STA442s04 Overheads
Set Four (Multiple Regression)

```
/* tvread1.sas  */
options linesize = 79 noovp formdlim='_';
title 'TV Data simulated with SURVEY: Exploratory sample';

proc format;
     value locfmt  1 = 'Rural'
                   2 = 'Small   Town'
                   3 = 'Urban';
data tv;
     infile 'tv1fixed.dat';
     input dist hsehold value q1-q9;
     label value = 'Value of house in $US'
           q1 = 'Number of persons 12 and older'
           q2 = 'Number of persons 11 and younger'
           q3 = 'Numbr TV sets in Household'
           q4 = 'Price willing to pay for cable TV'
           q5 = 'Total TV hours watched last week'
           q6 = 'Hours Public Affairs watched last week'
           q7 = 'Hours Sports watched last week'
           q8 = 'Hours Children''s programming last week'
           q9 = 'Hours Movies watched last week';
     people = q1+q2;
     label people = 'Number of persons in household';
     if 1 <= dist <= 25 then location=1;
        else if 26 <= dist <= 50 then location=2;
        else if 51 <= dist <= 75 then location=3;
        else location = . ;
     format location locfmt.;
     /* Dummy Variables for Location */
     if location = 1 then loc1 = 1;
        else if location = . then loc1 = .;
        else loc1 = 0;
     if location = 2 then loc2 = 1;
        else if location = . then loc2 = .;
        else loc2 = 0;

proc freq;
     tables (loc1 loc2) * location / norow nocol nopercent missing;
```

TABLE OF LOC1 BY LOCATION

LOC1	LOCATION			
Frequency	Rural	Small Town	Urban	Total
0	0	98	311	409
1	91	0	0	91
Total	91	98	311	500

TABLE OF LOC2 BY LOCATION

LOC2	LOCATION			
Frequency	Rural	Small Town	Urban	Total
0	91	0	311	402
1	0	98	0	98
Total	91	98	311	500

```
/* 442s04tvreg1.sas */

title2 'Multiple Regression with TV data';
%include 'tvread1.sas';

/* If you control for number of children in house, does number of TV sets
predict amount of kid's programming watched? */

proc reg;
     model q8 = q2 q3;
     sets: test q3=0;

/* Is location related to total number of TV hourse watched? Do it with both
proc glm and proc reg to check.  */

proc glm;
     class location;
     model q5 = location;
     means location;

proc reg;
     model q5 = loc1 loc2;

/* Controlling for number of people in household, is location related to total
number of TV hours watched?  */

proc reg;
     model q5 = people loc1 loc2;
     loctest: test loc1=loc2=0;

/* Re-do the preceding question to answer some additional questions, in more
than one way.

     1. Using proc reg, fit a full and a reduced model to find the proportion
of remaining variation explained by location, once number of people in the
household is taken into account.

               (0.5836-0.5273)/(1-0.5273) = 0.1191030

     2. Obtain the same information from Type I (sequential) sums of squares.

                [SS1(loc1)+SS1(loc2)]/(SSTO-SS1(people)) =
                (20938+28687)/(882258.52830-465254) = 0.1190035

     3. Obtain the same information from the F statistic and degrees of
freedom.

       F*s / ( F*s + n - p ) = 31.9456*2 / (31.9456*2 + 473) = 0.1190021
*/
```

```
proc reg simple;
     model q5 = people;
     model q5 = people loc1 loc2 / ss1;

/*
     4. Using output from proc reg, find the mean number of TV hours watched
for each location, CORRECTED for number of people in the household. Use proc
iml as a calculator.
*/

proc iml;
     b0 = -5.851272 ; b1 = 15.852576;
     b2 = 21.967611 ; b3 = 20.362185;
     xbar = 3.4088;
     rural = b0 + b1*xbar + b2*1 + b3*0;
     small = b0 + b1*xbar + b2*0 + b3*1;
     urban = b0 + b1*xbar + b2*0 + b3*0;
     print rural small urban;

/*
     5. Obtain the same information from proc glm, using lsmeans.
*/

proc glm;
     class location;
     model q5 = people location;
     lsmeans location;

/* If we allow for number of children, number of adults, number of TVs,
location and value of house, is amount of kids TV related to amount(s) of
other TV watched?  */

proc reg;
     model q8 = q2 q1 q3 loc1 loc2 value
                q6 q7 q9 / ss1;
     otherTV: test q6=q7=q9=0;

/* Using proc iml as a calculator, find the proportion of remaining variation
explained by hours of Public Affairs, Sports and Movies, once we control for
the other variables in the model.  */

proc iml;
     F = 18.6178; ndf = 3; ddf = 468;
     a = F*ndf / (F*ndf + ddf);
     print a;
```

```
/* If you control for number of children in house, does number of TV sets
predict amount of kid's programming watched? */

proc reg;
     model q8 = q2 q3;
     sets: test q3=0;
```

```
              TV Data simulated with SURVEY: Exploratory sample              1
                                          13:21 Monday, February 2, 2004

Model: MODEL1
Dependent Variable: Q8         Hours Children's programming last week

                           Analysis of Variance

                                Sum of         Mean
      Source           DF      Squares       Square      F Value       Prob>F

      Model             2  15272.06369   7636.03184      590.441       0.0001
      Error           474   6130.12499     12.93275
      C Total         476  21402.18868

          Root MSE       3.59621     R-square       0.7136
          Dep Mean       4.78616     Adj R-sq       0.7124
          C.V.          75.13771

                           Parameter Estimates

                    Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1     -0.148654    0.32377355        -0.459        0.6464
     Q2         1      4.790295    0.15114550        31.693        0.0001
     Q3         1      0.672812    0.15038853         4.474        0.0001

                  Variable
     Variable  DF     Label

     INTERCEP   1  Intercept
     Q2         1  Number of persons 11 and younger
     Q3         1  Numbr TV sets in Household
```

```
Dependent Variable: Q8
Test: SETS     Numerator:    258.8506  DF:    1   F value:  20.0151
               Denominator:  12.93275  DF:  474   Prob>F:    0.0001
```

```
/* Is location related to total number of TV hourse watched? Do it with both
proc glm and proc reg to check.  */

proc glm;
     class location;
     model q5 = location;
     means location;

proc reg;
     model q5 = loc1 loc2;
```

General Linear Models Procedure
Class Level Information

Class	Levels	Values
LOCATION	3	Rural Small Town Urban

Number of observations in data set = 500

NOTE: Due to missing values, only 477 observations can be used in this analysis.

General Linear Models Procedure

Dependent Variable: Q5 Total TV hours watched last week

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	100416.66787	50208.33393	30.44	0.0001
Error	474	781841.86043	1649.45540		
Corrected Total	476	882258.52830			

R-Square	C.V.	Root MSE	Q5 Mean
0.113818	72.13522	40.613488	56.301887

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LOCATION	2	100416.66787	50208.33393	30.44	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LOCATION	2	100416.66787	50208.33393	30.44	0.0001

General Linear Models Procedure

Level of LOCATION	N	Q5 Mean	Q5 SD
Rural	90	72.8333333	45.9017637
Small Town	93	76.3870968	47.5936222
Urban	294	44.8877551	36.2926715

Model: MODEL1
Dependent Variable: Q5 Total TV hours watched last week

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	100416.66787	50208.33393	30.439	0.0001
Error	474	781841.86043	1649.45540		
C Total	476	882258.52830			

Root MSE	40.61349	R-square	0.1138
Dep Mean	56.30189	Adj R-sq	0.1101
C.V.	72.13522		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	44.887755	2.36862672	18.951	0.0001
LOC1	1	27.945578	4.89261431	5.712	0.0001
LOC2	1	31.499342	4.83181872	6.519	0.0001

```
/* Controlling for number of people in household, is location related to total
number of TV hours watched?  */

proc reg;
     model q5 = people loc1 loc2;
     loctest: test loc1=loc2=0;
```

Model: MODEL1
Dependent Variable: Q5 Total TV hours watched last week

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	514878.66820	171626.22273	220.968	0.0001
Error	473	367379.86010	776.70161		
C Total	476	882258.52830			

Root MSE	27.86937	R-square	0.5836
Dep Mean	56.30189	Adj R-sq	0.5810
C.V.	49.49988		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-5.851272	2.73246377	-2.141	0.0328
PEOPLE	1	15.852576	0.68625352	23.100	0.0001
LOC1	1	21.967611	3.36731784	6.524	0.0001
LOC2	1	20.362185	3.35050985	6.077	0.0001

Variable	DF	Variable Label
INTERCEP	1	Intercept
PEOPLE	1	Number of persons in household
LOC1	1	
LOC2	1	

Dependent Variable: Q5
Test: LOCTEST Numerator: 24812.1707 DF: 2 F value: 31.9456
Denominator: 776.7016 DF: 473 Prob>F: 0.0001

```
/* Re-do the preceding question to answer some additional questions, in more
than one way.

     1. Using proc reg, fit a full and a reduced model to find the proportion
of remaining variation explained by location, once number of people in the
household is taken into account.

               (0.5836-0.5273)/(1-0.5273) = 0.1191030

     2. Obtain the same information from Type I (sequential) sums of squares.

                [SS1(loc1)+SS1(loc2)]/(SSTO-SS1(people)) =
                (20938+28687)/(882258.52830-465254) = 0.1190035

     3. Obtain the same information from the F statistic and degrees of
freedom.

       F*s / ( F*s + n - p ) = 31.9456*2 / (31.9456*2 + 473) = 0.1190021
*/


proc reg simple;
     model q5 = people;
     model q5 = people loc1 loc2 / ss1;
```

TV Data simulated with SURVEY: Exploratory sample 9
13:21 Monday, February 2, 2004

Descriptive Statistics

Variables	Sum	Mean	
INTERCEP	477	1	Intercept
PEOPLE	1626	3.4088050314	Number of persons in household
Q5	26856	56.301886792	Total TV hours watched last week
LOC1	90	0.1886792453	
LOC2	93	0.1949685535	

Variables	Uncorrected SS	Variance	
INTERCEP	477	0	Intercept
PEOPLE	7230	3.5447122245	Number of persons in household
Q5	2394302	1853.4843032	Total TV hours watched last week
LOC1	90	0.153400983	
LOC2	93	0.1572855557	

```
Variables    Std Deviation

INTERCEP                0   Intercept
PEOPLE         1.8827406153   Number of persons in household
Q5             43.052111483   Total TV hours watched last week
LOC1           0.3916643755
LOC2           0.3965924303
```

(Reduced model - just people)

```
              TV Data simulated with SURVEY: Exploratory sample            10
                                              13:21 Monday, February 2, 2004

Model: MODEL1
Dependent Variable: Q5         Total TV hours watched last week

                           Analysis of Variance

                              Sum of         Mean
  Source           DF      Squares       Square      F Value       Prob>F

  Model             1 465254.32676 465254.32676      529.961       0.0001
  Error           475 417004.20154    877.90358
  C Total         476 882258.52830

      Root MSE      29.62944     R-square       0.5273
      Dep Mean      56.30189     Adj R-sq       0.5263
      C.V.          52.62601

                          Parameter Estimates

                   Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1     -0.302932    2.80827350        -0.108        0.9141
PEOPLE     1     16.605473    0.72132244        23.021        0.0001

               Variable
Variable  DF     Label

INTERCEP   1  Intercept
PEOPLE     1  Number of persons in household
```

Full model - people and location

TV Data simulated with SURVEY: Exploratory sample 11
13:21 Monday, February 2, 2004

Model: MODEL2
Dependent Variable: Q5 Total TV hours watched last week

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	514878.66820	171626.22273	220.968	0.0001
Error	473	367379.86010	776.70161		
C Total	476	882258.52830			

Root MSE	27.86937	R-square	0.5836
Dep Mean	56.30189	Adj R-sq	0.5810
C.V.	49.49988		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-5.851272	2.73246377	-2.141	0.0328
PEOPLE	1	15.852576	0.68625352	23.100	0.0001
LOC1	1	21.967611	3.36731784	6.524	0.0001
LOC2	1	20.362185	3.35050985	6.077	0.0001

Variable	DF	Type I SS	Variable Label
INTERCEP	1	1512043	Intercept
PEOPLE	1	465254	Number of persons in household
LOC1	1	20938	
LOC2	1	28687	

```
/*
     4. Using output from proc reg, find the mean number of TV hours watched
for each location, CORRECTED for number of people in the household. Use proc
iml as a calculator.
*/

proc iml;
     b0 = -5.851272 ; b1 = 15.852576;
     b2 = 21.967611 ; b3 = 20.362185;
     xbar = 3.4088;
     rural = b0 + b1*xbar + b2*1 + b3*0;
     small = b0 + b1*xbar + b2*0 + b3*1;
     urban = b0 + b1*xbar + b2*0 + b3*0;
     print rural small urban;

/*
```

```
               TV Data simulated with SURVEY: Exploratory sample             12
                                          13:21 Monday, February 2, 2004

                            RURAL     SMALL     URBAN
                          70.1546 68.549174 48.186989
```

```
     5. Obtain the same information from proc glm, using lsmeans.

*/

proc glm;
     class location;
     model q5 = people location;
     lsmeans location;
```

```
               TV Data simulated with SURVEY: Exploratory sample             13
                                          13:21 Monday, February 2, 2004

                      General Linear Models Procedure
                          Class Level Information

                   Class    Levels    Values

                   LOCATION      3    Rural Small   Town Urban


                    Number of observations in data set = 500
```

NOTE: Due to missing values, only 477 observations can be used in this
      analysis.

TV Data simulated with SURVEY: Exploratory sample 14
13:21 Monday, February 2, 2004

General Linear Models Procedure

Dependent Variable: Q5 Total TV hours watched last week

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	514878.66820	171626.22273	220.97	0.0001
Error	473	367379.86010	776.70161		
Corrected Total	476	882258.52830			

R-Square	C.V.	Root MSE	Q5 Mean
0.583592	49.49988	27.869367	56.301887

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PEOPLE	1	465254.32676	465254.32676	599.01	0.0001
LOCATION	2	49624.34144	24812.17072	31.95	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PEOPLE	1	414462.00034	414462.00034	533.62	0.0001
LOCATION	2	49624.34144	24812.17072	31.95	0.0001

General Linear Models Procedure
Least Squares Means

LOCATION	Q5 LSMEAN
Rural	70.1546800
Small Town	68.5492539
Urban	48.1870687

```
/* If we allow for number of children, number of adults, number of TVs,
location and value of house, is amount of kids TV related to amount(s) of
other TV watched?  */

proc reg;
     model q8 = q2 q1 q3 loc1 loc2 value
                q6 q7 q9 / ss1;
     otherTV: test q6=q7=q9=0;
```

Model: MODEL1
Dependent Variable: Q8 Hours Children's programming last week

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	9	15763.21411	1751.46823	145.050	0.0001
Error	467	5638.97457	12.07489		
C Total	476	21402.18868			

Root MSE	3.47489	R-square	0.7365
Dep Mean	4.78616	Adj R-sq	0.7314
C.V.	72.60292		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.288128	0.79658863	1.617	0.1065
Q2	1	4.367102	0.16914647	25.818	0.0001
Q1	1	-0.791774	0.16207500	-4.885	0.0001
Q3	1	0.372331	0.16861744	2.208	0.0277
LOC1	1	-1.111007	0.47197706	-2.354	0.0190
LOC2	1	-0.805423	0.49157217	-1.638	0.1020
VALUE	1	-0.000006785	0.00001151	-0.590	0.5557
Q6	1	0.174388	0.09668258	1.804	0.0719
Q7	1	0.022382	0.01593602	1.404	0.1608
Q9	1	0.075467	0.03546210	2.128	0.0339

Variable	DF	Type I SS	Variable Label
INTERCEP	1	10927	Intercept
Q2	1	15013	Number of persons 11 and younger
Q1	1	6.195326	Number of persons 12 and older
Q3	1	307.119363	Numbr TV sets in Household
LOC1	1	0.132850	
LOC2	1	8.199659	
VALUE	1	4.336034	Value of house in $US
Q6	1	332.562166	Hours Public Affairs watched last week
Q7	1	36.770619	Hours Sports watched last week
Q9	1	54.685042	Hours Movies watched last week

Dependent Variable: Q8

Test: OTHERTV	Numerator:	141.3393	DF:	3	F value:	11.7052
	Denominator:	12.07489	DF:	467	Prob>F:	0.0001

```
/* Using proc iml as a calculator, find the proportion of remaining variation
explained by hours of Public Affairs, Sports and Movies, once we control for
the other variables in the model.  */

proc iml;
     F = 18.6178; ndf = 3; ddf = 468;
     a = F*ndf / (F*ndf + ddf);
     print a;
```

A
0.1066203