For problems **1.** and **2.**, Perform regression diagnostics on the model to answer following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

(a) Check the constant variance assumption for the errors.

(b) Check the normality assumption.

(c) Check for large leverage points.

(d) Check for outliers.

(e) Check for influential points.

(f) Check the structure of the relationship between the predictors and the response.

1. Faraway 4.4. (pg. 75) For the `swiss` data, fit a model with `Fertility` as the response and the other variables as predictors. Also construct a half-normal plot of the leverages and interpret the plot.

2. Faraway 4.5. (pg. 75) For the `divusa` data, fit a model with `divorce` as the response and the other variables as predictors.

3. Consider the table below of regression diagnostics on the wood data from problem 5 of Homework #3.

Obs	h_i	r_i	t_i	Cooks D	Obs.	h_i	r_i	ti	Cooks D
1	.085	-.25	-.25	.001	29	.069	.27	.26	.001
2	.055	1.34	1.35	.021	30	.029	.89	.89	.005
3	.021	.57	.57	.001	31	.204	.30	.30	.005
4	.031	.35	.35	.001	32	.057	.38	.37	.002
5	.032	2.19	2.28	.032	33	.057	.05	.05	.000
6	.131	.20	.19	.001	34	.085	-2.43	-2.56	.109
7	.027	1.75	1.79	.017	35	.186	-2.17	-2.26	.215
8	.026	1.23	1.24	.008	36	.184	1.01	1.01	.046
9	.191	.52	.52	.013	37	.114	.85	.85	.019
10	.082	.47	.46	.004	38	.022	.19	.19	.000
11	.098	-3.39	-3.82	.250	39	.022	-.45	-.45	.001
12	.066	.32	.32	.001	40	.053	-1.15	-1.15	.015
13	.070	-.09	-.09	.000	41	.053	.78	.78	.007
14	.059	.08	.08	.000	42	.136	-.77	-.76	.018
15	.058	-.91	-.91	.010	43	.072	-.78	-.77	.009
16	.085	-.09	-.09	.000	44	.072	-.27	-.26	.001
17	.113	1.28	1.29	.042	45	.072	-.40	-.40	.002
18	.077	-1.05	-1.05	.018	46	.063	-.62	-.62	.005
19	.167	.38	.38	.006	47	.025	.46	.46	.001
20	.042	.24	.23	.000	48	.021	.18	.18	.000
21	.314	-.19	-.19	.003	49	.050	-.44	-.44	.002
22	.099	.56	.55	.007	50	.161	-.66	-.66	.017
23	.093	.47	.46	.004	51	.042	-.44	-.43	.002
24	.039	-.60	-.60	.003	52	.123	-.26	-.26	.002
25	.098	-1.07	-1.07	.025	53	.460	1.81	1.86	.558
26	.033	.14	.13	.000	54	.055	.50	.50	.003
27	.042	1.19	1.19	.012	55	.093	-1.03	-1.03	.022
28	.185	-1.41	-1.42	.090					

(a) Are there any unusual values in the predictor variables? In other words, do any of the values have unusually large influence on the model? Identify all such cases and explain your reasoning.

(b) Identify any model outliers in these data, clearly explaining your reasoning.

(c) Which observations are most influential in terms of the predictive ability of the model? Explain your reasoning clearly

4. There have been numerous efforts to collect data which support or refute the theory of global warming. This problem considers a data set containing the temperature in degrees Celsius averaged for the northern hemisphere over a full year, from 1881 through 2005. The data can be found in the data file `warming2005.txt` on the course webpate. [Data from K.M. Lugina et al., 2006. Monthly surface In Trends Online: A compendium of Data on Global Change. Carbon Diode Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy. Oak Ridge, Tennessee, U.S.A.doi:10.3334/CDIAC/cli.003.]

(a) Make a scatterplot of temperature vs. time and fit a single linear regression model with temperature as the response variable. Test for an *increase* in temperature over time. Explain your conclusions.

(b) Inherent with the test in part (a) are a number of assumptions. One of these is that of normality of the erros. Make a normal quantile plot of the residuals. Just looking at the plot, do these data appear to violate the normality assumption? Test for normality of the residuals using a normal scores correlation (Shapiro-Francia) test.

(c) Using the unstandardized residuals found in part (b), make a plot of the residuals vs. time. Just looking at this plot, does there appear to be any serial correlation? Explain. Test for the presence of serial correlation using a Durbin-Watson test. [In `R`, you can request the Durbin-Watson test using the `dwtest` function in the package `lmtest` as illustrated in class.

(d) Test for serial correlation using a runs test. What is the p-value for this test? [In `R`, you can perform a runs test using the `runs.test` function in the `tseries` package as shown in the `R`-code from the Diagnostics handout on the course webpage.]

5. Data were collected from 60 major US cities to study the relationship between mortality rates and air pollution. Letting the response variable be total age-adjusted mortality rate per 100,000 people for a city, the following predictor variables were considered: mean annual precipitation in inches (x_1), mean January temperature in degrees F (x_2), mean July temperature in degrees F (x_3), population per household (x_4), median school years completed by those over the age of 25 (x_5), percent of housing units that are sound and with all facilities (x_6), population per square mile in urbanized areas (x_7), percent non-white population in urbanized areas (x_8), relative pollution potential of sulphur dioxide (x_9), and annual average of percent relative humidity at 1pm (x_{10}). These data can be found in the file `pollution.txt` on the course webpage. Using only the variables y, x_2, x_4, x_5 and x_7 (to keep things manageable), carry out the following.

(a) Make pairwise scatterplots and compute the matrix of sample correlations between the 5 variables. In a paragraph, briefly describe any apparent relationships between the variables and the nature of those relationships, based on the plots and correlations.

(b) Construct a table similar to the one given in the Chapter 8 class notes containing the R^2, $R^2_{adj}, C_p, \sqrt{\text{MSE}}$, AIC, and PRESS statistics for *all* possible first-order models using only ????? [Note: there are 4, 6, 4, & 1 models with 1, 2, 3, & 4 predictors in them respectively, for a total of 15 models.] Interpret this table. Do the selection criteria for finding the "best model" agree? Which model seems best (based on the four explanatory variables used)?

(c) Perform some residual diagnostics, such as examining residual plots and normal quantile plots for the "best model" chosen from part (b). Draw some conclusions about assumptions on the errors.

(d) In addition to performing residual diagnostics, it is important to conduct various regression diagnostics to look for outliers or influential data values. Probably the most popular statistics for identifying outlying or influential values are leverage, Cook's D, DFFits, or DFBetas, discussed in class. Each of these four diagnostic tools can be requested in `R` as explained in class or in Section 4.2.3 of the text. Examine the values for the residuals, the leverages, and the three influence statistics resulting from the final model above, and identify any points which would be considered outliers or influential by more than one of the statistics.

6. Faraway 8.3. Use the `divusa` dataset with `divorce` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

(a) Backward Elimination

(b) AIC

(c) Adjusted R^2

(d) Mallows C_p