This data set will be posted in the StatCrunch group.

*Air Pollution in U.S. Cities*

For 41 cities in the United States, the following 7 variables were recorded:

Y: Sulphur dioxide content of air in micrograms per cubic meter (SO2)

X1: Average annual temperature in degrees F (Temp)

X2: Number of manufacturing enterprises employing 20 or more workers (Manuf)

X3: Population size in thousands (Pop)

X4: Average annual wind speed in miles per hour (Wind)

X5: Average annual precipitation in inches (Precip)

X6: Average number of days with precipitation per year (Days)

This data has been collected to investigate the determinants of air pollution.

1. Look at the correlation matrix of all the explanatory variables. Which two explanatory variables are so highly correlated that they may give cause for concern?
2. Now conduct a regression analysis using **all** theexplanatory variables. If using StatCrunch store the standardized residuals and the predicted values. Interpret in detail your printout.
3. Examine the residual plots (or construct a residual plot of the standardized residuals and the predicted values if using StatCrunch). What do the plots indicate?

(d) Look at the standardized residuals for each city. You will notice from these and your plots that two cities stand out in the model fit as being outliers. Locate these cities and comment.

(e) Using the information in the ANOVA table (part b) and the correlation matrix (part a) comment on if any variables should be eliminated from the regression model.

*Air Pollution in U.S. Cities*

For 41 cities in the United States, the following 7 variables were recorded:

Y: Sulphur dioxide content of air in micrograms per cubic meter (SO2)

X1: Average annual temperature in degrees F (Temp)

X2: Number of manufacturing enterprises employing 20 or more workers (Manuf)

X3: Population size in thousands (Pop)

X4: Average annual wind speed in miles per hour (Wind)

X5: Average annual precipitation in inches (Precip)

X6: Average number of days with precipitation per year (Days)

This data has been collected to investigate the determinants of air pollution.

1. Look at the correlation matrix of all the variables. Which two explanatory variables are so highly correlated that they may give cause for concern?

SO2 Temp Manuf Pop Wind Precip

Temp -0.434

Manuf 0.645 -0.190

Pop 0.494 -0.063 0.955

Wind 0.095 -0.350 0.238 0.213

Precip 0.054 0.386 -0.032 -0.026 -0.013

Days 0.370 -0.430 0.132 0.042 0.164 0.496

The correlation between population and manufacturing is 0.955. This shows a very strong positive relationship between this two explanatory variables. Due to the strong correlation between these two variables it may be unnecessary to include both variables in the final model.

1. Now conduct a regression analysis using **all** theexplanatory variables. Store the standardized residuals. Interpret in detail your printout.

The regression equation is

SO2 = 112 - 1.27 Temp + 0.0649 Manuf - 0.0393 Pop - 3.18 Wind + 0.512 Precip

- 0.052 Days

Predictor Coef SE Coef T P

Constant 111.73 47.32 2.36 0.024

Temp -1.2679 0.6212 -2.04 0.049

Manuf 0.06492 0.01575 4.12 0.000

Pop -0.03928 0.01513 -2.60 0.014

Wind -3.181 1.815 -1.75 0.089

Precip 0.5124 0.3628 1.41 0.167

Days -0.0521 0.1620 -0.32 0.750

S = 14.6360 R-Sq = 67.0% R-Sq(adj) = 61.1%

Analysis of Variance

Source DF SS MS F P

Regression 6 14754.6 2459.1 11.48 0.000

Residual Error 34 7283.3 214.2

Total 40 22037.9

Source DF Seq SS

Temp 1 4143.3

Manuf 1 7230.8

Pop 1 2125.2

Wind 1 447.9

Precip 1 785.4

Days 1 22.1

Unusual Observations

Obs Temp SO2 Fit SE Fit Residual St Resid

11 50.6 110.00 110.54 12.22 -0.54 -0.07 X

30 50.4 61.00 30.93 4.61 30.07 2.16R

31 50.0 94.00 45.24 5.71 48.76 3.62R

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

1. R-Sq = 67.0%. This tells us that 67% of the variation in S02 is explained by the regression on the chosen explanatory while the r^square adjusted value which takes into account both the sample size and the number of parameters in the model is 61.1%
2. The F-test in the “Analysis of Variance” section which gives a F statistic of 11.48 and a p-value near 0 tells us that at least one of the regression coefficients will be significantly different from zero.
3. Three of the explanatory variables, temperature, manufacturing and population are shown to be significant, at  with the t-test while the other three variables are not significant. However, we know that manufacturing and population are highly correlation and this may have to be looked at further.
4. Notice that the signs of the coefficients are positive for manufacturing and precipitation but negative for temperature, population, wind and days. An interpretation for the manufacturing variable would be, “Holding all other variables constant, a one unit increase in manufacturing results in an average increase of 4.12 in S02 levels.”
5. Examine the residual plots. What do the plots indicate?



1. The normal probability plot shows that the residuals are Normally distributed which is one of the assumptions needed to be satisfied in order to conduct this analysis. Possible one outlier.
2. The assumption of constant variance is checked by the residual vs fitted value plot. This plot should show random constant scatter if this assumption has been satisfied. However, this plot shows cause for concern as variance does not appear constant.

(d) Look at the standardized residuals for each city. You will notice from these and your plots that two cities stand out in the model fit as being outliers. Locate these cities and comment.

Chicago has the lowest standardized residual value of -0.0673 but it is a highly influential point. Chicago has a very high S02 value and very high values of x2 and x3 which pull the regression fit towards this point.

Providence has the highest standardized residual value of 3.61842. This city has a smaller than average population and manufacturing but a relatively high S02 level. The actual value for the S02 level in Providence is 94 while the fitted value is 45.2424. Researchers would need to determine why this city has a lot more pollution than predicted.

Pittsburgh has the second highest standardized residual value of 2.16471. This city has the same problem as Providence – more polluted than it should be.