HW3

Duc Le

12/7/2020

Problem 1

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
library(mlbench)

rm(list = ls())
d = get(data(BostonHousing))
str(d)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
```

```
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
q = quantile(d$medv, c(0.25, 0.75))

d$medv = ifelse(d$medv < q[1], 0, ifelse(d$medv >= q[1] & d$medv <= q[2], 1, 2))
table(d$medv)
```

```
##
##   0   1   2
## 127 255 124
```

```
model1 = vglm(medv~., multinomial(refLevel = 1), data = d)
summary(model1)
```

```
##
## Call:
## vglm(formula = medv ~ ., family = multinomial(refLevel = 1),
##     data = d)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  24.156526   4.974550   4.856 1.20e-06 ***
## (Intercept):2  29.487035   6.911503   4.266 1.99e-05 ***
## crim:1         -0.223224   0.074894  -2.981 0.002877 **
## crim:2         -0.064667   0.064968  -0.995 0.319561
## zn:1            0.046785   0.042013   1.114 0.265450
## zn:2            0.057202   0.043325   1.320 0.186735
## indus:1         0.102091   0.060672   1.683 0.092437 .
## indus:2        -0.016965   0.084062  -0.202 0.840063
## chas1:1         0.344743   0.754352   0.457 0.647667
## chas1:2         1.314876   1.086850   1.210 0.226354
## nox:1         -12.902011   3.229133  -3.996 6.46e-05 ***
## nox:2         -18.831263   5.589106  -3.369 0.000754 ***
## rm:1           -0.161492   0.423880  -0.381 0.703214
## rm:2            1.762873   0.600001   2.938 0.003302 **
## age:1          -0.039121   0.016202  -2.415 0.015757 *
## age:2          -0.036698   0.019530  -1.879 0.060244 .
## dis:1          -0.663647   0.251272  -2.641 0.008262 **
## dis:2          -1.216773   0.295898  -4.112 3.92e-05 ***
## rad:1           0.182666   0.058094   3.144 0.001665 **
## rad:2           0.461493   0.098197   4.700 2.61e-06 ***
## tax:1          -0.006587   0.003290  -2.002 0.045288 *
## tax:2          -0.017814   0.005194  -3.430 0.000604 ***
## ptratio:1      -0.375092   0.135338  -2.772 0.005580 **
## ptratio:2      -0.714513   0.181660  -3.933 8.38e-05 ***
## b:1             0.005880   0.002010   2.925 0.003440 **
## b:2             0.001375   0.004295   0.320 0.748942
## lstat:1        -0.241131   0.048098  -5.013 5.35e-07 ***
## lstat:2        -0.590055   0.089672  -6.580 4.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])
##
## Residual deviance: 405.7153 on 984 degrees of freedom
##
## Log-likelihood: -202.8577 on 984 degrees of freedom
##
## Number of Fisher scoring iterations: 9
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1', 'nox:2', 'lstat:2'
##
##
## Reference group is level  1  of the response
```

Looking at the summary of the coefficients, we can eliminate variables that aren't statistically significant to simplify the model such as: crim, zn, indus, cha1. Since my reference level = 1, these variables don't contribute much to the log odds of being between the 25th & 75th percentile Similarly, they aren't significant when it comes to predicting the log odds of being above the 75th percentile.

In addition, we can view the effects of these coefficients the same way we did in normal logistic regression, with the exception that each Beta's will have different effects on the odds depending on what you selected for the reference level.

Problem 2

```
model2 = vglm(medv~., family = propodds(), data = d)
summary(model2)
```

```
##
## Call:
## vglm(formula = medv ~ ., family = propodds(), data = d)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  1.509e+01  3.125e+00   4.830 1.37e-06 ***
## (Intercept):2  9.112e+00  2.970e+00   3.068 0.002156 **
## crim          -1.469e-01  6.004e-02  -2.448 0.014380 *
## zn             1.571e-02  8.038e-03   1.954 0.050706 .
## indus          6.629e-04  3.383e-02   0.020 0.984368
## chas1          9.262e-01  4.932e-01   1.878 0.060387 .
## nox           -1.117e+01  2.327e+00  -4.800 1.59e-06 ***
## rm             9.751e-01  2.578e-01   3.782 0.000155 ***
## age           -4.037e-03  7.432e-03  -0.543 0.587048
## dis           -6.369e-01  1.199e-01  -5.311 1.09e-07 ***
## rad            1.670e-01  4.349e-02   3.840 0.000123 ***
## tax           -6.424e-03  2.273e-03  -2.826 0.004714 **
## ptratio       -3.411e-01  7.793e-02  -4.377 1.20e-05 ***
## b              4.785e-03  1.797e-03   2.663 0.007755 **
## lstat         -2.923e-01  3.801e-02  -7.689 1.48e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
##
## Residual deviance: 469.3142 on 997 degrees of freedom
```

```
##
## Log-likelihood: -234.6571 on 997 degrees of freedom
##
## Number of Fisher scoring iterations: 8
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##         crim           zn        indus        chas1          nox           rm
## 8.633389e-01 1.015830e+00 1.000663e+00 2.524953e+00 1.411562e-05 2.651328e+00
##          age          dis          rad          tax      ptratio            b
## 9.959717e-01 5.289203e-01 1.181742e+00 9.935963e-01 7.109641e-01 1.004797e+00
##        lstat
## 7.465597e-01
```

```
lrtest(model1, model2)
```

```
## Likelihood ratio test
##
## Model 1: medv ~ .
## Model 2: medv ~ .
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1 984 -202.86
## 2 997 -234.66 13 63.599  1.183e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the likelihood ratio statistical test, we are able to reject H0 due to the statistical significance, thus the model from part a is correct.