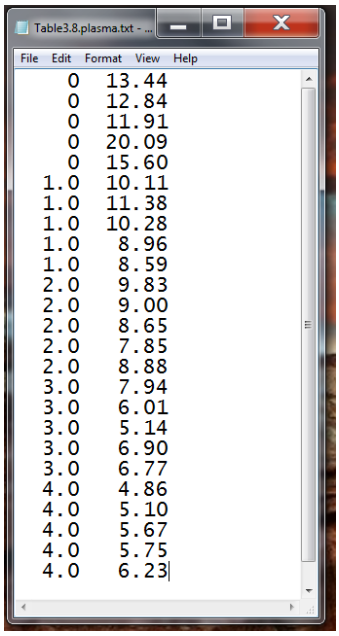
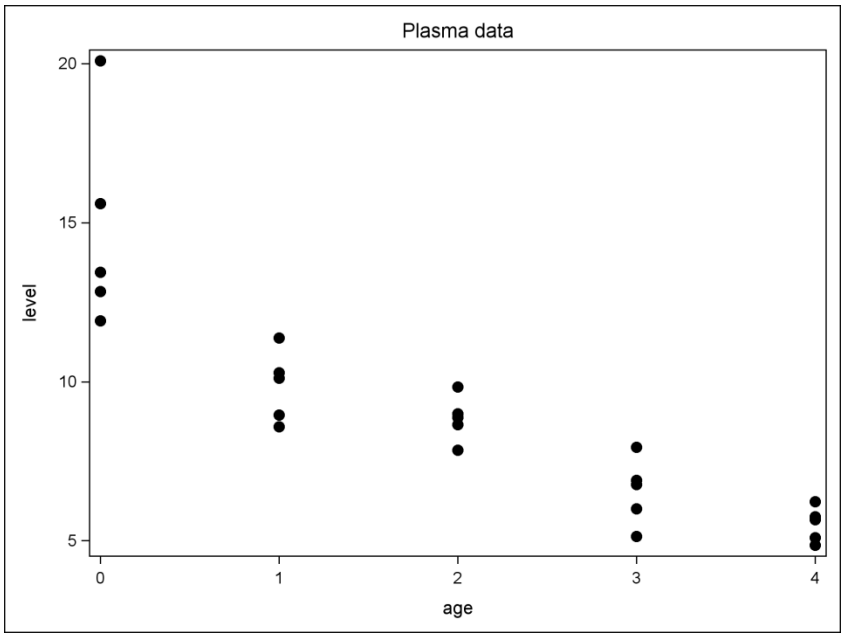
2.2.2: SAS - Linear Regression Remedial Measures

Dr. Bean – Stat 5100

Example: Age and plasma level for 25 healthy children in a study are reported. Of interest is how plasma level depends on age. (Text Table 3.8 – first column is age; second column is plasma level)





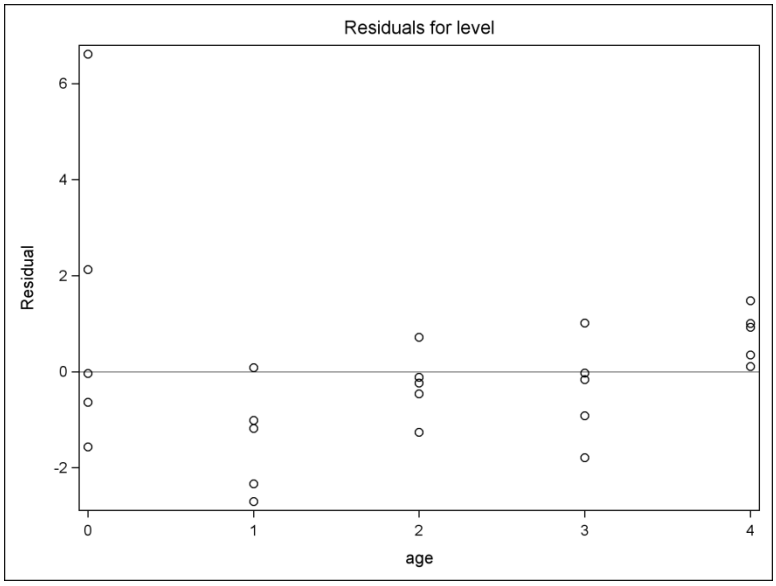

```
/*
data plasma;
   infile "[File Path]/Table3.8.plasma.txt";
   input age level;
run;
*/
data plasma; input age level @@; cards;
   0  13.44    0  12.84    0  11.91    0  20.09    0  15.60
 1.0  10.11  1.0  11.38  1.0  10.28  1.0   8.96  1.0   8.59
 2.0   9.83  2.0   9.00  2.0   8.65  2.0   7.85  2.0   8.88
 3.0   7.94  3.0   6.01  3.0   5.14  3.0   6.90  3.0   6.77
 4.0   4.86  4.0   5.10  4.0   5.67  4.0   5.75  4.0   6.23
  ;

/* Fit regression model and check assumptions */
proc reg data=plasma;
  model level = age;
  output out=out1 r=resid p=pred;
  title1 'Simple model for plasma data';
run;
```

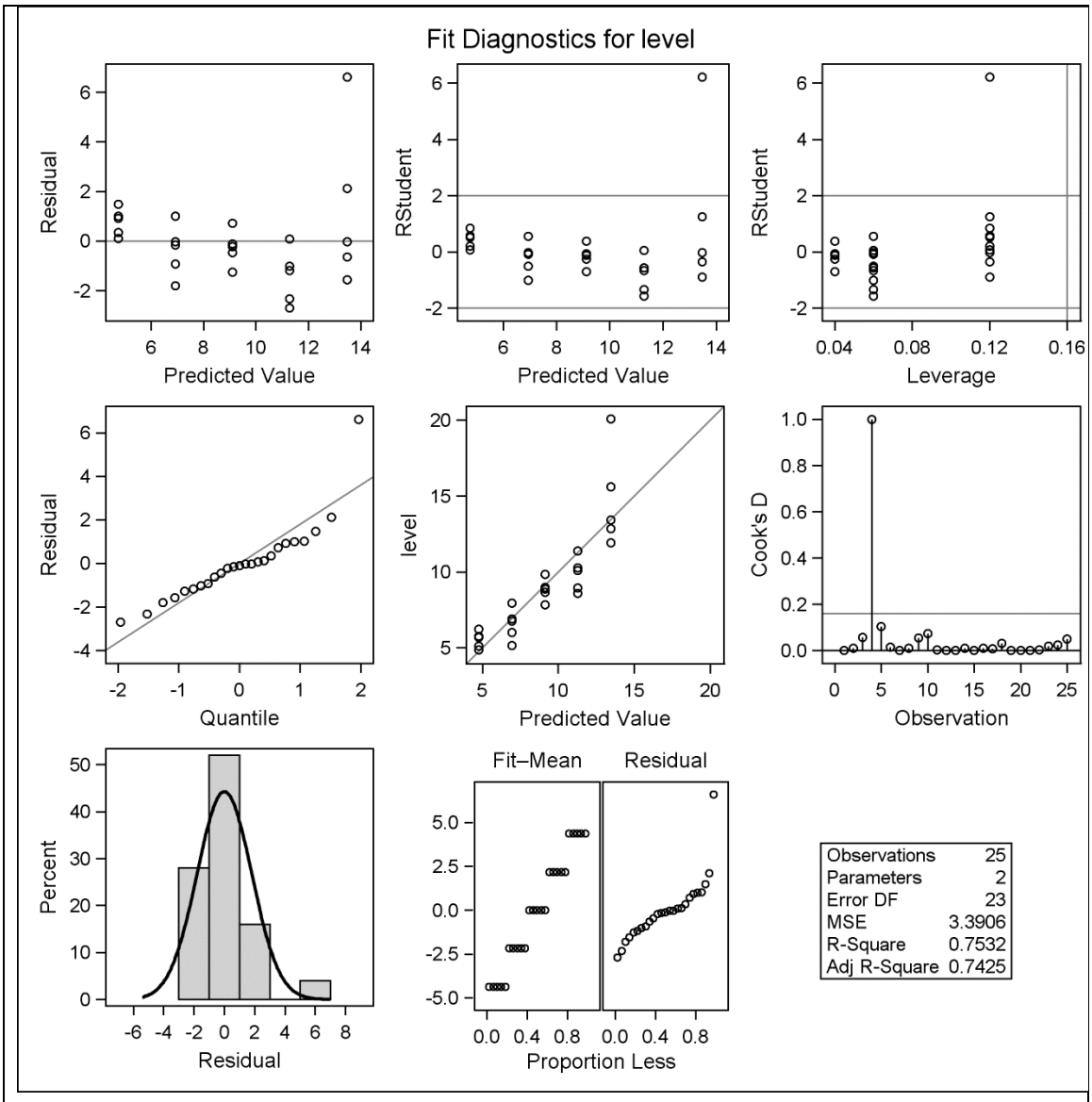
Simple model for plasma data

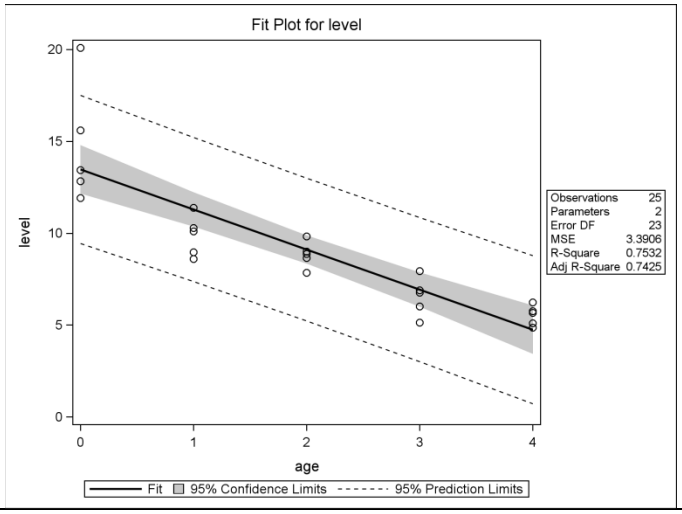
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	238.05620	238.05620	70.21	<.0001
Error	23	77.98306	3.39057		
Corrected Total	24	316.03926			

Root MSE	1.84135	**R-Square**	0.7532
Dependent Mean	9.11120	**Adj R-Sq**	0.7425
Coeff Var	20.20974		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	13.47520	0.63786	21.13	<.0001
age	**1**	-2.18200	0.26041	-8.38	<.0001




Fit Diagnostics for level
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
level
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 25
Parameters 2
Error DF 23
MSE 3.3906
R-Square 0.7532
Adj R-Square 0.7425


Fit Plot for level
level
age
Observations 25
Parameters 2
Error DF 23
MSE 3.3906
R-Square 0.7532
Adj R-Square 0.7425
Fit
95% Confidence Limits
95% Prediction Limits

```
/* Define shortcut macro, using line copied from
     [File Path]/resid_num_diag_1line.sas
 */
%macro resid_num_diag(dataset,...
/* Call shortcut macro */
%resid_num_diag(dataset=out1, datavar=resid, label='Residual',
   predvar=pred, predlabel='Predicted Value');
```

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

Obs	t_BF	BF_pvalue
1	1.50583	0.14572

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

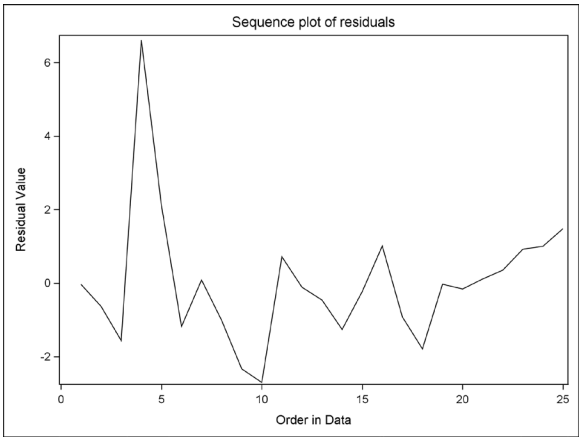
Pearson Correlation Coefficients, N = 25
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid Residual	1.00000	0.90360 <.0001
expectNorm	0.90360 <.0001	1.00000

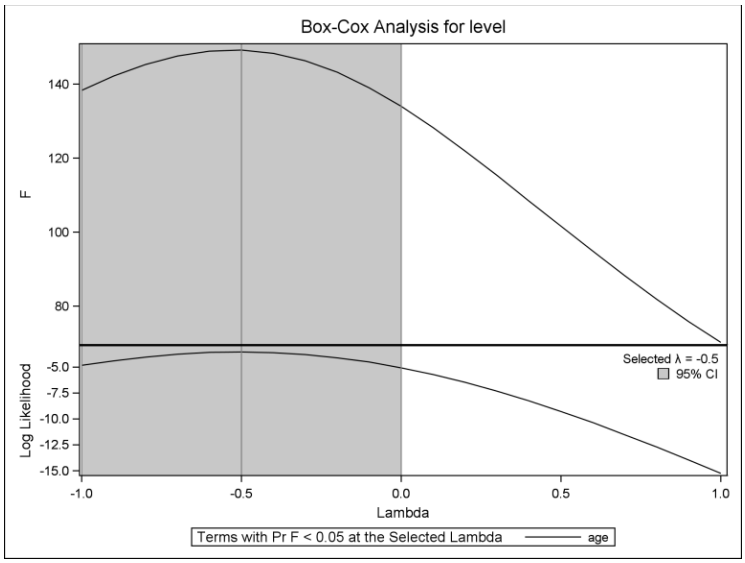
```
/* F-test for lack of fit
Options:
Covar=1 - specifies the first variable on the right
          Hand side of the equation is linear
          (default is quadratic)
Noopt   - suppresses rsreg output not associated
          with the F-test for lack of fit */
proc rsreg data=plasma;
  model level = age / 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	3	22.748784	7.582928	2.75	0.0699
Pure Error	20	55.234280	2.761714		
Total Error	23	77.983064	3.390568		

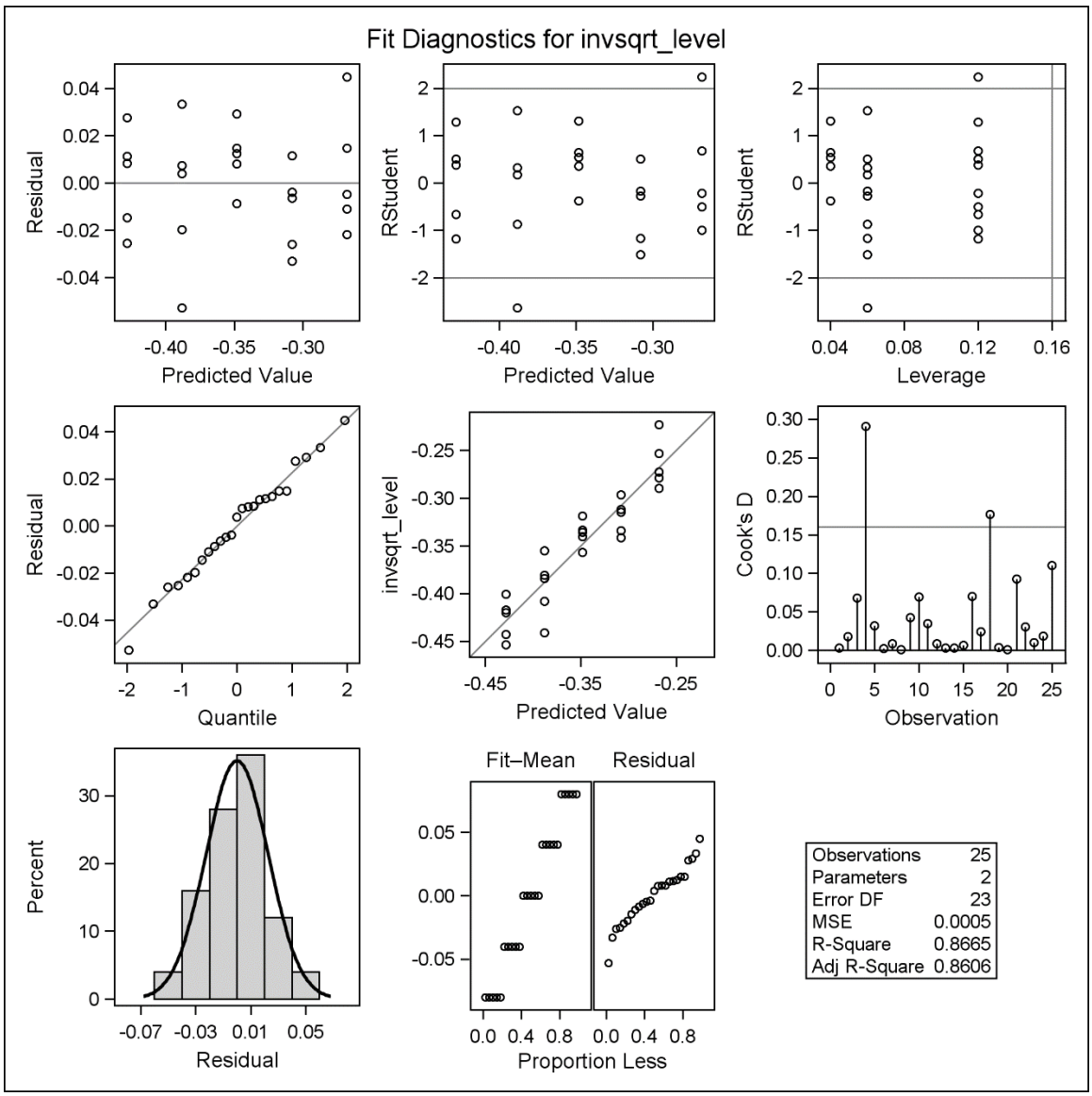
```
/* Look at sequence plot */
data temp; set out1;
  order = _n_;
proc sgplot data=temp;
  series x=order y=resid / lineattrs=(pattern=solid) ;
  xaxis label='Order in Data';
  yaxis label='Residual Value';
  title1 'Sequence plot of residuals';
run;
```




```
/********** Consider Transformations ***************/

proc transreg data=plasma;
 model boxcox(level / lambda=-1 to 1 by 0.1)
       = identity(age);
 title1 'Box-Cox Transformation: Regressing Level on Age';
run;
```




```
data plasma; set plasma;
  log_level = log(level);
  invsqrt_level = -1/sqrt(level);
run;

/* Inverse square root */
proc reg data=plasma;
  model invsqrt_level = age;
  output out=out2 r=resid p=pred;
  title1 'Simple model for negative inverse root plasma
data';
run;
```


Fit Diagnostics for invsqrt_level
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
invsqrt_level
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 25
Parameters 2
Error DF 23
MSE 0.0005
R-Square 0.8665
Adj R-Square 0.8606

```
%resid_num_diag(dataset=out2, datavar=resid,
   label='Residual (neg. inverse root)',
   predvar=pred, predlabel='Predicted Value (neg. inverse
root)');
```

P-value for Brown-Forsythe test of constant variance in Residual (neg. inverse root) vs. Predicted Value (neg. inverse root)

Obs	t_BF	BF_pvalue
1	0.16654	0.86918

Output for correlation test of normality of Residual (neg. inverse root) (Check text Table B.6 for threshold)

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

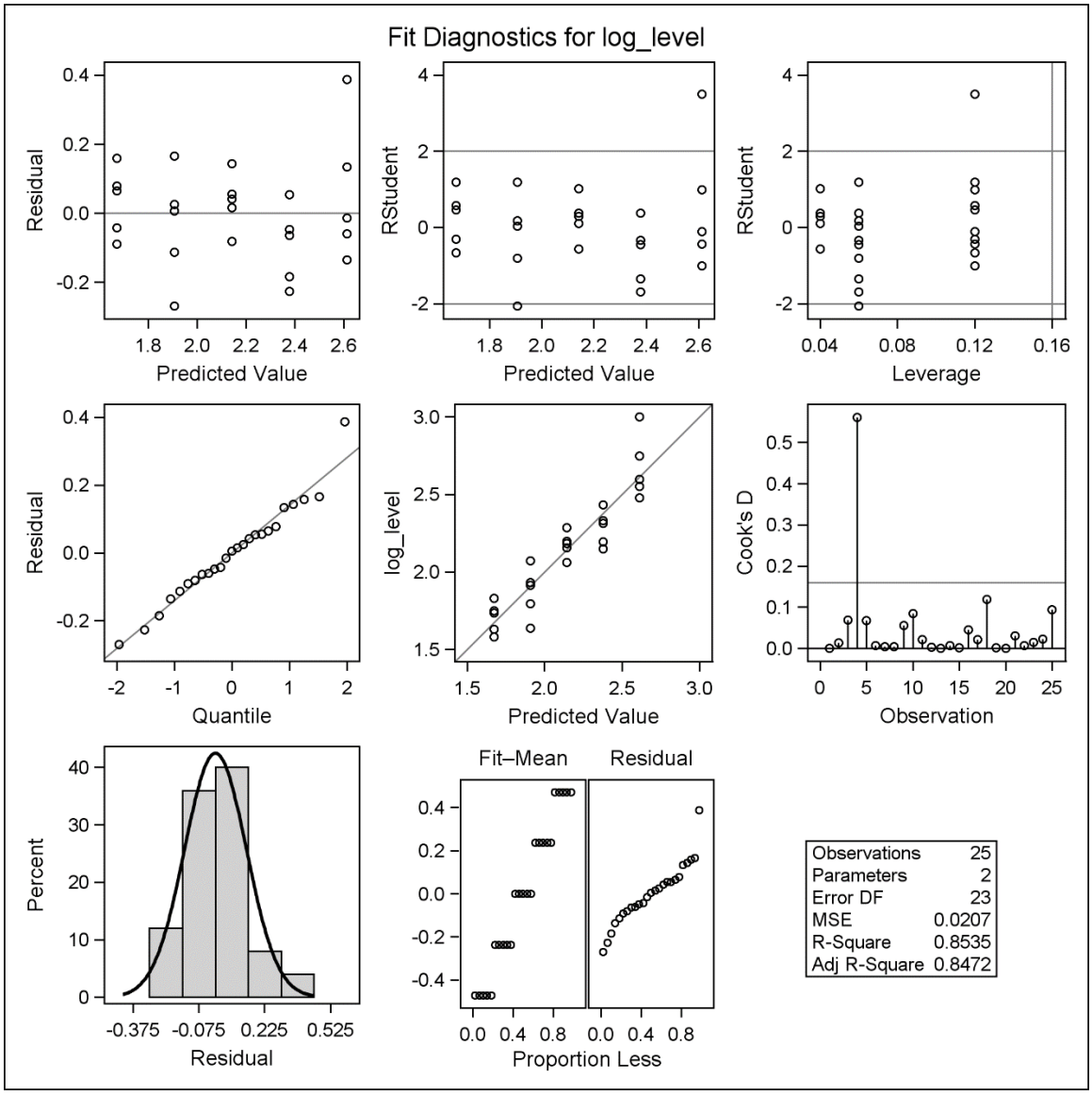
	resid	expectNorm
resid Residual (neg. inverse root)	1.00000	0.99188 <.0001
expectNorm	0.99188 <.0001	1.00000

```
proc rsreg data=plasma;
  model invsqrt_level = age / lackfit covar=1 noopt;
  title1 'F-test for lack of fit (neg. inverse root)';
run;
```

F-test for lack of fit (neg. inverse root)

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	0.001556	0.000519	0.96	0.4312
Pure Error	20	0.010813	0.000541		
Total Error	23	0.012369	0.000538		

```
/* Log */
proc reg data=plasma;
  model log_level = age;
  output out=out3 r=resid p=pred;
  title1 'Simple model for log plasma data';
run;
```



```
%resid_num_diag(dataset=out3, datavar=resid,
label='Residual (log)',
   predvar=pred, predlabel='Predicted Value (log)');
```

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

Obs	t_BF	BF_pvalue
1	0.95179	0.35110

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

Pearson Correlation Coefficients, N = 25 Prob > |r| under H0: Rho=0		
	resid	expectNorm
resid Residual (log)	1.00000	0.98071 <.0001
expectNorm	0.98071 <.0001	1.00000

```
proc rsreg data=plasma;
  model log_level = age / lackfit covar=1 noopt;
  title1 'F-test for lack of fit (log)';
run;
```

F-test for lack of fit (log)

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	0.081944	0.027315	1.39	0.2758
Pure Error	20	0.394004	0.019700		
Total Error	23	0.475948	0.020693		

```
/* Probably go with inverse square root */
proc reg data=plasma;
  model invsqrt_level = age;
  title1 'Negative inverse root plasma data';
run;
```

Negative inverse root plasma data

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	0.08025	0.08025	149.22	<.0001
Error	23	0.01237	0.00053778		
Corrected Total	24	0.09262			

Root MSE	0.02319	**R-Square**	0.8665

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	-0.26803	0.00803	-33.36	<.0001
age	**1**	-0.04006	0.00328	-12.22	<.0001