Answers logistic regression, day 2

Exercise 1

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

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

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

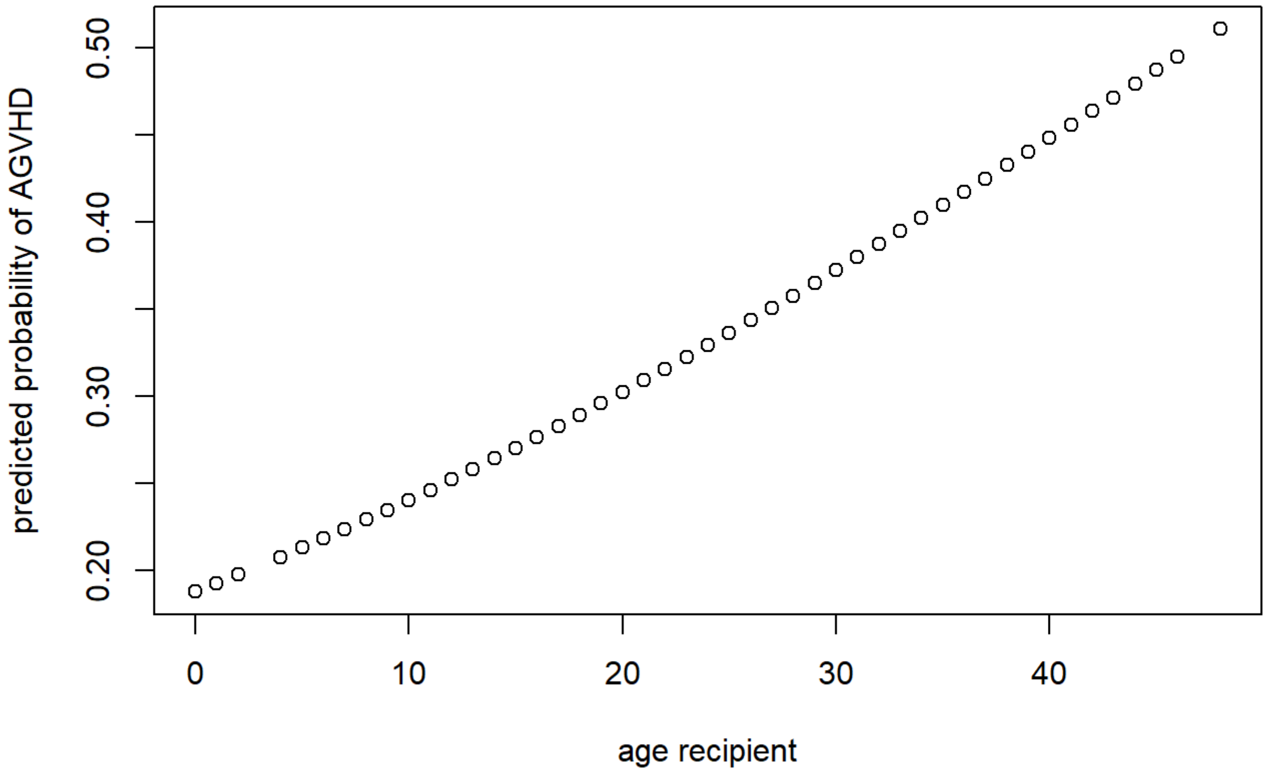
```
View(bonemarrow)
```

a Start performing a logistic regression analysis with agvhd as dependent and agerec as independent variable. Plot the predicted probabilities on the y-axis and the age of the recipient on the x-axis. What do you see?

```
model.lr1 <- glm(agvhd~agerec, family=binomial, data=bonemarrow)
summary(model.lr1 )
```

```
## 
## 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
```

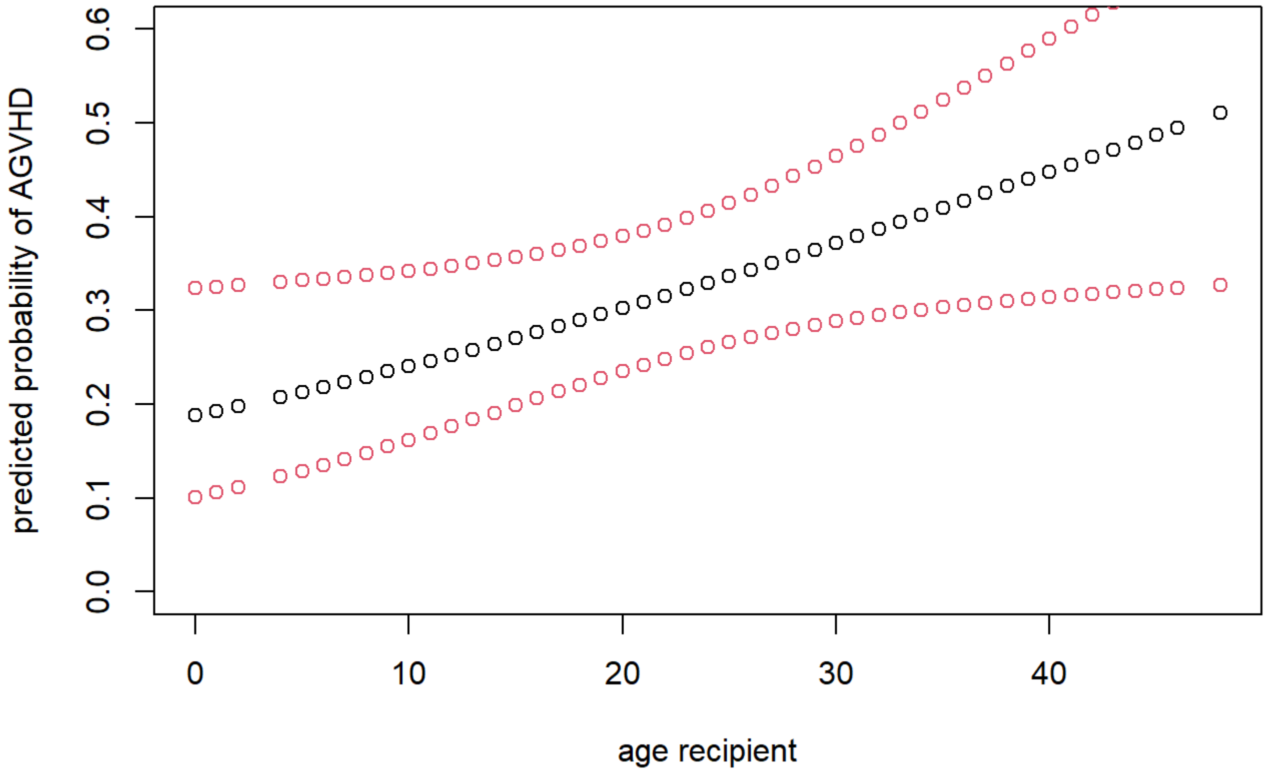
```
pred <- predict(model.lr1, type="response")
plot(bonemarrow$agerec, pred, xlab="age recipient", ylab="predicted probability of AGVHD")
```



We observe that the probability of AGVHD increases with age from about 0.20 at age 0 to about 0.50 at age 50.

b. Calculate a 95% confidence bound around the predicted probabilities.

```
# first calculate logit with standard error
logit <- predict(model.lr1, se.fit = TRUE)
# lowerbound 95% confidence interval logit
logit.lwb <- logit$fit-1.96*logit$se.fit
# upperbound 95% confidence interval logit
logit.upb <- logit$fit+1.96*logit$se.fit
# lowerbound 95% confidence interval pred
pred.lwb <- exp(logit.lwb)/(1+exp(logit.lwb))
# upperbound 95% confidence interval pred
pred.upb <- exp(logit.upb)/(1+exp(logit.upb))
plot(bonemarrow$agerec, pred, xlab="age recipient", ylab="predicted probability of AGVHD", yl
im=c(0,0.60))
points(bonemarrow$agerec,pred.lwb, col=2)
points(bonemarrow$agerec,pred.upb, col=2)
```

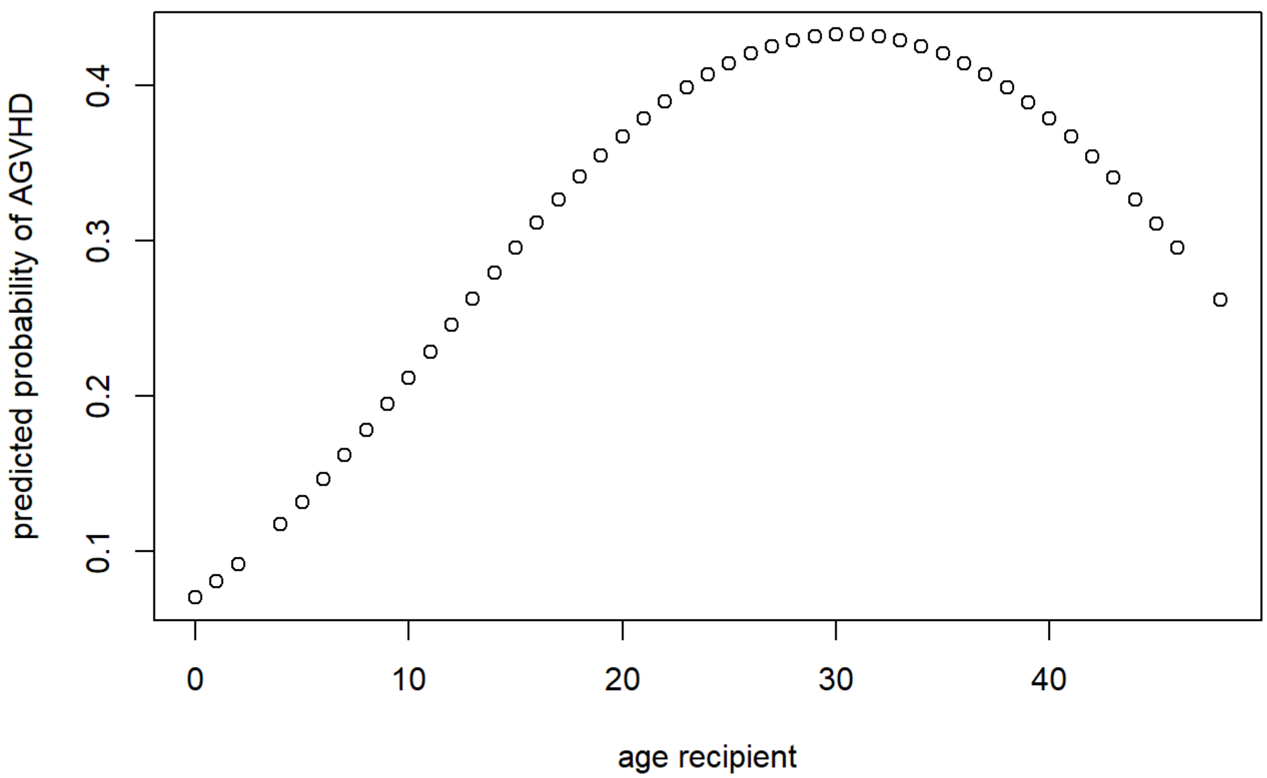


c. Add a quadratic term to the model

```
bonemarrow$agerec2 <- bonemarrow$agerec**2
model.lr2 <- glm(agvhd~agerec+agerec2, family=binomial, data=bonemarrow)
summary(model.lr2)
```

```
##
## Call:
## glm(formula = agvhd ~ agerec + agerec2, family = binomial, data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0644  -0.9755  -0.6890   1.3248   2.0154
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.587789   0.751641  -3.443 0.000576 ***
## agerec       0.151981   0.067905   2.238 0.025212 *
## agerec2     -0.002493   0.001355  -1.839 0.065910 .
## ---
## 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: 199.32  on 163  degrees of freedom
## AIC: 205.32
##
## Number of Fisher Scoring iterations: 4
```

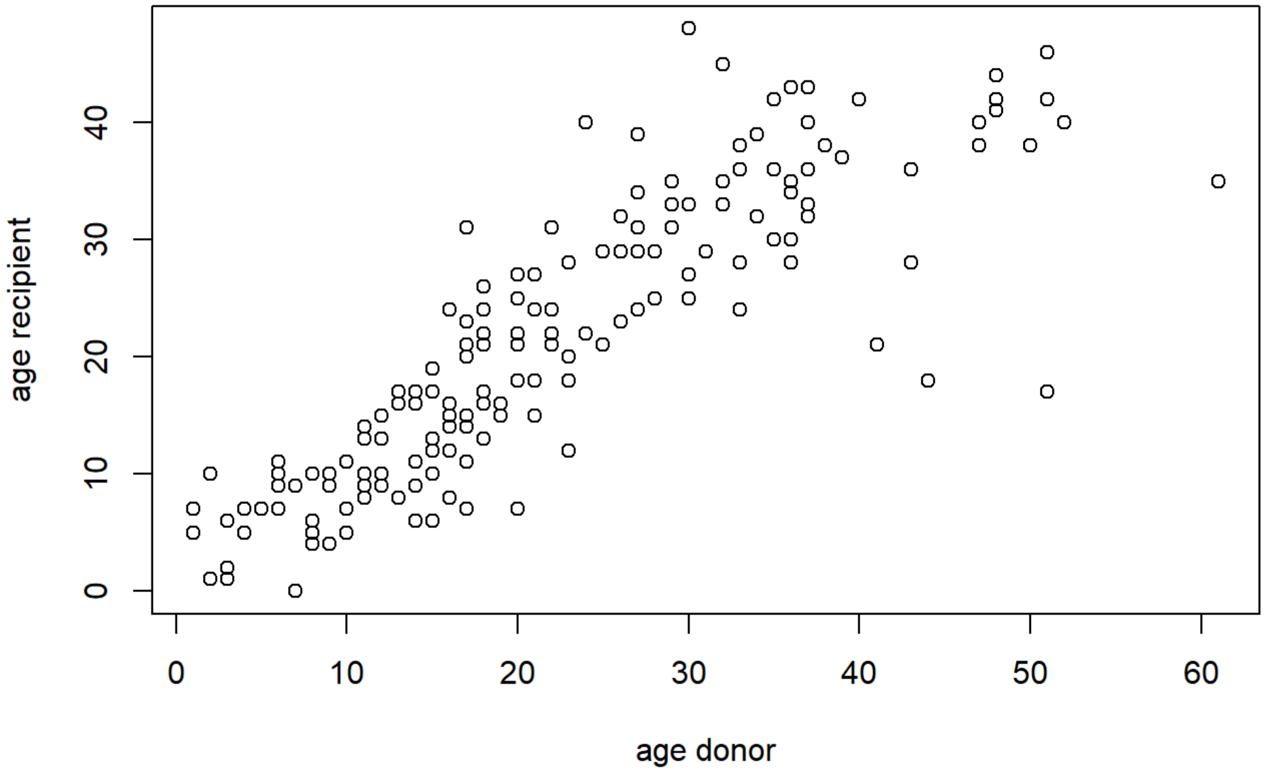
```
pred2 <- predict(model.lr2, type="response")
plot(bonemarrow$agerec, pred2, xlab="age recipient", ylab="predicted probability of AGVHD")
```



The quadratic term in age recipient (agerec2) is just not statistically significant (p=0.066). Not statistically significant does not apply that the linear model fits perfectly. If we consider the plot of the predicted probabilities of this model, it suggest that it may be that the risk increases until adulthood and then remains more or less constant. Modelling with splines or other non linear models would be a next step to explore this further.

d. A plot of agedon versus agerec.

```
plot(bonemarrow$agedon,bonemarrow$agerec, ylab="age recipient", xlab="age donor")
```



A strong correlation between age donor and age recipient

e. model with age donor and age recipient

```
model.lr3 <- glm(agvhd~agerec+agedon, family=binomial, data=bonemarrow)
summary(model.lr3)
```

```
## 
## Call:
## glm(formula = agvhd ~ agerec + agedon, family = binomial, data = bonemarrow)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.3636  -0.8606  -0.7387   1.3075   1.7896  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.555822   0.380882  -4.085 4.41e-05 ***
## agerec       0.004143   0.025938   0.160    0.873    
## agedon       0.030141   0.024058   1.253    0.210    
## ---
## 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: 201.38  on 163  degrees of freedom
## AIC: 207.38
## 
## Number of Fisher Scoring iterations: 4
```

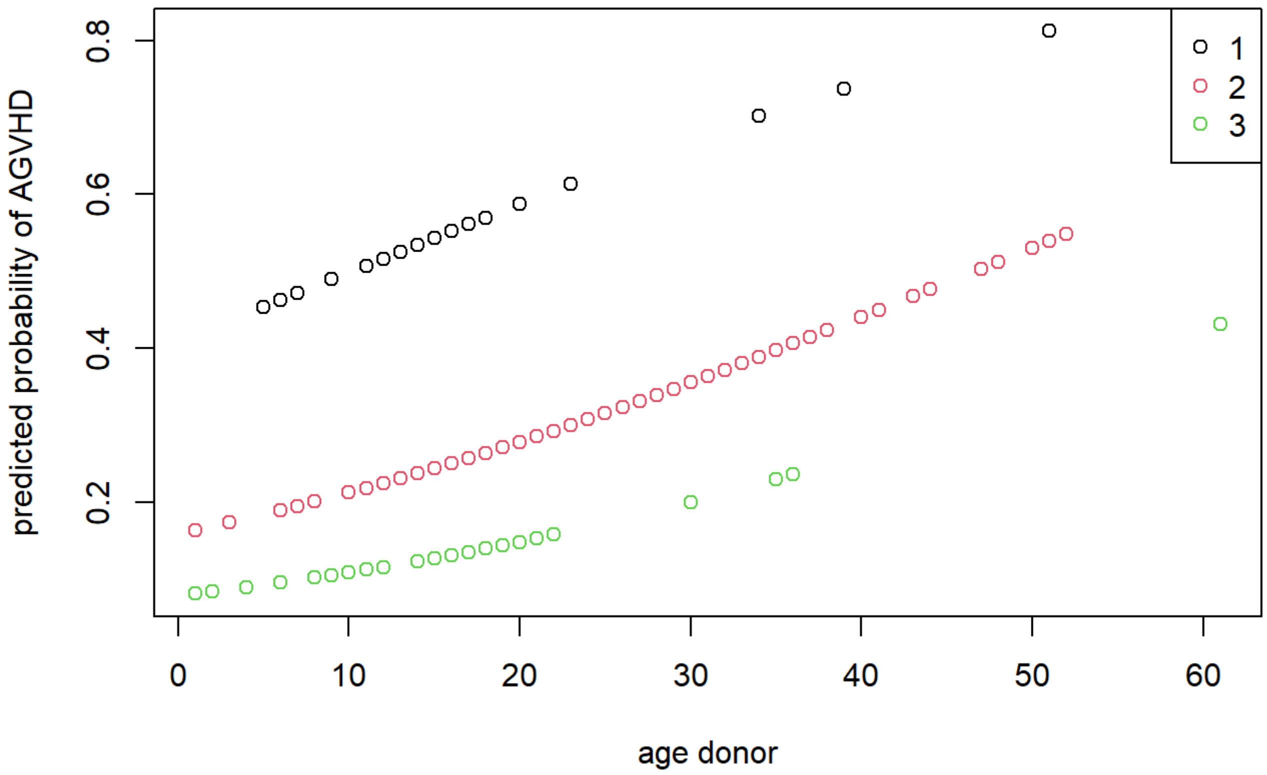
After adding agedon to the model the regression coefficient for agerec is very small and is no longer statistically different from 0. This can be explained by the fact that the two age variables are very correlated.

f. A model with AGEDON and the diagnosis (as categorical variable).

```
model.lr4 <- glm(agvhd~agedon + as.factor(diag), family=binomial, data=bonemarrow)
summary(model.lr4)
```

```
##
## Call:
## glm(formula = agvhd ~ agedon + as.factor(diag), family = binomial,
##     data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5547  -0.8934  -0.6165   1.1178   2.1386
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.36525    0.49537  -0.737  0.46093
## agedon            0.03586    0.01480   2.422  0.01543 *
## as.factor(diag)2 -1.30641    0.50182  -2.603  0.00923 **
## as.factor(diag)3 -2.10133    0.64924  -3.237  0.00121 **
## ---
## 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: 189.22  on 162  degrees of freedom
## AIC: 197.22
##
## Number of Fisher Scoring iterations: 4
```

```
pred4 <- predict(model.lr4, type="response")
plot(bonemarrow$agedon, pred4, xlab="age donor", ylab="predicted probability of AGVHD", col=b
onemarrow$diag)
# Adding a legend
legend("topright", legend = unique(bonemarrow$diag), col = unique(bonemarrow$diag), pch = 1)
```



The regression output indicates that patients with diag=1 have a higher risk to develop AGVHD than diag =2 (because the regression coefficient for diag=2 is negative), the difference is even larger for patients with diag=3. This can also been seen in the plot.

g. Add sex donor and sex recipient to the model.

```
model.lr5 <- glm(agvhd~agedon+ as.factor(diag) +sexdon + sexrec, family=binomial, data=bonema
rrow)
summary(model.lr5)
```

```
##
## Call:
## glm(formula = agvhd ~ agedon + as.factor(diag) + sexdon + sexrec,
##     family = binomial, data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5088  -0.8816  -0.6181   1.1359   2.2114
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.37903    0.52284  -0.725  0.46848
## agedon            0.03635    0.01495   2.431  0.01505 *
## as.factor(diag)2 -1.41051    0.52662  -2.678  0.00740 **
## as.factor(diag)3 -2.16142    0.66309  -3.260  0.00112 **
## sexdon           -0.10469    0.35887  -0.292  0.77049
## sexrec            0.25860    0.36945   0.700  0.48395
## ---
## 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: 188.65  on 160  degrees of freedom
## AIC: 200.65
##
## Number of Fisher Scoring iterations: 4
```

```
# odds ratios + 95% CIs
cbind(exp(coefficients(model.lr5)), exp(confint(model.lr5)))
```

```
## Waiting for profiling to be done...
```

```
##                                 2.5 %    97.5 %
## (Intercept)      0.6845261 0.24395654 1.9302198
## agedon           1.0370145 1.00751177 1.0687766
## as.factor(diag)2 0.2440193 0.08387694 0.6715422
## as.factor(diag)3 0.1151617 0.02886741 0.4002322
## sexdon           0.9006013 0.44306314 1.8188203
## sexrec           1.2951125 0.63084068 2.7030914
```

```
# Likelihood ratio test
lr_stat <-   -2 * (logLik(model.lr4) - logLik(model.lr5))
#Calculate the difference in number of parameters.
df <- df.residual(model.lr4) - df.residual(model.lr5)
#Calculate the p-value using the chi-square distribution:
p_value <- 1 - pchisq(lr_stat, df)
```

We observe that both for sexdon and sexrec the adjusted odds ratios are close to 1, p-values for both variables are large.The deviance is 188.65, number of fitted parameters is 6. That yields as AIC 188.65 + 2× 6= 200.65. That value is also given in the output.

h. Compare this model to the model without sexdon and sexrec.

```
# Likelihood ratio test
lr_stat <-   -2 * (logLik(model.lr4) - logLik(model.lr5))
#Calculate the difference in number of parameters.
df <- df.residual(model.lr4) - df.residual(model.lr5)
#Calculate the p-value using the chi-square distribution:
p_value <- 1 - pchisq(lr_stat, df)
```

The deviance of the model without sexdon and sexrec (189.22) is close to the log-likelihood of the extended model 188.65 . The p-value of the likelihood ratio test is 0.75, .

i. Add an interaction term between sex donor and sex recipient to the model.

```
model.lr6 <- glm(agvhd~agedon+ as.factor(diag) +sexdon + sexrec + sexdon*sexrec, family=binom
ial, data=bonemarrow)
summary(model.lr6)
```

```
##
## Call:
## glm(formula = agvhd ~ agedon + as.factor(diag) + sexdon + sexrec +
##     sexdon * sexrec, family = binomial, data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6932  -0.8557  -0.5808   1.0310   2.2486
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.70008    0.55629  -1.258  0.20821
## agedon            0.03672    0.01515   2.423  0.01538 *
## as.factor(diag)2 -1.43816    0.53595  -2.683  0.00729 **
## as.factor(diag)3 -2.20872    0.67311  -3.281  0.00103 **
## sexdon            0.61243    0.51447   1.190  0.23389
## sexrec            0.90014    0.50119   1.796  0.07249 .
## sexdon:sexrec    -1.41596    0.72550  -1.952  0.05098 .
## ---
## 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: 184.75  on 159  degrees of freedom
## AIC: 198.75
##
## Number of Fisher Scoring iterations: 4
```

The p-value for the interaction term is borderline significant. Compared to the reference category (sexdon = male, sexrec=male), the odds of AGVHD is exp(0.612) = 1.845 times larger when sexdon = female and sexrec=male. For sexdon = male, sexrec=female, the odds is exp(0.9) = 2.46 times larger. For sexdon = female, sexrec=female, the odds is exp(0.612 + 0.9+ -1.416) = 1.101 time larger. This shows that when there is a sex mismatch between donor and recipient that the odds of AGVHD is larger.

j. Calculate a mismatch variable.

```
bonemarrow$mismatch <- abs(bonemarrow$sexdon-bonemarrow$sexrec)
table(bonemarrow$mismatch)
```

```
## 
##  0  1 
## 84 82
```

```
model.lr7 <- glm(agvhd~agedon+ as.factor(diag) +mismatch, family=binomial, data=bonemarrow)
summary(model.lr7)
```

```
## 
## Call:
## glm(formula = agvhd ~ agedon + as.factor(diag) + mismatch, family = binomial, 
##     data = bonemarrow)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7336  -0.8604  -0.5895   1.0496   2.2769  
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)   
## (Intercept)      -0.70224    0.53104  -1.322  0.18604   
## agedon            0.03617    0.01503   2.408  0.01606 * 
## as.factor(diag)2 -1.35782    0.51219  -2.651  0.00803 **
## as.factor(diag)3 -2.17382    0.66223  -3.283  0.00103 **
## mismatch          0.72336    0.36105   2.003  0.04513 * 
## ---
## 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: 185.11  on 161  degrees of freedom
## AIC: 195.11
## 
## Number of Fisher Scoring iterations: 4
```

When there is a sex-mismatch between donor and recipient, the odds of AGVHD is exp(0.723) = 2.061 times larger than when donor and recipient are of the same sex.

k.Perform a stepwise selection procedure

```
initial_model.lr <- glm(agvhd~agedon+ agerec+ as.factor(diag) +sexdon+sexrec+mismatch, family
=binomial, data=bonemarrow)
stepwise_model.lr <- step(initial_model.lr, direction = "both")
```

```
## Start:  AIC=200.28
## agvhd ~ agedon + agerec + as.factor(diag) + sexdon + sexrec + 
##     mismatch
## 
##                   Df Deviance    AIC
## - sexdon           1   184.40 198.40
## - sexrec           1   184.59 198.59
## - agerec           1   184.75 198.75
## - agedon           1   185.04 199.04
## <none>                 184.28 200.28
## - mismatch         1   188.26 202.26
## - as.factor(diag)  2   197.34 209.34
## 
## Step:  AIC=198.4
## agvhd ~ agedon + agerec + as.factor(diag) + sexrec + mismatch
## 
##                   Df Deviance    AIC
## - sexrec           1   184.75 196.75
## - agerec           1   184.82 196.82
## - agedon           1   185.21 197.21
## <none>                 184.40 198.40
## + sexdon           1   184.28 200.28
## - mismatch         1   188.40 200.40
## - as.factor(diag)  2   197.66 207.66
## 
## Step:  AIC=196.75
## agvhd ~ agedon + agerec + as.factor(diag) + mismatch
## 
##                   Df Deviance    AIC
## - agerec           1   185.11 195.11
## - agedon           1   185.63 195.63
## <none>                 184.75 196.75
## + sexrec           1   184.40 198.40
## + sexdon           1   184.59 198.59
## - mismatch         1   188.94 198.94
## - as.factor(diag)  2   197.67 205.67
## 
## Step:  AIC=195.11
## agvhd ~ agedon + as.factor(diag) + mismatch
## 
##                   Df Deviance    AIC
## <none>                 185.11 195.11
## + agerec           1   184.75 196.75
## + sexrec           1   184.82 196.82
## + sexdon           1   185.01 197.01
## - mismatch         1   189.22 197.22
## - agedon           1   191.15 199.15
## - as.factor(diag)  2   197.71 203.71
```

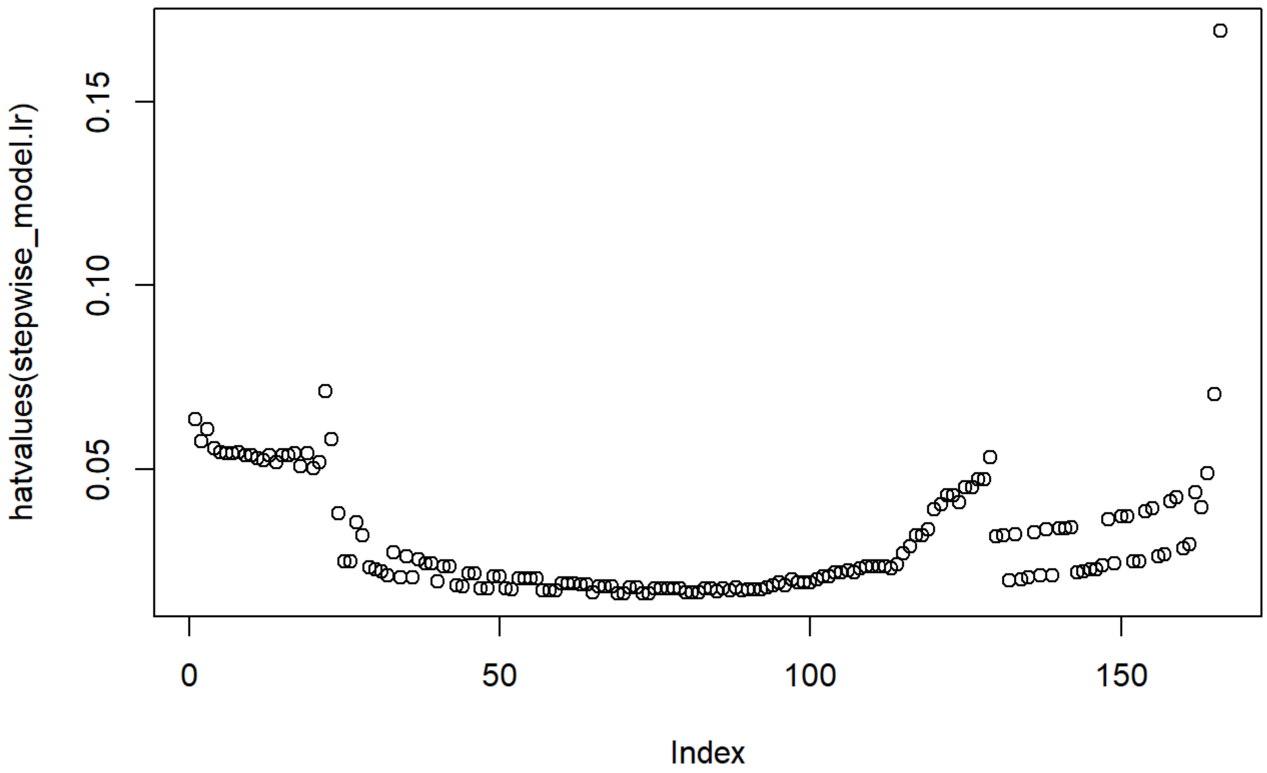
```
summary(stepwise_model.lr)
```

```
##
## Call:
## glm(formula = agvhd ~ agedon + as.factor(diag) + mismatch, family = binomial,
##     data = bonemarrow)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7336  -0.8604  -0.5895   1.0496   2.2769
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.70224    0.53104  -1.322  0.18604
## agedon            0.03617    0.01503   2.408  0.01606 *
## as.factor(diag)2 -1.35782    0.51219  -2.651  0.00803 **
## as.factor(diag)3 -2.17382    0.66223  -3.283  0.00103 **
## mismatch          0.72336    0.36105   2.003  0.04513 *
## ---
## 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: 185.11  on 161  degrees of freedom
## AIC: 195.11
##
## Number of Fisher Scoring iterations: 4
```

The AIC was used as selection criterion. The final model has agedon, diagnososis and mismatch as variables in the model.

l. Check if there any points with high leverage.

```
plot(hatvalues(stepwise_model.lr))
```



```
bonemarrow$hatvalues <- hatvalues(stepwise_model.lr)
# print information of the observation with the largest hatvalue
print(bonemarrow[bonemarrow$hatvalues>0.10, ] )
```

```
## # A tibble: 1 x 10
##     ...1 agvhd  diag agedon agerec sexdon sexrec agerec2 mismatch hatvalues
##    <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>    <dbl>     <dbl>
## 1   166     0     3     61     35      1      0    1225        1     0.169
```

```
    We observe one observation with a very high leverage. Exploring the data shows that the d
onor of this case is 61 years old, much older than the other donors. The next step would be e
xplore how influential this observation is by leaving it out and see how much the estimates o
f the parameters change.
```

Exercise 2

a. Looking at the code you can see that the probability that Y=1 is generated by probY <- 0.01 + 0.05*X. This implies that the true risk for X=0, P(Y=1|X=0) = 0.01 and for X=1, P(Y=1|X=1)= 0.06. This yields as true risk difference: RD=0.06-0.01= 0.05, true risk ratio RR=0.06/0.01 = 6 and odds ratio OR= (0.06/0.94)/ (0.01/0.99)=6.32

b. run the code and calculate observed RD RR and OR

```
nsim<-100000
set.seed(2468)
probX <- 0.5
X <- rbinom(n=nsim, size=1,prob=probX)
probY <- 0.01 + 0.05*X
Y<-rbinom(n=nsim, size=1, prob=probY)
out <- cbind(X,Y)
cohort<-data.frame(out)

# table
xtabs(~X+Y, data=cohort)
```

```
##    Y
## X       0     1
##   0 49589   507
##   1 46921  2983
```

```
pY1 <- mean(cohort$Y[cohort$X==1])
pY0 <- mean(cohort$Y[cohort$X==0])
# risk difference
pY1-pY0
```

```
## [1] 0.0496542
```

```
# Relative risk
pY1/pY0
```

```
## [1] 5.906266
```

```
# Oddsratio
(pY1/(1-pY1))/(pY0/(1-pY0))
```

```
## [1] 6.218181
```

The estimated values are close to the true values

c. Perform outcome dependent sampling (a case control study)

```
# Select all cases
cases <- cohort[cohort$Y ==1, ]
# and  1% of controls
select <- rbinom(n=nsim, size=1, prob=0.01)
controls <- cohort[cohort$Y ==0 & select==1, ]
case.control <- rbind(cases,controls)
```

d. And calculate RD RR and OR in the generated case-control data

```
# table
xtabs(~X+Y, data=case.control)
```

```
##    Y
## X      0    1
##   0  495  507
##   1  465 2983
```

```
# risk difference
pY1 <- mean(case.control$Y[case.control$X==1])
pY0 <- mean(case.control$Y[case.control$X==0])
pY1-pY0
```

```
## [1] 0.3591512
```

```
# Relative risk
pY1/pY0
```

```
## [1] 1.709802
```

```
# Oddsratio
(pY1/(1-pY1))/(pY0/(1-pY0))
```

```
## [1] 6.263218
```

We observe that risk difference and relative risk are far away from the true value, the odds ratio is still close to the true values, as it is invariant under the sampling scheme.