Part A. Data were collected to predict the **time** it takes to service a beverage vending machine from the number of **cases** of beverage that must be stocked and the **distance** the service man has to walk to the machine. Below is a matrix of scatterplots for 25 vending machines.

1. (16 points) Circle *all* the letters below corresponding to true statements.

a. \( X_1 \) and \( X_2 \) are positively correlated in this data set.

b. There is an observation with high leverage.

c. The possible curvature with respect to \( X_2 \) can be examined with a partial residual plot

d. The possible curvature with respect to \( X_2 \) can be examined by testing an \( X_2^2 \) term.

e. Possible influence problems can be investigated by inspecting \( R^2 \).

f. Possible influence problems can be investigated by inspecting Cook’s Distances

g. Possible influence problems can be investigated by inspecting fits with and without the suspected influential observation.

h. Possible influence problems can be investigated by inspecting the \( Cp \) statistic.
2. (6 points) The following is a partial residual plot for $X_1$ from the fit to the multiple regression of $Y$ on $X_1$ and $X_2$. Circle the letter below corresponding to the best assessment from this plot.

![Partial Residual Plot](image)

a. The observation with the largest $X_1$ value has a large leverage and a large Cook’s Distance.

b. The observation with the largest $X_1$ value has a small leverage and a large Cook’s Distance.

c. The observation with the largest $X_1$ value has a large leverage and a small Cook’s Distance.

The following are fits to 2 models for the linear regression of $Y$ on $X_1$ and $X_2$.

**Model I.**

|        | Value   | Std. Error | t value | Pr(>|t|) |
|--------|---------|------------|---------|---------|
| (Intercept) | 1.7357  | 0.1184     | 14.6535 | 0.0000  |
| distance ($X_1$) | 0.0004  | 0.0003     | 1.4183  | 0.1708  |
| Log cases ($X_2$) | 0.5072  | 0.0663     | 7.6528  | 0.0000  |

Residual standard error: 0.1212 on 21 degrees of freedom;  Multiple R-Squared: 0.9535

**Model II.**

|        | Value   | Std. Error | t value | Pr(>|t|) |
|--------|---------|------------|---------|---------|
| (Intercept) | 1.6886  | 0.0828     | 20.4040 | 0.0000  |
| distance ($X_1$) | 0.0006  | 0.0001     | 5.1751  | 0.0000  |
| Log cases ($X_2$) | 0.5264  | 0.0559     | 9.4106  | 0.0000  |

Residual standard error: 0.1193 on 22 degrees of freedom;  Multiple R-Squared: 0.9528
3. (6 points) What is a 2-sided $p$-value for the significance of the interactive effect of $distance$ and $\text{Log}(cases)$?

4. (6 points) What proportion of the variation in $Y$ is explained by the regression on $X_1$ and $X_2$ (Model II)?

5. (8 points) Using Model II, describe the estimated change in the distribution of servicing time associated with each extra 1 foot of distance to the machine (for a fixed number of $cases$).

6. (8 points) Using Model II, describe the estimated change in the distribution of servicing time associated with each doubling of the number of cases of beverage (for a fixed $distance$).

**Part B.** The concentration of polychlorinated biphenyl (PCB, in parts per million) was measured in 28 fish of known ages from Cayuga Lake in New York State. Shown to the right is a scatterplot of the log of the PCB concentration versus the log of the age of the fish.

The sum of squared residuals from the fit to the simple linear regression is 694.4 on 26 degrees of freedom. The sum of squared residuals from the fit of the one-way analysis of variance model, with 11 groups, is 322.2 on 17 degrees of freedom.
7. (5 points) What is the estimate of $\sigma^2$ from the simple linear regression fit?

8. (8 points) What is the F-statistic for the lack-of-fit F-test?

**Part C.** The weekly gas usage (in cubic feet) and the average outside temperature (in degrees Celsius) were recorded for 26 weeks *before* and 18 weeks *after* insulation had been installed in a gas-heated house in England. The house thermostat was set at 20°C throughout.

Several fits are shown below for the regression of gas usage ($\text{gas}$) on outside temperature ($\text{temp}$) and an indicator variable ($\text{Ind.before}$) taking the value 1 for those weeks before the insulation had been installed and 0 for those weeks after it had been installed. **Standard errors are shown in parentheses below the estimates.**

Model I: $3889 + 861\times\text{Ind.before}$; $R^2 = .17$; 
(223) (290)

Model II: $4922 + 1795\times\text{Ind.before} - 368\times\text{temp}$; $R^2 = .92$ 
(88) (104) (19)

Model III: $4951 + 2263\times\text{Ind.before} - 249\times\text{temp} - 144\times\text{Ind.before}\times\text{temp}$; $R^2 = .94$ 
(130) (173) (40) (45)
9. (8 points) What is the t-statistic for a two-sample t-test comparing mean gas usage for those weeks before the insulation was installed to the mean gas usage for those weeks after the insulation was installed?

10. (6 points) How much more was the mean gas usage before insulation than after insulation, after accounting for the effect of temperature? (Use Model II to answer this.)

11. (6 points) How much more was the mean gas usage before insulation than after insulation if the outside temperature was 4 degrees? (Use Model III to answer this.)

12. (5 points) True or False? Since $R^2$ is increased by only .02 when the interaction term is included in the model, the interaction term is not statistically significant.
   a. True
   b. False

13. (12 points) The method of estimation used in St 512 to estimate regression coefficients is called (circle one):
   a. Least Squares
   b. Psychic Hotline
   c. Guess and Hope
   d. Most Triangles