Sta442/1008f05 Overheads 5: Regression I

```
/* mathreg2.sas */
%include 'mathexread2.sas';
title2 'Illustrate regression with exploratory math data';

/* Make dummy variables */

if ethnic = . then e1=. ;
   else if ethnic = 1 then e1 = 1; else e1=0; /* Asian vs. Ref cat*/
if ethnic = . then e2=. ;
   else if ethnic = 2 then e2 = 1; else e2=0; /* Euro vs. Ref cat*/
if ethnic = . then e3=. ;
   else if ethnic = 3 then e3 = 1; else e3=0;  /* Mid-East vs Ref cat */
if ethnic = . then e4=. ;
   else if ethnic = 4 then e4 = 1; else e4=0;  /* East Ind. vs Ref cat */
if ethnic = . then e5=. ;
   else if ethnic = 5 then e5 = 1; else e5=0;  /* Other vs Ref cat */

/* Course=2 is the reference category */
if course = . then c1 = .;
   else if course = 4 then c1 = .;
   else if course = 1 then c1 = 1; else c1 = 0;
if course = . then c3 = .;
   else if course = 4 then c3 = .;
   else if course = 3 then c3 = 1; else c3 = 0;

if sex = 'Female' then gender=1;
   else if sex = 'Male' then gender=0;

if tongue=3 then tongue=0; /* Make tongue a dummy for English */

options pagesize=100;

/* First just illustrate that dummy variable regression gives same results as
   oneway ANOVA */

proc glm;
     class ethnic;
     model grade = ethnic;

/* ethnic: European has the highest freq, so make it reference category */
proc reg;
     model grade = e1 e3 e4 e5;
```

```
proc reg;
     title3 'Predict grade: Try some different models';
     model grade = gpa english hscalc;
     model grade = gpa english hscalc tongue;
     model grade = gpa english hscalc tongue
                   precalc calc / ss1;
           diagtest: test precalc = calc = 0;
     model grade = gpa english hscalc
                   precalc calc
                   gender tongue e1 e3 e4 e5;
           diagtest: test precalc = calc = 0;
           ethtest:  test e1=e3=e4=e5=0;
           demogr:   test gender=tongue=e1=e3=e4=e5=0;

/* Now it looks like the only demographic variable that matters is
   tongue. Here is the model I like most */

proc reg;
     title3 'Jerry''s favorite model';
     model grade = gpa english hscalc totscore tongue;

/* But maybe demographics are affecting HS performance, diagnostic test
performance and performance in university calculus. This is a different
model. */

proc reg;
     title3 'Multivariate regression';
     model gpa english hscalc totscore = gender tongue e1 e3 e4 e5;
         /* mtest gives multivariate test */
     ethnic: mtest e1=e3=e4=e5=0;
     gender: mtest gender=0;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
ethnic	5	Asian East Indian European Middle-Eastern Other or unknown

Number of observations 579

NOTE: Due to missing values, only 393 observations can be used in this analysis.

The GLM Procedure

Dependent Variable: grade Mark in university calculus

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3631.5326	907.8832	2.50	0.0418
Error	388	140622.4724	362.4291		
Corrected Total	392	144254.0051			

R-Square	Coeff Var	Root MSE	grade Mean
0.025175	32.41104	19.03757	58.73791

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ethnic	4	3631.532642	907.883161	2.50	0.0418

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ethnic	4	3631.532642	907.883161	2.50	0.0418

The REG Procedure
Model: MODEL1
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3631.53264	907.88316	2.50	0.0418
Error	388	140622	362.42905		
Corrected Total	392	144254			

Root MSE	19.03757	R-Square	0.0252
Dependent Mean	58.73791	Adj R-Sq	0.0151
Coeff Var	32.41104		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	56.15426	1.38846	40.44
e1		1	3.90322	2.46854	1.58
e3		1	3.20574	3.02926	1.06
e4		1	9.03442	2.96076	3.05
e5		1	2.44574	5.10781	0.48

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	<.0001
e1		1	0.1147
e3		1	0.2906
e4		1	0.0024
e5		1	0.6323

The REG Procedure
Model: MODEL1
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	41108	13703	69.59	<.0001
Error	283	55721	196.89337		
Corrected Total	286	96829			

Root MSE	14.03187	R-Square	0.4245
Dependent Mean	60.59582	Adj R-Sq	0.4184
Coeff Var	23.15650		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	-75.66788	11.38624	-6.65
gpa	High School GPA	1	1.70881	0.22629	7.55
english	Mark in HS English	1	-0.38163	0.12390	-3.08
hscalc	Mark in HS Calculus	1	0.34860	0.10091	3.45

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	<.0001
gpa	High School GPA	1	<.0001
english	Mark in HS English	1	0.0023
hscalc	Mark in HS Calculus	1	0.0006

The REG Procedure
Model: MODEL2
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	42127	10532	54.29	<.0001
Error	282	54702	193.97892		
Corrected Total	286	96829			

Root MSE	13.92763	R-Square	0.4351
Dependent Mean	60.59582	Adj R-Sq	0.4271
Coeff Var	22.98448		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	-71.30149	11.46113	-6.22
gpa	High School GPA	1	1.67559	0.22508	7.44
english	Mark in HS English	1	-0.32389	0.12553	-2.58
hscalc	Mark in HS Calculus	1	0.32193	0.10083	3.19
tongue	Mother Tongue	1	-4.98859	2.17680	-2.29

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	<.0001
gpa	High School GPA	1	<.0001
english	Mark in HS English	1	0.0104
hscalc	Mark in HS Calculus	1	0.0016
tongue	Mother Tongue	1	0.0227

The REG Procedure
Model: MODEL3
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	45177	7529.47663	40.82	<.0001
Error	280	51652	184.47234		
Corrected Total	286	96829			

Root MSE	13.58206	R-Square	0.4666
Dependent Mean	60.59582	Adj R-Sq	0.4551
Coeff Var	22.41419		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	-67.81236	11.26953	-6.02
gpa	High School GPA	1	1.57276	0.22101	7.12
english	Mark in HS English	1	-0.29798	0.12258	-2.43
hscalc	Mark in HS Calculus	1	0.22495	0.10213	2.20
tongue	Mother Tongue	1	-4.74223	2.12444	-2.23
precalc	Number precalculus correct	1	1.60480	0.55970	2.87
calc	Number calculus correct	1	0.61933	0.37553	1.65

Parameter Estimates

Variable	Label	DF	Pr > |t|	Type I SS
Intercept	Intercept	1	<.0001	1053822
gpa	High School GPA	1	<.0001	34656
english	Mark in HS English	1	0.0157	4101.93295
hscalc	Mark in HS Calculus	1	0.0284	2349.96005
tongue	Mother Tongue	1	0.0264	1018.76587
precalc	Number precalculus correct	1	0.0045	2548.06067
calc	Number calculus correct	1	0.1002	501.74077

The REG Procedure
Model: MODEL3

Test diagtest Results for Dependent Variable grade

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1524.90072	8.27	0.0003
Denominator	280	184.47234		

The REG Procedure
Model: MODEL4
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	46349	4213.50090	22.95	<.0001
Error	275	50481	183.56584		
Corrected Total	286	96829			

Root MSE	13.54865	R-Square	0.4787
Dependent Mean	60.59582	Adj R-Sq	0.4578
Coeff Var	22.35905		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	-65.51956	11.46394	-5.72
gpa	High School GPA	1	1.58254	0.22205	7.13
english	Mark in HS English	1	-0.36228	0.12803	-2.83
hscalc	Mark in HS Calculus	1	0.22203	0.10254	2.17
precalc	Number precalculus correct	1	1.73222	0.56567	3.06
calc	Number calculus correct	1	0.64614	0.38284	1.69
gender		1	1.81072	1.71680	1.05
tongue	Mother Tongue	1	-4.59977	2.16894	-2.12

e1	1	-1.04782	2.20041	-0.48
e3	1	-1.55625	2.77189	-0.56
e4	1	4.41561	2.45437	1.80
e5	1	3.35044	5.29918	0.63

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	<.0001
gpa	High School GPA	1	<.0001
english	Mark in HS English	1	0.0050
hscalc	Mark in HS Calculus	1	0.0312
precalc	Number precalculus correct	1	0.0024
calc	Number calculus correct	1	0.0926
gender		1	0.2925
tongue	Mother Tongue	1	0.0348
e1		1	0.6343
e3		1	0.5750
e4		1	0.0731
e5		1	0.5277

The REG Procedure
Model: MODEL4

Test diagtest Results for Dependent Variable grade

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1641.89631	8.94	0.0002
Denominator	275	183.56584		

The REG Procedure
Model: MODEL4

Test ethtest Results for Dependent Variable grade

Source	DF	Mean Square	F Value	Pr > F
Numerator	4	245.80285	1.34	0.2556
Denominator	275	183.56584		

The REG Procedure
Model: MODEL4

Test demogr Results for Dependent Variable grade

Source	DF	Mean Square	F Value	Pr > F
Numerator	6	348.47452	1.90	0.0812
Denominator	275	183.56584		

The REG Procedure
Model: MODEL1
Dependent Variable: grade Mark in university calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44881	8976.29511	48.56	<.0001
Error	281	51948	184.86704		
Corrected Total	286	96829			

Root MSE	13.59658	R-Square	0.4635
Dependent Mean	60.59582	Adj R-Sq	0.4540
Coeff Var	22.43815		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-66.75516	11.25053
gpa	High School GPA	1	1.58918	0.22087
english	Mark in HS English	1	-0.30024	0.12270
hscalc	Mark in HS Calculus	1	0.21759	0.10208
totscore	Total # right on diagnostic test	1	0.97213	0.25185
tongue	Mother Tongue	1	-4.69657	2.12640

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-5.93	<.0001
gpa	High School GPA	1	7.20	<.0001
english	Mark in HS English	1	-2.45	0.0150
hscalc	Mark in HS Calculus	1	2.13	0.0339
totscore	Total # right on diagnostic test	1	3.86	0.0001
tongue	Mother Tongue	1	-2.21	0.0280

The REG Procedure
Model: MODEL1
Dependent Variable: gpa High School GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	340.41454	56.73576	1.58	0.1515
Error	361	12954	35.88316		
Corrected Total	367	13294			

Root MSE	5.99026	R-Square	0.0256
Dependent Mean	79.99620	Adj R-Sq	0.0094
Coeff Var	7.48818		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	79.41162	0.92995
gender		1	0.01528	0.63837
tongue	Mother Tongue	1	0.49229	0.79175
e1		1	-0.19311	0.84755
e3		1	-0.40706	1.05468
e4		1	2.23858	0.92800
e5		1	-2.53655	2.16774

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	85.39	<.0001
gender		1	0.02	0.9809
tongue	Mother Tongue	1	0.62	0.5345
e1		1	-0.23	0.8199
e3		1	-0.39	0.6998
e4		1	2.41	0.0164
e5		1	-1.17	0.2427

The REG Procedure
Model: MODEL1
Dependent Variable: english Mark in HS English

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	3316.81454	552.80242	9.06	<.0001
Error	361	22016	60.98594		
Corrected Total	367	25333			

Root MSE	7.80935	R-Square	0.1309
Dependent Mean	76.61957	Adj R-Sq	0.1165
Coeff Var	10.19237		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	73.12162	1.21235
gender		1	3.25392	0.83223
tongue	Mother Tongue	1	3.05651	1.03219
e1		1	-3.22493	1.10493
e3		1	-0.67669	1.37496
e4		1	2.26714	1.20981
e5		1	-2.18009	2.82603

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	60.31	<.0001
gender		1	3.91	0.0001
tongue	Mother Tongue	1	2.96	0.0033
e1		1	-2.92	0.0037
e3		1	-0.49	0.6229
e4		1	1.87	0.0617
e5		1	-0.77	0.4410

The REG Procedure
Model: MODEL1
Dependent Variable: hscalc Mark in HS Calculus

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1548.51604	258.08601	1.70	0.1191
Error	361	54695	151.51013		
Corrected Total	367	56244			

Root MSE	12.30894	R-Square	0.0275
Dependent Mean	75.90489	Adj R-Sq	0.0114
Coeff Var	16.21627		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	74.69566	1.91088
gender		1	-0.76429	1.31175
tongue	Mother Tongue	1	-0.15370	1.62692
e1		1	4.06529	1.74156
e3		1	2.98335	2.16718
e4		1	3.32382	1.90687
e5		1	-2.78482	4.45434

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	39.09	<.0001
gender		1	-0.58	0.5605
tongue	Mother Tongue	1	-0.09	0.9248
e1		1	2.33	0.0201
e3		1	1.38	0.1695
e4		1	1.74	0.0822
e5		1	-0.63	0.5322

The REG Procedure
Model: MODEL1
Dependent Variable: totscore Total # right on diagnostic test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	225.63638	37.60606	2.94	0.0081
Error	361	4613.48319	12.77973		
Corrected Total	367	4839.11957			

Root MSE	3.57488	R-Square	0.0466
Dependent Mean	8.20109	Adj R-Sq	0.0308
Coeff Var	43.59026		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	8.11125	0.55498
gender		1	-0.91030	0.38097
tongue	Mother Tongue	1	0.01588	0.47250
e1		1	1.16384	0.50580
e3		1	0.86651	0.62941
e4		1	1.17287	0.55381
e5		1	-0.17198	1.29367

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	14.62	<.0001
gender		1	-2.39	0.0174
tongue	Mother Tongue	1	0.03	0.9732
e1		1	2.30	0.0220
e3		1	1.38	0.1695
e4		1	2.12	0.0349
e5		1	-0.13	0.8943

The REG Procedure
Model: MODEL1
Multivariate Test: ethnic

Multivariate Statistics and F Approximations

S=4 M=-0.5 N=178

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.90887726	2.17	16	1094.3	0.0047
Pillai's Trace	0.09330831	2.16	16	1444	0.0050
Hotelling-Lawley Trace	0.09786839	2.18	16	710.02	0.0047
Roy's Greatest Root	0.06445782	5.82	4	361	0.0002

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

The REG Procedure
Model: MODEL1
Multivariate Test: gender

Multivariate Statistics and Exact F Statistics

S=1 M=1 N=178

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.93111152	6.62	4	358	<.0001
Pillai's Trace	0.06888848	6.62	4	358	<.0001
Hotelling-Lawley Trace	0.07398521	6.62	4	358	<.0001
Roy's Greatest Root	0.07398521	6.62	4	358	<.0001