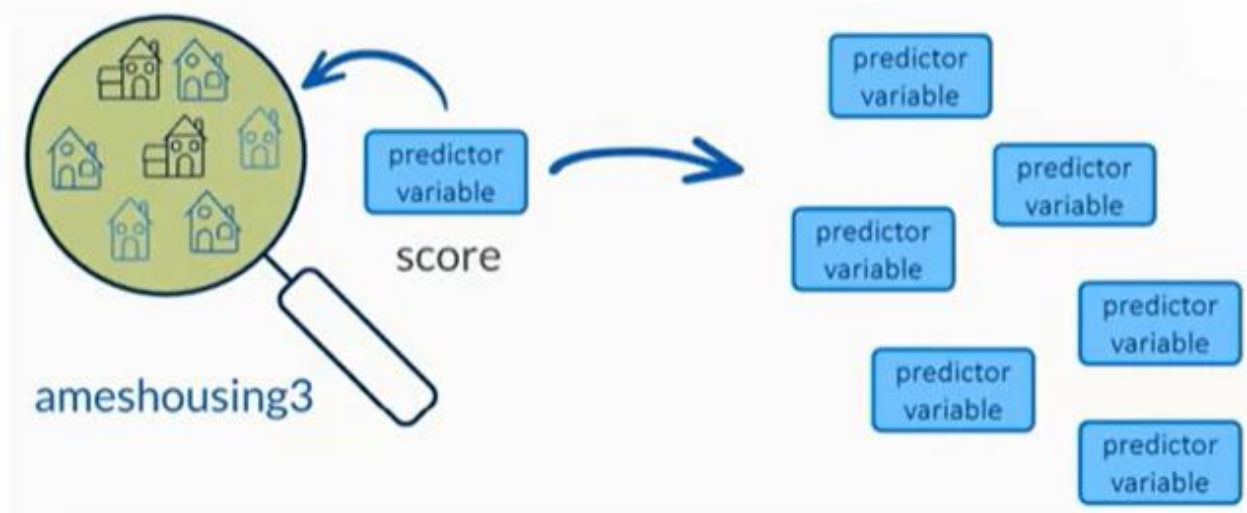

Y
X1
X2

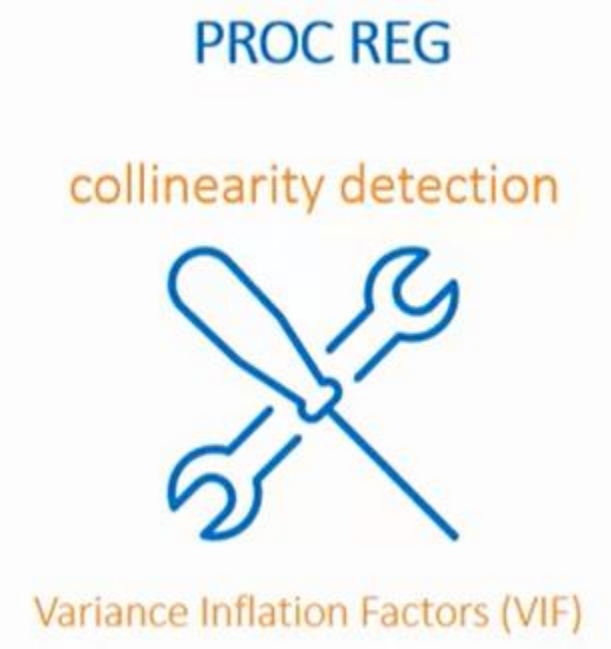

Y
X1
X2


Y
X1
X2


Y
X1
X2

Demo Calculating Collinearity Diagnostics Using PROC REG






```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d03.sas*/  /* Part A*/
proc sort data=STAT1.ameshousing3 out=STAT1.ames_sorted;
    by PID;
run;
proc sort data=STAT1.amesaltuse;
    by PID;
run;

data amescombined;
    merge STAT1.ames_sorted STAT1.amesaltuse;
    by PID;
run;

title;
proc corr data=amescombined nosimple;
    var &interval;
    with score;
run;
```

```
PROC CORR DATA=SAS-data-set <options>;
    VAR variables;
    WITH variables;
    ID variables;
RUN;
```

1 With Variables:	score
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Gr_Liv_Area	Basement_Area	Garage_Area	Deck_Porch_Area	Lot_Area	Age_Sold	Bedroom_AbvGr	Total_Bathroom
score	-0.61394 <.0001 300	-0.97894 <.0001 300	-0.38872 <.0001 300	-0.35979 <.0001 300	-0.29249 <.0001 300	0.39125 <.0001 300	-0.28357 <.0001 300	-0.51877 <.0001 300

```
22
23 /*st105d03.sas*/  /*Part B*/
24 proc reg data=amescombined;
25     model SalePrice = &interval score / vif;
26     title 'Collinearity Diagnostics';
27 run;
28 quit;
29
30 proc reg data=amescombined;
31     NOSCORE: model SalePrice = &interval / vif;
32     title2 'Removing Score';
33 run;
34 quit;
35
36
```

$$VIF_i = \frac{1}{1 - R_i^2}$$

$$VIF_i > 10$$

PROC REG DATA=*SAS-data-set <options>***;**
MODEL *dependents = <regressors> < / options>***;**
RUN;

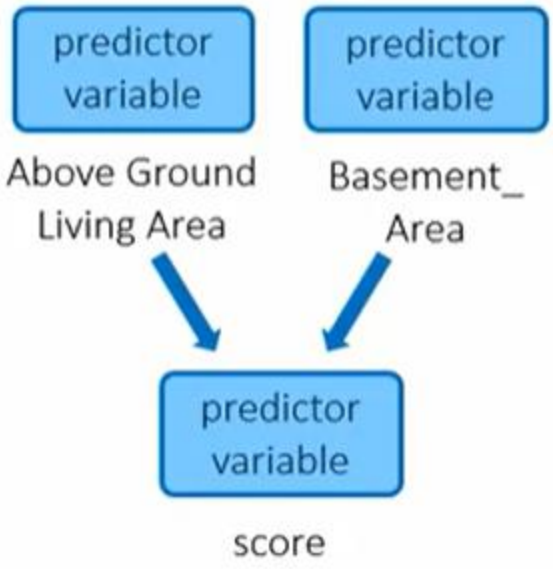
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3.435663E11	38174038666	138.98	<.0001
Error	290	79657171514	274679902		
Corrected Total	299	4.232235E11			

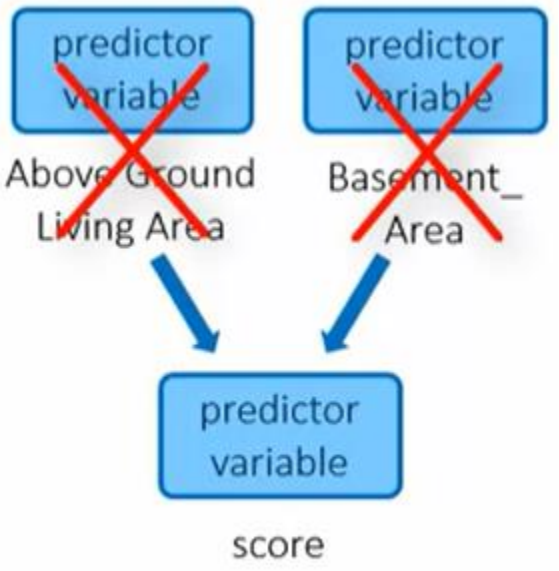
Root MSE	16573	R-Square	0.8118
Dependent Mean	137525	Adj R-Sq	0.8059
Coeff Var	12.05125		

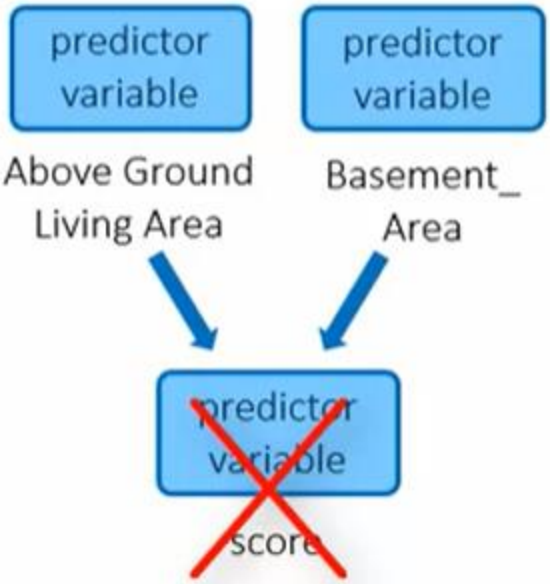
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-4254871	3419274	-1.24	0.2144	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	923.25717	684.04818	1.35	0.1782	27569
Basement_Area	Basement area in square feet	1	2178.35639	1709.68175	1.27	0.2036	411868
Garage_Area	Size of garage in square feet	1	35.01213	6.46640	5.41	<.0001	1.41398
Deck_Porch_Area	Total area of decks and porches in square feet	1	30.64725	7.97228	3.84	0.0001	1.21667
Lot_Area	Lot size in square feet	1	0.69964	0.31644	2.21	0.0278	1.20422
Age_Sold	Age of house when sold, in years	1	-422.21228	44.18905	-9.55	<.0001	1.60476
Bedroom_AbvGr	Bedrooms above grade	1	-4888.35244	1687.71153	-2.90	0.0041	1.48233
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3047.94315	1919.03449	1.59	0.1133	1.73073
score		1	429.97552	341.96962	1.26	0.2096	533085

```
score=round(10000 – (2 * Gr_Liv_Area + 5 * Basement_Area), 10);
```





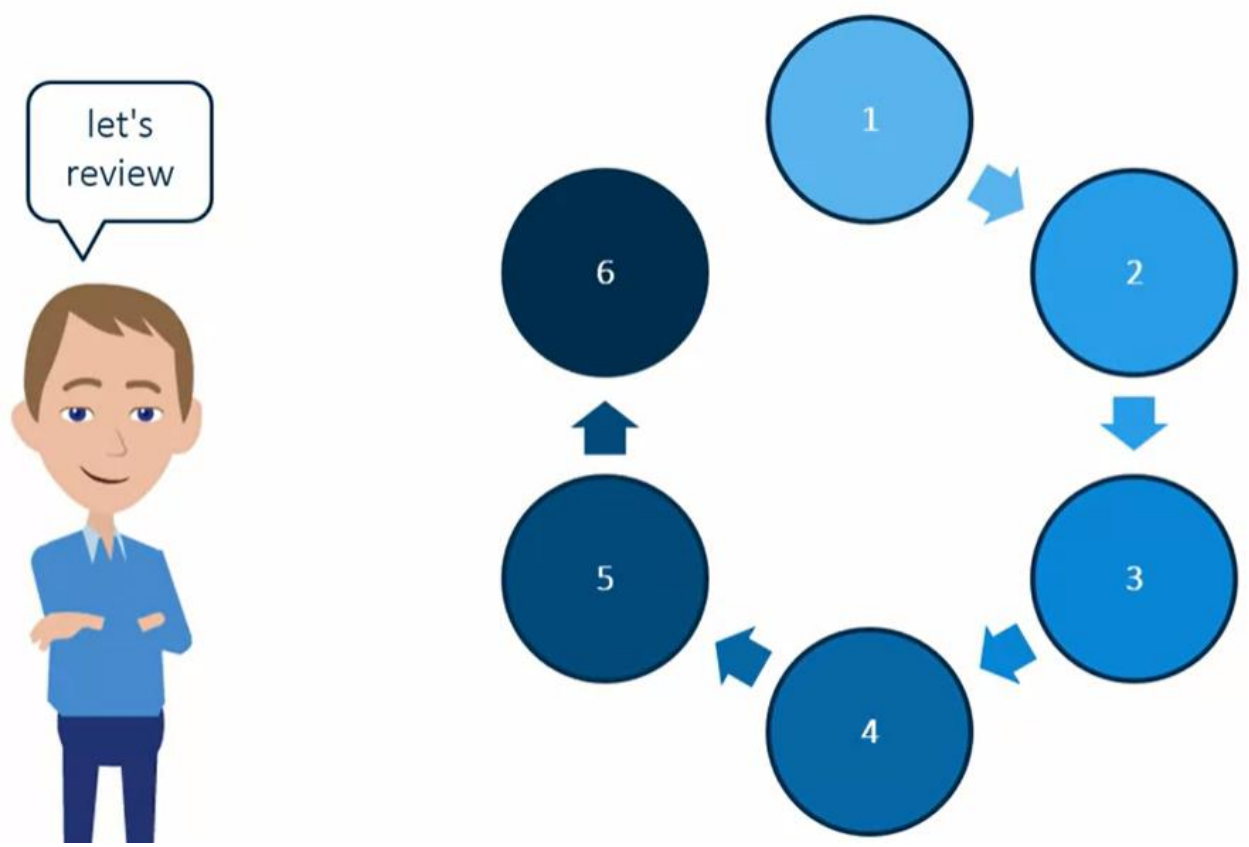


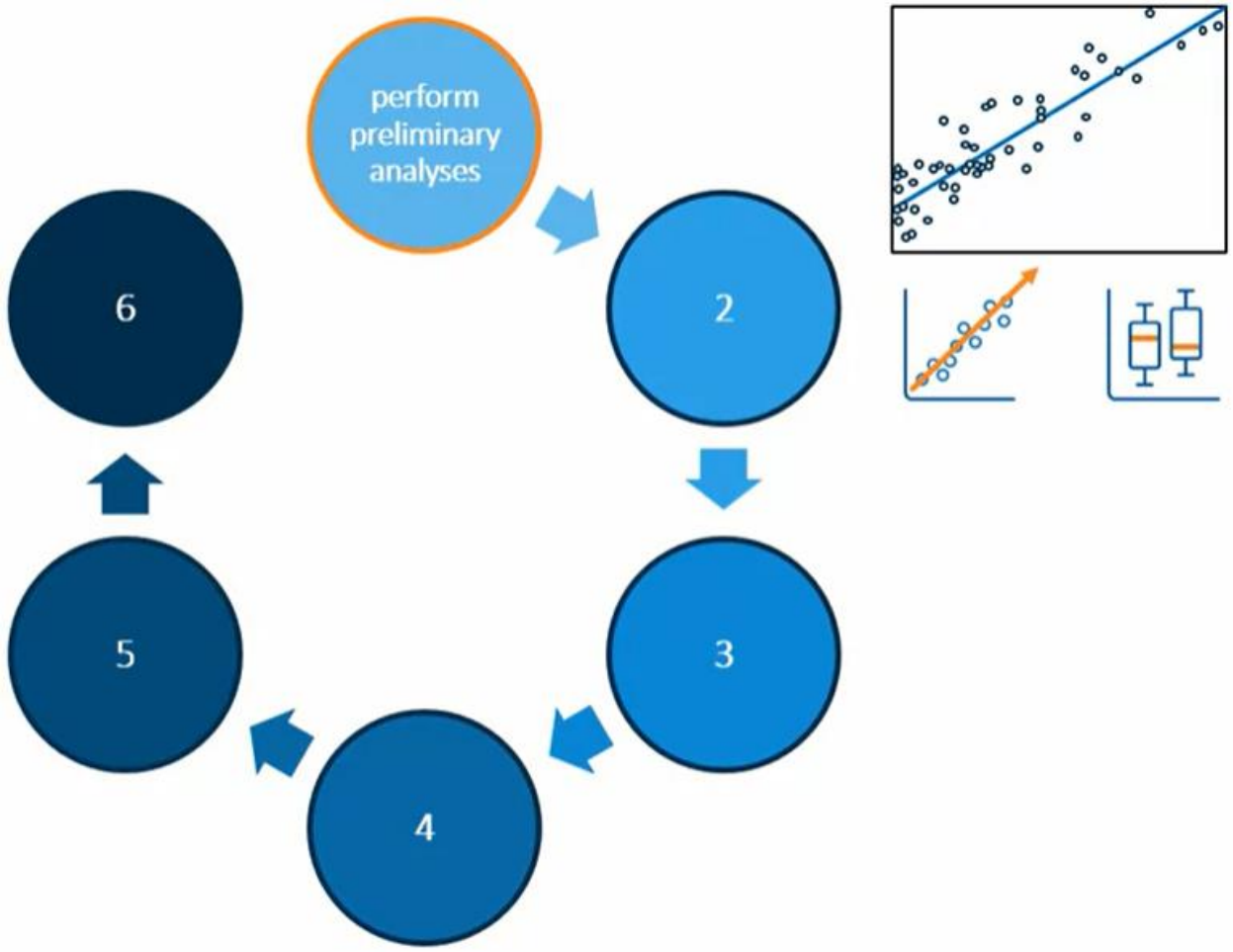
Root MSE	16590	R-Square	0.8108
Dependent Mean	137525	Adj R-Sq	0.8056
Coeff Var	12.06328		

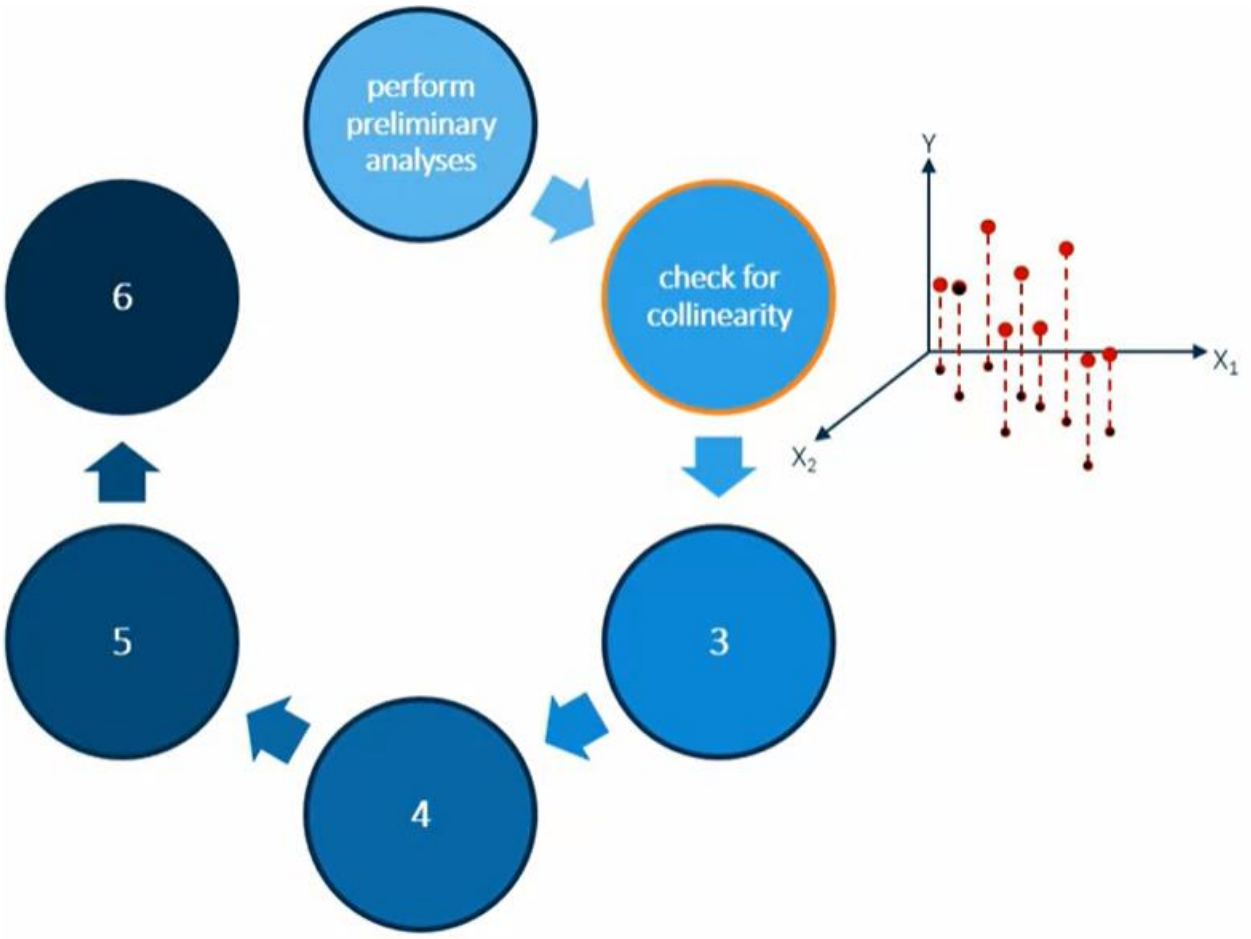
Parameter Estimates

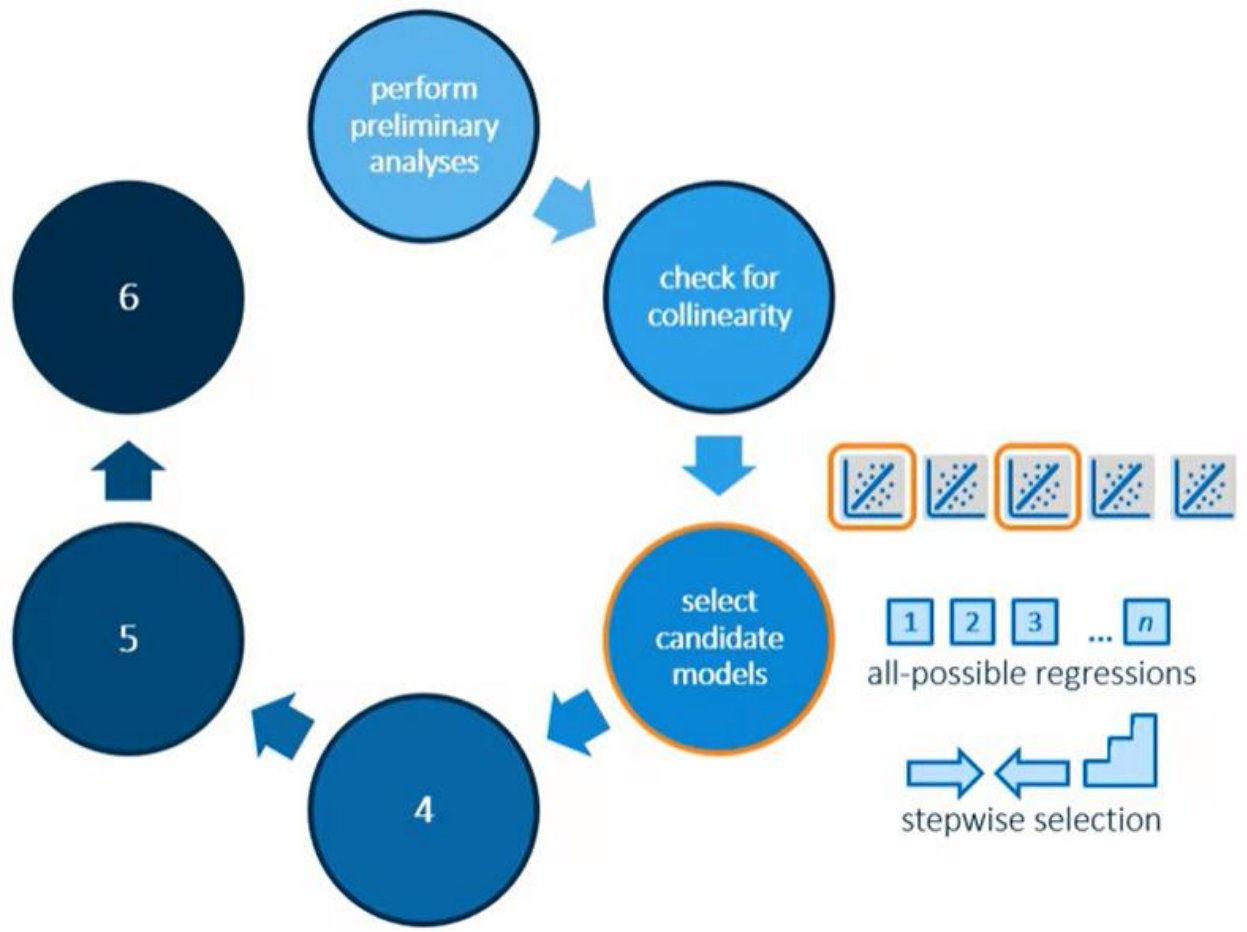
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19778	5.58574	11.31	<.0001	1.83461
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001	1.84195
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001	1.40220
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001	1.21034
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280	1.20422
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001	1.60375
Bedroom_AbvGr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045	1.48138
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167	1.73053

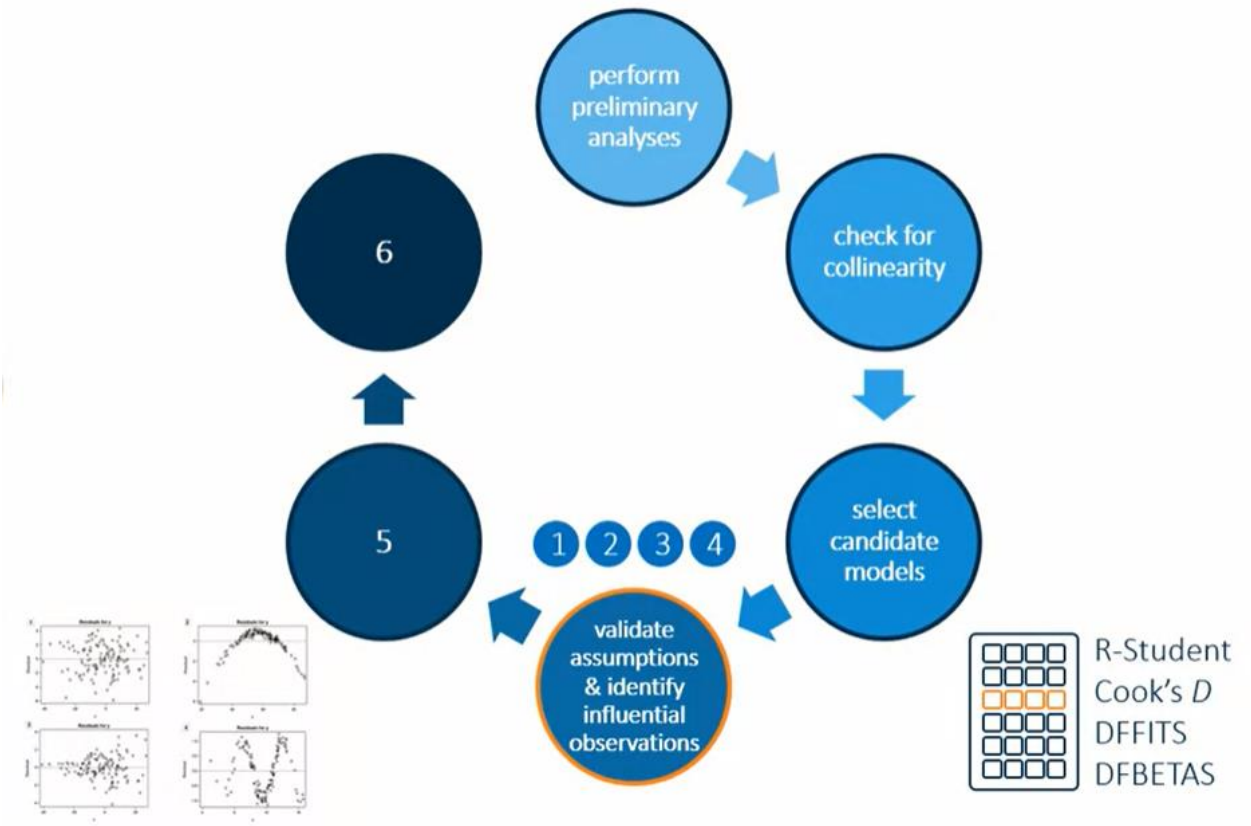
Using an Effective Modelling Cycle

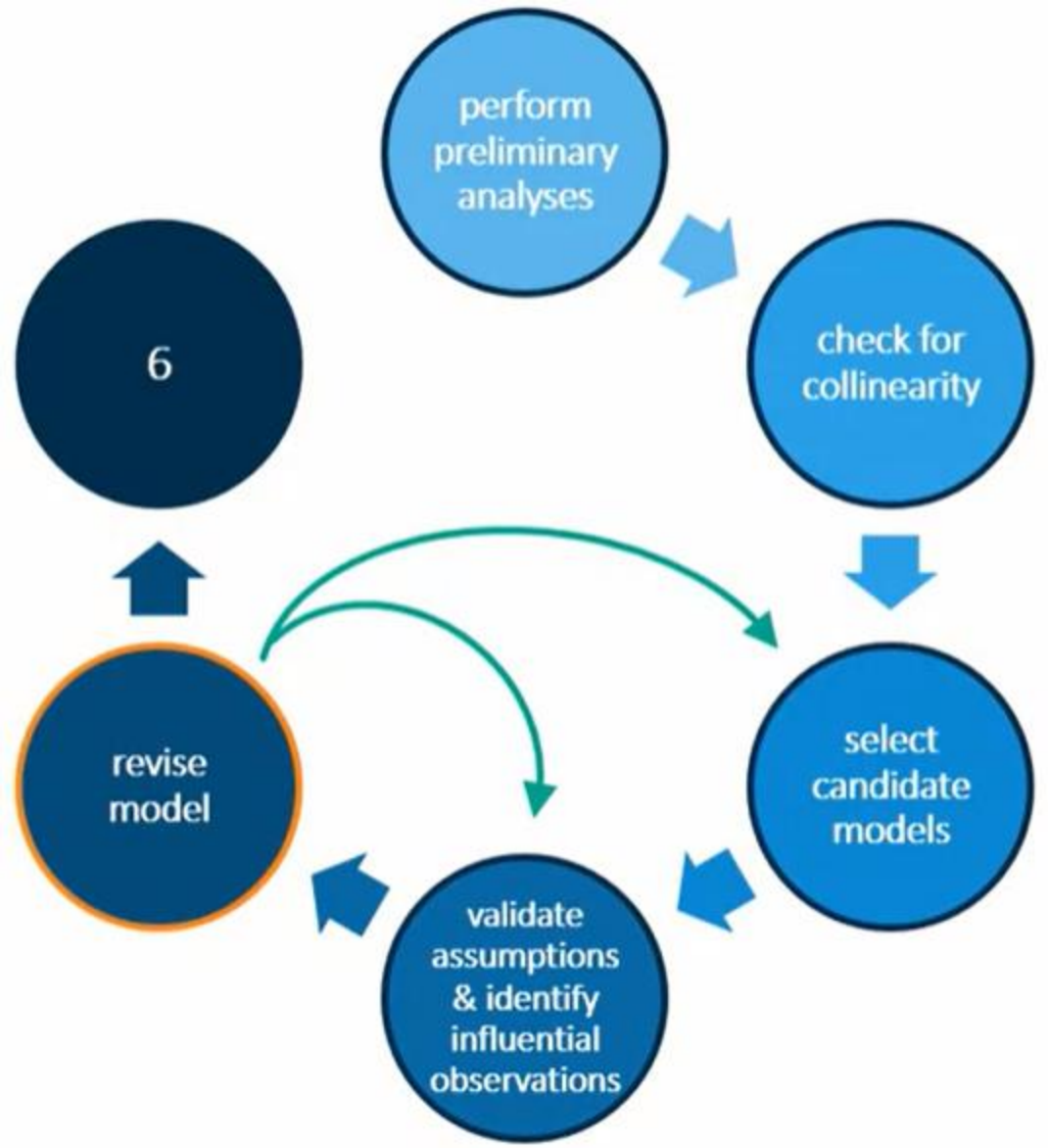


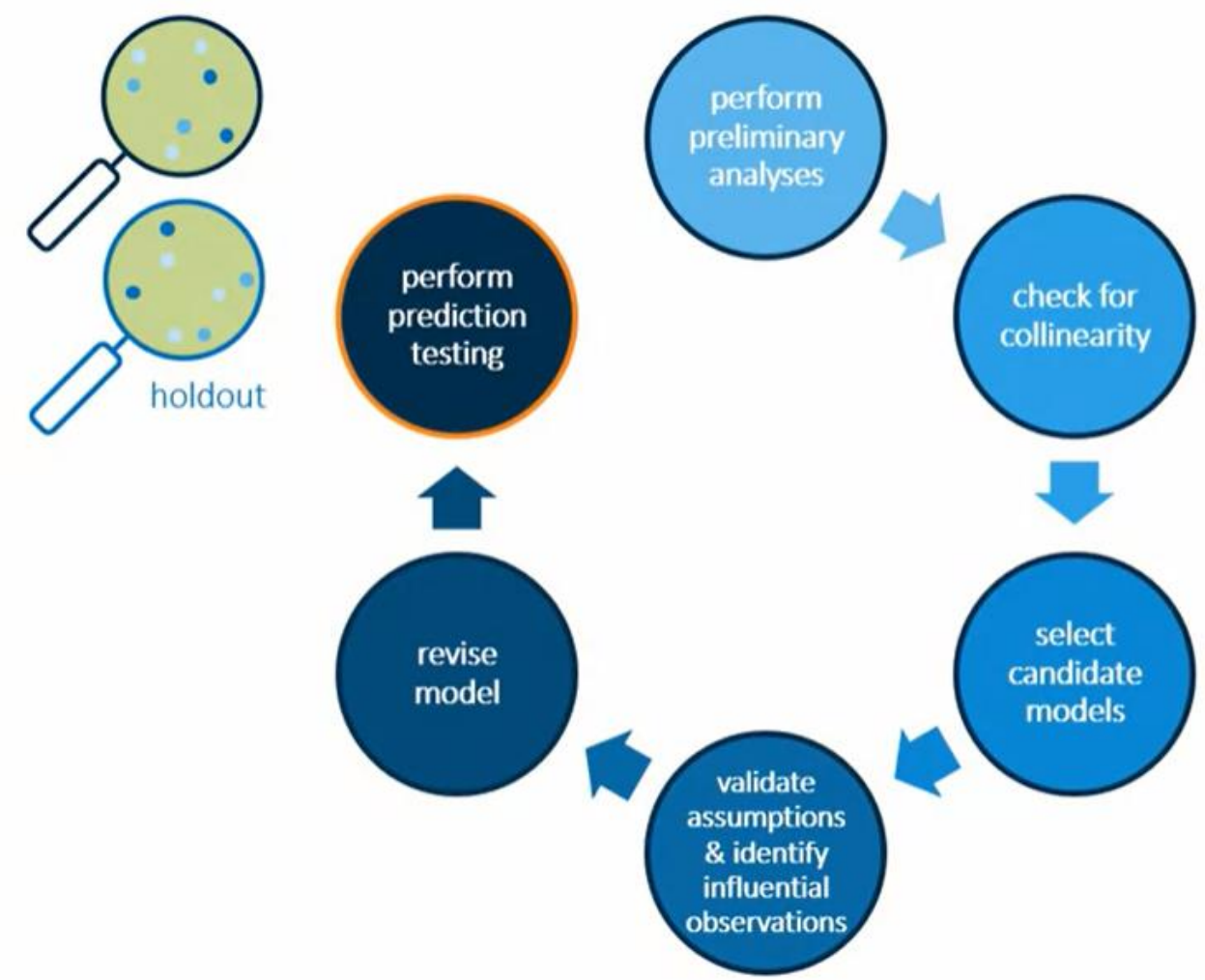

perform preliminary analyses
2
3
4
5
6


perform preliminary analyses
check for collinearity
3
4
5
6
Y
X1
X2


perform preliminary analyses
check for collinearity
select candidate models
4
5
6
1 2 3 ... n
all-possible regressions
stepwise selection


perform preliminary analyses
check for collinearity
select candidate models
validate assumptions & identify influential observations
5
6
1 2 3 4
R-Student
Cook's D
DFFITS
DFBETAS


perform preliminary analyses
check for collinearity
select candidate models
validate assumptions & identify influential observations
revise model
6




```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d03.sas*/  /* Part A*/
proc sort data=STAT1.ameshousing3 out=STAT1.ames_sorted;
   by PID;
run;
proc sort data=STAT1.amesaltuse;
   by PID;
run;
```

```
data amescombined;
   merge STAT1.ames_sorted STAT1.amesaltuse;
   by PID;
run;


title;
proc corr data=amescombined nosimple;
   var &interval;
   with score;
run;
```

The CORR Procedure

1 With Variables:	score
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Gr_Liv_Area	Basement_Area	Garage_Area	Deck_Porch_Area	Lot_Area	Age_Sold	Bedroom_AbvGr	Total_Bathroom
score	-0.61394 <.0001 300	-0.97894 <.0001 300	-0.38872 <.0001 300	-0.35979 <.0001 300	-0.29249 <.0001 300	0.39125 <.0001 300	-0.28357 <.0001 300	-0.51877 <.0001 300

```
/*st105d03.sas*/  /*Part B*/
proc reg data=amescombined;
   model SalePrice = &interval score / vif;
   title 'Collinearity Diagnostics';
run;
quit;
```

Collinearity Diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	2930
Number of Observations Used	300
Number of Observations with Missing Values	2630

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3.435663E11	38174038666	138.98	<.0001
Error	290	79657171514	274679902		
Corrected Total	299	4.232235E11			

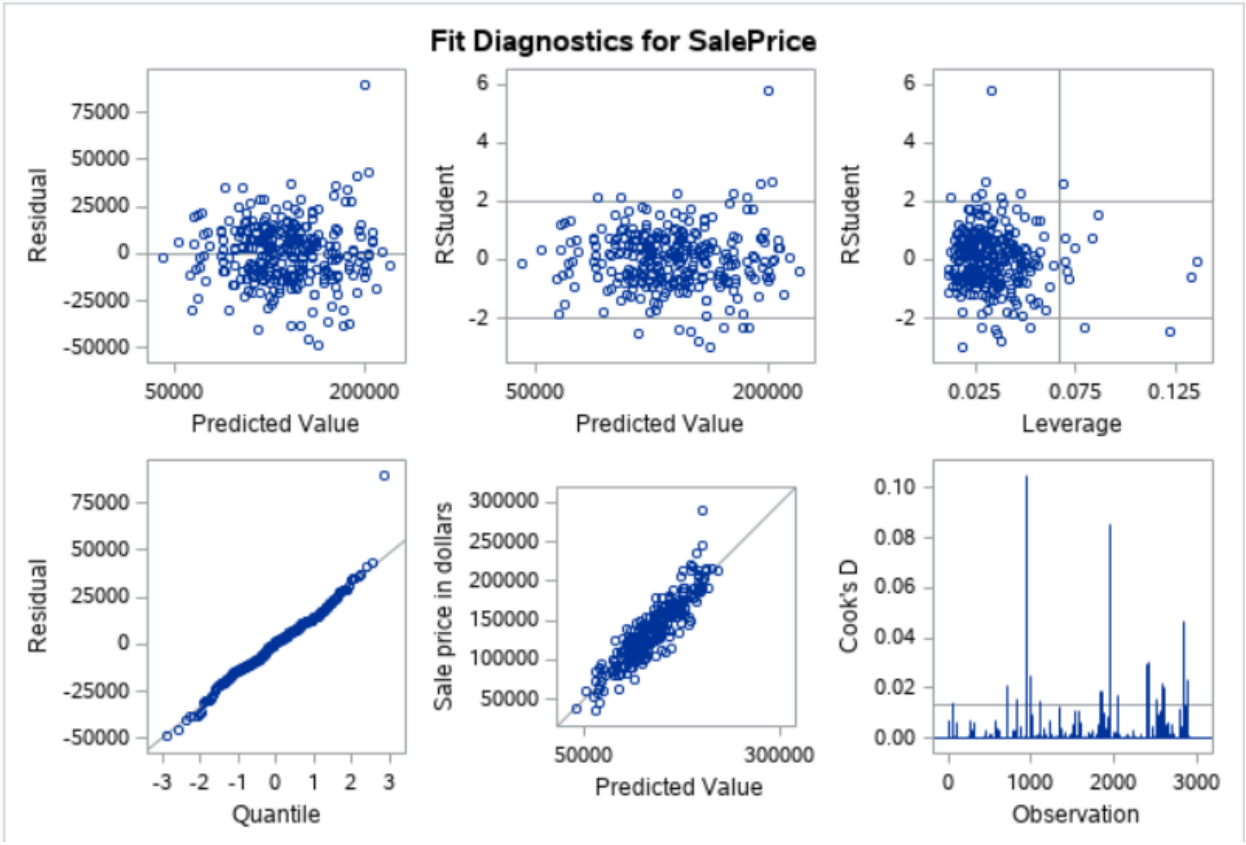
Root MSE	16573	R-Square	0.8118
Dependent Mean	137525	Adj R-Sq	0.8059
Coeff Var	12.05125		

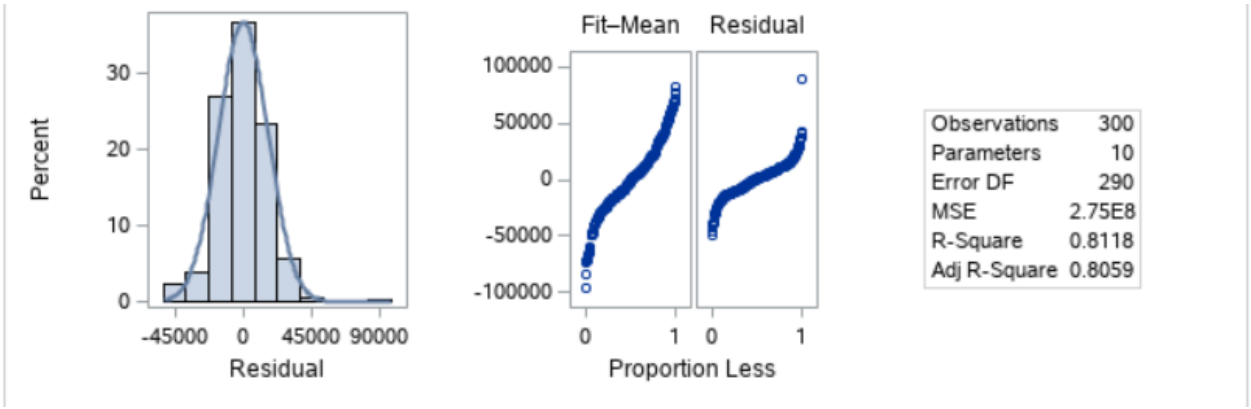
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-4254871	3419274	-1.24	0.2144	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	923.25717	684.04818	1.35	0.1782	27569
Basement_Area	Basement area in square feet	1	2178.35638	1709.68175	1.27	0.2036	411868
Garage_Area	Size of garage in square feet	1	35.01213	6.46640	5.41	<.0001	1.41398
Deck_Porch_Area	Total area of decks and porches in square feet	1	30.64725	7.97228	3.84	0.0001	1.21667
Lot_Area	Lot size in square feet	1	0.69964	0.31644	2.21	0.0278	1.20422
Age_Sold	Age of house when sold, in years	1	-422.21228	44.18905	-9.55	<.0001	1.60476
Bedroom_AbvGr	Bedrooms above grade	1	-4888.35244	1687.71153	-2.90	0.0041	1.48233
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3047.94315	1919.03449	1.59	0.1133	1.73073
score		1	429.97552	341.96962	1.26	0.2096	533085

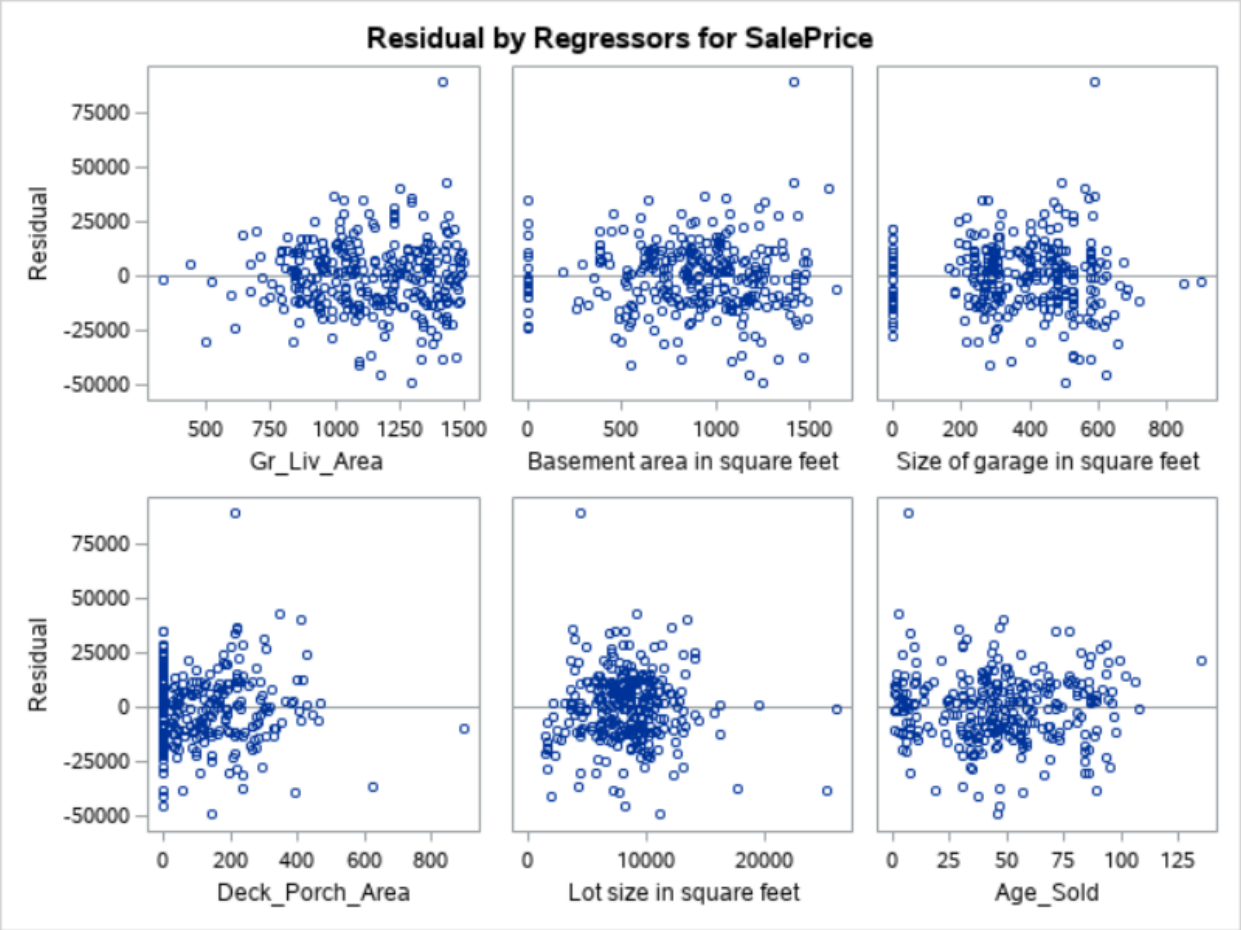
Collinearity Diagnostics

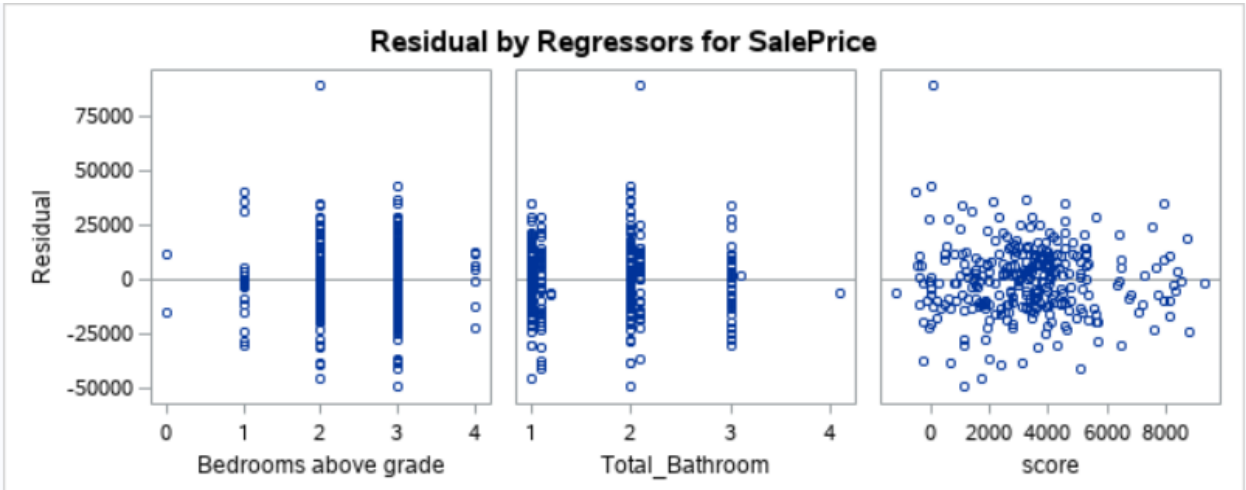
The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars





Observations	300
Parameters	10
Error DF	290
MSE	2.75E8
R-Square	0.8118
Adj R-Square	0.8059






```
proc reg data=amescombined;
   NOSCORE: model SalePrice = &interval / vif;
   title2 'Removing Score';
run;
```

quit;

Removing Score

The REG Procedure
Model: NOSCORE
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	2930
Number of Observations Used	300
Number of Observations with Missing Values	2630

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3.431321E11	42891512314	155.84	<.0001
Error	291	80091420996	275228251		
Corrected Total	299	4.232235E11			

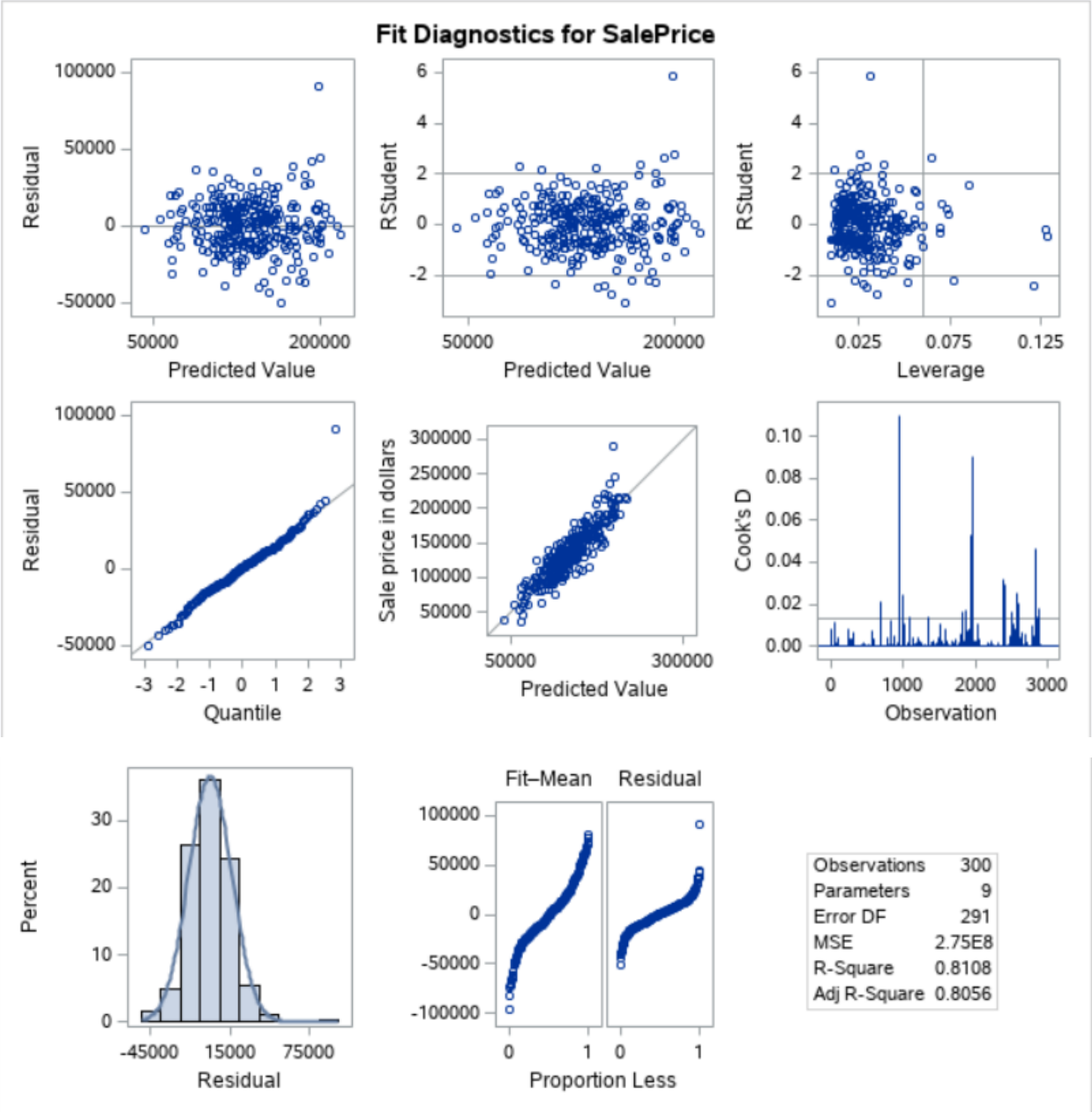
Root MSE	16590	R-Square	0.8108
Dependent Mean	137525	Adj R-Sq	0.8056
Coeff Var	12.06328		

Parameter Estimates

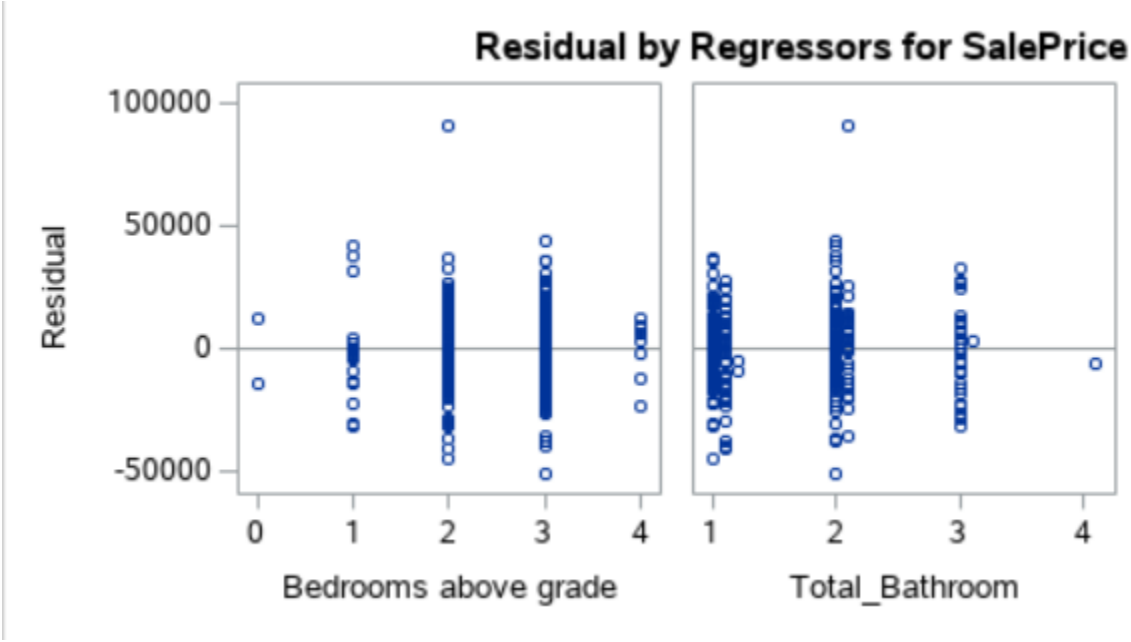
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19776	5.58574	11.31	<.0001	1.83461
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001	1.64195
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001	1.40220
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001	1.21034
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280	1.20422
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001	1.60375
Bedroom_AbvGr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045	1.48138
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167	1.73053

Removing Score

The REG Procedure
Model: NOSCORE
Dependent Variable: SalePrice Sale price in dollars




Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
500
750
1000
1250
1500
Gr_Liv_Area
0
500
1000
1500
Basement area in square feet
0
200
400
600
800
Size of garage in square feet
0
200
400
600
800
Deck_Porch_Area
0
10000
20000
Lot size in square feet
0
25
50
75
100
125
Age_Sold


Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
0
1
2
3
4
Bedrooms above grade
1
2
3
4
Total_Bathroom

```
/*st105s03.sas*/  /*Part A*/
ods graphics off;
proc reg data=STAT1.BodyFat2;
   FULLMODL: model PctBodyFat2
             = Age Weight Height
               Neck Chest Abdomen Hip Thigh
               Knee Ankle Biceps Forearm Wrist
             / vif;
   title 'Collinearity -- Full Model';
run;
quit;

ods graphics on;
```

Collinearity -- Full Model

The REG Procedure
Model: FULLMODL
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-21.35323	22.18616	-0.96	0.3368	0
Age	1	0.06457	0.03219	2.01	0.0460	2.22447
Weight	1	-0.09638	0.06185	-1.56	0.1205	44.65251
Height	1	-0.04394	0.17870	-0.25	0.8060	2.93911
Neck	1	-0.47547	0.23557	-2.02	0.0447	4.43192
Chest	1	-0.01718	0.10322	-0.17	0.8679	10.23469
Abdomen	1	0.95500	0.09016	10.59	<.0001	12.77553
Hip	1	-0.18859	0.14479	-1.30	0.1940	14.54193
Thigh	1	0.24835	0.14617	1.70	0.0906	7.95866
Knee	1	0.01395	0.24775	0.06	0.9552	4.82530
Ankle	1	0.17788	0.22262	0.80	0.4251	1.92410
Biceps	1	0.18230	0.17250	1.06	0.2917	3.67091
Forearm	1	0.45574	0.19930	2.29	0.0231	2.19193
Wrist	1	-1.65450	0.53316	-3.10	0.0021	3.34840

```
/*st105s03.sas*/  /*Part B*/
ods graphics off;
proc reg data=STAT1.BodyFat2;
   NOWT: model PctBodyFat2
         = Age Height
           Neck Chest Abdomen Hip Thigh
           Knee Ankle Biceps Forearm Wrist
         / vif;
   title 'Collinearity -- No Weight';
run;
quit;


ods graphics on;
```

Collinearity -- No Weight

The REG Procedure
Model: NOWT
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	13114	1092.81860	58.49	<.0001
Error	239	4465.16664	18.68271		
Corrected Total	251	17579			

Root MSE	4.32235	R-Square	0.7460
Dependent Mean	19.15079	Adj R-Sq	0.7332
Coeff Var	22.57008		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.44939	8.73019	1.20	0.2325	0
Age	1	0.07376	0.03174	2.32	0.0210	2.14988
Height	1	-0.22096	0.13836	-1.60	0.1116	1.75152
Neck	1	-0.60041	0.22217	-2.70	0.0074	3.91858
Chest	1	-0.09400	0.09096	-1.03	0.3025	7.90070
Abdomen	1	0.91038	0.08575	10.62	<.0001	11.48744
Hip	1	-0.30384	0.12485	-2.43	0.0157	10.74814
Thigh	1	0.21896	0.14538	1.51	0.1334	7.82619
Knee	1	-0.02664	0.24711	-0.11	0.9142	4.77198
Ankle	1	0.10706	0.21858	0.49	0.6247	1.84391
Biceps	1	0.12481	0.16901	0.74	0.4610	3.50299
Forearm	1	0.45808	0.19989	2.29	0.0228	2.19181
Wrist	1	-1.77201	0.52937	-3.35	0.0009	3.28143

Practice: Using PROC REG to Assess Collinearity

Question 1

Run a regression of **PctBodyFat2** on all the other numeric variables in the data set **stat1.bodyfat2**.

1. Write a PROC REG step to determine whether a collinearity problem exists in your model.
2. Submit the code and view the results.

Is there a collinearity problem in your model?

There seems to be high collinearity with **Weight**, **Hip**, and **Abdomen**. **Chest** and **Thigh** are below the cut off but are larger than the others that do not exceed 5.

```
/*st105s03.sas*/  /*Part A*/

ods graphics off;
proc reg data=STAT1.BodyFat2;
   FULLMODL: model PctBodyFat2 =
                  Age Weight Height
                  Neck Chest Abdomen Hip Thigh
                  Knee Ankle Biceps Forearm Wrist
                  / vif;
   title 'Collinearity -- Full Model';
run;
quit;
ods graphics on;
```

Question 2

If there is a collinearity problem, what would you like to do about it? Will you remove any variables? Why or why not?

The answer is not so easy. **Weight** is collinear with some of the other variables, but as you saw before in your model-building process, **Weight** is a relatively significant predictor in the "best" models. A subject-matter expert should determine the answer. If you want to remove **Weight**, simply run that model again without that variable.

```
/*st105s03.sas*/  /*Part B*/

ods graphics off;
proc reg data=STAT1.BodyFat2;
   NOWT: model PctBodyFat2 =
               Age Height
               Neck Chest Abdomen Hip Thigh
               Knee Ankle Biceps Forearm Wrist
               / vif;
   title 'Collinearity -- No Weight';
run;
quit;

ods graphics on;
```