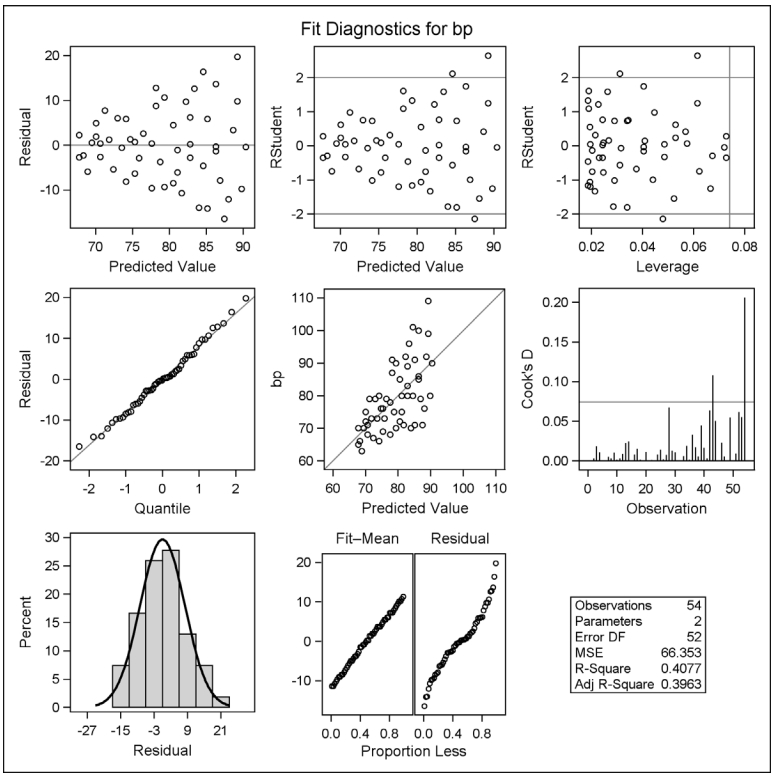
Stat 5100 Handout #21 – SAS: Variations on Ordinary Least Squares
(Weighted Least Squares, Robust Regression, Nonlinear Regression)

Example 1: (Weighted Least Squares) A health researcher is interested in studying the relationship between diastolic blood pressure (bp) and age in adult women. Data are reported on 54 healthy adult women.

```
/* Read in the data (Table 11.1) */
data bpexample; input age bp @@; cards;
  27   73   21   66   22   63   24   75   25   71   23   70
  20   65   20   70   29   79   24   72   25   68   28   67
  26   79   38   91   32   76   33   69   31   66   34   73
  37   78   38   87   33   76   35   79   30   73   31   80
  37   68   39   75   46   89   49  101   40   70   42   72
  43   80   46   83   43   75   44   71   46   80   47   96
  45   92   49   80   48   70   40   90   42   85   55   76
  54   71   57   99   52   86   53   79   56   92   52   85
  50   71   59   90   50   91   52  100   58   80   57  109
;

/* Try OLS */
proc reg data=bpexample;
  model bp = age;
  title1 'OLS model fit';
output out=out1 p=pred r=resid;
run;
```



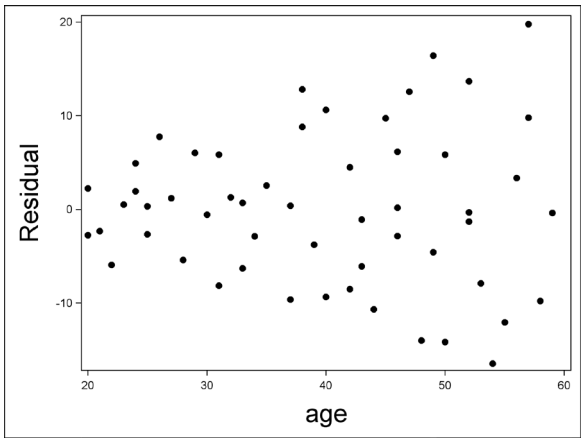

```
/* Use resid_num_diag macro from
   http://www.stat.usu.edu/jrstevens/stat5100/resid_num_diag_1line.sas
 */
 %macro resid_num_diag(dataset,datavar, ...

%resid_num_diag(dataset=out1, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');
run;
```

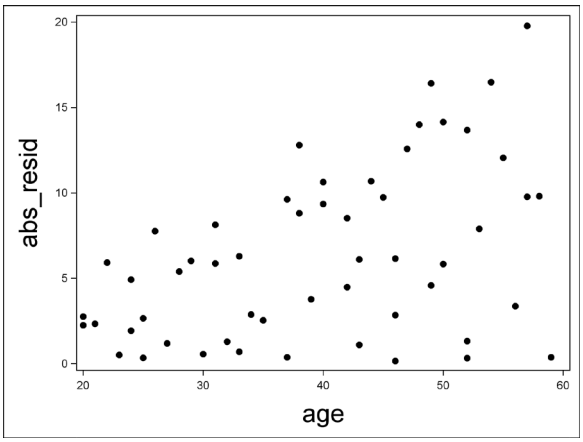
P-value for Brown-Forsythe test of constant variance in residual vs. predicted

Obs	t_BF	BF_pvalue
1	2.78547	.007440565

```
/* Look for relationship between SD of resid and X */
data out1; set out1;
  abs_resid = abs(resid);
proc sgplot data=out1;
  scatter x=age y=resid / markerattrs=(symbol=CIRCLEFILLED);
  xaxis labelattrs=(size=20pt);
```

```
  yaxis labelattrs=(size=20pt);
run;
proc sgplot data=out1;
  scatter x=age y=abs_resid / markerattrs=(symbol=CIRCLEFILLED);
  xaxis labelattrs=(size=20pt);
  yaxis labelattrs=(size=20pt);
run;
```





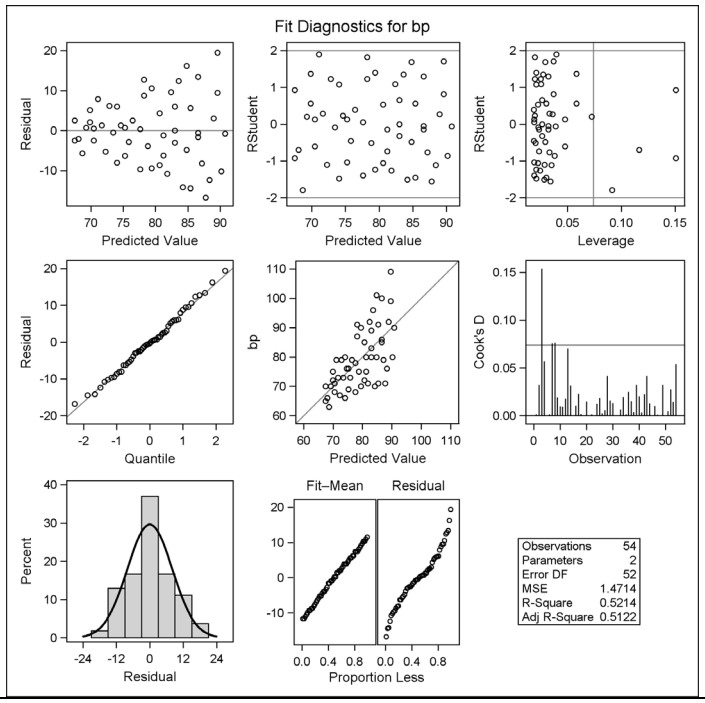

```
/* Get estimate of SD of resid based on X */
proc reg data=out1 noprint;
  model abs_resid = age;
  output out=out2 p=estSD;
run;

/* Define weight */
data out2; set out2;
  useWeight = 1/estSD**2;
run;

/* Fit WLS model */
proc reg data=out2;
  model bp = age;
  weight useWeight;
  title1 'WLS model fit';
run;
```

WLS model fit

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.56577	2.52092	22.04	<.0001

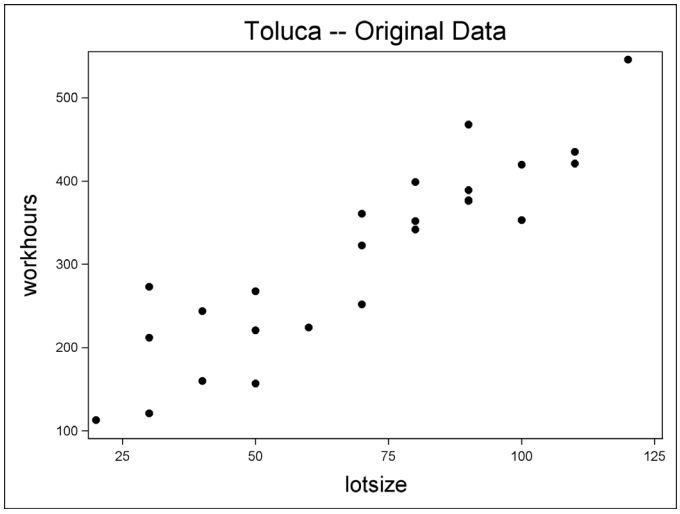
age	**1**	0.59634	0.07924	7.53	<.0001

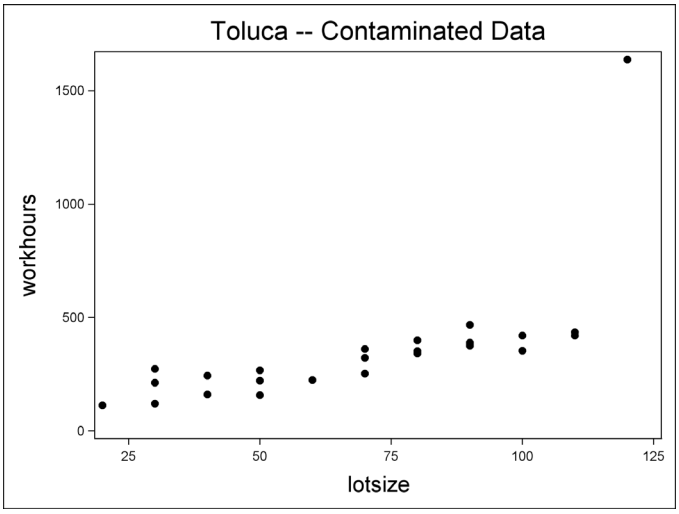


Example 2: (IRLS; recall Handout #2 example) As part of a cost improvement program, the Toluca company wished to better understand the relationship between the lot size (X) and the total work hours (Y).

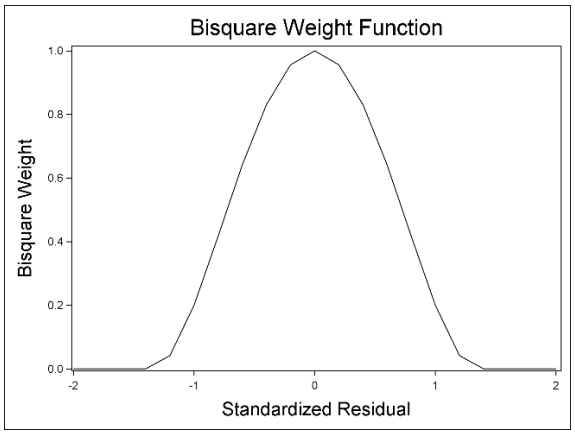
```
/* Input data -- recall Ch. 1 example */
data toluca; input lotsize workhours @@; cards;
   80   399    30   121    50   221    90   376    70   361    60   224
  120   546    80   352   100   353    50   157    40   160    70   252
   90   389    20   113   110   435   100   420    30   212    50   268
   90   377   110   421    30   273    90   468    40   244    80   342
   70   323
  ;
run;

/* Look at original data */
proc sgplot data=toluca;
  scatter x=lotsize y=workhours / markerattrs=(symbol=CIRCLEFILLED);
  xaxis labelattrs=(size=15pt);
  yaxis labelattrs=(size=15pt);
  title1 height=2 'Toluca -- Original Data';
run;

/* To show effect of robust regression, look at
   'contaminated' data */
data contam; set toluca;
  if workhours > 500 then workhours = workhours*3;
```

```
proc sgplot data=contam;
  scatter x=lotsize y=workhours / markerattrs=(symbol=CIRCLEFILLED);
  xaxis labelattrs=(size=15pt);
  yaxis labelattrs=(size=15pt);
  title1 height=2 'Toluca -- Contaminated Data';
run;
```






```
/* Look at shape of bisquare weighting curve */
data temp; input u @@; cards;
 -2.0 -1.8 -1.6 -1.4 -1.2
 -1.0 -0.8 -0.6 -0.4 -0.2
    0  0.2  0.4  0.6  0.8
  1.0 1.2 1.4 1.6 1.8 2.0
 ;
data temp; set temp;
 c = 1.345;
 w = (1-(u/c)**2)**2;
 if abs(u) >= c then w = 0;
proc sgplot data=temp;
 series x=u y=w;
 yaxis label='Bisquare Weight' labelattrs=(size=15pt);
 xaxis label='Standardized Residual' labelattrs=(size=15pt);
 title1 height=2 'Bisquare Weight Function';
run;
```




```
/* OLS regression on original data */
proc reg data=toluca;
 model workhours = lotsize;
 output out=out2 p=pred2;
 title1 'Regression on original data';
run;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
lotsize	1	3.57020	0.34697	10.29	<.0001

```
/* OLS regression on response-contaminated data */
proc reg data=contam;
 model workhours = lotsize;
 output out=out3 p=pred3;
 title1 'Regression on response-contaminated data';
run;
```

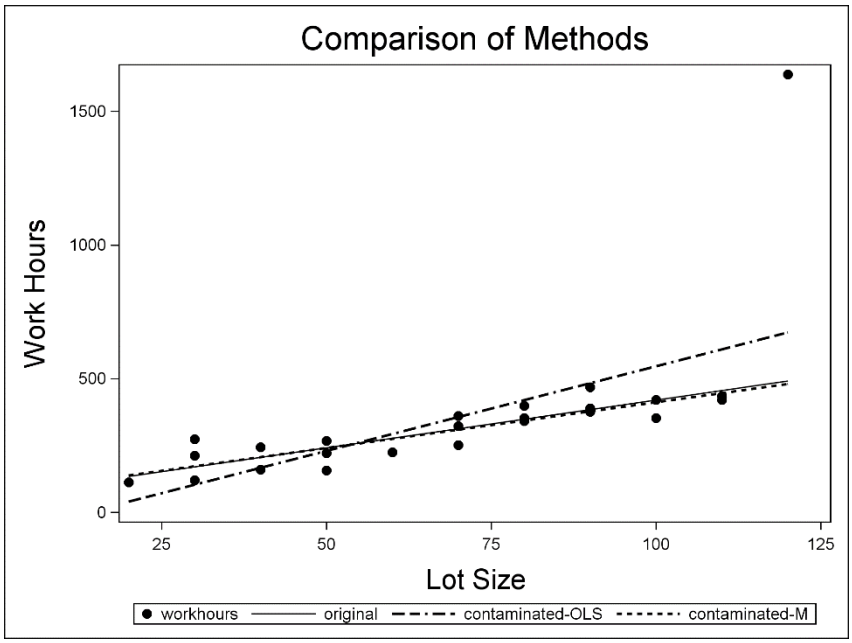
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-86.98444	120.90818	-0.72	0.4791
lotsize	1	6.32778	1.60259	3.95	0.0006

```
/* Robust (M) regression on response-contaminated data */
proc robustreg data=contam method=M (wf=bisquare);
 model workhours = lotsize;
 output out=out4 p=pred4;
 title1 'Robust (M) regression on response-contaminated data';
run;
```

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	69.2426	27.3941	15.5511	122.9340	6.39	0.0115
lotsize	1	3.4207	0.3631	2.7091	4.1324	88.75	<.0001
Scale	1	56.2335					

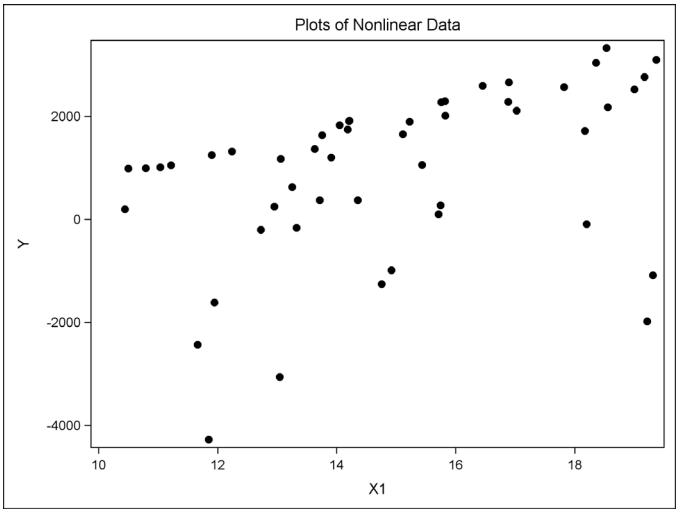
```
/* Visualize comparison of methods */
data out2; set out2; keep pred2;
data out3; set out3; keep pred3;
data out4; set out4; keep pred4;
data comp; merge contam out2 out3 out4;
  label pred2 = 'original'
```

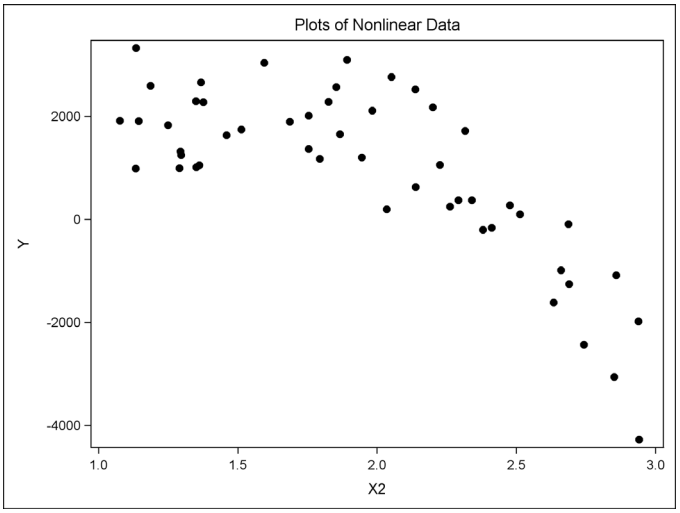
```
        pred3 = 'contaminated-OLS'
        pred4 = 'contaminated-M';
proc sort data=comp;  by lotsize;
proc sgplot data=comp;
 scatter x=lotsize y=workhours /
   markerattrs=(symbol=CIRCLEFILLED);
 series x=lotsize y=pred2 /
   lineattrs=(pattern=1
              thickness=1);
 series x=lotsize y=pred3 /
   lineattrs=(pattern=14
              thickness=2);
 series x=lotsize y=pred4 /
   lineattrs=(pattern=2
              thickness=2);
 xaxis label='Lot Size'
   labelattrs=(size=15pt);
 yaxis label='Work Hours'
   labelattrs=(size=15pt);
 title1 height=2
  'Comparison of Methods';
run;
```



Example 3.1: (Nonlinear Regression) Suppose $Y = \beta_0 + \beta_1 X_1^{\beta 2} - \beta_3 \exp(\beta_4 X_2) + \varepsilon$

```
/* Generate random data */
data temp;
 do i=1 to 50;
   X1 = 10+10*uniform(i);
   X2 = 1+2*uniform(i+2);
   error = 10*normal(2*i);
   output;
 end;
run;
/* uniform(A) --> U[0,1]
   normal(A) --> N(0,1)
   with seed A
 */

/* Define relation */
data temp1; set temp;
 Y=50+10*X1**2-16*exp(2*X2)+error;
run;

/* Look at plots */
proc sgplot data=temp1;
 scatter x=X1 y=Y /
markerattrs=(symbol=CIRCLEFILLED);
 title1 'Plots of Nonlinear Data';
run;
proc sgplot data=temp1;
 scatter x=X2 y=Y /
markerattrs=(symbol=CIRCLEFILLED);
run;
```






```
 /* Try proc nlin using the default loss function.
    The result would be the same if the pred and _LOSS_
    lines were deleted from the code. */
 proc nlin data=temp1 noitprint maxiter=500;
  pred = b0 + b1*X1**b2 + b3*exp(b4*X2);
  _LOSS_ = (Y-pred)**2;
  model Y = b0 + b1*X1**b2 + b3*exp(b4*X2);
  parameters b0=100 b1=8 b2=3 b3=-20 b4=4;
  title1 'proc nlin with [default] squared error loss function';
  title2 'truth: b0=50, b1=10, b2=2, b3=-16, b4=2';
  output out=out1 r=resid p=pred;
 run;
 /* What if we wanted better fits for smaller predicted values? */
  *_LOSS_ = ((Y-pred)/pred)**2;
```

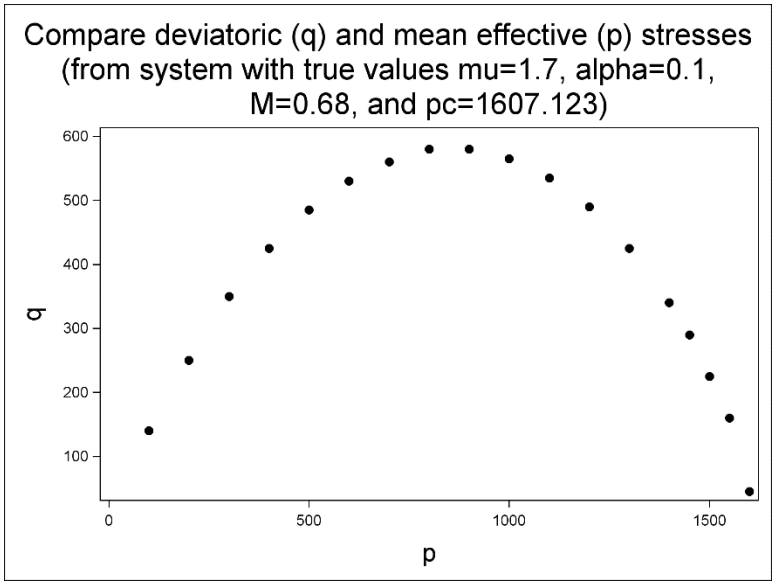
proc nlin with [default] squared error loss function
truth: b0=50, b1=10, b2=2, b3=-16, b4=2

NOTE: Convergence criterion met.

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
b0	32.9411	23.1548	-13.6950	79.5773
b1	10.1254	0.6771	8.7617	11.4891
b2	1.9970	0.0207	1.9554	2.0387
b3	-15.5777	0.2049	-15.9904	-15.1650
b4	2.0090	0.00450	1.9999	2.0180

Example 3.2: (Nonlinear Regression) A nonlinear curve to describe sand compression

```
data ex2; input p q @@; cards;
  100 140 200 250 300 350 400 425 500 485 600 530 700
  560 800 580 900 580 1000 565 1100 535 1200 490 1300
  425 1400 340 1450 290 1500 225 1550 160 1600 45
 ;
proc sgplot data=ex2;
 scatter x=p y=q / markerattrs=(symbol=CIRCLEFILLED);
 xaxis labelattrs=(size=15pt);
 yaxis labelattrs=(size=15pt);
 title1 h=2 'Compare
   deviatoric (q) and
   mean effective (p)
   stresses';
 title2 h=2 '(from system
   with true values mu=1.7,
   alpha=0.1,';
 title3 h=2 'M=0.68,
   and pc=1607.123)';
run;
```



```
proc model data=ex2;
 parms mu 1.7 alpha .2 M .7 ;
 bounds M mu > 0;
 control pc 1607.123;
 k1 = mu*(1-alpha)/(2*(1-mu)) *
                        (1+sqrt(1-4*alpha*(1-mu)/(mu*(1-alpha)**2)));
 k2 = mu*(1-alpha)/(2*(1-mu)) *
                        (1-sqrt(1-4*alpha*(1-mu)/(mu*(1-alpha)**2)));
 eq.f = p/pc - ((1+q/p/M/k2)**(k2/(1-mu)/(k1-k2)) /
                 (1+q/p/M/k1)**(k1/(1-mu)/(k1-k2)));
 fit f  / method=marquardt prl=lr corrb;
 title1 'Sand stress example';
 title2 '(truth: mu=1.7, alpha=0.1, M=0.68)';
run;

/*
   parms -- sets initial starting estimates of parameters
            to be estimated in model

   bounds -- sets boundaries on parameter values

   control -- define fixed [known] constants

   k1, k2  -- functions of parameters to be estimated

   eq.f  -- expression that equals 0 (i.e., want to find
            parameter values to make eq.f=0)

   method -- specify estimation routine

   prl=lr -- requests CI on parameter estimates

   corrb  -- requests correlation matrix among parameter estimates

 */
```

Sand stress example
(truth: mu=1.7, alpha=0.1, M=0.68)

Nonlinear OLS Parameter Estimates				
Parameter	**Estimate**	**Approx Std Err**	**t Value**	**Approx Pr > |t|**
mu	1.67184	0.0181	92.49	<.0001
alpha	0.110909	0.00762	14.56	<.0001
M	0.677976	0.00215	314.83	<.0001

Parameter Likelihood Ratio 95% Confidence Intervals			
Parameter	**Value**	**Lower**	**Upper**
mu	1.6718	1.6352	1.7061
alpha	0.1109	0.0967	0.1267
M	0.6780	0.6736	0.6821

Correlations of Parameter Estimates			
	mu	**alpha**	**M**
mu	1.0000	-0.9117	0.7978
alpha	-0.9117	1.0000	-0.8644
M	0.7978	-0.8644	1.0000