1. A study of 53 Florida lakes was conducted to determine factors that affect the amount of mercury in largemouth bass. The following variables were considered:

\[ x_1 = \log(\text{alkalinity}) \]
\[ x_2 = \log(\text{pH}) \]
\[ x_3 = \log(\text{calcium}) \]
\[ x_4 = \log(\text{chlorophyll}) \]
\[ y = \log(\text{average mercury content in a sample of fish}) \]

The four independent variables were measurements taken from the water of a given lake. Use the accompanying SPSS output to answer the following questions (a)-(f).

(a) (12) Does it appear that any of the four independent variables is useful in explaining variability in the logarithm of average mercury content? Answer using the appropriate test with \( \alpha = 0.05 \).

The F test in the ANOVA table for the full model is highly significant (\( F = 36.862 \) and \( p = 0.000 \)). Therefore we may conclude that at least one of the four variables is useful.

(b) (15) Does the model containing \( x_1, x_2, x_3, \) and \( x_4 \) improve significantly upon one containing only \( x_1 \) and \( x_4 \)? (Use \( \alpha = 0.05 \)).

\[
F = \frac{(105.496 - 101.963) / 2}{0.715} = 2.47
\]

\[
F_{2,48,.05} = 3.19
\]

Since \( 2.47 < 3.19 \), we cannot conclude that the full model improves significantly on the one that contains only \( x_1 \) and \( x_4 \).
(c) (7) In the full model (i.e., the one with all four independent variables), the value of $R^2$ is $\frac{105.496}{139.839} = 0.75$.

(d) (7) An estimate of the error variance for the full model is $0.715$.

(e) (14) The standard error of $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ when $x_1 = 4$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 2.25$ is 0.244. Let $\mu$ be the average value of $y$ for all lakes with $x_1 = 4$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 2.25$. Find a 95% confidence interval for $\mu$.

$$\hat{\mu}_x = 2.98 + 4(-1.1) + 2(-0.16) + 3(0.321) + 2.25(-0.323)$$

$$= -1.216$$

$$-1.216 \pm (2.01)(0.244) = -1.216 \pm 0.49044, \text{ or}$$

$$(-1.706, -0.726)$$

(f) (8) Give a practical interpretation of $\hat{\beta}_1$, the estimated regression coefficient for the variable log(alkalinity) in the full model.

Holding the other three variables constant, we estimate that an increase of 1 in log(alkalinity) results, on average, in a decrease of 1.1 in log(aver. mercury).
2. A study was done to investigate the relationship between the "percentage of a state's population living in metropolitan areas" and "a state's per capita public expenditures." We'll call these variables MET and EXPEND, respectively. MET is the independent variable and EXPEND is the dependent variable. Scatterplots of the data along with fitted models are shown on the next page.

(a) (15) Polynomial models of orders one through four were fitted to these data, and the following information was obtained from "General Linear Model" in SPSS:

<table>
<thead>
<tr>
<th>Order of Model</th>
<th>Type I Sum of Squares</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>332.5</td>
<td>0.12</td>
<td>0.729</td>
</tr>
<tr>
<td>2</td>
<td>37504.0</td>
<td>13.74</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>7244.5</td>
<td>2.65</td>
<td>0.111</td>
</tr>
<tr>
<td>4</td>
<td>13.1</td>
<td>0.005</td>
<td>0.945</td>
</tr>
</tbody>
</table>

On the basis of this information, which order model would you choose? Explain your reasoning. Using the testing method discussed in class, the most reasonable model seems to be order 2. The second degree term is highly sig. (p = 0.001) but neither the third or fourth is sig. for any p ≤ 0.111.

(b) (14) What are the values of $R^2$ and adjusted-$R^2$ for the third order model?

\[
SSR_3 = 332.5 + 37504 + 7244.5 = 45,081
\]

\[
R^2 = \frac{45,081}{45,094.1} = 0.9997 \quad \text{adj.-}R^2 = 1 - (0.003)^{\frac{47}{44}} = 0.9997
\]

(c) (8) Consider the graph of the model you chose in (a). (This is found on the next page.) Give an interpretation, in the language of this problem, of the fitted curve.

Expenditures are very similar on aver., in states with the highest and lowest percentages of their populations living in metropolitan areas. States with intermediate percentages have lower aver. expend.

* Actually, SST for the data plotted on p. 4 is 162,470.979, implying that $R^2$ in (b) is 0.277.