U Wroclaw, Fall 2015
Applied Stats

DISCUSSION/LAB 8: MULTIPLE LINEAR REGRESSION

Influential observations and multicollinearity

You will work with data in MINITAB project: wastedata.MPJ. The data contains information on y= energy content of waste (in kcal.kg), and three composition variables for waste: Plastics=% plastics by weight, Paper=%paper by weight, Garbage=%garbage by weight, and Water=% water content per weight. We will look for the best MLR model for energy as a linear function of the explanatory variables: plastic, paper, garbage and water.

1. For all measures of leverage, outliers and influence (standardized and deleted-t residuals, h-leverages, Cooks' D and DFFITS), find the "critical" values of those measures that separate OK values from the high ones. Use the table format below.

Measure	Critical number
Standardized residuals	
deleted t-residuals	
leverages h_i	
Cook's distance	
DFITTS	

There is an influential observation in this data set. Which one is it? Explain why do you think this observation is influential.

2. Is there multicollinearity in the data set? If yes, explain why you think so and which variables seem to be problematic. If no, explain why you think so.
3. Remove the influential observation.
4. Run Forward selection and Backward elimination procedures on this data (with removed influential obs) with no forcing of variables in/out of the regression equation. Do you get the same "best" models? Why?
5. If necessary, reduce the data further by removing variable(s) that might be collinear with other variable(s). Be careful with removing too many variables at a time, I would suggest to start with one, see if that improved the model. If not, try another etc. Write what you did, be very concise. Write the final set of variables you decided to keep in the reduced set.
6. Find the best model for the reduced (if you reduced it) or original data set with influential obs removed (if you did not find multicollinearity present). Use any method you like. Report the method you used and the results.
7. Explain why the model you decided is best is good from (a) practical i.e. prediction/fit and from (b) statistical i.e. inference point of views.

Solution

1. Influential observation?

Regression Analysis: Energy versus Plastics, Paper, Garbage, Water

```
Coefficients

Term         Coef  SE Coef  T-Value  P-Value    VIF
Constant     2526      138    18.29    0.000
Plastics    27.85     2.94     9.47    0.000   1.11
Paper        4.87     9.35     0.52    0.607  26.46
Garbage     -0.64     8.93    -0.07    0.944  46.09
Water      -36.91     8.72    -4.23    0.000  22.20

Regression Equation

Energy = 2526 + 27.85 Plastics + 4.87 Paper - 0.64 Garbage - 36.91 Water
```

```
Fits and Diagnostics for Unusual Observations

Obs  Energy     Fit  Resid  Std Resid
  7  1466.0  1401.6   64.4       2.01  R
 30  1155.0  1158.7   -3.7      -0.92     X

R  Large residual
X  Unusual X
```

Leverage and influence stats for observation number 19:

```
           SRES    TRES          HI          COOK      DFIT
 Obs 7:  2.00850  2.14892     0.0811835   0.0712873   0.638763
Obs 30: -0.920519 -0.917605   0.985747    11.7207     -7.63105
```

In the table below fill the “critical numbers” for the leverage and influence statistics.

n=30, k=4

Measure	Critical number
Standardized residuals	
deleted t-residuals	
leverages h_i	
Cook’s distance	
DFITTS	

Do we have an influential observation? Why? Which one?

2. Is there multicollinearity in the data set? If yes, explain why you think so and which variables seem to be problematic. If no, explain why you think so.

VIFinfo:

Correlations: Plastics, Paper, Garbage, Water

```
           Plastics     Paper   Garbage
Paper        -0.163
              0.390

Garbage      -0.286     0.716
              0.126     0.000

Water        -0.207    -0.045     0.649
              0.272     0.812     0.000
```

3. Remove influential observation

4. Because there is multicollinearity in the data, stepwise procedures may give different results. Here are the results of forward selection and backward elimination.

Stepwise Regression (forward selection)

```
 F-to-Enter:      4.00    F-to-Remove:      0.00

 Response is Energy c on  4 predictors, with N =   29

    Step        1        2        3
Constant      3410     2653     2523

Water        -42.1    -37.7    -37.5
T-Value     -10.73   -17.91   -19.14

Plastics              26.9     27.9
T-Value               8.60     9.48

Paper                           4.2
T-Value                        2.25

S             69.2     36.0     33.5
R-Sq         80.99    95.05    95.89
```

Stepwise Regression (backward elimination)

```
 F-to-Enter:   1000.00    F-to-Remove:      4.00

 Response is Energy c on  4 predictors, with N =   29

    Step        1        2
Constant      2500     2511

Plastics      27.9     27.9
T-Value       9.46     9.62

Paper         68.8     41.5
T-Value       0.98    16.00

Garbage      -64.5    -37.3
T-Value      -0.92   -19.43

Water           27
T-Value       0.39

S             33.6     33.0
R-Sq         96.03    96.00
```

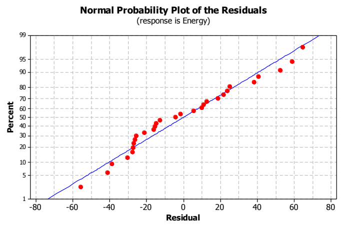
Conclusion? Same or different models? Is there collinearity?

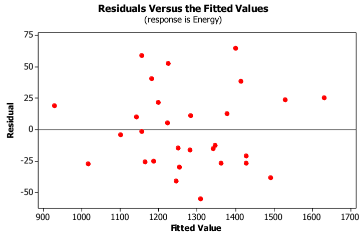
5. REMOVE GARBAGE!

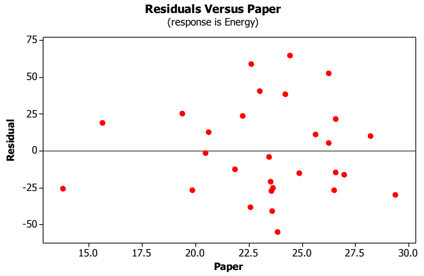
Regression Analysis without GARBAGE variable AND without observation 30

```
The regression equation is
Energy c = 2523 + 27.9 Plastics - 37.5 Water + 4.20 Paper

Predictor        Coef       StDev          T        P       VIF
Constant       2523.4       138.8      18.18    0.000
Plastics       27.906       2.943       9.48    0.000       1.1
Water         -37.496       1.959     -19.14    0.000       1.1
Paper           4.202       1.864       2.25    0.033       1.0

S = 33.47       R-Sq = 95.9%     R-Sq(adj) = 95.4%
```

How are the results now? Check the diagnostics.







6. Find the best model. I used Best Subsets Regression to find the best model because it compares all models for the given data.

```
Response is Energy c

                                                P
                                                l
                                                a
                                                s P W
                                                t a a
                                                i p t
              Adj.                              c e e
Vars   R-Sq   R-Sq    C-p        s              s r r

   1   81.0   80.3   90.6   69.247                  X
   1   34.0   31.6  376.3   129.03              X
   2   95.1   94.7    7.1   36.003              X   X
   2   81.1   79.6   91.9   70.363                X X
   3   95.9   95.4    4.0   33.469              X X X  BEST MODEL!
```

7. The model Energy c = 2523 + 27.9 Plastics - 37.5 Water + 4.20 Paper **is good from the stat point of view because:**