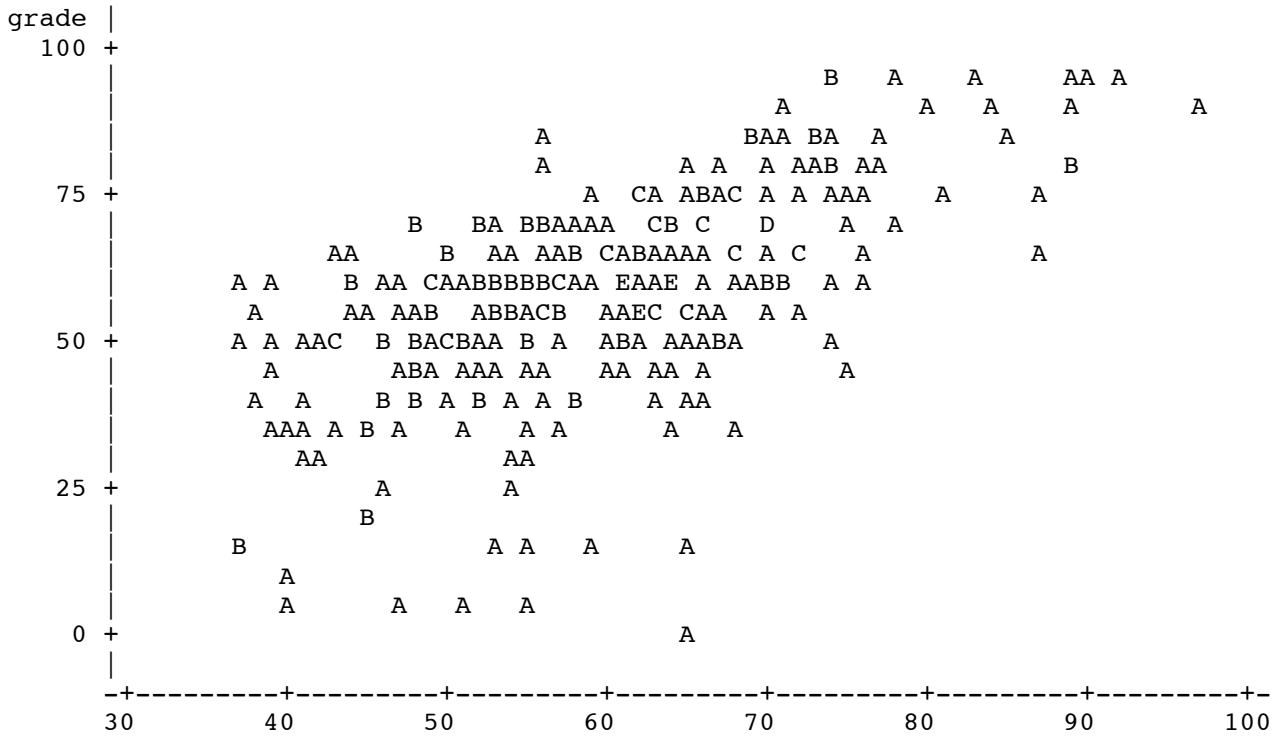
Sta442/1008f05 Overheads 7

Replication of a prediction model

```
/* mathreplic1.sas */
%include 'mathrepread.sas';
title2 'Check prediction of grade';

     if 80 le grade le 100 then Lgrade = 'A';
        else if 70 le grade le 79 then Lgrade = 'B';
        else if 60 le grade le 69 then Lgrade = 'C';
        else if 50 le grade le 59 then Lgrade = 'D';
        else if grade < 50 then Lgrade = 'F';
     label Lgrade = 'Letter Grade Received';

     premark1 = -66.75516 + 1.58918*gpa - 0.30024*english + 0.21759*hscalc +
                0.97213*totscore - 4.69657*tongue;
     premark1 = round(premark1);
     label premark1 = 'Mark Predicted by Jerry''s Model';
     dmark1 = grade-premark1; admark1 = abs(dmark1);
     if 80 le premark1 le 100 then preLG1 = 'A';
        else if 70 le premark1 le 79 then preLG1 = 'B';
        else if 60 le premark1 le 69 then preLG1 = 'C';
        else if 50 le premark1 le 59 then preLG1 = 'D';
        else if premark1 < 50 then preLG1 = 'F';
     label preLG1 = 'Letter Grade Predicted by Jerry''s Model';

     premark2 = -64.60808 + 1.58717*gpa - 0.32445*english + 0.20766 *hscalc +
                0.95113*totscore - 4.48110*tongue + 4.57904*e4;
     premark2 = round(premark2);
     label premark2 = 'Mark Predicted by Stepwise Model';
     dmark2 = grade-premark2; admark2 = abs(dmark2);
     if 80 le premark2 le 100 then preLG2 = 'A';
        else if 70 le premark2 le 79 then preLG2 = 'B';
        else if 60 le premark2 le 69 then preLG2 = 'C';
        else if 50 le premark2 le 59 then preLG2 = 'D';
        else if premark2 < 50 then preLG2 = 'F';
     label preLG2 = 'Letter Grade Predicted by Stepwise Model';


     premark3 = -70.49310 + 1.60612*gpa - 0.35057*english + 0.24685*hscalc +
                0.98964*totscore;
     premark3 = round(premark3);
     label premark3 = 'Mark Predicted by Politically Safe Model';
     dmark3 = grade-premark3; admark3 = abs(dmark3);
     if 80 le premark2 le 100 then preLG3 = 'A';
        else if 70 le premark3 le 79 then preLG3 = 'B';
        else if 60 le premark3 le 69 then preLG3 = 'C';
        else if 50 le premark3 le 59 then preLG3 = 'D';
        else if premark3 < 50 then preLG3 = 'F';
     label preLG3 = 'Letter Grade Predicted by Safe Model';

     pre1 = premark1; pre2 = premark2 ; pre3 = premark3 ;
     realmark = grade;

proc plot;
```

```
     plot grade*premark1;
run;

options pagesize = 200;

proc corr;
     var pre1 -- realmark;
proc freq;
     tables Lgrade preLG1 preLG2 preLG3;
proc means;
     var admark1 admark2 admark3
          dmark1  dmark2  dmark3;

proc freq;
     tables (preLG1 preLG2 preLG3) * Lgrade / nocol nopercent;

/* Just a check, commented out once it works

proc freq;
     tables grade*Lgrade / norow nocol nopercent;
     tables premark1*preLG1 / norow nocol nopercent;
     tables premark2*preLG2 / norow nocol nopercent;
     tables premark3*preLG3 / norow nocol nopercent;
```

Plot of grade*premark1. Legend: A = 1 obs, B = 2 obs, etc.



Mark Predicted by Jerry's Model

NOTE: 291 obs had missing values.

The CORR Procedure

4 Variables: pre1 pre2 pre3 realmark

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pre1	382	57.31152	11.70144	21893	31.00000	97.00000
pre2	382	57.28796	11.50400	21884	35.00000	100.00000
pre3	390	56.92564	11.57282	22201	31.00000	94.00000
realmark	380	56.44211	18.25745	21448	1.00000	97.00000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	pre1	pre2	pre3	realmark
pre1	1.00000	0.99049	0.98551	0.60611
		<.0001	<.0001	<.0001
	382	382	382	288
pre2	0.99049	1.00000	0.97646	0.60956
	<.0001		<.0001	<.0001
	382	382	382	288
pre3	0.98551	0.97646	1.00000	0.61690
	<.0001	<.0001		<.0001
	382	382	390	293
realmark	0.60611	0.60956	0.61690	1.00000
	<.0001	<.0001	<.0001	
	288	288	293	380

The FREQ Procedure

Letter Grade Received

Lgrade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	35	6.04	35	6.04
B	62	10.71	97	16.75
C	82	14.16	179	30.92
D	105	18.13	284	49.05
F	295	50.95	579	100.00

Letter Grade Predicted by Jerry's Model

pre LG1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	14	2.42	14	2.42
B	44	7.60	58	10.02
C	99	17.10	157	27.12
D	119	20.55	276	47.67
F	303	52.33	579	100.00

Letter Grade Predicted by Stepwise Model

pre LG2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	14	2.42	14	2.42
B	36	6.22	50	8.64
C	104	17.96	154	26.60
D	121	20.90	275	47.50
F	304	52.50	579	100.00

Letter Grade Predicted by Safe Model

pre LG3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	14	2.42	14	2.42
B	37	6.40	51	8.82
C	101	17.47	152	26.30
D	123	21.28	275	47.58
F	303	52.42	578	100.00

Frequency Missing = 1

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
admark1	288	10.7500000	9.5552674	0	64.0000000
admark2	288	10.6944444	9.5101470	0	63.0000000
admark3	293	10.5324232	9.5553438	0	66.0000000
dmark1	288	-1.8888889	14.2719291	-64.0000000	31.0000000
dmark2	288	-1.7638889	14.2158503	-63.0000000	32.0000000
dmark3	293	-1.4744027	14.1575114	-66.0000000	31.0000000

The FREQ Procedure

Table of preLG1 by Lgrade

preLG1(Letter Grade Predicted by Jerry's Model)
Lgrade(Letter Grade Received)

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 38.64	10 22.73	10 22.73	5 11.36	2 4.55	44
C	3 3.03	18 18.18	27 27.27	29 29.29	22 22.22	99
D	2 1.68	11 9.24	22 18.49	24 20.17	60 50.42	119
F	3 0.99	20 6.60	22 7.26	47 15.51	211 69.64	303
Total	35	62	82	105	295	579

Table of preLG2 by Lgrade

preLG2(Letter Grade Predicted by Stepwise Model)
Lgrade(Letter Grade Received)

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	15 41.67	9 25.00	7 19.44	3 8.33	2 5.56	36
C	5 4.81	19 18.27	28 26.92	31 29.81	21 20.19	104
D	2 1.65	11 9.09	24 19.83	25 20.66	59 48.76	121
F	3 0.99	20 6.58	22 7.24	46 15.13	213 70.07	304
Total	35	62	82	105	295	579

Table of preLG3 by Lgrade

preLG3(Letter Grade Predicted by Safe Model)
Lgrade(Letter Grade Received)

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	16 43.24	9 24.32	6 16.22	4 10.81	2 5.41	37
C	4 3.96	20 19.80	29 28.71	26 25.74	22 21.78	101
D	2 1.63	10 8.13	21 17.07	28 22.76	62 50.41	123
F	2 0.66	20 6.60	25 8.25	47 15.51	209 68.98	303
Total	34	62	82	105	295	578

Frequency Missing = 1

For any regression model, the squared correlation between Y and predicted Y ($\hat{Y}$) is exactly R^2. In the table below, the first column is the usual , which we interpret as the proportion of variation in the DV that is explained by the independent variables.

In the second column we see the proportion of variation that is explained by the prediction equation when we apply it to the replication sample.

	Squared correlation between predicted and observed	
	Exploratory Data	Replication Data
Jerry's Model	0.4635	0.3674
Stepwise Model	0.4710	0.3716
Politically Safe Model	0.4532	0.3806