Question 1

a) Two response variables are mentioned. The first is breast cancer recurrence and the second is death. It is not clear from the abstract how exactly these two variables were measured.

b) The article focuses on insulin level, but also mentions weight, stage of tumor, and treatment. So these could all potentially be factors. However, since weight in general is hard to control it is possible that it was used as a covariate as oppose to a factor and the same stage of tumor.

c) Based on the information contained in the abstract, does it appear that replication and randomization were used in the design and conduct of this study? Explain.
Given that over 500 women were included in the study it appears that replication was part of the design. However, the article does not mention whether the women were randomized to different treatment (or other) groups. The use of randomization is not mentioned at all.

d) The study was unlikely a survey, but since randomization is not discussed it could be either a designed experiment or an observational study.

Question 2 – Does Kudzu increase bone density?

a) Table with summary statistics for each treatment group is given below.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>0.218867</td>
<td>0.21900</td>
<td>0.20300</td>
<td>0.24500</td>
<td>0.0115873</td>
</tr>
<tr>
<td>Low dose</td>
<td>15</td>
<td>0.215933</td>
<td>0.21600</td>
<td>0.19800</td>
<td>0.23300</td>
<td>0.0115107</td>
</tr>
<tr>
<td>High dose</td>
<td>15</td>
<td>0.235067</td>
<td>0.23200</td>
<td>0.20600</td>
<td>0.26700</td>
<td>0.0187711</td>
</tr>
</tbody>
</table>

As we can see the High dose group had the highest mean BMD.

Further, the box-plot of the data is given below
b) The statistical model we would use to analyze this data is the 1-factor fixed effect model. It is given by the following equation:

\[ Y_{ij} = \mu + \tau_i + \epsilon_{ij} \]

Where \( \mu \) is the overall mean BMD, \( \tau_i \) is the effect of the \( i \)-th treatment (\( i = 1, 2, 3 \)) and \( \epsilon_{ij} \) is the error. The model assumes that \( \epsilon_{ij} \) are iid \( N(0, \sigma^2) \) and that \( \sum_{i=1}^{3} \tau_i = 0 \). Note, the levels of the factor here were fixed by the researchers and were not selected at random from a bigger population of levels, therefore this is a fixed effect model.

c) The ANOVA table is given below:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>0.00318564</td>
<td>0.00159282</td>
<td>7.72</td>
<td>0.0014</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>0.00866760</td>
<td>0.00020637</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>0.01185324</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d) The hypothesis of interests is:  
\( H_0 : \tau_i = 0 \) for all \( i \)

\( H_a : \) at least one \( \tau_i \neq 0 \)

The test statistics is \( F_{obs} = MS_{treat} / MSE = 7.72 \). From the ANOVA table on part c we see that the P-value = 0.0014 < \( \alpha = 0.05 \), therefore we reject the null hypothesis and conclude that we have evidence that the groups differ with respect to BMD.

e) As shown in lecture, a CI for the treatment means is calculated using the formula

\[ \bar{Y}_i \pm t_{a/2} \left( \frac{MS_E}{r_i} \right) \sqrt{\frac{MS_E}{n-a}} \]

These CI can also be obtained from SAS by adding the command CLM to Proc GLM (see SAS code for more details). The CI are given in the table below

<table>
<thead>
<tr>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>95% Lower limit</th>
<th>95% Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>0.2188667</td>
<td>0.211381</td>
<td>0.226352</td>
</tr>
<tr>
<td>Low dose</td>
<td>15</td>
<td>0.2159333</td>
<td>0.208448</td>
<td>0.223419</td>
</tr>
<tr>
<td>High dose</td>
<td>15</td>
<td>0.2350667</td>
<td>0.227581</td>
<td>0.242552</td>
</tr>
</tbody>
</table>

f) The output for the LSD comparisons is given below. The LSD is 0.0106. There are statistically significant differences between the high dose and the low dose and between the high does and the control where the differences are 0.0019 and 0.0162, respectively.
The pairwise comparisons were done in part (f) giving us indication which means differ. Further, looking and the numerical and graphical description of the data in part (a) the only reasonable contrast that we need to test here is the contrast comparing the high dose with the low dose and control combined. The hypothesis to test is:

\[ H_0 : \left( \mu_1 + \mu_2 \right)/2 - \mu_3. \]

From the SAS output we get that the contrast SS is 0.00312111 resulting in a test statistic \( F_{obs} = 15.12 \) with a P-value = 0.0004. We can conclude that there is a significant difference between the mean BMD in the high dose group and the average BMD in the low dose group and the control group combined.

Since the mean BMD in the high group was found to be statistically significantly different then the mean BMD in any of the other two groups. We can conclude that the high dose treatment is preferred to the others.

h) There are four assumptions that need to be checked. The first assumption is about the form of the model. It is checked by examining the plot of residuals versus treatment. The plot is given below. As we can see there is no clear pattern and the residuals are randomly spread around the 0 line.

The next assumptions we need to check is the assumption about the constant variance of the residuals. It is done by first looking at the ratio of the highest standard deviation and the lowest one. This ratio is \( s_{max} / s_{min} = 1.3374129 / 0.8201193 = 1.63 < 3 \), therefore the constant variance assumption is valid. This can be justified by examining the plot of
residuals versus fitted values which does not show major differences in the dispersion of the points around the 0 line. The plot is given below.

![Plot of Residuals versus Fitted values](image)

Finally, the last assumption we need to check is the normality of residuals. This is done by examining the normal probability plot of the residuals. In our case, this plot does not reveal any major departure from normality. The plot is given below.

![Normal Probability plot of Residuals](image)

Note, the assumption about the independence of the residuals can not be checked here since we do not know the order in which the observation where taken. However, we can plot the residuals versus the observation number as it appears in the data set. The plot is given below and does not reveal any major pattern suggesting that the independence assumption is valid as well.

![Residuals Versus the Order of the Data](image)
Marking Scheme

• Question 1 – total 26 marks
  a) 6 marks – part marks if mentioned only one response variable.
  b) 10 marks – part marks if mentioned some of the factors.
  c) 5 marks - Do not give full marks to someone who assumes that randomization was used without indicating that this information is missing from the article.
  d) 5 marks – Do not give full marks if the student concludes that it was definitely one or the other, since it’s not possible to tell from the article.

• Question 2 – total 70 marks
  a) 10 marks – part marks for answers that are missing the box-plot or table of means.
  b) 5 marks – part marks for answer without the model assumptions.
  c) 10 marks
  d) 5 marks – part marks for answers without the null and alternative hypothesis.
  e) 10 marks – part marks for answers without the formula. Only 0.5 marks off for any calculation mistake.
  f) 10 marks – part marks for answers without conclusions.
  g) 10 marks – if more then 1 contrast is mentioned, give full marks only if it was properly justified.
  h) 10 marks – do not take off marks if students did not plot the last plot (residual vs order) but they should say something about the independence assumption.