GR5291_FINAL

Nguyen Thuy Linh, tn2382

5/2/2019

```
library(survival)
library(KMsurv)
library(MASS)
library(nnet)
library('arm')
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /Users/Linh/Desktop/GR5291/Final Project
```

```
heart <- read.csv("Framingham Heart Data.csv", header = TRUE, na.strings=c("","NA"))
summary(heart)
```

```
##    Status                         DeathCause     AgeCHDdiag
##  Alive:3218   Cancer                   : 539   Min.   :32.0
##  Dead :1991   Cerebral Vascular Disease: 378   1st Qu.:57.0
##               Coronary Heart Disease   : 605   Median :63.0
##               Other                    : 357   Mean   :63.3
##               Unknown                  : 112   3rd Qu.:70.0
##               NA's                     :3218   Max.   :90.0
##                                                NA's   :3760
##      Sex         AgeAtStart        Height          Weight
##  Female:2873   Min.   :28.00   Min.   :51.50   Min.   : 67.0
##  Male  :2336   1st Qu.:37.00   1st Qu.:62.25   1st Qu.:132.0
##                Median :43.00   Median :64.50   Median :150.0
##                Mean   :44.07   Mean   :64.81   Mean   :153.1
##                3rd Qu.:51.00   3rd Qu.:67.50   3rd Qu.:172.0
##                Max.   :62.00   Max.   :76.50   Max.   :300.0
##                                NA's   :6       NA's   :6
##    Diastolic         Systolic
##  Min.   : 50.00   Min.   : 82.0
##  1st Qu.: 76.00   1st Qu.:120.0
##  Median : 84.00   Median :132.0
##  Mean   : 85.36   Mean   :136.9
##  3rd Qu.: 92.00   3rd Qu.:148.0
##  Max.   :160.00   Max.   :300.0
##
##  MRW..Metropolitan.Relative.desired..Weight.    Smoking
##  Min.   : 67                                 Min.   : 0.000
##  1st Qu.:106                                 1st Qu.: 0.000
##  Median :118                                 Median : 1.000
##  Mean   :120                                 Mean   : 9.367
##  3rd Qu.:131                                 3rd Qu.:20.000
##  Max.   :268                                 Max.   :60.000
```

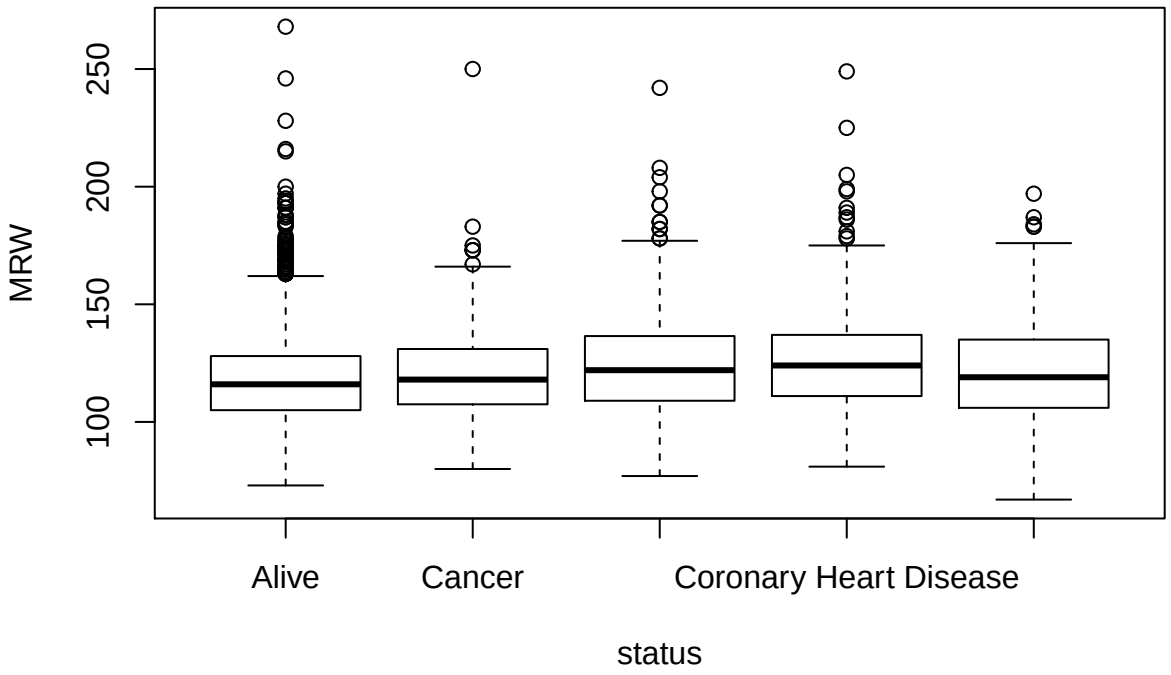
```
##  NA's   :6                                      NA's   :36
##    AgeAtDeath     Cholesterol        Chol_Status     BP_Status
##  Min.   :36.00   Min.   : 96.0   Borderline:1861   High   :2267
##  1st Qu.:63.00   1st Qu.:196.0   Desirable :1405   Normal :2143
##  Median :71.00   Median :223.0   High      :1791   Optimal: 799
##  Mean   :70.54   Mean   :227.4   NA's      : 152
##  3rd Qu.:79.00   3rd Qu.:255.0
##  Max.   :93.00   Max.   :568.0
##  NA's   :3218    NA's   :152
##      Weight_Status            Smoking_Status
##  Normal     :1472   Heavy (16-25)    :1046
##  Overweight :3550   Light (1-5)      : 579
##  Underweight: 181   Moderate (6-15)  : 576
##  NA's       :   6   Non-smoker       :2501
##                     Very Heavy (> 25): 471
##                     NA's             :  36
##
```

```
heart$status <- ifelse(is.na(heart$DeathCause), "Alive", as.character(heart$DeathCause))
heart <- heart[heart$status != "Unknown",]
heart$status <- as.factor(heart$status)
names(heart)[10] <- "MRW"
heart2 <- heart[,c(18, 4,5,6,7,8,9,10,11,13,14,15,16,17)]
```

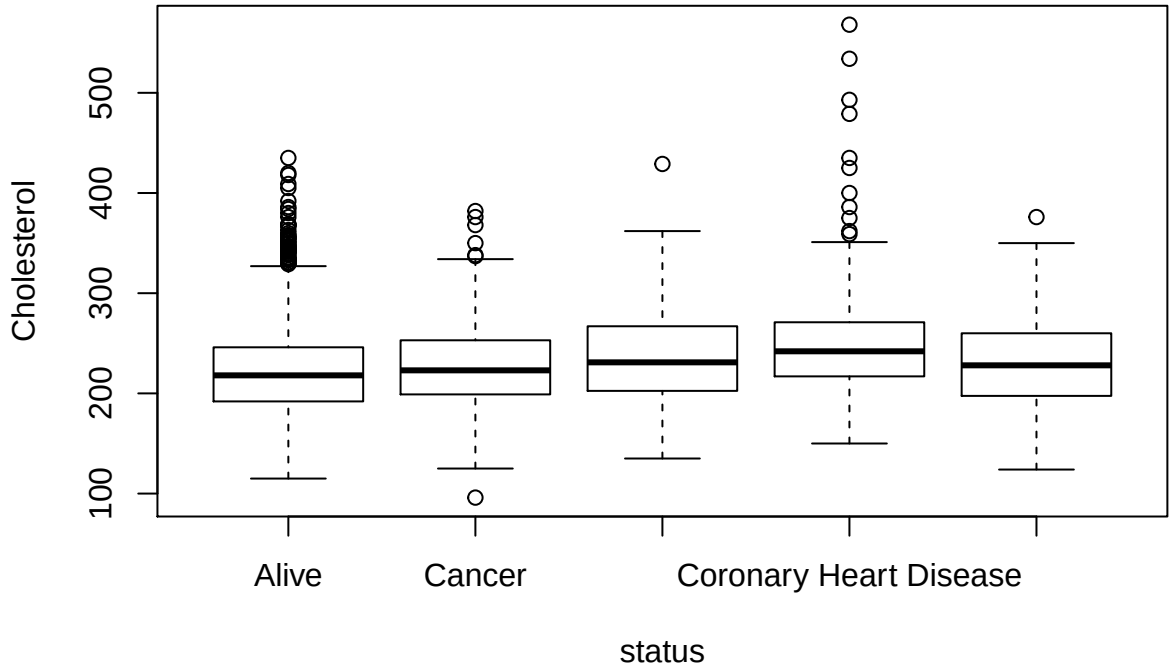
```
heart2 <- na.omit(heart2) # Omit NA values
cor(heart2[,c(3:10)]) # Correlation table
```

```
##              AgeAtStart      Height     Weight   Diastolic    Systolic
## AgeAtStart   1.00000000 -0.13316288 0.09171753  0.27808219  0.38107757
## Height      -0.13316288  1.00000000 0.52540450 -0.01082595 -0.06823295
## Weight       0.09171753  0.52540450 1.00000000  0.32899896  0.26280187
## Diastolic    0.27808219 -0.01082595 0.32899896  1.00000000  0.79703187
## Systolic     0.38107757 -0.06823295 0.26280187  0.79703187  1.00000000
## MRW          0.20425669 -0.13013071 0.76528950  0.38650883  0.36174515
## Smoking     -0.17092204  0.28970750 0.09148598 -0.06694752 -0.09143971
## Cholesterol  0.28015156 -0.07724940 0.07648347  0.18493532  0.19991789
##                     MRW     Smoking Cholesterol
## AgeAtStart   0.2042567 -0.17092204  0.28015156
## Height      -0.1301307  0.28970750 -0.07724940
## Weight       0.7652895  0.09148598  0.07648347
## Diastolic    0.3865088 -0.06694752  0.18493532
## Systolic     0.3617451 -0.09143971  0.19991789
## MRW          1.0000000 -0.12552686  0.14053776
## Smoking     -0.1255269  1.00000000 -0.01411134
## Cholesterol  0.1405378 -0.01411134  1.00000000
```

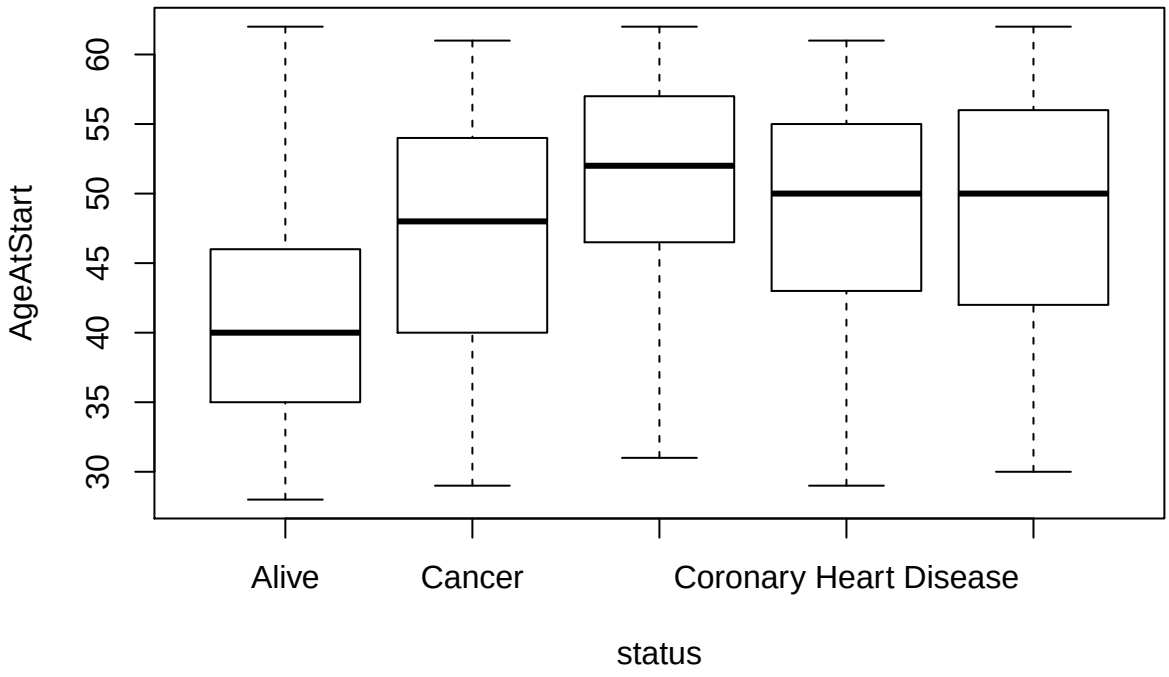
```
# Create boxplots for exploratory data analysis:
boxplot(MRW ~ status, data = heart2)
```



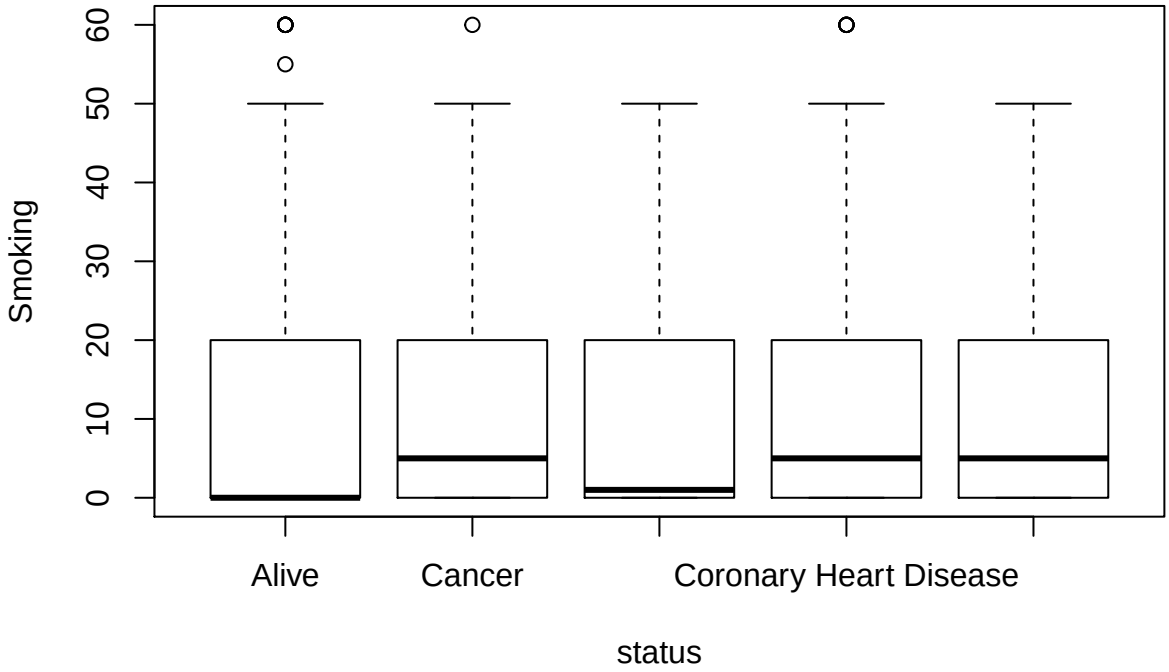

```
boxplot(Cholesterol ~ status, data = heart2)
```



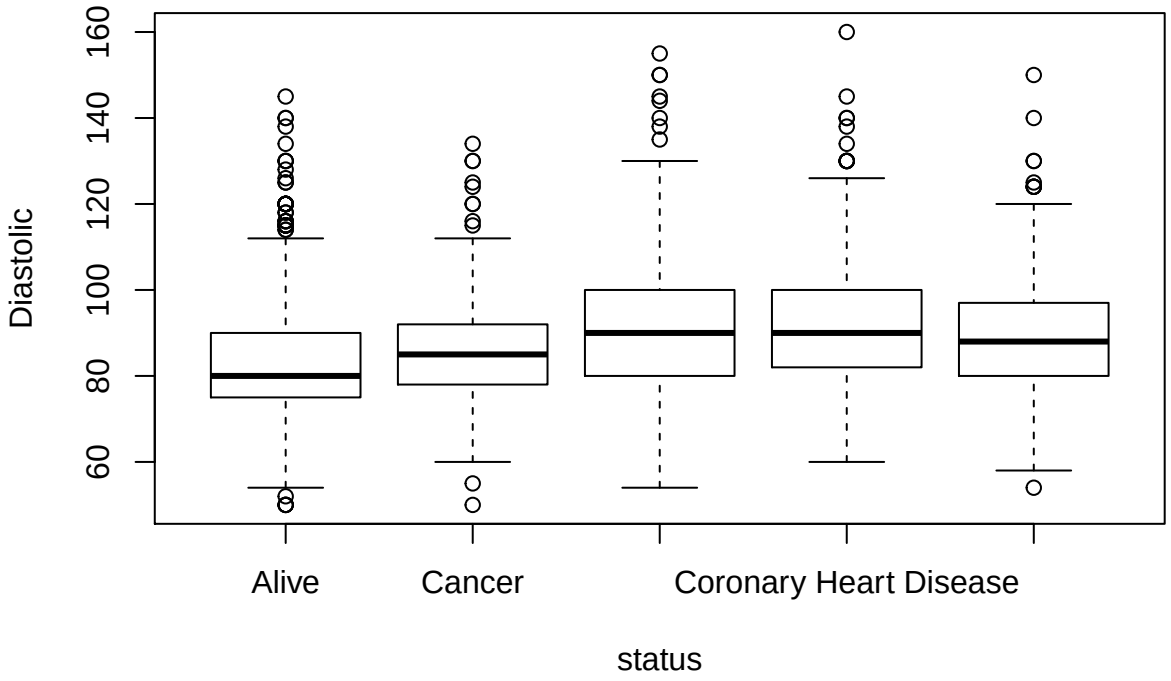

```
boxplot(AgeAtStart ~ status, data = heart2)
```



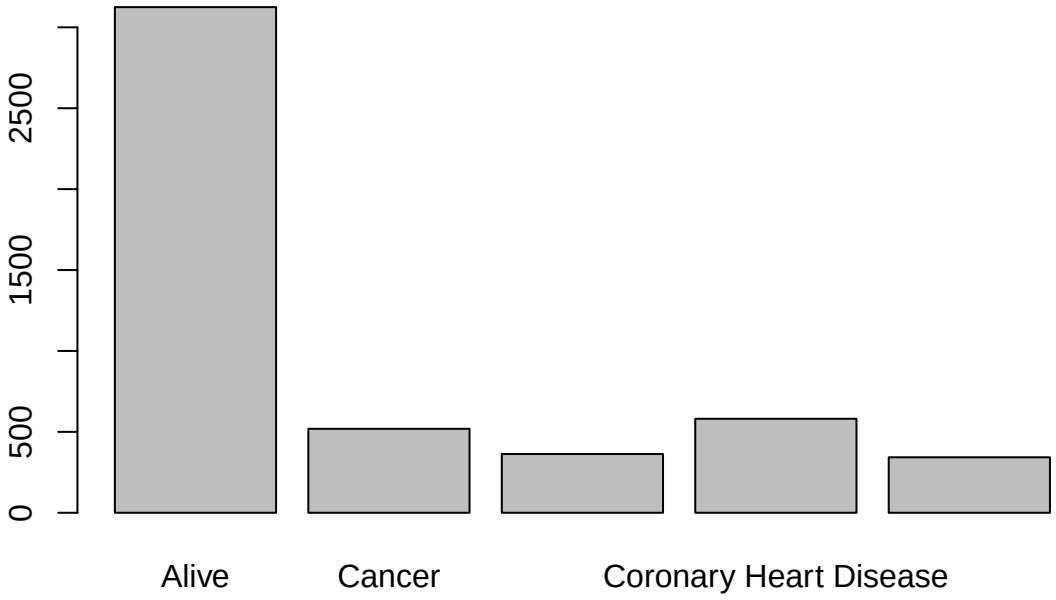

```
boxplot(Smoking ~ status, data = heart2)
```




```
boxplot(Diastolic ~ status, data = heart2)
```




```
barplot(summary(heart2$status))
```




```
# Full model
model11 <- multinom(status ~ Sex + AgeAtStart + Height + Weight + Diastolic + Systolic + MRW + Smoking +
```

```
## # weights:  55 (40 variable)
## initial  value 7936.138346
## iter  10 value 5497.078574
## iter  20 value 5374.453766
## iter  30 value 5328.390048
## iter  40 value 4897.941611
## iter  50 value 4889.712582
## iter  60 value 4889.349632
## final  value 4889.348250
```

```
## converged
```

```
summary(model11)
```

```
## Call:
## multinom(formula = status ~ Sex + AgeAtStart + Height + Weight +
##     Diastolic + Systolic + MRW + Smoking + Cholesterol, data = heart2)
##
## Coefficients:
##                           (Intercept)   SexMale AgeAtStart      Height
## Cancer                     -16.147367 0.4343035  0.1033085  0.13431189
## Cerebral Vascular Disease  -11.542571 0.5198131  0.1467815 -0.02414057
## Coronary Heart Disease     -12.105201 1.3011514  0.1150143 -0.02674811
## Other                       -1.233795 0.4950045  0.1202305 -0.13728295
##                                 Weight    Diastolic    Systolic
## Cancer                    -2.621938e-02 0.0002476929 0.007946151
## Cerebral Vascular Disease  6.678207e-05 0.0011289532 0.025201629
## Coronary Heart Disease     3.411793e-03 0.0124479938 0.017799458
## Other                      1.819632e-02 0.0188749804 0.008943226
##                                   MRW    Smoking   Cholesterol
## Cancer                     0.032478873 0.03219868 -0.0018134825
## Cerebral Vascular Disease  0.001301373 0.03447400 -0.0008125236
## Coronary Heart Disease     0.001585064 0.02761757  0.0071768010
## Other                     -0.026629571 0.03049415 -0.0019466613
##
## Std. Errors:
##                           (Intercept)   SexMale  AgeAtStart      Height
## Cancer                    0.001655168 0.1580599 0.006867805 0.007840582
## Cerebral Vascular Disease 0.002215839 0.1906947 0.008941314 0.010395780
## Coronary Heart Disease    0.001861678 0.1603625 0.007078583 0.008796396
## Other                     0.002257614 0.1878832 0.008441306 0.009767205
##                                Weight   Diastolic    Systolic         MRW
## Cancer                    0.004581351 0.006482056 0.003674504 0.005696090
## Cerebral Vascular Disease 0.005448565 0.007168026 0.003795288 0.006679880
## Coronary Heart Disease    0.004398936 0.006169173 0.003420983 0.005544198
## Other                     0.005370231 0.007509527 0.004178412 0.006790535
##                               Smoking Cholesterol
## Cancer                    0.004357224 0.001186356
## Cerebral Vascular Disease 0.005453306 0.001398769
## Coronary Heart Disease    0.004304711 0.001088523
## Other                     0.005302649 0.001412753
##
## Residual Deviance: 9778.697
## AIC: 9858.697
```

```
# Calculate p-values for slopes:
z11 <- summary(model11)$coefficients/summary(model11)$standard.errors
p11 <- (1-pnorm(abs(z11), 0, 1)) * 2
p11
```

```
##                           (Intercept)      SexMale AgeAtStart      Height
## Cancer                              0 6.001210e-03          0 0.000000000
## Cerebral Vascular Disease           0 6.412802e-03          0 0.020224821
## Coronary Heart Disease              0 4.440892e-16          0 0.002359482
## Other                               0 8.422663e-03          0 0.000000000
```

```
##                                Weight  Diastolic     Systolic
## Cancer                    1.046187e-08 0.96951859 3.057891e-02
## Cerebral Vascular Disease 9.902207e-01 0.87485201 3.131695e-11
## Coronary Heart Disease    4.379882e-01 0.04361486 1.960718e-07
## Other                     7.030989e-04 0.01195495 3.232724e-02
##                                   MRW      Smoking  Cholesterol
## Cancer                    1.184381e-08 1.472156e-13 1.263596e-01
## Cerebral Vascular Disease 8.455340e-01 2.587532e-10 5.613180e-01
## Coronary Heart Disease    7.749577e-01 1.402123e-10 4.305734e-11
## Other                     8.797320e-05 8.885429e-09 1.682279e-01
```

```
# Reduced model
model12 <- multinom(status ~ Sex + AgeAtStart + Systolic + Height + MRW + Smoking + Cholesterol, data =
```

```
## # weights:  45 (32 variable)
## initial  value 7936.138346
## iter  10 value 5743.924488
## iter  20 value 5631.413021
## iter  30 value 5084.800037
## iter  40 value 4896.932985
## final  value 4896.932882
## converged
```

```
summary(model12)
```

```
## Call:
## multinom(formula = status ~ Sex + AgeAtStart + Systolic + Height +
##     MRW + Smoking + Cholesterol, data = heart2)
##
## Coefficients:
##                           (Intercept)   SexMale AgeAtStart    Systolic
## Cancer                      -8.299286 0.3010519  0.1039421 0.008156203
## Cerebral Vascular Disease  -11.521405 0.5278792  0.1468859 0.025896448
## Coronary Heart Disease     -12.983975 1.3447085  0.1134744 0.023003048
## Other                       -6.214146 0.6267248  0.1175332 0.016839568
##                                 Height           MRW    Smoking
## Cancer                     0.012441442 -0.0001048864 0.03222578
## Cerebral Vascular Disease -0.024767205  0.0014955233 0.03435014
## Coronary Heart Disease    -0.008931127  0.0068133651 0.02723489
## Other                     -0.052539395 -0.0029897156 0.02989480
##                            Cholesterol
## Cancer                    -0.0019294432
## Cerebral Vascular Disease -0.0008190402
## Coronary Heart Disease     0.0072909763
## Other                     -0.0017364261
##
## Std. Errors:
##                           (Intercept)   SexMale  AgeAtStart    Systolic
## Cancer                    0.007376249 0.1166085 0.006609117 0.002466293
## Cerebral Vascular Disease 0.008209092 0.1489668 0.008464860 0.002435518
## Coronary Heart Disease    0.006559977 0.1298508 0.006641972 0.002218073
## Other                     0.008392981 0.1426355 0.008035251 0.002652016
##                                Height         MRW     Smoking Cholesterol
## Cancer                    0.006850242 0.002675037 0.004315014 0.001176406
## Cerebral Vascular Disease 0.008906952 0.002982596 0.005423290 0.001384913
```

```
## Coronary Heart Disease    0.007585577 0.002607722 0.004281150 0.001075783
## Other                     0.008416116 0.003149545 0.005301206 0.001398042
##
## Residual Deviance: 9793.866
## AIC: 9857.866
```

```
# Calculate p-values for slopes:
z12 <- summary(model12)$coefficients/summary(model12)$standard.errors
p12 <- (1-pnorm(abs(z12), 0, 1))*2
p12
```

```
##                           (Intercept)      SexMale AgeAtStart     Systolic
## Cancer                              0 9.830600e-03          0 9.427738e-04
## Cerebral Vascular Disease           0 3.946996e-04          0 0.000000e+00
## Coronary Heart Disease              0 0.000000e+00          0 0.000000e+00
## Other                               0 1.113394e-05          0 2.157015e-10
##                                 Height         MRW      Smoking
## Cancer                    6.933898e-02 0.968723486 8.126833e-14
## Cerebral Vascular Disease 5.424851e-03 0.616077912 2.391674e-10
## Coronary Heart Disease    2.390428e-01 0.008981298 1.996860e-10
## Other                     4.300462e-10 0.342491782 1.707974e-08
##                            Cholesterol
## Cancer                    1.009809e-01
## Cerebral Vascular Disease 5.542511e-01
## Coronary Heart Disease    1.223888e-11
## Other                     2.142213e-01
```

```
# Test which model is a better fit (equivalent to ANOVA):
1-pchisq(model12$deviance-model11$deviance, model11$edf-model12$edf)
```

```
## [1] 0.05593662
```

```
# Keep only as outcome variable causes of death:
heart3 <- heart2[heart2$status != "Alive",]
heart3$status <- droplevels(heart3$status)
# Full model:
model21 <- multinom(as.factor(status) ~ Sex + AgeAtStart + Height + Weight + Diastolic + Systolic + MRW
```

```
## # weights:  44 (30 variable)
## initial  value 2503.647616
## iter  10 value 2425.697966
## iter  20 value 2406.879208
## iter  30 value 2346.407019
## iter  40 value 2344.350692
## final  value 2344.228576
## converged
```

```
summary(model21)
```

```
## Call:
## multinom(formula = as.factor(status) ~ Sex + AgeAtStart + Height +
##     Weight + Diastolic + Systolic + MRW + Smoking + Cholesterol,
##     data = heart3)
##
## Coefficients:
##                           (Intercept)     SexMale AgeAtStart     Height
## Cerebral Vascular Disease    3.950559 -0.00849094 0.04335961 -0.1461075
```

```
## Coronary Heart Disease       3.126397  0.78984238 0.01197511 -0.1418803
## Other                       12.986890  0.04535573 0.01751991 -0.2421454
##                               Weight   Diastolic     Systolic         MRW
## Cerebral Vascular Disease 0.02316471 0.003807816 0.0152083618 -0.02795960
## Coronary Heart Disease    0.02505860 0.014516312 0.0076633890 -0.02625344
## Other                     0.03797201 0.019427548 0.0002504285 -0.05092670
##                                Smoking  Cholesterol
## Cerebral Vascular Disease  0.003059073 1.135162e-03
## Coronary Heart Disease    -0.004151674 8.819169e-03
## Other                     -0.001291760 1.897679e-05
##
## Std. Errors:
##                           (Intercept)   SexMale  AgeAtStart     Height
## Cerebral Vascular Disease 0.002515141 0.2240481 0.009998029 0.01178283
## Coronary Heart Disease    0.002119418 0.2012465 0.008605591 0.01054983
## Other                     0.002500539 0.2232505 0.009636713 0.01129415
##                                Weight   Diastolic    Systolic         MRW
## Cerebral Vascular Disease 0.006403173 0.008442112 0.004415424 0.007846730
## Coronary Heart Disease    0.005622373 0.007647880 0.004129946 0.007024723
## Other                     0.006403785 0.008787392 0.004751963 0.008009903
##                              Smoking Cholesterol
## Cerebral Vascular Disease 0.006282941 0.001649083
## Coronary Heart Disease    0.005397188 0.001444183
## Other                     0.006227140 0.001665850
##
## Residual Deviance: 4688.457
## AIC: 4748.457
```

```
# Calculate p-values for slopes:
z21 <- summary(model21)$coefficients/summary(model21)$standard.errors
p21 <- (1-pnorm(abs(z21), 0, 1))*2
p21
```

```
##                           (Intercept)      SexMale   AgeAtStart Height
## Cerebral Vascular Disease           0 9.697691e-01 1.445619e-05      0
## Coronary Heart Disease              0 8.681937e-05 1.640588e-01      0
## Other                               0 8.390094e-01 6.905826e-02      0
##                                 Weight  Diastolic     Systolic
## Cerebral Vascular Disease 2.972409e-04 0.65195340 0.0005723879
## Coronary Heart Disease    8.313658e-06 0.05768513 0.0635154165
## Other                     3.036354e-09 0.02704668 0.9579709312
##                                    MRW   Smoking  Cholesterol
## Cerebral Vascular Disease 3.663370e-04 0.6263394 4.912263e-01
## Coronary Heart Disease    1.860129e-04 0.4417573 1.017213e-09
## Other                     2.044418e-10 0.8356660 9.909110e-01
```

```
# Reduced model:
model22 <- multinom(as.factor(status) ~ Sex + AgeAtStart + Height + Weight  + Systolic + MRW  + Cholest
```

```
## # weights:  36 (24 variable)
## initial  value 2503.647616
## iter  10 value 2402.747498
## iter  20 value 2379.033364
## iter  30 value 2348.888524
## iter  40 value 2348.458336
## final  value 2348.458265
```

```
## converged
```

```
summary(model22)
```

```
## Call:
## multinom(formula = as.factor(status) ~ Sex + AgeAtStart + Height +
##     Weight + Systolic + MRW + Cholesterol, data = heart3)
##
## Coefficients:
##                            (Intercept)    SexMale AgeAtStart     Height
## Cerebral Vascular Disease    4.269743 0.02451405 0.04160652 -0.1485473
## Coronary Heart Disease       4.582787 0.75035615 0.01169754 -0.1605918
## Other                       14.972928 0.03095414 0.01506856 -0.2657741
##                               Weight   Systolic         MRW  Cholesterol
## Cerebral Vascular Disease 0.02355698 0.01688737 -0.02826523 0.0011775915
## Coronary Heart Disease    0.02918275 0.01368542 -0.03020565 0.0088324166
## Other                     0.04329627 0.00829191 -0.05610135 0.0001175059
##
## Std. Errors:
##                            (Intercept)   SexMale  AgeAtStart      Height
## Cerebral Vascular Disease 0.002396450 0.2155611 0.009538427 0.010819902
## Coronary Heart Disease    0.002046684 0.1938642 0.008184829 0.009799939
## Other                     0.002379492 0.2144093 0.009160504 0.010324129
##                                Weight    Systolic         MRW Cholesterol
## Cerebral Vascular Disease 0.006379320 0.002860790 0.007777336 0.001645460
## Coronary Heart Disease    0.005588289 0.002703405 0.006939367 0.001438877
## Other                     0.006355171 0.003050951 0.007908993 0.001657603
##
## Residual Deviance: 4696.917
## AIC: 4744.917
```

```
# Calculate p-values for slopes:
z22 <- summary(model22)$coefficients/summary(model22)$standard.errors
p22 <- (1-pnorm(abs(z22), 0, 1))*2
p22
```

```
##                            (Intercept)      SexMale   AgeAtStart Height
## Cerebral Vascular Disease            0 0.9094581245 1.288851e-05      0
## Coronary Heart Disease               0 0.0001086018 1.529543e-01      0
## Other                                0 0.8852087909 9.998046e-02      0
##                                 Weight     Systolic          MRW
## Cerebral Vascular Disease 2.218772e-04 3.568565e-09 2.787287e-04
## Coronary Heart Disease    1.768805e-07 4.142500e-07 1.344117e-05
## Other                     9.574119e-12 6.571531e-03 1.308953e-12
##                            Cholesterol
## Cerebral Vascular Disease 4.742008e-01
## Coronary Heart Disease    8.335248e-10
## Other                     9.434860e-01
```

```
# Check which model is a better fit:
1-pchisq(model22$deviance-model21$deviance, model21$edf-model22$edf)
```

```
## [1] 0.2063418
```

```
# We can do the same with ANOVA:
anova(model21, model22)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: as.factor(status)
##                                                                                 Model
## 1                       Sex + AgeAtStart + Height + Weight + Systolic + MRW + Cholesterol
## 2 Sex + AgeAtStart + Height + Weight + Diastolic + Systolic + MRW + Smoking + Cholesterol
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1      5394   4696.917
## 2      5388   4688.457 1 vs 2     6 8.459377 0.2063418
```

```r
# Change model12 to the name of your reduced model
# The code below plots the residuals of odds(cancer) against expected value of odds(cancer)
# You can change the '2' to '3' or '4' based on which status you like best
binnedplot(x=model12$fitted.values[,2],y=model12$residuals[,2],
           xlab='Expected values for cancer', main=NULL)
```