# Answers generalized linear models

## Exercise 1

```
library(readr)
bonemarrow <- read_csv("bonemarrow.csv", show_col_types = FALSE)
```

```
## New names:
## • `` -> `...1`
```

```
View(bonemarrow)
```

a. Perform a linear regression with age donor as dependent variable and age recipient as independent variable.

```
model.lm <- lm(agedon~agerec,data=bonemarrow)
summary(model.lm)
```

```
##
## Call:
## lm(formula = agedon ~ agerec, data = bonemarrow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.306  -4.231  -0.722   2.652  32.797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.79221    1.09575   2.548   0.0117 *
## agerec       0.90653    0.04416  20.526   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.789 on 164 degrees of freedom
## Multiple R-squared:  0.7198, Adjusted R-squared:  0.7181
## F-statistic: 421.3 on 1 and 164 DF,  p-value: < 2.2e-16
```

The regression coefficient for age is 0.907 and the standard error is 0.044 . The estimate of $\sigma$ is 6.789476.

b. Perform the same linear regression using the glm function.

```
model.lm.glm <- glm(agedon~agerec,family=gaussian, data=bonemarrow)
summary(model.lm.glm)
```

```
##
## Call:
## glm(formula = agedon ~ agerec, family = gaussian, data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.306   -4.231   -0.722    2.652   32.797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.79221    1.09575   2.548   0.0117 *
## agerec       0.90653    0.04416  20.526   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 46.09698)
##
##     Null deviance: 26981.5  on 165  degrees of freedom
## Residual deviance:  7559.9  on 164  degrees of freedom
## AIC: 1111
##
## Number of Fisher Scoring iterations: 2
```

We observe exactly the same estimate of the regression coefficient of agerec, with the same standard error. The estimate of the dispersion parameter $\phi$ = 46.097 can be used to estimate $\sigma$ by calculating sqrt(46.097)= 6.7895. Note that this the same (except for possible rounding errors) as obtained using the lm function.

c. Perform a logistic regression with AGVHD as dependent and AGEREC as independent varaible.

```
model.lr <- glm(agvhd~agerec,family=binomial , data=bonemarrow)
summary(model.lr)
```

```
##
## Call:
## glm(formula = agvhd ~ agerec, family = binomial, data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1952  -0.8939  -0.7417   1.3190   1.7586
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.46369    0.37124  -3.943 8.06e-05 ***
## agerec       0.03136    0.01420   2.208   0.0272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207.94  on 165  degrees of freedom
## Residual deviance: 202.95  on 164  degrees of freedom
## AIC: 206.95
##
## Number of Fisher Scoring iterations: 4
```

The model uses a logit link. Parameters estimates reflect effects on the log-odds scale.

d. Fit a model for a binary outcome with a identity link

```
model.lr.id <- glm(agvhd~agerec,family=binomial(link = "identity"), data=bonemarrow)
summary(model.lr.id)
```

```
##
## Call:
## glm(formula = agvhd ~ agerec, family = binomial(link = "identity"),
##     data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2058  -0.9014  -0.7182   1.3043   1.8247
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.151175   0.067947   2.225  0.02609 *
## agerec      0.007613   0.002935   2.594  0.00948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207.94  on 165  degrees of freedom
## Residual deviance: 202.32  on 164  degrees of freedom
## AIC: 206.32
##
## Number of Fisher Scoring iterations: 5
```

Here $\pi_i$ is modelled as a linear function of $\beta$: $\pi_i = \beta_0 + \beta_1 x_{i1}$. The coefficients reflect change in risk. An increase of one year in age of the recipient increases the expected risk of agvhd with 0.008.

# Exercise 2

```
library(readr)
pill <- read_csv("pill.csv", show_col_types = FALSE)
View(pill)
```

a. Calculate the expected number of thrombosis cases per month separately for second and third generation pill.

```
#second generation
sum(pill$users[pill$type_pill ==2])
```

```
## [1] 30033222
```

```
sum(pill$thrombosis[pill$type_pill ==2])
```

```
## [1] 1367
```

```
# third generation
sum(pill$users[pill$type_pill ==3])
```

```
## [1] 4579452
```

```
sum(pill$thrombosis[pill$type_pill ==3])
```

```
## [1] 343
```

b. Calculate the expected number of thrombosis cases per month separately for second and third generation pill.

```
sum(pill$thrombosis[pill$type_pill ==2])/sum(pill$users[pill$type_pill ==2])
```

```
## [1] 4.551626e-05
```

```
sum(pill$thrombosis[pill$type_pill ==3])/sum(pill$users[pill$type_pill ==3])
```

```
## [1] 7.489979e-05
```

c. A Poisson model with type pill as factor and log(users as offset.

```
model.pois <- glm(thrombosis~as.factor(type_pill)+offset(log(users)) ,family=poisson , data=p
ill)
summary(model.pois)
```

```
##
## Call:
## glm(formula = thrombosis ~ as.factor(type_pill) + offset(log(users)),
##     family = poisson, data = pill)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3016  -1.3071  -0.4743   0.8608   7.2835
##
## Coefficients:
##                       Estimate Std. Error  z value Pr(>|z|)
## (Intercept)           -9.99744    0.02705 -369.645   <2e-16 ***
## as.factor(type_pill)3  0.49808    0.06039    8.248   <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: 2418.1  on 887  degrees of freedom
## Residual deviance: 2356.7  on 886  degrees of freedom
## AIC: 3955.3
##
## Number of Fisher Scoring iterations: 5
```

The model is log(expected count) = $\beta_0 + \beta_1 * typepill + log(users)$. The thrombosis rate per month is $exp(\beta_0 + \beta_1 * typepill)$, which is exp(-9.997) = 4.552^{-5} for the second generation pill and exp(-9.997+ 0.498) = 7.49^{-5} for the third generation pill. Same results as found in 2.b.The rate ratio on thrombosis for the third versus the second generation pill is now exp(0.498)= 1.646.

d . Age and month added to the model

```
model.pois2 <- glm(thrombosis~as.factor(type_pill)+age+ month + offset(log(users)) ,family=po
isson , data=pill)
summary(model.pois2)
```

```
##
## Call:
## glm(formula = thrombosis ~ as.factor(type_pill) + age + month +
##     offset(log(users)), family = poisson, data = pill)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8868  -0.8568  -0.2123   0.4325   3.6680
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## (Intercept)           -12.902753   0.106346 -121.328  < 2e-16 ***
## as.factor(type_pill)3   0.175066   0.060538    2.892  0.00383 **
## age                     0.093567   0.002605   35.914  < 2e-16 ***
## month                  -0.010222   0.007247   -1.410  0.15840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2418.08  on 887  degrees of freedom
## Residual deviance:  941.77  on 884  degrees of freedom
## AIC: 2544.4
##
## Number of Fisher Scoring iterations: 5
```

There is a strong effect of age, the risk of thrombosis increases with age (very small p-value). The rate ratio for the third versus the second generation pill is now exp(0.175)= 1.191.

f. The deviance of the model is 941.77, we can compare the value to a chi-square distribution with 884 degrees of freedom. This yields a p-value of 0.09. No strong indication for lack of fit.

g Model with overdispersion

```
model.pois.disp <- glm(thrombosis~as.factor(type_pill)+age+ month + offset(log(users)) ,famil
y=quasipoisson , data=pill)
summary(model.pois.disp)
```

```
##
## Call:
## glm(formula = thrombosis ~ as.factor(type_pill) + age + month +
##     offset(log(users)), family = quasipoisson, data = pill)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8868  -0.8568  -0.2123   0.4325   3.6680
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           -12.902753   0.113254 -113.928  < 2e-16 ***
## as.factor(type_pill)3   0.175066   0.064470    2.715  0.00675 **
## age                     0.093567   0.002775   33.723  < 2e-16 ***
## month                  -0.010222   0.007718   -1.324  0.18570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.134138)
##
##     Null deviance: 2418.08  on 887  degrees of freedom
## Residual deviance:  941.77  on 884  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

The parameter for overdispersion is 1.13 which is close to the value 1. We can conclude that there is no substantial overdispersion in these data.