STA431s09 SAS Handout 1: Basics with the SENIC data

Initially, there is nothing in the subdirectory 431.

```
tuzo.utm.utoronto.ca:~/431% ls
```

Now I go to the data file with my web browser, copy the URL (address), and

```
tuzo.utm.utoronto.ca:~/431% curl
http://fisher.utstat.toronto.edu/~brunner/431s09/code_n_data/senic.data > senic.data
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  5989  100  5989    0     0   479k      0 --:--:-- --:--:-- --:--:-- 4710k
tuzo.utm.utoronto.ca:~/431% ls
senic.data
```

Always take a look! In less, spacebar gives another page and q quits.

```
tuzo.utm.utoronto.ca:~/431% less senic.data
    1  7.13 55.7 4.1  9.0  39.6 279 2 4 207 241 60.0
    2  8.82 58.2 1.6  3.8  51.7  80 2 2  51  52 40.0
    3  8.34 56.9 2.7  8.1  74.0 107 2 3  82  54 20.0
    4  8.95 53.7 5.6 18.9 122.8 147 2 4  53 148 40.0
    5 11.20 56.5 5.7 34.5  88.9 180 2 1 134 151 40.0

                          Skipping ...

  111  7.70 56.9 4.4 12.2  67.9 129 2 4  85 136 62.9
  112 17.94 56.2 5.9 26.4  91.8 835 1 1 791 407 62.9
  113  9.41 59.5 3.1 20.6  91.7  29 2 3  20  22 22.9
```

Now create the program `senic1.sas` with `emacs`. Type “emacs senic1.sas” and Enter. If the file `senic1.sas` already existed, you'd be editing it. Since it does not yet exist, you get an empty file. Start typing. If you are unfamiliar with `emacs`, print a copy of the shorter handout and use it while you work.

After you are done, take a look.

```
tuzo.utm.utoronto.ca:~/431% less senic1.sas
```

```
/*****  senic1.sas  Read and describe SENIC data  (No missing values) ******/
options linesize=79 noovp formdlim='_' ;
title 'Read and Describe SENIC data';

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;

data infect;
     infile 'senic.data';
     input id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

     label id       = 'Hospital identification number'
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infrisk  = 'Prob of acquiring infection in hospital'
           culratio = '# cultures  / # no hosp acq infect'
           xratio   = '# x-rays / # no signs of pneumonia'
           nbeds    = 'Average # beds during study period'
           medschl  = 'Medical school affiliation'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;
        /* Associating variables with their value labels */
     format medschl yesnofmt.;
     format region  regfmt.;

  /* Dummy variables (There are no missing values) */
     if region = 1 then r1=1; else r1=0;
     if region = 2 then r2=1; else r2=0;
     if region = 3 then r3=1; else r3=0;
     if region = 4 then r4=1; else r4=0;

     if medschl = 2 then mschool = 0; else mschool = medschl;
     /* mschool is an indicator for medical school = yes */

proc means;
     title2 'Basic Descriptive Stats for Quantitative Vars';
     var stay age infrisk culratio xratio nbeds census nurses service;

proc freq;
     title2 'Frequency distributions for Categorical Variables';
     tables region medschl;

proc freq;
     title2 'Check Dummy Variables';
     tables (r1-r4) * region / norow nocol nopercent missing;
     tables mschool * medschl / norow nocol nopercent missing;
```

Now run SAS:

```
tuzo.utm.utoronto.ca:~/431% ls
senic1.sas  senic.data
tuzo.utm.utoronto.ca:~/431% sas senic1
tuzo.utm.utoronto.ca:~/431% ls
senic1.log  senic1.lst  senic1.sas  senic.data
```

Look at the log file

```
tuzo.utm.utoronto.ca:~/431% cat senic1.log
1                                                          The SAS System

NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M3)
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987001.
NOTE: This session is executing on the Linux 2.6.9-67.ELsmp platform.



NOTE: SAS 9.1.3 Service Pack 3

You are running SAS 9. Some SAS 8 files will be automatically converted
by the V9 engine; others are incompatible.  Please see
http://support.sas.com/rnd/migration/planning/platform/64bit.html

PROC MIGRATE will preserve current SAS file attributes and is
recommended for converting all your SAS libraries from any
SAS 8 release to SAS 9.  For details and examples, please see
http://support.sas.com/rnd/migration/index.html


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the file "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.11 seconds
      cpu time            0.02 seconds

1          /*****  senic1.sas  Read and describe SENIC data  (No missing values) ******/
2          options linesize=79 noovp formdlim='_' ;
3          title 'Read and Describe SENIC data';
4
5          proc format;
5        !               /* value labels used in data step below */
6                value yesnofmt 1 = 'Yes'   2 = 'No' ;
NOTE: Format YESNOFMT has been output.
7                value regfmt  1 =  'Northeast'
8                              2 =  'North Central'
9                              3 =  'South'
10                             4 =  'West' ;
```

```
NOTE: Format REGFMT has been output.
11

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds


12         data infect;
13              infile 'senic.data';
14              input id stay age infrisk culratio xratio nbeds medschl
15                       region census nurses service;
16
17              label id       = 'Hospital identification number'
18                    stay     = 'Av length of hospital stay, in days'

2                                The SAS System

19                    age      = 'Average patient age'
20                    infrisk  = 'Prob of acquiring infection in hospital'
21                    culratio = '# cultures  / # no hosp acq infect'
22                    xratio   = '# x-rays / # no signs of pneumonia'
23                    nbeds    = 'Average # beds during study period'
24                    medschl  = 'Medical school affiliation'
25                    region   = 'Region of country (usa)'
26                    census   = 'Aver # patients in hospital per day'
27                    nurses   = 'Aver # nurses during study period'
28                    service  = '% of 35 potential facil. & services' ;
29                 /* Associating variables with their value labels */
30              format medschl yesnofmt.;
31              format region  regfmt.;
32
33            /* Dummy variables (There are no missing values) */
34              if region = 1 then r1=1; else r1=0;
35              if region = 2 then r2=1; else r2=0;
36              if region = 3 then r3=1; else r3=0;
37              if region = 4 then r4=1; else r4=0;
38
39              if medschl = 2 then mschool = 0; else mschool = medschl;
40              /* mschool is an indicator for medical school = yes */
41

NOTE: The infile 'senic.data' is:
      File Name=/student/cslec/brunnerj/431/senic.data,
      Owner Name=brunnerj,Group Name=lecturers,
      Access Permission=rw-------,
      File Size (bytes)=5989

NOTE: 113 records were read from the infile 'senic.data'.
      The minimum record length was 52.
      The maximum record length was 52.
NOTE: The data set WORK.INFECT has 113 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.16 seconds
      cpu time            0.01 seconds
```

```
42         proc means;
43              title2 'Basic Descriptive Stats for Quantitative Vars';
44              var stay age infrisk culratio xratio nbeds census nurses
44       ! service;
45

NOTE: There were 113 observations read from the data set WORK.INFECT.
NOTE: The PROCEDURE MEANS printed page 1.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.12 seconds
      cpu time            0.03 seconds


46         proc freq;
47              title2 'Frequency distributions for Categorical Variables';
48              tables region medschl;
49

NOTE: There were 113 observations read from the data set WORK.INFECT.

3                               The SAS System

NOTE: The PROCEDURE FREQ printed page 2.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


50         proc freq;
51              title2 'Check Dummy Variables';
52              tables (r1-r4) * region / norow nocol nopercent missing;
53              tables mschool * medschl / norow nocol nopercent missing;
NOTE: There were 113 observations read from the data set WORK.INFECT.
NOTE: The PROCEDURE FREQ printed pages 3-4.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           0.52 seconds
      cpu time            0.08 seconds
```

Now look at the list file.

```
tuzo.utm.utoronto.ca:~/431% cat senic1.lst
```

The MEANS Procedure

Variable	Label	N	Mean
stay	Av length of hospital stay, in days	113	9.6483186
age	Average patient age	113	53.2318584
infrisk	Prob of acquiring infection in hospital	113	4.3548673
culratio	# cultures / # no hosp acq infect	113	15.6840708
xratio	# x-rays / # no signs of pneumonia	113	81.6300885
nbeds	Average # beds during study period	113	252.1769912
census	Aver # patients in hospital per day	113	191.3716814
nurses	Aver # nurses during study period	113	173.2477876
service	% of 35 potential facil. & services	113	43.1548673

Variable	Label	Std Dev	Minimum
stay	Av length of hospital stay, in days	1.9114560	6.7000000
age	Average patient age	4.4616074	38.8000000
infrisk	Prob of acquiring infection in hospital	1.3409080	1.3000000
culratio	# cultures / # no hosp acq infect	10.1830441	1.6000000
xratio	# x-rays / # no signs of pneumonia	19.3667373	39.6000000
nbeds	Average # beds during study period	192.8451558	29.0000000
census	Aver # patients in hospital per day	153.7595639	20.0000000
nurses	Aver # nurses during study period	139.2653897	14.0000000
service	% of 35 potential facil. & services	15.2001879	5.7000000

Variable	Label	Maximum
stay	Av length of hospital stay, in days	19.5600000
age	Average patient age	65.9000000
infrisk	Prob of acquiring infection in hospital	7.8000000
culratio	# cultures / # no hosp acq infect	60.5000000
xratio	# x-rays / # no signs of pneumonia	133.5000000
nbeds	Average # beds during study period	835.0000000
census	Aver # patients in hospital per day	791.0000000
nurses	Aver # nurses during study period	656.0000000
service	% of 35 potential facil. & services	80.0000000

Frequency distributions for Categorical Variables

The FREQ Procedure

Region of country (usa)

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Northeast	29	25.66	29	25.66
North Central	32	28.32	61	53.98
South	36	31.86	97	85.84
West	16	14.16	113	100.00

Medical school affiliation

medschl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	17	15.04	17	15.04
No	96	84.96	113	100.00

Check Dummy Variables

The FREQ Procedure

Table of r1 by region

r1 region(Region of country (usa))

Frequency	Northeast	North Central	South	West	Total
0	0	32	36	16	84
1	29	0	0	0	29
Total	29	32	36	16	113

Table of r2 by region

r2 region(Region of country (usa))

Frequency	Northeast	North Central	South	West	Total
0	29	0	36	16	81
1	0	32	0	0	32
Total	29	32	36	16	113

Table of r3 by region

r3 region(Region of country (usa))

Frequency	Northeast	North Central	South	West	Total
0	29	32	0	16	77
1	0	0	36	0	36
Total	29	32	36	16	113

Table of r4 by region

r4 region(Region of country (usa))

Frequency	Northeast	North Central	South	West	Total
0	29	32	36	0	97
1	0	0	0	16	16
Total	29	32	36	16	113

Table of mschool by medschl

mschool medschl(Medical school affiliation)

Frequency	Yes	No	Total
0	0	96	96
1	17	0	17
Total	17	96	113

Mail yourself the log and list files

```
tuzo.utm.utoronto.ca:~/431% mail brunner@utstat.toronto.edu < senic1.log
tuzo.utm.utoronto.ca:~/431% mail brunner@utstat.toronto.edu < senic1.lst
```

Now do some regression. Always avoid re-typing when possible.

```
tuzo.utm.utoronto.ca:~/431% cp senic1.sas senic2.sas
tuzo.utm.utoronto.ca:~/431% emacs senic2.sas


/*****  senic2.sas  Basic multiple regression ******/
options linesize=79 pagesize=100 noovp formdlim='_' nodate;
title 'Mulltiple regression on SENIC data';

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;

data infect;
     infile 'senic.data';
     input id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

     label id       = 'Hospital identification number'
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infrisk  = 'Prob of acquiring infection in hospital'
           culratio = '# cultures  / # no hosp acq infect'
           xratio   = '# x-rays / # no signs of pneumonia'
           nbeds    = 'Average # beds during study period'
           medschl  = 'Medical school affiliation'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;
        /* Associating variables with their value labels */
     format medschl yesnofmt.;
     format region  regfmt.;

  /* Dummy variables (There are no missing values) */
     if region = 1 then r1=1; else r1=0;
     if region = 2 then r2=1; else r2=0;
     if region = 3 then r3=1; else r3=0;
     if region = 4 then r4=1; else r4=0;

     if medschl = 2 then mschool = 0; else mschool = medschl;
     /* mschool is an indicator for medical school = yes */
```

```
proc reg;
     model infrisk = stay age culratio xratio nbeds census nurses service
                     mschool r1 r2 r3;
     region: test r1=r2=r3=0;
             /* Test the corresponding regression coefficients */
```

Mulltiple regression on SENIC data 1

The REG Procedure
Model: MODEL1
Dependent Variable: infrisk Prob of acquiring infection in hospital

Number of Observations Read	113
Number of Observations Used	113

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	119.41108	9.95092	12.14	<.0001
Error	100	81.96875	0.81969		
Corrected Total	112	201.37982			

Root MSE	0.90537	R-Square	0.5930
Dependent Mean	4.35487	Adj R-Sq	0.5441
Coeff Var	20.78975		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-0.60602	1.20324
stay	Av length of hospital stay, in days	1	0.21771	0.06954
age	Average patient age	1	0.01616	0.02164
culratio	# cultures / # no hosp acq infect	1	0.05710	0.01052
xratio	# x-rays / # no signs of pneumonia	1	0.01060	0.00521
nbeds	Average # beds during study period	1	-0.00321	0.00265
census	Aver # patients in hospital per day	1	0.00388	0.00342
nurses	Aver # nurses during study period	1	0.00150	0.00168
service	% of 35 potential facil. & services	1	0.01993	0.00997
mschool		1	-0.66234	0.31874

```
r1                                              1       -0.98705        0.33065
r2                                              1       -0.69907        0.29510
r3                                              1       -0.82653        0.28739
```

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-0.50	0.6156
stay	Av length of hospital stay, in days	1	3.13	0.0023
age	Average patient age	1	0.75	0.4570
culratio	# cultures / # no hosp acq infect	1	5.43	<.0001
xratio	# x-rays / # no signs of pneumonia	1	2.04	0.0444
nbeds	Average # beds during study period	1	-1.21	0.2288
census	Aver # patients in hospital per day	1	1.14	0.2584
nurses	Aver # nurses during study period	1	0.89	0.3749
service	% of 35 potential facil. & services	1	2.00	0.0484
mschool		1	-2.08	0.0403
r1		1	-2.99	0.0036
r2		1	-2.37	0.0198
r3		1	-2.88	0.0049

The REG Procedure
Model: MODEL1

Test region Results for Dependent Variable infrisk

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	2.87046	3.50	0.0182
Denominator	100	0.81969		

Mail yourself the log and list files

```
tuzo.utm.utoronto.ca:~/431% mail brunner@utstat.toronto.edu < senic2.log
tuzo.utm.utoronto.ca:~/431% mail brunner@utstat.toronto.edu < senic2.lst
```