Biostat 200C Homework 1

Due Apr 12 @ 11:59PM

Yenlin Lai

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 $Var(Y_i)$ is larger than the Binomial variance (unless $n_i = 1$ or $\phi = 0$).

Solution: See the pdf file attached.

Q2. (ELMR Chapter 3 Exercise 1)

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

```
data(esoph)
help(esoph)
```

```
## starting httpd help server ... done
```

```
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
```

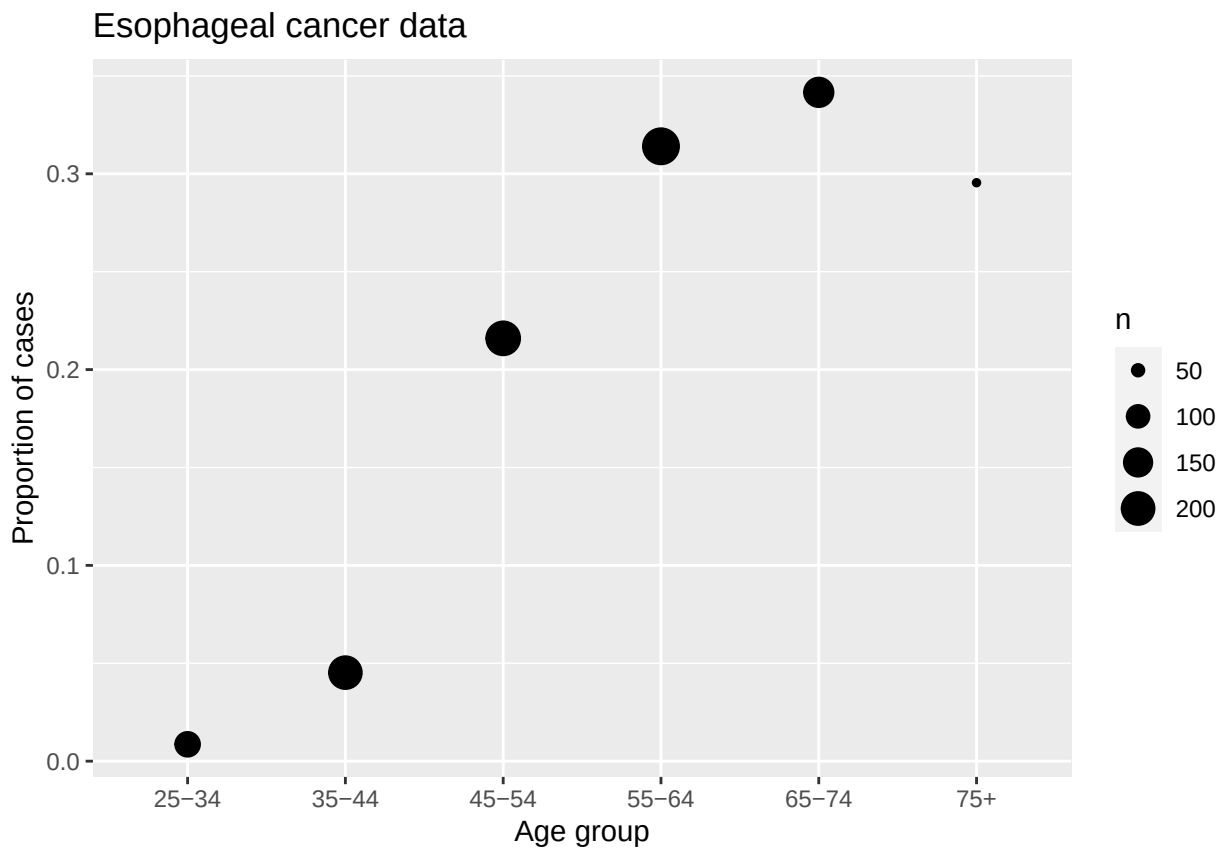
```
esoph <- esoph %>%
  as_tibble() %>%
  print()
```

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

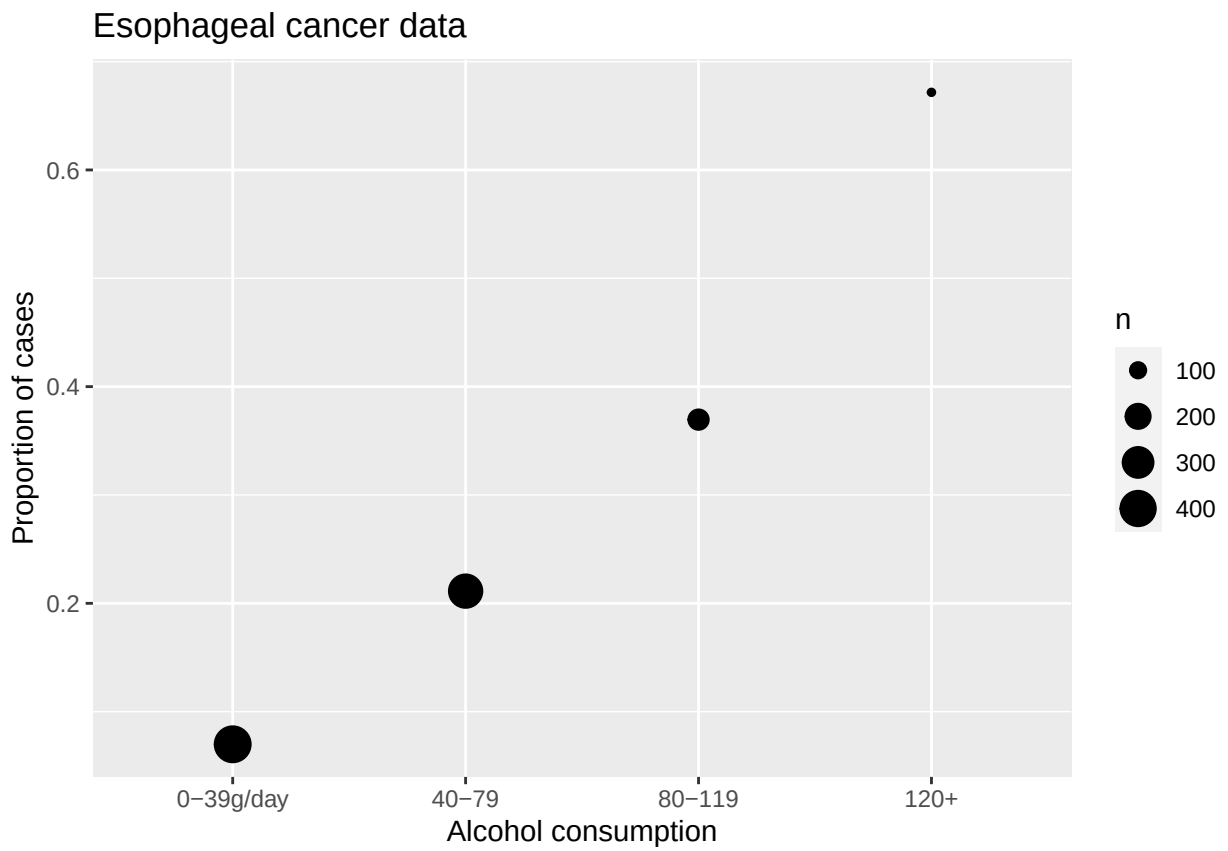
Solution:

```
library(ggplot2)
esoph %>%
  group_by(agegp) %>%
  summarise(ncases = sum(ncases), ncontrols = sum(ncontrols)) %>%
  mutate(n = ncases + ncontrols,
         rate_agegp = ncases / n) %>%
  ggplot(mapping = aes(x = agegp, y = rate_agegp, size = n)) +
  geom_point() +
  labs(x = "Age group", y = "Proportion of cases",
       title = "Esophageal cancer data")
```



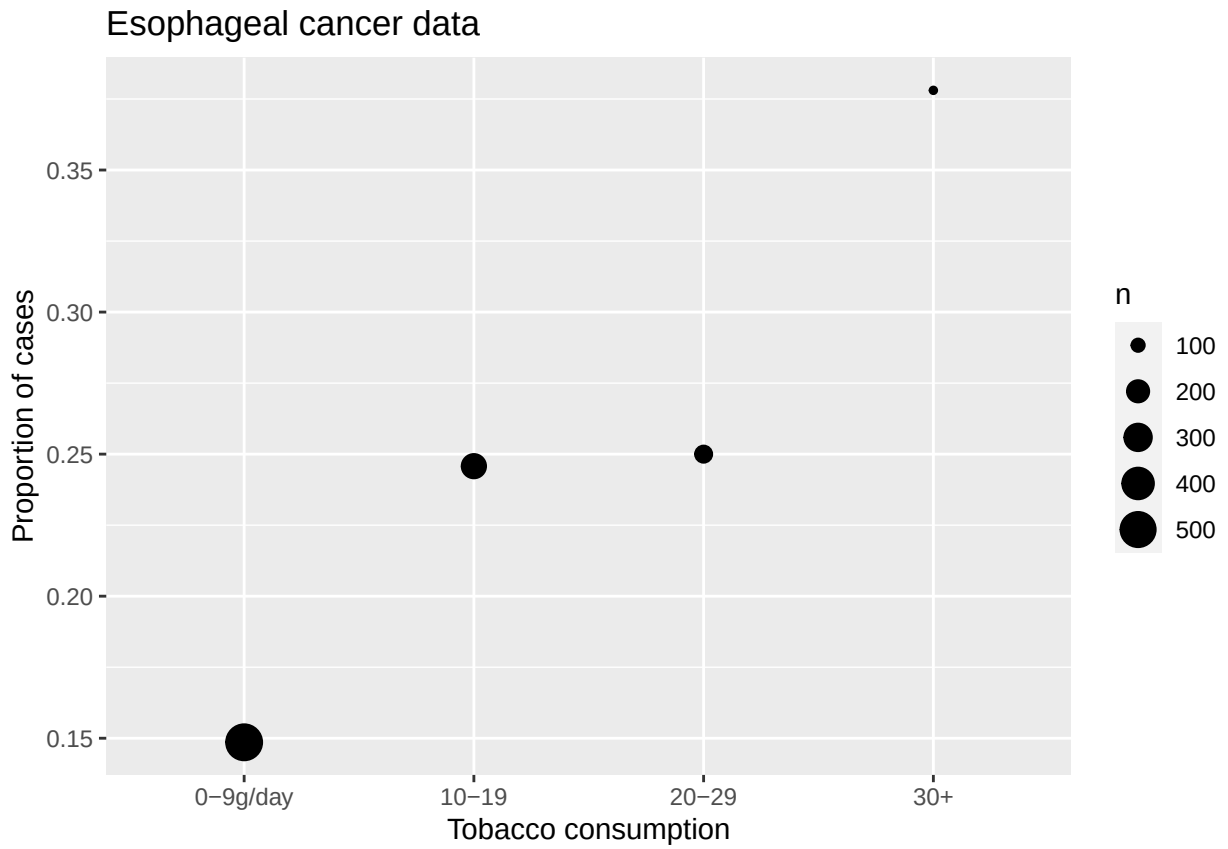
From the proportion of cases against age group plot shown above, higher age could lead to higher proportion of cases, i.e., older people were more likely to get esophageal cancer. Note that for age 75+ group, the proportion of cases dropped due to smaller sample size.

```
esoph %>%
  group_by(alcgp) %>%
  summarise(ncases = sum(ncases), ncontrols = sum(ncontrols)) %>%
  mutate(n = ncases + ncontrols,
         rate_alcgp = ncases / n) %>%
  ggplot(mapping = aes(x = alcgp, y = rate_alcgp, size = n)) +
  geom_point() +
  labs(x = "Alcohol consumption", y = "Proportion of cases",
       title = "Esophageal cancer data")
```



From the proportion of cases against alcohol consumption plot shown above, higher alcohol consumption per day could lead to higher proportion of cases, i.e., consuming more alcohol was more likely to get a esophageal cancer.

```
esoph %>%
  group_by(tobgp) %>%
  summarise(ncases = sum(ncases), ncontrols = sum(ncontrols)) %>%
  mutate(n = ncases + ncontrols,
         rate_tobgp = ncases / n) %>%
  ggplot(mapping = aes(x = tobgp, y = rate_tobgp, size = n)) +
  geom_point() +
  labs(x = "Tobacco consumption", y = "Proportion of cases",
       title = "Esophageal cancer data")
```



From the proportion of cases against tobacco consumption plot shown above, higher tobacco consumption per day could lead to higher proportion of cases, i.e., consuming more tobacco was more likely to get a esophageal cancer. Note that there seems to be no differences in proportion of cases between consuming 10-19 and 20-29 gm per day.

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:

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

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ (agegp + alcgp + tobgp)^2,
##     family = "binomial", data = esoph)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.92598  -0.32291  -0.00004   0.32199   1.49642
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)     -7.909e+00  1.913e+03  -0.004    0.997
## agegp.L          3.105e+01  6.604e+03   0.005    0.996
## agegp.Q         -1.686e+01  5.909e+03  -0.003    0.998
## agegp.C          9.987e+00  4.507e+03   0.002    0.998
## agegp^4          1.358e-01  3.003e+03   0.000    1.000
## agegp^5         -5.981e-01  1.400e+03   0.000    1.000
## alcgp.L          8.885e+00  2.529e+03   0.004    0.997
## alcgp.Q          4.603e+00  2.805e+03   0.002    0.999
## alcgp.C          2.348e+00  3.057e+03   0.001    0.999
## tobgp.L         -3.560e+00  2.057e+03  -0.002    0.999
## tobgp.Q         -2.849e+00  2.520e+03  -0.001    0.999
## tobgp.C          6.216e+00  2.910e+03   0.002    0.998
## agegp.L:alcgp.L -2.933e+00  9.000e+03   0.000    1.000
## agegp.Q:alcgp.L  2.343e+01  8.185e+03   0.003    0.998
## agegp.C:alcgp.L -3.667e+00  5.716e+03  -0.001    0.999
## agegp^4:alcgp.L  1.147e+01  3.119e+03   0.004    0.997
## agegp^5:alcgp.L -2.893e+00  1.190e+03  -0.002    0.998
## agegp.L:alcgp.Q -1.522e+01  9.755e+03  -0.002    0.999
## agegp.Q:alcgp.Q  8.917e+00  8.763e+03   0.001    0.999
## agegp.C:alcgp.Q -1.794e+00  6.547e+03   0.000    1.000
## agegp^4:alcgp.Q -2.618e+00  4.201e+03  -0.001    1.000
## agegp^5:alcgp.Q  2.095e+00  1.907e+03   0.001    0.999
## agegp.L:alcgp.C -1.558e+01  1.045e+04  -0.001    0.999
## agegp.Q:alcgp.C -2.583e+00  9.304e+03   0.000    1.000
## agegp.C:alcgp.C  8.464e-01  7.284e+03   0.000    1.000
## agegp^4:alcgp.C -8.098e+00  5.057e+03  -0.002    0.999
## agegp^5:alcgp.C  2.167e+00  2.420e+03   0.001    0.999
## agegp.L:tobgp.L  7.471e+00  6.357e+03   0.001    0.999
## agegp.Q:tobgp.L -6.707e+00  5.278e+03  -0.001    0.999
## agegp.C:tobgp.L -5.028e+00  5.414e+03  -0.001    0.999
## agegp^4:tobgp.L  5.445e+00  4.787e+03   0.001    0.999
## agegp^5:tobgp.L -3.663e+00  2.536e+03  -0.001    0.999
## agegp.L:tobgp.Q  2.202e+01  8.595e+03   0.003    0.998
## agegp.Q:tobgp.Q -3.549e+00  7.638e+03   0.000    1.000
## agegp.C:tobgp.Q  5.695e+00  6.024e+03   0.001    0.999
## agegp^4:tobgp.Q  4.773e+00  4.228e+03   0.001    0.999
## agegp^5:tobgp.Q -1.378e+00  2.036e+03  -0.001    0.999
## agegp.L:tobgp.C -4.815e+00  1.036e+04   0.000    1.000
## agegp.Q:tobgp.C  2.222e+01  9.424e+03   0.002    0.998
## agegp.C:tobgp.C -5.600e+00  6.577e+03  -0.001    0.999
## agegp^4:tobgp.C  1.050e+01  3.582e+03   0.003    0.998
## agegp^5:tobgp.C -1.673e+00  1.363e+03  -0.001    0.999
## alcgp.L:tobgp.L -5.630e-01  6.863e-01  -0.820    0.412
## alcgp.Q:tobgp.L  3.234e-02  6.590e-01   0.049    0.961
## alcgp.C:tobgp.L -2.149e-01  5.583e-01  -0.385    0.700
## alcgp.L:tobgp.Q  7.099e-01  6.383e-01   1.112    0.266
## alcgp.Q:tobgp.Q -3.413e-01  5.920e-01  -0.577    0.564
## alcgp.C:tobgp.Q  1.587e-01  4.987e-01   0.318    0.750
## alcgp.L:tobgp.C -2.793e-01  5.684e-01  -0.491    0.623
## alcgp.Q:tobgp.C -6.690e-02  5.119e-01  -0.131    0.896
## alcgp.C:tobgp.C -3.457e-01  4.402e-01  -0.785    0.432
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  30.824  on 37  degrees of freedom
## AIC: 247.88
##
## Number of Fisher Scoring iterations: 20
```

```
step(modelb)
```

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

```
##
## Call:  glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = "binomial", data = esoph)
##
## Coefficients:
## (Intercept)      agegp.L      agegp.Q      agegp.C      agegp^4      agegp^5
##    -1.19039      3.99663     -1.65741      0.11094      0.07892     -0.26219
##     alcgp.L      alcgp.Q      alcgp.C      tobgp.L      tobgp.Q      tobgp.C
##     2.53899      0.09376      0.43930      1.11749      0.34516      0.31692
##
```

```
## Degrees of Freedom: 87 Total (i.e. Null);  76 Residual
## Null Deviance:       368
## Residual Deviance: 82.34     AIC: 221.4
```

The model below is selected using AIC as the criterion (smallest AIC).

```
modelb.final <-
  glm(cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
      data = esoph, family = "binomial")
summary(modelb.final)
```

```
##
## 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
```

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:

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

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(tobgp) +
##     unclass(alcgp), family = binomial(), data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6478  -0.9246  -0.4338   0.6740   2.4568
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.16395    0.50931 -14.066  < 2e-16 ***
## unclass(agegp)  0.74375    0.08179   9.094  < 2e-16 ***
## unclass(tobgp)  0.43085    0.09394   4.587 4.51e-06 ***
## unclass(alcgp)  1.10255    0.10317  10.687  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.95  on 87  degrees of freedom
## Residual deviance: 108.78  on 84  degrees of freedom
## AIC: 231.83
##
## Number of Fisher Scoring iterations: 4
```

The model has been simplified when we convert the factors to the numerical representation. The quadratic, cubic and other terms are droppped.

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: From the factor model in part b, we can observe that the estimated coefficients for agegp.L, agegp.Q, alcgp.L, alcgp.C, and tobgp.L are significant. We then include the agegp, agegp^2, alcgp, and tobgp in the new linear model which is slightly more complex than the previous one. Note that the predictor alcgp^3 is not included since its lower effect alcgp^2 is not significant; in that case, include alcgp^3 would not make sense.

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

```
##
## Call:
```

```
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + I(unclass(agegp)^2) +
##     unclass(tobgp) + unclass(alcgp), 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 ***
## I(unclass(agegp)^2)  -0.23417    0.06402  -3.658 0.000255 ***
## unclass(tobgp)        0.43951    0.09559   4.598 4.27e-06 ***
## unclass(alcgp)        1.06511    0.10458  10.185  < 2e-16 ***
## ---
## 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(modeld$deviance, modeld$df.residual, lower = FALSE)
```

```
## [1] 0.2087964
```

By the analysis of deviance, we get the p-value to be 0.209, which is greater than the significant level α 0.05. Thus, we do not reject the null hypothesis, and conclude that the final model fits the data. In this data, the distribution is binomial, and the sample size for each group is not always large (≥ 5). Therefore, the χ^2 test is inappropriate for this data.

```
esoph <-
  esoph %>% mutate(n = ncases + ncontrols)
predprob <- predict(modeld, type = "response")
px2 <- sum((esoph$ncases - esoph$n * predprob)^2 / (esoph$n * predprob * (1 - predprob)))
pchisq(px2, modeld$df.residual, lower.tail = FALSE)
```

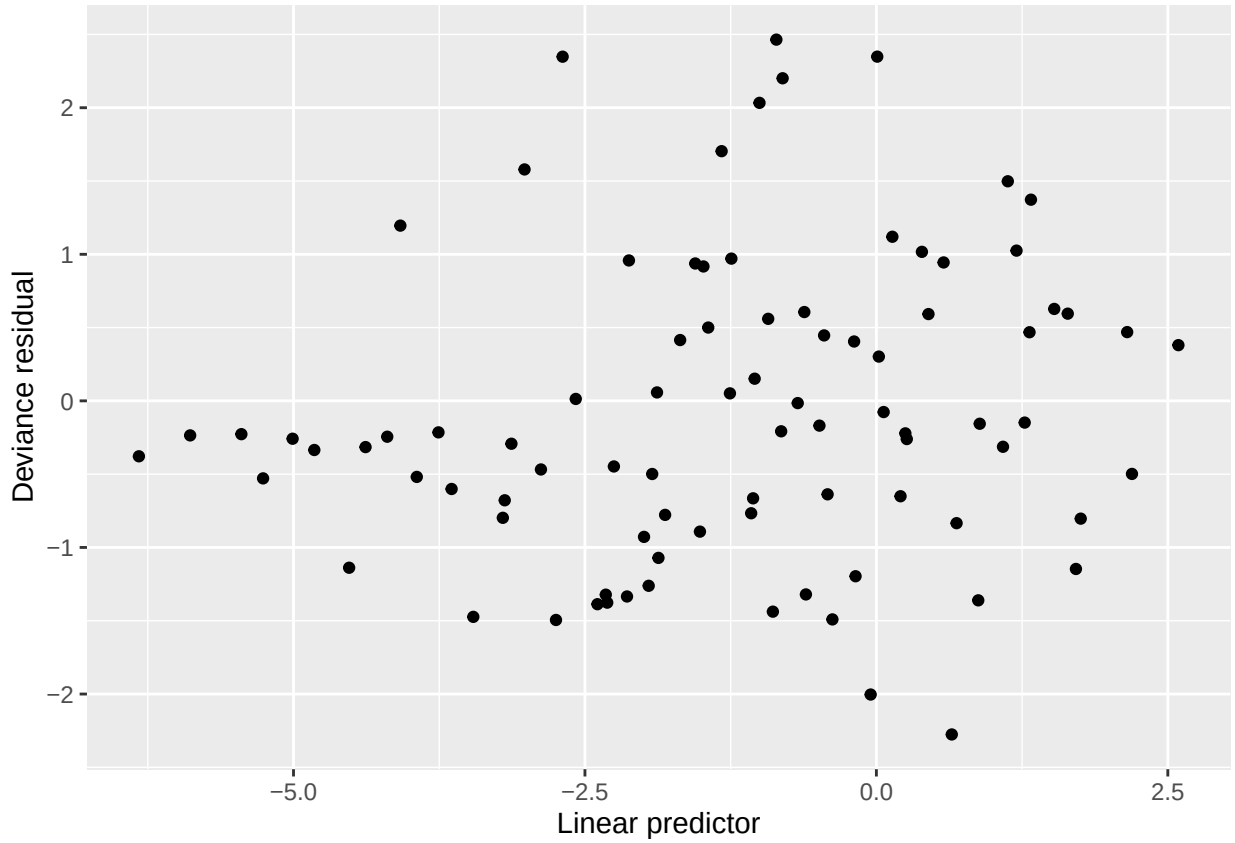
```
## [1] 0.2450936
```

We then use Pearson χ^2 test to test the goodness of fit for this data. The p-value is 0.245, which is greater than the significant level α 0.05. We do not reject the null hypothesis (the fitted model equals to saturated model), and conclude that the final model fits the data.

f. Check for outliers in your final model.

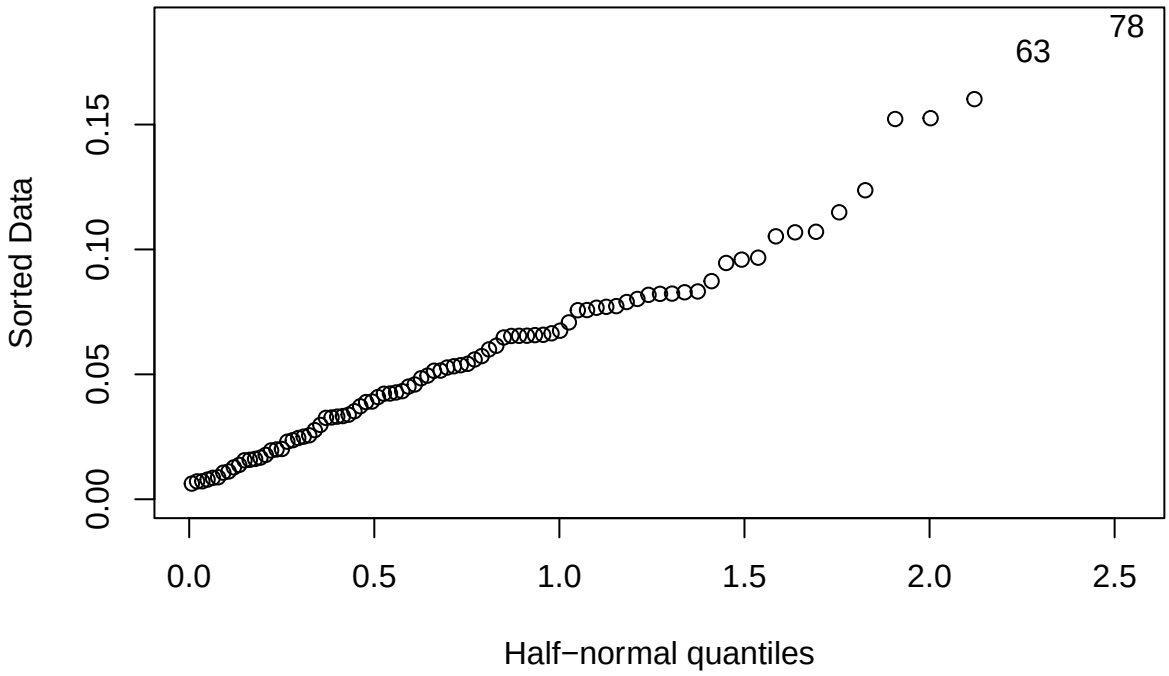
Solution:

```
esoph %>%
  mutate(devres = residuals(modeld, type = "deviance"),
         linpred = predict(modeld, type = "link")) %>%
  ggplot(mapping = aes(x = linpred, y = devres)) +
  geom_point() +
  labs(x = "Linear predictor", y = "Deviance residual")
```



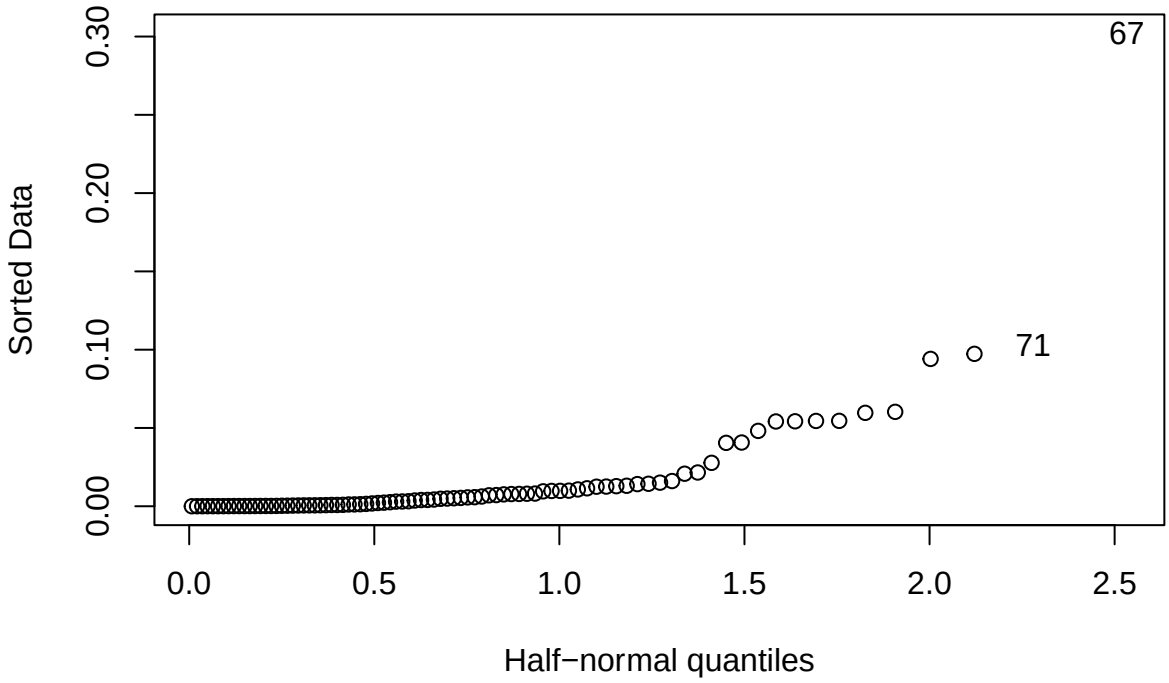
From the deviance residual plot, we can identify potential high residual observations. We do not find outliers here. We further determine which observations have high leverage or influence by using the following two plots.

```
library(faraway)
halfnorm(hatvalues(modeld))
```



From the plot sorted hat values against half-normal quantiles, we can identify potential high leverage observations are 63(age: 65-74 years, alcohol consumption: 0-39 gm/day, tobacco consumption: 0-9 gm/day) and 78(age: 75+ years, alcohol consumption: 0-39 gm/day, tobacco consumption: 0-9 gm/day).

```
halfnorm(cooks.distance(modeld))
```



From the plot sorted Cook distances against the half-normal quantiles, we can identify potential high influential observations are 71(age: 65-74 years, alcohol consumption: 80-119 gm/day, tobacco consumption: 10-19 gm/day) and 67(age: 65-74 years, alcohol consumption: 40-79 gm/day, tobacco consumption: 0-9 gm/day).

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

Solution:

```
library(gtsummary)
modeld %>%
  tbl_regression(intercept = TRUE, exponentiate = TRUE)
```

```
## 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	OR	95% CI	p-value
(Intercept)	0.00	0.00, 0.00	<0.001
unclass(agegp)	12.3	4.81, 34.6	<0.001
I(unclass(agegp)^2)	0.79	0.69, 0.89	<0.001
unclass(tobgp)	1.55	1.29, 1.87	<0.001
unclass(alcgp)	2.90	2.37, 3.58	<0.001

From the table above, we can conclude that moving one category higher in alcohol consumption has the odds of being more likely to get a esophageal cancer is multiplied 2.90 times (i.e., increases 190%), holding constant all other variables.

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

Solution:

```
modeld %>%
  tbl_regression(intercept = TRUE, exponentiate = TRUE)
```

```
## 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	OR	95% CI	p-value
(Intercept)	0.00	0.00, 0.00	<0.001
unclass(agegp)	12.3	4.81, 34.6	<0.001
I(unclass(agegp)^2)	0.79	0.69, 0.89	<0.001
unclass(tobgp)	1.55	1.29, 1.87	<0.001
unclass(alcgp)	2.90	2.37, 3.58	<0.001

From the same table above, we can conclude that the 95% confidence interval for the odds of being more likely to get a esophageal cancer is multiplied 2.37 to 3.58 times when moves one category higher in alcohol consumption, holding constant all other variables.