Impact of Lifestyle and Physical Indicators on First Diagnosis of Cancer

Cong Lyu Christine Xing Shahzab Hussain

Import data

```
library(ipumsr)

ddi <- read_ipums_ddi("nhis_00003.xml")
data <- read_ipums_micro(ddi)
```

```
## Use of data from IPUMS NHIS is subject to conditions including that users should cite the data appro
```

```
names(data)
```

```
##  [1] "YEAR"         "SERIAL"       "STRATA"       "PSU"          "NHISHID"
##  [6] "PERNUM"       "NHISPID"      "HHX"          "SAMPWEIGHT"   "LONGWEIGHT"
## [11] "PARTWEIGHT"   "ASTATFLG"     "CSTATFLG"     "AGE"          "SEX"
## [16] "EDUC"         "MAXEDUC"      "SPOUSEDUC"    "EMPSTAT"      "HOURSWRK"
## [21] "EMPHI"        "EMPFT"        "SPOUSWKFT"    "WRKADLTNO"    "QTCINCFAM"
## [26] "IMPINCFAM"    "CPI2009"      "HEALTH"       "HEIGHT"       "WEIGHT"
## [31] "BMICALC"      "CNBRAN"       "CNBRANAG"     "CNCOLN"       "CNCOLNAG"
## [36] "CNCOLRECT"    "CNCOLRECTAG"  "CNESOP"       "CNESOPAG"     "CNHDNCK"
## [41] "CNHDNCKAG"    "CNLIVR"       "CNLIVRAG"     "CNPANC"       "CNPANCAG"
## [46] "CNSTOM"       "CNSTOMAG"     "DIABETICAGE"  "DIABTYPE"     "ALCDRINKEV"
## [51] "ALC5UPEVYR"   "ALCEV30D"     "ALC5UPOCC30D" "ALCANYTP"     "MOD10FNO"
## [56] "MOD10FWK"
```

```
col.ages <- c("CNBRANAG", "CNCOLNAG", "CNCOLRECTAG", "CNESOPAG", "CNHDNCKAG",
              "CNLIVRAG", "CNPANCAG", "CNSTOMAG")
col.cancers <- c("CNBRAN", "CNCOLN", "CNCOLRECT", "CNESOP", "CNHDNCK",
                 "CNLIVR", "CNPANC", "CNSTOM")
col.numeric.categorical <- c("EDUC", "MAXEDUC", "SPOUSEDUC")
col.categorical <- c("SEX", "QTCINCFAM")
col.key <- c("SERIAL", "NHISHID", "NHISPID", "HHX")
```

Topic 1

Response: The age of having cancer for the first time among these kinds of cancer

EDA

```
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
```

```
# Data cleaning: figure out missing values
data.clean <- data %>%
  mutate(across(col.ages, ~if_else(. >= 96, NA_real_, .))) %>%
  # mutate(across(c("EMPSTAT"), ~if_else(. >= 900, NA_real_, .))) %>%
  # mutate(across(c("HOURSWRK"), ~if_else(. >= 97, NA_real_, .))) %>%
  # mutate(across(c("EMPHI", "EMPFT", "SPOUSWKFT", "HEALTH", "ALCDRINKEV"),
  #               ~if_else(. >= 7, NA_real_, .))) %>%
  mutate(across(c("DIABTYPE"), ~if_else(. >= 7, NA_real_, .))) %>%
  mutate(across(c("HEIGHT"), ~if_else(. >= 95, NA_real_, .))) %>%
  mutate(across(c("WEIGHT", "BMICALC"), ~if_else(. >= 995, NA_real_, .))) %>%
  mutate(across(c("ALCDRINKEV", "ALCEV30D"), ~if_else(. >= 7, NA_real_, .))) %>%
  mutate(across(c("ALCEV30D"), ~if_else(. ==0, 1, .))) %>%
  mutate(across(c("ALC5UPOCC30D"), ~if_else(. >= 97, NA_real_, .))) %>%
  mutate(across(c("ALC5UPEVYR"),  ~if_else((. >= 7) | (. ==0), NA_real_, .))) %>%
  mutate(across(c("MOD10FNO"), ~if_else(. >= 995, NA_real_, .))) %>%
  mutate(across(c("MOD10FWK"), ~if_else((. >= 94) & (. <=96), 0, .))) %>%
  mutate(across(c("MOD10FWK"), ~if_else((. ==93) | (. >= 97), NA_real_, .))) %>%
  mutate(across(c("EMPSTAT"), ~if_else(. >= 900, NA_real_, .))) %>%
  mutate(across(c("EMPHI"), ~if_else((. >= 7) | (. ==0), NA_real_, .))) %>%
  mutate(across(c("SPOUSEDUC"), ~if_else((. >= 97) | (. ==0), NA_real_, .)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(col.ages, ~if_else(. >= 96, NA_real_, .))`.
## Caused by warning:
## ! Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(col.ages)
##
##   # Now:
##   data %>% select(all_of(col.ages))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

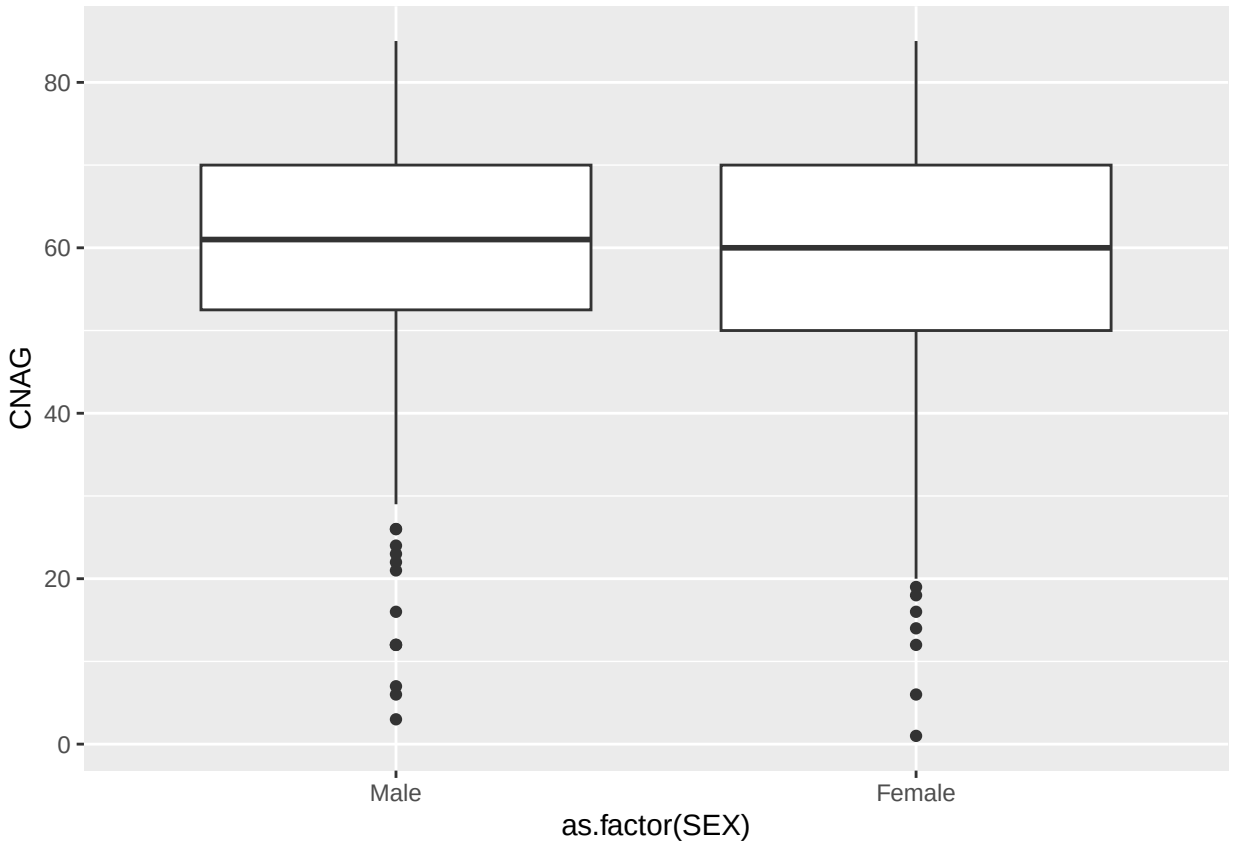
```r
### Check missing values
colSums(is.na(data.clean))
```

```
##         YEAR       SERIAL      STRATA          PSU      NHISHID       PERNUM
##            0            0           0            0            0            0
##      NHISPID          HHX  SAMPWEIGHT  LONGWEIGHT  PARTWEIGHT     ASTATFLG
##            0            0           0        35115        35115            0
##     CSTATFLG          AGE         SEX         EDUC     MAXEDUC    SPOUSEDUC
##            0            0           0            0            0        47010
##      EMPSTAT     HOURSWRK       EMPHI       EMPFT    SPOUSWKFT    WRKADLTNO
##         2159            0       38568            0            0            0
##    QTCINCFAM    IMPINCFAM     CPI2009      HEALTH       HEIGHT       WEIGHT
##            0            0           0            0         4367         5501
##      BMICALC      CNBRAN    CNBRANAG      CNCOLN    CNCOLNAG    CNCOLRECT
##        12795            0       72420            0        72098            0
##  CNCOLRECTAG      CNESOP    CNESOPAG    CNHDNCK    CNHDNCKAG       CNLIVR
##        72054            0       72447            0        72363            0
##     CNLIVRAG      CNPANC    CNPANCAG     CNSTOM     CNSTOMAG  DIABETICAGE
##        72422            0       72433            0        72430            0
##     DIABTYPE   ALCDRINKEV  ALC5UPEVYR    ALCEV30D ALC5UPOCC30D     ALCANYTP
##          339         1334       36274           13          129            0
##     MOD10FNO    MOD10FWK
##        15506         1869
```

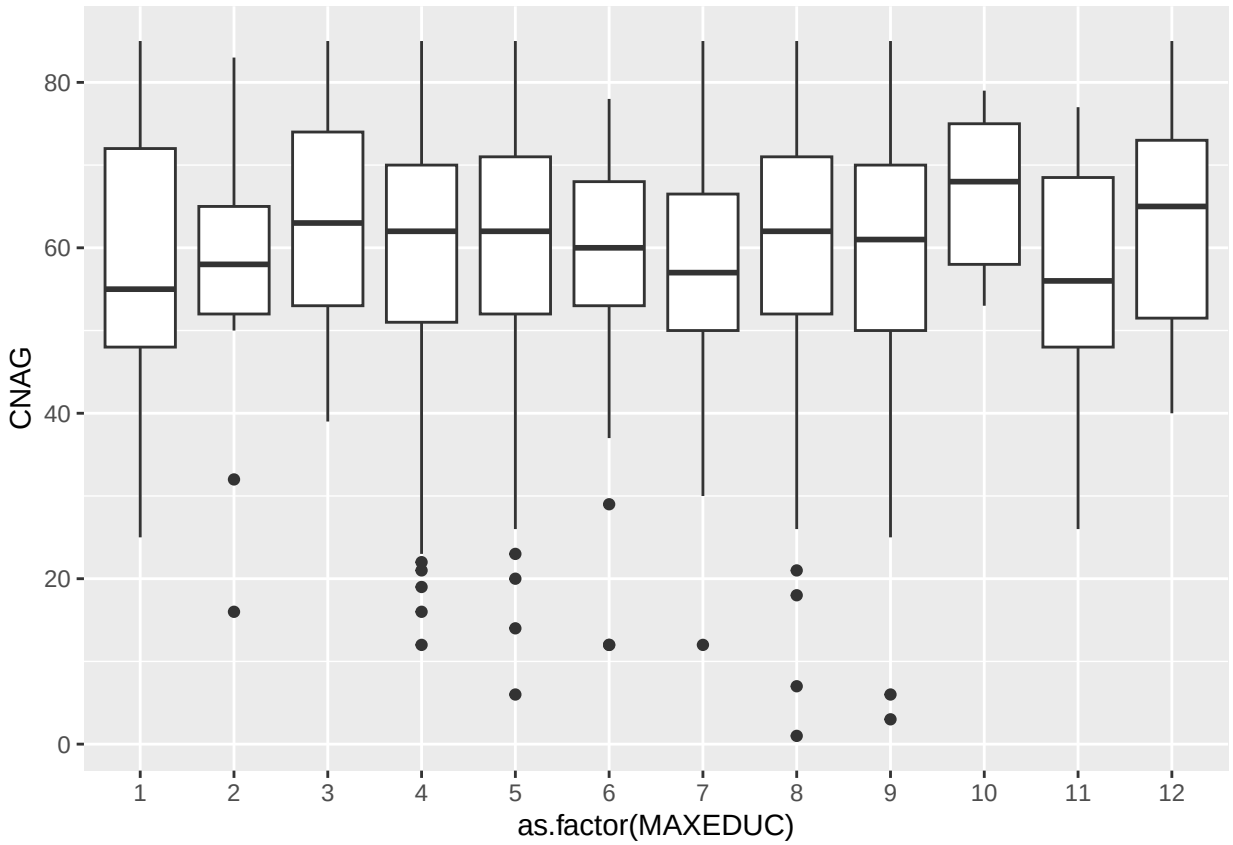
```r
subset.age.clean =  data.clean %>%
  filter(!is.na(CNBRANAG) | !is.na(CNCOLNAG) | !is.na(CNCOLRECTAG) |
           !is.na(CNESOPAG) | !is.na(CNHDNCKAG) | !is.na(CNLIVRAG) |
           !is.na(CNPANCAG) | !is.na(CNSTOMAG)) %>%
  mutate(CNAG = pmin(as.numeric(CNBRANAG), as.numeric(CNCOLNAG),
                     as.numeric(CNCOLRECTAG), as.numeric(CNESOPAG),
                     as.numeric(CNHDNCKAG), as.numeric(CNLIVRAG),
                     as.numeric(CNPANCAG), as.numeric(CNSTOMAG),
                     na.rm = TRUE)) %>%
  mutate(across(c(SEX, EMPSTAT, EMPHI, QTCINCFAM,
                  ALCDRINKEV, ALC5UPEVYR, ALCEV30D, # Binary variables
                  DIABTYPE), # non-numerical factor
                ~as_factor(as_factor(.))))
```

Check info and labels

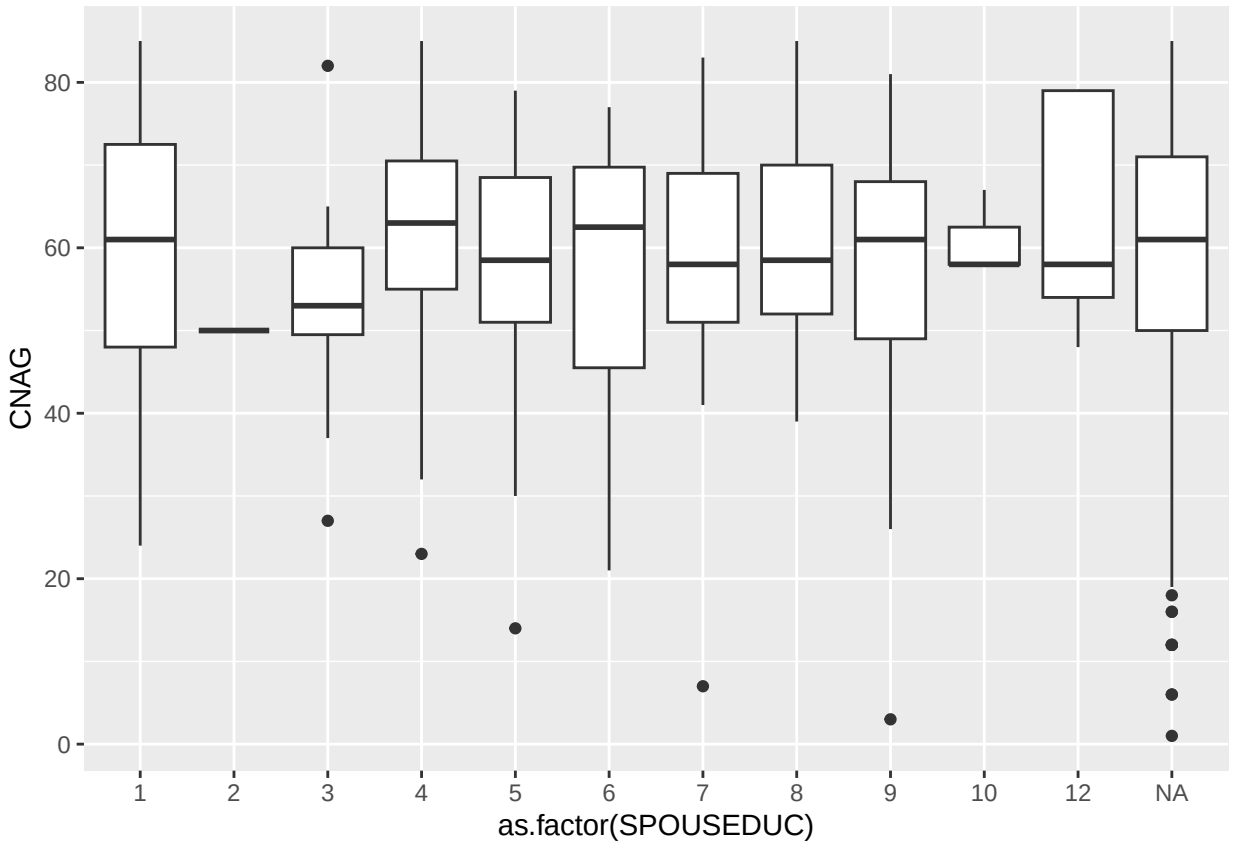
```r
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(SEX), y=CNAG)) + ggplot2::geom_boxplot()
```



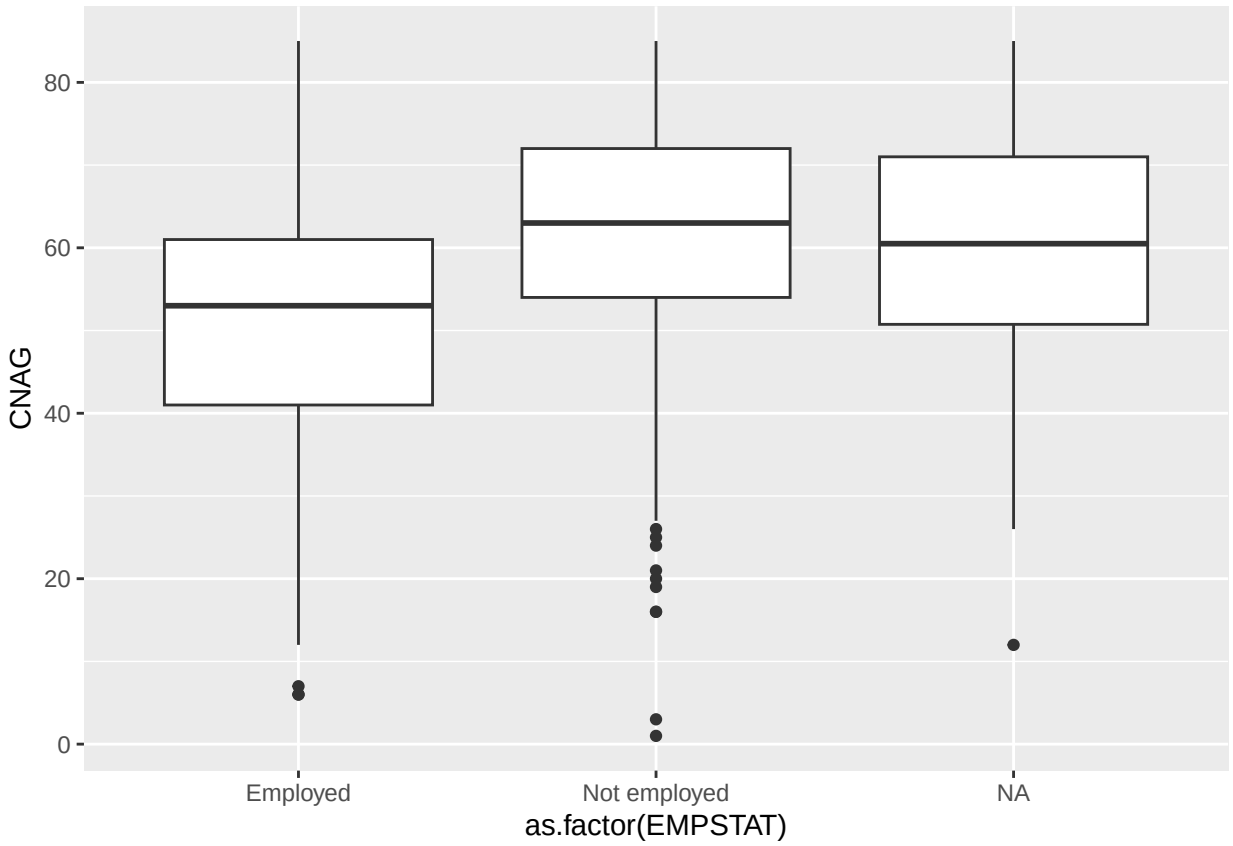

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(MAXEDUC), y=CNAG)) + ggplot2::geom_boxplot
```



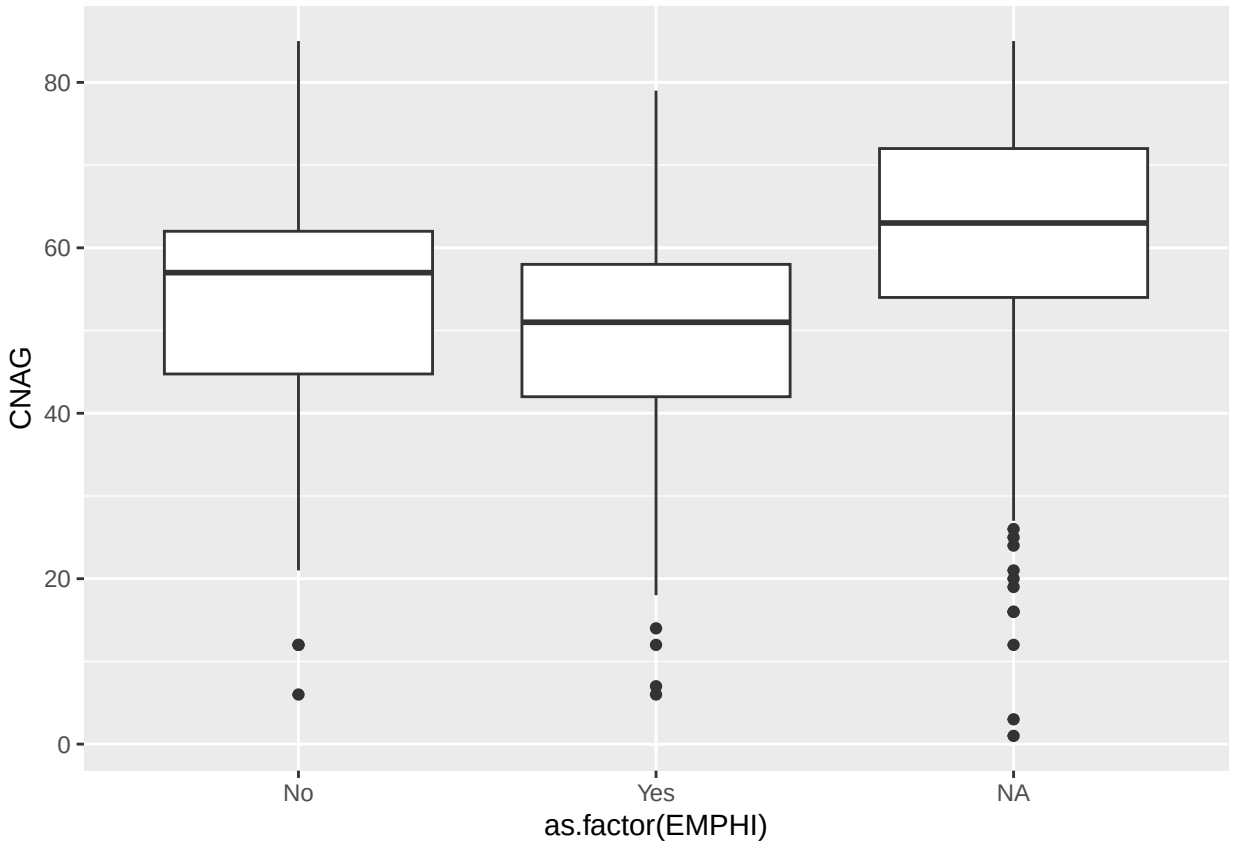

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(SPOUSEDUC), y=CNAG)) + ggplot2::geom_boxpl
```



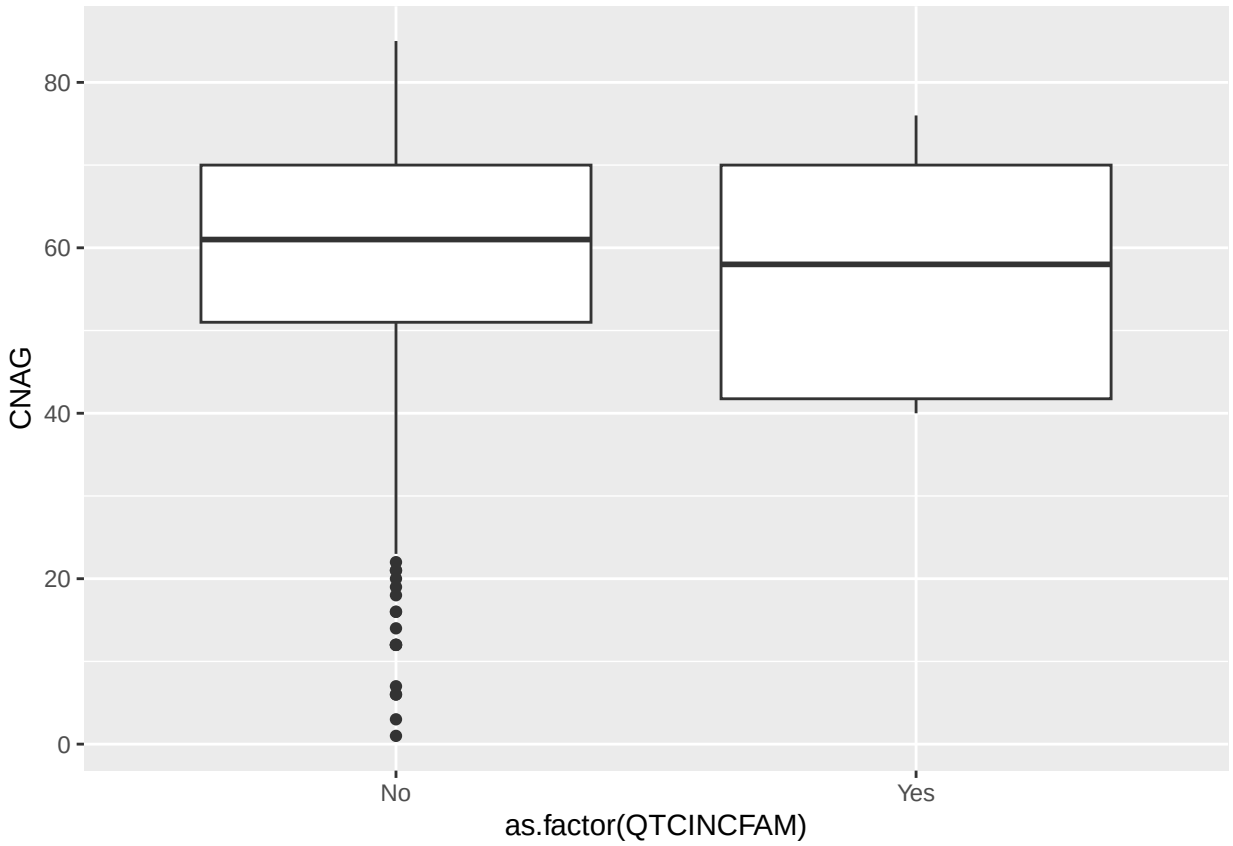

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(EMPSTAT), y=CNAG)) + ggplot2::geom_boxplot
```



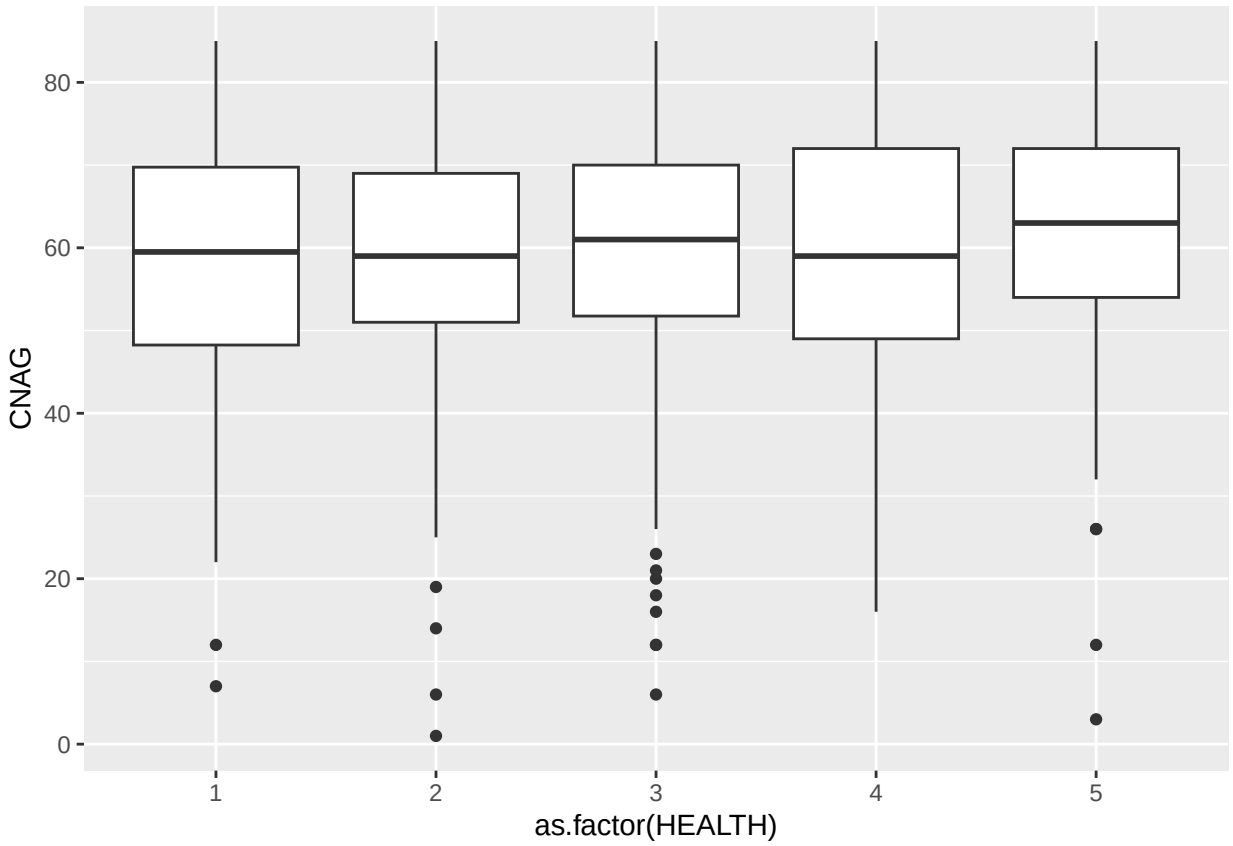

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(EMPHI), y=CNAG)) + ggplot2::geom_boxplot()
```




```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(QTCINCFAM), y=CNAG)) + ggplot2::geom_boxpl
```

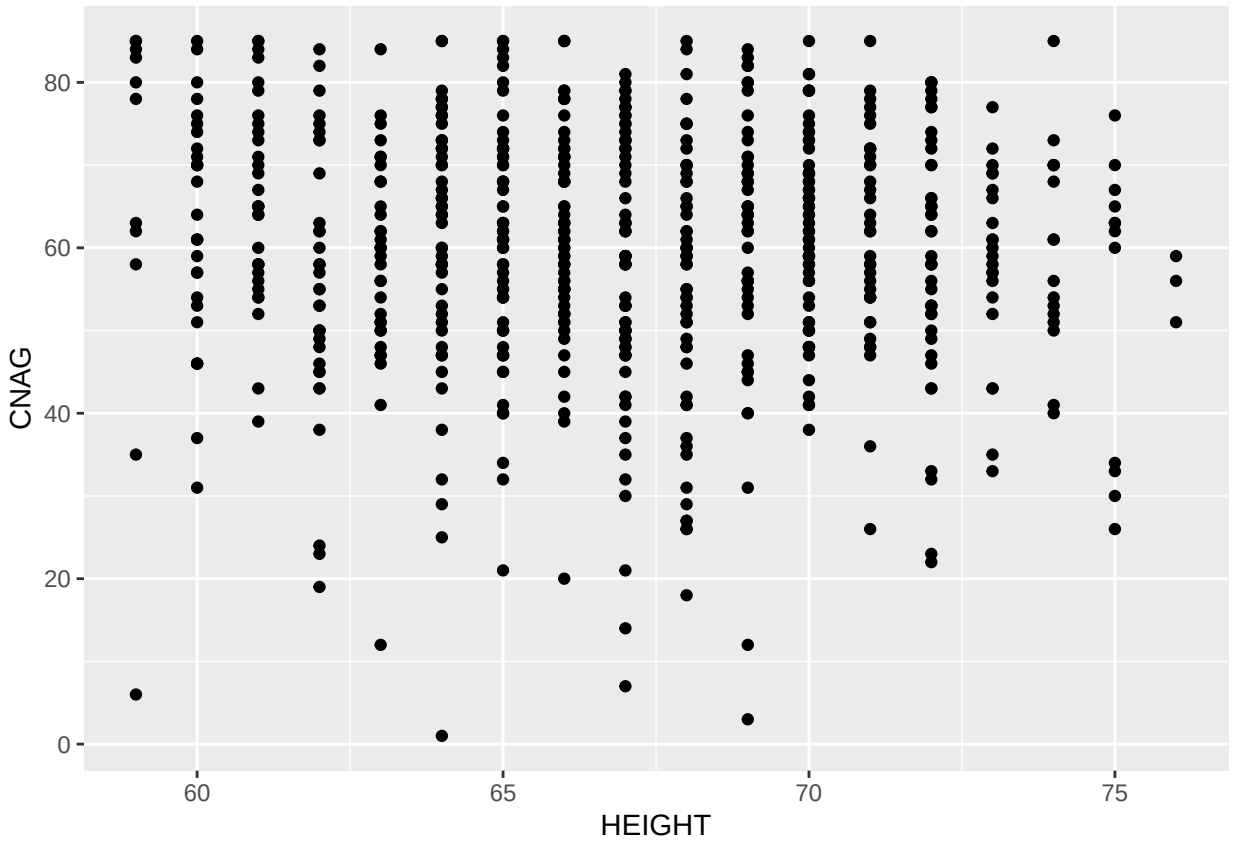



```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(HEALTH), y=CNAG)) + ggplot2::geom_boxplot()
```



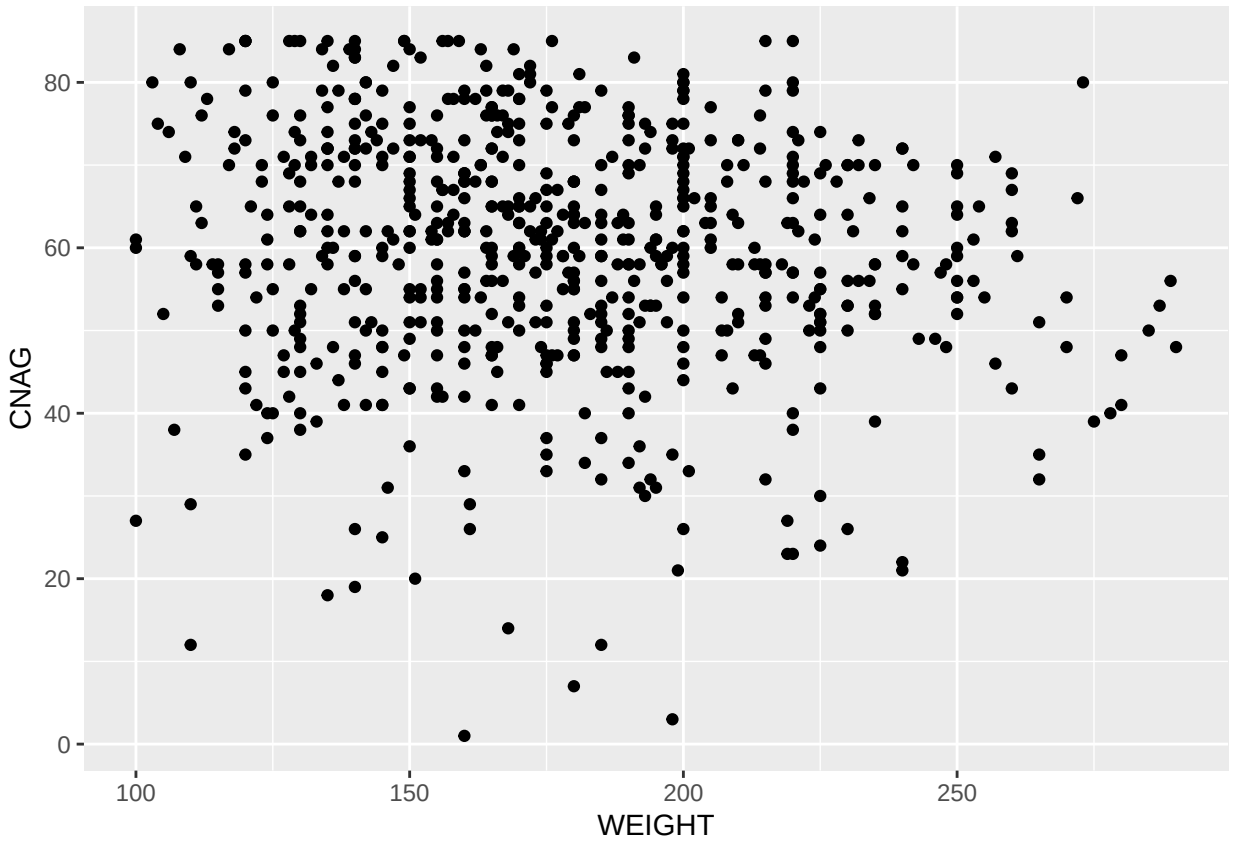

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = HEIGHT, y=CNAG)) + ggplot2::geom_point()
```

```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



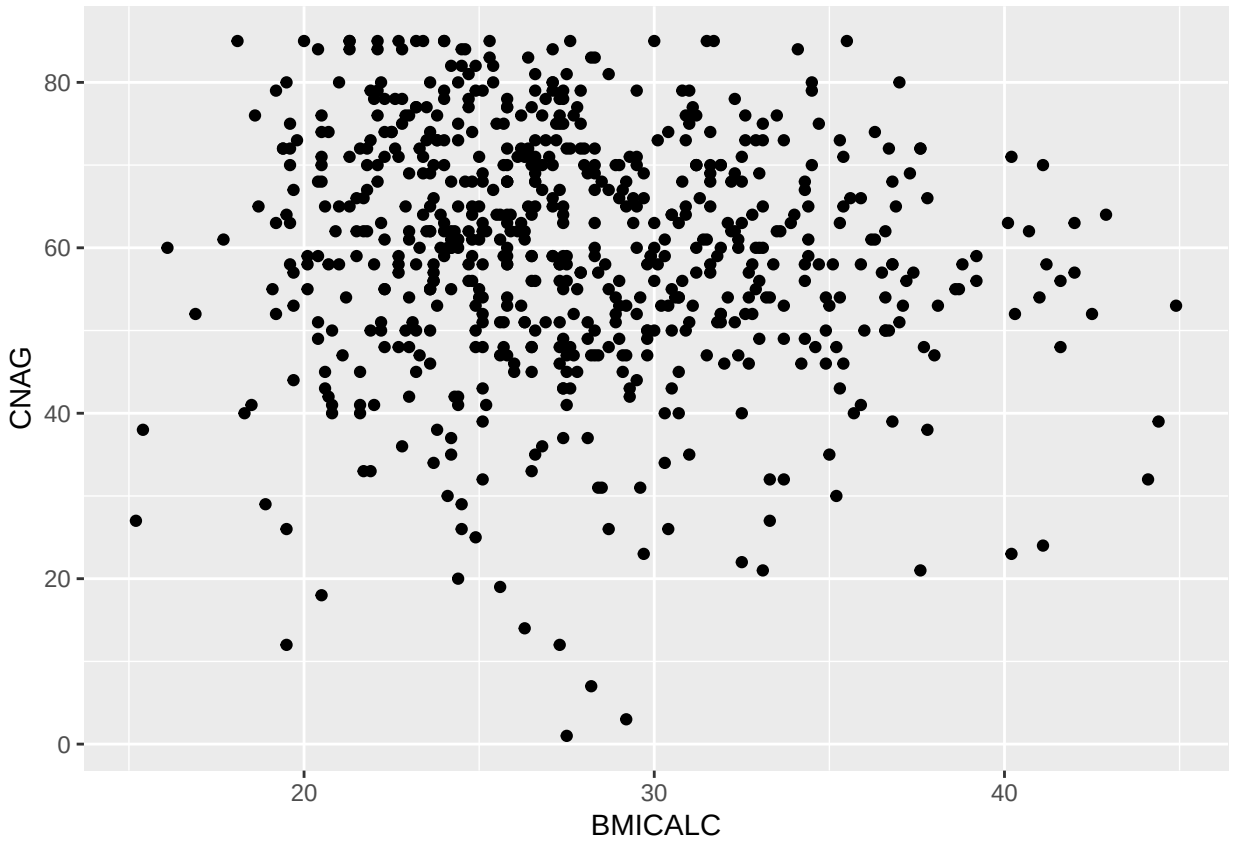

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = WEIGHT, y=CNAG)) + ggplot2::geom_point()
```

```
## Warning: Removed 53 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

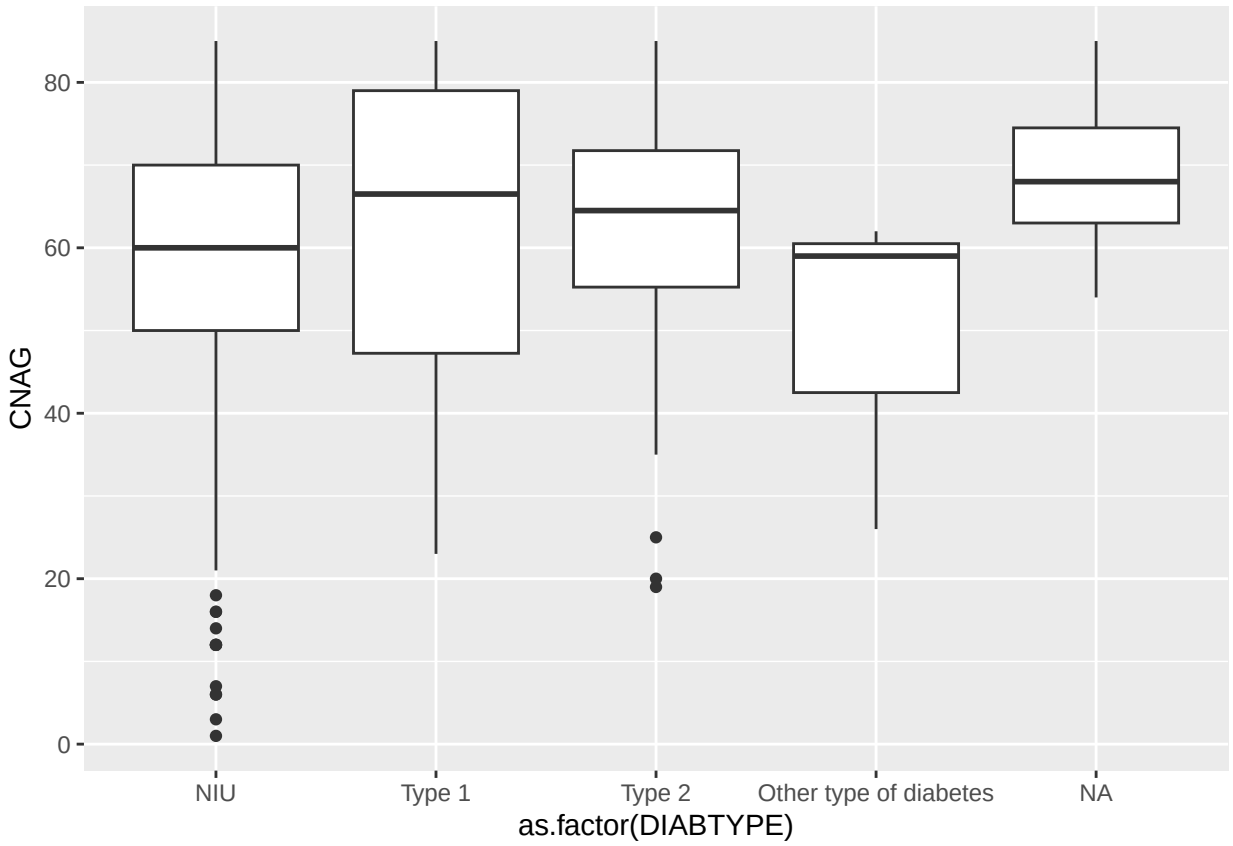



```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = BMICALC, y=CNAG)) + ggplot2::geom_point()
```

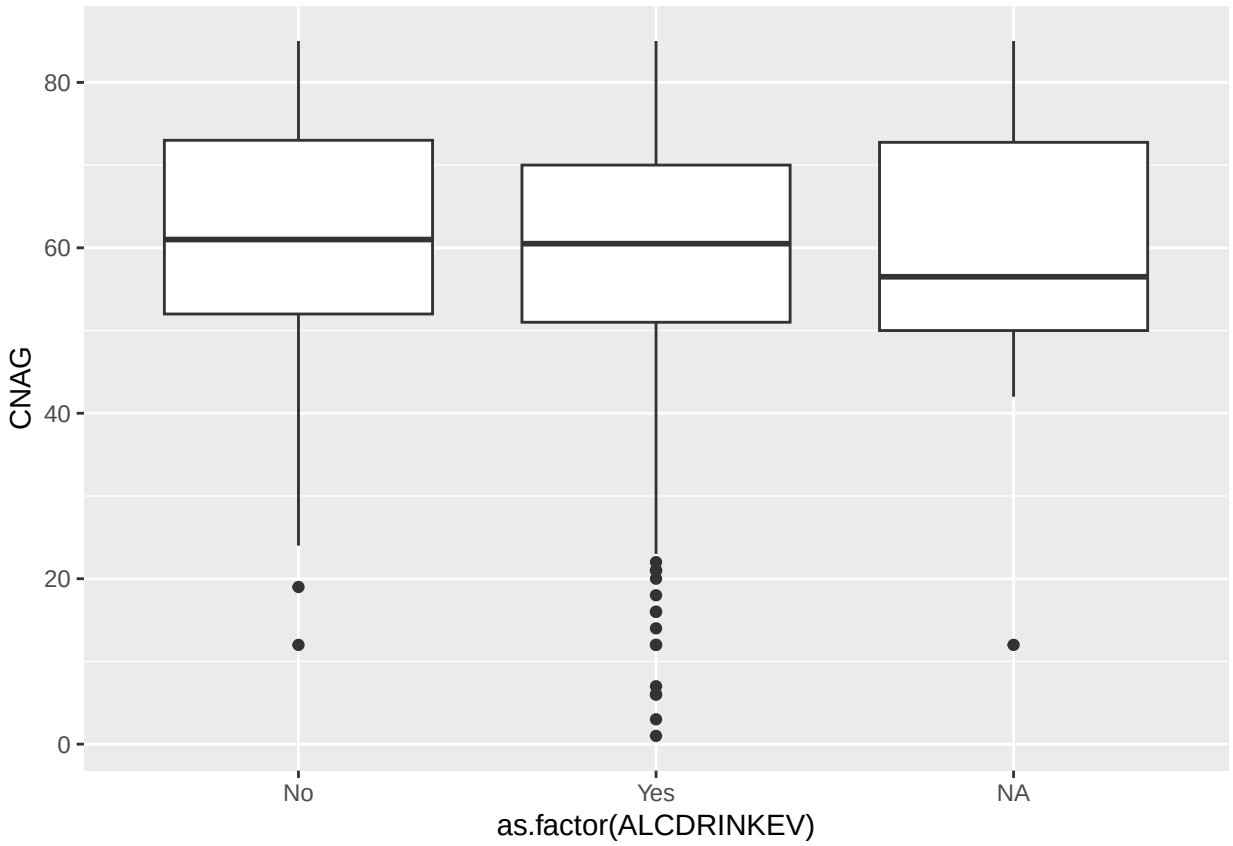
```
## Warning: Removed 54 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



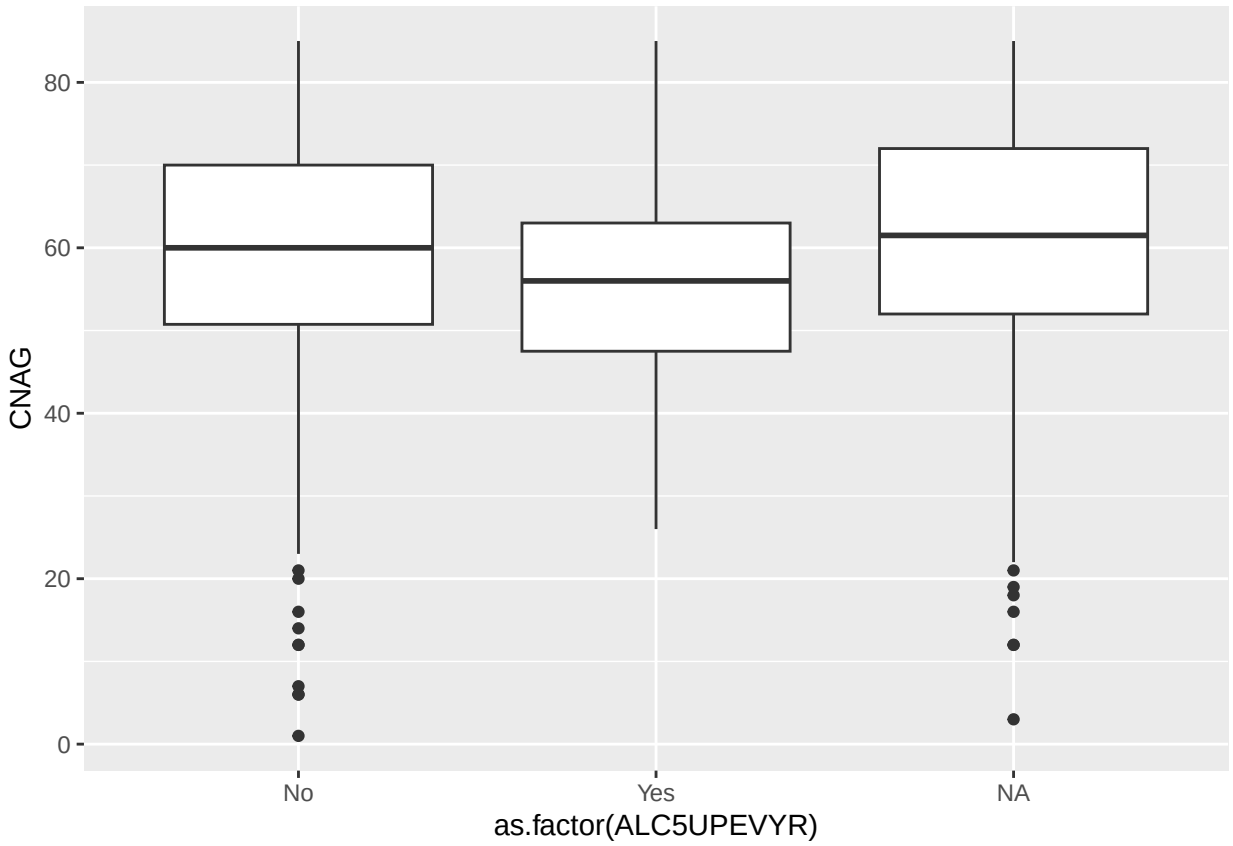

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(DIABTYPE), y=CNAG)) + ggplot2::geom_boxplo
```



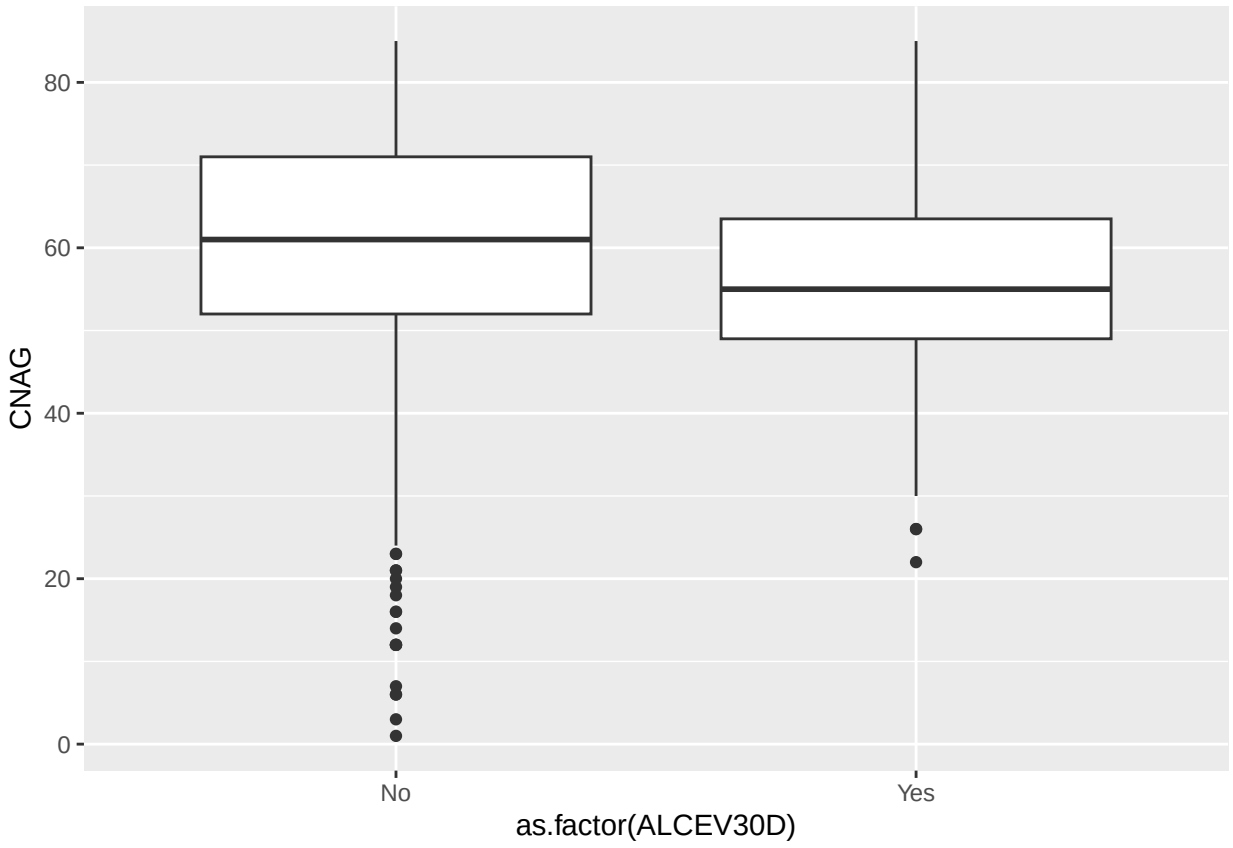

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(ALCDRINKEV), y=CNAG)) + ggplot2::geom_boxpl
```



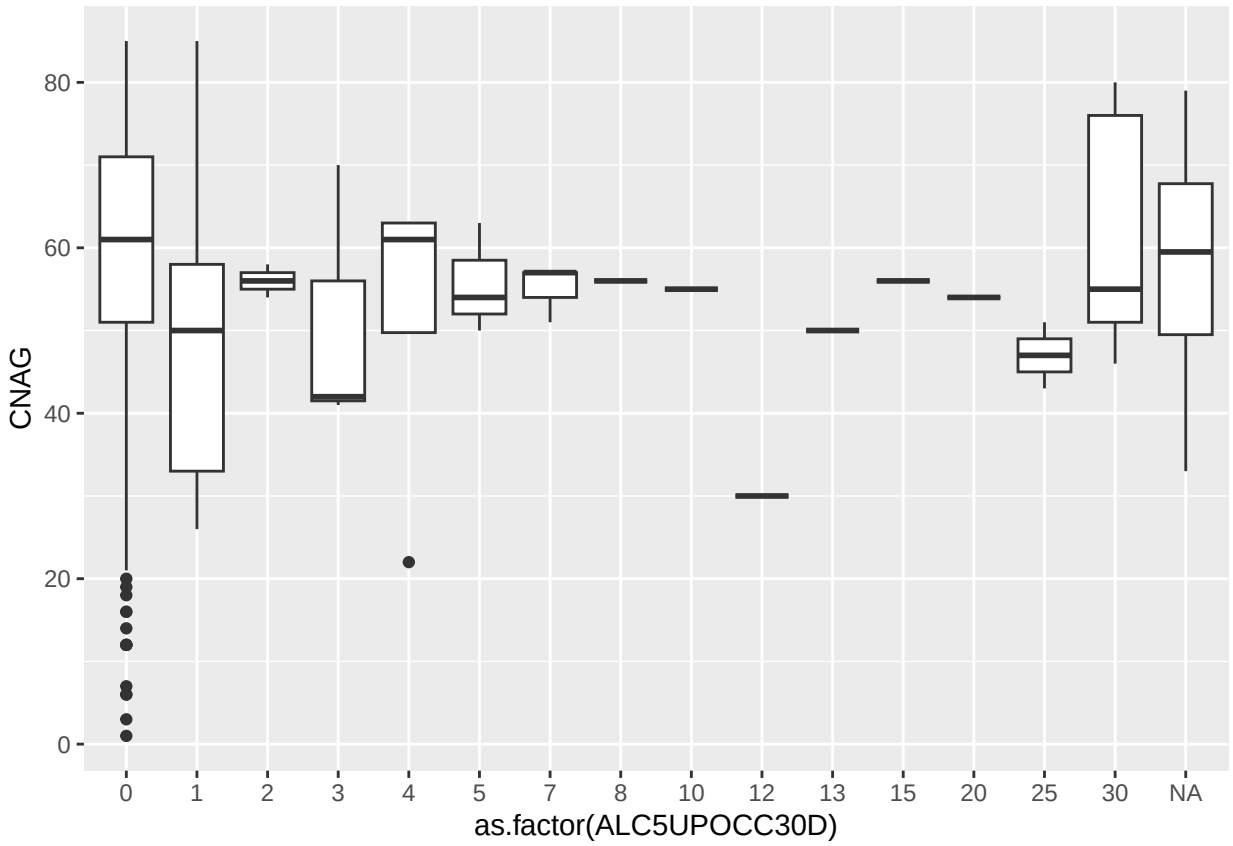

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(ALC5UPEVYR), y=CNAG)) + ggplot2::geom_boxpl
```



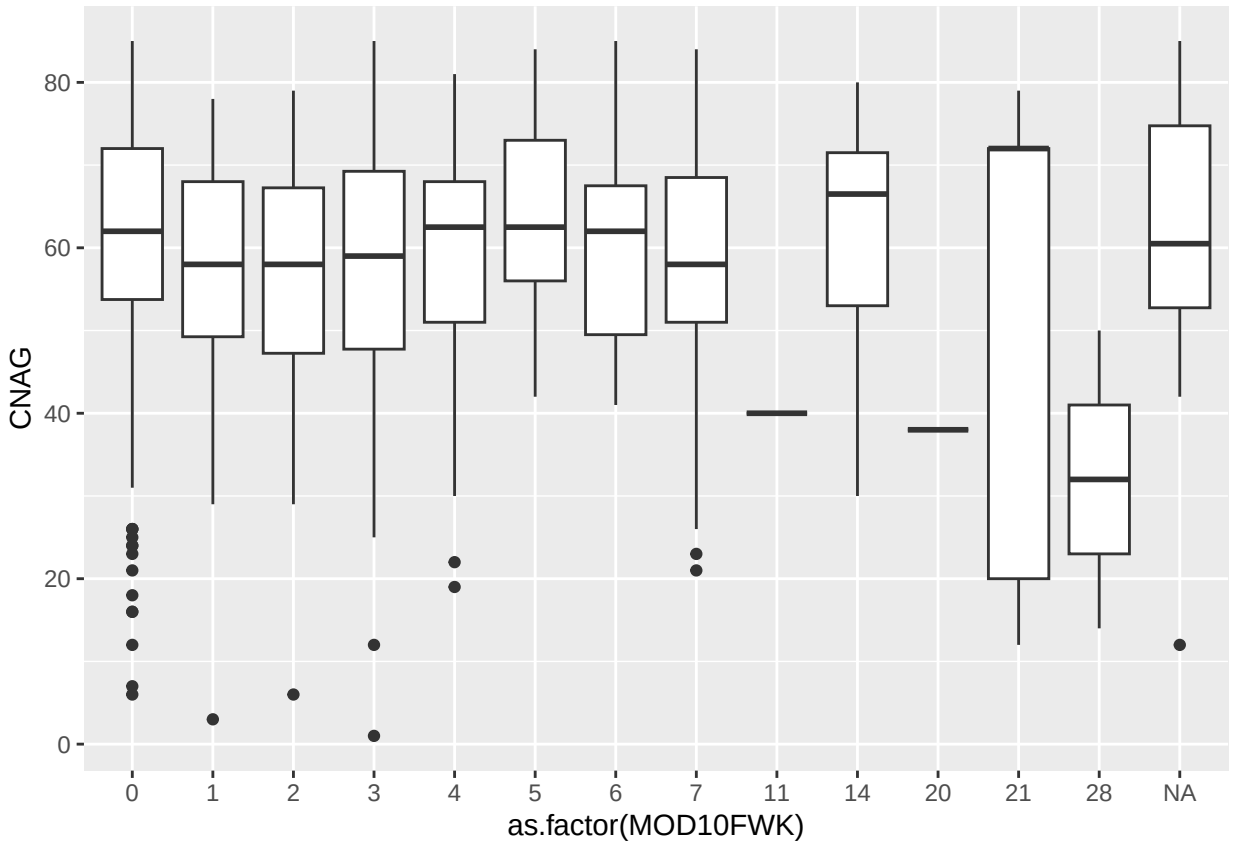

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(ALCEV30D), y=CNAG)) + ggplot2::geom_boxplo
```



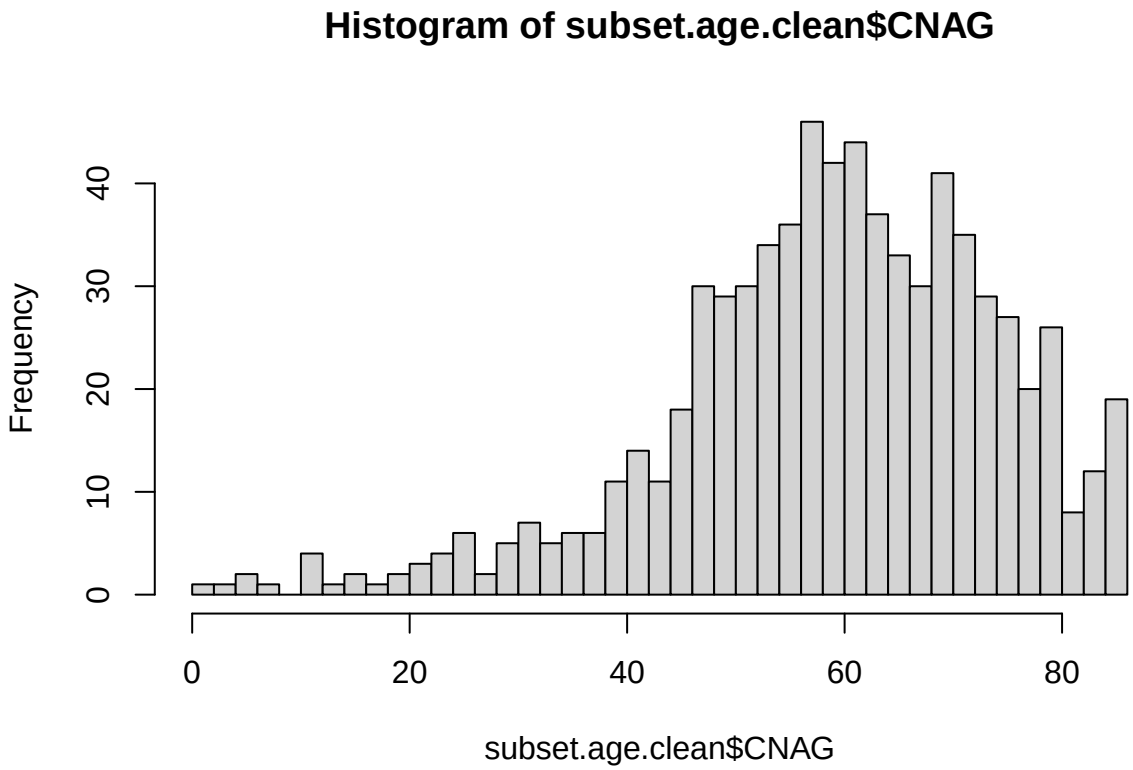

```
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(ALC5UPOCC30D), y=CNAG)) + ggplot2::geom_bo
```




```
# ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(MOD1OFNO), y=CNAG)) + ggplot2::geom_boxp
ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(MOD1OFWK), y=CNAG)) + ggplot2::geom_boxplo
```




```
hist(subset.age.clean$CNAG, breaks=43)
```



```
head(data$MOD1OFWK   )
```

```
## <labelled<double>[6]>: Frequency of moderate activity 10+ minutes: Times per week
## [1] 95  7  2 95 98  3
##
## Labels:
##  value                      label
##      0          Not in Universe
##     93            Extreme value
##     94  Less than once per week
##     95                      Never
##     96 Unable to do this activity
##     97          Unknown-refused
##     98  Unknown-not ascertained
##     99       Unknown-don't know
```

```
summary(data$MOD1OFWK)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    5.00   29.08   95.00   99.00
```

Models

```
library(MASS)
```

Saturated model

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

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

```
subset.age.clean.nb <- subset.age.clean %>%
  dplyr::select(SEX, MAXEDUC, EMPSTAT, QTCINCFAM, HEALTH, HEIGHT,
                WEIGHT, BMICALC, DIABTYPE, ALCEV30D, MOD1OFWK, CNAG) %>%
  na.omit()

fit0.poisson = glm(CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM
           + HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE
           + ALCEV30D + MOD1OFWK,
           data=subset.age.clean.nb, family=poisson(link="log"),
           na.action = na.omit)
anova(fit0.poisson)
```

```
## Analysis of Deviance Table
##
```

```
## Model: poisson, link: log
##
## Response: CNAG
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                        624     2494.7
## SEX        1    6.867       623     2487.9
## MAXEDUC    1    0.356       622     2487.5
## EMPSTAT    1  277.570       621     2209.9
## QTCINCFAM  1    0.668       620     2209.3
## HEALTH     1    1.834       619     2207.4
## HEIGHT     1   44.556       618     2162.9
## WEIGHT     1   25.471       617     2137.4
## BMICALC    1    3.982       616     2133.4
## DIABTYPE   3   17.158       613     2116.3
## ALCEV30D   1    6.745       612     2109.5
## MOD10FWK   1   29.512       611     2080.0
```

```
summary(fit0.poisson)
```

```
##
## Call:
## glm(formula = CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH +
##     HEIGHT + WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK,
##     family = poisson(link = "log"), data = subset.age.clean.nb,
##     na.action = na.omit)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     5.533581   0.493262  11.218  < 2e-16 ***
## SEXFemale                      -0.114292   0.015462  -7.392 1.45e-13 ***
## MAXEDUC                         0.005426   0.002068   2.624 0.008686 **
## EMPSTATNot employed             0.184088   0.013477  13.659  < 2e-16 ***
## QTCINCFAMYes                   -0.033689   0.040949  -0.823 0.410678
## HEALTH                         -0.001721   0.004794  -0.359 0.719643
## HEIGHT                         -0.020414   0.007381  -2.766 0.005679 **
## WEIGHT                          0.001418   0.001364   1.040 0.298477
## BMICALC                        -0.015525   0.008645  -1.796 0.072501 .
## DIABTYPEType 1                  0.016241   0.049368   0.329 0.742179
## DIABTYPEType 2                  0.053634   0.013914   3.855 0.000116 ***
## DIABTYPEOther type of diabetes  0.101605   0.092847   1.094 0.273811
## ALCEV30DYes                    -0.049074   0.017516  -2.802 0.005083 **
## MOD10FWK                       -0.007758   0.001444  -5.372 7.77e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2494.7  on 624  degrees of freedom
## Residual deviance: 2080.0  on 611  degrees of freedom
## AIC: 5786.8
```

```
##
## Number of Fisher Scoring iterations: 4
```

Model selection Simplify: Backward stepwise

```
fit0.poisson.backward <- step(fit0.poisson, direction = "backward")
```

```
## Start:  AIC=5786.76
## CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##
##             Df Deviance    AIC
## - HEALTH     1   2080.1 5784.9
## - QTCINCFAM  1   2080.7 5785.4
## - WEIGHT     1   2081.1 5785.8
## <none>           2080.0 5786.8
## - BMICALC    1   2083.2 5788.0
## - MAXEDUC    1   2086.9 5791.6
## - HEIGHT     1   2087.7 5792.4
## - ALCEV30D   1   2087.9 5792.7
## - DIABTYPE   3   2095.4 5796.2
## - MOD10FWK   1   2109.5 5814.3
## - SEX        1   2134.6 5839.3
## - EMPSTAT    1   2272.2 5976.9
##
## Step:  AIC=5784.89
## CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEIGHT + WEIGHT +
##     BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##
##             Df Deviance    AIC
## - QTCINCFAM  1   2080.8 5783.5
## - WEIGHT     1   2081.2 5784.0
## <none>           2080.1 5784.9
## - BMICALC    1   2083.4 5786.2
## - MAXEDUC    1   2087.4 5790.2
## - HEIGHT     1   2087.9 5790.6
## - ALCEV30D   1   2088.0 5790.8
## - DIABTYPE   3   2095.6 5794.3
## - MOD10FWK   1   2109.9 5812.7
## - SEX        1   2134.6 5837.4
## - EMPSTAT    1   2275.7 5978.4
##
## Step:  AIC=5783.54
## CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + WEIGHT + BMICALC +
##     DIABTYPE + ALCEV30D + MOD10FWK
##
##            Df Deviance    AIC
## - WEIGHT    1   2081.8 5782.6
## <none>          2080.8 5783.5
## - BMICALC   1   2084.0 5784.7
## - MAXEDUC   1   2087.7 5788.4
## - HEIGHT    1   2088.4 5789.1
## - ALCEV30D  1   2089.2 5790.0
```

```
## - DIABTYPE  3   2096.1 5792.9
## - MOD1OFWK  1   2110.7 5811.5
## - SEX       1   2135.1 5835.8
## - EMPSTAT   1   2277.0 5977.8
##
## Step:  AIC=5782.6
## CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + DIABTYPE +
##     ALCEV3OD + MOD1OFWK
##
##            Df Deviance    AIC
## <none>          2081.8 5782.6
## - MAXEDUC   1   2088.8 5787.5
## - ALCEV3OD  1   2090.5 5789.2
## - DIABTYPE  3   2097.0 5791.7
## - MOD1OFWK  1   2112.5 5811.2
## - BMICALC   1   2122.6 5821.4
## - HEIGHT    1   2124.4 5823.2
## - SEX       1   2136.4 5835.1
## - EMPSTAT   1   2278.0 5976.7
```

```
summary(fit0.poisson.backward)
```

```
##
## Call:
## glm(formula = CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC +
##     DIABTYPE + ALCEV3OD + MOD1OFWK, family = poisson(link = "log"),
##     data = subset.age.clean.nb, na.action = na.omit)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    5.034128   0.145634  34.567  < 2e-16 ***
## SEXFemale                     -0.114105   0.015447  -7.387 1.50e-13 ***
## MAXEDUC                        0.005345   0.002033   2.629 0.008561 **
## EMPSTATNot employed            0.183567   0.013315  13.787  < 2e-16 ***
## HEIGHT                        -0.012977   0.001989  -6.525 6.81e-11 ***
## BMICALC                       -0.006569   0.001032  -6.366 1.94e-10 ***
## DIABTYPEType 1                 0.016091   0.049349   0.326 0.744370
## DIABTYPEType 2                 0.052250   0.013700   3.814 0.000137 ***
## DIABTYPEOther type of diabetes 0.098822   0.092272   1.071 0.284179
## ALCEV3ODYes                   -0.050852   0.017398  -2.923 0.003468 **
## MOD1OFWK                      -0.007765   0.001419  -5.474 4.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2494.7  on 624  degrees of freedom
## Residual deviance: 2081.8  on 614  degrees of freedom
## AIC: 5782.6
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(fit0.poisson, fit0.poisson.backward)
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + DIABTYPE +
##     ALCEV30D + MOD10FWK
##   Resid. Df Resid. Dev Df Deviance
## 1       611     2080.0
## 2       614     2081.8 -3  -1.8369
```

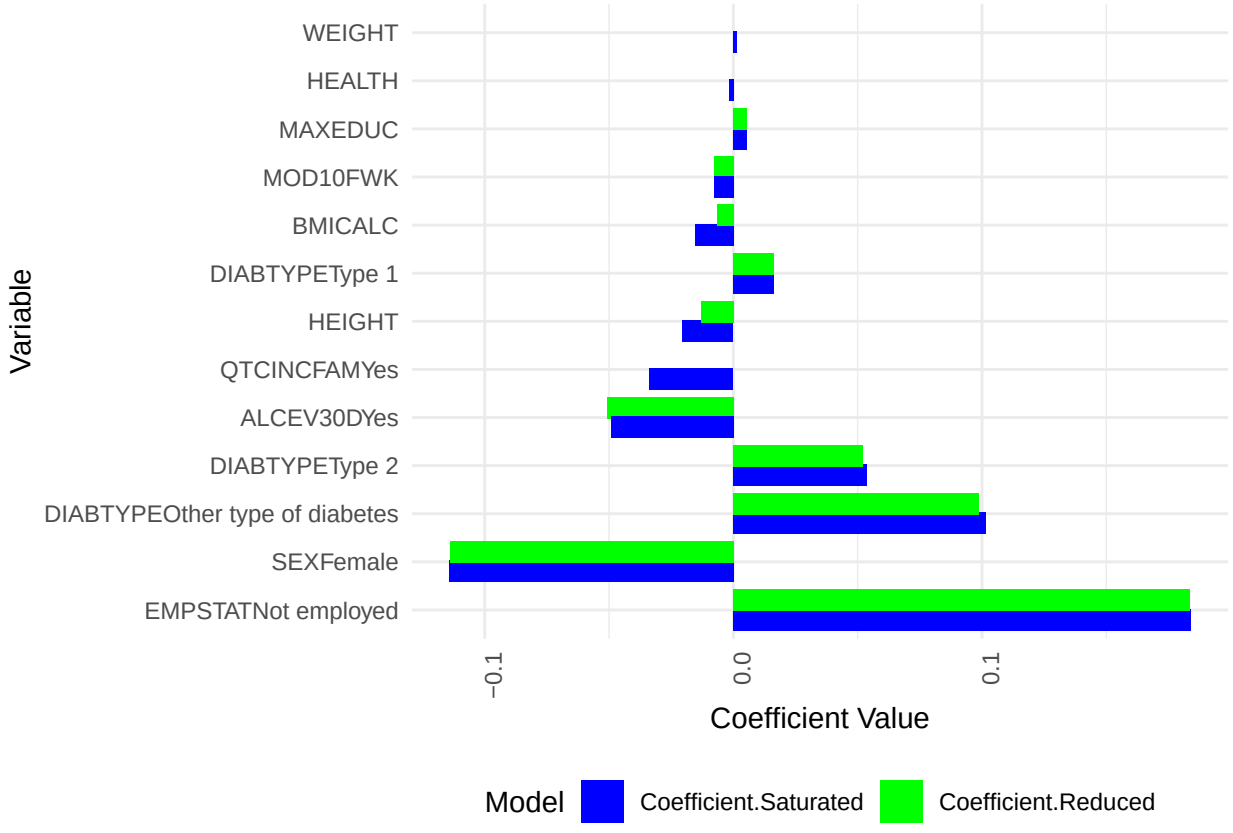
```r
anova(fit0.poisson, fit0.poisson.backward, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + DIABTYPE +
##     ALCEV30D + MOD10FWK
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       611     2080.0
## 2       614     2081.8 -3  -1.8369   0.6069
```

```r
# Load the necessary libraries
library(ggplot2)
library(reshape2)

# Extract coefficients from both models
coefficients_saturated <- coef(fit0.poisson)
coefficients_reduced <- coef(fit0.poisson.backward)

# Remove the intercepts
coefficients_saturated <- coefficients_saturated[-1]
coefficients_reduced <- coefficients_reduced[-1]

# Create data frames for plotting
coeff_df_saturated <- data.frame(
  Variable = names(coefficients_saturated),
  Coefficient = coefficients_saturated,
  Model = 'Saturated'
)

coeff_df_reduced <- data.frame(
  Variable = names(coefficients_reduced),
  Coefficient = coefficients_reduced,
  Model = 'Reduced'
)

# Combine the two data frames, allowing NAs for the variables not in the reduced model
coeff_df_combined <- merge(coeff_df_saturated, coeff_df_reduced, by = "Variable", all = TRUE, suffixes =
```

```r
# Create a long format data frame for plotting with ggplot2
coeff_df_long <- melt(coeff_df_combined, id.vars = "Variable",
                      measure.vars = c("Coefficient.Saturated", "Coefficient.Reduced"),
                      variable.name = "Model", value.name = "Coefficient")

# Sort the data frame by the value of the coefficients from the saturated model
coeff_df_long <- coeff_df_long[order(-abs(coeff_df_long$Coefficient)),]

# Reorder the variables based on the sorted coefficients for plotting
coeff_df_long$Variable <- factor(coeff_df_long$Variable, levels = unique(coeff_df_long$Variable))

# Plot using ggplot2
ggplot(coeff_df_long, aes(x = Variable, y = Coefficient, fill = Model)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
  coord_flip() +  # Flipping coordinates to make it easier to read variable names
  scale_fill_manual(values = c("Coefficient.Saturated" = "blue", "Coefficient.Reduced" = "green", "NA" =
  labs(x = "Variable", y = "Coefficient Value", fill = "Model") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")  # Adjust the an
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



Findings:

```
# library(lme4)
#
# fit0.mixed <- glmer(CNAG ~ SEX + MAXEDUC + SPOUSEDUC + EMPSTAT + EMPHI +
#   QTCINCFAM + HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE +
#      ALCEV30D +  MOD10FWK +
#     (1 | EMPSTAT),
#   data = subset.age.clean, family = poisson(link = "log"),
#   na.action = na.omit)
# summary(fit0.mixed)
# anova(fit0.mixed)
#
# fit0.mixed.simplified <- glmer(CNAG ~ SEX+ BMICALC + DIABTYPE +
#     MOD10FWK   + (1 | EMPSTAT),
#   data = subset.age.clean, family = poisson(link = "log"),
#   na.action = na.omit)
# summary(fit0.mixed.simplified)
# anova(fit0.mixed, fit0.mixed.simplified)
```

Mixed model with clusters regarding of EMPSTAT (employed vs. unemployed)

```
fit1.AL = glm(CNAG ~ ALCEV30D,
           data=subset.age.clean.nb, family=poisson(link="log"),
           na.action = na.omit)
anova(fit1.AL)
```

Drinking behavior

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: CNAG
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                       624     2494.7
## ALCEV30D  1   23.716       623     2471.0
```

```
summary(fit1.AL)
```

```
##
## Call:
## glm(formula = CNAG ~ ALCEV30D, family = poisson(link = "log"),
##     data = subset.age.clean.nb, na.action = na.omit)
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.097427   0.005467 749.538  < 2e-16 ***
## ALCEV30DYes -0.082220   0.017068  -4.817 1.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2494.7  on 624  degrees of freedom
## Residual deviance: 2471.0  on 623  degrees of freedom
## AIC: 6153.8
##
## Number of Fisher Scoring iterations: 4
```

```
anova(fit0.poisson, fit1.AL, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ ALCEV30D
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1       611       2080
## 2       623       2471 -12     -391 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit1.MOD = glm(CNAG ~ MOD10FWK,
           data=subset.age.clean.nb, family=poisson(link="log"),
           na.action = na.omit)
anova(fit1.MOD)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: CNAG
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                       624     2494.7
## MOD10FWK  1   41.356       623     2453.4
```

```
summary(fit1.MOD)
```

```
##
## Call:
## glm(formula = CNAG ~ MOD10FWK, family = poisson(link = "log"),
##     data = subset.age.clean.nb, na.action = na.omit)
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.112813   0.006373 645.333  < 2e-16 ***
## MOD10FWK    -0.008926   0.001406  -6.349 2.17e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2494.7  on 624  degrees of freedom
## Residual deviance: 2453.4  on 623  degrees of freedom
## AIC: 6136.1
##
## Number of Fisher Scoring iterations: 4
```

```
anova(fit0.poisson, fit1.MOD, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ MOD10FWK
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1       611     2080.0
## 2       623     2453.4 -12  -373.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Negative binomial GLM

The first time to success after Y years is a negative binomial problem.

```
library(dplyr)
library(MASS)


colSums(is.na(subset.age.clean.nb))
```

```
##       SEX   MAXEDUC   EMPSTAT QTCINCFAM    HEALTH    HEIGHT    WEIGHT   BMICALC
##         0         0         0         0         0         0         0         0
##  DIABTYPE  ALCEV30D  MOD10FWK      CNAG
##         0         0         0         0
```

```
fit.NB = glm.nb(CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
  WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK,
  data = subset.age.clean.nb)

summary(fit.NB)
```

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

```
## glm.nb(formula = CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM +
##     HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE + ALCEV30D +
##     MOD10FWK, data = subset.age.clean.nb, init.theta = 23.81494453,
##     link = log)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    5.6076946  0.9250989   6.062 1.35e-09 ***
## SEXFemale                     -0.1184204  0.0290172  -4.081 4.48e-05 ***
## MAXEDUC                        0.0055987  0.0038837   1.442   0.1494
## EMPSTATNot employed            0.1865337  0.0243172   7.671 1.71e-14 ***
## QTCINCFAMYes                  -0.0257516  0.0755195  -0.341   0.7331
## HEALTH                        -0.0018121  0.0089851  -0.202   0.8402
## HEIGHT                        -0.0215670  0.0138403  -1.558   0.1192
## WEIGHT                         0.0015458  0.0025457   0.607   0.5437
## BMICALC                       -0.0162195  0.0161480  -1.004   0.3152
## DIABTYPEType 1                 0.0009637  0.0928450   0.010   0.9917
## DIABTYPEType 2                 0.0555003  0.0264685   2.097   0.0360 *
## DIABTYPEOther type of diabetes 0.1017760  0.1746070   0.583   0.5600
## ALCEV30DYes                   -0.0498623  0.0321475  -1.551   0.1209
## MOD10FWK                      -0.0082398  0.0026392  -3.122   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(23.8149) family taken to be 1)
##
##     Null deviance: 807.39  on 624  degrees of freedom
## Residual deviance: 684.61  on 611  degrees of freedom
## AIC: 5169.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  23.81
##           Std. Err.:  2.01
##
##  2 x log-likelihood:  -5139.215
```

```
anova(fit.NB)
```

```
## Warning in anova.negbin(fit.NB): tests made without re-estimating 'theta'

## Analysis of Deviance Table
##
## Model: Negative Binomial(23.8149), link: log
##
## Response: CNAG
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        624     807.39
```

```
## SEX        1    1.960       623     805.43 0.1615623
## MAXEDUC    1    0.098       622     805.33 0.7540320
## EMPSTAT    1   81.548       621     723.79 < 2.2e-16 ***
## QTCINCFAM  1    0.106       620     723.68 0.7452829
## HEALTH     1    0.598       619     723.08 0.4391768
## HEIGHT     1   13.779       618     709.30 0.0002056 ***
## WEIGHT     1    6.733       617     702.57 0.0094643 **
## BMICALC    1    1.166       616     701.40 0.2802565
## DIABTYPE   3    5.266       613     696.14 0.1533458
## ALCEV30D   1    2.104       612     694.03 0.1468845
## MOD10FWK   1    9.427       611     684.61 0.0021383 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.NB.reduced = stepAIC(fit.NB, direction="backward")
```

```
## Start:  AIC=5167.21
## CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##
##             Df    AIC
## - HEALTH     1 5165.3
## - QTCINCFAM  1 5165.3
## - WEIGHT     1 5165.6
## - DIABTYPE   3 5165.8
## - BMICALC    1 5166.2
## <none>         5167.2
## - MAXEDUC    1 5167.3
## - ALCEV30D   1 5167.6
## - HEIGHT     1 5167.6
## - MOD10FWK   1 5174.6
## - SEX        1 5181.5
## - EMPSTAT    1 5220.9
##
## Step:  AIC=5165.25
## CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEIGHT + WEIGHT +
##     BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##
##             Df    AIC
## - QTCINCFAM  1 5163.4
## - WEIGHT     1 5163.6
## - DIABTYPE   3 5163.8
## - BMICALC    1 5164.3
## <none>         5165.3
## - MAXEDUC    1 5165.5
## - ALCEV30D   1 5165.6
## - HEIGHT     1 5165.7
## - MOD10FWK   1 5172.7
## - SEX        1 5179.5
## - EMPSTAT    1 5219.7
##
## Step:  AIC=5163.36
## CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + WEIGHT + BMICALC +
##     DIABTYPE + ALCEV30D + MOD10FWK
```

```
##
##            Df    AIC
## - WEIGHT    1 5161.7
## - DIABTYPE  3 5161.9
## - BMICALC   1 5162.4
## <none>        5163.4
## - MAXEDUC   1 5163.5
## - HEIGHT    1 5163.8
## - ALCEV30D  1 5163.9
## - MOD10FWK  1 5170.8
## - SEX       1 5177.5
## - EMPSTAT   1 5218.1
##
## Step:  AIC=5161.73
## CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + DIABTYPE +
##     ALCEV30D + MOD10FWK
##
##            Df    AIC
## - DIABTYPE  3 5160.2
## <none>        5161.7
## - MAXEDUC   1 5161.8
## - ALCEV30D  1 5162.3
## - MOD10FWK  1 5169.4
## - BMICALC   1 5170.7
## - HEIGHT    1 5172.5
## - SEX       1 5175.9
## - EMPSTAT   1 5216.3
##
## Step:  AIC=5160.2
## CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + ALCEV30D +
##     MOD10FWK
##
##            Df    AIC
## <none>        5160.2
## - MAXEDUC   1 5160.6
## - ALCEV30D  1 5161.1
## - BMICALC   1 5166.6
## - MOD10FWK  1 5168.6
## - HEIGHT    1 5170.9
## - SEX       1 5175.0
## - EMPSTAT   1 5219.9
```

```
summary(fit.NB.reduced)
```

```
##
## Call:
## glm.nb(formula = CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC +
##     ALCEV30D + MOD10FWK, data = subset.age.clean.nb, init.theta = 23.51947934,
##     link = log)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.038414   0.274241  18.372  < 2e-16 ***
## SEXFemale           -0.120446   0.029025  -4.150 3.33e-05 ***
```

```
## MAXEDUC              0.005919   0.003831   1.545 0.122342
## EMPSTATNot employed  0.193127   0.023798   8.115 4.85e-16 ***
## HEIGHT              -0.013424   0.003745  -3.584 0.000338 ***
## BMICALC             -0.005470   0.001872  -2.922 0.003483 **
## ALCEV30DYes         -0.054695   0.032008  -1.709 0.087494 .
## MOD10FWK            -0.008562   0.002603  -3.290 0.001003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(23.5195) family taken to be 1)
##
##     Null deviance: 800.79  on 624  degrees of freedom
## Residual deviance: 684.05  on 617  degrees of freedom
## AIC: 5162.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  23.52
##           Std. Err.:  1.98
##
##  2 x log-likelihood:  -5144.205
```

```
anova(fit.NB, fit.NB.reduced, test="Chisq")
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: CNAG
##
## 1                                          SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + ALCEV30D + M
## 2 SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE + ALCEV30D + M
##      theta Resid. df    2 x log-lik.   Test    df LR stat.  Pr(Chi)
## 1 23.51948       617       -5144.205
## 2 23.81494       611       -5139.215 1 vs 2     6 4.990151 0.545077
```

```
anova(fit0.poisson, fit.NB, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       611    2080.01
## 2       611     684.61  0   1395.4
```

```
anova(fit0.poisson.backward, fit.NB.reduced, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + DIABTYPE +
```

```
##     ALCEV30D + MOD10FWK
## Model 2: CNAG ~ SEX + MAXEDUC + EMPSTAT + HEIGHT + BMICALC + ALCEV30D +
##     MOD10FWK
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       614    2081.85
## 2       617     684.05 -3   1397.8
```

```
# Perform the likelihood ratio test
lrt_result <- anova(fit0.poisson, fit.NB, test = "LRT")

# Print the results of the likelihood ratio test
print(lrt_result)
```

```
## Analysis of Deviance Table
##
## Model 1: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
## Model 2: CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM + HEALTH + HEIGHT +
##     WEIGHT + BMICALC + DIABTYPE + ALCEV30D + MOD10FWK
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       611    2080.01
## 2       611     684.61  0   1395.4
```

Topic 2

Response: binary CNLIVR

```
head(data$CNLIVR)
```

```
## <labelled<integer>[6]>: Ever had cancer: Liver
## [1] 0 0 1 0 0 0
##
## Labels:
##  value                   label
##      0                     NIU
##      1           Not mentioned
##      2               Mentioned
##      7         Unknown-refused
##      8 Unknown-not ascertained
##      9      Unknown-don't know
```

```
subset.CNLIVR.clean =  data.clean %>%
  # mutate(across(c("CNLIVR"), ~if_else(. >= 7, NA_real_, .))) %>%
  filter(CNLIVR==1 | CNLIVR==2) %>%
  mutate(CNLIVR=CNLIVR-1) %>% # Change label: Positive 2->1, negative 1->0
  mutate(across(c(SEX, EMPSTAT, EMPHI, QTCINCFAM,
                  ALCDRINKEV, ALC5UPEVYR, ALCEV30D, # Binary variables
                  DIABTYPE), # non-numerical factor
                ~as_factor(as_factor(.))))
fit.cnliver = glm(CNLIVR ~ AGE + SEX + MAXEDUC + SPOUSEDUC +
                    EMPSTAT + EMPHI + QTCINCFAM +
                    HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE +
```

```
                    ALCDRINKEV +  ALCEV30D + ALC5UPOCC30D + MOD10FWK,
                  data=subset.CNLIVR.clean, family=binomial(link = "logit"),
                  na.action = na.omit)
```

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

```
anova(fit.cnliver)
```

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

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

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

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

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

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

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

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

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

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: CNLIVR
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                         1143     53.234
## AGE           1   1.2396      1142     51.994
## SEX           1   0.0488      1141     51.946
## MAXEDUC       1   5.7187      1140     46.227
## SPOUSEDUC     1   0.0157      1139     46.211
## EMPSTAT       1   0.5412      1138     45.670
## EMPHI         1   0.0241      1137     45.646
## QTCINCFAM     1   0.3094      1136     45.336
## HEALTH        1   5.8618      1135     39.475
## HEIGHT        1   0.0899      1134     39.385
## WEIGHT        1   0.8460      1133     38.539
## BMICALC       1   0.1320      1132     38.407
## DIABTYPE      3   8.0608      1129     30.346
## ALCDRINKEV    1   0.4505      1128     29.895
## ALCEV30D      1   0.0001      1127     29.895
## ALC5UPOCC30D  1   0.0065      1126     29.889
## MOD10FWK      1   1.4840      1125     28.405
```

```r
summary(fit.cnliver)
```

```
##
## Call:
## glm(formula = CNLIVR ~ AGE + SEX + MAXEDUC + SPOUSEDUC + EMPSTAT +
##     EMPHI + QTCINCFAM + HEALTH + HEIGHT + WEIGHT + BMICALC +
##     DIABTYPE + ALCDRINKEV + ALCEV30D + ALC5UPOCC30D + MOD10FWK,
##     family = binomial(link = "logit"), data = subset.CNLIVR.clean,
##     na.action = na.omit)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.405e+01  7.467e+03  -0.005  0.99636
## AGE                            -2.811e-03  5.640e-02  -0.050  0.96025
## SEXFemale                      -1.498e-01  2.043e+00  -0.073  0.94154
## MAXEDUC                        -6.529e-01  5.695e-01  -1.146  0.25163
## SPOUSEDUC                       1.330e-01  5.680e-01   0.234  0.81491
## EMPSTATNot employed            -1.734e+01  7.747e+03  -0.002  0.99821
## EMPHIYes                        2.454e-01  1.498e+00   0.164  0.86987
## QTCINCFAMYes                   -1.469e+01  5.909e+03  -0.002  0.99802
## HEALTH                          1.949e+00  8.175e-01   2.385  0.01710 *
## HEIGHT                          5.327e-02  1.085e+00   0.049  0.96083
## WEIGHT                         -3.628e-03  1.558e-01  -0.023  0.98143
## BMICALC                         1.483e-01  9.872e-01   0.150  0.88062
## DIABTYPEType 1                  5.477e+00  2.043e+00   2.681  0.00734 **
## DIABTYPEType 2                 -1.838e+01  6.249e+03  -0.003  0.99765
## DIABTYPEOther type of diabetes -2.124e+01  3.899e+04  -0.001  0.99957
## ALCDRINKEVYes                   1.745e+01  7.466e+03   0.002  0.99814
## ALCEV30DYes                    -2.030e-01  1.604e+00  -0.127  0.89929
## ALC5UPOCC30D                    1.915e-02  1.306e-01   0.147  0.88343
## MOD10FWK                        1.723e-01  1.216e-01   1.417  0.15641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.234  on 1143  degrees of freedom
## Residual deviance: 28.405  on 1125  degrees of freedom
##   (6373 observations deleted due to missingness)
## AIC: 66.405
##
## Number of Fisher Scoring iterations: 22
```

```r
########################################
## Reference Only, Not for final version
########################################

#install.packages("glmnet")

library(glmnet)
```

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

```
## Loaded glmnet 4.1-8
```

```r
x <- model.matrix(~ AGE + SEX + MAXEDUC + SPOUSEDUC +
                    EMPSTAT + EMPHI + QTCINCFAM +
                    HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE +
                    ALCDRINKEV + ALCEV30D + ALC5UPOCC30D + MOD10FNO - 1,
                  data=subset.CNLIVR.clean)

yx <- model.matrix(~CNLIVR+ AGE + SEX + MAXEDUC + SPOUSEDUC +
                     EMPSTAT + EMPHI + QTCINCFAM +
                     HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE +
                     ALCDRINKEV + ALCEV30D + ALC5UPOCC30D + MOD10FNO - 1,
                   data=subset.CNLIVR.clean)
y <- yx[,1]

# Lasso regression
fit.cnliver.glmnet <- glmnet(x, y, family="binomial", alpha=1)
```

```
## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground
```

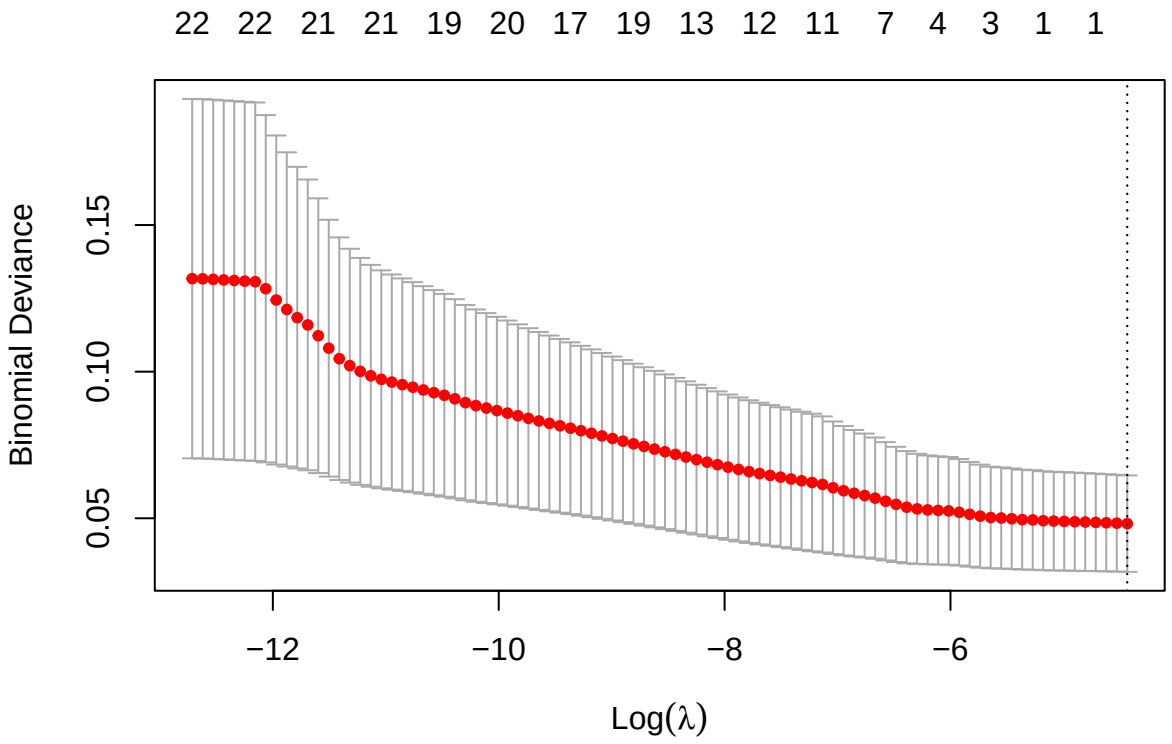
```r
# You can then use cv.glmnet to find the optimal lambda value (regularization strength)
fit.optimal <- cv.glmnet(x, y, family="binomial", alpha=1)
```

```
## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground
```

```
## Warning: from glmnet C++ code (error code -88); Convergence for 88th lambda
## value not reached after maxit=100000 iterations; solutions for larger lambdas
## returned
```

```
## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground
```

```
# Plot the cross-validation curve
plot(fit.optimal)
```



Correlation check

```
head(data$MOD10FWK)
```

```
## <labelled<double>[6]>: Frequency of moderate activity 10+ minutes: Times per week
## [1] 95  7  2 95 98  3
##
## Labels:
##  value                       label
##      0            Not in Universe
##     93              Extreme value
##     94    Less than once per week
##     95                      Never
##     96 Unable to do this activity
##     97            Unknown-refused
##     98    Unknown-not ascertained
##     99         Unknown-don't know
```

```r
table(data$MOD10FWK)
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
## 13254  4475  6601  7705  3703  3748   953 10456    10     8    23     8    18
##    13    14    15    16    17    18    20    21    22    23    24    28    93
##    10   765     6     2     3     1     5   280     5     5     3   108   259
##    94    95    96    97    98    99
##  1920 16146   383    48  1398   164
```

```r
# library(corrplot)
# corr_matrix <- cor(subset.age, use=)
# corrplot(corr_matrix, method = "color", addCoef.col = "black",
#          tl.col="black", tl.srt=45, cl.pos='b', type="upper")
```

```r
#
# hist(subset.age.clean$ALC5UPEVYR)
# ggplot2::ggplot(subset.age.clean, ggplot2::aes(x = as.factor(ALC5UPEVYR), y=CNAG)) + ggplot2::geom_bo
```