Regression on the Math Data
Part 3: Prediction

```
/* MathReg3.sas  */
%include 'readexplor.sas'; /* Creates data set explore */
%include 'readreplic.sas'; /* Creates data set replic  */
title2 'Predict Grade for Replication Sample';

/* Plan:

1. Findings from the exploration (based on Model 9, which predicts grade
   from hsgpa hscalc hsengl totscore mtongue) were
      a. HS Engl neg
      b. mtongue neg
      c. totscore positive (diagnostic test matters)
   Test these on the replication with a Bonferroni correction for 3 tests.

2. See if prediction intervals work as advertised for Model 8, which
   predicts grade from hsgpa hscalc hsengl totscore.

3. Compare prediction of letter grade for the models with and without
   the diagnostic test.


First, just illustrate use of different data sets in the same run.  */

proc freq data = explore;
     title3 'Exploratory Sample';
     tables outcome;
proc freq data =replic;
     title3 'Replication Sample';
     tables outcome;

/* Now test the three findings: Point 1 above */

proc reg data = replic;
     title3 'Try to replicate HS Engl neg, mtongue neg, totscore pos';
     title4 'with a Bonferroni correction (check p < 0.05/3 = 0.01666667)';
     model grade = hsgpa hscalc hsengl totscore mtongue;

/* Make combined data set, look at prediction intervals: Point 2 */

data predict;
     set explore replic;
     keeper =  grade+hsgpa+hscalc+hsengl+totscore;
     /* keeper will be missing if any of the vars are missing */
     if keeper ne .; /* Discards all other cases */
     grade2 = grade; /* Save value of grade! */
     if sample=2 then grade=. ; /* DV is now missing for replication sample.
                                   But it is preserved in grade2 */

proc reg data = predict; /* Data set predict is the default anyway */
     title3 'Model 8: hsgpa hscalc hsengl totscore: R-sq = 0.4532';
     model grade = hsgpa hscalc hsengl totscore;
     output out = predata8   predicted = predgrade8
                             LCL       = lowpred
                             UCL       = hipred;
options pagesize=2000;
proc print;
     var id sample grade2 predgrade8 lowpred hipred;
```

```
/* Does 95 Percent Prediction Interval really contain 95 percent of grades? */

data predata8b;
     set predata8;
     if (lowpred < grade2 < hipred) then ininterval='Yes';
        else ininterval='No';

proc freq;
     title3 'Does 95 Percent Prediction Interval Work?';
     tables sample * ininterval / nocol nopercent chisq;

/* Keep trying. Try to predict letter grade. */

data pre8c;
     set predata8b;
     if         80 <= grade2 <= 100 then lgrade = 'A';
        else if 70 <= grade2 <=  79 then lgrade = 'B';
        else if 60 <= grade2 <=  69 then lgrade = 'C';
        else if 50 <= grade2 <=  59 then lgrade = 'D';
        else if  0 <= grade2 <=  49 then lgrade = 'F';
     label lgrade = 'Letter Grade';
     pregrade = round(predgrade8);
     if         80 <= pregrade <= 100 then prelgrade = 'A';
        else if 70 <= pregrade <=  79 then prelgrade = 'B';
        else if 60 <= pregrade <=  69 then prelgrade = 'C';
        else if 50 <= pregrade <=  59 then prelgrade = 'D';
        else if  0 <= pregrade <=  49 then prelgrade = 'F';
     label prelgrade = 'Predicted Letter Grade';

proc freq;
     title3 'Accuracy of predicting Letter Grades From Model 8';
     tables sample*prelgrade*lgrade / nocol nopercent;
     /* Separate table for each sample */

/* That's not completely useless. Is it better than predictions based
   only on High School Information? Re-do everything with Model 1,
   which has just hsgpa hscalc hsengl.  */

data predict1;
     set explore replic;
     keeper =  grade+hsgpa+hscalc+hsengl; /* No totscore this time */
     /* keeper will be missing if any of the vars are missing */
     if keeper ne .; /* Discards all other cases */
     grade2 = grade; /* Save value of grade! */
     if sample=2 then grade=. ; /* DV is now missing for replication sample.
                                   But it is preserved in grade2 */

proc reg data = predict1;
     title3 'Model 1: hsgpa hscalc hsengl: R-sq = 0.4078';
     model grade = hsgpa hscalc hsengl ;
     output out = predata1   predicted = predgrade1;
```

```
data pre1;
     set predata1;
     if         80 <= grade2 <= 100 then lgrade = 'A';
        else if 70 <= grade2 <=  79 then lgrade = 'B';
        else if 60 <= grade2 <=  69 then lgrade = 'C';
        else if 50 <= grade2 <=  59 then lgrade = 'D';
        else if  0 <= grade2 <=  49 then lgrade = 'F';
     label lgrade = 'Letter Grade';
     pregrade = round(predgrade1);
     if         80 <= pregrade <= 100 then prelgrade = 'A';
        else if 70 <= pregrade <=  79 then prelgrade = 'B';
        else if 60 <= pregrade <=  69 then prelgrade = 'C';
        else if 50 <= pregrade <=  59 then prelgrade = 'D';
        else if  0 <= pregrade <=  49 then prelgrade = 'F';
     label prelgrade = 'Predicted Letter Grade';

proc freq;
     title3 'Accuracy of predicting Letter Grades From Model 1';
     tables sample*prelgrade*lgrade / nocol nopercent;
     /* Separate table for each sample */
```

MathReg3.lst

```
              Prediction of Performance in First-year Calculus              1
                    Predict Grade for Replication Sample
                             Exploratory Sample

                             The FREQ Procedure

                                                 Cumulative    Cumulative
      outcome        Frequency     Percent     Frequency      Percent
      ---------------------------------------------------------------------
      Disappeared          186       32.12           186        32.12
      Failed                88       15.20           274        47.32
      Passed               305       52.68           579       100.00
```

```
              Prediction of Performance in First-year Calculus              2
                    Predict Grade for Replication Sample
                             Replication Sample

                             The FREQ Procedure

                                                 Cumulative    Cumulative
      outcome        Frequency     Percent     Frequency      Percent
      ---------------------------------------------------------------------
      Disappeared          199       34.37           199        34.37
      Failed                96       16.58           295        50.95
      Passed               284       49.05           579       100.00
```

Predict Grade for Replication Sample
Try to replicate HS Engl neg, mtongue neg, totscore pos
with a Bonferroni correction (check p < 0.05/3 = 0.01666667)

The REG Procedure
Model: MODEL1
Dependent Variable: grade Final mark (if any)

Number of Observations Read	579
Number of Observations Used	288
Number of Observations with Missing Values	291

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	36155	7230.94549	36.45	<.0001
Error	282	55948	198.39605		
Corrected Total	287	92102			

Root MSE	14.08531	R-Square	0.3925
Dependent Mean	57.73264	Adj R-Sq	0.3818
Coeff Var	24.39749		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-57.59065	12.01535
hsgpa	High School GPA	1	1.15692	0.25532
hscalc	HS Calculus	1	0.17755	0.10404
hsengl	HS English	1	-0.04691	0.13560
totscore	Total # right on diagnostic test	1	1.58048	0.26400
mtongue	English vs. Other	1	-1.34224	1.96918

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-4.79	<.0001
hsgpa	High School GPA	1	4.53	<.0001
hscalc	HS Calculus	1	1.71	0.0890
hsengl	HS English	1	-0.35	0.7297
totscore	Total # right on diagnostic test	1	5.99	<.0001
mtongue	English vs. Other	1	-0.68	0.4960

hsengl, mtongue not replicated

Replication raises a moral dilemma, because High School Calculus was non-significant! Don't change prediction equation. Let this be a lesson.

Predict Grade for Replication Sample

Model 8: hsgpa hscalc hsengl totscore: R-sq = 0.4532

The REG Procedure

Model: MODEL1

Dependent Variable: grade Final mark (if any)

Number of Observations Read	582
Number of Observations Used	289
Number of Observations with Missing Values	293

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	43951	10988	58.86	<.0001
Error	284	53019	186.68704		
Corrected Total	288	96970			

Root MSE	13.66335	R-Square	0.4532
Dependent Mean	60.57785	Adj R-Sq	0.4455
Coeff Var	22.55502		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-70.49310	11.11446
hsgpa	High School GPA	1	1.60612	0.22085
hscalc	HS Calculus	1	0.24685	0.10148
hsengl	HS English	1	-0.35057	0.12060
totscore	Total # right on diagnostic test	1	0.98964	0.25277

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-6.34	<.0001
hsgpa	High School GPA	1	7.27	<.0001
hscalc	HS Calculus	1	2.43	0.0156
hsengl	HS English	1	-2.91	0.0039
totscore	Total # right on diagnostic test	1	3.92	0.0001

Predict Grade for Replication Sample

Model 8: hsgpa hscalc hsengl totscore: R-sq = 0.4532

Obs	id	sample	grade2	predgrade8	lowpred	hipred
1	1	1	39	44.7631	17.6303	71.896
2	2	1	57	30.4651	3.1031	57.827
3	3	1	62	60.7022	33.7130	87.691
4	4	1	76	63.9140	36.8352	90.993
5	5	1	86	70.2654	43.2550	97.276

Skipping ...

282	568	1	57	65.6756	38.6031	92.748
283	570	1	80	72.4965	45.3954	99.597
284	571	1	56	53.2341	26.1009	80.367
285	572	1	70	68.6775	41.6112	95.744
286	574	1	60	57.3580	30.1966	84.519
287	576	1	50	61.2562	34.2514	88.261
288	577	1	50	50.4860	23.2451	77.727
289	579	1	61	58.0563	31.0303	85.082
290	580	2	56	52.6957	25.5933	79.798
291	584	2	50	68.3759	41.3845	95.367
292	586	2	70	71.0902	44.0049	98.175
293	587	2	50	65.3331	38.3070	92.359
294	588	2	8	40.5222	13.4496	67.595
295	589	2	44	64.6459	37.5533	91.738

Skipping ...

296	590	2	60	60.4670	33.5018	87.432
297	593	2	45	53.3053	26.2441	80.367
298	595	2	66	54.0232	27.0466	81.000
299	601	2	67	72.2857	45.2524	99.319
300	603	2	53	47.6901	20.3970	74.983
301	605	2	57	56.9820	29.6573	84.307

Skipping ...

576	1147	2	33	52.1708	25.1675	79.174
577	1149	2	76	67.1041	40.0618	94.146
578	1150	2	84	69.7338	42.7267	96.741
579	1151	2	86	70.3894	43.0777	97.701
580	1153	2	75	74.6985	47.5499	101.847
581	1154	2	60	57.8540	30.8700	84.838
582	1155	2	62	61.7287	34.7303	88.727

The prediction intervals are almost useless. This actually should not be too surprising. Statistical methods may be able to detect trends, but those trends are often not strong enough to allow accurate predictions about individuals. Do the intervals at least capture the observation 95% of the time? Maybe not, since the data fail all tests for normality, and the prediction intervals are based on normal theory.

The FREQ Procedure

Table of sample by ininterval

sample	ininterval		
Frequency / Row Pct	No	Yes	Total
1	15 5.19	274 94.81	289
2	15 5.12	278 94.88	293
Total	30	552	582

Statistics for Table of sample by ininterval

Statistic	DF	Value	Prob
Chi-Square	1	0.0015	0.9692
Likelihood Ratio Chi-Square	1	0.0015	0.9692
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0015	0.9692
Phi Coefficient		0.0016	
Contingency Coefficient		0.0016	
Cramer's V		0.0016	

Fisher's Exact Test

Cell (1,1) Frequency (F)	15
Left-sided Pr <= F	0.5894
Right-sided Pr >= F	0.5589
Table Probability (P)	0.1482
Two-sided Pr <= P	1.0000

Sample Size = 582

The prediction intervals may be almost useless, but technically, they work beautifully.

The FREQ Procedure

Table 1 of prelgrade by lgrade
Controlling for sample=1

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency / Row Pct	A	B	C	D	F	Total
A	20	3	0	0	1	24
	83.33	12.50	0.00	0.00	4.17	
B	16	15	7	1	1	40
	40.00	37.50	17.50	2.50	2.50	
C	5	22	28	19	3	77
	6.49	28.57	36.36	24.68	3.90	
D	4	11	23	31	25	94
	4.26	11.70	24.47	32.98	26.60	
F	0	1	11	15	27	54
	0.00	1.85	20.37	27.78	50.00	
Total	45	52	69	66	57	289

Table 2 of prelgrade by lgrade
Controlling for sample=2

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency / Row Pct	A	B	C	D	F	Total
A	10	3	1	0	0	14
	71.43	21.43	7.14	0.00	0.00	
B	17	9	6	4	1	37
	45.95	24.32	16.22	10.81	2.70	
C	4	20	29	26	12	91
	4.40	21.98	31.87	28.57	13.19	
D	2	10	21	28	21	82
	2.44	12.20	25.61	34.15	25.61	
F	0	3	14	20	32	69
	0.00	4.35	20.29	28.99	46.38	
Total	33	45	71	78	66	293

That's not completely useless. Is it better than predictions based only on High School Information? Redo everything with Model 1, which has just `hsgpa hscalc hsengl`.

Predict Grade for Replication Sample
Model 1: hsgpa hscalc hsengl: R-sq = 0.4078

The REG Procedure
Model: MODEL1
Dependent Variable: grade Final mark (if any)

Number of Observations Read	647
Number of Observations Used	323
Number of Observations with Missing Values	324

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	45188	15063	73.23	<.0001
Error	319	65616	205.69147		
Corrected Total	322	110803			

Root MSE	14.34195	R-Square	0.4078
Dependent Mean	59.79257	Adj R-Sq	0.4022
Coeff Var	23.98617		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-77.07643	10.92142	-7.06	<.0001
hsgpa	High School GPA	1	1.76845	0.22097	8.00	<.0001
hscalc	HS Calculus	1	0.30594	0.09750	3.14	0.0019
hsengl	HS English	1	-0.38461	0.11815	-3.26	0.0013

The FREQ Procedure

Table 1 of prelgrade by lgrade
Controlling for sample=1

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency Row Pct	A	B	C	D	F	Total
A	18 78.26	2 8.70	2 8.70	1 4.35	0 0.00	23
B	17 39.53	15 34.88	7 16.28	2 4.65	2 4.65	43
C	7 8.14	24 27.91	27 31.40	23 26.74	5 5.81	86
D	6 5.66	12 11.32	26 24.53	34 32.08	28 26.42	106
F	0 0.00	2 3.08	11 16.92	21 32.31	31 47.69	65
Total	48	55	73	81	66	323

Table 2 of prelgrade by lgrade
Controlling for sample=2

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency Row Pct	A	B	C	D	F	Total
A	11 78.57	2 14.29	1 7.14	0 0.00	0 0.00	14
B	12 32.43	12 32.43	8 21.62	5 13.51	0 0.00	37
C	8 7.55	21 19.81	29 27.36	33 31.13	15 14.15	106
D	3 3.00	11 11.00	24 24.00	26 26.00	36 36.00	100
F	0 0.00	2 2.99	14 20.90	20 29.85	31 46.27	67
Total	34	48	76	84	82	324

Just compare results for Sample 2 (the replication Sample)

Prediction based on Model 1 (HS Only)

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency Row Pct	A	B	C	D	F	Total
A	11 78.57	2 14.29	1 7.14	0 0.00	0 0.00	14
B	12 32.43	12 32.43	8 21.62	5 13.51	0 0.00	37
C	8 7.55	21 19.81	29 27.36	33 31.13	15 14.15	106
D	3 3.00	11 11.00	24 24.00	26 26.00	36 36.00	100
F	0 0.00	2 2.99	14 20.90	20 29.85	31 46.27	67
Total	34	48	76	84	82	324

Prediction based on Model 8 (HS + Diagnostic Test)

prelgrade(Predicted Letter Grade) lgrade(Letter Grade)

Frequency Row Pct	A	B	C	D	F	Total
A	10 71.43	3 21.43	1 7.14	0 0.00	0 0.00	14
B	17 45.95	9 24.32	6 16.22	4 10.81	1 2.70	37
C	4 4.40	20 21.98	29 31.87	26 28.57	12 13.19	91
D	2 2.44	10 12.20	21 25.61	28 34.15	21 25.61	82
F	0 0.00	3 4.35	14 20.29	20 28.99	32 46.38	69
Total	33	45	71	78	66	293

```
/* MathReg4.sas  */
%include 'readexplor.sas'; /* Creates data set explore */
title2 'Predict Grade for a New Student';

/*         Predict grade for a new student with
           hsgpa=80 hscalc=90 hsengl=70 totscore=15
           For just a prediction, proc glm is easier */

proc glm;
     model grade = hsgpa hscalc hsengl totscore;
     estimate 'New Student' intercept 1  hsgpa 80  hscalc 90  hsengl 70
                            totscore 15;

/* Prediction for Y_{n+1} is the same as estimate of E[Y|X]. CI from proc glm
   is for E[Y|X]. PREDICTION interval for Y_{n+1} is wider. */

data student;
     hsgpa=80; hscalc=90; hsengl=70; totscore=15; id = -27;

data together;
     set explore student;
     /* All variables not assigned will be missing for observation -27 */

proc reg;
     title3 'Model 8: hsgpa hscalc hsengl totscore: R-sq = 0.4532';
     model grade = hsgpa hscalc hsengl totscore;
     output out = guess      predicted = PredictedY
                             LCL       = LowerLimit
                             UCL       = UpperLimit;
data newguess;
     set guess;
     if id < 0; /* Discard all other cases */

proc print;
     title3 'hsgpa=80 hscalc=90 hsengl=70 totscore=15';
     var predictedY LowerLimit UpperLimit;
```

The GLM Procedure

Number of Observations Read	579
Number of Observations Used	289

The GLM Procedure

Dependent Variable: grade Final mark (if any)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	43951.37758	10987.84440	58.86	<.0001
Error	284	53019.12069	186.68704		
Corrected Total	288	96970.49827			

R-Square	Coeff Var	Root MSE	grade Mean
0.453245	22.55502	13.66335	60.57785

Source	DF	Type I SS	Mean Square	F Value	Pr > F
hsgpa	1	34512.00462	34512.00462	184.87	<.0001
hscalc	1	4744.09530	4744.09530	25.41	<.0001
hsengl	1	1833.66807	1833.66807	9.82	0.0019
totscore	1	2861.60960	2861.60960	15.33	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
hsgpa	1	9873.322719	9873.322719	52.89	<.0001
hscalc	1	1104.643344	1104.643344	5.92	0.0156
hsengl	1	1577.482988	1577.482988	8.45	0.0039
totscore	1	2861.609600	2861.609600	15.33	0.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
New Student	70.5174682	1.89565305	37.20	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-70.49309680	11.11446120	-6.34	<.0001
hsgpa	1.60611878	0.22085289	7.27	<.0001
hscalc	0.24684591	0.10147802	2.43	0.0156
hsengl	-0.35056599	0.12059923	-2.91	0.0039
totscore	0.98963670	0.25277126	3.92	0.0001

Predict Grade for a New Student

Model 8: hsgpa hscalc hsengl totscore: R-sq = 0.4532

The REG Procedure

Model: MODEL1

Dependent Variable: grade Final mark (if any)

Number of Observations Read	580
Number of Observations Used	289
Number of Observations with Missing Values	291

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	43951	10988	58.86	<.0001
Error	284	53019	186.68704		
Corrected Total	288	96970			

Root MSE	13.66335	R-Square	0.4532
Dependent Mean	60.57785	Adj R-Sq	0.4455
Coeff Var	22.55502		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-70.49310	11.11446
hsgpa	High School GPA	1	1.60612	0.22085
hscalc	HS Calculus	1	0.24685	0.10148
hsengl	HS English	1	-0.35057	0.12060
totscore	Total # right on diagnostic test	1	0.98964	0.25277

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-6.34	<.0001
hsgpa	High School GPA	1	7.27	<.0001
hscalc	HS Calculus	1	2.43	0.0156
hsengl	HS English	1	-2.91	0.0039
totscore	Total # right on diagnostic test	1	3.92	0.0001

Predict Grade for a New Student

hsgpa=80 hscalc=90 hsengl=70 totscore=15

Obs	Predicted Y	Lower Limit	Upper Limit
1	70.5175	43.3656	97.6694