Biostats VHM 802 Course
Winter 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Final examination, 22 April 2021 (for students with VHM 8120)

This exam is for students who are completing VHM 8120 in the same semester or have completed VHM 8120 in a previous semester. Students not taking, nor having already taken VHM 8120 for credit should *not* answer this exam.

The final exam is a 24-hour take home exam, and worth 30% of the course mark. The exam is open book, but must be completed by each student individually without assistance from other people. By handing in the exam you implicitly acknowledge to have read, accepted, and agreed to comply with the instructions for the exam, as presented in the page entitled "Instructions for home assignments and exam" at the VHM 802 course homepage.

The exam consists of two equally-weighted questions that should both be answered. Further weights are indicated for subquestions within the two questions.

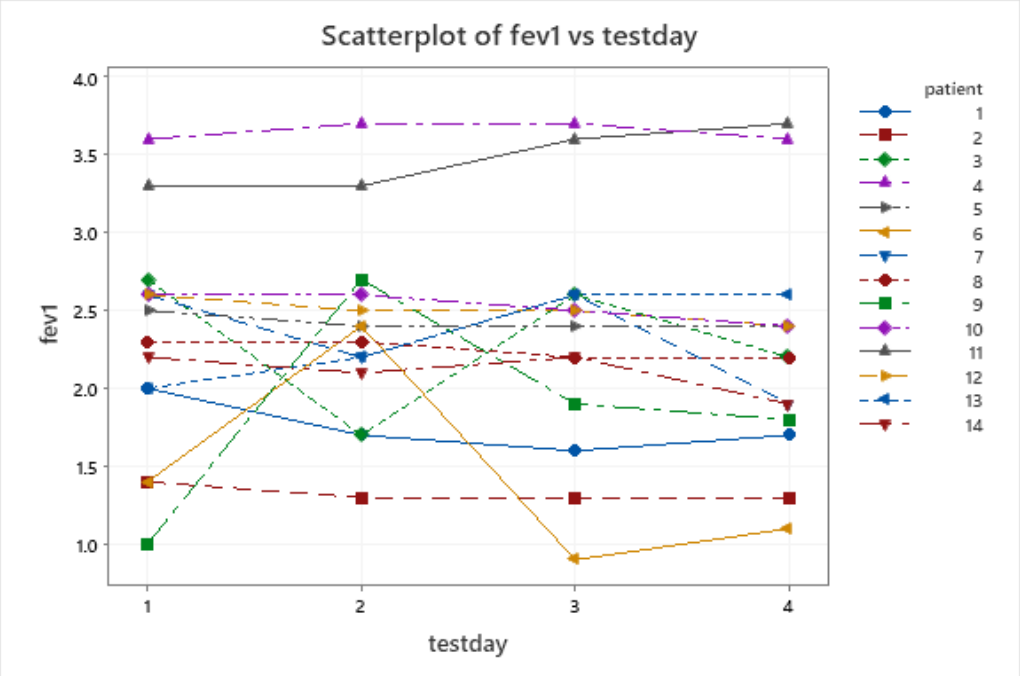
Question 1.

A trial was carried out to compare different administrations of a drug, formoterol, against asthma in humans. The drug could be administered by one of two formulations (a suspension formulation and a solution formulation) and in two doses (12 μg and 24 μg). The suspension and solution canisters are indistinguishable but deliver only 12 μg per "puff", and thus to take 24 μg two puffs are needed. For each treatment, the patient was provided with two canisters and instructed to take one puff from each, and the treatments were obtained from the following combinations:

Treatment	Description
A	one puff 12 μg formoterol suspension + one puff placebo
B	one puff 12 μg formoterol suspension + another identical puff
C	one puff 12 μg formoterol solution + one puff placebo
D	one puff 12 μg formoterol solution + another identical puff

There were 16 patients, nine men and seven women, who tried the four treatments at four separate test days. These days are labeled $1-4$ although they were different for different patients and separated by several (at least 4) days. Various lung function measures were obtained throughout the test day, and we consider here the Forced Expiratory Volume in the first second (FEV1) measured 6 hours after the treatment. FEV1 measures the volume of air that can be forced out in one second after taking a deep breath, and is an important measure of pulmonary function. The table on the next page shows for each patient and test day the FEV1 value and the treatment given on that day. Some values have been masked (by $*$) in the table so as to discourage attempts to enter the data into statistical software for analysis; instead you should use the Minitab and Stata listings provided below. The data are also displayed graphically, however including only the first 14 of the 16 patients.

FEV1 (Treatment)	Test day			
Patient	1	2	3	4
1	2.0 (D)	1.7 (A)	1.6 (C)	1.7 (B)
2	1.4 (C)	1.3 (D)	1.3 (B)	1.3 (A)
3	2.7 (A)	1.7 (B)	2.6 (D)	2.2 (C)
4	3.6 (B)	3.7 (C)	3.7 (A)	3.6 (D)
5	2.5 (A)	2.4 (B)	2.4 (D)	2.4 (C)
6	1.4 (B)	2.4 (C)	0.9 (A)	1.1 (D)
7	2.6 (D)	∗ (A)	2.6 (C)	1.9 (B)
8	2.3 (C)	2.3 (D)	2.2 (B)	2.2 (A)
9	1.0 (C)	2.7 (D)	1.9 (B)	1.8 (A)
10	2.6 (B)	2.6 (C)	∗ (A)	2.4 (D)
11	3.3 (D)	3.3 (A)	3.6 (C)	3.7 (B)
12	2.6 (A)	2.5 (B)	2.5 (D)	2.4 (C)
13	2.0 (A)	2.2 (B)	2.6 (D)	2.6 (C)
14	2.2 (C)	2.1 (D)	2.2 (B)	1.9 (A)
15	∗ (D)	∗ (A)	∗ (C)	∗ (B)
16	∗ (B)	∗ (C)	∗ (A)	∗ (D)



Use the data listing, the graph and the information contained in the subsequent Minitab and Stata listings (for Question 1) to answer the following questions.

A) (*6 points*) Describe the experimental design in statistical terms, and explain for each of the properties (descriptors) you list for the design also the advantage(s) associated with that particular property. Outline briefly how randomisation could be done for this type of design, or seems to have been done in the present study. In addition, describe the data structure obtained in the study, if possible in terms of a suitable diagram. Finally, review briefly the information you can extract from the above graph.

B) (*6 points*) Explain the statistical model used in the listing; you may either give a model formula with the relevant model assumptions, or give a verbal description of the model and its assumptions. Discuss, based on the information provided, whether the assumptions behind the model seem to be met to a satisfactory degree. If you think this is not the case, describe how you would continue the analysis to obtain a satisfactory model. Include in the discussion any suggestions you might have for supplementary analysis, either with the purpose of checking model assumptions or with the purpose of examining additional effects of potential interest. Supplement your suggestions with relevant details on how you would carry out the proposed analysis (in statistical software of your choice).

C) (*3 points*) Irrespective of any critique you might have had about the model in part B), assume for this part the model to be "acceptable" for statistical inference. Draw conclusions about all effects included in the model/analysis. Include an assessment of the effects of drug dose and drug formulation based on the information provided. If you think that additional analysis is needed to assess some effects of interest, outline how the additional analysis should be carried out (again, in statistical software of your choice) and indicate what information (e.g. estimates or tests) you would be able to obtain from it.

Minitab listing and plots for Question 1:

Factor Information

Factor	Type	Levels	Values
patient	Fixed	16	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
testday	Fixed	4	1, 2, 3, 4
tx	Fixed	4	A, B, C, D

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
patient	15	22.2098	1.48066	14.69	0.000
testday	3	0.1380	0.04599	0.46	0.714
tx	3	0.5167	0.17224	1.71	0.180
Error	42	4.2328	0.10078		
Total	63	27.0973			

Fits and Diagnostics for Unusual Observations

Obs	fev1	Fit	Resid	Std Resid	
9	1.000	1.922	-0.922	-3.58	R
19	1.700	2.284	-0.584	-2.27	R
22	2.400	1.591	0.809	3.15	R
25	2.700	2.022	0.678	2.64	R

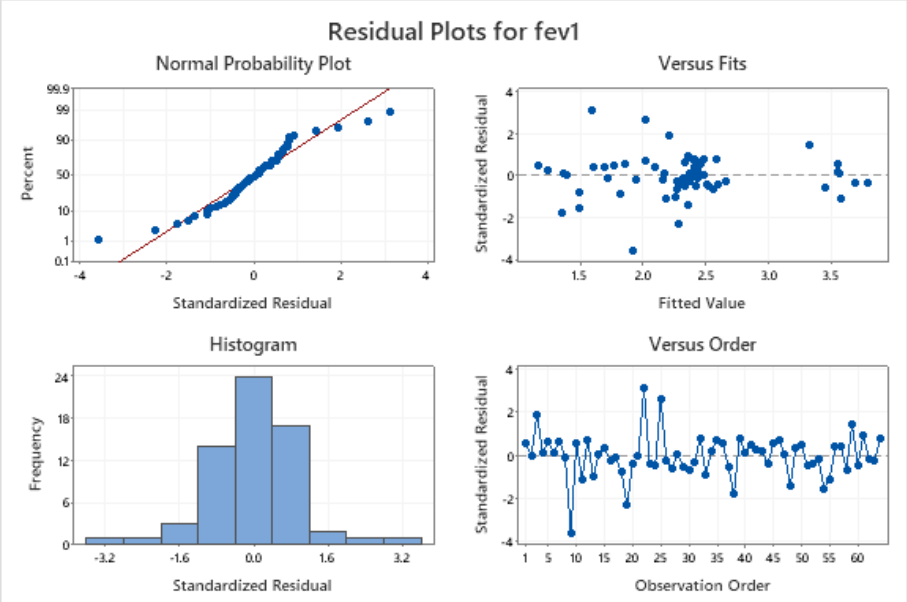
R Large residual

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.317461	84.38%	76.57%	63.73%

Means

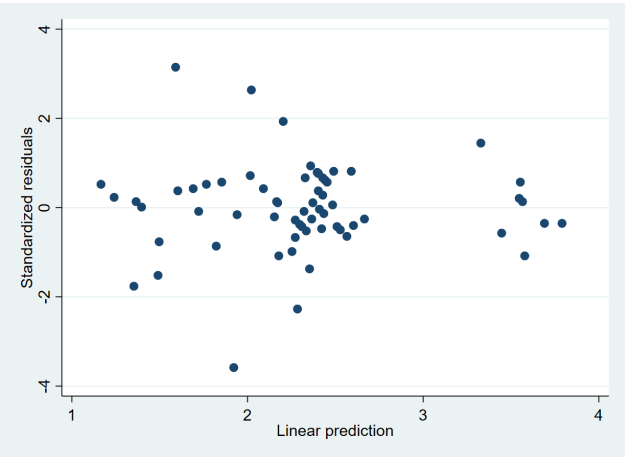
Term	Fitted Mean	SE Mean
tx		
A	2.2188	0.0794
B	2.2313	0.0794
C	2.3875	0.0794
D	2.4188	0.0794

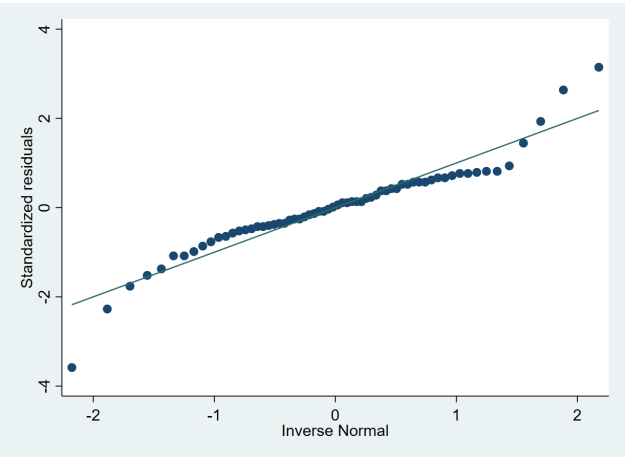


Stata listing and plots for Question 1:

```
. anova fev1 patient tx testday

                          Number of obs =         64    R-squared     =  0.8438
                          Root MSE      =    .317461    Adj R-squared =  0.7657

                   Source | Partial SS         df         MS        F     Prob>F
                ----------+----------------------------------------------------
                    Model |  22.864531         21   1.0887872     10.80  0.0000
                          |
                  patient |  22.209843         15   1.4806562     14.69  0.0000
                       tx |   .5167187          3   .17223957      1.71  0.1797
                  testday |  .13796876          3   .04598959      0.46  0.7142
                          |
                 Residual |  4.2328125         42   .10078125
                ----------+----------------------------------------------------
                    Total |  27.097343         63   .43011656

. regress fev1 i.patient i.tx i.testday

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(21, 42)       =     10.80
       Model |  22.8645307        21  1.08878718   Prob > F        =    0.0000
    Residual |  4.23281253        42  .100781251   R-squared       =    0.8438
-------------+----------------------------------   Adj R-squared   =    0.7657
       Total |  27.0973432        63   .43011656   Root MSE        =    .31746

------------------------------------------------------------------------------
        fev1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     patient |
           2 |  -.4250001   .2244786    -1.89   0.065    -.8780162     .028016
           3 |        .55   .2244786     2.45   0.019     .0969839    1.003016
         ... |         ...   ...          ...    ...       ...          ...
```

```
            |
         tx |
          B |      .0125   .1122393     0.11   0.912     -.214008      .239008
          C |     .16875   .1122393     1.50   0.140     -.057758      .395258
          D |         .2   .1122393     1.78   0.082    -.0265081      .426508
            |
    testday |
          2 |     .06875   .1122393     0.61   0.543     -.157758     .2952581
          3 |   2.24e-08   .1122393     0.00   1.000     -.226508     .2265081
          4 |     -.0625   .1122393    -0.56   0.581     -.289008     .1640081
            |
      _cons |   1.653125   .1861278     8.88   0.000     1.277504     2.028746
------------------------------------------------------------------------------

. margins tx, asbalanced

Adjusted predictions                            Number of obs     =         64
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : patient          (asbalanced)
               tx               (asbalanced)
               testday          (asbalanced)
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |
          A  |    2.21875   .0793652    27.96   0.000     2.058585    2.378915
          B  |    2.23125   .0793652    28.11   0.000     2.071085    2.391415
          C  |     2.3875   .0793652    30.08   0.000     2.227335    2.547665
          D  |    2.41875   .0793652    30.48   0.000     2.258585    2.578915
------------------------------------------------------------------------------

. predict fit, xb
. predict stdres, rstandard
. scatter stdres fit
. qnorm stdres
```





Question 2.

Data on average protein consumption from different food sources for the inhabitants of a number of European countries are shown in Table 6.7 of the Manly textbook. For Table 10.4, these data are combined with the data on employment in different sectors of the European countries (also studied in the book). A version of this combined dataset, comprising 24 European countries, has been put together for the question. In addition to the variables of the European employment data (including the four political groups of the countries), the data contain average protein consumption (g per person per day) by food type categories, and represented in the following variables (as in the Manly text):

Variable	Description
`rm`	red meat
`wm`	white meat
`egg`	eggs
`milk`	milk products
`fish`	fish
`cer`	cereals (as a food group)
`stch`	starchy foods (suitably defined to avoid overlap with cereals)
`pno`	pulses, nuts and oilseeds
`fveg`	fruits and vegetables

An extra grouping variable (`group2`) is included in the data as well, where the countries are labelled as either Western or Eastern European. Note that the protein source data are older than the employment data, and the grouping from the employment data does not correctly describe the countries at the earlier time. On the other hand, the division into Western and Eastern European countries is valid for both sets of variables. You are not expected to discuss the construction of the data, just to work with the data as provided to you. A datafile in comma-separated (`.csv`) file format for import into any statistical software is made available for analysis.

Although the Chapter 10 exercise of the Manly text focuses on a canonical correlation analysis for these combined data, such an analysis is **not** requested here. Alternative objectives for data analysis that are of interest for this question are listed below (in no prioritized order):

a1. explore relationships/patterns among protein source variables,

a2. explore relationships/patterns among employment variables,

b1. explore relationships/patterns among countries based on protein source variables, with interpretation in the context of relevant groups,

b2. explore relationships/patterns among countries based on employment variables, with interpretation in the context of relevant groups,

b3. explore relationships/patterns among countries based on both protein source and employment variables, with interpretation in the context of relevant groups,

c1. explore the ability to determine relevant group membership based on protein source variables,

c2. explore the ability to determine relevant group membership based on employment variables,

c3. explore the ability to determine relevant group membership based on both protein source and employment variables,

(*continues on the next page*)

d1. explore the ability to describe most of (or suitably selected) information about employment by most of (or suitably selected) information about protein sources, in a reasonably succinct way[1] and with meaningful interpretations,

d2. explore the ability to describe most of (or suitably selected) information about protein sources by most of (or suitably selected) information about employment, in a reasonably succinct way[1] and with meaningful interpretations.

Your task for the exam is to conduct full (statistical) analyses according to **two of the above objectives**. You are free to choose any two objectives from the list, but they must be from different letter categories on the list. For example, **a1** and **b2** is a valid choice, but **b1** and **b2** is not (because both from the **b.** categories). For each objective, it may be natural and helpful to include several methods (of analysis) but they will still only count for one objective. Your selected two analyses/objectives will be worth **7.5 points each**, for a total of 15 points for this question. It is allowed (though not recommended) to include one extra objective, in which case the mark will be for the two best analyses of the three. If more than three objectives are included, only the first three as they appear in the text will be considered. *Important note:* For all objectives, your analyses must **not** *essentially replicate* analyses already done in the course (in the lectures, the lab exercises or the Manly text).[2]

Recall that a full (statistical) analysis will typically include (beyond the specific objective): relevant descriptive analyses for the variables involved; specification of either a statistical model or a method of analysis that describes the assumptions involved, as well as assessment of these assumptions; estimates for the (relevant) model parameters or method characteristics with an indication of uncertainty (whenever applicable); any relevant statistical inference or comparison of settings of the method; conclusions worded both in statistical/methodological terms and in non-technical terms; and a discussion of your approach, e.g. related to its suitability or any potential analyses you chose to not include with your answers (e.g., you could indicate what extra information such analyses might be able to offer).

[1] It is perhaps easiest to explain by examples of what is **not** wished for: a table of all regression coefficients (or correlation coefficients) between the two sets of variables is clearly not succinct; on the other hand, a canonical correlation is obviously succinct but outside the course syllabus.

[2] An analysis will **not count**, if it is the same analysis of the European employment data as already performed/ discussed in the course, except by the present data including a smaller set of countries.