Regression on the Math Data
Part 2: Residual Analysis

```
/* MathReg2.sas  */
%include 'readmath2b.sas';
title2 'Residual Analysis for Predicting Grade';

proc reg noprint; /* No output - We've already seen it. */
     title3 'Model 8: hsgpa hscalc hsengl totscore';
     model grade = hsgpa hscalc hsengl totscore;
     output out = Explor predicted = yhat
                         residual  = resid
                         rstudent  = delstud;
                                  /* Deleted Studentized Residual */

/* Could have included LCL and UCL for upper and lower limits of a
   95% prediction interval for each case in the file */

proc sort data=Explor;
     by delstud;

options pagesize=1000;
proc print data = Explor; /* Explor is the default anyway */
     var grade yhat resid delstud;


/* What is a big (Studentized deleted) residual? If the model is correct,
   each one has a t distribution with n-p-1 = 283 df (practically standard
   normal), so the Studentized deleted residual can be treated directly as
   a t-test statistic. Values that are too big in absolute value will cause
   H0: mu=0 to be rejected. Tests are NOT independent, but use a Bonferroni
   correction for n = 289 tests. Get the critical value from proc iml. */

proc iml;
     title3 'Critical value for Joint t-test on Studentized Residuals';
     Alpha = 0.05/289; print Alpha;
     Critval = tinv(1-Alpha/2,283); print Critval;

options pagesize=35;
proc univariate normal plot;
     title3 'Close look at the Studentized deleted residuals';
     var delstud;

/* Tests for normality indicate residuals are not normal. No st resids
greater than crit val. Still, two biggest residuals correspond to grades of
1 and 2 out of 100 -- unnatural. */

proc plot;
     plot grade*yhat;

/* Based on this plot, I would consider setting 6 cases aside and
re-estimating. Probably it would be best to involve the client in the
decision. I wish we had recorded which students took the final exam. */

proc plot;
     title3 'Plot deleted studentized residuals against vars in the model';
     plot delstud * (hsgpa hscalc hsengl totscore);

proc plot;
     title3 'Plot deleted studentized residuals against vars not in model';
     plot delstud * (calc precalc mtongue);
```

Prediction of Performance in First-year Calculus 1
Residual Analysis for Predicting Grade
Model 8: hsgpa hscalc hsengl totscore

Obs	grade	yhat	resid	delstud
1	60	.	.	.
2	61	.	.	.
3	54	.	.	.
4	.	37.5318	.	.
5	14	.	.	.
6	.	.	.	.

Skipping ...

289	.	.	.	.
290	78	.	.	.
291	1	48.2471	-47.2471	-3.56992
292	2	48.0637	-46.0637	-3.47647
293	39	80.2925	-41.2925	-3.09479
294	18	57.0384	-39.0384	-2.90412
295	17	51.9520	-34.9520	-2.63924
296	12	47.2569	-35.2569	-2.61808
297	13	46.8726	-33.8726	-2.53054
298	19	49.7737	-30.7737	-2.30441
299	19	49.7453	-30.7453	-2.27498
300	20	50.2779	-30.2779	-2.26311
301	46	76.5033	-30.5033	-2.26263

Skipping ...

574	82	56.5998	25.4002	1.88228
575	82	55.9114	26.0886	1.94864
576	80	53.3529	26.6471	1.96744
577	57	30.4651	26.5349	1.98728
578	90	63.4272	26.5728	1.99076
579	84	52.6872	31.3128	2.32317

Prediction of Performance in First-year Calculus 2
Residual Analysis for Predicting Grade
Critical value for Joint t-test on Studentized Residuals

Alpha
0.000173

Critval
3.8061889

No Studentized deleted residuals are beyond the Bonferroni-corrected critical value. Still, the two biggest residuals correspond to grades of 1 and 2 out of 100 -- unnatural.

The UNIVARIATE Procedure
Variable: delstud (Studentized Residual without Current Obs)

Moments

N	289	Sum Weights	289
Mean	-0.0011771	Sum Observations	-0.3401753
Std Deviation	1.00731084	Variance	1.01467513
Skewness	-0.713887	Kurtosis	0.96443899
Uncorrected SS	292.226839	Corrected SS	292.226439
Coeff Variation	-85577.292	Std Error Mean	0.05925358

Basic Statistical Measures

Location		Variability	
Mean	-0.00118	Std Deviation	1.00731
Median	0.11240	Variance	1.01468
Mode	.	Range	5.89310
		Interquartile Range	1.20788

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-0.01987	Pr > |t|	0.9842
Sign	M	11.5	Pr >= |M|	0.1955
Signed Rank	S	1596.5	Pr >= |S|	0.2623

The UNIVARIATE Procedure
Variable: delstud (Studentized Residual without Current Obs)

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.969104	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.074472	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.337202	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.194223	Pr > A-Sq	<0.0050

Conclude that the (standardized deleted) residuals are not normally distributed. (With 284 df, the *t* distribution is almost standard normal.)

Quantiles (Definition 5)

Quantile	Estimate
100% Max	2.323174
99%	1.987284
95%	1.486004
90%	1.176567
75% Q3	0.673189
50% Median	0.112403
25% Q1	-0.534687
10%	-1.372680
5%	-1.940896
1%	-3.094794
0% Min	-3.569924

Residual Analysis for Predicting Grade
Close look at the Studentized deleted residuals

The UNIVARIATE Procedure
Variable: delstud (Studentized Residual without Current Obs)

Extreme Observations

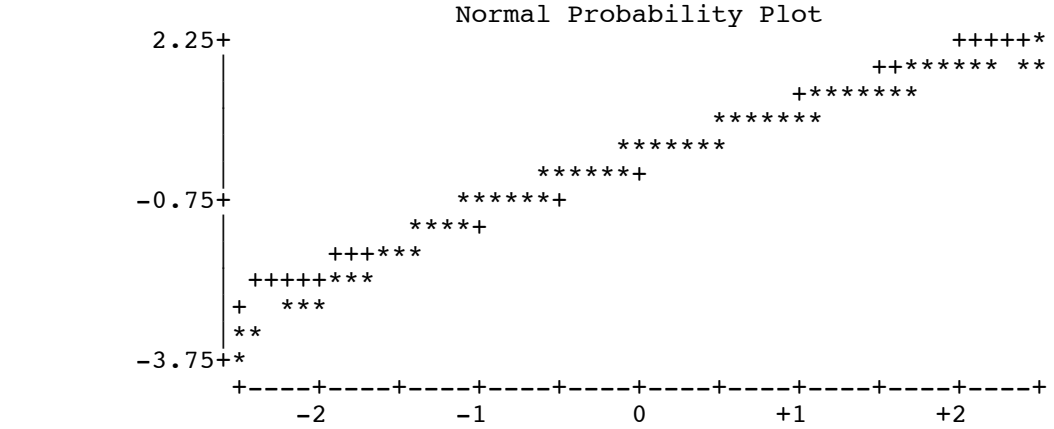
------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
-3.56992	291	1.94864	575
-3.47647	292	1.96744	576
-3.09479	293	1.98728	577
-2.90412	294	1.99076	578
-2.63924	295	2.32317	579

Missing Values

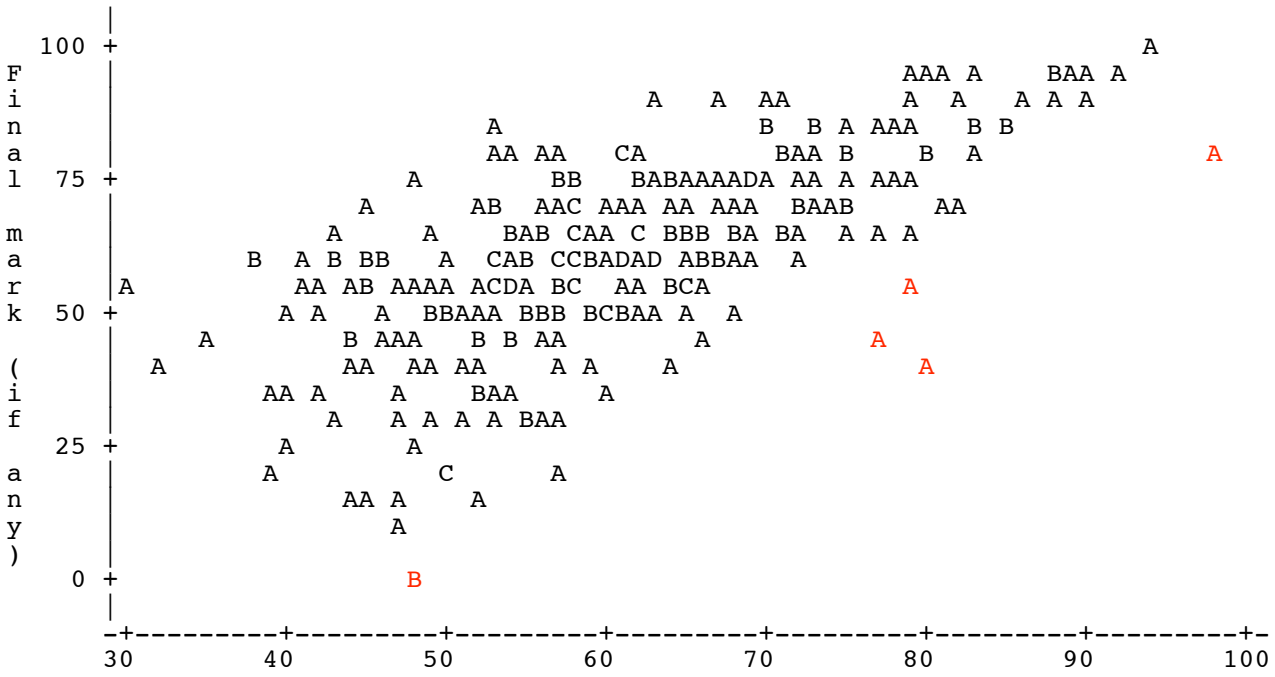
Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	290	50.09	100.00

```
        Prediction of Performance in First-year Calculus                6
             Residual Analysis for Predicting Grade
         Close look at the Studentized deleted residuals

                    The UNIVARIATE Procedure
Variable:  delstud  (Studentized Residual without Current Obs)

                 Histogram                  #             Boxplot
 2.25+*                                     1                |
     .******                                12                |
     .*************                         26                |
     .***************************           53             +-----+
     .********************************      64             *-----*
     .*****************************         58             |  +  |
-0.75+******************                    36             +-----+
     .********                              15                |
     .*****                                 10                |
     .****                                   7                |
     .**                                     4                0
     .*                                      2                0
-3.75+*                                      1                0
      ----+----+----+----+----+----+--
      * may represent up to 2 counts
```

```
        Prediction of Performance in First-year Calculus                7
             Residual Analysis for Predicting Grade
         Close look at the Studentized deleted residuals

                    The UNIVARIATE Procedure
Variable:  delstud  (Studentized Residual without Current Obs)

                         Normal Probability Plot
     2.25+                                                 +++++*
         |                                            ++****** **
         |                                       +*******
         |                                  *******
         |                             *******
         |                        ******+
    -0.75+                  ******+
         |               ****+
         |          +++***
         | +++++***
         |+  ***
         |**
    -3.75+*
          +----+----+----+----+----+----+----+----+----+----+
              -2        -1         0        +1        +2
```



Plot of grade*yhat. Legend: A = 1 obs, B = 2 obs, etc.



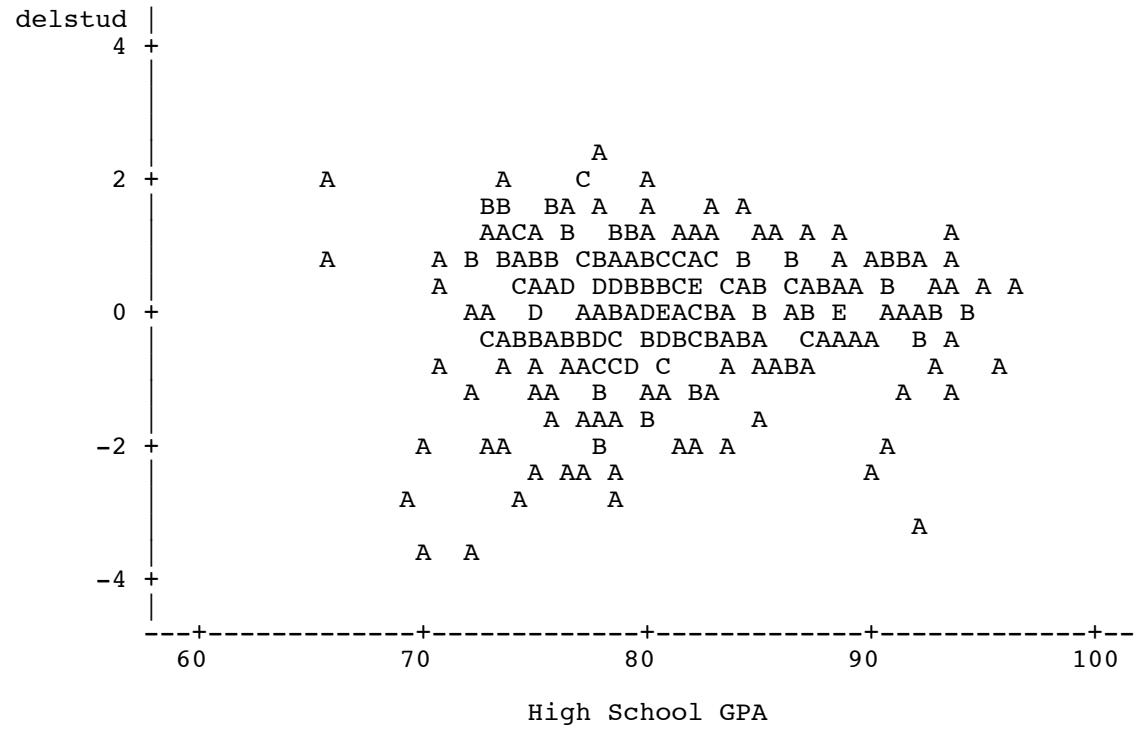
NOTE: 290 obs had missing values.

Based on this plot, I would consider setting 6 cases aside and re-estimating. Probably it would be best to involve the client in the decision. I wish we had recorded which students took the final exam. They will not be dropped for this analysis.

Residual Analysis for Predicting Grade

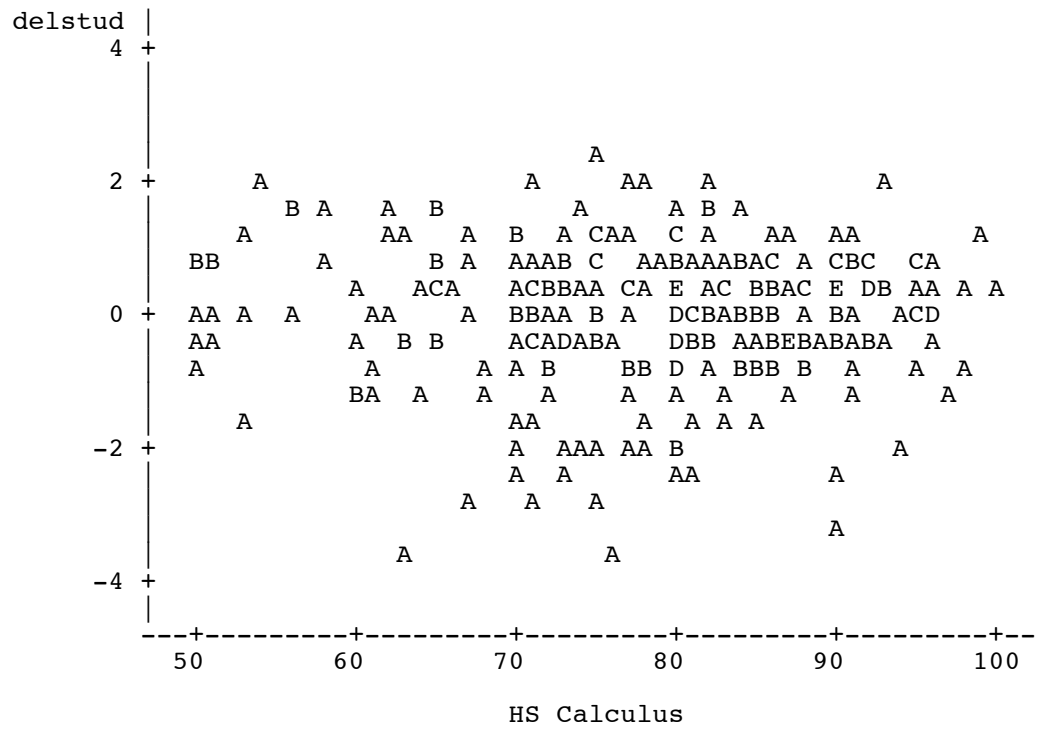
Plot deleted studentized residuals against vars in the model

Plot of delstud*hsgpa. Legend: A = 1 obs, B = 2 obs, etc.



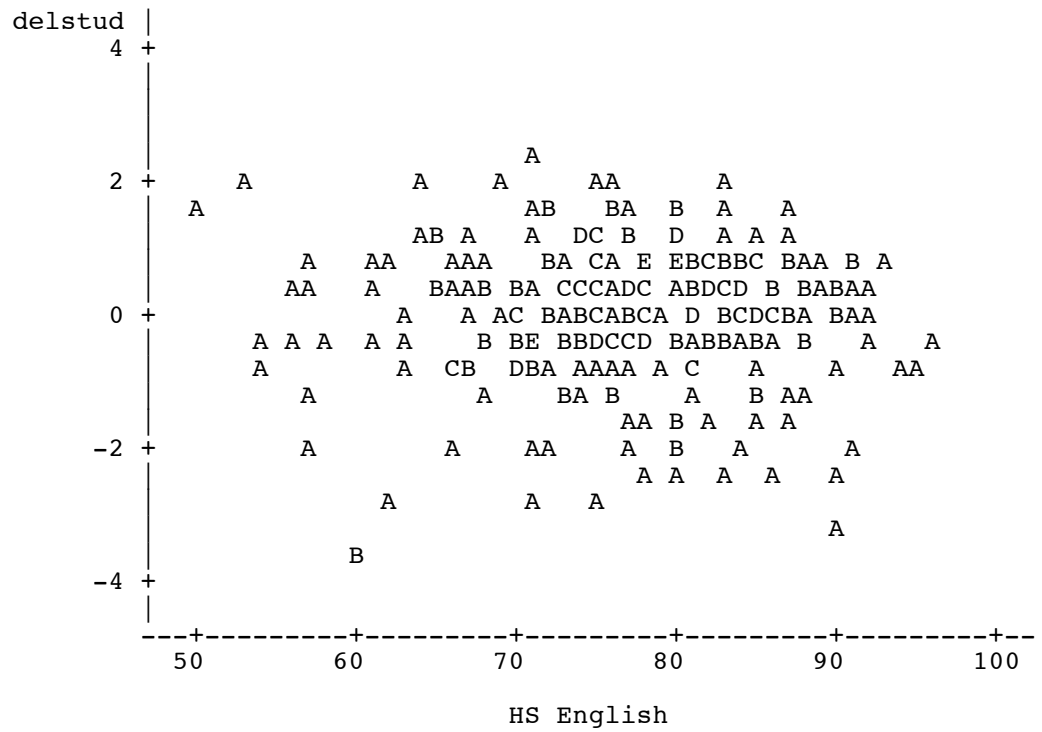
NOTE: 290 obs had missing values.

Plot of delstud*hscalc. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 290 obs had missing values.

Plot of delstud*hsengl. Legend: A = 1 obs, B = 2 obs, etc.

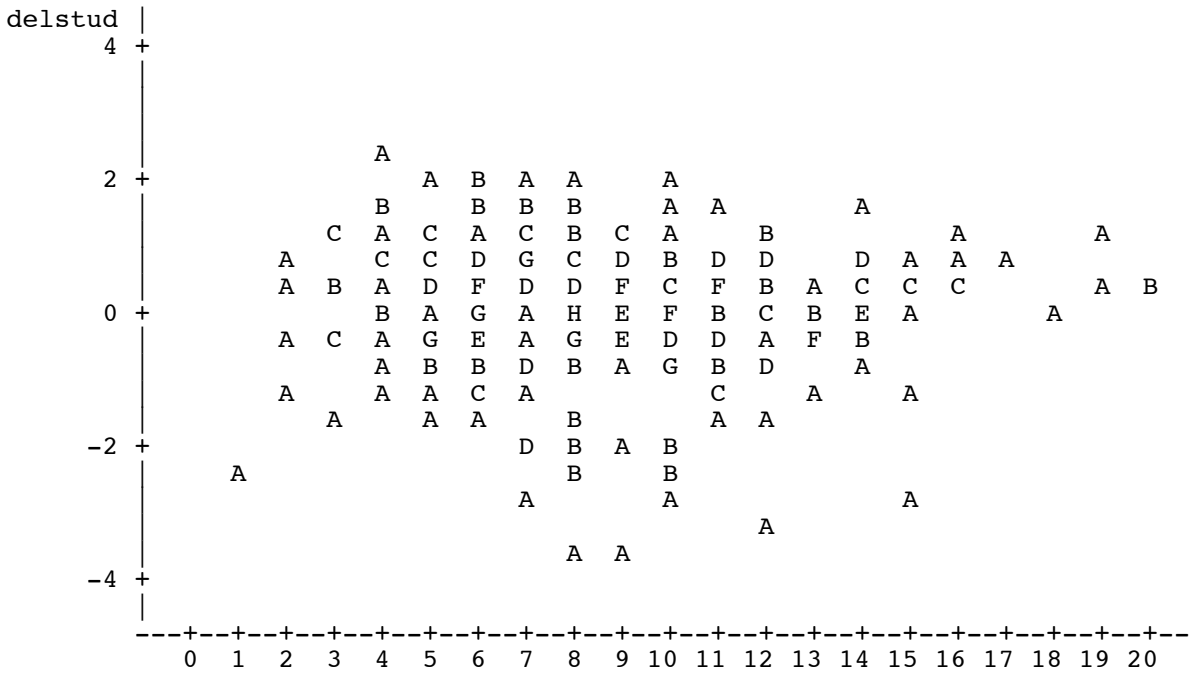


NOTE: 290 obs had missing values.

Residual Analysis for Predicting Grade

Plot deleted studentized residuals against vars in the model

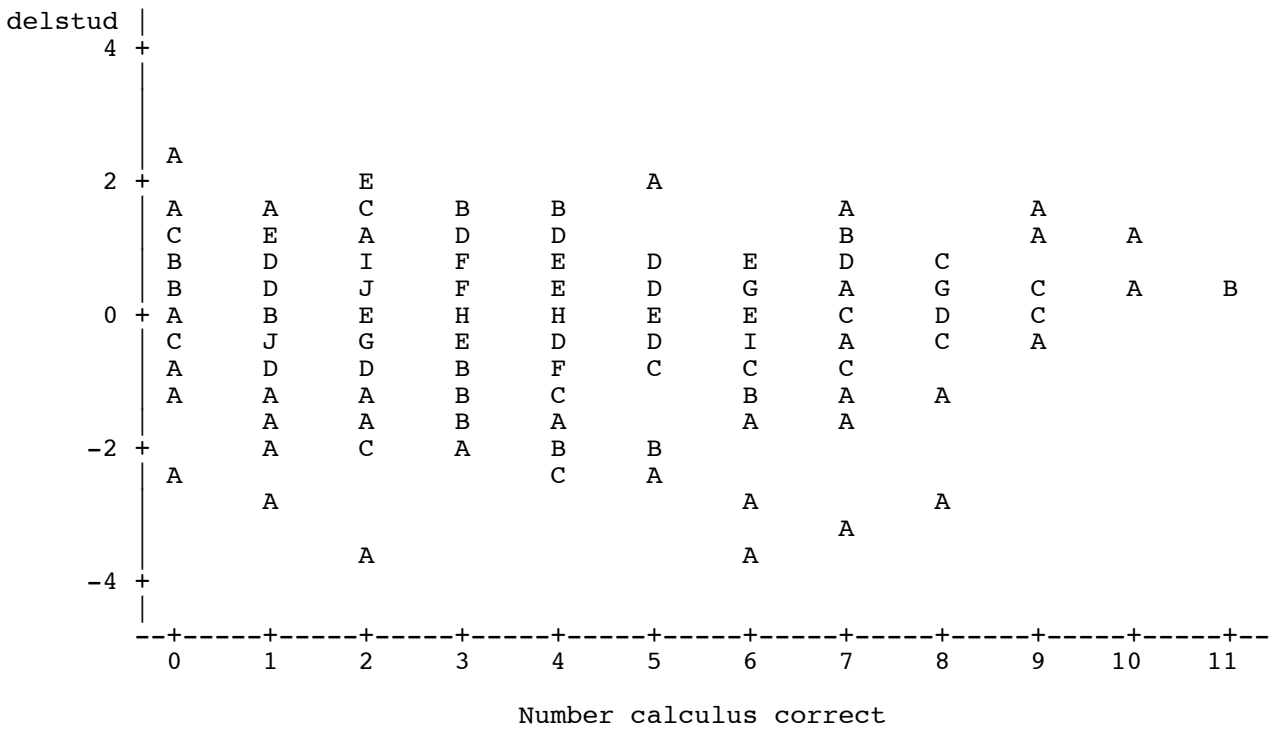
Plot of delstud*totscore. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 290 obs had missing values.

Residual Analysis for Predicting Grade
Plot deleted studentized residuals against vars not in model

Plot of delstud*calc. Legend: A = 1 obs, B = 2 obs, etc.

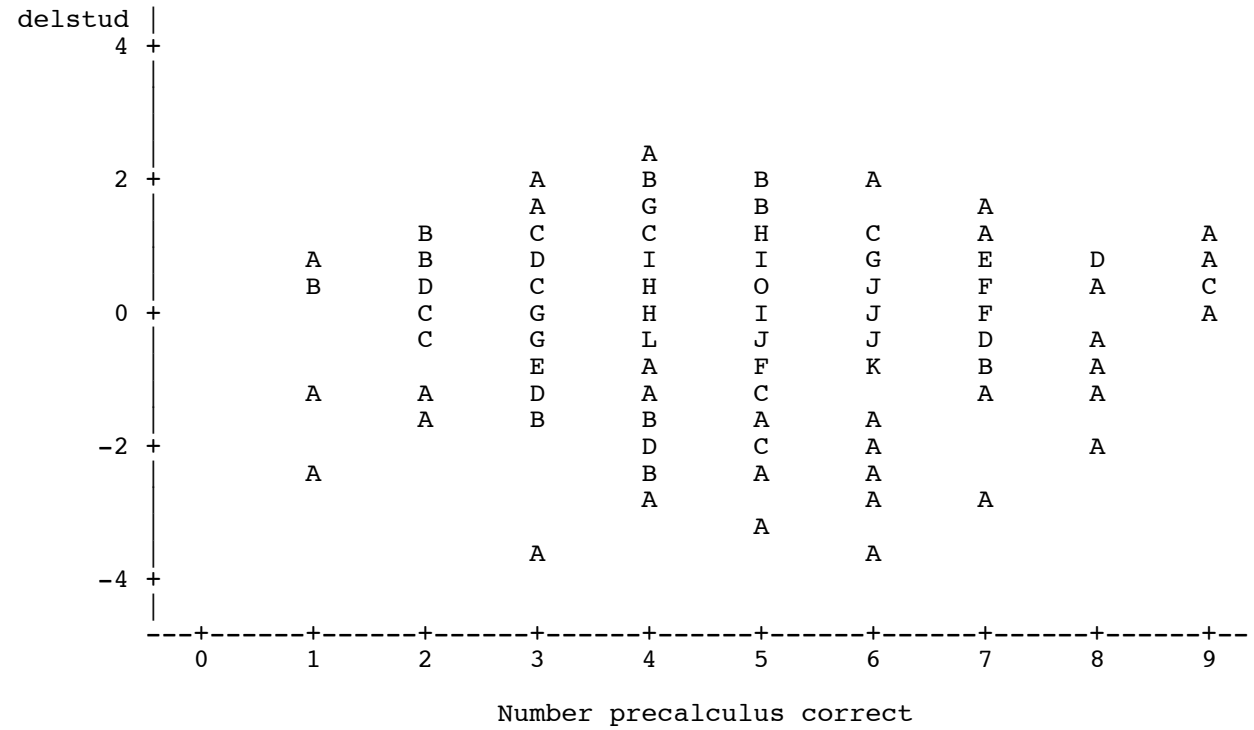


NOTE: 290 obs had missing values.

Residual Analysis for Predicting Grade

Plot deleted studentized residuals against vars not in model

Plot of delstud*precalc. Legend: A = 1 obs, B = 2 obs, etc.

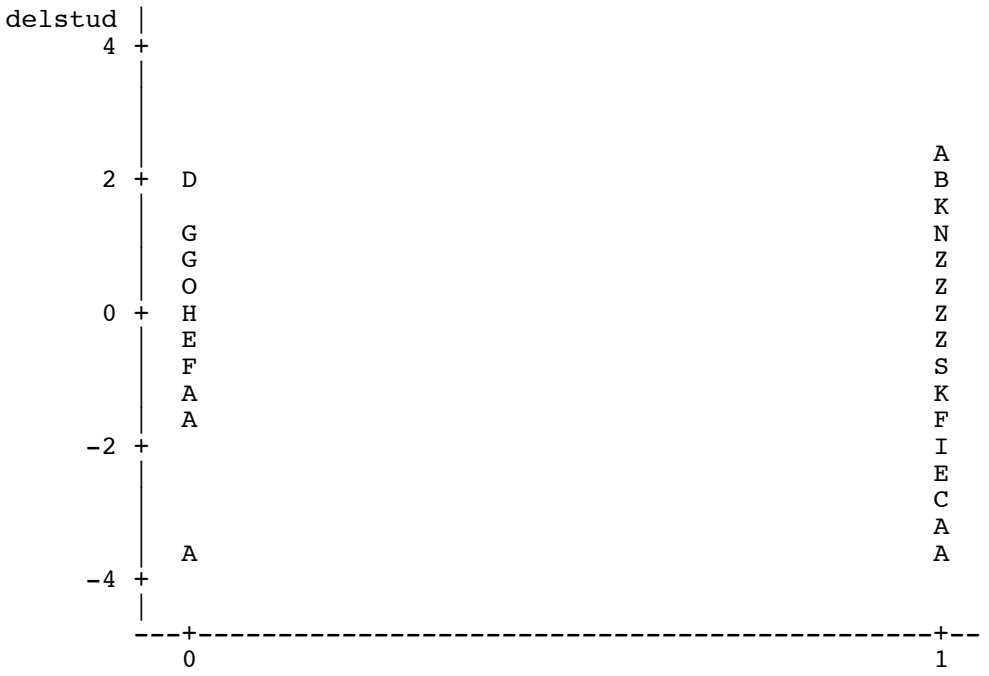


NOTE: 290 obs had missing values.

Residual Analysis for Predicting Grade

Plot deleted studentized residuals against vars not in model

Plot of delstud*mtongue. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 292 obs had missing values. 45 obs hidden.