**Assignment 1, Part 3, SCMA 632 (20 points)**

**Due 9 AM, Saturday, November 26**

**E-mail Excel File to scma.stat@gmail.com with A1P3 in the subject line**

R.L. Andrews

The assignment will use the same groups as those for Quiz 1, part 2 and will be submitted electronically in an Excel spreadsheet by 9 AM, Saturday, November 26. You are to use the data found on my homepage http://www.people.vcu.edu/~randrews/ under the title [**U.S. Census Bureau Data (Excel 2007)**](http://www.people.vcu.edu/~randrews/fast_track/US_Census_Bureau_DataSet_2006.xlsx) .

You are to **use the final best model your group submitted for Assignment 1, part 2 and apply it to the data for the localities that are recognized subdivisions of the states** (counties and cities in Virginia and counties in other states) in the Excel file for U.S. Census Bureau Data (Excel 2007). This data set has the same variables as the one you used to build your best model in Assignment 1, part 2. Hence it has values for the response/dependent variable in column 49, People of all ages in poverty - percent 2004 and all other variables you used to build your best model. You can calculate values for any of the new predictors you created. You are to assess how well your model works by applying it to this data set by obtaining summary measures of how close the predicted value $\hat{y}\_{i}$ is to the actual value yi for the units in the data set.

One measure of closeness is R2 with

$R^{2}=\frac{SS\_{Regression}}{SS\_{Total}}=1- \frac{SS\_{Error}}{SS\_{Total}}$**,**

where **SSError =** $\sum\_{i=1}^{n}\left(y\_{i}-\hat{y}\_{i}\right)^{2}$, **SSTotal =** $\sum\_{i=1}^{n}\left(y\_{i}- \overbar{y}\_{i}\right)^{2}$ and

**SSRegression = SSTotal – SSError**.

Another measure that is often used with forecasting models is Mean Absolute Error, MAE, with

 $MAE=\frac{1}{n}∙\sum\_{i=1}^{n}\left|y\_{i}-\hat{y}\_{i}\right| $.

Yet another measure is the Mean Absolute Percent Error or Mean Absolute Proportion Error, MAPE, with

$MAPE= \frac{1}{n} ∙ \sum\_{i=1}^{n}\left|\frac{y\_{i}-\hat{y}\_{i}}{y\_{i}}\right|$.

The R2 measure has a range of 0 to 1 while the MAE and MAPE measures have no definite upper limit. Using MAPE can be problematic for small values or negative values of y. For a small y value, the error in the numerator may exceed the value of y, especially if y is less than 1. Of course, MAPE can’t be calculated if one or more values of y = 0.

Using the Excel file for the U.S. Census Bureau Data (Excel 2007) data you are to

1. Individually (10 points) each student will screen these data and recommend removing any rows you think should be removed for evaluation data set. By 9 AM Saturday November 12, each group member will e-mail to all of the group an Excel sheet that specifies the rows that are recommended for removal. A list of specific reasons for removing rows will be given and for each reason a list will be provided for the rows recommended for removal for that reason.

As a group (10 points) you will

1. Determine the actual data set you will use for the evaluation of your model. On one group tab, you will provide a list of specific reasons you chose for removing rows and for each reason a list will be provided for the rows recommended for removal for that reason. On another tab you will give your data set that only contains the columns for the predictor and response variables in your final model.
2. Write out your best model in a text box on another tab with values for each coefficient. Use the column number to define each variable in model [ For example: (#49) = 10.34 + .456 \* (#10)]. Provide definitions of any variables you created using the original set of variables and give each one a column number larger than 82.
3. Put the screened data on separate tab removing any variables (columns) that are not part of your best model placing y in the 1st column, $\hat{y}$ in the 2nd column, the residual = $y\_{i}-\hat{y}\_{i}$ in the third column, in the following columns put in computations you need to calculate R2, MAE and MAPE for your best model, then after these computation columns put in the values of the predictor variables in your best model.
4. On the final tab, give your calculated values for R2, MAE and MAPE for your best model. Using these measures and any other information you think is relevant, give a group statement describing how well your best works for predicting the value of the response variable, #49, for localities in the US.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group 1** | **Group 2** | **Group 3** | **Group 4** | **Group 5** | **Group 6** | **Group 7** | **Group 8** | **Group 9** |
| **Lnu** | **Jain** | **Boyce** | **Bates** | **Erskine** | **Ahmad** | **Gordon** | **Shah** | **Howse** |
| **Wiley** | **Mann** | **Ciccolo** | **Rapp** | **Janjua** | **Halun** | **Harris** | **Stells** | **Pandya** |
| **Wood** | **Mattauch** | **Sharma** | **Zhu** | **Mahoney** | **Madsen** | **Rancic** | **Tunney** |  |