Final examination, 16 April 2025 (for students without VHM 8120)

All aids are allowed, except a fully operational computer and personal assistance. Restricted use of some computer-like devices (including laptops, tablets and smartphones) is permitted under the rules described at the VHM 802 course homepage. 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 was conducted to establish conditions under which a certain insect could best be used for biological control of the growth of a certain aquatic plant (a plant growing in water). The aquatic plant in question is non-native to the ecological system, and competes with a related native aquatic plant species. The insects feed on both native and non-native water plants. The experimental conditions studied were: whether an insect's parents were raised on non-native or native plants (coded as 0/1, where 0 corresponds to non-native plants, and 1 corresponds to native plants), whether the insect was hatched on non-native or native plants (0/1), and whether the insect grew to maturity on non-native or native plants (0/1). The study was carried out using 8 tanks (big aquariums), each of which were subdivided into four sections. The subdivision was done with a fine mesh that let water through but not insects. The tanks were planted with equal amounts of the non-native plant in the spring, whereupon the insects were introduced. Although uniformity between tanks was attempted, some tank to tank variation was expected due to differences in light and temperature. The biomass of the non-native plants in each section was measured in the fall. The data listing below shows the layout of the tanks.

		Insects on native plants (0/1)			
Tank	Section	Parents	Hatching	Growth	Biomass
1	1	0	0	0	10.4
1	2	1	1	0	17.5
1	3	1	0	1	22.2
1	4	0	1	1	27.7
2	1	1	0	0	4.8
2	2	0	1	0	8.9
2	3	0	0	1	6.8
2	4	1	1	1	17.6
3	1	0	0	0	16.8
3	2	1	1	0	19.6
3	3	0	0	1	16.4
3	4	1	1	1	35.6

(*Note:* the table continues on the next page)

(*Note:* the table is continued from the previous page)

Tank	Section	Insects on native plants (0/1) Parents	Hatching	Growth	Biomass
4	1	1	0	0	12.3
4	2	0	1	0	17.1
4	3	1	0	1	13.3
4	4	0	1	1	19.5
5	1	0	0	0	7.7
5	2	1	0	1	13.3
5	3	0	1	0	12.4
5	4	1	1	1	17.7
6	1	1	0	0	6.3
6	2	0	0	1	7.3
6	3	1	1	0	11.2
6	4	0	1	1	25.0
7	1	0	0	0	14.9
7	2	0	1	1	34.0
7	3	1	0	0	16.9
7	4	1	1	1	36.8
8	1	0	1	0	7.1
8	2	0	0	1	8.3
8	3	1	1	0	15.3
8	4	1	0	1	7.0

Use the description of the study and the information contained in the subsequent Minitab (version 21) listings (roughly matching Stata listings available upon request) to answer the following questions.

A) (*4 points*) Describe the experimental design in statistical terms, using standard descriptors such as factors, treatments, replication, blocks, balancedness, completeness, treatments, experimental units etc. You may also display the experimental design or data structure by a sketch/diagram. Discuss briefly whether the design matches one of the specific designs covered in the course. Use your characterization of the design to motivate a statistical model for the data (possibly, but not necessarily, the same model considered in the subsequent questions). Explain briefly the meaning of the different terms in your model.

B) (*3 points*) Use the attached output for part B) from statistical software to assess and interpret, by means of estimates and statistical tests, the different effects of the model. Draw conclusions about the question of interest that motivated the study: which (if any) of the experimental conditions should be recommended for reducing growth of the non-native plants?

C) (*2 point*) For these data it has also been suggested to carry out the analysis after square-root transforming the biomass values. The listings for part C) include the model output and the residual panel for analysis on square-root transformed scale. Based on this information, discuss (briefly) the choice of scale for the analysis and its implications for the results.

D) (*1 point*) The experimental design includes some features that distinguish it from the standard designs discussed in the course. Try to find some details in the listing of results that indicate these differences (that is, results you would not expect when carrying out a standard analysis).

Minitab prints for Question 1, part B):

INSECT.CSV

General Linear Model: biomass versus tank, parents, hatch, growth

Factor Information

Factor	Type	Levels	Values
tank	Fixed	8	1, 2, 3, 4, 5, 6, 7, 8
parents	Fixed	2	0, 1
hatch	Fixed	2	0, 1
growth	Fixed	2	0, 1

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
tank	7	1001.81	44.56%	685.806	97.972	8.84	0.000
parents	1	22.95	1.02%	22.950	22.950	2.07	0.168
hatch	1	597.72	26.59%	597.715	597.715	53.93	0.000
growth	1	373.33	16.61%	373.328	373.328	33.69	0.000
parents*hatch	1	8.28	0.37%	8.284	8.284	0.75	0.399
parents*growth	1	3.08	0.14%	3.082	3.082	0.28	0.605
hatch*growth	1	52.22	2.32%	52.215	52.215	4.71	0.044
parents*hatch*growth	1	0.48	0.02%	0.482	0.482	0.04	0.837
Error	17	188.41	8.38%	188.407	11.083		
Total	31	2248.27	100.00%				

Model Summary

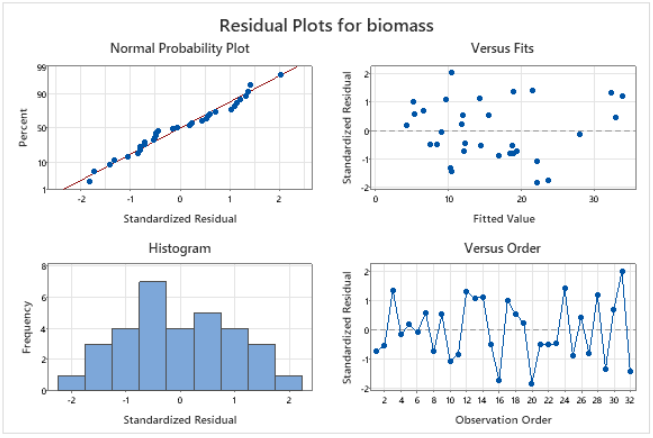
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
3.32908	91.62%	84.72%	667.574	70.31%	215.81	203.00

Fits and Diagnostics for Unusual Observations

Obs	biomass	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
31	15.30	10.37	2.28	(5.57, 15.18)	4.93	2.03	2.26	0.46875	0.24

Obs	DFITS
31	2.12555 R

R Large residual



Means

Term	Fitted Mean	SE Mean
parents		
0	15.019	0.832
1	16.712	0.832
hatch		
0	11.544	0.832
1	20.188	0.832
growth		
0	12.450	0.832
1	19.281	0.832
parents*hatch		
0 0	10.11	1.23
0 1	19.93	1.23
1 0	12.98	1.23
1 1	20.45	1.23
parents*growth		
0 0	11.96	1.23
0 1	18.08	1.23
1 0	12.94	1.23
1 1	20.49	1.23
hatch*growth		
0 0	9.60	1.23
0 1	13.48	1.23
1 0	15.30	1.23
1 1	25.08	1.23

INSECT.CSV

Comparisons for biomass

Fisher Pairwise Comparisons: parents*hatch

Fisher Individual Tests for Differences of Means

Difference of parents*hatch Levels	Difference of Means	SE of Difference	95% CI	T-Value	P-Value
(0 1) - (0 0)	9.82	1.80	(6.03, 13.61)	5.46	0.000
(1 0) - (0 0)	2.87	1.80	(-0.92, 6.66)	1.60	0.129
(1 1) - (0 0)	10.34	1.66	(6.83, 13.85)	6.21	0.000
(1 0) - (0 1)	-6.95	1.66	(-10.46, -3.44)	-4.18	0.001
(1 1) - (0 1)	0.52	1.80	(-3.27, 4.31)	0.29	0.776
(1 1) - (1 0)	7.47	1.80	(3.68, 11.26)	4.15	0.001

Simultaneous confidence level = 81.03%

Fisher Pairwise Comparisons: parents*growth

Fisher Individual Tests for Differences of Means

Difference of parents*growth Levels	Difference of Means	SE of Difference	95% CI	T-Value	P-Value
(0 1) - (0 0)	6.11	1.80	(2.32, 9.91)	3.40	0.003
(1 0) - (0 0)	0.98	1.80	(-2.82, 4.77)	0.54	0.594
(1 1) - (0 0)	8.53	1.66	(5.01, 12.04)	5.12	0.000
(1 0) - (0 1)	-5.14	1.66	(-8.65, -1.63)	-3.09	0.007
(1 1) - (0 1)	2.41	1.80	(-1.38, 6.20)	1.34	0.198
(1 1) - (1 0)	7.55	1.80	(3.75, 11.34)	4.20	0.001

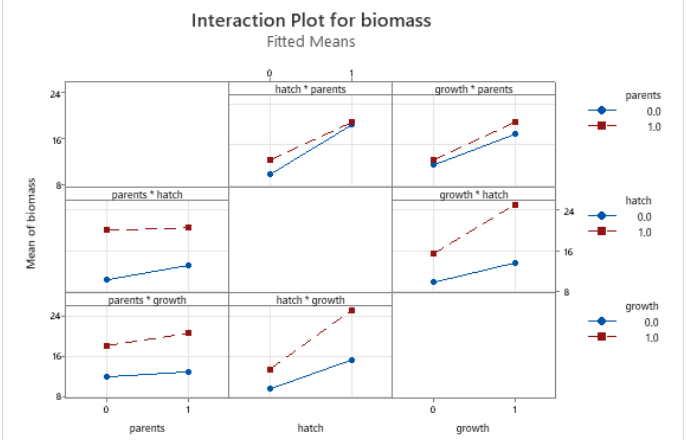
Simultaneous confidence level = 81.03%

Fisher Pairwise Comparisons: hatch*growth

Fisher Individual Tests for Differences of Means

Difference of hatch*growth Levels	Difference of Means	SE of Difference	95% CI	T-Value	P-Value
(0 1) - (0 0)	3.88	1.80	(0.09, 7.67)	2.16	0.045
(1 0) - (0 0)	5.69	1.80	(1.90, 9.49)	3.17	0.006
(1 1) - (0 0)	15.48	1.66	(11.96, 18.99)	9.30	0.000
(1 0) - (0 1)	1.81	1.66	(-1.70, 5.32)	1.09	0.291
(1 1) - (0 1)	11.59	1.80	(7.80, 15.39)	6.45	0.000
(1 1) - (1 0)	9.78	1.80	(5.99, 13.57)	5.44	0.000

Simultaneous confidence level = 81.03%



Minitab prints for Question 1, part C):

INSECT.CSV

General Linear Model: rootbio versus tank, parents, hatch, growth

Factor Information

Factor	Type	Levels	Values
tank	Fixed	8	1, 2, 3, 4, 5, 6, 7, 8
parents	Fixed	2	0, 1
hatch	Fixed	2	0, 1
growth	Fixed	2	0, 1

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
tank	7	15.3510	46.93%	10.5357	1.50510	9.75	0.000
parents	1	0.3323	1.02%	0.3323	0.33231	2.15	0.161
hatch	1	8.9774	27.45%	8.9774	8.97735	58.16	0.000
growth	1	4.7941	14.66%	4.7941	4.79411	31.06	0.000
parents*hatch	1	0.0522	0.16%	0.0522	0.05217	0.34	0.569
parents*growth	1	0.0280	0.09%	0.0280	0.02802	0.18	0.675
hatch*growth	1	0.4796	1.47%	0.4796	0.47961	3.11	0.096
parents*hatch*growth	1	0.0687	0.21%	0.0687	0.06866	0.44	0.514
Error	17	2.6242	8.02%	2.6242	0.15437		
Total	31	32.7075	100.00%				

Model Summary

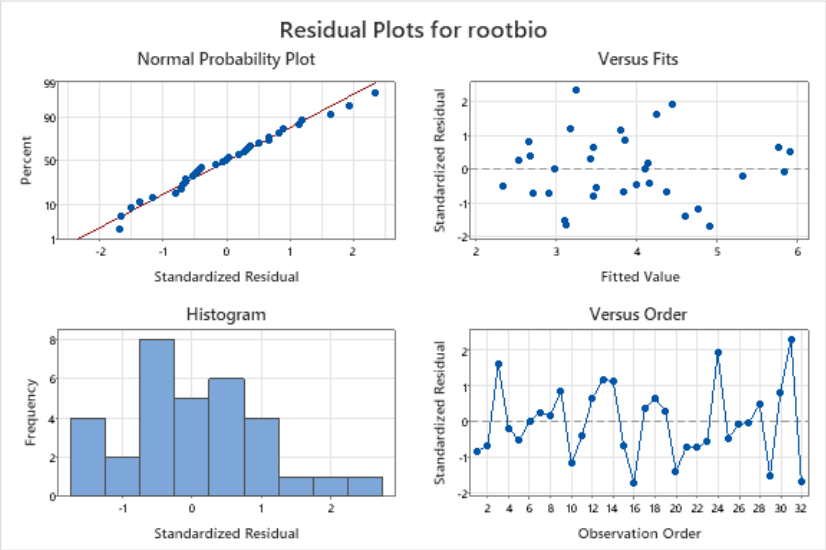
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.392895	91.98%	85.37%	9.29832	71.57%	79.05	66.23

Fits and Diagnostics for Unusual Observations

Obs	rootbio	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
31	3.912	3.243	0.269	(2.675, 3.810)	0.669	2.33	2.75	0.46875	0.32

Obs	DFITS	
31	2.58143	R

R Large residual



Question 2.

A methodology called functional Magnetic Resonance Imaging (fMRI) is used to determine the amount of the brain that is "activated" (in use) during certain activities. A study is planned with 12 right-handed subjects. Each subject will conduct two trials while positioned inside the imaging equipment. In each trial, the subject will, on a visual signal, perform an action (tapping of fingers in a certain order) using either the left or the right hand, depending on the signal. The measured response is the number of pixels in the image of the brain obtained by fMRI that are activated. The signals obtained from different subjects may vary considerably, and there may be a consistent difference between the first trial and the second trial. Six subjects are chosen at random to be measured first for the left hand and afterwards for the right hand, and the other six subjects are measured in right-left order. Thus, responses are obtained for each subject under both right- and left-hand tapping, and the interest is in comparing those.

A) (*3 points*) Outline the data to be obtained in the study: explain what is the experimental unit, give the number of observations and the important variables with their possible values. Describe the statistical design; include a diagram if the data structure is hierarchical.

B) (*3 points*) Suggest a statistical model to analyse the data to be obtained in the study, and explain the model's parameters. If several obvious modelling possibilities exist, you should motivate your choice. Explain briefly how you would fit the model in statistical software, without going through all steps of the statistical analysis.

C) (*2 points*) Explain how to carry out a calculation to determine whether the sample size is sufficient to detect an anticipated difference between left-hand and right-hand tapping responses. State explicitly what information you need for your calculation and what assumptions it relies on.

D) (*2 points*) There is interest in expanding the experiment to include also left-handed subjects, with the specific aim of comparing the differences between left-hand and right-hand responses between the two types of subjects (right-handed and left-handed subjects). Characterize the resulting experimental design, and explain how you would obtain statistical inference corresponding to the specific aim described above.

Question 3.

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic. A research team from the United States and Bangladesh measured the arsenic content in all (deep) wells used for drinking water in a rural area in Bangladesh. At the time, arsenic levels in drinking water below 0.5 in units of hundreds of micrograms per liter was considered as "safe". The wells were labelled as "safe" or "unsafe" based on this criterion, and people with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. A few years later, the researchers returned to find out who had switched wells. One question of interest was to better understand factors going into decisions in these households about changing their source of drinking water. For a total of 3020 households with unsafe arsenic levels at the first visit, the variables in the table below were recorded.

Variable	Description	Values
switch	whether the household switched to a new water source	0 (no) / 1 (yes)
arsenic	the initial arsenic level in the household's well	$0.51 - 9.65$
dist	the distance (in meters) to the closest known safe well	$0.4 - 339.5$
assoc	whether any members of the household were active in community organizations	0 (no) / 1 (yes)
educ	years of education of the household head	$0 - 17$

To further describe the data, a list of the first 10 observations in the dataset and some descriptive statistics of the variables are provided, from Stata (version 18) statistical software.

```
. list switch-educ in 1/10, sep(10)

     +-----------------------------------------+
     | switch   arsenic     dist   assoc   educ |
     |-----------------------------------------|
  1. |      1      2.36   16.826       0      0 |
  2. |      1       .71   47.322       0      0 |
  3. |      0      2.07   20.967       0     10 |
  4. |      1      1.15   21.486       0     12 |
  5. |      1       1.1   40.874       1     14 |
  6. |      1       3.9   69.518       1      9 |
  7. |      1      2.97   80.711       1      4 |
  8. |      1      3.24   55.146       0     10 |
  9. |      1      3.28   52.647       1      0 |
 10. |      1      2.52   75.072       1      0 |
     +-----------------------------------------+

. codebook, c

Variable     Obs Unique      Mean   Min      Max
-------------------------------------------------
switch      3020      2  .5751656     0        1
arsenic     3020    424   1.65693   .51     9.65
dist        3020   2938  48.33186  .387  339.531
assoc       3020      2  .4228477     0        1
educ        3020     18  4.828477     0       17
-------------------------------------------------
```

```
. tabstat switch, by(educ) stats(n sum mean)

Summary for variables: switch
Group variable: educ

    educ |         N       Sum      Mean
---------+----------------------------
       0 |       889       503  .5658043
       1 |         6         2  .3333333
       2 |        52        28  .5384615
       3 |       121        70  .5785124
       4 |       174        86  .4942529
       5 |       725       380  .5241379
       6 |       130        67  .5153846
       7 |       126        76  .6031746
       8 |       210       123  .5857143
       9 |        90        58  .6444444
      10 |       254       174  .6850394
      11 |        21        14  .6666667
      12 |       147       109  .7414966
      13 |         5         3        .6
      14 |        32        20      .625
      15 |        21        13  .6190476
      16 |        16        10      .625
      17 |         1         1          1
---------+----------------------------
   Total |      3020      1737  .5751656
-------------------------------------
```

```
. tabstat arsenic-educ, stats(mean sd min p25 p50 p75 max) columns(statistics)

    Variable |      Mean        SD       Min       p25       p50       p75       Max
-------------+--------------------------------------------------------------------------
     arsenic |   1.65693  1.107387       .51       .82       1.3       2.2       9.65
        dist |  48.33186  38.47867      .387   21.1165   36.7615    64.057   339.531
       assoc |  .4228477  .4940935         0         0         0         1         1
        educ |  4.828477  4.017317         0         0         5         8        17
--------------------------------------------------------------------------------------

. tab assoc switch

           |        switch        |
     assoc |         0          1 |     Total
-----------+----------------------+----------
         0 |       714      1,029 |     1,743
         1 |       569        708 |     1,277
-----------+----------------------+----------
     Total |     1,283      1,737 |     3,020

. pwcorr switch-educ

             |   switch  arsenic     dist    assoc     educ
-------------+---------------------------------------------
      switch |   1.0000
     arsenic |   0.1839   1.0000
        dist |  -0.1179   0.1781   1.0000
       assoc |  -0.0359  -0.0249  -0.0035   1.0000
        educ |   0.0764  -0.0296  -0.0267  -0.0314   1.0000
```

A) (*3 points*) Explain the statistical model used in the Stata listings presented for Question 3.A and interpret the parameter estimates in the model. Your interpretations should include, for each effect, both a quantitative statement of its size and a statement about its statistical significance. You may include a suitable graphical representation to aid the interpretation if you want (it is optional). Defer any discussion about the validity of the model to part B).

B) (*5 points*) Briefly review the assumptions behind the statistical model. Use the additional Stata listings and graphs provided to discuss any concerns you might have about the validity of the model and/or the results. Explain what information about model validity you obtain from each of the two additional models shown in the listings. Discuss also any further steps and procedures you would want to explore in order to assess the model assumptions.

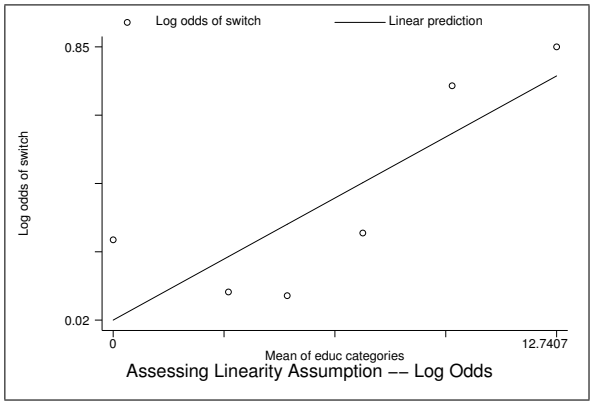
C) (*2 points*) Use the information provided in all the Stata listings to outline a strategy for developing a good model for understanding the decisions to switch or not switch from unsafe wells. Explain and motivate the different steps you would go through in developing the model, while explaining how you use the information from the analyses already carried out in the process.

Stata listings for Question 3.A and 3.B:

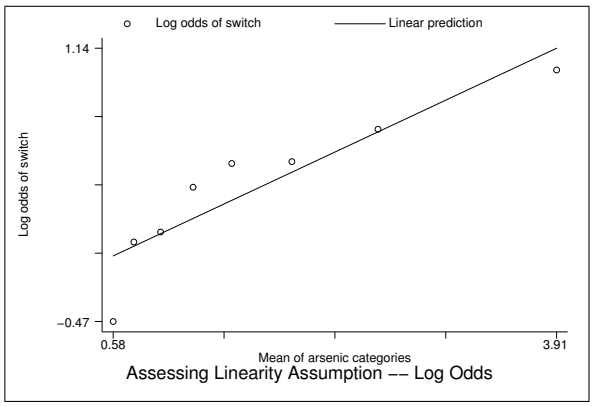
```
. logit switch arsenic dist assoc educ
...

Logistic regression                                     Number of obs =  3,020
                                                        LR chi2(4)    = 210.27
                                                        Prob > chi2   = 0.0000
Log likelihood = -1953.913                              Pseudo R2     = 0.0511

------------------------------------------------------------------------------
      switch | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     arsenic |   .4670216   .0416023    11.23   0.000     .3854825    .5485606
        dist |  -.0089611   .0010458    -8.57   0.000    -.0110108   -.0069114
       assoc |     -.1243   .0769661    -1.61   0.106    -.2751507    .0265507
        educ |   .0424466   .0095876     4.43   0.000     .0236552    .0612381
       _cons |  -.1567117   .0996009    -1.57   0.116    -.3519258    .0385025
------------------------------------------------------------------------------

. estat gof

Goodness-of-fit test after logistic model

      Number of observations =   3,020
Number of covariate patterns =   3,020
          Pearson chi2(3015) = 3048.39
                Prob > chi2 =  0.3311

. estat gof, g(10)
note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model


 Number of observations =  3,020
       Number of groups =     10
Hosmer-Lemeshow chi2(8) =  10.33
            Prob > chi2 = 0.2425
```

Additional Stata listings for Question 3.B and 3.C:

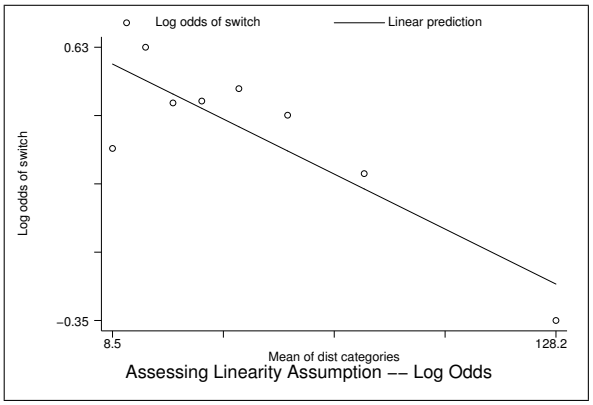
```
. lintrend switch educ, g(8) plot(log)

The proportion and log odds of switch by categories of educ
  (Note: 8 educ categories of equal sample size;
     Uses mean educ value for each category)
  +-------------------------------------------------------+
  |     educ   min   max    d   total  switch  logodds |
  |-------------------------------------------------------|
  |        0     0     0  503     889    0.57     0.26 |
  | 3.311615     1     4  186     353    0.53     0.11 |
  |        5     5     5  380     725    0.52     0.10 |
  | 7.171674     6     8  266     466    0.57     0.29 |
  | 9.738372     9    10  232     344    0.67     0.73 |
  | 12.74074    11    17  170     243    0.70     0.85 |
  +-------------------------------------------------------+
```



```
. lintrend switch arsenic, g(8) plot(log)

The proportion and log odds of switch by categories of arsenic
  (Note: 8 arsenic categories of equal sample size;
     Uses mean arsenic value for each category)
  +-------------------------------------------------------+
  | arsenic    min    max    d   total  switch  logodds |
  |-------------------------------------------------------|
  |    0.58    .51    .64   147    382    0.38    -0.47 |
  |    0.73    .65    .82   188    376    0.50     0.00 |
  |    0.93    .83   1.05   193    375    0.51     0.06 |
  |    1.18   1.06    1.3   221    381    0.58     0.32 |
  |    1.47   1.31   1.66   232    378    0.61     0.46 |
  |    1.92   1.67    2.2   233    378    0.62     0.47 |
  |    2.57   2.21   2.98   247    374    0.66     0.67 |
  |    3.91   2.99   9.65   276    376    0.73     1.02 |
  +-------------------------------------------------------+
```




```
. lintrend switch dist, g(8) plot(log)

The proportion and log odds of switch by categories of dist
  (Note: 8 dist categories of equal sample size;
     Uses mean dist value for each category)
  +-------------------------------------------------------+
  |  dist      min       max    d   total  switch  logodds |
  |-------------------------------------------------------|
  |   8.5     .387    13.573   214    378    0.57     0.27 |
  |  17.5   13.578    21.115   246    377    0.65     0.63 |
  |  24.9   21.118    28.654   229    378    0.61     0.43 |
  |  32.6   28.671     36.75   229    377    0.61     0.44 |
  |  42.6   36.773    48.605   233    377    0.62     0.48 |
  |  55.8   48.616    64.025   225    378    0.60     0.39 |
  |  76.5   64.089    92.176   205    377    0.54     0.18 |
  | 128.2    92.49   339.531   156    378    0.41    -0.35 |
  +-------------------------------------------------------+
```




```
. logit switch c.arsenic##c.arsenic c.dist##c.dist assoc c.educ##c.educ
...
Logistic regression                                     Number of obs =  3,020
                                                        LR chi2(7)    = 239.91
                                                        Prob > chi2   = 0.0000
Log likelihood = -1939.0971                             Pseudo R2     = 0.0583
------------------------------------------------------------------------------
              switch | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
---------------------+--------------------------------------------------------------
             arsenic |   .8740755    .101527     8.61   0.000     .6750863    1.073065
                     |
c.arsenic#c.arsenic |  -.0851448   .0183621    -4.64   0.000    -.1211339   -.0491558
                     |
                dist |  -.0070431   .0029534    -2.38   0.017    -.0128316   -.0012546
                     |
      c.dist#c.dist |  -.0000162    .000018    -0.90   0.370    -.0000515    .0000192
                     |
               assoc |  -.1023913   .0775611    -1.32   0.187    -.2544083    .0496257
                educ |  -.0353104   .0266561    -1.32   0.185    -.0875555    .0169346
                     |
      c.educ#c.educ |   .0070969   .0022753     3.12   0.002     .0026374    .0115564
                     |
               _cons |  -.4416385   .1406246    -3.14   0.002    -.7172576   -.1660193
------------------------------------------------------------------------------
```

```
. logit switch c.arsenic##c.dist c.educ##assoc c.arsenic#c.educ c.dist#c.educ
...

Logistic regression                                     Number of obs =  3,020
                                                        LR chi2(8)    = 230.75
                                                        Prob > chi2   = 0.0000
Log likelihood = -1943.6755                             Pseudo R2     = 0.0560

------------------------------------------------------------------------------
      switch | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     arsenic |   .4651448   .0864899     5.38   0.000     .2956277    .6346618
        dist |   -.011044   .0026016    -4.25   0.000     -.016143    -.005945
             |
c.arsenic#c.dist |   -.001175   .0010356    -1.13   0.257    -.0032047    .0008547
             |
        educ |  -.0139514   .0214924    -0.65   0.516    -.0560758     .028173
     1.assoc |  -.0175597   .1208133    -0.15   0.884    -.2543495    .2192301
             |
 assoc#c.educ |
          1  |  -.0242307   .0197807    -1.22   0.221    -.0630002    .0145388
             |
c.arsenic#c.educ |   .0174646   .0110035     1.59   0.112    -.0041019    .0390311
             |
c.dist#c.educ |   .0008284   .0002675     3.10   0.002     .0003041    .0013526
             |
       _cons |   .0037881   .1643555     0.02   0.982    -.3183427    .3259189
------------------------------------------------------------------------------
```