1. Scatterplot Matrix for Navy Hospital Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>Load</th>
<th>Xray</th>
<th>Beddays</th>
<th>Pop</th>
<th>Length</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load</td>
<td>Average daily patient load</td>
<td>1.0000</td>
<td>0.9074</td>
<td>0.9999</td>
<td>0.9357</td>
<td>0.6712</td>
<td>0.9856</td>
</tr>
<tr>
<td>Xray</td>
<td>Monthly X-ray Exposures</td>
<td>0.9074</td>
<td>1.0000</td>
<td>0.9071</td>
<td>0.9105</td>
<td>0.4466</td>
<td>0.9452</td>
</tr>
<tr>
<td>Beddays</td>
<td>Monthly Occupied Bed Days</td>
<td>0.9999</td>
<td>0.9071</td>
<td>1.0000</td>
<td>0.9332</td>
<td>0.6711</td>
<td>0.9860</td>
</tr>
<tr>
<td>Pop</td>
<td>Eligible Population</td>
<td>0.9357</td>
<td>0.9105</td>
<td>0.9332</td>
<td>1.0000</td>
<td>0.4629</td>
<td>0.9404</td>
</tr>
<tr>
<td>Length</td>
<td>Average Length of Stay</td>
<td>0.6712</td>
<td>0.4466</td>
<td>0.6711</td>
<td>0.4629</td>
<td>1.0000</td>
<td>0.5786</td>
</tr>
<tr>
<td>Hours</td>
<td>Monthly labor Hours Required</td>
<td>0.9856</td>
<td>0.9452</td>
<td>0.9860</td>
<td>0.9404</td>
<td>0.5786</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Figure 1**
From the correlation matrix, the three largest correlations are .9999 between LOAD and BEDDAYS, .9357 between POP and LOAD, and .9332 between POP and BEDDAYS. LOAD and BEDDAYS seem to be the variables most strongly involved in multicollinearity. The scatterplot matrix (see Fig. 1) confirms these conclusions.

2.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>490177488</td>
<td>98035498</td>
<td>237.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>4535052</td>
<td>412277</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>16</td>
<td>494712540</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root MSE</th>
<th>R-Square</th>
<th>Dependent Mean</th>
<th>Adj R-Sq</th>
<th>Coeff Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>642.08838</td>
<td>0.9908</td>
<td>4978.48000</td>
<td>0.9867</td>
<td>12.89728</td>
</tr>
</tbody>
</table>

R squared is pretty large indicating that this model explains most of the variability in monthly labor hours. However the analysis below show that there are problems with the model as well as cases in the data set that may not fit this model well.

Since all the explanatory variables have strong positive correlation with dependent variable, intuitively we expect that all the parameter estimates would be positive sign. However, from the following Parameter Estimates table (see below), we can see that the coefficients for LOAD, POP and LENGTH have negative point estimates. The high multicollinearity causes large variance of the parameter estimates, as seen with the Standard Errors of LOAD and LENGTH, indicating that these coefficients are poorly estimated. VIF’s of LOAD and BEDDAYS are extremely large, pointing to the fact that they are involved in multicollinearity.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Variance Inflation</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>1962.94816</td>
<td>1071.36170</td>
<td>1.83</td>
<td>0.0941</td>
<td>0</td>
<td>-395.10304 4320.99935</td>
</tr>
<tr>
<td>Load</td>
<td>Average daily patient load</td>
<td>1</td>
<td>-15.85167</td>
<td>97.65299</td>
<td>-0.16</td>
<td>0.8740</td>
<td>9597.57076</td>
<td>-230.78446 199.08111</td>
</tr>
<tr>
<td>Xray</td>
<td>Monthly X-ray Exposures</td>
<td>1</td>
<td>0.05593</td>
<td>0.02126</td>
<td>2.63</td>
<td>0.0234</td>
<td>7.94059</td>
<td>0.00914 0.10272</td>
</tr>
<tr>
<td>Beddays</td>
<td>Monthly Occupied Bed Days</td>
<td>1</td>
<td>1.58962</td>
<td>3.09208</td>
<td>-0.51</td>
<td>0.6174</td>
<td>8933.08650</td>
<td>-5.21601 8.39525</td>
</tr>
<tr>
<td>Pop</td>
<td>Eligible Population</td>
<td>1</td>
<td>-4.21867</td>
<td>7.17656</td>
<td>-0.59</td>
<td>0.5685</td>
<td>23.29386</td>
<td>-20.01416 11.57683</td>
</tr>
<tr>
<td>Length</td>
<td>Average Length of Stay</td>
<td>1</td>
<td>-394.31412</td>
<td>209.63954</td>
<td>-1.88</td>
<td>0.0867</td>
<td>4.27984</td>
<td>-855.72764 67.09940</td>
</tr>
</tbody>
</table>

The residual statistics and diagnostic plots below indicate that there is at least one possible y-outlier that is influential, one possible y-outlier that is not influential, and several possible x-outliers. (See Output Statistics table below and Figure 2 below.)
<table>
<thead>
<tr>
<th>Obs</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Std Error Mean Predict</th>
<th>Residual</th>
<th>Std Error Residual</th>
<th>Student Residual</th>
<th>Cook’s D</th>
<th>RStudent</th>
<th>Hat Diag H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>566.5200</td>
<td>775.0251</td>
<td>241.2323</td>
<td>-208.5051</td>
<td>595.0</td>
<td>-0.350</td>
<td>0.003</td>
<td>-0.3360</td>
<td>0.1412</td>
</tr>
<tr>
<td>2</td>
<td>696.8200</td>
<td>740.6702</td>
<td>331.1402</td>
<td>-43.8502</td>
<td>550.1</td>
<td>-0.0797</td>
<td>0.000</td>
<td>-0.0760</td>
<td>0.2660</td>
</tr>
<tr>
<td>3</td>
<td>1033</td>
<td>1104</td>
<td>278.5116</td>
<td>-70.7734</td>
<td>578.5</td>
<td>-0.122</td>
<td>0.001</td>
<td>-0.1167</td>
<td>0.1881</td>
</tr>
<tr>
<td>4</td>
<td>1604</td>
<td>1240</td>
<td>268.1298</td>
<td>363.1244</td>
<td>583.4</td>
<td>0.622</td>
<td>0.014</td>
<td>0.6042</td>
<td>0.1744</td>
</tr>
<tr>
<td>5</td>
<td>1611</td>
<td>1564</td>
<td>211.2372</td>
<td>46.9483</td>
<td>606.3</td>
<td>0.0774</td>
<td>0.000</td>
<td>0.0738</td>
<td>0.1082</td>
</tr>
<tr>
<td>6</td>
<td>1613</td>
<td>2151</td>
<td>279.9293</td>
<td>-538.0017</td>
<td>577.9</td>
<td>-0.931</td>
<td>0.034</td>
<td>-0.9249</td>
<td>0.1901</td>
</tr>
<tr>
<td>7</td>
<td>1854</td>
<td>1690</td>
<td>218.9976</td>
<td>164.4696</td>
<td>603.6</td>
<td>0.272</td>
<td>0.002</td>
<td>0.2607</td>
<td>0.1163</td>
</tr>
<tr>
<td>8</td>
<td>2161</td>
<td>1736</td>
<td>468.9903</td>
<td>424.3145</td>
<td>438.5</td>
<td>0.968</td>
<td>0.178</td>
<td>0.9645</td>
<td>0.5335</td>
</tr>
<tr>
<td>9</td>
<td>2306</td>
<td>2737</td>
<td>290.4749</td>
<td>-431.4090</td>
<td>572.6</td>
<td>-0.753</td>
<td>0.024</td>
<td>-0.7376</td>
<td>0.2047</td>
</tr>
<tr>
<td>10</td>
<td>3504</td>
<td>3682</td>
<td>585.2517</td>
<td>-177.9234</td>
<td>264.1</td>
<td>-0.674</td>
<td><strong>0.371</strong></td>
<td>-0.6560</td>
<td><strong>0.8308</strong></td>
</tr>
<tr>
<td>11</td>
<td>3572</td>
<td>3239</td>
<td>189.0989</td>
<td>332.6011</td>
<td>613.6</td>
<td>0.542</td>
<td>0.005</td>
<td>0.5239</td>
<td>0.0867</td>
</tr>
<tr>
<td>12</td>
<td>3741</td>
<td>4353</td>
<td>328.8507</td>
<td>-611.9330</td>
<td>551.5</td>
<td>-1.110</td>
<td>0.073</td>
<td>-1.1227</td>
<td>0.2623</td>
</tr>
<tr>
<td>13</td>
<td>4027</td>
<td>4257</td>
<td>314.0481</td>
<td>-230.5684</td>
<td>560.0</td>
<td>-0.412</td>
<td>0.009</td>
<td>-0.3956</td>
<td>0.2392</td>
</tr>
<tr>
<td>14</td>
<td>10344</td>
<td>8768</td>
<td>252.2617</td>
<td>1576</td>
<td>590.5</td>
<td>2.669</td>
<td>0.217</td>
<td><strong>4.2878</strong></td>
<td>0.1544</td>
</tr>
<tr>
<td>15</td>
<td>11732</td>
<td>12237</td>
<td>573.9168</td>
<td>-504.8574</td>
<td>287.9</td>
<td>-1.753</td>
<td><strong>2.036</strong></td>
<td>-1.9697</td>
<td><strong>0.7989</strong></td>
</tr>
<tr>
<td>16</td>
<td>15415</td>
<td>15038</td>
<td>585.7046</td>
<td>376.5491</td>
<td>263.1</td>
<td>1.431</td>
<td>1.692</td>
<td>1.5126</td>
<td>0.8321</td>
</tr>
<tr>
<td>17</td>
<td>18854</td>
<td>19321</td>
<td>599.9780</td>
<td>-466.2470</td>
<td>228.7</td>
<td>-2.039</td>
<td><strong>4.767</strong></td>
<td><strong>-2.4643</strong></td>
<td>0.8731</td>
</tr>
</tbody>
</table>
We see that R-squared remains the same in this model showing that the contributions for the prediction of labor hours from the two variables left out, are very small in the presence of the three variables left in the model: XRAY, BEDDAYS and LENGTH, making this smaller model a viable model.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root MSE</td>
<td>614.77942</td>
</tr>
<tr>
<td>R-Square</td>
<td>0.9901</td>
</tr>
<tr>
<td>Dependent Mean</td>
<td>4978.48000</td>
</tr>
<tr>
<td>Adj R-Sq</td>
<td>0.9878</td>
</tr>
<tr>
<td>Coeff Var</td>
<td>12.34874</td>
</tr>
</tbody>
</table>
In the following parameter estimates table for this model, the largest VIF is seen to be 11.3, which is comparably smaller. Thus the multicollinearity effect doesn’t seem to be quite as large as in the previous model. The t-statistic are all significant at .1 the p-values being all close to the .05 significance level.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Variance</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>1523.3892</td>
<td>786.8977</td>
<td>1.94</td>
<td>0.0749</td>
<td>0</td>
<td>-176.599</td>
</tr>
<tr>
<td>Xray</td>
<td>Month. X-ray Exposure</td>
<td>1</td>
<td>0.05299</td>
<td>0.02009</td>
<td>2.64</td>
<td>0.0205</td>
<td>7.737</td>
<td>0.00958</td>
</tr>
<tr>
<td>Beddays</td>
<td>Month. Occ. Bed Days</td>
<td>1</td>
<td>0.97848</td>
<td>0.10515</td>
<td>9.31</td>
<td>&lt;.0001</td>
<td>11.269</td>
<td>0.7513</td>
</tr>
<tr>
<td>Length</td>
<td>Average Length of Stay</td>
<td>1</td>
<td>-320.9508</td>
<td>153.1922</td>
<td>-2.10</td>
<td>0.0563</td>
<td>2.493</td>
<td>-651.9025</td>
</tr>
</tbody>
</table>

4.

From the plots of residual versus predicted value and other variables (see Fig. 2 below), no apparent curvature pattern exists and the model seems to fit well except one possible outlier. The normal probability plot (see Fig. 3) shows that the normal assumption of the residuals is plausible, but one extreme value seems to be affecting the plot. With F-statistic =431.97 and P-value<.0001, the model fit seems to be adequate. R-Square=0.9901 means that 99% of the total variation can be explained by the model. The t-statistics for x1 and x2 are significant with .05 significant level. 95% confidence intervals of the coefficients for these two variables don't include 0. The x3 variable is not significant as shown by both t-statistic and confidence interval but quiet close to significant level.

There is no strong evidence from residual plots (Figure. 4) that the variance is not constant as there is no clear pattern shown.
Hospital 14 is a y-outlier because studentized residual = 4.5584 > 3.72. (Similar conclusion can be reached using RStudent.) It’s also influential, given Cook’s D = 0.353 > 4/17 = .235. These two statistic provide evidence that hospital 14 might not fit the model well because it’s a clear y-outlier and also influential.

Hospital 17 is also highly influential but does appear to fit model better than Hospital 14. In general, this model fits the data fairly well, except that we need to investigate the outlier hospital data values further.
Figure 4

5.

i. With 2 explanatory variables, the model with BEDDAYS and LENGTH is selected because R-square is the largest, SSE is smallest and Cp value is closer to 3. With 3 variables, the model with XRAY, BEDDAYS, and LENGTH is chosen because this model has the largest R-square, smallest SSE and Cp value (see Figure. 4) closer to number of parameters 4. In the four-variable case, the R-square are quite similar for several models, but the model with LOAD, XRAY, POP, and LENGTH has Cp value closer to number of parameters (see Fig. 4), 5, thus it’s selected as the good model. In the above 3 models, I would probably choose the 3-variable model, because it is more parsimonious, all the variables are significant, and was the one chosen by the subset selection procedures. Note also that it has the smallest AIC of all models considered here.
<table>
<thead>
<tr>
<th>Model Index</th>
<th>Number in Model</th>
<th>R-Square</th>
<th>C(p)</th>
<th>MSE</th>
<th>SSE</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>3</td>
<td>0.9934</td>
<td>11.1047</td>
<td>254363</td>
<td>3052355</td>
<td>Beddays Pop Length</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.9934</td>
<td>11.1098</td>
<td>254430</td>
<td>3053163</td>
<td>Load Beddays Length</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0.9965</td>
<td>4.0234</td>
<td>145585</td>
<td>1601430</td>
<td>Xray Beddays Pop Length</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.9964</td>
<td>4.5190</td>
<td>152784</td>
<td>1680621</td>
<td>Load Xray Beddays Length</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>0.9964</td>
<td>4.5385</td>
<td>153066</td>
<td>1683724</td>
<td>Load Xray Pop Length</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>0.9934</td>
<td>13.0374</td>
<td>276509</td>
<td>3041597</td>
<td>Load Beddays Pop Length</td>
</tr>
</tbody>
</table>

![Fit Criterion for Hours with Model Index](image)

**Figure 4**

5. **ii.** and **iii.**

The detailed SAS output from the selection methods used for these are not reported here.

From the summary results, (ii) and (iii) selected the same model. The model is the same as that resulted in Problem 3, which is \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \), with three explanatory variables, \( x_1(\text{XRAY}) \), \( x_2(\text{BEDDAYS}) \), and \( x_3(\text{LENGTH}) \).
### Backward Elimination Method Final Step:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xray</td>
<td>0.0028</td>
<td>0.9933</td>
<td>8.75</td>
<td>0.0120</td>
</tr>
<tr>
<td>Beddays</td>
<td>0.0765</td>
<td>0.9197</td>
<td>236.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Length</td>
<td>0.0057</td>
<td>0.9904</td>
<td>17.61</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

### Summary of Backward Elimination

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable Removed</th>
<th>Label</th>
<th>Number Vars In</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>C(p)</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Load</td>
<td>Average daily patient load</td>
<td>4</td>
<td>0.0000</td>
<td>0.9965</td>
<td>4.0234</td>
<td>0.02</td>
<td>0.8815</td>
</tr>
<tr>
<td>2</td>
<td>Pop</td>
<td>Eligible Population</td>
<td>3</td>
<td>0.0004</td>
<td>0.9961</td>
<td>3.2582</td>
<td>1.36</td>
<td>0.2690</td>
</tr>
</tbody>
</table>

### Stepwise Selection Method Final Step:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xray</td>
<td>0.0028</td>
<td>0.9933</td>
<td>8.75</td>
<td>0.0120</td>
</tr>
<tr>
<td>Beddays</td>
<td>0.0765</td>
<td>0.9197</td>
<td>236.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Length</td>
<td>0.0057</td>
<td>0.9904</td>
<td>17.61</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Tolerance</th>
<th>Model R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load</td>
<td>0.000195</td>
<td>0.9964</td>
<td>0.77</td>
<td>0.3981</td>
</tr>
<tr>
<td>Pop</td>
<td>0.079216</td>
<td>0.9965</td>
<td>1.36</td>
<td>0.2690</td>
</tr>
</tbody>
</table>

*All variables left in the model are significant at the 0.1000 level.*

*No other variable met the 0.2000 significance level for entry into the model.*

### Summary of Stepwise Selection

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable Entered</th>
<th>Variable Removed</th>
<th>Label</th>
<th>Number Vars In</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>C(p)</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beddays</td>
<td>Monthly Occupied Bed Days</td>
<td>1</td>
<td>0.9779</td>
<td>0.9779</td>
<td>52.3131</td>
<td>618.37</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Length</td>
<td>Average Length of Stay</td>
<td>2</td>
<td>0.0154</td>
<td>0.9933</td>
<td>9.4667</td>
<td>29.95</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Xray</td>
<td>Monthly X-ray Exposures</td>
<td>3</td>
<td>0.0028</td>
<td>0.9961</td>
<td>3.2582</td>
<td>8.75</td>
<td>0.0120</td>
<td></td>
</tr>
</tbody>
</table>
The parameter estimates and the analysis of variance are shown below. The model selected has dropped both LOAD and POP, while the full model had both variables. From Cp value perspective, model is better. In the backward elimination procedure, since all the p-values of the variables are far away from .05, a slight change of cut-off level would not change the selected model. In the stepwise procedure, if the entry level is increased to (say) .25, the POP variable might be included since the p-value for POP is .26 is not far away from .2 entry level.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1946.80204</td>
<td>504.18193</td>
<td>2234856</td>
<td>14.91</td>
<td>0.0023</td>
</tr>
<tr>
<td>Xray</td>
<td>0.03858</td>
<td>0.01304</td>
<td>1311468</td>
<td>8.75</td>
<td>0.0120</td>
</tr>
<tr>
<td>Beddays</td>
<td>1.03939</td>
<td>0.06756</td>
<td>35482722</td>
<td>236.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Length</td>
<td>-413.75780</td>
<td>98.59828</td>
<td>2639576</td>
<td>17.61</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

**SAS PROGRAMS**

**Problems 1,2**

```sas
proc sgscatter data=labor;
  title "Scatterplot Matrix for Navy Hospital Data";
  matrix Hours Load Xray Beddays Pop Length;
run;
```

```sas
proc reg corr data=labor;
  model Hours = Load Xray Beddays Pop Length/clb vif r p influence;
  title "Navy Hospital Labor Needs: Model fit with all 17 hospitals";
run;
```

**Problem 3.4**

```sas
proc reg corr data=labor;
  model Hours = Xray Beddays Length/clb vif r p influence;
  title "Three Variable Model fit with all 17 hospitals";
run;
```

**Problem 5**

```sas
proc reg plots(only)=cp(label);
  model Hours = Load Xray Beddays Pop Length /selection=b sls=.05 details = all ;
  model Hours = Load Xray Beddays Pop Length/selection=stepwise
               sle=.2 sls=.1 details = all;
  model Hours = Load Xray Beddays Pop Length/selection=rsquare
               start=2 stop=4 best=4 cp sse mse;
  title "Model Selection Using 16 Navy Hospitals";
run;
```