2.2.1: SAS - Residual Diagnostics

Dr. Bean – Stat 5100

Example: (The Toluca Company data from Handout #2).

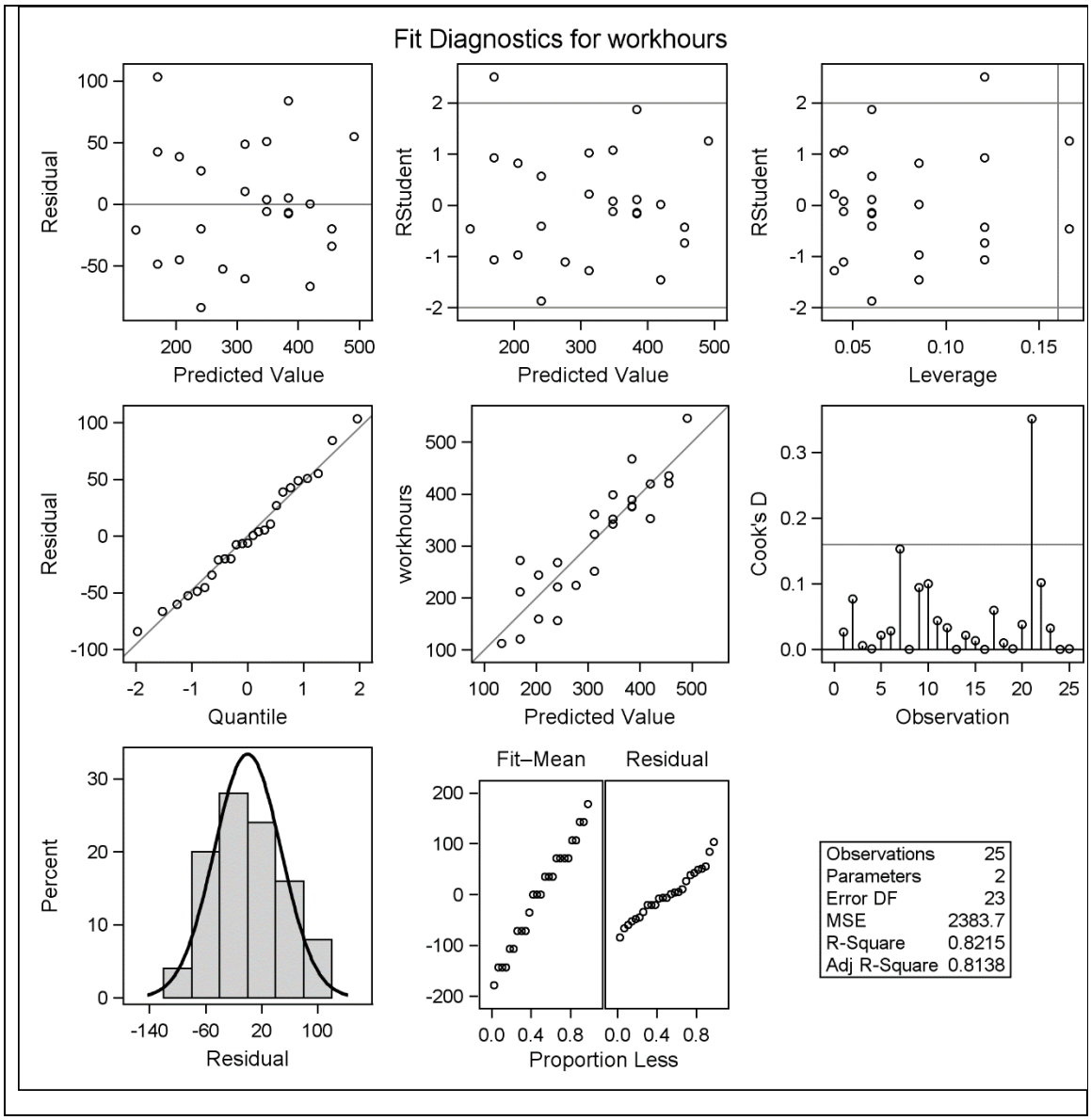
```
/* Input Toluca data (recall Ch. 1 example) */
data toluca; input lotsize workhours @@; cards;
   80  399   30  121   50  221   90  376   70  361   60  224
  120  546   80  352  100  353   50  157   40  160   70  252
   90  389   20  113  110  435  100  420   30  212   50  268
   90  377  110  421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

/* Now fit simple linear model with Y=workhours and X=lotsize,
   with residuals and predicted values saved in data set
   tolucaout */
proc reg data=toluca;
  model workhours = lotsize;
  output out=tolucaout r=resid p=pred;
  title1 'Simple linear model';
run;
```

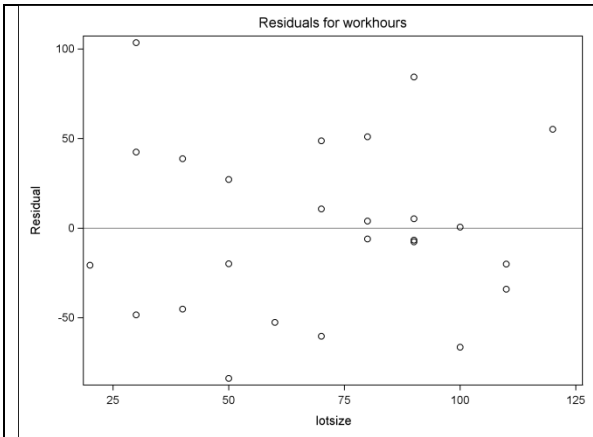
Analysis of Variance

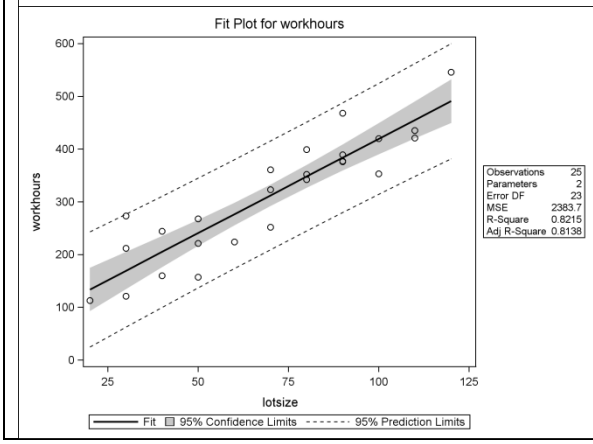
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001
Error	23	54825	2383.71562		
Corrected Total	24	307203			

Parameter Estimates

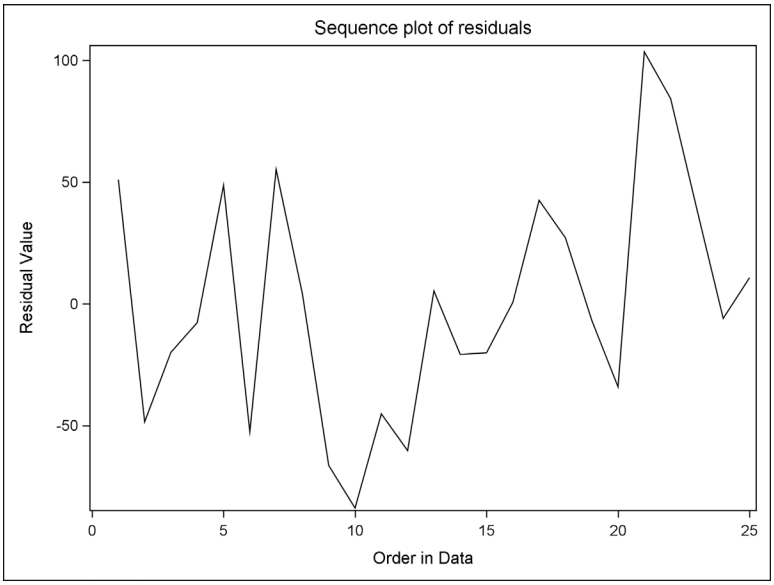
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
lotsize	1	3.57020	0.34697	10.29	<.0001


Fit Diagnostics for workhours
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
workhours
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 25
Parameters 2
Error DF 23
MSE 2383.7
R-Square 0.8215
Adj R-Square 0.8138


Residuals for workhours
Residual
lotsize
100
50
0
-50
25
50
75
100
125


Fit Plot for workhours
workhours
lotsize
600
500
400
300
200
100
0
25
50
75
100
125
Observations 25
Parameters 2
Error DF 23
MSE 2383.7
R-Square 0.8215
Adj R-Square 0.8138
Fit
95% Confidence Limits
95% Prediction Limits

```
/* Look at sequence plot */
data temp; set tolucaout;
  order = _n_;
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;
```




```
/**************  Numerical Diagnostics  *************/

/* F-test for lack of fit */
proc rsreg data=toluca;
  model workhours = lotsize / lackfit covar=1 noopt;
  title1 'F-test for lack of fit';
run;
```

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	9	17245	1916.069540	0.71	0.6893
Pure Error	14	37581	2684.345238		
Total Error	23	54825	2383.715617		

```
/** Brown-Forsythe and Correlation Test of Normality (shortcut)
 **/
 /* Two [unused] ways to access shortcut:
     filename macrourl "C:\[filepath]\resid_num_diag.sas";
     %include macrourl;
  */
%macro resid_num_diag(dataset,datavar,label= ...


/*
   This resid_num_diag.sas file provides a convenient shortcut
   to obtaining numerical checks of residuals from
   a fitted linear regression model.

   The macro takes five arguments:
     dataset is the name of the data set
     datavar is the name of the variable in the data set
             for which numerical diagnostics are desired
             (usually a residual)
     label is a character string for detail in output
     predvar is the name of the variable (usually predicted
           value) on which to sort for the Brown-Forsythe test
          (t-statistic and p-value reported)
     predlabel is the character string for detail in output
           related to the predvar variable
 */
```

```
/* Call the shortcut: */
%resid_num_diag(dataset=tolucaout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');
```

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

Obs	t_BF	BF_pvalue
1	1.31648	0.20098

Output for correlation test of normality of residual (Check text Table B.6 for threshold)

Pearson Correlation Coefficients, N = 25
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid residual	1.00000	0.99151 <.0001
expectNorm	0.99151 <.0001	1.00000