Impact of new housing projects on per-capita expenditures of New York State municipalities.

Dataset: nym.txt

VARIABLES:

1. State Code
2. County Code
3. Expenditure per person
4. Wealth per person
5. Population
6. Percent intergovernmental
7. Density
8. Mean Income per person
9. id # (for matching)
10. Growth rate

VARIABLE NAMES: DATASET project1.txt

ST CO EXPEN WEALTH POP PINTERG DENS INCOME ID GROWR

Dataset click here: NY Municipalities data

TOWNS OF INTEREST:

<table>
<thead>
<tr>
<th>Town</th>
<th>ST</th>
<th>CO</th>
<th>EXP</th>
<th>WEALTH</th>
<th>POP</th>
<th>PINTERG</th>
<th>DENS</th>
<th>INCOME</th>
<th>ID</th>
<th>GROWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Town of Warwick</td>
<td>36</td>
<td>33</td>
<td>237</td>
<td>78908</td>
<td>16225</td>
<td>24.7</td>
<td>170</td>
<td>19044</td>
<td>8730</td>
<td>30.3</td>
</tr>
<tr>
<td>Town of Monroe</td>
<td>36</td>
<td>33</td>
<td>159</td>
<td>55067</td>
<td>9338</td>
<td>8.8</td>
<td>599</td>
<td>16726</td>
<td>5420</td>
<td>30.0</td>
</tr>
<tr>
<td>Town of Tuxedo</td>
<td>36</td>
<td>33</td>
<td>926</td>
<td>155034</td>
<td>2328</td>
<td>6.1</td>
<td>52</td>
<td>30610</td>
<td>8400</td>
<td>2.5</td>
</tr>
</tbody>
</table>

QUESTION: PREDICT EXPENDITURE FOR YEARS 2010 and 2025

TOWN OF WARWICK

<table>
<thead>
<tr>
<th>Year</th>
<th>EXP</th>
<th>WEALTH</th>
<th>POP</th>
<th>PINTERG</th>
<th>DENS</th>
<th>INCOME</th>
<th>ID</th>
<th>GROWTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>237</td>
<td>78908</td>
<td>16225</td>
<td>24.7</td>
<td>170</td>
<td>19044</td>
<td>8730</td>
<td>30.3</td>
</tr>
<tr>
<td>2010</td>
<td>NA</td>
<td>85000</td>
<td>20424</td>
<td>24.7</td>
<td>214</td>
<td>19500</td>
<td>8730</td>
<td>20</td>
</tr>
<tr>
<td>2025</td>
<td>NA</td>
<td>89000</td>
<td>31018</td>
<td>26.0</td>
<td>325</td>
<td>20000</td>
<td>8730</td>
<td>10</td>
</tr>
</tbody>
</table>

TOWN OF MONROE

<table>
<thead>
<tr>
<th>Year</th>
<th>EXP</th>
<th>WEALTH</th>
<th>POP</th>
<th>PINTERG</th>
<th>DENS</th>
<th>INCOME</th>
<th>ID</th>
<th>GROWTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>159</td>
<td>55067</td>
<td>9338</td>
<td>8.8</td>
<td>599</td>
<td>16726</td>
<td>5420</td>
<td>30.0</td>
</tr>
<tr>
<td>2010</td>
<td>NA</td>
<td>58000</td>
<td>10835</td>
<td>8.8</td>
<td>695</td>
<td>17100</td>
<td>5420</td>
<td>20</td>
</tr>
<tr>
<td>2025</td>
<td>NA</td>
<td>60000</td>
<td>14950</td>
<td>10.1</td>
<td>959</td>
<td>18000</td>
<td>5420</td>
<td>10</td>
</tr>
</tbody>
</table>

TOWN OF TUXEDO

<table>
<thead>
<tr>
<th>Year</th>
<th>EXP</th>
<th>WEALTH</th>
<th>POP</th>
<th>PINTERG</th>
<th>DENS</th>
<th>INCOME</th>
<th>ID</th>
<th>GROWTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>926</td>
<td>155034</td>
<td>2328</td>
<td>6.1</td>
<td>52</td>
<td>30610</td>
<td>8400</td>
<td>2.5</td>
</tr>
<tr>
<td>2010</td>
<td>NA</td>
<td>116000</td>
<td>11148</td>
<td>6.1</td>
<td>249</td>
<td>28300</td>
<td>8400</td>
<td>20</td>
</tr>
<tr>
<td>2025</td>
<td>NA</td>
<td>115000</td>
<td>29369</td>
<td>7.0</td>
<td>656</td>
<td>25000</td>
<td>8400</td>
<td>10</td>
</tr>
</tbody>
</table>

Please read chapter 7 of the text book where you can find more detailed information about the report. This case study was motivated by a proposal of new developments in the above municipalities.
The report may be addressed as having a group of questions in mind and answering them. The central question is if there will be a large increase in per-capita expenditures as a result of the new developments? By how much?

**HINTS! Just to get you started.**

Of course you need an introduction with the basic facts. It should not be too long you do not need to describe the problem in detail because everybody should know it, but just the fact that are relevant to the analysis, the variables and some graphs of the data.

Here are some basic comments.

1. Prepare the data for analysis. It is convenient to add the prediction data to the dataset.

2. Transformations. Graph the data and see what kind of relationships you observe. Think of using log transformations or any other ones, since the variables in this dataset are very skewed, as they will be used for fitting a linear model.

```plaintext
1ex = log(expen);
1w = log(wealth);
lpop = log(pop);
ldens = log(dens);
lincome = log(income);
pint2 = pinterg**2;
pint3 = pinterg**3;
lpop2 = lpop**2;
ldens2 = ldens**2;
lpop3 = lpop**3;
ldens3 = ldens**3;

if growr < 0 then lgrowr = - log(1-growr);
if growr > 0 then lgrowr =  log(1+growr);
```

3. Subsetting. One way to deal with nonlinear relationships is to select a subset of interest where we want to calculate the predictions. For example, lets define a subset that might be of interest.

```plaintext
data a;
  set NYEXP;
  if pop < 40000;
  if lgrowr > -1 and lgrowr < 2;
  if income > 15000;
run;
```

To justify your subset you need to show that the points where you will perform the prediction are withing the limits of the subset and are as interior as possible. The above subset contains Tuxedo's points for predicting expenditures for 2010 and 2025.

4. Regression models: Use PROC REG to model the data. A simple equation is:

```plaintext
proc reg;
  model lEX = lW lPOP PINTERG ldens lINCOME lgROWR/ P R;
  output out=b cookd=cd p=p r =r;
run;
```
You can add more terms to the model formula by defining new variables in the data step.

\[
\text{model lEX = lW lPOP lPOP2 lPOP3 PINTERG lDENS lINCOME lGROWR/ P R;}
\]

Look at the Adjusted R^2 to decide if it improves much.

5. Plots are also important to visualize residuals and predicted values.

```
proc gplot data = b;
pplot cd*r p*r;
run;
```

will produce a plot of Cook’s D vs Residuals

6. Outliers. Outliers should be eliminated when they change the output model very much. Also if relations are too complicated it may be useful to consider only parts of the data, that is subsets generated by various variable constraints.

Example: If \text{LWEALTH} > \text{value}; or things like that.

Let’s say that high expenditures happen for different reasons and their relation to population and to wealth or income is complicated...

It is easier to just look at a part of the data where we have to do the prediction and hope that the complicated relationship has become simpler for that subset.

We will continue in the next lecture by building a shell SAS program for implementing the analysis.

7. Select your variables

```
proc reg data = a;
model lEX = lW lPOP lpop2 lpop3 PINTERG pint2 pint3 lDENS ldens2 ldens3 lINCOME lGROWR/ METHOD=adjrsq;
run;
```

8. Forecast Expenditures in 2010 and 2025

Suppose the model resulting from the previous step is:

\[
lEX = lW lPOP lpop2 lpop3 PINTERG pint2 pint3 lDENS ldens2 ldens3 lINCOME lGROWR/ \text{P R;}
\]

then use the following code to predict the results:

```
proc reg data = a;
model lEX = lW lPOP lpop2 lpop3 PINTERG pint2 pint3 lDENS / P R;
run;
/* USE THE DATA FROM THE OUTPUT TO CONSTRUCT THE PREDICTIONS */
data c;
input z zh se zr;
yh = exp( zh + se**se/2);
ym = exp(zh);
y = exp(z);
yr = y - yh;
cards;
6.8309 6.5597 0.0730 0.2712
. 6.3169 0.0562 .
. 6.5243 0.0566 .
;
run;
```