MSc track Statistical Science for the Life and Behavioural Sciences
Answers to exam Linear Models, Generalized Linear Models and Linear Algebra

January 13, 2017; 14.00 - 17.00 h

Note: you are not allowed to use your computer; you can either read/interpret given output, or calculate "by hand" (where the use of a calculator may be handy). In the last case bare answers are *not* enough. It should be possible to see how you arrived at the solution.
In total 90 points can be earned, divided over 3 questions (with 54, 28 and 8 points). In question 2 a bonus question can bring an extra 4 points.
The end result of the exam is $(x+10)/10$ rounded to halfs. The end result of the course is $\frac{2}{3}$exam+$\frac{1}{3}$report, provided that the score for the written exam is at least 5.
For hypothesis tests use $\alpha = 0.05$ unless stated otherwise.
If you are working on a question that uses results from an earlier question, but you were not able to answer it, then either proceed using a hypothetical answer on the former question or describe in words how you would proceed.
To give an indication of available time per point: the exam takes 180 minutes, and there are 90 points to be earned. So try to spend not more than 2 minutes per point. You will not have time to look up every detail.

Question 1 (54) Glyphosate

Glyphosate (Roundup) is a herbicide that is used worldwide to kill weeds. There are concerns about its toxicity for humans. In an experiment the breakdown of glyphosate in a specific (loess) soil is studied at three different temperatures (T =10°C, 20°C, or 30°C). Six containers were filled with equal amounts of soil, and spiked with 16 mg of glyphosate per kg soil. The containers were randomly assigned to one of the three treatments and incubated for two months. Each container was regularly sampled over time to measure the remaining glyphosate concentration. Per container the half-life time $DT50$ (in days) was calculated from the time series of concentrations. The $DT50$ is the time at which 50% of the glyphosate has been broken down.

T	$y = DT50$
10	25
10	21
20	12
20	10
30	5
30	7

Does incubation temperature influences the half-life time of glyphosate? To test this we fit a one-way ANOVA model for DT50.

1a1 (3) Write down the one-way ANOVA model using effects model notation (μ, α_i), including the error assumptions.

Answer: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ with $i = 1 \ldots 3$ index for temperature, and $j = 1 \ldots 2$ index for replicates within temperature; y_{ij} is $DT50$ for temperature i and replicate j; ϵ_{ij} are error terms which are assumed to be independent, have constante variance σ^2_ϵ, and are normally distributed with expectation 0.

1a2 (4) Construct the corresponding ANOVA table. As you know, the ANOVA table contains the columns: source of variation, sum of squares, degrees of freedom, mean squares, and F-statistic (skip the P-value); the rows are: corrected model, residual, corrected total.

Answer:

T_{ij}	y_{ij}	$\bar{y}_{i.}$	$y_{ij} - \bar{y}_{i.}$	$(y_{ij} - \bar{y}_{i.})^2$	$\bar{y}_{..}$	$y_{ij} - \bar{y}_{..}$	$(y_{ij} - \bar{y}_{..})^2$
10	25	23	-2	4	13.33	11.67	136.11
10	21	23	2	4	13.33	7.67	58.78
20	12	11	1	1	13.33	-1.33	1.78
20	10	11	1	1	13.33	-3.33	11.11
30	5	6	-1	1	13.33	-8.33	69.44
30	7	6	1	1	13.33	-6.33	40.11
	80		0	$SS_{within} = 12$	80	0	$SST = 317.33$

So, $SS_{within} = 12$ and $SST = 317.33$, hence $SS_{Between} = 317.33 - 12 = 305.33$

ANOVA table:

Source	*SS*	*df*	*MS*	*F*
Between	305.33	2	152.67	38.17
Within	12	3	4	
Corrected Total	317.33	5		

1a3 (1) Which basic model comparison is made in the ANOVA table? Why is this useful?

Answer:
The current (full) model is compared with the null (reduced) model, which contains the intercept only. In this way it is assessed whether the explanatory variables are able to explain substantially (significantly more than the model without any explanatory variables (i.e. intercept only model).

1a4 (2) Give the estimate of the residual standard deviation and its interpretation.

Answer:
Estimator residual variance is MS_{within} from ANOVA table, so $\hat{\sigma}_\epsilon = \sqrt{4} = 2$. It represents the average deviation of observations from the predicted value according to fitted model.

1a5 (2) Give the null hypothesis that is tested with the F-test in the ANOVA table. How do you judge the significance of the F-test here?

Answer:
$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. The F-test statistic equal 38.17 here, which is much larger than 1, so will be signficant.

1a6 (2) Calculate R^2_{adj}.

Answer:
$R^2_{adj} = 1 - \frac{MS_{within}}{MS_{Ctotal}} = 1 - \frac{4}{317.33/5} = 0.937.$

The one-way ANOVA model in effects model notation is overparameterized. R solves this by taking parameter $\alpha_1 = 0$, essentially removing this parameter from the model.

1b1 (1) What does overparameterization mean?

Answer:
Overparameterization means that the model contains more parameters than can be estimated; there are superfluous parameters in the model.

1b2 (2) What is the interpretation of μ, α_2 and α_3 using the default parameterization of R? Using this knowledge, give quickly the least-squares estimates of the parameters.

Answer:
μ is the expected value of $DT50$ in the reference group, here at temperature 10°; $\hat{\mu} = 23$.
α_2 is the difference between expected $DT50$ for temperature 20 °and 10°; $\hat{\alpha}_2 = 11 - 23 = -12$.
α_3 is the difference between expected $DT50$ for temperature 30 °and 10°; $\hat{\alpha}_3 = 6 - 23 = -17$.

We want to write the one-way ANOVA model for the 6 observations in matrix notation:

$$y = X\beta + \epsilon \tag{1}$$

for the default parameterization of R with $\beta' = (\mu, \alpha_2, \alpha_3)$.

1b3 (2) Give model matrix X.

Answer:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Maybe a simple linear regression model is enough to describe the relationship between DT50 and temperature. We are going to look into this by reparameterizing the model. We want to form a new parameter vector $\gamma = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$ instead of $\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$.

In vector γ we want element γ_0 to represent the average of the mean $DT50$'s at the three temperatures: $\gamma_0 = (\mu + (\mu + \alpha_2) + (\mu + \alpha_3))/3 = \mu + \alpha_2/3 + \alpha_3/3$; γ_1 is a coefficient for the linear trend: $\gamma_1 = (((\mu+\alpha_2) - \mu) + ((\mu+\alpha_3) - (\mu+\alpha_2)))/2 = \alpha_3/2$; γ_2 represents the deviation from the linear trend, here by comparing differences: $\gamma_2 = ((\mu + \alpha_2) - \mu) - ((\mu + \alpha_3) - (\mu + \alpha_2) = 2\alpha_2 - \alpha_3$.

1b4 (2) Give matrix C such that $\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = C \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$.

Answer:

$$C = \begin{pmatrix} 1 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} \\ 0 & 2 & -1 \end{pmatrix}.$$

We want to change the linear model (1) such that we directly get least-squares estimates of γ. For that, the inverse of matrix C is needed.

1b5 (3) Show that the inverse of C is $C^{-1} = \begin{pmatrix} 1 & -1 & -\frac{1}{6} \\ 0 & 1 & \frac{1}{2} \\ 0 & 2 & 0 \end{pmatrix}$.

Answer:

E.g. by Gaussian elimination: $\left(\begin{array}{ccccccc} 1 & \frac{1}{3} & \frac{1}{3} & | & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & | & 0 & 1 & 0 \\ 0 & 2 & -1 & | & 0 & 0 & 1 \end{array}\right)$.

$$\xrightarrow[\text{swap r2 - r3}]{\text{r3/2}} \left(\begin{array}{ccccccc} 1 & \frac{1}{3} & \frac{1}{3} & | & 1 & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & | & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & | & 0 & 1 & 0 \end{array}\right) \xrightarrow{\text{r1-r2/3}} \left(\begin{array}{ccccccc} 1 & 0 & \frac{1}{2} & | & 1 & 0 & -\frac{1}{6} \\ 0 & 1 & -\frac{1}{2} & | & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & | & 0 & 1 & 0 \end{array}\right)$$

$$\xrightarrow{\text{2r3}} \left(\begin{array}{ccccccc} 1 & 0 & \frac{1}{2} & | & 1 & 0 & -\frac{1}{6} \\ 0 & 1 & -\frac{1}{2} & | & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & | & 0 & 2 & 0 \end{array}\right) \xrightarrow[\text{r2+r3/2}]{\text{r1-r3/2}} \left(\begin{array}{ccccccc} 1 & 0 & 0 & | & 1 & -1 & -\frac{1}{6} \\ 0 & 1 & 0 & | & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & | & 0 & 2 & 0 \end{array}\right).$$

If you were not able to obtain C in question b4 find the inverse of $\begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix}$.

Inverse of this matrix is $\begin{pmatrix} -\frac{1}{2} & -\frac{5}{2} & 3 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

Continuing, with $\gamma = C\beta$, then $\beta = C^{-1}\gamma$, and $X\beta = XC^{-1}\gamma = X_{nw}\gamma$. Form the new parameterized model:

$$y = X_{nw}\gamma + \epsilon \tag{2}$$

.

1b6 (2) Show that X_{nw} is $\begin{pmatrix} 1 & -1 & -\frac{1}{6} \\ 1 & -1 & -\frac{1}{6} \\ 1 & 0 & \frac{1}{3} \\ 1 & 0 & \frac{1}{3} \\ 1 & 1 & -\frac{1}{6} \\ 1 & 1 & -\frac{1}{6} \end{pmatrix}$.

Answer:

$X_{nw} = XC^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & -\frac{1}{6} \\ 0 & 1 & \frac{1}{2} \\ 0 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -\frac{1}{6} \\ 1 & -1 & -\frac{1}{6} \\ 1 & 0 & \frac{1}{3} \\ 1 & 0 & \frac{1}{3} \\ 1 & 1 & -\frac{1}{6} \\ 1 & 1 & -\frac{1}{6} \end{pmatrix}$ using matrix multiplication.

1b7 (4) Form the matrix $X'_{nw}Xnw$ and construct the normal equations. You may want to use that $X'_{nw}y = \begin{pmatrix} 80 \\ -34 \\ -2\frac{1}{3} \end{pmatrix}$.

Solve the normal equations to obtain the least-squares estimates $\hat{\gamma} = \begin{pmatrix} 13.33 \\ -8.5 \\ -7 \end{pmatrix}$.

Answer:

$X'_{nw}X_{nw} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 0 & 0 & 1 & 1 \\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{6} & -\frac{1}{6} \end{pmatrix} \begin{pmatrix} 1 & -1 & -\frac{1}{6} \\ 1 & -1 & -\frac{1}{6} \\ 1 & 0 & \frac{1}{3} \\ 1 & 0 & \frac{1}{3} \\ 1 & 1 & -\frac{1}{6} \\ 1 & 1 & -\frac{1}{6} \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}$ which is a diagonal matrix, so very easily inverted.

Normal equations are: $X_{nw}'X_{nw}\gamma = X_{nw}'y$. Filling in the numbers:

$$\begin{pmatrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \hat{\gamma} = \begin{pmatrix} 80 \\ -34 \\ -2\frac{1}{3} \end{pmatrix} \iff \hat{\gamma} = (X_{nw}'X_{nw})^{-1}X_{nw}'y = \begin{pmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 80 \\ -34 \\ -2\frac{1}{3} \end{pmatrix} = \begin{pmatrix} 13\frac{1}{3} \\ -8\frac{1}{2} \\ -7 \end{pmatrix}$$

1b8 (2) Show how the variance-covariance matrix $\hat{Var}(\hat{\gamma}) = \begin{pmatrix} 0.67 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 12 \end{pmatrix}$ is obtained. What are the off-diagonal zero's telling you?

Answer:

$\hat{Var}(\hat{\gamma}) = \hat{\sigma}_\epsilon^2 (X_{nw}'X_{nw})^{-1} = 4 \begin{pmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 12 \end{pmatrix}$. The off-diagonal zero's tell that estimators of parameters γ_i $(i = 0, 1, 2)$ are uncorrelated.

Returning to the question whether a simple linear regression model is good enough, we test $H_0 : \gamma_2 = 0$ vs $H_a : \gamma_2 \neq 0$.

1b9 (3) Use the estimates shown in b7 and b8 to calculate the outcome of an appropriate test statistic. Perform the test. What is the conclusion? Motivate your answer.

Answer:
A t-test may be performed. Outcome $t = \frac{\hat{\gamma}_2 - 0}{\hat{se}(\hat{\gamma}_2)} = \frac{-7}{\sqrt{(12)}} = 2.02$. Under H_0 t has a t_3 distribution. This outcome is not very extreme in the t_3 distribution, so probably H_0 will not be rejected.

1b10 (2) Describe another way of testing for goodness of fit of the simple linear regression model that could be applied here (no calculations needed).

Answer:
The F-test for goodness of fit may be used. In this test the Full Model is the one-way ANOVA model, and the Reduced Model is the simple linear regression model.

In reality the experiment was larger. Not only were there more (5) temperature levels and more replicates, but also a second experimental variable was involved: light regime fL. Soil containers were either kept under dark (D) or light (L) conditions. Break-down of glyphosate is largely by bacterial activity, and it is expected that under light conditions break-down is faster. For each of the 5 incubation temperatures, 5 containers were incubated in a dark condition and 5 in a light condition. The first 5 observations of the dataframe are shown here:

```
head(glyphosate, n=5)

##    T fL     y
## 1 10  D 19.37
## 2 10  D 22.37
## 3 10  D 16.32
## 4 10  D 24.26
## 5 10  D 28.14
```

We fit a model with linear and quadratic effects of T and main effects of factor fL:

$$y_{ijk} = \mu + \alpha_i + \beta_1 T_j + \beta_2 T_j^2 + \epsilon_{ijk} \tag{3}$$

with index $i = 1, 2$ for the two light levels (D and L), index $j = 1 \ldots 5$ for the five temperature levels (10, 15, 20, 25, 30) and $k = 1 \ldots 5$ for the replicates per treatment combination.

```
lmo1 <- lm(y ~ fL + T + I(T^2), data=glyphosate)
summary(lmo1)

##
## Call:
## lm(formula = y ~ fL + T + I(T^2), data = glyphosate)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.552 -1.183 -0.304  1.074  5.516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.95161    3.34285   13.15  < 2e-16
## fLL         -2.76510    0.74748   -3.70  0.00058
## T           -2.49685    0.36125   -6.91  1.2e-08
## I(T^2)       0.03888    0.00893    4.35  7.4e-05
##
## Residual standard error: 2.64 on 46 degrees of freedom
## Multiple R-squared:  0.884,Adjusted R-squared:  0.876
## F-statistic:  117 on 3 and 46 DF,  p-value: <2e-16

deviance(lmo1)

## [1] 321.272
```

1c1 (2) Write down the fitted model equation for containers under the light condition.

Answer:
$\hat{\mu}_y = (43.95 - 2.77) - 2.50T + 0.039T^2 = 41.18 - 2.50T + 0.039T^2$

1c2 (4) Test the null hypothesis of no effect of light regime against the suggested alternative hypothesis. Mention: 1) H_0 and H_a formulated in parameters; 2) definition of the test-statistic; 3) distribution of the test-statistic under H_0; 4) outcome of the test-statistic; 5) P-value; 6) conclusion.

Answer:

1. $H_0 : \alpha_2 = 0$ vs $H_a : \alpha_2 < 0$ (taking R's paramerization $\alpha_1 = 0$); notice left-sided alternative hypothesis, because we want to show that light increases break-down, i.e. DT50 is smaller.
2. $t = (\hat{\alpha}_2 - 0)/\hat{se}(\hat{\alpha}_2)$
3. Under H_0 $t \sim t_{46}$
4. Outcome $t = -3.70$
5. P-value is $0.00058/2 = 0.00029$. We divide the P-value from the R output by two, because we need a left-tailed P-value, whereas R took a two-sided alternative hypothesis.
6. Conclusion: reject H_0 because $P < 0.05$. We conclude that light leads to systematically lower DT50-values.

```
anova(lmo1)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## fL         1   95.6    95.6   13.68 0.000576
## T           1 2216.2  2216.2  317.31  < 2e-16
## I(T^2)      1  132.3   132.3   18.94 7.44e-05
## Residuals 46  321.3     7.0
```

1c3 (3) Comment on the sums of squares produced by the `anova()` function. Why are the P-values for the t-tests (`summary` output) and corresponding F-tests (`anova` output) identical for factor fL and T^2, but not for T? Applying the principle of marginality, which tests make most sense?

Answer:
The `anova()` function produces sequential sums of squares. If regressors are correlated, these sequential sums of squares are order dependent. The t-tests are partial tests, which correspond to partial sums of squares (terms being added to the model last). Here, the dummy for fL is uncorrelated both with T and T^2, so the P-values for the F-test from the `anova` function and t-test are identical for fL. T and T^2 are correlated, so T before T^2 will give a different SS compared to T after T^2.
The principle of marginality states that we should first consider higher order terms (like T^2), and if the higher order terms are unimportant (so they can be removed) the lower order terms (like T). This corresponds most to the tests obtained with the `anova` function here.

The effect of light on the $DT50$ may be different at different temperatures. To check this we add interaction terms to the model.

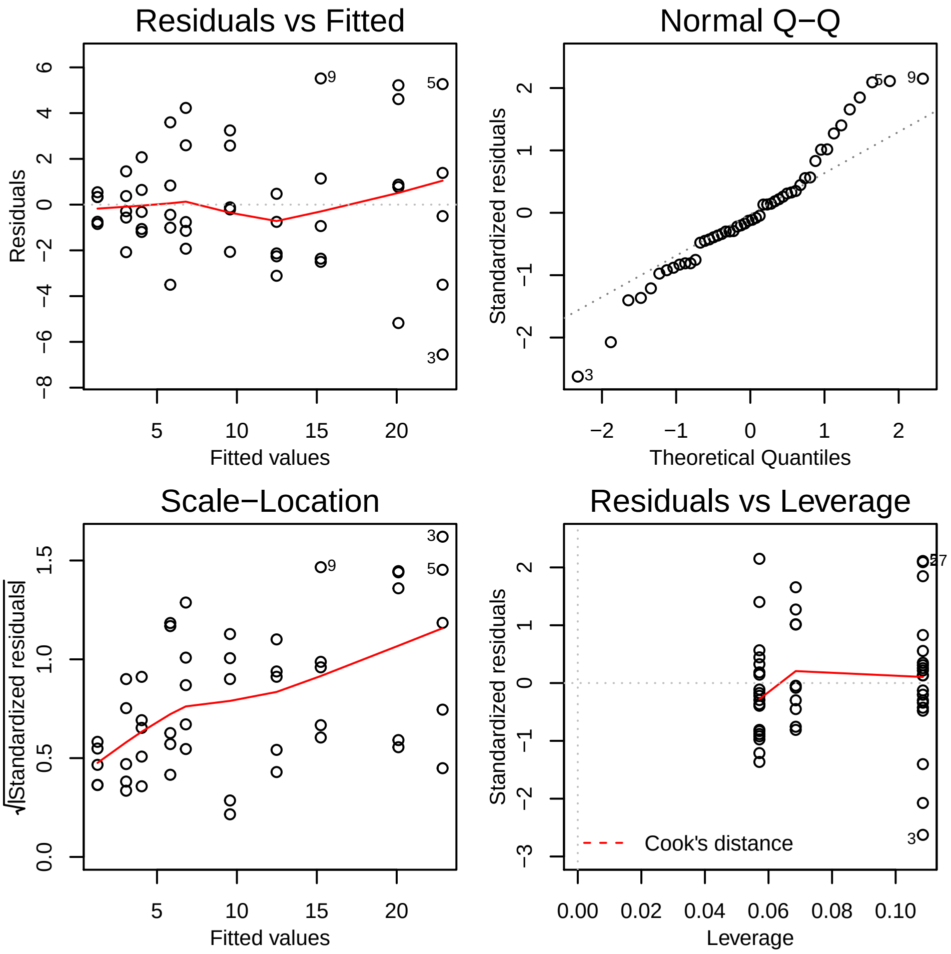
```
lmo2 <- lm(y ~ fL + T + I(T^2) + fL:T + fL:I(T^2), data=glyphosate)
deviance(lmo2)

## [1] 313.281

lmo2$df.residual

## [1] 44
```

1c4 (3) Construct the test statistic to test the null hypothesis stating that the effect of light regime does not vary with temperature. What do you conclude (roughly)?

Answer:
F-test comparing FM and RM: $F = \dfrac{(321.3 - 313.3)/(46 - 44)}{313.3/44} = 0.56$. F is even smaller than 1, so there is no indication whatsoever that interaction is important.

1d1 (4) Comment on the fit of model (3) based on the array of residual plots shown below.

```
par(mfrow=c(2,2), mar=c(3,3,2,1), mgp=c(2,1,0), cex=0.7)
plot(lmo1)
```



Answer:
Plot left-top shows residuals versus predicted values, mainly meant to check the constant variance assumption. Here we see that the variance increases with the mean (”loudspeaker” shaped pattern).
Plot left-bottom shows the same: square roots of absolute residuals tend to increase with increasing value of $\hat{y}$.
Plot right-top is a normal QQ-plot, meant to check normality of the (studentized) residuals. The points are deviating a bit from the lined plotted in the graph, but the studentized residuals are all within the range (-3,3), so indications of non-normality are not very strong.
Plot right-bottom shows leverages on x-axis versus studentized residuals. In this experimental setting we do not expect to see outliers on the x-axis (as indicated by leverages). The threshold for extreme leverages ($2 * p/n = 2 * 4/50 = 0.16$) is not reached here. All studentized residuals are in the range -3 to 3, so no regression outliers.

1d2 (1) Suggest a transformation of y that could possibly solve the problems we see here.

Answer:
A transformation ”down the ladder” could work, like the *log* or *sqrt* transformation.

Question 2 (28) Lemurs

An observational study of the infection of lemurs on Madagascar by helminths (parasitic worms) is performed. Observations on droppings of wild lemurs are obtained during two seasons (factor $Season$, levels Dry and Wet) and two locations (factor Loc, levels $East$ and $West$, indicating whether lemurs lived at the Eastern or Western side of Madagascar). The two factors are combined into a single factor $Comb$ with 4 levels.
Lemurs live in family groups with group sizes n ranging roughly between 1 and 9. For each individual of a group droppings are collected and infection with the parasite (yes/no) is determined. The data are aggregated per group, leading to counts k of infected animals out of n.
Below the first five lines of the dataframe are shown:

```
##    Loc Season     Comb n k
## 1 East    Dry East.Dry 2 2
## 2 East    Dry East.Dry 5 5
## 3 East    Dry East.Dry 4 2
## 4 East    Dry East.Dry 4 1
## 5 East    Dry East.Dry 4 2
```

The data are analyzed using logistic regression. We take the combination factor as (single) explanatory variable, so that the linear predictor η_{ij} is $\mu + \alpha_i$, like in one-way ANOVA ($i = 1 \dots 4$, corresponding to the 4 combinations: $East.Dry$, $East.Wet$, $West.Dry$, $West.Wet$; j is an index for lemur group within combination; numbers of family groups are 12, 15,10 and 16 for the four combinations).

```
glmo1 <- glm(cbind(k, n-k) ~ Comb, family = binomial(link=logit), data=lemurs)
summary(glmo1)

##
## Call:
## glm(formula = cbind(k, n - k) ~ Comb, family = binomial(link = logit),
##     data = lemurs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.049  -0.614  -0.103   0.621   2.257
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.619      0.271    2.29    0.022
## CombEast.Wet   0.362      0.361    1.00    0.317
## CombWest.Dry   1.046      0.493    2.12    0.034
## CombWest.Wet  -0.773      0.353   -2.19    0.029
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 76.423  on 52  degrees of freedom
## Residual deviance: 54.473  on 49  degrees of freedom
## AIC: 144.1
##
## Number of Fisher Scoring iterations: 4
```

2a (3) Write down the model, paying attention to the three components of a generalized linear model (GLM).

Answer: Three components of a generalized linear model:

1. Random part of model $k_{ij} \sim Bin(n_{ij}, p_{ij})$, independent, with index $i = 1 \dots 4$ for combinations, and index j for family groups within combinations.

2. Systematic part of model: linear predictor $\eta_{ij} = \mu + \alpha_i$, with α_i difference in linear predictor value of combination i compared to the first combination ($East.Dry$), and μ the mean linear predictor value of the first combination.

3. Link function: $logit(p_{ij}) = \eta_{ij}$

In binomial GLM's overdispersion is a serious topic that should be considered first.

2b1 (2) Explain what overdispersion is about. What happens if overdispersion is present, but ignored?

Answer:
Overdispersion is the phenomenon that counts vary more than can be explained by the binomial distribution. For a $Bin(n, p)$ distribution the variance is $np(1 - p)$. In case of overdispersion the counts (at constant explanatory variables values) show larger variance. If overdispersion is ignored, the standard errors (calculated based upon the binomial variance) will be too small, leading to confidence intervals which are too narrow and P-values which are too small.

2b2 (2) Can you expect overdispersion on biological grounds in the lemur example? Is there conclusive evidence for overdispersion here?

Answer:
Yes, you can. The individuals within a group may infest each other, leading to correlated binary scores. For a binomial distribution the binary (individual) scores need to be independent. Correlation among binary scores may lead to overdispersion.
Here the residual deviance is 54.5 on 49 df. As the residual deviance has approximately a chi-square distribution with $49df$ if there was no overdispersion, we expected values around 49 with standard deviation $\sqrt{2 * 49} \approx 10$. The value 54.5 is well within one standard deviation from the expected value, so no indication of overdispersion here.

2c (5) Use a Likelihood Ratio test to test the null hypothesis that the infection prevalences (proportions) are the same in the four combinations of seasons and locations, using the `summary` output above. Mention: 1) H_0 and H_a formulated in parameters; 2) definition of the test statistic; 3) distribution of the test statistic if H_0 is true; 4) outcome of the test statistic; 5) conclusion (roughly) with its motivation.

Answer:

1. H_0: $(\alpha_1 =)\alpha_2 = \alpha_3 = \alpha_4 = 0$ versus H_a: not all $\alpha_i = 0$.

2. LR test-statistic: LRT = deviance(null model) - deviance(full model).

3. Under H_0 $LRT \sim \chi^2_3$ with $df = 3 = 52 - 49$.

4. Outcome: $LRT = 76.423 - 54.473 = 21.95$

5. Conclusion: reject H_0 because 21.95 is far in the right tail of the χ^2_3 distribution (for reasoning see 2b2). We conclude that infection prevalences are systematically different among the four combinations.

2d (2) How can the estimated coefficient $\hat{\alpha}_2 = 0.362$ (as found in the `summary` output) be used to estimate the infection Odds Ratio during the wet compared to the dry season in the Eastern part of Madagascar?

Answer:
Infection Odds Ratio of wet versus dry season in Eastern part is $exp(0.362) = 1.44$, because the coefficient 0.362 for combination $East.Wet$ measures the difference (on logit scale) between $East.Wet$ and the reference group, which is $East.Dry$.

2e1 (1) Estimate the mean response (on the linear predictor scale) during the dry season in the Eastern part.

Answer:
The linear predictor value for this combination is simply μ with estimate $\hat{\mu} = 0.619$.

2e2 (2) Give an approximate 0.95 confidence interval for the mean response (on the linear predictor scale) of lemurs during the dry season in the Eastern part.

Answer:
Approximate 0.95 c.i. for μ is $0.619 \pm 1.96 \times 0.271 = (0.088, 1.150)$.

2e3 (3) Give 1) an estimate of the mean infection probability of lemurs during the dry season in the Eastern part, and 2) an approximate 0.95 confidence interval for the mean infection probability of a lemur during the dry season in the Eastern part (using the inverse link function).

Answer:

1. $\hat{p} = \dfrac{1}{1+e^{-\hat{\mu}}} = \dfrac{1}{1+e^{-0.619}} = 0.65$
2. Approximate 0.95 c.i. for p by backtransforming lower and upper bounds of the interval from 2e2: $(\dfrac{1}{1+e^{-0.088}}, \dfrac{1}{1+e^{-1.150}}) = (0.52, 0.76)$.

2f1 (1) Estimate the difference in mean response (on the linear predictor scale) between the wet and dry season in the Western part of Madagascar (this is the linear combination $\alpha_4 - \alpha_3$).

Answer:
$\hat{\alpha_4} - \hat{\alpha_3} = -0.773 - 1.046 = -1.819$.

To obtain a confidence interval for this difference, we need to estimate its standard error.

2f2 (3) Calculate the approximate standard error of $\hat{\alpha}_4 - \hat{\alpha}_3$. You may want to use the following.

```
vcov(glmo1)

##                (Intercept) CombEast.Wet CombWest.Dry CombWest.Wet
## (Intercept)        0.07326     -0.07326     -0.07326     -0.07326
## CombEast.Wet      -0.07326      0.13055      0.07326      0.07326
## CombWest.Dry      -0.07326      0.07326      0.24314      0.07326
## CombWest.Wet      -0.07326      0.07326      0.07326      0.12485
```

Answer:
$\hat{se}(\hat{\alpha}_4-\hat{\alpha}_3) = \sqrt{v\hat{a}r(\hat{\alpha}_4 - \hat{\alpha}_3)} = \sqrt{v\hat{a}r(\hat{\alpha}_4) + v\hat{a}r(\hat{\alpha}_3) - 2cov(\hat{\alpha}_4, \hat{\alpha}_3)} = \sqrt{0.12485 + 0.24314 - 2 \times 0.07326} = \sqrt{0.22147} = 0.47$

Bonus question (skip first, answer if time allows):
2f3 (4) Calculate the approximate standard error of $exp(\hat{\alpha}_4 - \hat{\alpha}_3)$.

Answer:
Use the delta-method with function $f(\alpha_4, \alpha3) = exp(\alpha_4 - \alpha_3)$. The partial derivatives are $\frac{\partial f}{\partial \alpha_4} = exp(\alpha_4 - \alpha_3)$ (with realized value $exp{-1.819} = 0.162$) and $\frac{\partial f}{\partial \alpha_3} = -exp(\alpha_4 - \alpha_3)$, with realized value -0.162.
Then $v\hat{a}r(exp(\hat{\alpha}_4 - \hat{\alpha}_3)) \approx (\frac{\partial f}{\partial \alpha_4})^2 v\hat{a}r(\hat{\alpha}_4) + (\frac{\partial f}{\partial \alpha_3})^2 v\hat{a}r(\hat{\alpha}_3) + 2\frac{\partial f}{\partial \alpha_4}\frac{\partial f}{\partial \alpha_3} c\hat{o}v(\hat{\alpha}_4, \hat{\alpha}_3) = 0.162^2 \times 0.12485 + (-0.162)^2 \times 0.24314 + 2 \times 0.162 \times -0.162 \times 0.07326 = 0.005812$. So, the approximate standard error is $\sqrt{0.005812} = 0.0762$.

The western part of Madagascar is much drier than the eastern part. The difference in rainfall between wet and dry season is relatively large in the Western part, but small in the Eastern part. This may affect food availability and infection susceptibility over the year. It might be that in the Western part the seasonal difference in infection prevalence is higher than in the Eastern part.
We try to study this by introducing main effects for *Loc*, *Season* and their interaction.

```
glmo2 <- glm(cbind(k, n-k) ~ Loc + Season + Loc:Season, family = binomial(link=logit), data=lemurs)
summary(glmo2)

##
## Call:
## glm(formula = cbind(k, n - k) ~ Loc + Season + Loc:Season, family = binomial(link = logit),
##     data = lemurs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.049  -0.614  -0.103   0.621   2.257
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.619      0.271    2.29  0.02219
## Locwest              1.046      0.493    2.12  0.03390
## SeasonWet            0.362      0.361    1.00  0.31668
## Locwest:SeasonWet   -2.181      0.593   -3.68  0.00024
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 76.423  on 52  degrees of freedom
## Residual deviance: 54.473  on 49  degrees of freedom
## AIC: 144.1
##
## Number of Fisher Scoring iterations: 4
```

2g1 (3) Test the hypothesis that, on the logit-scale, the effect of season is the same, regardless the location, using the output given above. Give 1) the name of the test, 2) the distribution of the test statistic under H_0, 3) the outcome of the test statistic, and 4) P-value and conclusion.

Answer:
Test for interaction parameter $Locwest.SeasonWet$. This parameter measures a difference of differences: the seasonal difference in infection in the West compared to the difference in the East.

1. Wald test
2. approximately $N(0, 1)$
3. Outcome: $z = -3.68$
4. One-sided P-value is $0.00024/2 = 0.00012 < 0.05$. So reject H_0. We conclude that in the Western part the seasonal difference is higher.

2g2 (1) If we would have used the first model, which linear combination of parameters would we need to estimate to test the null hypothesis of 2g1?

Answer:
$(\mu + \alpha_4 - (\mu + \alpha3)) - (\mu + \alpha_2 - (\mu + \alpha1)) = \alpha_4 - \alpha_3 - \alpha_2$

Question 3 (8) Maximum likelihood and generalized linear models

3a (4) Suppose we have 3 random observations y_1, y_2 and y_3 from a Poisson(λ) distribution (density $f(y) = \lambda^y e^{-\lambda}/y!$) with realizations 3, 5, and 10. Write down the log-likelihood for these data, and calculate the maximum likelihood estimate of λ.

Answer:
$LL = log((\lambda^3 e^{-\lambda}/3!)(\lambda^5 e^{-\lambda}/5!)(\lambda^{10} e^{-\lambda}/10!)) = log(\lambda^{18} e^{-3\lambda}/(3!5!10!)) = 18log(\lambda) - 3\lambda - log(3!5!10!)$.
Mle of λ obtained equating the derivative to zero: $\frac{dLL}{d\lambda} = 18/\lambda - 3 = 0 \Rightarrow \lambda = 18/3 = 6$.

3b (4) Write the Poisson distribution in exponential family form, thereby specifying: 1) the canonical parameter θ, 2) the function $b(\theta)$, 3) the dispersion parameter ϕ, 4) the expectation $E(y)$, 5) the variance function $V(\mu)$, 6) var(y), and 7) the canonical link function.

Answer:
Exponential family: $log(f(y)) = log(\lambda^y e^{-\lambda}/y!) = ylog(\lambda) - \lambda - log(y!) = \frac{ylog(\lambda) - exp(log(\lambda))}{1} - log(y!)$

1. $\theta = log(\lambda)$
2. $b(\theta) = exp(\theta)$
3. $\phi = 1$
4. $\mu = b'(\theta) = exp'(\theta) = exp(\theta) = exp(log(\lambda)) = \lambda$
5. $V(\mu) = b''(\theta) = exp''(\theta) = exp(\theta) = exp(log(\lambda)) = \lambda$
6. var$(y) = V(\mu)\phi = \lambda 1 = \lambda$
7. canonical link by inverting b': $\theta = (b')^{-1}(\mu) = exp^{-1}(\mu) = log(\mu)$.