Answers to the exam of Linear Models, Generalized Linear Models and Linear Algebra, dd 29 January 2010

Gerrit Gort

The answers given below are rather elaborate to give you some extra background information. Generally, you can give shorter answers during the examination!

1 (30). We study the relationship between amount of lime applied to a soil and the pH of the soil. We have 3 soils and apply no lime, 1 unit of lime, and 2 units of lime. After application of lime, the pH values are measured: 5.8, 6.2, 6.3. The relationship is modeled by simple linear regression, where variable rate represents the amount of lime (with values 0,1,2): pH $= \alpha + \beta$ rate$+\epsilon$; $var(\epsilon) = \sigma^2_\epsilon$ (constant).

a (2) Calculate by hand the least-squares estimates of intercept α and slope β.
Using $\sum x = 3$, $\sum y = 18.3$, $\sum x^2 = 5$, $\sum xy = 0 \times 5.8 + 1 \times 6.2 + 2 \times 6.3 = 18.8$, $n = 3$, we get $\hat{\beta} = \dfrac{\sum xy - \sum x \sum y/n}{\sum x^2 - (\sum x)^2/n/} = \dfrac{18.8 - 3 \times 18.3/3}{5 - 3^2/3} = 0.25$, and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 6.1 - 0.25 \times 1 = 5.85$.

In reality a larger experiment was done. In total 25 soils were examined, using 5 levels of lime (0, 1, 2, 4, and 8 units), each applied to 5 randomly selected soils. Below we fit the simple linear regression model, regressing pH on rate.

```
> lime.reg <- lm(pH ~ rate, data=lime)
> summary(lime.reg)

Call:
lm(formula = pH ~ rate, data = lime)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14663 -0.03783 -0.00128  0.03217  0.15545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.73783    0.02119  270.83  < 2e-16
rate         0.03172    0.00514    6.17  2.7e-06

Residual standard error: 0.0727 on 23 degrees of freedom
Multiple R-squared: 0.624,        Adjusted R-squared: 0.607
F-statistic: 38.1 on 1 and 23 DF,  p-value: 2.68e-06
```

b (1) Give the estimates of α, β, and σ_ϵ.
$\hat{\alpha} = 5.738$, $\hat{\beta} = 0.0317$, and $\hat{\sigma_\epsilon} = 0.0727$.

c (2) Give practical interpretations of α, β, and σ_ϵ.
α is the expected soil pH if no lime is applied; β is the expected change in soil pH for a unit change in lime amount; σ_ϵ may be described as the expected deviation of the pH of a soil from the population mean at constant lime rate.

d (2) In the output above, a number of hypothesis tests is performed. For all, give H_0, H_1, test-statistics, P-values, and conclusions.
Three tests:

1. "Omnibus" F-test: $H_0 : \beta = 0$, $H_1 : \beta \neq 0$, $F = MSR/MSE$ with outcome $F = 38.1$, P-value $P(F_{1,23} \geq 38.1) = 2.68e-06$; P-value < 0.05, so reject H_0, slope is not equal to zero, relationship lime and pH is proven.
2. t-test: $H_0 : \alpha = 0$, $H_1 : \alpha \neq 0$, $t = (\hat{\alpha} - 0)/se(\hat{\alpha})$ with outcome $t = 270.83$; P-value $<$ 2e-16, so H_0 is rejected, intercept is not zero; this is a test without meaning.
3. t-test: $H_0 : \beta = 0$, $H_1 : \beta \neq 0$, $t = (\hat{\beta} - 0)/se(\hat{\beta})$ with outcome $t = 6.17$; P-value $P(|t_{23}| > 6.17) = 2.7e-6$, so H_0 is rejected, slope deviates from zero; notice that $F = t^2$: in simple linear regression omnibus F-test and t-test for slope are equivalent.

e1 (2) Write down the model in matrix notation, making clear what is what.

Linear model is $y = X\beta + \epsilon$, here: $\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ .. \\ y_{24} \\ y_{25} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ .. & .. \\ 1 & 8 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ .. \\ \epsilon_{24} \\ \epsilon_{25} \end{bmatrix}$.
Second column of X contains the 25 values of lime.

e2 (3) Write down the normal equations, and solve them. You may want to use that $\sum_{i=1}^{25} X_i = 75$, $\sum_{i=1}^{25} X_i^2 = 425$, $\sum_{i=1}^{25} Y_i = 145.82$, and $\sum_{i=1}^{25} X_i Y_i = 443.82$.

Normal equations are $X'X\mathbf{b} = X'y$, here in matrix notation: $\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$.
Fill in the numbers: $\begin{bmatrix} 25 & 75 \\ 75 & 425 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 145.82 \\ 443.82 \end{bmatrix}$.
Hence two equations with two unknowns: $\begin{bmatrix} 25\hat{\alpha} + 75\hat{\beta} & = & 145.82 \\ 75\hat{\alpha} + 425\hat{\beta} & = & 443.82 \end{bmatrix}$.
Solving, second equation minus 3 times first equation gives $200\hat{\beta} = 6.36$, so $\hat{\beta} = 0.0318$. Then $\hat{\alpha} = (145.82 - 75 \times 0.0318)/25 = 5.7374$. Notice slight differences with R-output due to rounding.

e3 (3) Show how the standard error of $\hat{\beta}$ can be obtained using matrix calculation.

Let $\mathbf{b}$ be the least-squares solution $\mathbf{b} = (X'X)^{-1}X'y$ with X and y as in e2. Then variance-covariance matrix of $\mathbf{b}$ is $Var(\mathbf{b}) = \sigma_\epsilon^2 (X'X)^{-1}$.
Now, the inverse $X'X^{-1} = \begin{bmatrix} 25 & 75 \\ 75 & 425 \end{bmatrix}^{-1} = \frac{1}{25 \times 425 - 75^2} \begin{bmatrix} 425 & -75 \\ -75 & 25 \end{bmatrix}$. The second element of the diagonal is needed: $\frac{1}{25 \times 425 - 75^2} \times 25 = 25/5000 = 0.005$. Variance σ_ϵ^2 is estimated as $\hat{\sigma}_\epsilon^2 = MSE = 0.0727^2$. Then, $se(\hat{\beta}) = \sqrt{0.0727^2 \times 0.005} = 0.0051407$, as given by R.

e4 (1) Sketch the geometrical interpretation of least squares estimation.
For a geometrical interpretation look at figure 10.2 in Fox, page 222.

The simple linear regression may be too simple. Let's compare it with the model that gives a mean to each level of variable rate.
f (3) Calculate the F-statistic for the test of goodness of fit based on the residual sums of squares of the simple linear regression model and the one-way ANOVA model. How would you judge the significance of the outcome?

```
> deviance(lime.reg)

[1] 0.121455
```

```
> lime.anova <- lm(pH ~ as.factor(rate), data=lime)
> deviance(lime.anova)
```

```
[1] 0.09909
```

$F = \frac{(SSE_{RM} - SSE_{FM})/(dfe_{RM} - dfe_{FM})}{SSE_{FM}/dfe_{FM}}$ with RM=reduced model = simple linear regression model, and FM=full model=one-way ANOVA model. Fill in: $F = \frac{(0.121455 - 0.09909)/3}{0.09909/20} = 1.5047$.
The test-statistic behaves under H_0 like a random draw from a $F_{3,20}$ distribution. Because the ratio 1.5 is not far from 1, we can say (with some knowledge of F-distributions, without looking in the F-table) that H_0: "Extra parameters to go from regression to ANOVA model not needed" will not be rejected. In other words, the simple linear regression is good enough, goodness-of-fit is ok.

g (1) Write down the ANOVA model, used above, in matrix notation.

The linear model $y = X\beta + \epsilon$ for one-way ANOVA situation with overparameterization will be:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ .. \\ y_{24} \\ y_{25} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ .. & (x_1) & (x_2) & (x_3) & (x_4) & (x_5) \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ .. \\ \epsilon_{24} \\ \epsilon_{25} \end{bmatrix}.$$

Variables $x_1 \ldots x_5$ are five dummy variables for 5 levels of lime.

In the experiment two types of lime were used (`AL` and `GL`), both applied at rates 0, 1, 2, 4 and 8, in total to 50 soils. The effects of the two types of lime on soil pH may be different. We want to include both types of lime in the linear model, and make comparisons between them.

h1 (1) How are linear models of this type generally called?
ANCOVA models = Analysis of covariance models

We fit a linear model that allows two separate regression lines for the two types of lime, see output below.

```
> lime.reg2 <- lm(pH ~ type + rate + type*rate, data=lime.all)
> summary(lime.reg2)$coefficients
```

```
             Estimate Std. Error   t value    Pr(>|t|)
(Intercept)  5.780775 0.01765676 327.39733 4.11130e-79
typeGL      -0.042950 0.02497042  -1.72003 9.21482e-02
rate         0.085675 0.00428239  20.00634 2.47651e-24
typeGL:rate -0.053950 0.00605622  -8.90820 1.40758e-11
```

```
> anova(lime.reg2)
```

```
Analysis of Variance Table

Response: pH
          Df Sum Sq Mean Sq F value   Pr(>F)
type       1 0.5243  0.5243  142.94 1.03e-15
rate       1 1.3783  1.3783  375.78  < 2e-16
type:rate  1 0.2911  0.2911   79.36 1.41e-11
Residuals 46 0.1687  0.0037
```

h2 (2) Write down the fitted regression lines for the two types of lime.

typeAL: $\hat{y} = 5.781 + 0.0857 \times rate$
typeGL: $\hat{y} = 5.781 - 0.043 + (0.0857 - 0.0540) \times rate = 5.738 + 0.0317 \times rate$.

h3 (2) Test H_0: "two regression lines run parallel". Give test-statistic, P-value and conclusion.

Either F-test or t-test (because it concerns a single parameter):
F-test: TS is $F = ((SSE_{\text{parallel lines}} - SSE_{FM})/1)/MSE_{FM}$ with outcome $F = 79.4$, P-value= $P(F_{1,46} > 79.4) << 0.0001$, so reject H_0: different slopes.
t-test: TS is $t = ((\hat{\beta}_{\text{typeGL:rate}} - 0)/se_{\text{typeGL:rate}})$ with outcome $t = -0.054/0.006 = -8.91$, P-value= $2P(t_{46} < -8.91) << 0.0001$, so reject H_0: different slopes.

h4 (2) Compare the P-values for the t-tests for the individual coefficients, with the P-values for the F-tests in the anova-table. Which P-values are different and which are the same? Explain.

The ANOVA table contains sequential sums of squares for `type`, `rate`, and `type:rate`, in that order. In that case, the F-tests make the follow model comparisons, all concerning single degrees of freedom:
for `type`: `intercept` –> `intercept` + `type`
for `rate`: `intercept` + `type` –> `intercept` + `type` + `rate`
for `type:rate`: `intercept` + `type` + `rate` –> `intercept` + `type` + `rate` + `type:rate`
The t-tests for individual coefficients are partial tests, so that the effect of a variable is judged in presence of all other variables. So, for instance the coefficient of `type` is the difference in intercepts of the two lime types, but in the model with different slopes. So, this concerns the model comparison `intercept` + `rate` + `type:rate` –> `intercept` + `type` + `rate` + `type:rate`, which differs from the F-test done above.
Summarizing, the interaction-test for difference in slopes gives identical P-values for F and t (because it concerns identical model comparisons), but the tests for `type` and `rate` do not.
If all regressors were uncorrelated (which is not the case), would the t-tests and F-test produce identical results here.

h5 (3) Show, using the output above, that the intercepts are not significantly different. This suggests that a model with common intercept is sufficient here. Comment on this model in the light of the principle of marginality.

The P-value from the t-test for comparison of intercepts is 0.092. Testing at $\alpha = 0.05$, we do not reject the null hypothesis of equal intercepts. If we fit a model without `type`, hence forcing the intercepts to be equal, we violate the principle of marginality, because the model contains the interaction (`type:rate`), but only one of the "main effects" (`rate`). Still, this model would make perfect sense.

2 (11). In a study on sleeping behaviour of animals, the sleep duration is related to a number of explanatory variables: log(body weight), log(brain weight), log(lifespan), log(gestation), predation ((1=least likely to be preyed upon, 5=most likely to be preyed upon), and exposure (1=least exposed, 5=most exposed).
We fit a multiple linear regression model, explaining `totalsleep` from all 6 quantitative regressors.

```
> sleep.lm <- lm(totalsleep ~ logbodywgt + logbrainwgt + loglifespan + loggest + exposure + predation,
+ data=sleep)
> summary(sleep.lm)


Call:
lm(formula = totalsleep ~ logbodywgt + logbrainwgt + loglifespan +
    loggest + exposure + predation, data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max
-6.90186 -1.68891 -0.26267  1.61512  6.40439

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 22.053211   3.075389  7.1709 6.427e-09
logbodywgt  -0.063608   0.524492 -0.1213   0.90403
logbrainwgt -0.659306   0.781667 -0.8435   0.40353
loglifespan  0.584594   0.842920  0.6935   0.49162
loggest     -1.682804   0.738533 -2.2786   0.02760
exposure    -0.030104   0.611942 -0.0492   0.96099
predation   -1.208401   0.491140 -2.4604   0.01787

Residual standard error: 3.0496 on 44 degrees of freedom
Multiple R-squared: 0.62624,        Adjusted R-squared: 0.57528
F-statistic: 12.287 on 6 and 44 DF,  p-value: 4.5068e-08
```

a (2) Give the P-value of the "omnibus" F-test and compare it with the P-values of the t-tests for the individual coefficients. Comment on the difference. What could be the reason?

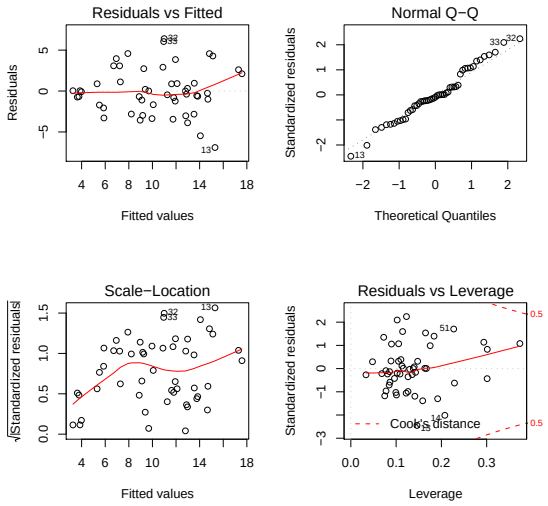
Omnibus F-test: P=4.5e-8. So, H_0 : "all regression coefficients zero" is strongly rejected, there is explanatory power in this model. Individual t-tests: most P-values are large, smallest (0.018) is for `predation`. Individual t-tests tell that effects are absent or just significant. Reason is multicollinearity: t-test measures effect of a regressor, partial to all other regressors. Due to correlation between regressors, an individual regressor is not contributing a lot extra.

b1 (3) Give the assumptions that underly linear models.

In linear models errors ϵ_i are assumed to
1) be independent
2) have constant variance
3) be normally distributed
4) have expectation zero
In short: ϵ_i are i.i.d.,$\sim N(0, \sigma_\epsilon^2)$

b2 (4) Check whether there are problems with those assumptions using the plots and output below. Comment on all residual plots (skip the plot left-below), but also on collinearity (see output next page).

```
> plot(sleep.lm)
```




```
> vif(sleep.lm)
```

```
 logbodywgt logbrainwgt loglifespan     loggest    exposure   predation
 14.8986711  20.9921439   3.5710519   3.1901759   5.0180650   2.8723461
```

*) Plot residuals vs fitted values is used for checking constant variance. There are slight indications that variance increases with mean: variability (in y-direction) seems to increase for larger fitted values.
*) Normal QQ-plot is used for checking normality. Points lay reasonably well on straight line, no indications of deviations from normality.
*) Plot of leverage vs standardized residuals can be used to check for potential influence (leverage), and regression outliers. All standardized residuals are roughly in the range (-2,2), so no problems here. There are 4 observations with leverage exceeding the threshold $2 \times p/n = 2 \times 7/51 = 0.27$. The standardized residuals are not large for these observations, though, does not look problematic.
*) Variables `logbodywgt` and `logbrainwgt` have VIF values exceeding the threshold value 10: each of the two is highly related to one or more other regressors. This will cause problems, like inflated standard errors. In subquestion 2a we already saw the effects of multicollinearity.

c (2) Using stepwise regression a smaller model is obtained. Interpret the first step of the backward elimination procedure below. Which criterion is used for selection and how is it defined? Which variable will be removed from the model after this step?

```
> step(sleep.lm)
```

```
Start:  AIC=120.2
totalsleep ~ logbodywgt + logbrainwgt + loglifespan + loggest +
    exposure + predation

              Df Sum of Sq     RSS     AIC
- exposure     1     0.023 409.237 118.206
- logbodywgt   1     0.137 409.351 118.220
- loglifespan  1     4.473 413.688 118.758
- logbrainwgt  1     6.616 415.831 119.021
<none>                    409.214 120.203
- loggest      1    48.286 457.501 123.892
- predation    1    56.300 465.514 124.777

Step:  AIC=118.21
totalsleep ~ logbodywgt + logbrainwgt + loglifespan + loggest +
    predation
```

```
              Df Sum of Sq     RSS     AIC
- logbodywgt   1     0.188 409.425 116.229
- loglifespan  1     4.867 414.104 116.809
- logbrainwgt  1     6.681 415.918 117.032
<none>                    409.237 118.206
- loggest      1    54.772 464.008 122.612
- predation    1   153.289 562.526 132.431

Step:  AIC=116.23
totalsleep ~ logbrainwgt + loglifespan + loggest + predation

              Df Sum of Sq     RSS     AIC
- loglifespan  1     5.506 414.931 114.911
<none>                    409.425 116.229
- logbrainwgt  1    42.383 451.809 119.253
- loggest      1    54.998 464.424 120.658
- predation    1   157.247 566.673 130.806

Step:  AIC=114.91
totalsleep ~ logbrainwgt + loggest + predation

              Df Sum of Sq     RSS     AIC
<none>                    414.931 114.911
- logbrainwgt  1    40.679 455.610 117.680
- loggest      1    51.611 466.542 118.890
- predation    1   177.374 592.305 131.062

Call:
lm(formula = totalsleep ~ logbrainwgt + loggest + predation,     data = sleep)

Coefficients:
(Intercept)  logbrainwgt      loggest    predation
   23.08575     -0.58799     -1.61560     -1.28089
```

The AIC is used for backward elimination. $AIC = -2\log(\text{likelihood}) + 2p$ with p the number of parameters in the model, smaller values point to better fitting models. Each variable is removed from the model in turn, and the resulting AIC's are reported. For 2 regressors the AIC deteriorates (becomes larger) by removal, so these variables are important. For 4 regressors removal makes the AIC smaller (better), so these regressors are candidates for removal. The largest improvement of AIC is obtained by the removal of `exposure`, so this will be the first variable to be removed.

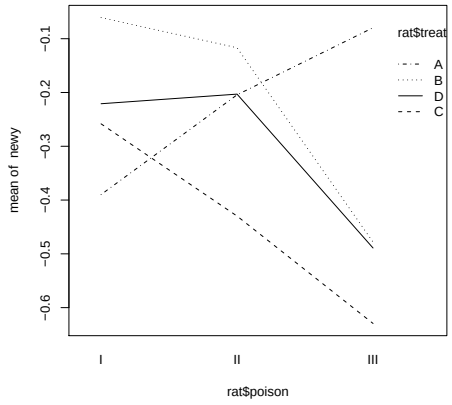
3 (14). In an experiment 48 rats were exposed to a poison (3 types: I, II, III) and a (not further specified) treatment (4 treatments, labeled A, B, C, D). Each combination of poison and treatment was applied to exactly 4 rats, so we have a balanced design. The survival time in tens of hours was measured. Here we analyze the $y = \log(\text{survival time})$ using a two-way ANOVA model.
The two-way ANOVA model with interaction in means notation is $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$, with $i = 1, 2, 3$ for poisons, $j = 1, \ldots, 4$ for treatments, and $k = 1, \ldots, 4$ for replicates; furthermore, $\text{var}(\epsilon_{ijk}) = \sigma^2_\epsilon$.

a (1) Write down this model in effects notation.

$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$

b (3) Explain what it means if factors `poison` and `treatment` interact. Sketch a "profile plot" that would indicate a situation with interaction.

Interaction between `poison` and `treatment` means that the effect of poison is not the same for all four treatments. An example of a profile plot with interaction is given below, showing poison A with different response pattern over the treatments (does not correspond to data in this question):



c (2) Use the output below to test the null-hypothesis of "no interaction". Give H_0, test statistic and outcome, P-value and conclusion.

```
> rat.aov <- aov(y ~ poison + treat + poison*treat, data=rat)
> anova(rat.aov)

Analysis of Variance Table

Response: y
            Df   Sum Sq  Mean Sq  F value     Pr(>F)
poison       2 0.987849 0.493924 48.43244 6.1953e-11
treat        3 0.670925 0.223642 21.92948 2.9868e-08
poison:treat  6 0.074642 0.012440  1.21986    0.31894
Residuals    36 0.367136 0.010198


> tapply(rat$y, rat$poison, mean)

          I          II         III
-0.23243351 -0.31350023 -0.56907825
```

H_0 : "no interaction" $\Leftrightarrow$ $H_0 : \alpha\beta_{11} = \alpha\beta_{12} = \ldots = \alpha\beta_{34}$
Test statistic $F = \frac{SS_{\text{poison:treat}}/6}{MSE}$ with outcome 1.22, and P-value $P(F_{6,36} \geq 1.22) = 0.32$. P-value > 0.05, so do not reject H_0, we cannot find evidence for interaction between poison and treatment.

d1 (1) Estimate the difference in expected log(surval time) ($\mu_{1.} - \mu_{2.}$) between poison types I and II (averaged over the treatments).

$y_{1..} - y_{2..} = -0.2324 - (-0.3135) = 0.081$, so log(survival time) is on average 0.0815 longer for poison I compared to poison II.

d2 (2) To determine whether this difference is significant, the standard error of difference (sed) is needed. Write the sed($y_{1..} - y_{2..}$) as a function of σ_ϵ^2.

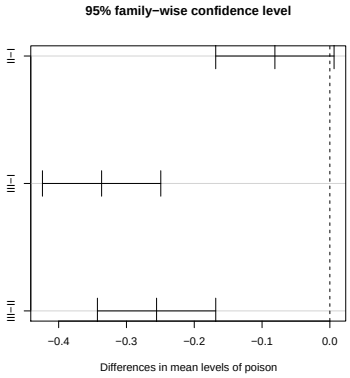
First look at the variance of the difference (sed^2): sed$^2(y_{1..} - y_{2..}) = \sigma_\epsilon^2/16 + \sigma_\epsilon^2/16 = \sigma_\epsilon^2/8$, because 16 rats were used for poisson I and 16 for II. So sed$(y_{1..} - y_{2..}) = \sigma_\epsilon/\sqrt{8}$.

d3 (3) Estimate this standard error, and use it get the Least Significant Difference LSD (using 5% significance level; you may use t-value 2.028). Is the difference significant?

$\widehat{\text{sed}}(y_{1..} - y_{2..}) = \sqrt{MSE}/\sqrt{8} = \sqrt{0.010198}/\sqrt{8} = 0.0357$.
$LSD = t_{36}(0.975) \times \text{sed}(y_{1..} - y_{2..}) = 2.028 \times 0.0357 = 0.0724$. The estimated difference is 0.0815, which is larger than $LSD = 0.0724$, so we would call the difference significant.

d4 (2) Tukey's multiple comparison approach is an alternative. Are poisons I and II significantly different according to the graph below? Compare with d3).

```
> plot(TukeyHSD(rat.aov, "poison"))
```



I and II are not signficantly different according to the Tukey approach, because the Tukey 95% confidence interval for $\mu_{1.} - \mu_{2.}$does contain the value 0.

4 (14). In a study on the research productivity of PhD students in biochemistry, the number of articles published by the student during the last three years (`art`) is related to a number of explanatory variables: number of kids five years or younger (`kid5`), number of articles published by the mentor (`ment`), gender (`fem`). Variable `fem` is an indicator variable for female. We assume an additive model at this moment.

a (2) Write down a generalized linear model that is appropriate for the data we have here. Pay attention to the three components of a glm.
1) Random part of the model: $\texttt{art}_i \sim Poisson(\mu_i)$, $\texttt{art}_i$ independent ($i = 1, \ldots, 915$)
2) Systematic part of the model: linear predictor $\eta_i = \beta_0 + \beta_1 \texttt{kid5}_i + \beta_2 \texttt{ment}_i + \beta_3 \texttt{fem}_i$
3) Link function: $log(\mu_i) = \eta_i$.

b (3) Interpret the coefficients given below (assuming that the model fits well), both in terms of significance as in terms of direction of the effect, and effect on the expected number of articles.

```
> head(articles)
```

```
  fem ment kid5 art
1   0    8    2   3
2   0    7    0   0
3   0   47    0   4
4   0   19    1   1
5   0    0    0   1
6   0    6    1   1
```

```
> articles.glm <- glm(art ~ kid5 + ment + fem, family=poisson(link=log), data=articles)
> summary(articles.glm)
```

```
Call:
glm(formula = art ~ kid5 + ment + fem, family = poisson(link = log),
    data = articles)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.57458  -1.56330  -0.36517   0.55834   5.55137

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  0.4367326  0.0468254  9.3268 < 2.2e-16
kid5        -0.1437870  0.0365734 -3.9315 8.443e-05
ment         0.0256452  0.0019523 13.1361 < 2.2e-16
fem         -0.2431065  0.0542073 -4.4848 7.300e-06

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1817.41  on 914  degrees of freedom
Residual deviance: 1640.85  on 911  degrees of freedom
AIC: 3316.59

Number of Fisher Scoring iterations: 5
```

All three regression coefficients are significantly different from zero, with the strongest effect (in terms of P-value) for `ment` (number of articles published by the mentor).
The higher the number of kids, the lower number of articles. For every extra kid the predicted number of articles has to be multiplied by $e^{-0.144} = 0.866$, so a relative decrease of 100-87%=13%.
The higher the output of the mentor, the larger the number of articles. Every extra article of the mentor leads to a multiplication factor of $e^{0.0256} = 1.026$, so a relative increase of 2.6%.
Females have lower number of articles than males: females produce on average $e^{-0.243} = 0.784$ times the number of articles of males, so a relative decrease of 100-78%=22%.

c (3) Explain the differences in "Analysis of deviance" tables obtained by `anova` and `drop1` statements, given on the next page.

```
> anova(articles.glm, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: art

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                   914    1817.405
kid5   1    4.196       913    1813.210      0.041
ment   1  152.049       912    1661.161 6.1814e-35
fem    1   20.309       911    1640.851 6.5884e-06


> drop1(articles.glm, test = "Chisq")


Single term deletions

Model:
art ~ kid5 + ment + fem
       Df Deviance     AIC     LRT    Pr(Chi)
<none>     1640.85 3316.59
kid5    1  1657.00 3330.74   16.15 5.8470e-05
ment    1  1781.51 3455.25  140.66 < 2.22e-16
fem     1  1661.16 3334.90   20.31 6.5884e-06
```

The `anova` function sequentially adds variables to the model, starting with an intercept-only model. The column labeled 'Deviance' shows the *change* in residual deviance by entering this variable to the model. The residual deviance itself is given in column 'Resid.Dev'.
The `drop1` function drops a single term, always from the full model. Notice that R is not consequent in naming the output, because the column labeled 'Deviance' now shows the residual deviance, corresponding to the column 'Resid. Dev' in the `anova` output. The column 'LRT' contains the differences in deviance, always starting from the full model.
Comparing column 'Deviance' from `anova` with column 'LRT' from `drop1` shows differences, because the model comparisons are different for sequential ("Type I" in SAS terminology) and partial ("Type II") approaches, as we saw earlier in question 1h4. Only for the last term `fem`, the model comparisons are identical, and hence are the deviances (20.31).

d (3) Check the goodness of fit of the model by interpretation of the deviance. What value would you expect if all is well, and how "significant" is the deviation here?

Look at the residual deviance in the full model, which should behave like a random draw from a χ^2_{911} distribution if model fits well. A χ^2_{911} distributed variable has expectation 911, and variance $2 \times 911 = 1822$, or standard deviation 42.7. Here, the residual deviance equals 1640.85 on 911 d.f. Notice that the deviance is much larger than would be expected by change (average value would be 911). Its value of 1640.95 is close to 20 s.d.'s above the mean, so the deviation is highly "significant". Conclusion: goodness of fit is not ok.

e (2) Describe how you would handle the problem (if needed at all), and what the consequences would be for the conclusions in b).

There may be different causes for a high deviance value, like non-fitting systematic part of model, outliers, or clustered responses (like the same mentor having different PhD students). It is not clear what the causes are here. A simple and rather standard approach is to allow for overdispersion by inclusion of an extra scale parameter. All standard errors are multiplied with factor $\sqrt{1640.85/911} = 1.34$ (and instead of χ^2 tests F-tests are done). As the denominators of the Wald test statistics become larger, the Wald test-statistics themselves are closer to zero, and P-values increase. E.g. for `kid5` Wald statistic z = 2.93. The results will remain significant here, though.

f (1) Mention another type of application of Poisson glm's.

Poisson glm's can also be used for analysis of contingency tables.

5 (10). Generalized Linear Models: general theory.

a (2) Write the Poisson distribution as a member of the exponential family.

Let $y \sim Poisson(\lambda)$, then probability distribution is $f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$. The log-likelihood is $LL(\lambda) = \log(\frac{e^{-\lambda}\lambda^y}{y!}) = -\lambda + y\log(\lambda) - \log(y!) = \frac{y\log(\lambda) - \lambda}{1} + (-\log(y!)) = \frac{y\log(\lambda) - \exp(\log(\lambda))}{1} + (-\log(y!))$
Compare this with the general formula for the exponential family: $LL = \frac{y\theta - b(\theta)}{\phi} + c(y, \phi)$.

b (4) Specify using this notation: 1) the canonical parameter; 2) the dispersion parameter; 3) the expectation μ; 4) the variance function $V(\mu)$; 5) the canonical link function.

1) canonical parameter is $\theta = \log(\lambda)$
2) dispersion parameter $\phi = 1$
3) function $b(\theta) = \exp(\theta)$, so that expectation $\mu = b'(\theta) = \frac{d}{d\theta}\exp(\theta) = \exp(\theta) = \exp(\log(\lambda)) = \lambda$
4) variance function $V(\mu) = b''(\theta) = \frac{d^2}{d\theta^2}\exp(\theta) = \exp(\theta) = \lambda$
5) canonical link function is link function which brings the mean to the scale of the canonical parameter; it can be obtained by starting with the expectation $\mu = b'(\theta)$, so that $\theta = (b')^{-1}(\mu)$, and the link function is $(b')^{-1}$. Here, $b'(\theta) = \exp(\theta)$, with inverse log. So, the canonical link function is the log function (natural logarithm).

c (1) Give the name of the estimation principle in glm.

Maximum likelihood estimation.

d (1) Different types of tests are used in glm. Mention at least two.

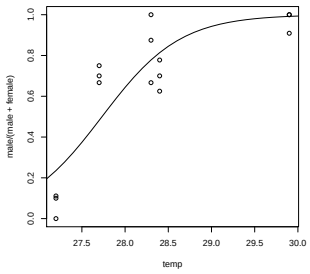
1) Likelihood ratio tests; 2) Wald tests; 3) score tests.

e (2) Explain the (residual) deviance.

The (residual) deviance of a model is twice the difference in maximized log-likelihood of the saturated model and the current model. The saturated model is the model with as many parameters as observations (and as such useless), resulting in perfect predictions. The maximized log-likelihood of this model is the highest possible likelihood obtainable for the given set of data. The deviance is twice the difference between this log-likelihood and that of the current model. In easier words, it tells how much can be gained by making the model larger than the current model. It is like a residual sum of squares, which measures how much the model can explain more by extending the current model until predictions are perfect (in which case the residual sum of squares is zero).

6 (11). In a study of the effect of incubation temperature on the sex of turtles, an experiment was conducted with 5 temperatures. The number of male and female turtles born was recorded. We analyze the data using logistic regression.

```
> glmturtle <- glm(cbind(male,female) ~ temp, family=quasibinomial (link=logit), data=turtle)
> summary(glmturtle)


Call:
glm(formula = cbind(male, female) ~ temp, family = quasibinomial(link = logit),
    data = turtle)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.07212  -1.02915  -0.27137   0.80867   2.55496

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.31832   17.08133 -3.5898 0.003296
temp          2.21103    0.61216  3.6119 0.003159

(Dispersion parameter for quasibinomial family taken to be 2.0186618)

    Null deviance: 74.5080  on 14  degrees of freedom
Residual deviance: 24.9425  on 13  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5


> plot(male/(male+female) ~ temp, data=turtle)
> ld <- data.frame(temp=seq(26.9,30,by=0.1))
> phat <- predict(glmturtle, ld, type="response")
> lines(phat ~ ld$temp)
```



a (2) Write down the fitted model.

$\log(\frac{p}{1-p}) = a + b\,\texttt{temp} = -61.32 + 2.211\,\texttt{temp}$, with p the probability of a male.

b (3) Use the fitted model to find the temperature T75 where 75% of the turtles are males.

For 75% males, logit is $\log(0.75/0.25) = 1.0986$. So, we want to find the temperature $T75$ for which $1.0986 = -61.32 + 2.211\,\texttt{temp}$. It then follows that $T75 = (1.0986 - (-61.32))/2.211 = 28.23°$.

c (4) Give an approximate standard error of T75 using the delta method. Below is the estimated variance-covariance matrix V of the parameters, that you may need.

```
> sigma2 <- sum(residuals(glmturtle,type="pearson")^2)/13
> (V <- sigma2 * summary(glmturtle)$cov.unscaled)
```

```
            (Intercept)      temp
(Intercept)     291.769 -10.45456
temp            -10.455   0.37473
```

T75 $= f(a,b) = (1.0986 - a)/b$, and we need an approximate variance of T75 by the delta method: $V(\text{T75}) \approx (\frac{\partial f(a,b)}{\partial a})^2 v_{11} + (\frac{\partial f(a,b)}{\partial b})^2 v_{22} + 2 \times \frac{\partial f(a,b)}{\partial a} \frac{\partial f(a,b)}{\partial b} v_{12}$, everything evaluated at maximum likelihood estimates.
First the partial derivative w.r.t. a: $\dfrac{\partial(1.0986 - a)/b}{\partial a} = -1/b$, evalutated at mle: $-1/2.21103 = -0.45228$.
Next the partial derivative w.r.t. b: $\dfrac{\partial(1.0986 - a)/b}{\partial b} = -(1.0986 - a)/b^2$, evaluated at mle: $-(1.0986 - (-61.318))/2.211^2 = -12.7677$.
Combining: $V(\text{T75}) \approx (-0.45228)^2 \times 291.769 + (-12.7677)^2 \times 0.37473 + 2 \times -0.45228 \times -12.7677 \times -10.45456 = 59.68345 + 61.0863 - 120.7413 = 0.02845$ and standard error $\sqrt{0.02845} = 0.169$.

d (2) Give an approximate 95% confidence interval for T75, using the approximate standard error from c). (Use standard error 0.15 if you don't have the answer from c).

Approximate 95 % c.i. for T75 is $28.23 \pm 1.96 \times 0.169 = (27.90, 28.56)$.