**Problem 1** A random variable $Y$ belongs to the univariate exponential family if the point mass (pmf) or probability density function (pdf) can be written on the form

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)\right) \tag{1}$$

where $\theta$ is the natural (or canonical) parameter, $b$ and $c$ are functions, and the support of the pmf/pdf does not depend on $\theta$. In the following you can assume that the dispersion parameter $\phi = 1$ and that the weight $w = 1$.

If $X \sim \text{Poisson}(\lambda)$ and $Y$ has the same distribution as $X$ conditional on $X > 0$, then $Y$ is zero-truncated Poisson distributed with point mass function

$$f(y|\lambda) = \frac{1}{1 - e^{-\lambda}} \cdot \frac{\lambda^y e^{-\lambda}}{y!} = \frac{\lambda^y}{y!(e^{\lambda} - 1)}$$

for $y = 1, 2, \ldots$.

**a)** Show that the zero-truncated Poisson distribution belongs to the exponential family and identify $\theta$ and the function $b(\theta)$.

**b)** Using the connection between $E(Y)$ and $b(\theta)$, derive expressions for $E(Y)$ both in terms of $\theta$ and in terms of $\lambda$.

As a sanity check, explain why $E(Y)$ should be asymptotically equal to $\lambda$ as $\lambda$ tends to $\infty$ and check that this is the case by veryfying that

$$\lim_{\lambda \to \infty} \frac{E(Y)}{\lambda} = 1.$$

**Problem 2** In this problem we will model the following data on the number of covid-19 cases (the variable `cases` below) among different school children at 8 different schools in Oslo, Norway, during a particular week in the autumn of 2021. The variable `pupils` is the total number pupils attending each school.

```
coviddata


##   cases pupils
## 1     4    201
## 2    15    304
## 3     7    150
## 4    17    653
```

```
## 5    17    444
## 6    10    200
## 7    59   1120
## 8    16    230
```

We first fit the following generalized linear model to the data.

```
cvd1 <- glm(cases ~ log(pupils), family=poisson(link="log"))
summary(cvd1)
```

```
##
## Call:
## glm(formula = cases ~ log(pupils), family = poisson(link = "log"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4000  -0.8455   0.3938   0.6775   1.8142
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.3562     0.7706  -4.355 1.33e-05 ***
## log(pupils)   1.0368     0.1221   8.494  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 87.942  on 7  degrees of freedom
## Residual deviance: 14.285  on 6  degrees of freedom
## AIC: 53.973
##
## Number of Fisher Scoring iterations: 4
```

**a)** State the model assumptions in suitable mathematical notation. What are the estimated values of the unknown model parameters? Explain why the Poisson assumption may be a reasonable approximation.

**b)** Using a level of significance of $\alpha = 0.05$, test if there is over-dispersion in the data and if so, estimate the dispersion parameter $\varphi$ and adjusted standard errors of the other parameters. Discuss possible mechanisms that may generate overdispersion in this kind of data.

**c)** According to the fitted model, what is the relationship between the expected number of cases and the number pupils attending each school? Suggest a possible way the model can be simplified such that the number of unknown parameters is reduced by 1. Give an interpretation of the parameter(s) of this simplified model. Also conduct a Wald-test of this simpler model against the model fitted above (`cvd1`), again using a significance level of $\alpha = 0.05$.

### Problem 3

In this problem we will use data from Lillard and Panis (2000) on 1060 births to 501 mothers. The first 15 observations are shown below. The outcome of interest is whether each birth was delivered in a hospital or elsewhere encoded as the variable `hosp` being equal to 1 and 0, respectively. The other covariates are the log of the family income (`loginc`), the distance to the nearest hospital in km (`distance`), whether the mother dropped out from school (`dropout`) and maternal education (college or not) after the last child was born (`college`). The variable `mother` (a categorical factor) has a unique value for each mother.

```
head(hosp,15)
```

```
##    hosp   loginc distance dropout college mother
## 1     0 4.330733      1.7       0       1      1
## 2     0 5.616771      7.9       0       0      2
## 3     1 5.298317      1.8       0       0      2
## 4     0 3.850148      6.2       0       0      2
## 5     1 7.417580      1.0       0       0      2
## 6     0 4.553877      4.8       0       0      2
## 7     1 5.278115      1.8       0       1      3
## 8     0 4.382027      3.7       0       0      4
## 9     1 5.129899     10.6       0       0      4
## 10    1 3.931826      3.2       0       0      4
## 11    0 5.247024      3.1       0       0      4
## 12    1 6.990257      0.7       0       0      5
## 13    1 5.771441      3.4       0       0      5
## 14    1 7.705713      0.6       0       0      5
## 15    0 3.610918      2.0       1       0      6
```

We first fit the following generalized linear mixed model to the data.

```
mod1 <- glmmTMB(hosp ~ loginc + distance + dropout + college + (1|mother),
                data=hosp,
                family=binomial(link="logit"),
                REML=FALSE)
summary(mod1)
```

```
##  Family: binomial  ( logit )
## Formula:          hosp ~ loginc + distance + dropout + college + (1 | mother)
## Data: hosp
##
##      AIC      BIC   logLik deviance df.resid
##   1061.2   1091.0   -524.6   1049.2     1054
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  mother (Intercept) 1.251    1.118
## Number of obs: 1060, groups:  mother, 501
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.29454    0.46670  -7.059 1.68e-12 ***
## loginc       0.55040    0.07095   7.758 8.63e-15 ***
## distance    -0.07741    0.03169  -2.443  0.01456 *
## dropout     -1.94731    0.24443  -7.967 1.63e-15 ***
## college      1.02325    0.37261   2.746  0.00603 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**a)** State the model assumptions in suitable mathemathical notation. What are the unknown model parameters and their estimates?

We next fit the following simpler generalized linear model.

```
mod0 <- glm(hosp ~ loginc + distance + dropout + college,
                data=hosp,
                family=binomial(link="logit"))
logLik(mod0)
```

```
## 'log Lik.' -537.4577 (df=5)
```

**b)** Suppose we want to test if the random intercept term in `mod` is statistically significant. Formulate the appropriate null and alternative hypothesis and describe the approximate distribution of a suitable test statistic under $H_0$. Find the critical value if using a level of significance of $\alpha = 0.005$ and compare this with observed value. What is the conclusion of the test?

**c)** Relying on the GLMM `mod1`, for a given mother, give an interpretation of the effect of having a college education on the outcome variable (birth given at the hospital or elsewhere). Is the interpretation exact or approximate?

Will this or a similar interpretation apply also for the average effect of college education on the outcome variable in the population as a whole? Is this interpretation of the average effect exact or approximate?

## Problem 4

In this problem we will consider the chess data you analysed in project 2. Below we have fitted the ordinal regression model assuming that

$$\text{probit}\, P(Y_i \leq r) = \theta_r + \alpha_{j(i)} - \alpha_{k(i)} \tag{2}$$

for $r = 1, 2$ and $i = 1, 2, \ldots, n$, where $Y_i = 1, 2, 3$ if a game ends with a win to white, a draw, and a win to black respectively and $j(i) = 1, 2, \ldots, 6$ and $k(i) = 1, 2, \ldots, 6$ encodes who plays with white and black pieces respectively. In the follow R code, this linear predictor is represented via the variables `friouzja`, `karjakin`, ..., `tari` taking values of 1 and -1 depending on whether these players play with white or black pieces in each particular game. The model have been fitted to all the data and in the following we will for simplicity assume that there is no difference between classic and armageddon games in the probabilities of the different outcomes.

```
mod0 <- vglm(y ~ firouzja + karjakin + nepomniachtchi + rapport + tari,
             cumulative(link="probitlink",parallel = TRUE),
             data=data2)
summary(mod0)


##
## Call:
## vglm(formula = y ~ firouzja + karjakin + nepomniachtchi + rapport +
##     tari, family = cumulative(link = "probitlink", parallel = TRUE),
##     data = data2)
##
```

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   -0.2894     0.2042  -1.417 0.156441
## (Intercept):2    0.8827     0.2297   3.842 0.000122 ***
## firouzja        -0.3593     0.3955  -0.908 0.363634
## karjakin        -0.6507     0.4083  -1.594 0.110953
## nepomniachtchi  -0.5668     0.3704  -1.530 0.125998
## rapport         -0.3183     0.4013  -0.793 0.427667
## tari            -1.0484     0.3901  -2.687 0.007203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: probitlink(P[Y<=1]), probitlink(P[Y<=2])
##
## Residual deviance: 85.8886 on 81 degrees of freedom
##
## Log-likelihood: -42.9443 on 81 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       firouzja       karjakin nepomniachtchi        rapport           tari
##      0.6981572      0.5216631      0.5673641      0.7273973      0.3505024
```

In the following, consider a hypothetical game $i$ between two equally strong players ($\alpha_{j(i)} = \alpha_{k(i)}$).

**a)** Based on the model fitted above, what are the probabilities of a win to white, a draw, and a win to black?

**b)** Explain how equation (2) corresponds to a certain latent variable formulation of the model assumptions.

Explain how the null hypothesis that white and black are equally likely to win is a linear hypothesis of the form

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$$

where $\boldsymbol{\beta} = (\theta_1, \theta_2, \alpha_2, \alpha_3, \ldots, \alpha_6)$. Carry out a test of this hypothesis using a significance level of 0.05. An estimate of variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is given below.

```
round(vcov(mod0),4)
```

```
##                (Intercept):1 (Intercept):2 firouzja karjakin nepomniachtchi
## (Intercept):1         0.0417        0.0176  -0.0028  -0.0002         0.0060
## (Intercept):2         0.0176        0.0528  -0.0094  -0.0124        -0.0047
## firouzja             -0.0028       -0.0094   0.1564   0.0831         0.0756
## karjakin             -0.0002       -0.0124   0.0831   0.1667         0.0798
## nepomniachtchi        0.0060       -0.0047   0.0756   0.0798         0.1372
## rapport              -0.0083       -0.0141   0.0720   0.0898         0.0664
## tari                 -0.0038       -0.0235   0.0634   0.0862         0.0680
##                rapport    tari
## (Intercept):1  -0.0083 -0.0038
## (Intercept):2  -0.0141 -0.0235
## firouzja        0.0720  0.0634
## karjakin        0.0898  0.0862
## nepomniachtchi  0.0664  0.0680
## rapport         0.1610  0.0702
## tari            0.0702  0.1522
```