MSc Statistical Science

Exam Linear Models, Generalized Linear Models and Linear Algebra (4433LAGLMT)

January 11, 2021; 10.15 - 13.15 h

Answers

Please write your name and student card number on every page.

Please write and sign the following statement on your exam at the top of the first page of solutions, before starting the exam:

"This exam will be solely undertaken by myself, without any assistance from others and without use of sources other than those explicitly allowed by the lecturer. Moreover, I have read and signed the declaration of integrity."

This is an open book exam. You are allowed to consult the following course material:

- clean hand-outs, as available on Brightspace; it is ok if the hand-outs contain a few notes, but worked-out exercises or exams are NOT allowed;
- Fox book and appendix, Faraway linear regression text (or book), Faraway Extending linear models;
- self-made summary of course material on single A4 (possibly two-sided).

This material may be consulted on paper or you may look it up in Brightspace.
No old exams with answers are allowed.
Use an ordinary hand calculator for calculations.
Do not give bare answers in case of calculations (unless this is explicitly mentioned), but show how you arrived at a solution. Bare answers in case of calculations will not be rewarded.

In total 90 points can be earned, divided over 3 main questions (with 52, 32 and 6 points). With x the number of scored points, the end result of the exam is $(x + 10)/10$ rounded to halfs. The final score for the course is $\frac{2}{3}$exam $+ \frac{1}{3}$case study report, provided that the result for the written exam is at least 5 and for the case study at least 6. Otherwise the score is the lowest of the two.

If you are working on a question that uses results from an earlier question, but you were not able to answer that earlier question, then either proceed using a hypothetical answer to the earlier question or describe in words how you would proceed.

For hypothesis tests use $\alpha = 0.05$ unless stated otherwise.

Please formulate answers in a compact and concise way.

Indication of available time per point: as the exam takes 180 minutes on a total of 90 points, on average 2 minutes per point are available. Realize that you will not have time to look up many details!

This exam is composed by dr. G. Gort.
Second reader of the exam is dr. E.J. Bakker.

Question 1 (52) Plastic pollution

In agriculture soils are often covered with plastic films to prevent water loss and to suppress the growing of weeds. However, plastic residues are afterwards found in the soils, which may have serious consequences. In an experimental study low-density polyethylene (LDPE) particles (of a specific size range) were mixed with soil at three difference weight concentrations w: 0% (no LDPE), 0.5% and 1%. The three concentrations, each with two replicates, were randomized over six containers. After a month the water repellency y of the contaminated soils was measured (using a measure called the water drop penetration time; higher values mean higher soil water repellency).
Below the results are shown:

container	w	y
1	0	4
2	0	5
3	0.5	6
4	0.5	6.5
5	1	6
6	1	7

For now we treat the concentration as a *factor*, so as a grouping variable. We study the relationship between response variable y and group variable w using a one-way ANOVA model.

1a1 (3) Write down this one-way ANOVA model using effects model notation. Pay attention both to the systematic part and to the random part of the model (error assumptions).

model with assumptions:
$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ with $i = 1, 2, 3$ for the three concentrations, $j = 1, 2$ for the two containers per concentration, y_{ij} the water repellency of container j of concentration i.
For the random part ϵ_i we assume that ϵ_i are independent, normally distributed with mean 0 and common variance σ^2_ϵ.

1.5pt for model description, 1.5pt for random part (0.5pt per assumption).

1a2 (4) Complete the ANOVA table as shown below for this example:

Source	*df*	*SS*	*MS*	*F*
Between groups	...	...	...	...
Within groups	...	...	...	
Corrected total	5	5.875		

Calculate $SS(within)$ first and $SS(between)$ by subtraction from corrected total SS:

$group_i$	y_{ij}	$\bar{y}_{i.}$	$y_{ij} - \bar{y}_{i.}$	$(y_{ij} - \bar{y}_{i.})^2$
1	4	4.5	−0.5	0.25
1	5	4.5	0.5	0.25
2	6	6.25	0.25	0.0625
2	6.5	6.25	−0.25	0.0625
3	6	6.5	−0.5	0.25
3	7	6.5	0.5	0.25
	34.5	34.5	0	1.125

So, $SS(within) = 1.125$, $SS(between) = SST - SS(within) = 5.875 - 1.125 = 4.75$.

Source	*df*	*SS*	*MS*	*F*
Between groups	$3 - 1 = 2$	4.75	2.375	6.333
Within groups	$6 - 3 = 3$	1.125	0.375	
Corrected total	$6 - 1 = 5$	5.875		

ANOVA table for regression instead of ANOVA: -2pt; $SS(within)$ (or $SS(between)$): 2pt; df : 1pt; remainder: 1pt.

1a3 (2) Give the estimate of the error standard deviation and its interpretation.

$\hat{\sigma}_\epsilon = \sqrt{0.375} = 0.6124$ (1 pt)

Interpretation: (type of) average deviation of observations from their predicted values (obtained using the fitted model). (1 pt)

1a4 (2) Which null hypothesis is tested with the F-statistic in the ANOVA table? Please formulate it using parameters. Judge roughly the significance of the F-test here.

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ (1 pt)

$F = 6.333$, so quite a bit larger than 1. P will be small, but unclear whether it is smaller than 0.05, given the low df of numerator and denominator. [Actual $P = 0.084$. (1 pt)

1a5 (2) Give R^2. What is it telling?

$R = SS(Between)/TSS = 4.75/5.875 = 0.81$. (1 pt)

It is the fraction variation, explained by the model, 81% here, i.e. quite high. (1 pt)

The model of question 1a1 is overparameterized.

1b1 (2) Explain what overparameterization is in terms of this example, and the default way it is solved in R.

Overparameterization means that there are more parameters in the model than are identifiable: the model has 4 parameters, but there are only 3 groups (plastic concentrations). (1pt)

R solves the problem using the cornerstone restriction: R makes the parameter for the first level (α_1) equal to 0. (1pt)

Below you find some output of the fitted one-way ANOVA model in R.

```
> wf <- factor(w)
> lmo1 <- lm(y ~ wf)
> coef(summary(lmo1))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.50     0.4330  10.392 0.001901
wf0.5           1.75     0.6124   2.858 0.064697
wf1             2.00     0.6124   3.266 0.046919
```

1b2 (2) Give the interpretation (if any) of the estimated coefficients labeled `(Intercept)` and `wf0.5`.

Parameter `(Intercept)` $= \mu$ is the expected water repellency at concentration 0 (reference level). (1pt)

Parameter `wf0.5` $= \alpha_2$ is the difference in expected water repellency at concentration 0.5 (second level) and concentration 0 (reference level). (1pt)

The one-way ANOVA model is an example of a linear model. So, it can be written in matrix notation as $y = X\beta + \epsilon$.

~~**1b3** (2) Give the model matrix X that R used to produce the output given above, and show how from~~

X the matrix $X'X = \begin{pmatrix} 6 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}$, containing sums of squares and cross-products, is obtained.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad (1\text{pt})$$

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} \text{ by matrix multiplication. (1pt)}$$

For least-squares estimation in linear models the inverse of $X'X$ plays an important role. To obtain $X'X^{-1}$ Gaussian elimination may be used.

1b4 (3) Show how Gaussian elimination works by sweeping the first column of $X'X$. (*Only* sweep the first column; sweeping all columns will be too time costly.)

$$X'X|I_3 = \left(\begin{array}{ccc|ccc} 6 & 2 & 2 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 2 & 0 & 2 & 0 & 0 & 1 \end{array}\right) \xrightarrow{(r1/6, r2/2, r3/2)} \left(\begin{array}{ccc|ccc} 1 & 1/3 & 1/3 & 1/6 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1/2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1/2 \end{array}\right)$$

$$\xrightarrow{(r2-r1, r3-r1)} \left(\begin{array}{ccc|ccc} 1 & 1/3 & 1/3 & 1/6 & 0 & 0 \\ 0 & 2/3 & -1/3 & -1/6 & 1/2 & 0 \\ 0 & -1/3 & 2/3 & -1/6 & 0 & 1/2 \end{array}\right)$$

The resulting inverse is $X'X^{-1} = \begin{pmatrix} 0.5 & -0.5 & -0.5 \\ -0.5 & 1.0 & 0.5 \\ -0.5 & 0.5 & 1.0 \end{pmatrix}$

1b5 (3) Show how the standard error of the coefficient labeled `(Intercept)` (=0.433) is obtained, using (amongst others) $X'X^{-1}$.

$\hat{Var}(b) = \hat{\sigma}_\epsilon X'X^{-1}$.
Parameter `(Intercept)` is $\hat{\mu}$, the first element of the coefficient vector b.
So, $\hat{var}(b[1]) = \hat{\sigma}^2_\epsilon X'X^{-1}[1,1] = 0.375 \times 0.5 = 0.1875$,
and $\hat{se}(b[1]) = \sqrt{0.1875} = 0.433$.

Until now the concentration of plastic was treated as a factor (with three levels 0%, 0.5% and 1%). But maybe the model can be simplified by treating concentration as a quantitative regressor, so that a simple linear regression model $y = \beta_0 + \beta_1 w + \epsilon$ would suffice. An F-test for goodness-of-fit may be used to check whether this is reasonable.

1c (4) Use the R-output below to (i) estimate the pure error variance, and (ii) calculate the F-test statistic for goodness of fit.

```
> lmo.anova <- lm(y ~ wf)
> lmo.regr  <- lm(y ~ w)
> deviance(lmo.anova); df.residual(lmo.anova)
```

```
[1] 1.125
```

```
[1] 3
```

```
> deviance(lmo.regr); df.residual(lmo.regr)
```

```
[1] 1.875
```

```
[1] 4
```

$\hat{\sigma}^2_{PE} = 1.125/3 = 0.375$ (1.5pt)

$$F_{GOF} = \frac{(SSE_{regression} - SSE_{anova})/(4-3)}{SSE_{anova}/3} = \frac{(1.875 - 1.125)/1}{1.125/3} = \frac{0.75}{0.375} = 2 \text{ (2.5pt)}$$

From now on we treat the concentration as a quantitative regressor.

The data shown earlier was part of a larger study, in which there was an extra concentration of plastic (not only 0%, 0.5% and 1%, but also 2%), and there were two types of plastic: not only LDPE, but also biodegradable plastic. Ten containers were used as control (concentration 0%), and five containers were used for each type of plastic (LDPE and bio) at each of the three concentrations (0.5%, 1%, 2%). In total there were $n = 40$ observations.
We define variable y_i $(i = 1, \ldots, 40)$ as the water repellency of the soil of container i, xL_i as the LDPE concentration of the soil of container i (0 for the control containers, and 0 for containers with biodegradable plastic), and xB_i as the concentration of biodegradable plastic of container i (0 for the control containers, and 0 for containers with LDPE plastic).
The following multiple regression model is fitted:

$$y_i = \alpha + \beta \cdot xL_i + \gamma \cdot xB_i + \epsilon_i$$

Below, the first two and final two observations of dataframe P are show, together with results on the fitted multiple linear regression model.

```
> head(P, n=3)
```

```
      y xL xB
1 3.440  0  0
2 3.770  0  0
3 5.559  0  0
```

```
> tail(P, n=3)
```

```
       y xL xB
38 6.438  0  2
39 6.194  0  2
40 6.120  0  2
```

```
> lmo2 <- lm(y ~ xL + xB, data=P)
> summary(lmo2)
```

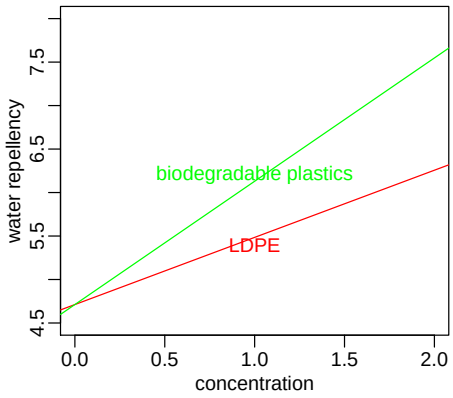
```
Call:
lm(formula = y ~ xL + xB, data = P)

Residuals:
   Min     1Q Median     3Q    Max
-1.977 -0.978 -0.370  0.938  2.803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.712      0.316   14.92  < 2e-16
xL             0.772      0.328    2.35  0.02399
xB             1.418      0.328    4.32  0.00011

Residual standard error: 1.29 on 37 degrees of freedom
Multiple R-squared:  0.34,        Adjusted R-squared:  0.304
F-statistic: 9.53 on 2 and 37 DF,  p-value: 0.000458
```

```
> vcov(lmo2)
```

```
            (Intercept)       xL       xB
(Intercept)     0.09973 -0.06649 -0.06649
```

```
xL            -0.06649  0.10765  0.04432
xB            -0.06649  0.04432  0.10765
```

1d1 (3) Make a sketch of the resulting fitted model, plotting the concentration on the x-axis, the water repellency on the y-axis, using separate lines for LDPE and biodegradable plastics.



no axis labels/scales: -0.5pt; if no scales but slopes are mentioned: -0pt

1d2 (1) Which null hypothesis H_0 is tested with the omnibus F-test shown in the `summary` output above? Formulate it using parameters from the multiple regression model.

$H_0 : \beta = \gamma = 0$

We wonder whether LDPE and biodegradable plastics have the same effect on water repellency. So, we want to compare β with γ.

1d3 (1) Estimate the difference in slopes for LDPE and biodegradable plastics.

$\hat{\beta} - \hat{\gamma} = 0.772 - 1.418 = -0.646$

To test whether the slopes are different, we need the standard error of $\hat{\beta} - \hat{\gamma}$.

1d4 (3) Estimate the standard error of $\hat{\beta} - \hat{\gamma}$.

$\hat{se}(\hat{\beta} - \hat{\gamma}) = \sqrt{\hat{var}(\hat{\beta} - \hat{\gamma})} = \sqrt{\hat{var}(\hat{\beta}) + \hat{var}(\hat{\gamma}) - 2\hat{cov}(\hat{\beta}, \hat{\gamma})} = \sqrt{0.10765 + 0.10765 - 2 * 0.04432} = \sqrt{0.1267} = 0.356$.
correct formula var of diff: 1 pt; var instead of se: -0.5pt

1d5 (4) Test with a t-test the null hypothesis that the slope difference is zero. Mention: 1) H_0 and H_a using parameter(s); 2) outcome of test statistic; 3) distribution of test statistic under H_0, including degrees of freedom (if any); 4) rejection region of the test (roughly) and conclusion.

1) $H_0 : \beta - \gamma = 0$ versus $H_a : \beta - \gamma \neq 0$
2) $t = ((\hat{\beta} - \hat{\gamma}) - 0)/\hat{se}(\hat{\beta} - \hat{\gamma}) = -0.646/0.356 = -1.815$
3) If H_0 is true $t \sim t_{37}$-distribution
4) Rejection region is roughly $(-\infty, -2] \cup [2, \infty)$; t is not in rejection region, so do not reject H_0.
1 pt for every part

1d6 (2) Suppose we want to use an F-test to test the same null hypothesis as in 1d5 by comparing a full and reduced model, what would be the full model and what would be the reduced model?

FM: $y_i = \alpha + \beta \cdot xL_i + \gamma \cdot xB_i + \epsilon_i$ (as given earlier) (0.5pt)
RM: $y_i = \alpha + \delta \cdot (xL_i + xB_i) + \epsilon_i$ (1.5pt)

It could be that the straight line relationships that we assume here are too simple.

1e1 (2) What type of residual plot is specifically meant to check this in a multiple regression setting?

Component-plus-residual plot (partial residual plot)
Alternative (but less usable for checking curvature): added-variable plot (partial regression plot): -0.5pt Plot of residuals vs regressor or vs predicted values: -1.5pt

To allow for possible curvature in the relationships, we may add quadratic terms into the model. Below we try this in this example. We show residual sums of squares of this model, but also for some other models.

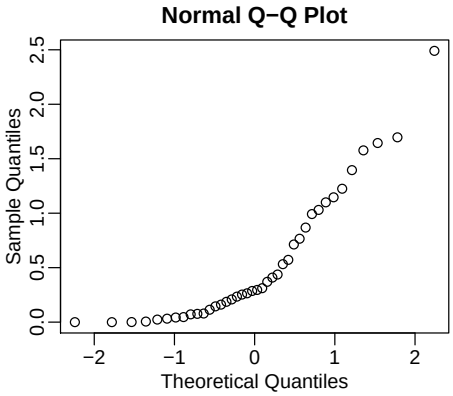
```
> P$xLsq <- P$xL^2
> P$xBsq <- P$xB^2
> lmo3 <- lm(y ~ xL + xB + xLsq + xBsq, data=P)
> deviance(lmo3)
[1] 34.64
> lmo4 <- lm(y ~ xLsq + xBsq, data=P)
> deviance(lmo4)
[1] 77.18
> lmo5 <- lm(y ~ 1, data=P)
> deviance(lmo5)
[1] 93.2
> deviance(lmo2) # lmo2 refers to the old model y = a + b * xL + g * xB + e
[1] 61.5
```

1e2 (3) Use the output above to test with an F-test whether the extension of the model with quadratic terms is needed. Mention (i) null hypothesis in words, (ii) definition and outcome of the F-test statistic (including degrees of freedom), (iii) (rough) judgment of significance and conclusion.

i) H_0: coefficients for quadratic terms are zero, or H_0: no curvature in relationships (0.5pt)
ii) $F = \dfrac{(SSE(RM) - SSE(FM))/2}{SSE(FM)/35}$ with FM the model with linear and quadratic terms (5 parameters) and RM is model with linear terms alone (3 parameters); outcome: $F = \dfrac{(61.50 - 34.64)/2}{34.64/35} = 13.57$ (2pt)
iii) F much larger than 1, outcome far in right tail of $F_{2,35}$ distribution, so P-value small. Conclusion: curvature in relationship(s) (0.5pt)

After the fitting of a linear model the assumptions should be checked.

1f1 (2) Suppose that there is a serious violation of the normality assumption: the residuals are heavily skewed to the right. Sketch how a normal QQ-plot would look like (plotting the residuals on the y-axis and theoretical quantiles on the x-axis, as R does).



The plot should show points that curve ”upwards”.

If scatter or points not monotonously ”increasing”: -0.5pt

1f2 (2) ~~What type of model would you consider if you find heteroscedasticity, more precisely a variance that is proportional to the square of the mean?~~

~~Generalized linear model with gamma distribution~~

~~Linear model for log transformed response y (-1pt)~~

~~If simply ”GLM”: 0 pt~~

Question 2 (32) Winter wheat

Winter wheat is a grain, planted in autumn, which needs cold to produce good yields. However, not all winter wheat cultivars are able to survive the extreme cold that occurs in some environments.
In an experiment three varieties of winter wheat are tested for survival under freezing conditions using two treatments: with or without a hardening treatment prior to the freeze. The hardening of the plants may prepare them for the freeze, and may result in higher survival fractions.
For each variety twelve pots, each containing $n = 5$ plants, are prepared. Per variety six randomly chosen pots receive the hardening treatment, and the remaining six do not. Next, all $3 \times 2 \times 6 = 36$ pots (each containing 5 plants) are placed at random locations in a freezer. After the freezing test, the pots are moved to a greenhouse, and after a while the number y of surviving plants per pot is counted.
Below the first few observations of dataframe `ww` are shown. Factor T represents the hardening treatment with levels N = not hardened and H = hardened. Factor V represents the variety of winter wheat with levels 1,2,3. Because of some mishap, two pots got lost, so in total there are 34 remaining pots.

```
> head(ww, n=7)
```

```
  V T y n
1 1 N 0 5
2 1 N 0 5
3 1 N 1 5
4 1 N 2 5
5 1 N 1 5
6 1 N 1 5
7 1 H 2 5
```

E.g. on the first row of data shown above you see the results of a pot containing 5 plants ($n = 5$) of variety 1 ($V = 1$), which did not receive the hardening treatment ($T =$ "N"). No plant survived ($y = 0$).

The fraction surviving plants is analyzed using logistic regression. The systematic part of the model comprises main effects of the factors variety and hardening, and their interaction.

2a (3) Write down the model in mathematical notation, paying attention to the three components of a generalized linear model (g.l.m.). (Do NOT fill in estimates of parameters yet.)

Three components of a generalized linear model:
part 1:
Random part of model $y_{ijk} \sim Bin(5, p_{ijk})$, independent, with y_{ijk} the number of surviving plants out of 5 for variety i, hardening treatment j and pot k, p_{ijk} is the probability that an individual plant in that group survives ($i = 1, 2, 3$ for variety, $j = 1, 2$ for hardening treatments `N` and `H`, $k = 1, \ldots, 6$ for replicates) (1.5pt)

part 2:
Systematic part of model: linear predictor $\eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ with α_i main effect parameters for variety, β_j main effect parameters for treatment and γ_{ij} for interaction (1pt)

part 3:
Link function: $logit(p_{ijk}) = \eta_{ijk}$ (0.5pt)

Below the g.l.m. is fitted to the data, and some results are shown.

```
> glmo <- glm(cbind(y, n-y) ~ V + T + V:T, family=binomial(link=logit), data=ww)
> summary(glmo)
```

```
Call:
glm(formula = cbind(y, n - y) ~ V + T + V:T, family = binomial(link = logit),
    data = ww)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.137  -0.939   0.154   0.757   1.479
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.609      0.490   -3.29    0.001
V2            -0.262      0.727   -0.36    0.718
V3             0.223      0.700    0.32    0.750
TH             1.529      0.633    2.42    0.016
V2:TH         -0.204      0.912   -0.22    0.823
V3:TH          2.496      1.089    2.29    0.022

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 90.364  on 33  degrees of freedom
Residual deviance: 28.551  on 28  degrees of freedom
AIC: 85.69

Number of Fisher Scoring iterations: 5
```

First check whether there are signs of binomial overdispersion.

2b (2) Explain what binomial overdispersion is and why it is important to take into consideration. Is there a problem in this case (motivate)?

Binomial overdispersion means that there is more variability on the fractions than can be explained by the binomial distribution. If overdispersion is not handled, R works with the binomial variation, which is too small, leading to standard errors which are too small. (1pt)
Problem here? Check the residual deviance: 28.55 on 28 df, ratio is almost 1; hence there is no overdispersion. (1pt)

Is there any effect at all of variety and/or hardening treatment?

2c1 (3) Use a Likelihood Ratio test to test the null hypothesis that there is no effect at all of hardening and variety on the probability of a plant to die. Mention: (i) definition and outcome of the test statistic; (ii) distribution of the test statistic underH_0; (iii) conclusion (roughly) with motivation.

(i) LRT test statistic TS = Null deviance − Residual deviance; outcome $TS = 90.364 - 28.551 = 61.81$ (1pt)
(ii) distribution of TS under H_0 with d.f.: $TS \sim \chi^2$ with 5 (=28-23) d.f. (1pt)
(iii) conclusion: outcome is very far in the right tail of the χ^2_5 distribution (recall that $E(\chi^2_5) = 5$ and $var(\chi^2_5) = 10$). Hence, the P-value is very small, and the null hypothesis of no effect of variety and/or hardening rejected (1pt)

Next, we continue by checking interaction.

2c2 (2) Describe in practical terms what interaction between the hardening treatment and variety means in this example about plant survival.

Interaction between hardening treatment and variety means that the difference in survival (on the logit-scale) between the hardening and control treatment is not the same for the three varieties.

Below an Analysis of Deviance table is produced with the `anova` function.

```
> anova(glmo, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit
```

```
Response: cbind(y, n - y)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    33       90.4
V     2     16.6        31       73.7  0.00024
T     1     37.0        30       36.7  1.2e-09
V:T   2      8.2        28       28.6  0.01668
```

2c3 (3) What type of hypothesis tests, based upon deviances, are shown here? Explain whether the given test results for the interaction `V:T`, but also for the main effects `V` and `T` can be used directly.

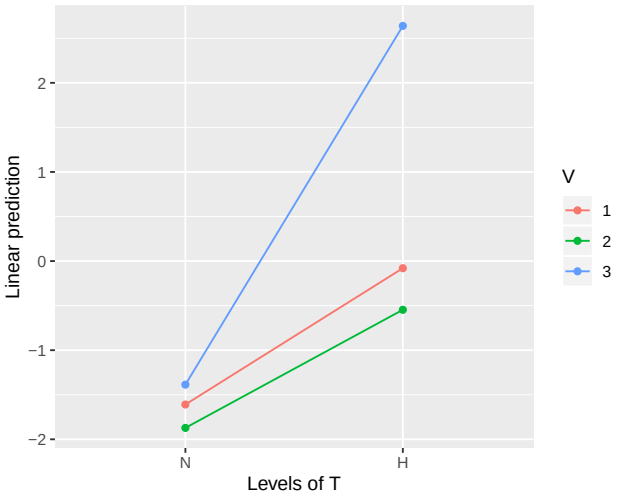
Shown tests are likelihood ratio tests, employing type I (sequential) deviance differences (1pt)
The test for interaction can be directly used, as the interaction is the last in the sequence, checking the effect of the interaction after both main effects (1pt)
The test for main effect of variety cannot be used, because it is the first in the sequence; hence no correction for treatment (or interaction); the test for main effect of hardening may be used, but realize that is a type II comparison of deviances: interaction is not (yet) in the model (1pt)

2c4 (3) What do you conclude with respect to the interaction? Mention (i) the outcome of the test statistic (TS), (ii) the distribution of TS under H_0 with df, (iii) the P-value and conclusion.

(i) outcome $TS = 8.2$ (1pt)
(ii) distribution of TS under H_0 with d.f.: $TS \sim \chi^2$ with 2 d.f. (1pt)
(iii) $P = 0.017 < 0.05$,so reject H_0: the effect of the hardening treatment is not the same for the different varieties(1pt)
If conclusion not in words: -0pt

Below a profile plot (on the linear predictor scale) is shown.

```
> emmip(glmo, V ~ T)
```



2c5 (2) What do you conclude based upon this profile plot?

The profile plot shows interaction, as lines for varieties do not run parallel. The difference in survival (in linear predictor value, i.e. on logit scale) between H and N for variety 1 is higher than for varieties 2 and 3. (2 pt)

Testing for main effects in presence of interaction is a topic that makes some statisticians frown.

2d1 (2) Why is that, and to which modeling principle does it refer?

Presence of interaction means that the effect of a factor cannot be seen separately from the level of the second factor. In other words, looking at main effects makes less sense. (1 pt) Principle of marginality: main effects are marginal to interactions; usually do not fit models excluding marginal terms (main effects) if higher order terms (interactions) are still in the model. (1pt) [Explanation of principle of marginality not needed]

2d2 (2) If you would choose to test for the main effect of the hardening treatment here, which type of model comparison would you use: type II or type III? Motivate your answer.

Type III comparisons, because interaction is significant. With type III main effect tests we average over the levels of the second factor, allowing interactions, whereas with type II main effect tests the interaction is removed from the modedl first. (2 pt) Note that some statisticians would always use type II comparisons, even if interactions are found important. If this answer is given with motivation, points may be given too.

Given the results so far, it would make sense to judge the effect of the hardening treatment separately for the three varieties.

Below, the maximum likelihood estimates of the coefficients are shown again.

```
> coef(summary(glmo))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6094     0.4899 -3.2853 0.001019
V2           -0.2624     0.7270 -0.3609 0.718167
V3            0.2231     0.7000  0.3188 0.749896
TH            1.5294     0.6327  2.4174 0.015631
V2:TH        -0.2041     0.9123 -0.2238 0.822940
V3:TH         2.4960     1.0890  2.2919 0.021910
```

2e (2) Which group does the parameter labeled `(Intercept)` refer to? Estimate the probability of a plant to survive in that group.

`(Intercept)` refers to the reference group (for both factors), i.e. variety 1 in the control treatment; it is the mean survival (on logit scale) $\hat{\mu}$ estimated as -1.609. (1pt) As $logit(p) = \mu$, we have $p = \frac{e^{\mu}}{1+e^{\mu}}$; hence $\hat{p} = \frac{e^{-1.609}}{1+e^{-1.609}} = 0.17$. (1 pt)

In the table with parameter estimates you see a parameter labeled `TH` with estimate 1.529.

2f1 (2) Show that exp(`TH`) represents the Odds Ratio (OR) for hardening versus not-hardening for variety 1.

`TH` is parameter β_2 in the earlier formulated model. Linear predictor value variety 1, hardened: $logit(p_{12k}) = \mu + \beta_2$ (as $\alpha_1 = 0, \gamma_{12} = 0$ for R's cornerstone parameterization). Linear predictor value variety 1, control: $logit(p_{11k}) = \mu$. Hence, difference $logit(p_{12k}) - logit(p_{11k}) = \beta_2$. This difference is $log(odds_{12k}) - log(odds_{11k}) = log(\frac{odds_{12k}}{odds_{11k}}) = log(OR) = \beta_2$. Finally: $OR = exp(\beta_2)$. Answer does not need to be so extensive. If R's cornerstone parameterization not mentioned: -0 pt.

2f2 (3) Estimate the (approximate) standard error of the OR from question 2f1 using the delta method.

$var(OR) = var(exp(\hat{\beta}_2)) \approx (\frac{dexp(\beta_2)}{d\beta_2})^2 var(\hat{\beta}_2) = exp(\beta_2)^2 var(\hat{\beta}_2)$.
Filling in $\hat{\beta}_2$ for β_2 we get $\hat{var}(OR) \approx exp(1.5294)^2 0.6327^2$.
Finally $\hat{se}(OR) \approx exp(1.5294)0.6327 = 2.920$.

2f3 (1) Give an approximate 95% confidence interval for this OR.

$OR \pm 1.96\hat{se}(OR)) = exp(1.5294) \pm 1.96 \times 2.920 = 4.615 \pm 5.723 = (-1.11, 10.34)$.
OR cannot be smaller than 0, so more logical interval is $(0, 10.34)$.
Better interval is obtained by exponentiating lower bound and upper bound of 95% ci for β_2: $1.5294 \pm 1.96 \times 0.6327$, giving $(exp(0.289), exp(2.769)) = (1.34, 15.95)$. This interval doesn't even contain 1!!

2f4 (2) Estimate the OR for hardening versus not-hardening for variety 3, based upon the estimated coefficients.

variety 3, hardened: $log(odds_{32k}) = \mu + \alpha_3 + \beta_2 + \gamma_{32}$.
variety 3, control: $log(odds_{31k}) = \mu + \alpha_3$.
Difference: $log(odds_{32k}) - log(odds_{31k}) = log(OR) = \beta_2 + \gamma_{32}$.
Hence: $OR = exp(\beta_2 + \gamma_{32})$ estimated as $exp(1.5295 + 2.496) = 56.0$.

Question 3 (6) Likelihoods

Suppose we have 3 independent counts y_1, y_2 and y_3 from a Poisson distribution with parameter λ with realizations $y_1 = 3$, $y_2 = 10$, $y_3 = 5$.

Recall that the probability $p(y)$ for outcome y for a Poisson(λ) distribution is given by $p(y) = \frac{\lambda^y e^{-\lambda}}{y!}$.

3a (3) Write down the log-likelihood for the vector $(y_1, y_2, y_3)'$ as a function of λ and find the m.l.e. of λ.

LL:
$LL = log(\frac{\lambda^3 e^{-\lambda}}{3!}) + log(\frac{\lambda^{10} e^{-\lambda}}{10!}) + log(\frac{\lambda^5 e^{-\lambda}}{5!}) = (3+10+5)log(\lambda) - 3\lambda - log(3!) - log(10!) - log(5!) = 18log(\lambda) - 3\lambda - log(3!) - log(10!) - log(5!)$. (1.5pt)

MLE of λ:
$dLL/d\lambda = 18/\lambda - 3 = 0$, so MLE $\hat{\lambda} = 18/3 = 6$. (1.5pt)

The deviance of the current model (i.e. the model that states that all 3 y_i's are independent, coming from the same Poisson distribution with parameter λ) is twice the difference of the maximized log-likelihoods of the current model and the saturated model. In question 3a you already calculated the log-likelhood of this current model.

3b (3) Calculate the deviance of the current model.

Deviance $D = 2(LL(SM) - LL(CM))$.
For the SM each observation has its own parameter:
$LL(SM) = 3log(\lambda_1) - \lambda_1 - log(3!) + 10log(\lambda_2) - \lambda_2 - log(10!) + 5log(\lambda_3) - \lambda_3 - log(5!)$.
Maximizing w.r.t. each λ_i gives MLE's $\hat{\lambda_i} = y_i$, i.e. the observations themselves.
Filling these into the expression for $LL(SM)$ gives:
$LL(SM) = 3log(3) - 3 + 10log(10) - 10 + 5log(5) - 5 - log(3!) - log(10!) - log(5!)$
$LL(CM) = 18log(6) - 3 \times 6 - log(3!) - log(10!) - log(5!)$ (fill MLE $\hat{\lambda} = 6$ into LL of 3a)

$D = 2(3log(3) - 3 + 10log(10) - 10 + 5log(5) - 5 - log(3!) - log(10!) - log(5!) - (18log(6) - 3 \times 6 - log(3!) - log(10!) - log(5!))) = 2(3log(3) - 3 + 10log(10) - 10 + 5log(5) - 5 - (18log(6) - 3 \times 6)) = 2(16.37 - 14.25) = 4.234$.

$LL(SM)$: 1.5pt
$LL(CM)$: 1pt
D: 0.5pt.