Biostats VHM 802 Course
Winter 2016, Atlantic Veterinary College, PEI
Henrik Stryhn

Final examination, 21 April 2016

All aids are allowed, except a computer-like device (including tablets and smartphones) and personal assistance. The exam consists of 3 questions, which have equal weight (*10 points each*) and should all be answered; further detail about the points is given for specific parts of each question. The duration of the exam is 3 hours.

Generally, all statistical models used should be specified, and to such detail that it is clear which terms are present and in which form. Your answers should generally (unless specified otherwise) be based on the information provided. Nevertheless, if at some point you think it is necessary to carry out additional analysis in statistical software, explain carefully the purpose of your proposed analysis and how you would implement it in the statistical software.

Question 1.

A study explored how different preparations and doses of a parathyroid extract (i.e. obtained from the parathyroid gland) affects the serum calcium levels of dogs. A total of 20 dogs were included in the study, and standard (S) or test (T) preparations were used, both in either low (L; 0.125 cubic-centimeter per kg) or high (H; 0.250 cc/kg) doses. Each dog was subjected to an injection of the extract in a certain preparation and dose on three days (March 15, March 24 and April 5; denoted here simply as 1–3), and had their serum calcium concentration measured 17 hours after injections. The actual allocations of injections and the measured serum calcium values are shown in the table below; each cell gives the injection type (e.g. TL) and the corresponding measurement (e.g. 14.7).

	Day		
Dog	1	2	3
1	TL, 14.7	TH, 15.4	SH, 14.8
2	TH, 15.8	TL, 14.3	SL, 14.8
3	SL, 13.8	SH, 17.0	TH, 16.0
4	SH, 15.0	SL, 14.5	TL, 14.0
5	TL, 15.1	TH, 15.0	SH, 15.8
6	TH, 17.0	TL, 16.5	SL, 15.0
7	SL, 12.0	SH, 13.8	TH, 14.0
8	SH, 15.0	SL, 14.0	TL, 14.6
9	TH, 14.4	SH, 13.8	TL, 14.4
10	TL, 13.6	SL, 15.3	TH, 17.2
11	SH, 14.6	TH, 15.4	SL, 14.0
12	SL, 15.8	TL, 15.0	SH, 15.2
13	TH, 16.2	TL, 14.0	SH, 13.0
14	TL, 14.0	TH, 13.8	SL, 14.0
15	SH, 13.0	SL, 14.0	TH, 14.0
16	SL, 13.2	SH, 16.0	TL, 14.9
17	TH, 15.8	SH, 16.0	TL, 15.0
10	TL, 13.0	SL, 13.4	TH, 13.8
19	SH, 15.2	TH, 16.2	SL, 15.0
20	SL, 14.2	TL, 14.1	SH, 15.0

Use the description of the study and the information contained in the Minitab and Stata listings on the following pages to answer the questions.

A) (*3 points*) Describe the experimental design in statistical terms, using standard descriptors such as factors, treatments, replication, blocks, balancedness, completeness, experimental units, hierarchical structure, nesting etc. Use your characterization of the design to motivate a statistical model for the data (possibly, but not necessarily, the model considered in the subsequent questions). Make sure to explain clearly the meaning of the terms in your model.

B) (*4 points*) Describe the statistical model used in the Minitab and Stata listings, and use the results to assess and interpret, by means of estimates and statistical tests, the different effects of the model. Note that the Minitab and Stata listings do not contain exactly the same information for the model. Draw conclusions about whether each of the factors affect the serum calcium concentrations — if yes, quantify such effect(s). If you need additional information to support you conclusions, describe how you would obtain such information and how you would use it (ideally you should use the information provided as much as you can).

C) (*1 point*) The results shown are for analysis of the measured serum concentrations. In order to assess the need for a transformation of the outcome, a Box-Cox analysis was carried out. The results (from Minitab) suggested to transform the outcome with a power of -1, based on an optimal λ-value of -1.35 and a 95% CI of $(-3.54, 1.04)$. Use this information as well as the Minitab plots shown (for Part C)) from analysis of both the transformed (`invcalc`$= 1/$`calcium`) and untransformed outcomes to discuss the need and usefulness of transforming the outcome.

D) (*2 points*) Motivated by the data structure and/or the experimental design, suggest at least two revisions of the model (on untransformed scale) to explore additional features of the data. Make sure to explain both the rationale behind your suggestions, and describe also their implementation in statistical software. (*Hint:* You should consider models/analyses that could provide new information, and irrespective of your conclusion from Part C) your suggestions should not be focused on possible transformations of the outcome.)

Minitab analysis and listing for Question 1:

```
MTB > GLM;
SUBC>   Response 'calcium';
SUBC>   Nodefault;
SUBC>   Categorical 'dog' 'prep' 'dose' 'day';
SUBC>   Terms dog prep dose day prep*dose prep*day dose*day prep*dose*day;
SUBC>   Means prep dose day prep*dose prep*day dose*day prep*dose*day;
SUBC>   TExpand;
SUBC>   TAnova;
SUBC>   TSummary;
SUBC>   TFactor;
SUBC>   TMeans;
SUBC>   TDiagnostics 0.
General Linear Model: calcium versus dog, prep, dose, day

Factor Information
Factor  Type   Levels  Values
dog     Fixed      20  1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
prep    Fixed       2  S, T
dose    Fixed       2  H, L
day     Fixed       3  1, 2, 3
```

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
dog	19	34.7273	51.21%	38.1437	2.0076	4.69	0.000
prep	1	1.4741	2.17%	1.1987	1.1987	2.80	0.105
dose	1	10.0920	14.88%	9.9159	9.9159	23.15	0.000
day	2	0.9303	1.37%	0.9303	0.4652	1.09	0.351
prep*dose	1	0.3101	0.46%	0.3029	0.3029	0.71	0.407
prep*day	2	6.0250	8.89%	6.0250	3.0125	7.03	0.003
dose*day	2	1.1913	1.76%	1.1913	0.5957	1.39	0.265
prep*dose*day	2	0.6369	0.94%	0.6369	0.3185	0.74	0.484
Error	29	12.4203	18.32%	12.4203	0.4283		
Total	59	67.8073	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.654435	81.68%	62.73%	53.3238	21.36%

Fits and Diagnostics for Unusual Observations

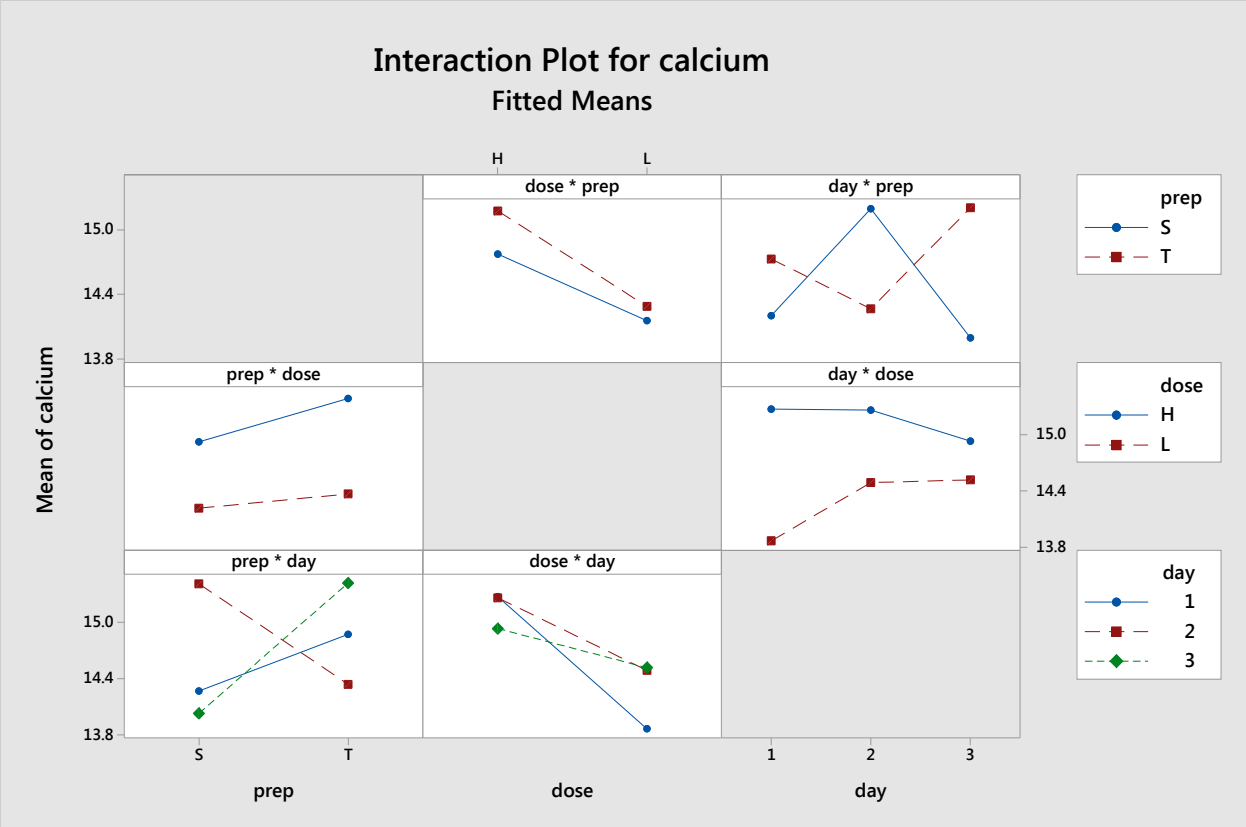
Obs	calcium	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS	
48	17.200	16.277	0.466	(15.323, 17.230)	0.923	2.01	2.13	0.507862	0.13	2.16422	R

R Large residual

Means

Term	Fitted Mean	SE Mean
prep		
S	14.569	0.125
T	14.877	0.125
dose		
H	15.156	0.123
L	14.291	0.123
day		
1	14.570	0.146
2	14.875	0.146
3	14.725	0.146
prep*dose		
S H	14.925	0.179
S L	14.214	0.179
T H	15.387	0.179
T L	14.367	0.179
prep*day		
S 1	14.268	0.234
S 2	15.410	0.256
S 3	14.030	0.256
T 1	14.872	0.234
T 2	14.340	0.256
T 3	15.420	0.256
dose*day		
H 1	15.274	0.239
H 2	15.261	0.232
H 3	14.933	0.239
L 1	13.866	0.239
L 2	14.489	0.232
L 3	14.517	0.239
prep*dose*day		

```
  S H 1        14.758    0.367
  S H 2        15.934    0.377
  S H 3        14.081    0.367
  S L 1        13.777    0.367
  S L 2        14.885    0.377
  S L 3        13.980    0.367
  T H 1        15.790    0.367
  T H 2        14.588    0.377
  T H 3        15.785    0.367
  T L 1        13.954    0.367
  T L 2        14.093    0.377
  T L 3        15.055    0.367

MTB > FacPlot 'calcium';
SUBC>   Factors dog prep dose day;
SUBC>   GInt;
SUBC>     Full.
Interaction Plot for calcium
```



Stata analysis and listings for Question 1:

```
. encode prep, g(Prep)
. encode dose, g(Dose)
. anova calcium dog Prep##Dose##day

                          Number of obs =         60    R-squared     =  0.8168
                          Root MSE      =    .654435    Adj R-squared =  0.6273
```

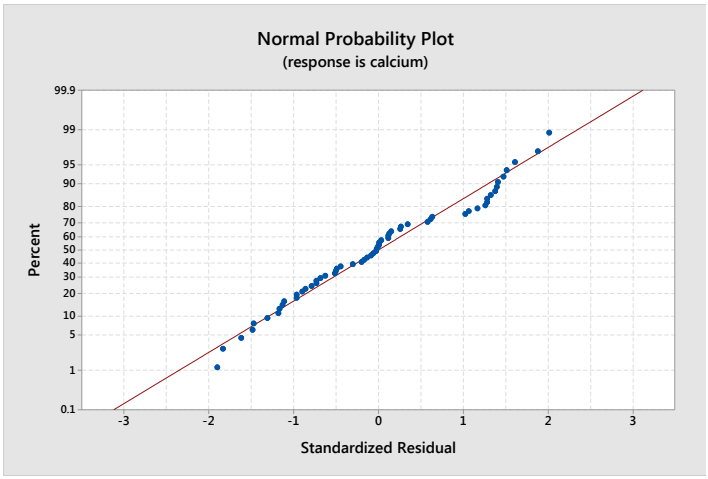
```
              Source | Partial SS         df         MS        F    Prob>F
-------------+------------------------------------------------------------
               Model |  55.387058         30   1.8462353      4.31  0.0001
                     |
                 dog |  38.143722         19   2.0075643      4.69  0.0001
                Prep |  1.1987148          1   1.1987148      2.80  0.1051
                Dose |  9.9158864          1   9.9158864     23.15  0.0000
           Prep#Dose |  .30292998          1   .30292998      0.71  0.4072
                 day |  .93033316          2   .46516658      1.09  0.3509
            Prep#day |   6.024973          2   3.0124865      7.03  0.0032
            Dose#day |  1.1913144          2   .59565722      1.39  0.2650
       Prep#Dose#day |  .63693142          2   .31846571      0.74  0.4843
                     |
            Residual |  12.420283         29   .42828563
-------------+------------------------------------------------------------
               Total |  67.807341         59    1.149277
```

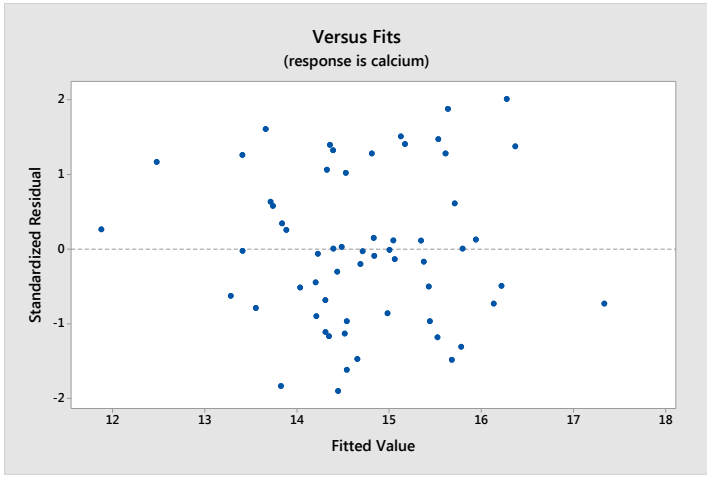
```
. regress
```

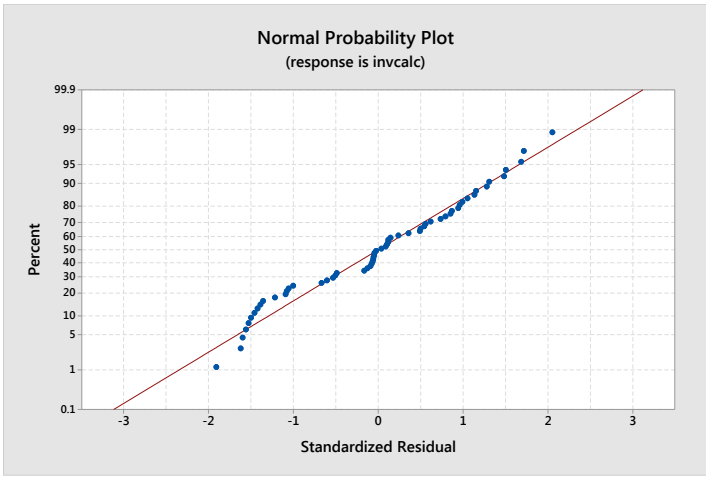
```
      Source |       SS           df       MS      Number of obs   =        60
-------------+----------------------------------   F(30, 29)       =      4.31
       Model |  55.3870578        30  1.84623526   Prob > F        =    0.0001
    Residual |  12.4202834        29  .428285634   R-squared       =    0.8168
-------------+----------------------------------   Adj R-squared   =    0.6273
       Total |  67.8073411        59  1.14927697   Root MSE        =    .65444

------------------------------------------------------------------------------
      calcium |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         dog |
           2  |  -.4129834   .6350358    -0.65   0.521    -1.711777    .8858106
           3  |  -.3243056   .6589644    -0.49   0.626    -1.672039    1.023428
           4  |  -1.158122   .6535014    -1.77   0.087    -2.494683    .1784381
           5  |   .3333337   .5343442     0.62   0.538     -.759523     1.42619
           6  |   .7870164   .6350358     1.24   0.225    -.5117776     2.08581
           7  |  -2.657639   .6589644    -4.03   0.000    -4.005372   -1.309905
           8  |  -1.124789   .6535014    -1.72   0.096    -2.461349    .2117716
           9  |  -2.151786   .6672939    -3.22   0.003    -3.516555   -.7870168
          10  |  -.2667529   .6118645    -0.44   0.666    -1.518156    .9846505
          11  |  -.5344244   .5967034    -0.90   0.378     -1.75482     .685971
          12  |   .5908854   .6118645     0.97   0.342     -.660518    1.842289
          13  |  -1.013219   .6144198    -1.65   0.110    -2.269849    .2434106
          14  |  -.9997642   .5622888    -1.78   0.086    -2.149774    .1502456
          15  |  -2.234747   .6474184    -3.45   0.002    -3.558866   -.9106276
          16  |  -.9810146   .6686058    -1.47   0.153    -2.348467    .3864378
          17  |  -.7517857   .6672939    -1.13   0.269    -2.116555    .6129834
          18  |   -2.23342   .6118645    -3.65   0.001    -3.484823   -.9820167
          19  |   .2655758   .5967034     0.45   0.660    -.9548197    1.485971
          20  |  -.3091145   .6118645    -0.51   0.617    -1.560518    .9422889
              |
        Prep  |
           T  |    1.03125   .5572316     1.85   0.074    -.1084163    2.170917
              |
        Dose  |
           L  |  -.9810637   .5644759    -1.74   0.093    -2.135547    .1734191
              |
   Prep#Dose  |
         T#L  |  -.8541672   .8406374    -1.02   0.318    -2.573464    .8651294
```

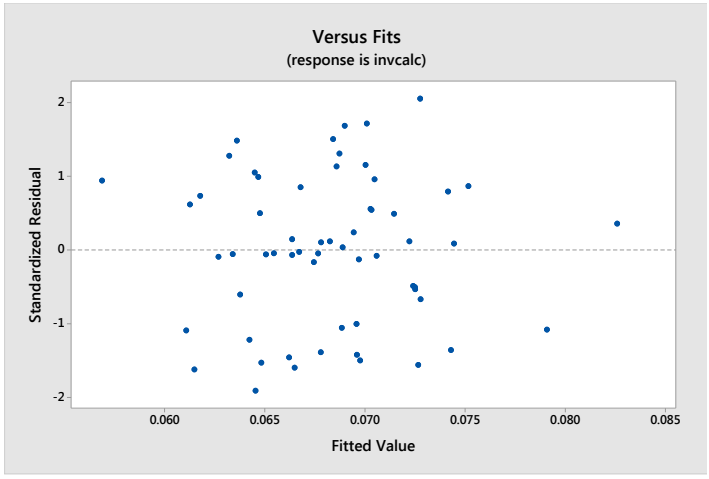
day						
2	1.176005	.5603965	2.10	0.045	.0298653	2.322144
3	-.6776785	.5626948	-1.20	0.238	-1.828519	.4731616
Prep#day						
T#2	-2.377501	.8901283	-2.67	0.012	-4.198018	-.5569839
T#3	.6724998	.8532793	0.79	0.437	-1.072652	2.417652
Dose#day						
L#2	-.0686763	.9098854	-0.08	0.940	-1.929601	1.792248
L#3	.8803568	.85507	1.03	0.312	-.8684577	2.629171
Prep#Dose#day						
T#L#2	1.408334	1.463362	0.96	0.344	-1.584577	4.401246
T#L#3	.2250004	1.206719	0.19	0.853	-2.243017	2.693018
_cons	15.5173	.5869395	26.44	0.000	14.31688	16.71773

Minitab plots for Question 1, Part C):









Question 2.

Studies of reproduction and breeding rely on valid measurements of the concentration of live sperm cells of the donors involved. Several investigations have been carried out at the Department of Reproduction at a Veterinary College to compare the precision of and agreement between different methods and instruments to assess semen quality. We consider here a dataset involving two methods (denoted simply as 1 and 2) that were applied to semen collections from 7 bulls. A semen collection was obtained from each of the bulls on three separate days (but these collections occurred on different days for different bulls). Furthermore, from each collection two samples were taken by different operators, and each of the samples was measured twice by the same operator. The structure of the dataset is shown in the Minitab listing below, where the two measurements on the same sample are distinguished by the variable `repl`; `conc1` and `conc2` denote the measurements (in 10^6 cells per ml) by methods 1 and 2, respectively; and `lnconc1` and `lnconc2` the natural log transformations of these values.

```
MTB > Print  'bull' 'day' 'sample' 'repl' 'conc1' 'conc2' 'lnconc1' 'lnconc2'.
Data Display

Row  bull  day  sample  repl  conc1  conc2  lnconc1  lnconc2
  1     1    1       1     1   3220   2960  8.07714  7.99294
  2     1    1       1     2   3130   3660  8.04879  8.20522
  3     1    1       2     1   3220   3800  8.07714  8.24276
  4     1    1       2     2   3280   3420  8.09560  8.13740
  5     1    2       1     1   2620   3040  7.87093  8.01961
  6     1    2       1     2   2500   2160  7.82405  7.67786
  7     1    2       2     1   2570   2680  7.85166  7.89357
  8     1    2       2     2   2600   2880  7.86327  7.96555
  9     1    3       1     1   2400   3140  7.78322  8.05198
 10     1    3       1     2   2370   2780  7.77065  7.93021
 11     1    3       2     1   2390   2960  7.77905  7.99294
 12     1    3       2     2   2430   3120  7.79565  8.04559
 13     2    1       1     1   1550   2100  7.34601  7.64969
 14     2    1       1     2   1690   1500  7.43248  7.31322
 15     2    1       2     1   1850   2340  7.52294  7.75791
 16     2    1       2     2   1890   2160  7.54433  7.67786
...
 81     7    3       1     1   1590   1220  7.37149  7.10661
 82     7    3       1     2   1510    940  7.31986  6.84588
 83     7    3       2     1   1630   2160  7.39634  7.67786
 84     7    3       2     2   1550   2020  7.34601  7.61085
```

Our main focus here is on comparing the precision of the two methods (other analyses have shown a fair agreement between the methods); therefore, the two sets of concentrations will be analyzed separately. Sperm concentration data are commonly analyzed and interpreted on logarithmic scale.

A) (*3 points*) The Minitab listing on the next pages shows analyses of the concentrations measured by the two methods. Analyses of the same models in Stata are shown as well; note however that the two listings do not contain exactly the same information. Describe the data structure and the statistical model on which these analyses are based, and explain how the model reflects the data structure. If you think the model is inadequate for the data structure, describe the issue(s) you identify and their potential impact, and suggest relevant improvements to the model.

B) (*4 points*) Compare the precision of the two methods in terms of the different variations (or variances/variability) estimated in the analysis, as well as their relative proportion of the total

variation. As part of the comparison, make sure to carefully interpret each of the variations. Compute also for each of the two methods the repeatability of duplicate measurements and the reproducibility of measurements by different operators. Estimate finally, in a similar way, the difference in (logarithmic) concentrations that can be expected between two collections from the same bull. Conclude briefly about how the two methods compare with respect to precision.

C) (*2 points*) For method 1 it was furthermore known that only two operators (say A and B) were involved in all analyses, and that all first samples (i.e., `sample=1`) were done by A and all second samples by B. Also, it was known that the recording for each sample listed as `repl=1` was done prior to the recording for `repl=2`, and it was hypothesized that the concentration of live sperm cells could decrease over the time between measurements; if needed, the actual times between measurements could be extracted from laboratory records. Explain, either by explicitly writing a statistical model or by indicating the specification of a model for analysis in statistical software, how you would take these two sources of variation into account and/or investigate whether there are any systematic differences between operators or recording times.

D) (*1 point*) After having completed the analysis for each method separately, it was also discussed whether the data for the two concentrations should be combined into a single dataset, in order to assess any systematic differences in concentrations between the two methods. If you think this is feasible and a good idea, give details about the steps involved and suggest a model to be used for the combined analysis. If you think this is either infeasible or not a good idea, explain the issue(s) you see with the proposed approach.

Minitab analysis listings for Question 2:

```
MTB > GLM;
SUBC>   Response 'lnconc1';
SUBC>   Nodefault;
SUBC>   Categorical 'bull' 'day' 'sample' 'repl';
SUBC>   Nest day(bull) sample(bull day);
SUBC>   Random day sample;
SUBC>   Terms bull day sample;
SUBC>   TExpand;
SUBC>   TAnova;
SUBC>   TSummary;
SUBC>   TFactor;
SUBC>   TEMS;
SUBC>   TVariance;
SUBC>   Rtype 2.
General Linear Model: lnconc1 versus bull, day, sample

Factor Information
Factor            Type    Levels  Values
bull              Fixed        7  1, 2, 3, 4, 5, 6, 7
day(bull)         Random      21  1(1), 2(1), 3(1), 1(2), 2(2), 3(2), 1(3), 2(3), 3(3),
                                  1(4), 2(4), 3(4), 1(5), 2(5), 3(5), 1(6), 2(6), 3(6),
                                  1(7), 2(7), 3(7)
sample(bull,day)  Random      42  1(1,1), 2(1,1), 1(1,2), 2(1,2), 1(1,3), 2(1,3), 1(2,1),
                                  2(2,1), 1(2,2), 2(2,2), 1(2,3), 2(2,3), 1(3,1), 2(3,1),
                                  1(3,2), 2(3,2), 1(3,3), 2(3,3), 1(4,1), 2(4,1), 1(4,2),
                                  2(4,2), 1(4,3), 2(4,3), 1(5,1), 2(5,1), 1(5,2), 2(5,2),
                                  1(5,3), 2(5,3), 1(6,1), 2(6,1), 1(6,2), 2(6,2), 1(6,3),
                                  2(6,3), 1(7,1), 2(7,1), 1(7,2), 2(7,2), 1(7,3), 2(7,3)
```

```
Analysis of Variance
Source              DF   Seq SS  Contribution   Adj SS    Adj MS  F-Value  P-Value
  bull               6  4.62242        64.54%  4.62242  0.770403     4.43    0.010
  day(bull)         14  2.43452        33.99%  2.43452  0.173894    49.33    0.000
  sample(bull,day)  21  0.07402         1.03%  0.07402  0.003525     4.78    0.000
Error               42  0.03097         0.43%  0.03097  0.000737
Total               83  7.16192       100.00%

Model Summary
        S    R-sq  R-sq(adj)     PRESS  R-sq(pred)
0.0271531  99.57%     99.15%  0.123865      98.27%

Expected Mean Squares, using Adjusted SS
   Source             Expected Mean Square for Each Term
1  bull               (4) + 2.0000 (3) + 4.0000 (2) + Q[1]
2  day(bull)          (4) + 2.0000 (3) + 4.0000 (2)
3  sample(bull,day)   (4) + 2.0000 (3)
4  Error              (4)

Variance Components, using Adjusted SS
Source              Variance  % of Total     StDev  % of Total
day(bull)          0.0425923   (omitted)  0.206379   (omitted)
sample(bull,day)   0.0013938           .  0.037334           .
Error              0.0007373           .  0.027153           .
Total              0.0447234           .  0.211479           .

MTB > GLM;
SUBC>   Response 'lnconc2';
...
General Linear Model: lnconc2 versus bull, day, sample
...
Analysis of Variance
Source              DF  Seq SS  Contribution  Adj SS   Adj MS  F-Value  P-Value
  bull               6  4.9173        57.74%  4.9173  0.81955     5.70    0.003
  day(bull)         14  2.0118        23.62%  2.0118  0.14370     2.93    0.013
  sample(bull,day)  21  1.0310        12.11%  1.0310  0.04910     3.71    0.000
Error               42  0.5560         6.53%  0.5560  0.01324
Total               83  8.5162       100.00%

Model Summary
       S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
0.115059  93.47%     87.10%  2.22409      73.88%

Expected Mean Squares, using Adjusted SS
   Source             Expected Mean Square for Each Term
1  bull               (4) + 2.0000 (3) + 4.0000 (2) + Q[1]
2  day(bull)          (4) + 2.0000 (3) + 4.0000 (2)
3  sample(bull,day)   (4) + 2.0000 (3)
4  Error              (4)

Variance Components, using Adjusted SS
Source              Variance  % of Total     StDev  % of Total
day(bull)          0.0236515   (omitted)  0.153790   (omitted)
sample(bull,day)   0.0179294           .  0.133901           .
Error              0.0132386           .  0.115059           .
Total              0.0548195           .  0.234136           .
```

Stata analysis and listings for Question 2:

```
. egen bull_day=group(bull day)
. egen bull_day_sample=group(bull day sample)
. mixed lnconc1 i.bull || bull_day: || bull_day_sample:, reml
...

Mixed-effects REML regression                   Number of obs     =         84
-----------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+------------------------------------------
       bull_day |         21          4        4.0          4
   bull_day_s~e |         42          2        2.0          2
-----------------------------------------------------------

                                                Wald chi2(6)      =      26.58
Log restricted-likelihood =  105.05584          Prob > chi2       =     0.0002
------------------------------------------------------------------------------
     lnconc1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bull |
          2  |  -.6507607    .170242    -3.82   0.000    -.9844289   -.3170925
          3  |  -.6198883    .170242    -3.64   0.000    -.9535565   -.2862201
          4  |  -.6773104    .170242    -3.98   0.000    -1.010979   -.3436422
          5  |  -.5519478    .170242    -3.24   0.001     -.885616   -.2182796
          6  |   -.731343    .170242    -4.30   0.000    -1.065011   -.3976748
          7  |  -.6888122    .170242    -4.05   0.000     -1.02248    -.355144
             |
       _cons |   7.903094   .1203793    65.65   0.000     7.667155    8.139033
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
bull_day: Identity           |
                  var(_cons) |   .0425923   .0164337      .0199943    .0907311
-----------------------------+------------------------------------------------
bull_day_s~e: Identity       |
                  var(_cons) |   .0013938   .0005498      .0006433    .0030198
-----------------------------+------------------------------------------------
               var(Residual) |   .0007373   .0001609      .0004807    .0011308
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 183.31                Prob > chi2 = 0.0000

. mixed lnconc2 i.bull || bull_day: || bull_day_sample:, reml
...
Mixed-effects REML regression                   Number of obs     =         84

-----------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+------------------------------------------
       bull_day |         21          4        4.0          4
   bull_day_s~e |         42          2        2.0          2
-----------------------------------------------------------

                                                Wald chi2(6)      =      34.22
Log restricted-likelihood =   18.08795          Prob > chi2       =     0.0000
```

```
------------------------------------------------------------------------------
     lnconc2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bull |
          2  |  -.5200365   .1547597    -3.36   0.001    -.8233599   -.2167131
          3  |  -.5703341   .1547597    -3.69   0.000    -.8736575   -.2670106
          4  |  -.7162567   .1547597    -4.63   0.000     -1.01958   -.4129333
          5  |  -.6612296   .1547597    -4.27   0.000     -.964553   -.3579062
          6  |  -.8068065   .1547597    -5.21   0.000     -1.11013   -.5034831
          7  |  -.5345612   .1547597    -3.45   0.001    -.8378847   -.2312378
             |
       _cons |   8.012969   .1094316    73.22   0.000     7.798487    8.227451
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
bull_day: Identity           |
                  var(_cons) |   .0236515   .0140971      .0073538    .0760686
-----------------------------+------------------------------------------------
bull_day_s~e: Identity       |
                  var(_cons) |   .0179294   .0077124      .0077165    .0416593
-----------------------------+------------------------------------------------
               var(Residual) |   .0132387   .0028889      .0086317    .0203044
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 36.22                 Prob > chi2 = 0.0000
```

Question 3.

A study was carried out to identify risk factors for human esophageal cancer (i.e., cancer related to the esophagus, the food pipe that runs between the throat and the stomach). Two hundred adult men that had been diagnosed with this form of cancer at one of the hospitals in a particular region of France were selected for the study, together with close to 800 other men drawn from electoral lists in the same region. All subjects were administered a dietary interview which contained questions about their consumption of tobacco and alcoholic beverages in addition to those about foods. Our focus here is on the impact of the tobacco and alcohol consumption as well as age on the cancer risk; these variables are summarized in the table below.

Variable	Description	Values
`cancer`	diagnosed with esophageal cancer	0/1 (no/yes)
`age`	age in years	$25-91$
`tob`	tobacco consumption (g/day)	$0-60$
`alc`	alcohol consumption (g/day)	$0-268$

The tobacco consumption was actually recorded in intervals: 0, $0-5$, $5-10$, $10-15$, $15-20$, $20-30$, $30-40$, $40-50$, >50, and the values of the `tob` variable are category midpoints (i.e., 0, 2.5, 7.5, 12.5, 17.5, 25, 35, 45, and 60 g/day). The two other variables were ungrouped, with integer values.

A) (*2 points*) Describe the study type, and discuss the epidemiological roles of the three predictors. For the latter, you may assume that the main hypothesis prior to the study was that the consumption of alcohol would be related to the risk of this type of cancer.

B) (*3 points*) Explain the statistical model behind the analysis shown in the Stata listings[1] shown for part B); you may either give a model formula or a verbal description of the model and the outcome/parameter being modelled. Interpret the parameters of the model; your interpretation should for each effect include both a quantitative statement of the size of the effect and a statement about its statistical significance. Review briefly the model assumptions, and discuss what the information provided tells you about their validity for these data.

In addition, all three variables were also categorized into groups based on selected cutpoints in their distributions, as shown below together with the count of observations in each group.

`agegp`	Count	`tobgp`	Count	`alcgp`	Count
$25-34$	116	$0-9$	526	$0-39$	414
$35-44$	199	$10-19$	236	$40-79$	355
$45-54$	213	$20-29$	131	$80-119$	139
$55-64$	242	≥ 30	82	≥ 120	67
$65-74$	161				
≥ 75	44				

C) (*3 points*) Explain the statistical model analyzed for Part C), and compare it with the previous model in terms of predictor effects and model fit. For each of the three predictors studied, compare the results obtained with the two models and suggest, based on the information obtained from both models, the "best" way (in your view) to model the effect of the predictor. If you need additional information to support your analysis/conclusions, describe how you would obtain such information and how you would use it (ideally you should use the information provided as much as you can).

[1] Minitab listings are available upon request.

D) (*2 points*) Suggest at least two supplementary analyses (or sets of analyses) you would want to carry out to determine causal or other roles of the model's predictors, or to support the model's validation or interpretation. Make sure to explain both the rationale behind your suggestions, and describe also the implementation in statistical software. (*Hint:* You should consider models/analyses that would provide new information, without including suggestions from Part C) related to the choice of the "best" way to model each predictor.)

Stata listing for Question 3, Part B):

```
. logit cancer age tob alc

Iteration 0:   log likelihood = -494.74421
Iteration 1:   log likelihood = -374.32138
Iteration 2:   log likelihood = -359.33245
Iteration 3:   log likelihood = -359.14054
Iteration 4:   log likelihood = -359.14052

Logistic regression                             Number of obs     =        975
                                                LR chi2(3)        =     271.21
                                                Prob > chi2       =     0.0000
Log likelihood = -359.14052                     Pseudo R2         =     0.2741

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0722892   .0082232     8.79   0.000      .056172    .0884064
         tob |    .038803   .0075749     5.12   0.000     .0239565    .0536495
         alc |   .0264904    .002525    10.49   0.000     .0215416    .0314393
       _cons |  -7.572919   .5864228   -12.91   0.000    -8.722287   -6.423551
------------------------------------------------------------------------------

. estat gof
Logistic model for cancer, goodness-of-fit test

       number of observations =       975
 number of covariate patterns =       955
           Pearson chi2(951) =       835.62
                 Prob > chi2 =         0.9970

. estat gof, g(10)
Logistic model for cancer, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       975
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        15.83
                 Prob > chi2 =         0.0449
```

(continues on next page)

Stata listing for Question 3, Part C):

```
. logit cancer i.agegp i.tobgp i.alcgp

Iteration 0:   log likelihood = -494.74421
Iteration 1:   log likelihood =  -375.9178
Iteration 2:   log likelihood =  -354.1739
Iteration 3:   log likelihood =  -352.2086
Iteration 4:   log likelihood = -352.13664
Iteration 5:   log likelihood = -352.13648
Iteration 6:   log likelihood = -352.13648

Logistic regression                             Number of obs     =        975
                                                LR chi2(11)       =     285.22
                                                Prob > chi2       =     0.0000
Log likelihood = -352.13648                     Pseudo R2         =     0.2882

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agegp |
      35-44  |   1.969249   1.104626     1.78   0.075    -.1957782    4.134276
      45-54  |   3.772413   1.068688     3.53   0.000     1.677823    5.867004
      55-64  |   4.328333   1.065716     4.06   0.000     2.239569    6.417098
      65-74  |   4.892692   1.077035     4.54   0.000     2.781741    7.003642
        75+  |   4.820212   1.121893     4.30   0.000     2.621342    7.019081
             |
       tobgp |
      10-19  |   .4430374    .228129     1.94   0.052    -.0040871     .890162
      20-29  |    .515671   .2729527     1.89   0.059    -.0193064    1.050648
        30+  |   1.644375    .344021     4.78   0.000     .9701059    2.318643
             |
       alcgp |
      40-79  |   1.438004   .2501848     5.75   0.000     .9476512    1.928358
     80-119  |   1.966517   .2843235     6.92   0.000     1.409253    2.523781
      120+   |   3.603085    .385185     9.35   0.000     2.848136    4.358034
             |
       _cons |  -6.891574   1.086532    -6.34   0.000    -9.021138    -4.76201
------------------------------------------------------------------------------

. estat gof
Logistic model for cancer, goodness-of-fit test

       number of observations =       975
 number of covariate patterns =        88
             Pearson chi2(76) =        85.98
                  Prob > chi2 =         0.2033

. estat gof, g(10)
Logistic model for cancer, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       975
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =         4.20
                  Prob > chi2 =         0.8383
```