Stat 231 Exam 3
Fall 2011

I have neither given nor received unauthorized assistance on this exam.

________________________________________________________
Name Signed                Date

________________________________________________________
Name Printed
This exam concerns the analysis of some data collected by the Norwegian Public Roads Administration (part of the "NO2" data set available from the "statlib" data sets archive, submitted to the archive by Magne Aldrin) concerning how air pollution at a road is related to traffic volume, and time-related and meteorological variables. The response variable is

\[ y = \text{hourly values of the logarithm of the concentration of NO}_2 \text{ particles} \]

measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The predictor variables in the original data set were

- \( x_1 = \text{logarithm of the number of cars per hour} \)
- \( x_2 = \text{temperature 2m above ground (°C)} \)
- \( x_3 = \text{wind speed (m/s)} \)
- \( x_4 = \text{temperature difference between 25m and 2m above ground (°C)} \)
- \( x_5 = \text{wind direction (in degrees between 0 and 360)} \)
- \( x_6 = \text{hour of day (from 0 to 24)} \)
- \( x_7 = \text{day number from October 1, 2001} \)

Vardeman also added predictors to the data set by defining the variables

\[ \sin \left( \frac{2\pi x_3}{360} \right), \cos \left( \frac{2\pi x_5}{360} \right), \sin \left( \frac{2\pi x_6}{24} \right), \cos \left( \frac{2\pi x_6}{24} \right), \sin \left( \frac{2\pi x_7}{365} \right), \text{and } \cos \left( \frac{2\pi x_7}{365} \right) \]

Here are the sample correlations between the variables in this problem.

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>sin(x3)</th>
<th>cos(x5)</th>
<th>sin(x6)</th>
<th>cos(x6)</th>
<th>sin(x7)</th>
<th>cos(x7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.0000</td>
<td>0.5354</td>
<td>-0.1071</td>
<td>-0.3067</td>
<td>0.4067</td>
<td>0.3280</td>
<td>0.3956</td>
<td>0.2592</td>
<td>-0.2259</td>
<td>-0.1888</td>
<td>-0.4270</td>
<td>-0.2956</td>
<td>0.2038</td>
<td>-0.0179</td>
</tr>
<tr>
<td>x1</td>
<td>0.5354</td>
<td>1.0000</td>
<td>0.1115</td>
<td>0.0059</td>
<td>0.0862</td>
<td>0.1741</td>
<td>0.7165</td>
<td>-0.1635</td>
<td>-0.2301</td>
<td>-0.0986</td>
<td>-0.8198</td>
<td>-0.5133</td>
<td>0.1595</td>
<td>-0.0022</td>
</tr>
<tr>
<td>x2</td>
<td>-0.1071</td>
<td>0.1115</td>
<td>1.0000</td>
<td>-0.0215</td>
<td>0.0566</td>
<td>0.1791</td>
<td>-0.0673</td>
<td>0.0428</td>
<td>-0.2235</td>
<td>-0.1924</td>
<td>-0.0064</td>
<td>0.0675</td>
<td>-0.7454</td>
<td>-0.2822</td>
</tr>
<tr>
<td>x3</td>
<td>-0.3067</td>
<td>-0.3067</td>
<td>-0.0215</td>
<td>1.0000</td>
<td>-0.2846</td>
<td>-0.4651</td>
<td>0.0212</td>
<td>0.1034</td>
<td>0.3100</td>
<td>0.2827</td>
<td>-0.0117</td>
<td>-0.1093</td>
<td>0.0954</td>
<td>0.1014</td>
</tr>
<tr>
<td>x4</td>
<td>0.4067</td>
<td>0.4067</td>
<td>0.0566</td>
<td>-0.2846</td>
<td>1.0000</td>
<td>0.5130</td>
<td>0.2715</td>
<td>0.2419</td>
<td>-0.4518</td>
<td>-0.4046</td>
<td>0.2160</td>
<td>0.3000</td>
<td>-0.0551</td>
<td>-0.1389</td>
</tr>
<tr>
<td>x5</td>
<td>0.3280</td>
<td>0.1741</td>
<td>0.1791</td>
<td>-0.4651</td>
<td>0.5130</td>
<td>1.0000</td>
<td>0.2781</td>
<td>0.1451</td>
<td>-0.9400</td>
<td>-0.5784</td>
<td>-0.1527</td>
<td>-0.1897</td>
<td>-0.0633</td>
<td>-0.0974</td>
</tr>
<tr>
<td>x6</td>
<td>0.3956</td>
<td>0.7166</td>
<td>-0.0673</td>
<td>0.0212</td>
<td>0.2715</td>
<td>0.2781</td>
<td>1.0000</td>
<td>-0.1271</td>
<td>-0.2565</td>
<td>-0.0663</td>
<td>-0.8261</td>
<td>0.0159</td>
<td>0.2126</td>
<td>0.0640</td>
</tr>
<tr>
<td>x7</td>
<td>0.2592</td>
<td>-0.1635</td>
<td>0.0428</td>
<td>0.1034</td>
<td>0.2419</td>
<td>0.1451</td>
<td>-0.1271</td>
<td>1.0000</td>
<td>-0.1383</td>
<td>-0.0643</td>
<td>0.1551</td>
<td>-0.7263</td>
<td>-0.1994</td>
<td></td>
</tr>
<tr>
<td>sin(x3)</td>
<td>-0.2259</td>
<td>-0.2301</td>
<td>-0.2235</td>
<td>0.3195</td>
<td>-0.4618</td>
<td>-0.9400</td>
<td>-0.2566</td>
<td>-0.1383</td>
<td>1.0000</td>
<td>0.5242</td>
<td>0.2023</td>
<td>-0.0687</td>
<td>0.0959</td>
<td>0.0993</td>
</tr>
<tr>
<td>cos(x5)</td>
<td>-0.1888</td>
<td>-0.0986</td>
<td>-0.1924</td>
<td>0.2827</td>
<td>-0.0408</td>
<td>-0.5874</td>
<td>-0.0063</td>
<td>-0.943</td>
<td>0.5242</td>
<td>1.0000</td>
<td>0.0904</td>
<td>-0.1212</td>
<td>0.2079</td>
<td>0.1891</td>
</tr>
<tr>
<td>sin(x6)</td>
<td>-0.4270</td>
<td>-0.8198</td>
<td>-0.0675</td>
<td>0.0212</td>
<td>0.1034</td>
<td>0.2781</td>
<td>0.2781</td>
<td>1.0000</td>
<td>-0.2094</td>
<td>0.0904</td>
<td>-0.1212</td>
<td>0.2079</td>
<td>0.1891</td>
<td></td>
</tr>
<tr>
<td>cos(x6)</td>
<td>-0.2956</td>
<td>-0.5133</td>
<td>0.0675</td>
<td>0.0212</td>
<td>0.1034</td>
<td>0.2781</td>
<td>0.2781</td>
<td>1.0000</td>
<td>-0.2094</td>
<td>0.0904</td>
<td>-0.1212</td>
<td>0.2079</td>
<td>0.1891</td>
<td></td>
</tr>
<tr>
<td>sin(x7)</td>
<td>0.2030</td>
<td>0.1585</td>
<td>-0.7454</td>
<td>0.0954</td>
<td>-0.0551</td>
<td>-0.533</td>
<td>0.2125</td>
<td>0.2583</td>
<td>-0.0595</td>
<td>0.2079</td>
<td>-0.0640</td>
<td>-0.2540</td>
<td>1.0000</td>
<td>-0.4534</td>
</tr>
<tr>
<td>cos(x7)</td>
<td>-0.0179</td>
<td>-0.0022</td>
<td>-0.2822</td>
<td>0.1614</td>
<td>-0.1369</td>
<td>-0.0374</td>
<td>0.0540</td>
<td>-0.0209</td>
<td>0.0093</td>
<td>0.1981</td>
<td>-0.0640</td>
<td>-0.0640</td>
<td>1.0000</td>
<td>-0.4534</td>
</tr>
</tbody>
</table>

8 pts  a) What single predictor will produce the largest value of \( R^2 \) for a simple linear regression analysis of \( y \)? Explain.

predictor:________________

explanation:
Below are two JMP reports using $x_i$ as a predictor of $y$. Use these in parts b) through f).

b) What is the value of a test statistic and a corresponding $p$-value for testing whether $x_i$ and $x_i^2$ together provide some statistically detectable predictive power for describing $y$?

8 pts
c) Is the quadratic model in $x_i$ a statistically significant improvement over the linear one in terms of ability to predict or explain $y$? (Answer "yes" or "no" and provide supporting values of a test statistic and its corresponding $p$-value.)

Yes/No (Circle one only)

test statistic = __________________________  $p$-value = __________________________

d) Give 95% two-sided confidence limits for the standard deviation of log concentration of NO$_2$ particles at a given traffic volume, assuming there is a linear relation between $y$ and $x_i$. (Plug in completely, but you need not simplify.)

e) As it turns out, the standard error of predicted $y$ for $x_i = 7.0$ is .1205 under the linear model. Use this fact and make 95% prediction limits for the log concentration of NO$_2$ particles the next day that $x_i = 7.0$ at this location. (Plug in completely, but you need not simplify.)

f) $y$ and $x_i$ are both expressed on log scales. The least squares line for $y$ and $x_i$ corresponds to another function

$$\text{concentration of NO}_2 \text{ particles} \approx g(\text{traffic volume})$$

where $\text{concentration of NO}_2 \text{ particles}$ and $\text{traffic volume}$ are on the scale of "counts" (not "log counts"). What is this function? (Give the approximate relationship between traffic count and particle count implied by the SLR analysis.)
Henceforth, return to the use of full set of predictors in the modeling of $y$.

4 pts **g)** Do the correlations on page 2 indicate that there is multicollinearity in the data set? Answer "yes" or "no" and explain carefully (say exactly which correlation(s) indicate what).

Yes/No (Circle one only)

explanation:

4 pts **h)** In the problem context, why is it not surprising that there is substantial correlation between $x_i$ and $x_6$ (and, for that matter, between $x_i$ and the two variables Vardeman made from $x_6$)?

Below is a JMP report regarding models with the largest $R^2$ for a given number of predictors.

![All Possible Models](image)

4 pts **i)** From the information above, which model size looks potentially most effective? (Pick a number of predictors and justify your choice.)

number = __________
Notice that from the earlier *JMP* reports we know that \( \text{SSTot} = 30.89 \). Give the value of an \( F \) statistic and degrees of freedom for testing whether after accounting for traffic volume (\( x_1 \)) the other (meteorological and time-related) predictor variables together account for a statistically significant part of the variation in \( y \).

\[
F = \text{______________________} \quad \text{d.f.} = \text{___________, ____________}
\]

Below is a *JMP* report for a \( k = 5 \) predictor model. Use it in what follows.
What is the sample correlation between $y$ and $\hat{y}$ for the 5-predictor model? (Give a number.)

Do the two plots on the right side of the previous page indicate any serious problems with the 5-predictor MLR model? (What should they look like if "all is well"?)

There are histograms for the predictors $x_1, x_2, x_4,$ and $x_7$ on the last page of this exam. They may be helpful for thinking about the predictors.

According to the fitted 5-predictor model, $\hat{y}$ depends upon $x_7$ through the function

$$0.00215x_7 + 0.46915 \sin \left( \frac{2\pi x_7}{365} \right)$$

Interpret what this says about how pollution is developing over time. (You'll need to say what both terms summed above "do.")

The set of conditions $x_1 = 7.0, x_3 = 2.8, x_4 = 0.27, x_7 = 1095$ represents more or less "average" traffic volume, wind speed, and temperature difference on a day about 3 years after the beginning the study (and 1 year beyond the last data point in the data set). For this set of conditions, using the 5-predictor model $\hat{y} \approx 5.06$ and $SE_{\hat{y}} \approx 0.4333$. Give 95% prediction limits for an observed $y$ on that day. (Plug in completely, but you need not simplify.)
What about the set of conditions in n) potentially calls into question the usefulness of the predictions made there?