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 'file' 'zle' 'physpp' 'tvpp'

  NOW  file  male  physpp  tvpp
  1   74  67    0.0027027  0.265000
  2   53  54    0.0001622  0.003175
  3   68  62    0.0014820  0.265000
  4   80  73    0.0022272  0.882356
  5   72  68    0.0015552  0.126000
  6   74  68    0.0006447  0.178771
  7   61  60    0.0016234  0.666667
  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  0.027277
 12   63  59    0.0001346  0.041667
 13   65  64    0.0003342  0.043473
 14   82  75    0.0042918  0.263158
```

Regression of Female Life Expectancy on Physicians Per Person

```
MTB > regress 'file' 1 'physpp'

The regression equation is

file = 62.9 + 502 physpp

Predictor  Coef  Sdev  t-ratio  p
Constant   62.948  1.594  39.50  0.000
physpp     502.34  834.7  6.25  0.000

R-sq = 50.2%  R-sq(adj) = 48.0%

Analysis of Variance

SOURCE    DF  SS   MS    F    p
Regression  1 1492.4 1492.4 39.22 0.000
Error      36 1423.8  39.5    
Total      37 2916.2

Unusual Observations

Obs  physpp  file  male  physpp  tvpp
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.
```

Scatterplot of Female Life Expectancy Versus Physicians Per Person

```
MTB > plot 'file' 'physpp'

10
  9
  8
  7
  6
  5
  4
  3
  2
  1
  0

0.0000  0.0005  0.0010  0.0015  0.0020  0.0025  0.0030  0.0035  0.0040

physpp

Residual Plot

```
MTB > regress 'file' 1 'physpp'

SUBO: residuals ci0.

MTB > plot ci0 '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 'life' 'tvpp'

Regression of Female Life Expectancy on Televisions Per Person

MTB > regress 'life' 1 'tvpp';
SUBC: residuals c1.

The regression equation is
life = 63.2 + 37.0 tvpp

Predictor Coef Stdev t-ratio p
Constant 63.223 1.266 46.62 0.000
tvpp 36.966 5.156 7.17 0.000

s = 5.716 R-seq = 58.6% R-seq(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 life StDevFit 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.
I denotes an obs. whose tvalue gives large influence.

Residual Plot

MTB > regress 'life' 1 'tvpp';
SUBC: residuals c1.
... MTB > plot c11 'tvpp'

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 'life' 'sqtvpp'

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

12b.2
A Regression Model Using the Transformed Variable

```
MTB > regress 'fle' 1 'sqvpp';
SUMO residuals c12.

The regression equation is
fle = 56.9 - 36.0 sqvpp

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

s = 4.561  R-sq = 73.6%  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  20.9
Total    37 2866.2

Unusual Observations

Obs.  sqvpp  fle  Fit  Stdev.Fit  Residual  St.Resid
  17  0.106 73.000 60.619  1.215  12.38R  2.82R
  30  0.206 64.000 64.224  0.956 -10.224 -2.23R
  36  0.877 79.000 87.596  1.870   8.596 -2.07R

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 physicians per person variable in the same way:

```
MTB > let c0 = sqrt('sqvpp')
MTB > name c0 'sqhypp'
MTB > plot 'fle' 'sqhypp'

```

A Regression Model Using the Transformed Variable

```
MTB > regress 'fle' 1 'sqhypp';
SUMO residuals c13.

The regression equation is
fle = 56.4 - 403 sqhypp

Predictor  Coef  Stdev  t-ratio  p
Constant  56.436  2.046  27.58  0.000
sqhypp  403.465  53.43  7.66  0.000

s = 5.641  R-sq = 61.3%  R-sq(adj) = 60.2%

Analysis of Variance

SOURCE  DF  SS    MS  F    p
Regression  1 1700.9 1700.9 57.02  0.000
Error    36 1105.3 30.7
Total    37 2806.2

Unusual Observations

Obs.  sqhypp  fle  Fit  Stdev.Fit  Residual  St.Resid
  7  0.0403 61.000 72.691  0.962 -11.691 -2.14R

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

Residual Plot

```
MTB > plot c13 'sqhypp'

```
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?

MNB > regress 'fie' 2 'sqvpp' 'sgphypp'

The regression equation is

\[ \text{fie} = 54.8 + 26.1 \text{sqvpp} + 171 \text{sgphypp} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdve</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>54.87</td>
<td>1.561</td>
<td>35.13</td>
<td>0.000</td>
</tr>
<tr>
<td>sqvpp</td>
<td>26.07</td>
<td>4.642</td>
<td>5.60</td>
<td>0.000</td>
</tr>
<tr>
<td>sgphypp</td>
<td>171.03</td>
<td>56.77</td>
<td>2.91</td>
<td>0.006</td>
</tr>
</tbody>
</table>

\[ s = 4.151 \quad \text{R}^2 = 76.9\% \quad \text{R}^2(\text{adj}) = 77.7\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>1253.2</td>
<td>626.6</td>
<td>112.6</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>35</td>
<td>503.0</td>
<td>14.4</td>
<td>17.2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>1756.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

Usb. sqvpp fie Fit Stdev.Fit Residual St.Resid
17 0.100 73.000 66.500 2.260 6.600 1.91 X

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

Including Another Explanatory Variable

MNB > regress 'fie' 3 'sqvpp' 'sgphypp' 'nie'

The regression equation is

\[ \text{fie} = -0.55 + 3.71 \text{sqvpp} + 66.6 \text{sgphypp} + 1.03 \text{nie} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdve</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.549</td>
<td>3.287</td>
<td>-0.14</td>
<td>0.907</td>
</tr>
<tr>
<td>sqvpp</td>
<td>3.71</td>
<td>2.72</td>
<td>1.03</td>
<td>0.112</td>
</tr>
<tr>
<td>sgphypp</td>
<td>66.64</td>
<td>23.18</td>
<td>2.87</td>
<td>0.007</td>
</tr>
<tr>
<td>nie</td>
<td>1.0397</td>
<td>0.07032</td>
<td>14.64</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 1.558 \quad \text{R}^2 = 97.1\% \quad \text{R}^2(\text{adj}) = 96.9\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>2773.69</td>
<td>924.56</td>
<td>360.95</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>34</td>
<td>62.52</td>
<td>1.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>2836.21</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

Usb. sqvpp fie Fit Stdev.Fit Residual St.Resid
17 0.100 73.000 66.500 2.260 6.600 1.91 X

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