Handout 1: Predicting GPA from SAT

```
appsrv01.srv.cquest.utoronto.ca>
appsrv01.srv.cquest.utoronto.ca> ls
Desktop  grades.data  grades.sas  oldstuff  sasuser.800
appsrv01.srv.cquest.utoronto.ca> cat grades.data
      verbal   math    gpa
   1     623    509    2.6
   2     454    471    2.3
   3     643    700    2.4
   4     585    719    3.0
   5     719    710    3.1
   6     693    643    2.9
   7     571    665    3.1
   8     646    719    3.3
   9     613    693    2.3
  10     655    701    3.3

   . . . Skipping . . .

 195     710    647    3.0
 196     509    538    3.0
 197     480    526    2.4
 198     487    672    2.9
 199     526    796    1.8
 200     532    710    2.1
appsrv01.srv.cquest.utoronto.ca> ls
Desktop  grades.data  grades.sas  oldstuff  sasuser.800
appsrv01.srv.cquest.utoronto.ca> cat grades.sas
/********************* grades.sas *************************/
options linesize=79 noovp formdlim=' ';
title 'Predict First-Year GPA from SAT Scores';

data sat;
     infile 'grades.data' firstobs=2 ;     /* Skipping the header on line 1 */
     input id verbal math gpa;
     sat = verbal+math;
     label gpa = 'First-year GPA'
           sat = 'Total SAT score';

proc means;
     var verbal math sat gpa;

proc univariate normal plot;
     var verbal math sat gpa;

proc plot;
     title2 'Rough Scatterplot of SAT vs GPA';
     plot gpa*sat;

proc corr;
     title2 'Product-moment Correlations (Normal Assumption for tests)';
     var verbal math sat gpa;

proc corr spearman nosimple;
     title2 'Rank Correlations (Non-parametric)';
     var verbal math sat gpa;
```

```
proc reg;
     title2 'Simple Regression';
     model gpa=sat;

proc glm;
     title2 'Specific predictions, the easy way';
     model gpa=sat;
     estimate 'For SAT=1400' intercept 1 sat 1400;
     estimate 'For SAT=1500' intercept 1 sat 1500;
     estimate 'For SAT=1600' intercept 1 sat 1600;

proc reg;
     title2 'Multiple Regression';
     model gpa = verbal math;

proc glm;
     title2 'Specific predictions for multiple regression';
     model gpa = verbal math;
     estimate 'Verbal=600, Math=600' intercept 1 verbal 600 math 600;
     estimate 'Verbal=400, Math=800' intercept 1 verbal 400 math 800;

proc reg; /* Could have said   proc reg noprint */
     title2 'Save residuals for further analysis';
     model gpa = verbal math;
     output out=sat2 residual=gparesid;
     /* Creates new SAS data set called sat2. It is just like the original
        data set sat, but it also has a new variable called gparesid, the
        residuals from the regression we just did. By default, SAS
        procedures use the most recently created data set, so it is easy to
        do further analysis of the residuals. */

proc univariate normal plot;
     title2 'Examine residuals, test for normality';
     var gparesid;

/* Some issues to consider:

     Should we be throwing out variables?
     Should we be throwing out cases (outliers)?
     Omitted variables
```

```
appsrv01.srv.cquest.utoronto.ca> ls
Desktop  grades.data  grades.sas  oldstuff  sasuser.800
appsrv01.srv.cquest.utoronto.ca>
appsrv01.srv.cquest.utoronto.ca> sas grades
appsrv01.srv.cquest.utoronto.ca> ls
Desktop      grades.log  grades.sas  sasuser.800
grades.data  grades.lst  oldstuff    sasuser.v91

appsrv01.srv.cquest.utoronto.ca> less grades.log

1                                                          The SAS System
                    09:32 Sunday, September 16, 2007

NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M3)
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987
001.
NOTE: This session is executing on the Linux 2.6.18-8.1.8.el5PAE platform.



NOTE: SAS 9.1.3 Service Pack 3

You are running SAS 9. Some SAS 8 files will be automatically converted
by the V9 engine; others are incompatible.  Please see
http://support.sas.com/rnd/migration/planning/platform/64bit.html

PROC MIGRATE will preserve current SAS file attributes and is
recommended for converting all your SAS libraries from any
SAS 8 release to SAS 9.  For details and examples, please see
http://support.sas.com/rnd/migration/index.html


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the file "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           1.06 seconds
      cpu time            0.07 seconds

1          /********************* grades.sas *************************/
2          options linesize=79 noovp formdlim=' ';
3          title 'Predict First-Year GPA from SAT Scores';
4
5          data sat;
6               infile 'grades.data' firstobs=2 ;     /* Skipping the header on
6        !  line 1 */
7               input id verbal math gpa;
8               sat = verbal+math;
9               label gpa = 'First-year GPA'
10                    sat = 'Total SAT score';
11
```

```
NOTE: The infile 'grades.data' is:
      File Name=/homes/staff/u0/stats/brunner/grades.data,
      Owner Name=brunner,Group Name=stats,
      Access Permission=rw-r--r--,
      File Size (bytes)=5427

NOTE: 200 records were read from the infile 'grades.data'.
      The minimum record length was 26.
      The maximum record length was 26.
NOTE: The data set WORK.SAT has 200 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.21 seconds
      cpu time            0.02 seconds

^L2                              The SAS System
                                            09:32 Sunday, September 16, 2007


12         proc means;
13              var verbal math sat gpa;
14

NOTE: There were 200 observations read from the data set WORK.SAT.
NOTE: The PROCEDURE MEANS printed page 1.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.53 seconds
      cpu time            0.04 seconds


15         proc univariate normal plot;
16              var verbal math sat gpa;
17

NOTE: The PROCEDURE UNIVARIATE printed pages 2-12.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.09 seconds
      cpu time            0.04 seconds


18         proc plot;
19              title2 'Rough Scatterplot of SAT vs GPA';
20              plot gpa*sat;
21

NOTE: There were 200 observations read from the data set WORK.SAT.
NOTE: The PROCEDURE PLOT printed page 13.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.02 seconds
      cpu time            0.00 seconds


22         proc corr;
23              title2 'Product-moment Correlations (Normal Assumption for
23       ! tests)';
24              var verbal math sat gpa;
25
```

```
NOTE: The PROCEDURE CORR printed page 14.
NOTE: PROCEDURE CORR used (Total process time):
      real time           0.07 seconds
      cpu time            0.02 seconds


26         proc corr spearman nosimple;
27              title2 'Rank Correlations (Non-parametric)';
28              var verbal math sat gpa;
29

NOTE: The PROCEDURE CORR printed page 15.
NOTE: PROCEDURE CORR used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


^L3                              The SAS System
                                              09:32 Sunday, September 16, 2007

30         proc reg;
31              title2 'Simple Regression';
32              model gpa=sat;
33

NOTE: The PROCEDURE REG printed page 16.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.20 seconds
      cpu time            0.03 seconds


34         proc glm;
35              title2 'Specific predictions, the easy way';
36              model gpa=sat;
37              estimate 'For SAT=1400' intercept 1 sat 1400;
38              estimate 'For SAT=1500' intercept 1 sat 1500;
39              estimate 'For SAT=1600' intercept 1 sat 1600;
40

NOTE: The PROCEDURE GLM printed pages 17-18.
NOTE: PROCEDURE GLM used (Total process time):
      real time           0.09 seconds
      cpu time            0.01 seconds


41         proc reg;
42              title2 'Multiple Regression';
43              model gpa = verbal math;
44

NOTE: The PROCEDURE REG printed page 19.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.01 seconds
      cpu time            0.03 seconds
```

```
45         proc glm;
46              title2 'Specific predictions for multiple regression';
47              model gpa = verbal math;
48              estimate 'Verbal=600, Math=600' intercept 1 verbal 600 math 600
48       ! ;
49              estimate 'Verbal=400, Math=800' intercept 1 verbal 400 math 800
49       ! ;
50
```

```
NOTE: The PROCEDURE GLM printed pages 20-21.
NOTE: PROCEDURE GLM used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

```
51         proc reg; /* Could have said   proc reg noprint */
52              title2 'Save residuals for further analysis';
53              model gpa = verbal math;
54              output out=sat2 residual=gparesid;
55              /* Creates new SAS data set called sat2. It is just like the
55       ! original
^L4                               The SAS System
                                                  09:32 Sunday, September 16, 2007

56                   data set sat, but it also has a new variable called
56       ! gparesid, the
57                   residuals from the regression we just did. By default, SAS
58                   procedures use the most recently created data set, so it is
58       ! easy to
59                   do further analysis of the residuals. */
60
```

```
NOTE: The data set WORK.SAT2 has 200 observations and 6 variables.
NOTE: The PROCEDURE REG printed page 22.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.02 seconds
      cpu time            0.02 seconds
```

```
61         proc univariate normal plot;
62              title2 'Examine residuals, test for normality';
63              var gparesid;
64
65         /* Some issues to consider:
66
67              Should we be throwing out variables?
68              Should we be throwing out cases (outliers)?
69              Omitted variables
NOTE: The PROCEDURE UNIVARIATE printed pages 23-25.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

```
NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           2.65 seconds
      cpu time            0.32 seconds
```

```
appsrv01.srv.cquest.utoronto.ca> cat grades.lst
```

```
                  Predict First-Year GPA from SAT Scores                    1
                                        09:32 Sunday, September 16, 2007

                            The MEANS Procedure

 Variable   Label                N           Mean        Std Dev        Minimum
 ------------------------------------------------------------------------------
 verbal                          200    595.6500000     73.2098759    361.0000000
 math                            200    649.5300000     66.3471084    441.0000000
 sat        Total SAT score      200        1245.18    111.4879609    925.0000000
 gpa        First-year GPA       200      2.6300000      0.5803309      0.3000000
 ------------------------------------------------------------------------------

                  Variable   Label                    Maximum
                  ----------------------------------------
                  verbal                         780.0000000
                  math                           800.0000000
                  sat        Total SAT score        1526.00
                  gpa        First-year GPA       3.9000000
                  ----------------------------------------
```

Skipping: Just show proc univariate output for gpa

```
                  Predict First-Year GPA from SAT Scores                   10
                                        09:32 Sunday, September 16, 2007

                        The UNIVARIATE Procedure
                      Variable:  gpa  (First-year GPA)

                                 Moments

      N                          200    Sum Weights               200
      Mean                      2.63    Sum Observations          526
      Std Deviation      0.58033087    Variance           0.33678392
      Skewness           -0.3927091    Kurtosis           0.58666505
      Uncorrected SS         1450.4    Corrected SS            67.02
      Coeff Variation    22.0658126    Std Error Mean     0.04103559


                      Basic Statistical Measures

            Location                      Variability

        Mean     2.630000     Std Deviation            0.58033
        Median   2.600000     Variance                 0.33678
        Mode     2.300000     Range                    3.60000
                              Interquartile Range      0.70000
```

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	64.09071	Pr > |t|	<.0001
Sign	M	100	Pr >= |M|	<.0001
Signed Rank	S	10050	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.978372	Pr < W	0.0035
Kolmogorov-Smirnov	D	0.079124	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.173051	Pr > W-Sq	0.0123
Anderson-Darling	A-Sq	1.05625	Pr > A-Sq	0.0090

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3.90
99%	3.80
95%	3.55
90%	3.40
75% Q3	3.00
50% Median	2.60
25% Q1	2.30
10%	2.00
5%	1.80
1%	1.15
0% Min	0.30

The UNIVARIATE Procedure
Variable: gpa (First-year GPA)

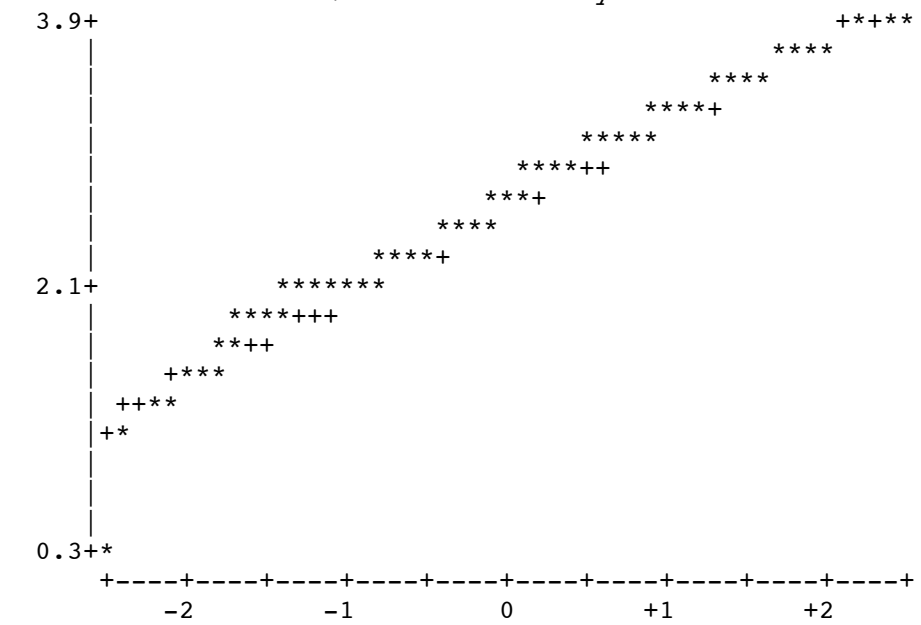
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0.3	131	3.7	35
1.1	136	3.7	174
1.2	40	3.8	62
1.3	121	3.8	105
1.4	127	3.9	79

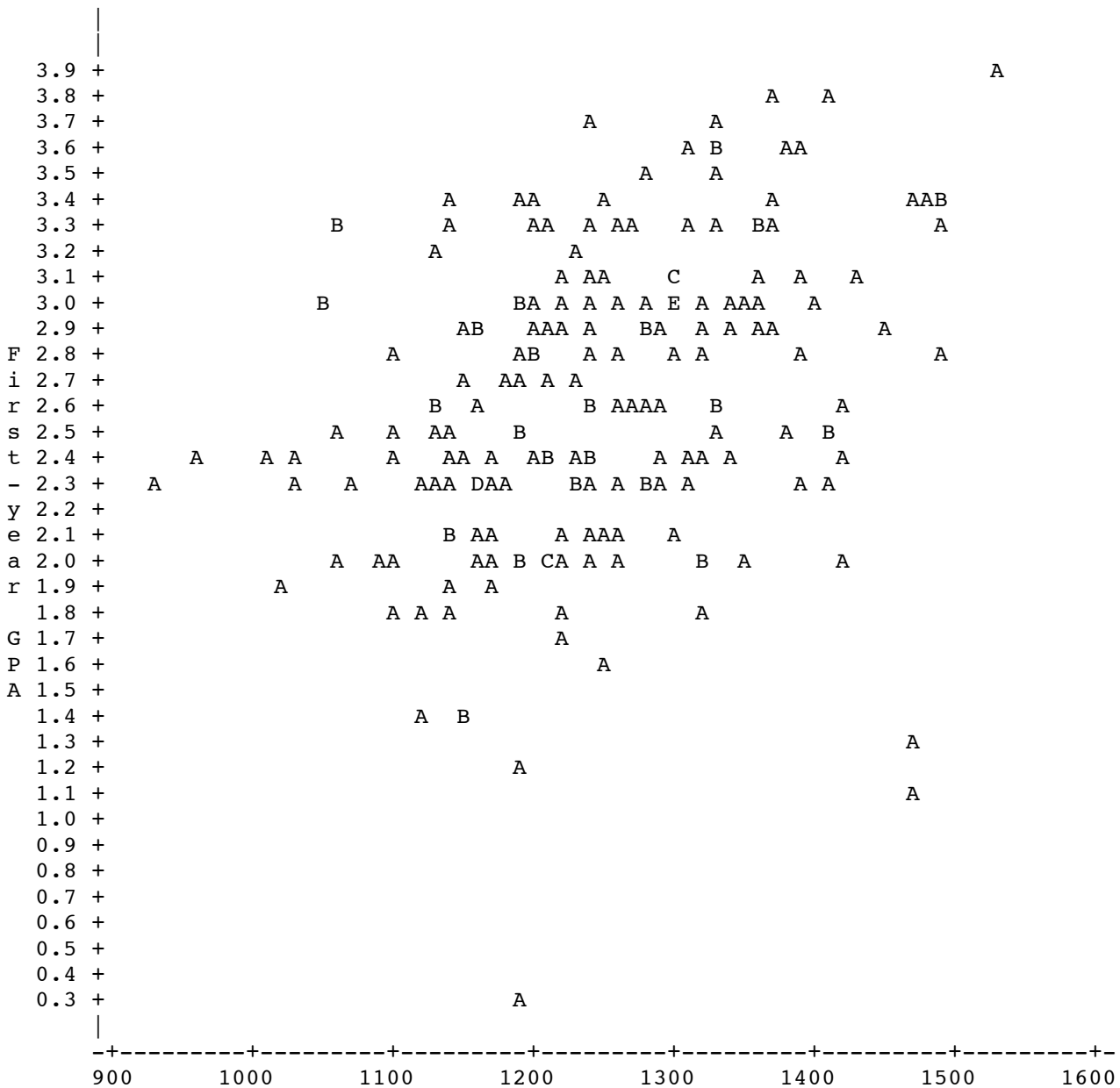
```
Stem Leaf                             #             Boxplot
  38 000                              3                |
  36 0000000                          7                |
  34 00000000000                     11                |
  32 0000000000000000                16                |
  30 00000000000000000000000000000   28             +-----+
  28 0000000000000000000000000       25             |     |
  26 00000000000000000               17             *--+--*
  24 000000000000000000000000000000  28             |     |
  22 0000000000000000000000          22             +-----+
  20 00000000000000000000000000      26                |
  18 00000000                         8                |
  16 00                               2                |
  14 000                              3                |
  12 00                               2                0
  10 0                                1                0
   8
   6
   4
   2 0                                1                0
     ----+----+----+----+----+---
 Multiply Stem.Leaf by 10**-1
```


The UNIVARIATE Procedure
Variable: gpa (First-year GPA)

Normal Probability Plot



Plot of gpa*sat. Legend: A = 1 obs, B = 2 obs, etc.



The CORR Procedure

4 Variables: verbal math sat gpa

Skipping some descriptive statistics we have already seen ...

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	verbal	math	sat	gpa
verbal	1.00000	0.27463 <.0001	0.82010 <.0001	0.32245 <.0001
math	0.27463 <.0001	1.00000	0.77545 <.0001	0.19424 0.0058
sat Total SAT score	0.82010 <.0001	0.77545 <.0001	1.00000	0.32733 <.0001
gpa First-year GPA	0.32245 <.0001	0.19424 0.0058	0.32733 <.0001	1.00000

Rank Correlations (Non-parametric)
09:32 Sunday, September 16, 2007

The CORR Procedure

4 Variables: verbal math sat gpa

Spearman Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	verbal	math	sat	gpa
verbal	1.00000	0.26580 0.0001	0.79788 <.0001	0.35320 <.0001
math	0.26580 0.0001	1.00000	0.76724 <.0001	0.22347 0.0015
sat Total SAT score	0.79788 <.0001	0.76724 <.0001	1.00000	0.37032 <.0001
gpa First-year GPA	0.35320 <.0001	0.22347 0.0015	0.37032 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: gpa First-year GPA

Number of Observations Read 200
Number of Observations Used 200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.18104	7.18104	23.76	<.0001
Error	198	59.83896	0.30222		
Corrected Total	199	67.02000			

Root MSE	0.54974	R-Square	0.1071
Dependent Mean	2.63000	Adj R-Sq	0.1026
Coeff Var	20.90276		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.50836	0.43698	1.16	0.2461
sat	Total SAT score	1	0.00170	0.00034955	4.87	<.0001

The GLM Procedure

Number of Observations Read 200
Number of Observations Used 200

The GLM Procedure

Dependent Variable: gpa First-year GPA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.18103787	7.18103787	23.76	<.0001
Error	198	59.83896213	0.30221698		
Corrected Total	199	67.02000000			

R-Square	Coeff Var	Root MSE	gpa Mean
0.107148	20.90276	0.549743	2.630000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sat	1	7.18103787	7.18103787	23.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sat	1	7.18103787	7.18103787	23.76	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
For SAT=1400	2.89379476	0.06663116	43.43	<.0001
For SAT=1500	3.06418279	0.09718439	31.53	<.0001
For SAT=1600	3.23457083	0.12997520	24.89	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.5083622486	0.43698071	1.16	0.2461
sat	0.0017038804	0.00034955	4.87	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: gpa First-year GPA

Number of Observations Read 200
Number of Observations Used 200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7.77792	3.88896	12.93	<.0001
Error	197	59.24208	0.30072		
Corrected Total	199	67.02000			

Root MSE	0.54838	R-Square	0.1161
Dependent Mean	2.63000	Adj R-Sq	0.1071
Coeff Var	20.85097		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.60630	0.44141	1.37	0.1711
verbal		1	0.00231	0.00055222	4.18	<.0001
math		1	0.00099985	0.00060934	1.64	0.1024

The GLM Procedure

Number of Observations Read 200
Number of Observations Used 200

The GLM Procedure

Dependent Variable: gpa First-year GPA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7.77791708	3.88895854	12.93	<.0001
Error	197	59.24208292	0.30072123		
Corrected Total	199	67.02000000			

R-Square	Coeff Var	Root MSE	gpa Mean
0.116054	20.85097	0.548381	2.630000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
verbal	1	6.96823854	6.96823854	23.17	<.0001
math	1	0.80967854	0.80967854	2.69	0.1024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
verbal	1	5.24922727	5.24922727	17.46	<.0001
math	1	0.80967854	0.80967854	2.69	0.1024

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Verbal=600, Math=600	2.59051345	0.04959917	52.23	<.0001
Verbal=400, Math=800	2.32904960	0.16439161	14.17	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.6062974824	0.44140615	1.37	0.1711
verbal	0.0023071729	0.00055222	4.18	<.0001
math	0.0009998537	0.00060934	1.64	0.1024

Skipping the proc reg output, which we have already seen ...

The UNIVARIATE Procedure
Variable: gparesid (Residual)

Moments

N	200	Sum Weights	200
Mean	0	Sum Observations	0
Std Deviation	0.54561791	Variance	0.29769891
Skewness	-0.7222441	Kurtosis	1.47719763
Uncorrected SS	59.2420829	Corrected SS	59.2420829
Coeff Variation	.	Std Error Mean	0.03858101

Basic Statistical Measures

Location		Variability	
Mean	0.000000	Std Deviation	0.54562
Median	0.046589	Variance	0.29770
Mode	.	Range	3.28403
		Interquartile Range	0.74154

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	7	Pr >= |M|	0.3580
Signed Rank	S	466	Pr >= |S|	0.5709

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.965243	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.044907	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.074753	Pr > W-Sq	0.2445
Anderson-Darling	A-Sq	0.693631	Pr > A-Sq	0.0730

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.0352721
99%	0.9680205
95%	0.8166440
90%	0.7162640
75% Q3	0.3900501
50% Median	0.0465888
25% Q1	-0.3514863
10%	-0.6379376
5%	-0.8383794
1%	-1.8837181

The UNIVARIATE Procedure
Variable: gparesid (Residual)

Quantiles (Definition 5)

Quantile	Estimate
0% Min	-2.2487543

Extreme Observations

------Lowest-----		------Highest-----	
Value	Obs	Value	Obs
-2.24875	131	0.888088	118
-1.96965	136	0.906373	18
-1.79779	121	0.963609	105
-1.23771	40	0.972432	41
-1.07023	113	1.035272	35

```
Stem Leaf                              #             Boxplot
  10 4                                 1                |
   8 011234689167                     12                |
   6 35780122237999                   14                |
   4 00011355556688899990159          23                |
   2 011122335789900122333466668      28             +-----+
   0 112334555555679011234455668 89   29             *--+--*
  -0 98877666655322209754 1111        24             |     |
  -2 99877555432211998765442 00       25             +-----+
  -4 99766543227764321000             20                |
  -6 77877544332                      11                |
  -8 604320                            6                |
 -10 740                               3                |
 -12 4                                 1                |
 -14
 -16
 -18 70                                2                0
 -20
 -22 5                                 1                0
     ----+----+----+----+----+----
 Multiply Stem.Leaf by 10**-1
```


The UNIVARIATE Procedure
Variable: gparesid (Residual)

Normal Probability Plot