Biostat 200C Homework 1

Due Apr 12 @ 11:59PM

To submit homework, please upload both RMD and pdf files to CCLE by the deadline.

Q1. Binomial Distribution

Let Y_i be the number of successes in n_i trials with

$$Y_i \sim Bin(n_i, \pi_i),$$

where the probabilities π_i have a Beta distribution

$$\pi_i \sim Beta(\alpha, \beta).$$

The probability density function for the Beta distribution is $f(x;\alpha,\beta) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ for $x \in [0,1], \alpha > 0, \beta > 0$, and the beta function $B(\alpha,\beta)$ defining the normalizing constant required to ensure that $\int_0^1 f(x;\alpha,\beta) = 1$. Let $\theta = \alpha/(\alpha+\beta)$, show that

a. $E(\pi_i) = \theta$

b. $Var(\pi_i) = \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)$

c. $E(Y_i) = n_i\theta$

d. $Var(Y_i) = n_i\theta(1-\theta)[1+(n_i-1)\phi]$ so that $Var(Y_i)$ larger than the Binomial variance (unless $n_i = 1$ or $\phi = 0$).

$E(x^n) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}$

$\phi = \frac{1}{\alpha+\beta+1}$

$Beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ $\left(\int_0^1 pdf\ Beta(\alpha+1,\beta) = 1\right)$

a) $E(\pi_i) = \int_0^1 \pi_i \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1} d\pi_i$

$= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(\alpha+1+\beta)} \int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)} \pi_i^{\alpha+1-1}(1-\pi_i)^{\beta-1} d\pi_i$

$= \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)}{\Gamma(\alpha)(\alpha+\beta)\Gamma(\alpha+\beta)} \cdot 1 = \frac{\alpha}{(\alpha+\beta)} = \theta.$

b)

$E(\pi_i^2) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+2)}{\Gamma(\alpha)\Gamma(\alpha+\beta+2)} = \frac{\Gamma(\alpha+\beta)(\alpha+1)\Gamma(\alpha+1)}{\Gamma(\alpha)(\alpha+\beta+1)\Gamma(\alpha+\beta+1)} = \frac{(\alpha+1)\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{(\alpha+\beta+1)\Gamma(\alpha)\Gamma(\alpha+\beta+1)}$

$= \frac{(\alpha+1)\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)}{(\alpha+\beta+1)\Gamma(\alpha)(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} \cdot \underbrace{\int_0^1 pdf\ Beta}_{\Downarrow 1}$

$= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$

$= \frac{\theta(\alpha+1)}{(\alpha+\beta+1)}$

$$\text{Var}(\bar{\pi}_i) = \frac{\theta(\alpha+1)}{(\alpha+\beta+1)} - \left(\bar{E}(\bar{\pi}_i)\right)^2$$

$$= \frac{(\alpha^2+\alpha)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{(\alpha\beta)}{(\alpha+\beta)^2} / (\alpha+\beta+1)$$

$$= \frac{\theta(1-\theta)}{(\alpha+\beta+1)} = \bar{\phi}\,\theta(1-\theta)$$

$\bar{E}(\bar{\pi}_i^2) - (\bar{E}(\bar{\pi}_i))^2$

$$= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \left(\frac{\alpha^2}{(\alpha+\beta)^2}\right)$$

$$= \frac{(\alpha+1)\alpha\,(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)(\alpha+\beta)} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{(\alpha^2+\alpha)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{\cancel{\alpha^3}+\cancel{\alpha^2\beta}+\cancel{\alpha^2}+\alpha\beta-\cancel{\alpha^3}-\cancel{\alpha^2\beta}-\cancel{\alpha^2}}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$ (always matter~)

c) $\bar{E}(Y_i) = \bar{E}_{\bar{\pi}_i}(\bar{E}_{Y_i}(Y_i/\bar{\pi}_i))$

$$= \bar{E}_{\bar{\pi}_i}(n_i \cdot \bar{\pi}_i)$$

$$= n_i\,\bar{E}(\bar{\pi}_i)$$

$$= n_i\,\theta$$

d) $\text{Var}(Y_i) = \bar{E}_{\bar{\pi}_i}(\text{Var}(Y_i|\bar{\pi}_i)) + \text{Var}_{\bar{\pi}_i}(\bar{E}(Y_i/\bar{\pi}_i))$

$$= \bar{E}_{\bar{\pi}_i}(n_i \cdot \bar{\pi}_i \cdot (1-\bar{\pi}_i)) + \text{Var}(n_i \cdot \bar{\pi}_i)$$

$$= n_i \cdot [\bar{E}(\bar{\pi}_i) - \bar{E}(\bar{\pi}_i^2)] + n_i^2 \cdot \bar{\phi}\,\theta(1-\theta)$$

$$= n_i \cdot \left(\theta - \frac{\theta(\alpha+1)}{(\alpha+\beta+1)}\right) + n_i^2 \cdot \bar{\phi}\,\theta(1-\theta)$$

$$= n_i \cdot (\theta\beta/(\alpha+\beta+1)) + n_i^2 \cdot \bar{\phi}\,\theta(1-\theta)$$

$$= n_i \cdot [\theta\cdot(1-\theta)(1-\bar{\phi})] + n_i^2 \cdot \bar{\phi}\,\theta(1-\theta)$$

$$= \cancel{n_i\theta(1-\theta)}(1-\bar{\phi}) + n_i^2 \cdot \bar{\phi}\,\cancel{\theta(1-\theta)}$$

$$= n_i\theta(1-\theta)[1-\bar{\phi}+n_i\bar{\phi}]$$

$$= n_i\theta(1-\theta)[1-(n_i-1)\bar{\phi}]$$

Biostat 200C Homework 1

Due Apr 12 @ 11:59PM

To submit homework, please upload both RMD and pdf files to CCLE by the deadline.

Q2. (ELMR Chapter 3 Exercise 1)

A case-control study of esophageal cancer in Ileet-Vilaine, France.

```
data(esoph)
help(esoph)
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(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

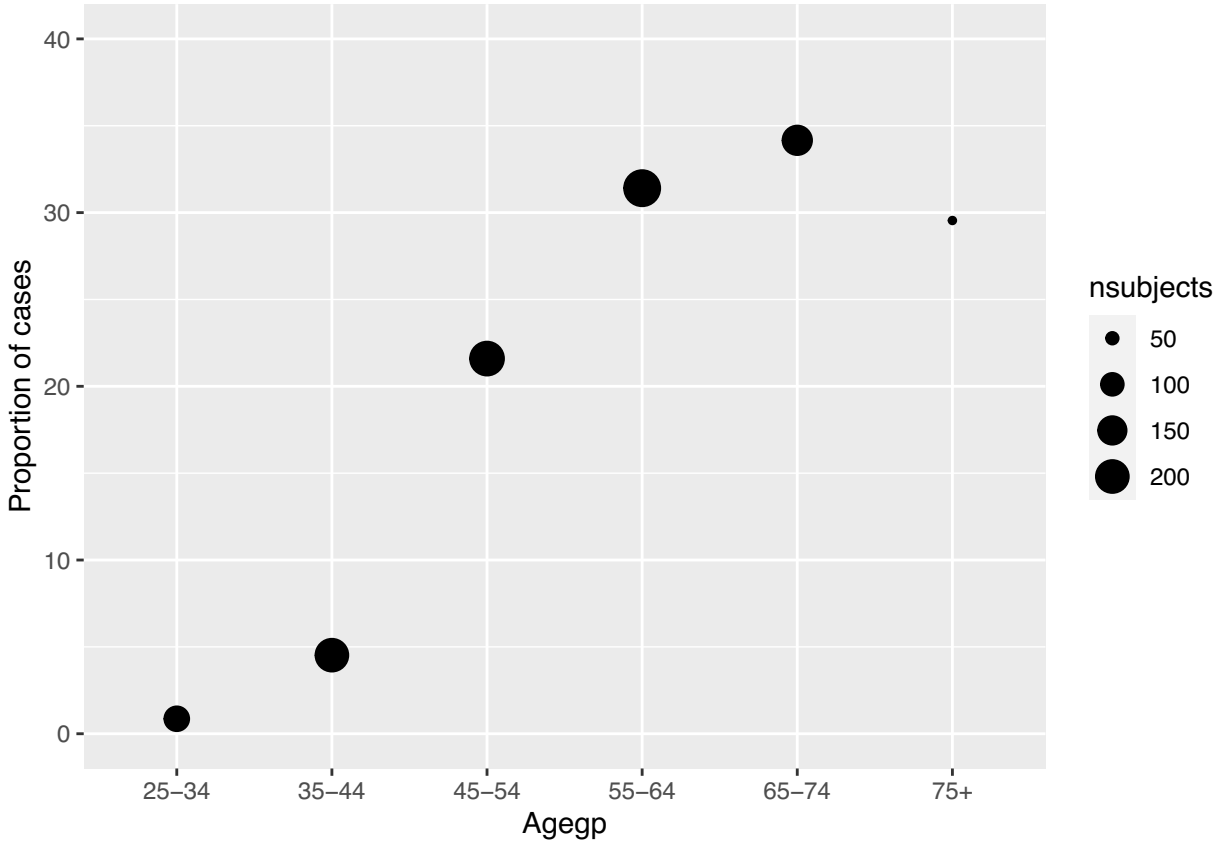
```
#conver to tibble for later use.
esoph <- esoph %>%
  as_tibble()

#creat # of subject.
esoph <- esoph %>%
  mutate(nsubjects = ncases + ncontrols) %>%
  print(width = Inf)
```

```
## # A tibble: 88 x 6
##    agegp alcgp     tobgp    ncases ncontrols nsubjects
##    <ord> <ord>     <ord>     <dbl>     <dbl>     <dbl>
##  1 25-34 0-39g/day 0-9g/day      0        40        40
##  2 25-34 0-39g/day 10-19         0        10        10
##  3 25-34 0-39g/day 20-29         0         6         6
##  4 25-34 0-39g/day 30+           0         5         5
##  5 25-34 40-79     0-9g/day      0        27        27
##  6 25-34 40-79     10-19         0         7         7
##  7 25-34 40-79     20-29         0         4         4
##  8 25-34 40-79     30+           0         7         7
##  9 25-34 80-119    0-9g/day      0         2         2
## 10 25-34 80-119    10-19         0         1         1
## # ... with 78 more rows
```

a. Plot the proportion of cases against each predictor using the size of the point to indicate the number of subject as seen in Figure 2.7. Comment on the realtionships seen in the plots.

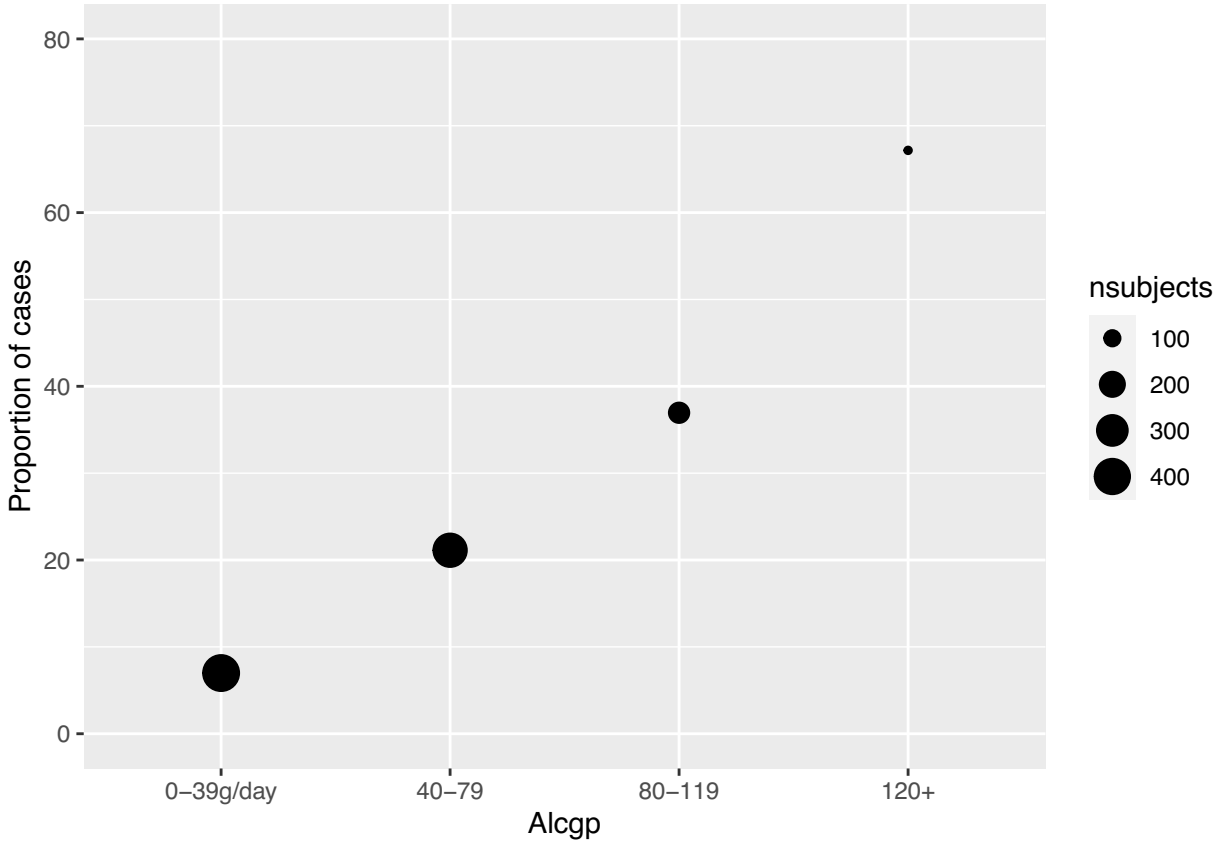
Solution:

```
esoph %>%
  group_by(agegp) %>%
  summarize(ncases = sum(ncases), nsubjects = sum(nsubjects)) %>%
  ggplot(aes(x = agegp, y = 100 * ncases / nsubjects, size = nsubjects)) +
  geom_point() +
  labs(x = "Agegp", y = "Proportion of cases") +
  scale_y_continuous(limits = c(0, 40))
```



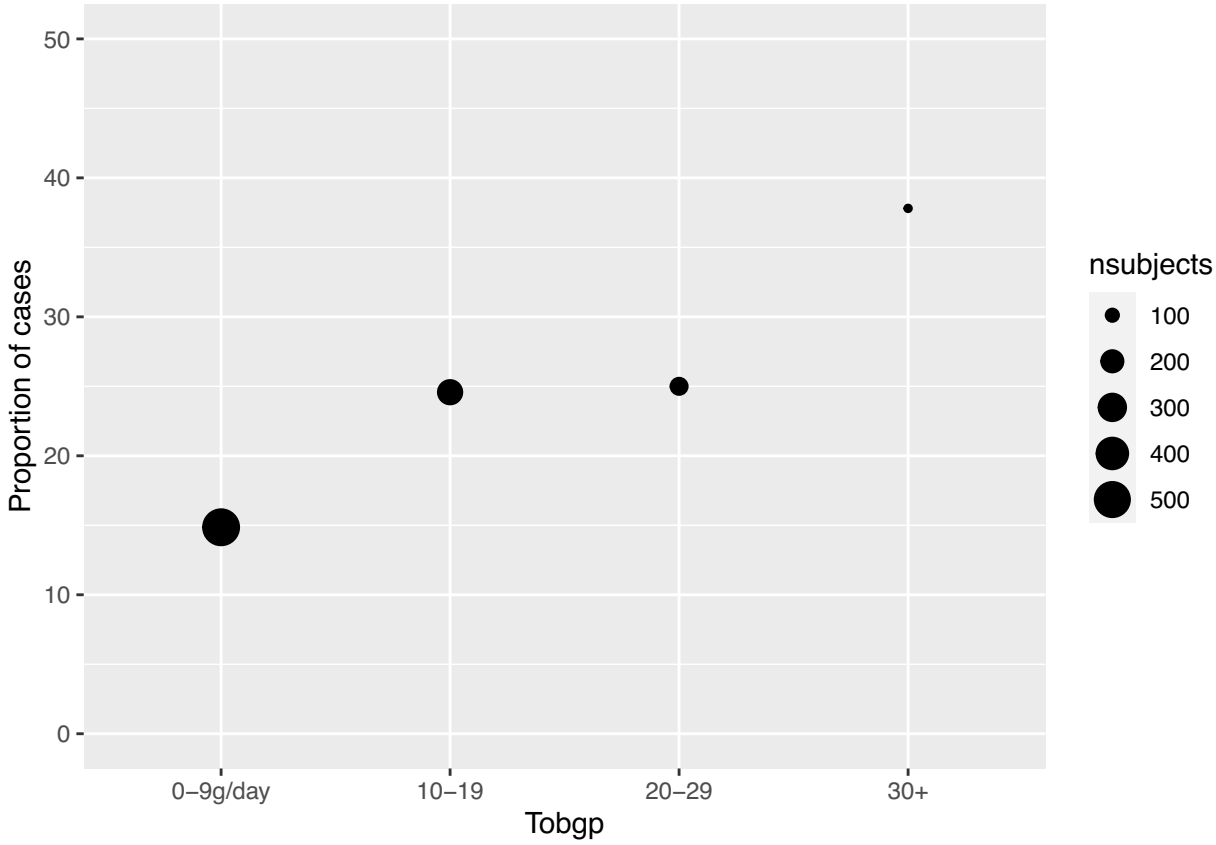
From the plot above, as the age increases, the proportion of cases increases, but at age 75 and above, the proportion of cases decreases. Also, the number of subjects only gets to 50. My guess is that some of them are dead.

```
esoph %>%
  group_by(alcgp) %>%
  summarize(ncases = sum(ncases), nsubjects = sum(nsubjects)) %>%
  ggplot(aes(x = alcgp, y = 100 * ncases / nsubjects, size = nsubjects)) +
  geom_point() +
  labs(x = "Alcgp", y = "Proportion of cases") +
  scale_y_continuous(limits = c(0, 80))
```



From the plot above, as people get more alcohol, the proportion of cases increases, but the number of subjects decreases.

```
esoph %>%
  group_by(tobgp) %>%
  summarize(ncases = sum(ncases), nsubjects = sum(nsubjects)) %>%
  ggplot(aes(x = tobgp, y = 100 * ncases / nsubjects, size = nsubjects)) +
  geom_point() +
  labs(x = "Tobgp", y = "Proportion of cases") +
  scale_y_continuous(limits = c(0, 50))
```



From the plot above, as the tobacco consumption increases the proportion of cases increases, but the number of subjects decreases.

b. Fit a binomial GLM with interactions between all three predictors. Use AIC as a criterion to select a model using the `step` function. Which model is selected?

Solution:

```
biglm <- glm(cbind(ncases, ncontrols) ~ (agegp + alcgp + tobgp)^2,
             family = binomial, data = esoph)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
smallm <- step(biglm, trace = TRUE)
```

```
## Start:  AIC=247.88
## cbind(ncases, ncontrols) ~ (agegp + alcgp + tobgp)^2

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                Df Deviance    AIC
## - alcgp:tobgp   9   37.535 236.59
## - agegp:tobgp  15   50.309 237.36
## - agegp:alcgp  15   56.807 243.86
## <none>             30.824 247.88
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Step:  AIC=236.59
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp
##
##               Df Deviance    AIC
## - agegp:tobgp 15   56.256 225.31
## - agegp:alcgp 15   62.776 231.83
## <none>             37.535 236.59
##
## Step:  AIC=225.31
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp
##
##               Df Deviance    AIC
## - agegp:alcgp 15   82.337 221.39
## <none>             56.256 225.31
## - tobgp        3   80.300 243.35
##
## Step:  AIC=221.39
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##         Df Deviance    AIC
## <none>       82.337 221.39
## - tobgp  3  105.881 238.94
## - agegp  5  208.825 337.88
## - alcgp  3  210.270 343.32
```

From the result, we chose the smallest value of AIC, which is AIC=221.39, cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp. Also, no intereaction term.

c. All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model may be possible by eliminating some of these terms. Use the `unclass` function to convert the factors to a numerical representation and check whether the model may be simplified.

Solution:

```
biglm2 <- glm(cbind(ncases, ncontrols) ~
               (unclass(agegp) + unclass(alcgp) + unclass(tobgp))^2,
             family = binomial, data = esoph)
smallm2 <- step(biglm2, trace = TRUE)
```

```
## Start:  AIC=236.12
## cbind(ncases, ncontrols) ~ (unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp))^2
##
##                                  Df Deviance    AIC
## - unclass(agegp):unclass(alcgp)  1   107.09 234.14
## - unclass(agegp):unclass(tobgp)  1   107.11 234.16
## - unclass(alcgp):unclass(tobgp)  1   108.49 235.54
```

```
## <none>                                107.07 236.12
##
## Step:  AIC=234.14
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + unclass(agegp):unclass(tobgp) + unclass(alcgp):unclass(tobgp)
##
##                                 Df Deviance    AIC
## - unclass(agegp):unclass(tobgp)  1   107.12 232.18
## - unclass(alcgp):unclass(tobgp)  1   108.50 233.56
## <none>                               107.09 234.14
##
## Step:  AIC=232.18
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + unclass(alcgp):unclass(tobgp)
##
##                                 Df Deviance    AIC
## - unclass(alcgp):unclass(tobgp)  1   108.78 231.83
## <none>                               107.12 232.18
## - unclass(agegp)                 1   208.95 332.00
##
## Step:  AIC=231.83
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp)
##
##                  Df Deviance    AIC
## <none>                108.78 231.83
## - unclass(tobgp)  1   129.81 250.87
## - unclass(agegp)  1   211.16 332.22
## - unclass(alcgp)  1   244.66 365.72
```

From the results, it is simplified, because the high AIC is not include here.

d. Use the summary output of the factor model to suggest a model that is slightly more complex than the linear model proposed in the previous question.

Solution:

```
summary(smallm)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9507  -0.7376  -0.2438   0.6130   2.4127
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19039    0.20737  -5.740 9.44e-09 ***
## agegp.L      3.99663    0.69389   5.760 8.42e-09 ***
## agegp.Q     -1.65741    0.62115  -2.668  0.00762 **
```

```
## agegp.C      0.11094    0.46815   0.237  0.81267
## agegp^4      0.07892    0.32463   0.243  0.80792
## agegp^5     -0.26219    0.21337  -1.229  0.21915
## alcgp.L      2.53899    0.26385   9.623  < 2e-16 ***
## alcgp.Q      0.09376    0.22419   0.418  0.67578
## alcgp.C      0.43930    0.18347   2.394  0.01665 *
## tobgp.L      1.11749    0.24014   4.653 3.26e-06 ***
## tobgp.Q      0.34516    0.22414   1.540  0.12358
## tobgp.C      0.31692    0.21091   1.503  0.13294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  82.337  on 76  degrees of freedom
## AIC: 221.39
##
## Number of Fisher Scoring iterations: 6
```

From the results, linear term for agegp, alcgp, and tobgp are significant. Quadratic term for agegp is significant. And cubic term for alcgp is significant.

```
finalm <- glm(cbind(ncases, ncontrols) ~
                (unclass(agegp)) +
                (unclass(alcgp)) +
                (unclass(tobgp)) +
                I(unclass(agegp)^2), family = binomial, data = esoph)
summary(finalm)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ (unclass(agegp)) + (unclass(alcgp)) +
##     (unclass(tobgp)) + I(unclass(agegp)^2), family = binomial,
##     data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2757  -0.7828  -0.2313   0.5679   2.4646
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -10.10233    1.03074  -9.801  < 2e-16 ***
## unclass(agegp)        2.50576    0.50188   4.993 5.95e-07 ***
## unclass(alcgp)        1.06511    0.10458  10.185  < 2e-16 ***
## unclass(tobgp)        0.43951    0.09559   4.598 4.27e-06 ***
## I(unclass(agegp)^2)  -0.23417    0.06402  -3.658 0.000255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
```

```
## Residual deviance:  93.172  on 83  degrees of freedom
## AIC: 218.23
##
## Number of Fisher Scoring iterations: 5
```

e. Does your final model fit the data? Is the test you use appropriate for this data?

Solution:

```
pchisq(finalm$null.deviance - finalm$deviance, finalm$df.residual, lower = FALSE)
```

```
## [1] 2.250667e-22
```

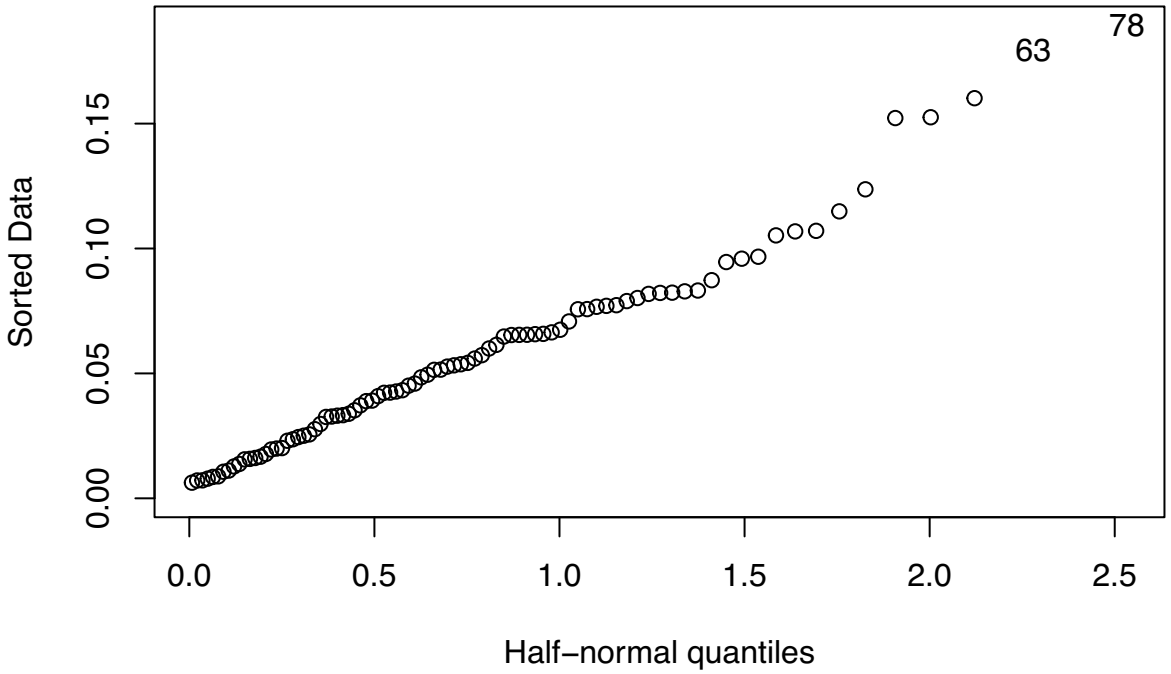
From the results, the final model fit the data better than the null model.

f. Check for outliers in your final model.

Solution:

```
library(faraway)
```

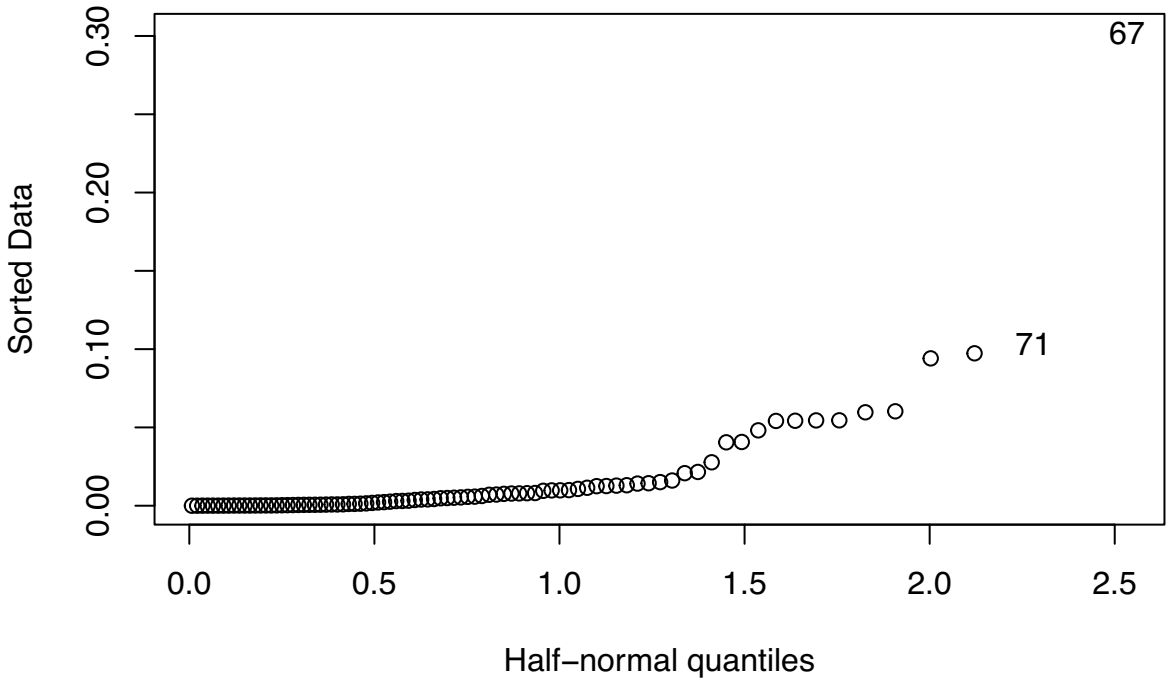
```
halfnorm(hatvalues(finalm))
```




```
esoph %>%
  slice(c(63, 78))
```

```
## # A tibble: 2 x 6
##   agegp alcgp     tobgp    ncases ncontrols nsubjects
##   <ord> <ord>     <ord>     <dbl>     <dbl>     <dbl>
## 1 65-74 0-39g/day 0-9g/day      5        43        48
## 2 75+   0-39g/day 0-9g/day      1        17        18
```

```
halfnorm(cooks.distance(finalm))
```




```
esoph %>%
  slice(c(67, 71))
```

```
## # A tibble: 2 x 6
##   agegp alcgp  tobgp    ncases ncontrols nsubjects
##   <ord> <ord>  <ord>     <dbl>     <dbl>     <dbl>
## 1 65-74 40-79  0-9g/day     17        17        34
## 2 65-74 80-119 10-19         4         8        12
```

From the plots, groups of 63 and 78 are high leverage. And for groups of 67 and 71 are high influential observations.

g. What is the predicted effect of moving one category higher in alcohol consumption?

Solution:

```
coef(finalm)
```

```
##         (Intercept)      unclass(agegp)      unclass(alcgp)      unclass(tobgp)
##         -10.1023336           2.5057645           1.0651087           0.4395077
## I(unclass(agegp)^2)
##          -0.2341676
```

```
exp(1.0651087 )
```

```
## [1] 2.901154
```

By looking at the coefficients, alcohol consumption with one category higher will increase the outcome by 1.065 higher with other predictor under control. And odds increase 2.901164.

h. Compute a 95% confidence interval for this predicted effect.

Solution:

```
library(gtsummary)
finalm %>%
  tbl_regression() %>%
  bold_labels() %>%
  bold_p(t = 0.05)
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	log(OR)	95% CI	p-value
unclass(agegp)	2.5	1.6, 3.5	**<0.001**
unclass(alcgp)	1.1	0.86, 1.3	**<0.001**
unclass(tobgp)	0.44	0.25, 0.63	**<0.001**
I(unclass(agegp)^2)	-0.23	-0.37, -0.11	**<0.001**