An Example of Regression Modeling

Here is part of a set of data for 38 countries on female and male life expectancy, the number of physicians per person, and the number of televisions per person:

```
MTB > print 'fle' 'mle' 'physpp' 'tvpp'

 ROW   fle   mle      physpp       tvpp

   1    74    67   0.0027027   0.250000
   2    53    54   0.0001622   0.003175
   3    68    62   0.0014620   0.250000
   4    80    73   0.0022272   0.588235
   5    72    68   0.0015552   0.125000
   6    74    68   0.0006447   0.178571
   7    61    60   0.0016234   0.066667
   8    53    50   0.0000273   0.001988
   9    82    74   0.0024814   0.384615
  10    79    73   0.0028902   0.384615
  11    58    57   0.0004047   0.022727
  12    63    59   0.0001346   0.041667
  13    65    64   0.0003342   0.043478
  14    82    75   0.0042918   0.263158
                      .
                      .
                      .
```

Scatterplot of Female Life Expectancy Versus Physicians Per Person

```
MTB > plot 'fle' 'physpp'

      -                      *         *             *        *
    80+                     *       *  *    *
      -              *        *   *
fle   -                      **                                *
      -         *   *                     2             *
      -    *       *       *
    70+
      -     *    *  2     *
      -    **
      -   2
      -                     *
    60+
      -      2
      -     *
      - ***
      -
    50+
        +---------+---------+---------+---------+---------+------
     0.00000   0.00080   0.00160   0.00240   0.00320   0.00400

                              physpp
```

Regression of Female Life Expectancy on Physicians Per Person

```
MTB > regress 'fle' 1 'physpp'

The regression equation is
fle = 62.9 + 5023 physpp

Predictor       Coef       Stdev    t-ratio        p
Constant      62.948       1.594      39.50    0.000
physpp        5023.4       834.7       6.02    0.000

s = 6.289       R-sq = 50.2%     R-sq(adj) = 48.8%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      1432.4      1432.4     36.22    0.000
Error        36      1423.8        39.5
Total        37      2856.2

Unusual Observations
Obs.  physpp       fle        Fit Stdev.Fit  Residual  St.Resid
 14  0.00429     82.00      84.51      2.57     -2.51    -0.44 X
 34  0.00442     75.00      85.18      2.67    -10.18    -1.79 X

X denotes an obs. whose X value gives it large influence.
```

Residual Plot

```
MTB > regress 'fle' 1 'physpp';
SUBC> residuals c10.
...
MTB > plot c10 'physpp'

C10   -                      *
      -              *
      -         *           *
    7.0+    *                  *        *
      -             *               *
      -            *         *    *    *
      -     *                 *
      -    *               *                *        *
    0.0+  2 *    *  *
      -             *
      -                   *               *                   *
      -                                   *
      -
   -7.0+     *
      -     **                                          *
      - **                  *                                  *
      -   *
        +---------+---------+---------+---------+---------+------
     0.00000   0.00080   0.00160   0.00240   0.00320   0.00400

                              physpp
```

Does this plot give you any reason to be concerned about using the linear regression model?

Scatterplot of Female Life Expectancy Versus Televisions Per Person

```
MTB > plot 'fle' 'tvpp'

       -                   *       2          *
     80+                       *   *            *           *
       -             *    *   *
fle    -           **          *
       -  *          * *  *   *
       -       * *    *
     70+
       -   *  ** *        *
       -    2
       -  * *
       -     *
     60+
       -  **
       - *
       - 2  *
       -
     50+
        +---------+---------+---------+---------+---------+------
      0.00      0.15      0.30      0.45      0.60      0.75

                                  tvpp
```

Regression of Female Life Expectancy on Televisions Per Person

```
MTB > regress 'fle' 1 'tvpp';
SUBC> residuals c11.

The regression equation is
fle = 63.2 + 37.0 tvpp

Predictor       Coef       Stdev    t-ratio        p
Constant      63.223       1.356      46.62    0.000
tvpp          36.966       5.156       7.17    0.000

s = 5.716       R-sq = 58.8%     R-sq(adj) = 57.7%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      1679.8      1679.8     51.40    0.000
Error        36      1176.4        32.7
Total        37      2856.2

Unusual Observations
Obs.    tvpp       fle       Fit Stdev.Fit  Residual  St.Resid
 36    0.769    79.000    91.658     3.118   -12.658    -2.64RX

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

Residual Plot

```
MTB > regress 'fle' 1 'tvpp';
SUBC> residuals c11.
...
MTB > plot c11 'tvpp'

       -  *
       -             *     *
    7.0+           *
       -            *
C11    -       * *   *    *        2
       -   *           *      **
       -    * *       *   *        *
    0.0+  * *  * *             *
       -    *                 *               *
       -
       -     *            *
       -   *                                    *
   -7.0+ **
       -
       - 2
       -    *
       -                                                    *
        +---------+---------+---------+---------+---------+------
      0.00      0.15      0.30      0.45      0.60      0.75

                                 tvpp
```

Where is the influential outlier? Is the relationship really linear? Are there outliers if you don't regard the relationship as linear?

Transforming Televisions Per Person

The relationship of female life expectancy to TVs per person is clearly not linear. We can try transforming the explanatory variable to get a better fit — eg, by taking the square root:

```
MTB > let c8 = sqrt('tvpp')
MTB > name c8 'sqtvpp'
MTB > plot 'fle' 'sqtvpp'

       -                                 *      2       *
     80+                                     *  *        *      *
       -                           *     *  *
fle    -                         * *         *
       -       *                   * *  *   *
       -                   *  *     *
     70+
       -            *    * * *         *
       -              **
       -      *       *
       -                 *
     60+
       -       * *
       -    *
       -    **        *
       -
     50+
        +---------+---------+---------+---------+---------+------
      0.00      0.16      0.32      0.48      0.64      0.80

                                sqtvpp
```

A Regression Model Using the Transformed Variable

```
MTB > regress 'fle' 1 'sqtvpp';
SUBC> residuals c12.

The regression equation is
fle = 56.9 + 35.0 sqtvpp

Predictor       Coef       Stdev    t-ratio        p
Constant      56.934       1.522      37.42    0.000
sqtvpp        34.960       3.473      10.06    0.000

s = 4.561       R-sq = 73.8%     R-sq(adj) = 73.1%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      2107.3      2107.3    101.30    0.000
Error        36       748.9        20.8
Total        37      2856.2

Unusual Observations
Obs.  sqtvpp       fle       Fit Stdev.Fit  Residual  St.Resid
 17    0.105    73.000    60.619     1.215    12.381     2.82R
 30    0.209    54.000    64.224     0.956   -10.224    -2.29R
 36    0.877    79.000    87.596     1.870    -8.596    -2.07RX

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

Transforming Physicians Per Person

We can try transforming the the physicians per person variable in the same way:

```
MTB > let c9 = sqrt('physpp')
MTB > name c9 'sqphyspp'
MTB > plot 'fle' 'sqphyspp'

       -                                 *       *       *    *
     80+                                 *    * *   *
       -                          *       *  *
 fle   -                                 **                   *
       -                    *    *                2        *
       -           *           *        *
     70+
       -              *       *  2     *
       -           *  *
       -         2
       -                                 *
     60+
       -                2
       -             *
       -   *  *   *
       -
     50+
         --------+---------+---------+---------+---------+--------
             0.012     0.024     0.036     0.048     0.060

                               sqphyspp
```

A Regression Model Using the Transformed Variable

```
MTB > regress 'fle' 1 'sqphyspp';
SUBC> residuals c13.

The regression equation is
fle = 56.4 + 403 sqphyspp

Predictor       Coef       Stdev    t-ratio        p
Constant      56.435       2.046      27.58    0.000
sqphyspp      403.45       53.43       7.55    0.000

s = 5.541       R-sq = 61.3%     R-sq(adj) = 60.2%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      1750.9      1750.9     57.02    0.000
Error        36      1105.3        30.7
Total        37      2856.2

Unusual Observations
Obs.sqphyspp       fle       Fit Stdev.Fit  Residual   St.Resid
  7   0.0403    61.000    72.691     0.952   -11.691     -2.14R

R denotes an obs. with a large st. resid.
```

Residual Plot

```
MTB > plot c13 'sqphyspp'

       -                                 *
       -           *              *
    7.0+                    *
       -                                 *       *
 C13   -           *           * *        *   *
       -              *                  *   *  *
       -         2    *                   *         *    *
    0.0+                      *         *
       -                         2                            *
       -                                          *
       -                               *          *
       -   *  *
   -7.0+             *  *                                  *
       -          *     *                                     *
       -
       -                                 *
       -
         --------+---------+---------+---------+---------+--------
             0.012     0.024     0.036     0.048     0.060

                                 sqphyspp
```

What Do These Models Mean?

- What (if anything) do these regression models say about the causal influences on female life expectancy?
- Why might televisions per person be a better predictor of female life expectancy than physicians per person?
- Does transforming the variables complicate the interpretation of the results?

A Multiple Regression Model

Can we predict female life expectancy better by looking at *both* televisions per person and physicians per person?

```
MTB > regress 'fle' 2 'sqtvpp' 'sqphyspp'

The regression equation is
fle = 54.8 + 25.1 sqtvpp + 171 sqphyspp

Predictor       Coef       Stdev    t-ratio        p
Constant      54.837       1.561      35.13    0.000
sqtvpp        25.067       4.642       5.40    0.000
sqphyspp      171.03       58.77       2.91    0.006

s = 4.151       R-sq = 78.9%     R-sq(adj) = 77.7%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2      2253.2      1126.6     65.39    0.000
Error        35       603.0        17.2
Total        37      2856.2

SOURCE       DF      SEQ SS
sqtvpp        1      2107.3
sqphyspp      1       145.9

Unusual Observations
Obs.  sqtvpp       fle       Fit Stdev.Fit  Residual   St.Resid
 17    0.105    73.000    66.370     2.265     6.630      1.91 X
X denotes an obs. whose X value gives it large influence.
```

Including Another Explanatory Variable

```
MTB > regress 'fle' 3 'sqtvpp' 'sqphyspp' 'mle'

The regression equation is
fle = - 0.55 + 3.71 sqtvpp + 66.6 sqphyspp + 1.03 mle

Predictor       Coef       Stdev    t-ratio        p
Constant      -0.548       3.827      -0.14    0.887
sqtvpp         3.711       2.272       1.63    0.112
sqphyspp       66.64       23.18       2.87    0.007
mle          1.02976     0.07032      14.64    0.000

s = 1.558       R-sq = 97.1%     R-sq(adj) = 96.9%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    3     2773.69      924.56    380.95    0.000
Error        34       82.52        2.43
Total        37     2856.21

SOURCE       DF      SEQ SS
sqtvpp        1     2107.33
sqphyspp      1      145.88
mle           1      520.48

Unusual Observations
Obs.  sqtvpp       fle       Fit Stdev.Fit  Residual   St.Resid
  2    0.056    53.000    56.116     0.492    -3.116     -2.11R
  7    0.258    61.000    64.880     0.470    -3.880     -2.61R
 17    0.105    73.000    72.301     0.942     0.699      0.56 X

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```