Problem 1. Refer to the pollution data output. A linear model was fit relating mortality to various pollution concentrations (HC, SO2, NOX) and other variables that might affect mortality.
The Goverment claims that pollution levels have no linear effect on mortatility (MORT) after adjusting for the other nonpollution variables.

a) What is the null hypothesis?.
Full Model: $E(Y) = \beta_0 + \beta_1 SO + \beta_2 NOX + \beta_3 HC + +\beta_4 NONW + \beta_5 HUMID + \beta_6 OVR65 + \beta_7 DENS$
Ho: $\beta_1 = \beta_2 = \beta_3 = 0$ Ha: at least one of β_1, β_2, or β_3 is non-zero.

b) Do you agree with the Goverment based on the data and analysis here? Explain.
We have a nested model, so use an F test to test Ho.

$$F = \frac{(102127.32 - 75713.51)/3}{75713.51/52}$$

so F = 6.047 with 3 and 52 df.
Using either the table entry with 40 df or 60 df in the denominator we would find that $F_{.01} < 6.047 < F_{.001}$, so that our p-value is between 0.01 and 0.001. Thus we would reject the null hypothesis for $\alpha > .01$, but not for $\alpha < .001$. This would imply that there is a linear relationship between Mortality and **at least one** of the pollution variables, HC, NOX or SO2, after adjusting for the other variables (i.e. including them in the regression)

c) Does this agree with the results of the t-tests? Explain.
The F-test indicates that the coefficient of at least one of the pollution variables is significantly different from 0, but not which ones. The p-values for the t-tests regarding Ho: $\beta_i = 0$ for i = 1,2, or 3 are indicating that any one of them could be dropped given that the others are kept in the model so that there is some simplification that could be obtained.

b	p-value
SO	0.1911
NOX	0.2107
HC	0.1133

Problem 2. Is there any indication of multicollinearity problems? Does this have any effect on the analysis in problem 1? Explain.

There are several indications of multicollinearity. One is that the simple correlation between NOX and HC is high (.9838). The second is that the correlation between the parameter estimates for HC and NOX is also high (-.9933). While the simple correlation does indicate that NOX and HC are highly correlated with each other, these results do not rule out that there may be high correlations among the other variables, and we should examine the tolerances. The low tolerances less than .1 are for NOX and HC:

```
Variable  DF     Tolerance

   NOX        1     0.01099830
   HC         1     0.01204649
```

indicating that these two variables can be explained very accurately by linear combinations of the other variables, and thus there is strong evidence of multicollinearity.

Other things to look at are the results of the F-test for hypothesis in problem 1 being highly significant while none of the t-tests for the same coefficients had significant values.

Please note: a significant F-statistic for the overall test of the model (Ho: all β's $= 0$) and SOME insignificant t-tests for individual β's does not necessarily imply multicollinearity. There simply could be nonimportant variables. Rejecting Ho with the overall F-test just means that at least ONE coefficient is significant, and does not rule out that SOME are insignificant, which would show up in the individual t-tests. If, however, the overall F test is significant AND ALL individual t-tests are insignificant, then multicollinearity is a likely culprit.

Judging standard errors can be tricky, since if one throws out a group of important variables (the second model), then the residual sum of squares will be much larger and the MSE will be an overestimate of σ. Since this enters into the standard error of the b_k's, one would expect the standard errors in the second model to be larger. While they are larger, it does not seem like a significant change. Note, none of the non-pollution variable coefficients change sign. Howeve, in the full model, the coefficient for HC is negative (-.7925) which is not what we would typically expect. Having a coefficient with a sign that is the opposite of what we expect could be an indication of multicollinearity. Fitting the model without NOX could clear this up.

Problem 3. Use the diagnostics output to answer the following questions.

a) Are there any data points that are "significant" outliers? Test at the $\alpha = .1$ level (clearly show how you come to your conclusions).
No. Using a Bonferroni procedure, we would compare the p-value from the studentized residuals (t_i) to $\alpha/n = .1/60 = .0016$. Since none of the p-values are less than this cutoff, we would declare that there are no significant outliers at the $\alpha = .1$ level.

b) Are there any "influential" observations? Explain.
Gobs, depending on your criterion.
For Cook's distance, $D_{43} = 1.6 > 1$ (the absolute cutoff). There are several other observations with Cook's distances that are larger than the size adjusted cutoff of $4/n$. Note, we would first delete case 43, and then see if in the new model there are still any influential cases. Sometimes deleting the most influential may make the other cases less influential. Having a large number of influential cases based on the size adjusted cutoff, could actually be a result of non-constant variance - where a number of points show up with large residuals. We should have definitely looked at a scatter plot of residuals versus fitted values to check for this before investigating measures of influence.
If we want to look at the influence of individual cases on the coefficients, we would use the DFBETAS. We find that none are larger than 2 in absolute value, so based on an external criterion there would not be any influential cases. Based on the size adjusted cutoff of $2/\sqrt(n)$ there are lots of influential cases. In particular, observation 29, 48, 48, 56, and 59 seem to have a high influence on several coefficients.

Problem 4. The output includes the Durbin-Watson statistic and the first order autocorrelation coefficient. Is it appropriate to test for first order autocorrelation? If you do not have enough information given in the problem to answer, what would you need to know from the researcher to decide if the test is appropriate?

We really do not have enough information to tell if the Durbin-Watson test is appropriate. We would need to investigate how the data were collected. For example, if observations were based on yearly measurments at the same location, the test would be o.k (the data would need to be sorted in time order). However, if the observations were made for one year (or several years), but at many locations, the test would not be appropriate. Note, if there is actually some seasonal effect or an effect that carries over to multiple years (higher order auto-correlations) then the Durbin-Watson test will not be able to pick this up. Looking at correlograms and plots of the residuals over time would be useful then.