GLM project

Impact of lifestyle on cancer

Christine Xing

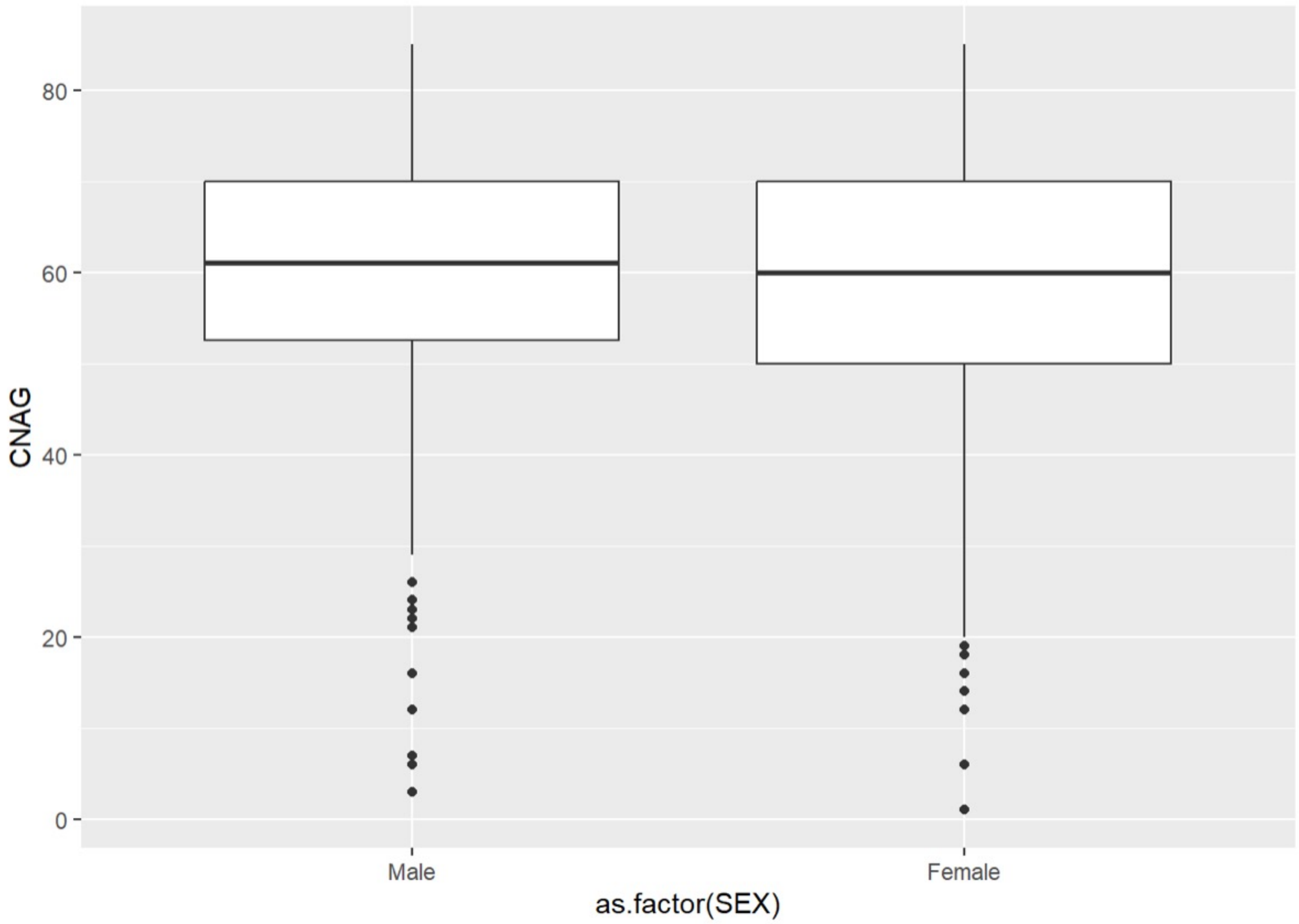
Shahzab Hussain

Cong Lyu

Project Outline

- Diving into when cancer first strikes
- We mixed biology with lifestyle in our stats blender to predict cancer's "first hello"
- Our GLM's recipe: age, gender, body metrics, education, jobs, family history, and more
- The result? A revealing picture of how factors like gender and diabetes type stir the pot

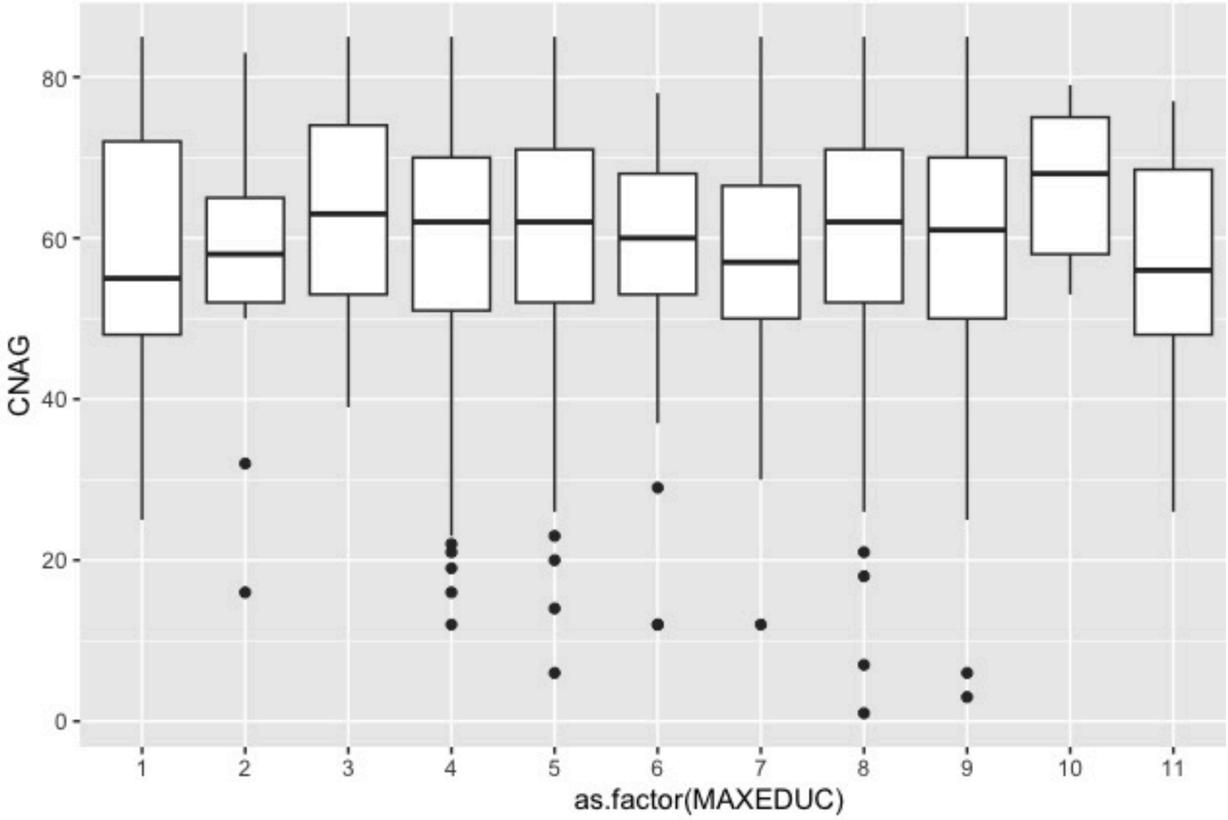
EDA on gender



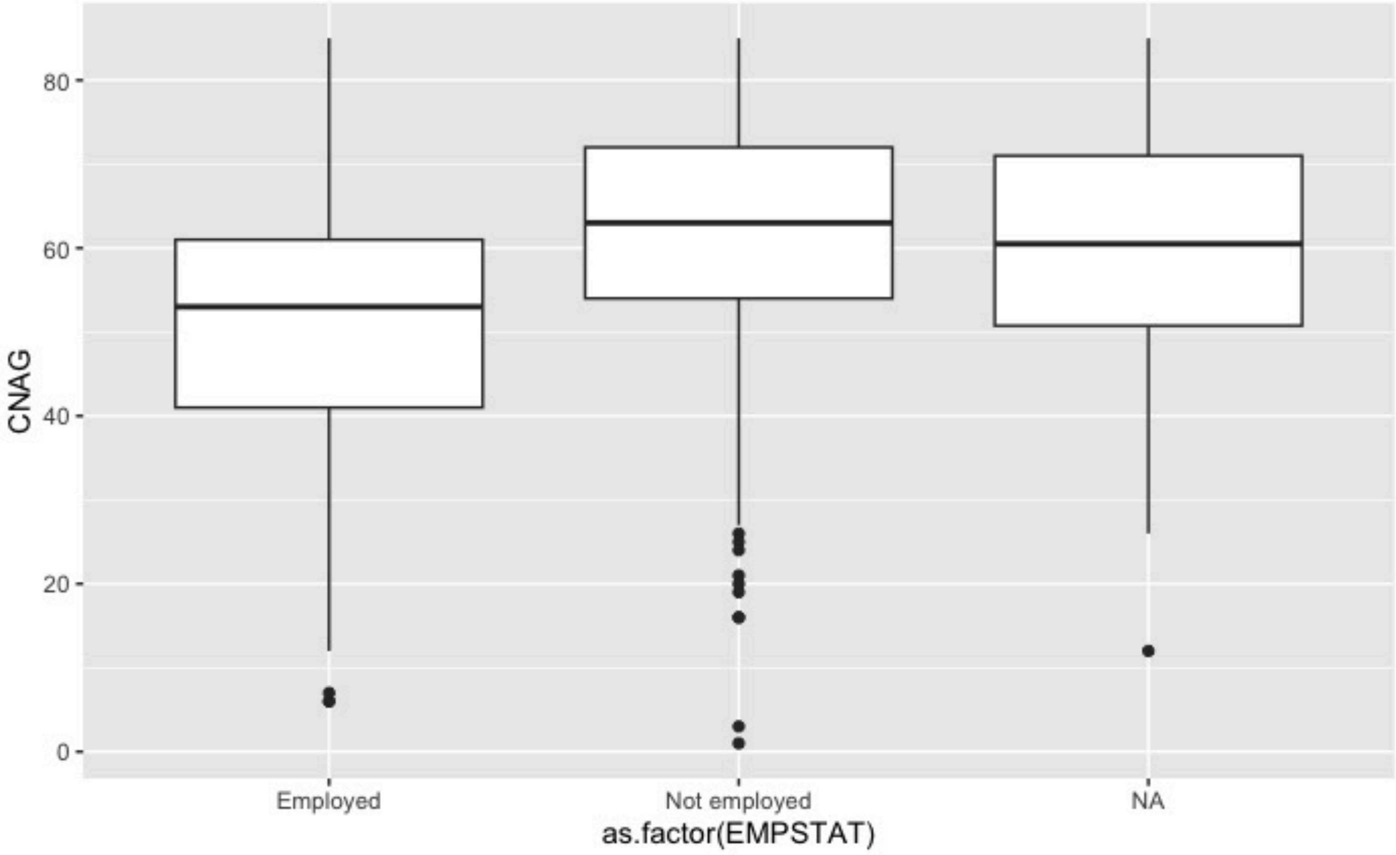
The median age at first diagnosis appears to be roughly similar between males and females.

Education level within patient's family

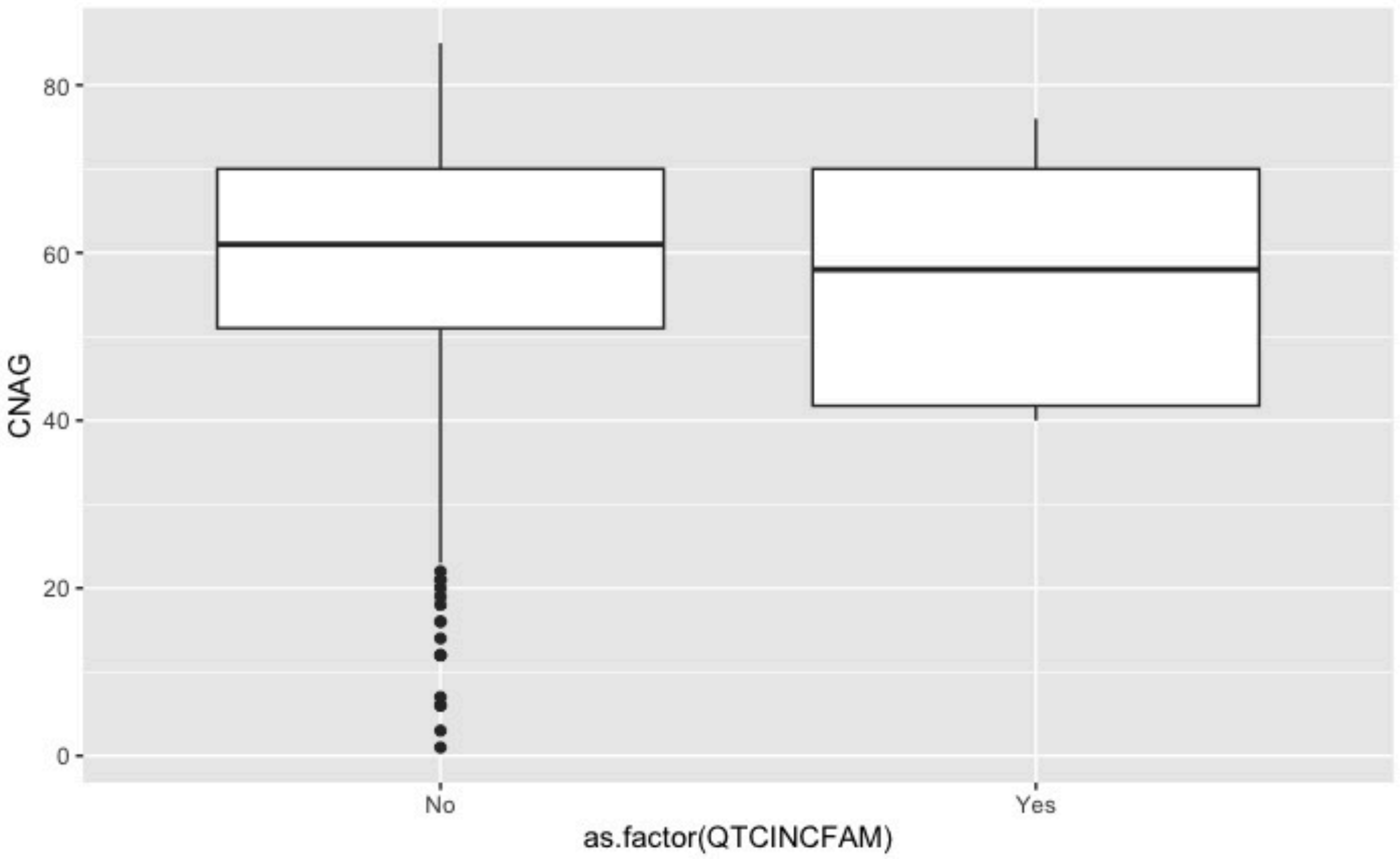
- 1. Grade 0-11
- 2. 12th grade, no diploma
- 3. GED or equivalent
- 4. High school graduate
- 5. Some college, no degree
- 6. Associate degree (occupational, technical, or vocational program)
- 7. Associate degree (academic program)
- 8. Bachelor's degree
- 9. Master's degree
- 10. Professional school degree
- 11. Doctoral degree



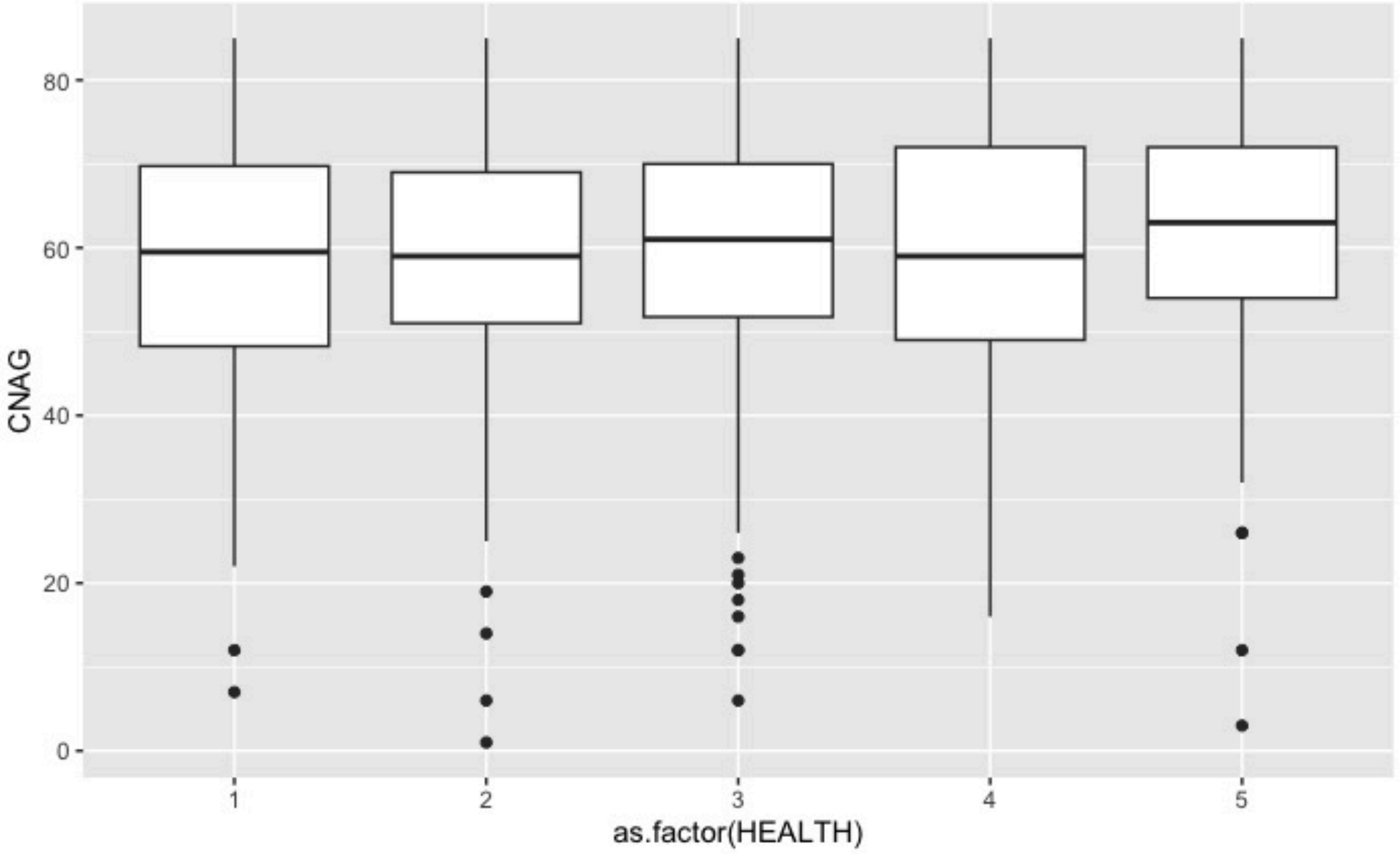
EMPSTAT: whether the adults were working last week



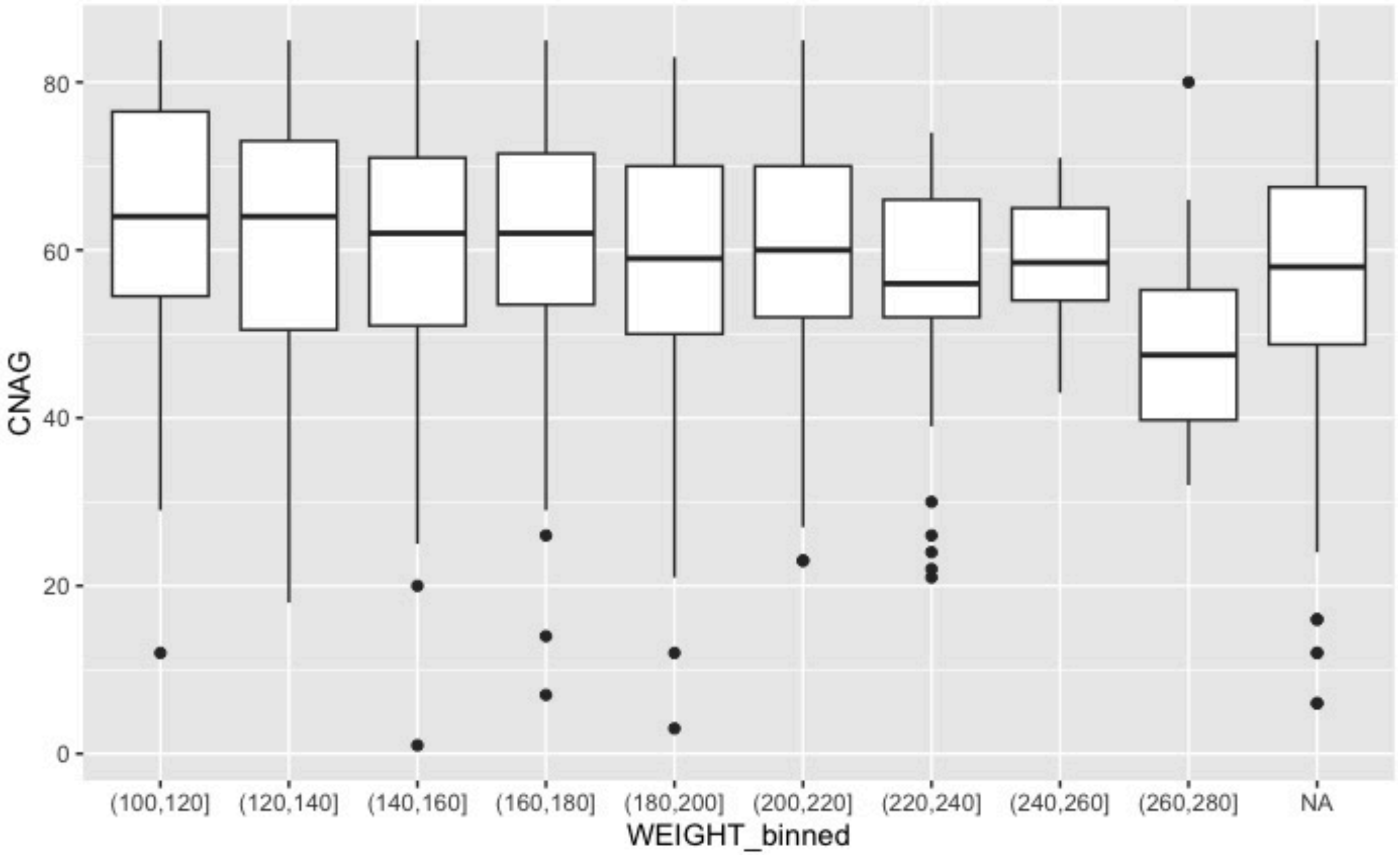
Whether the family income as reported was top coded at $220,000 or more.



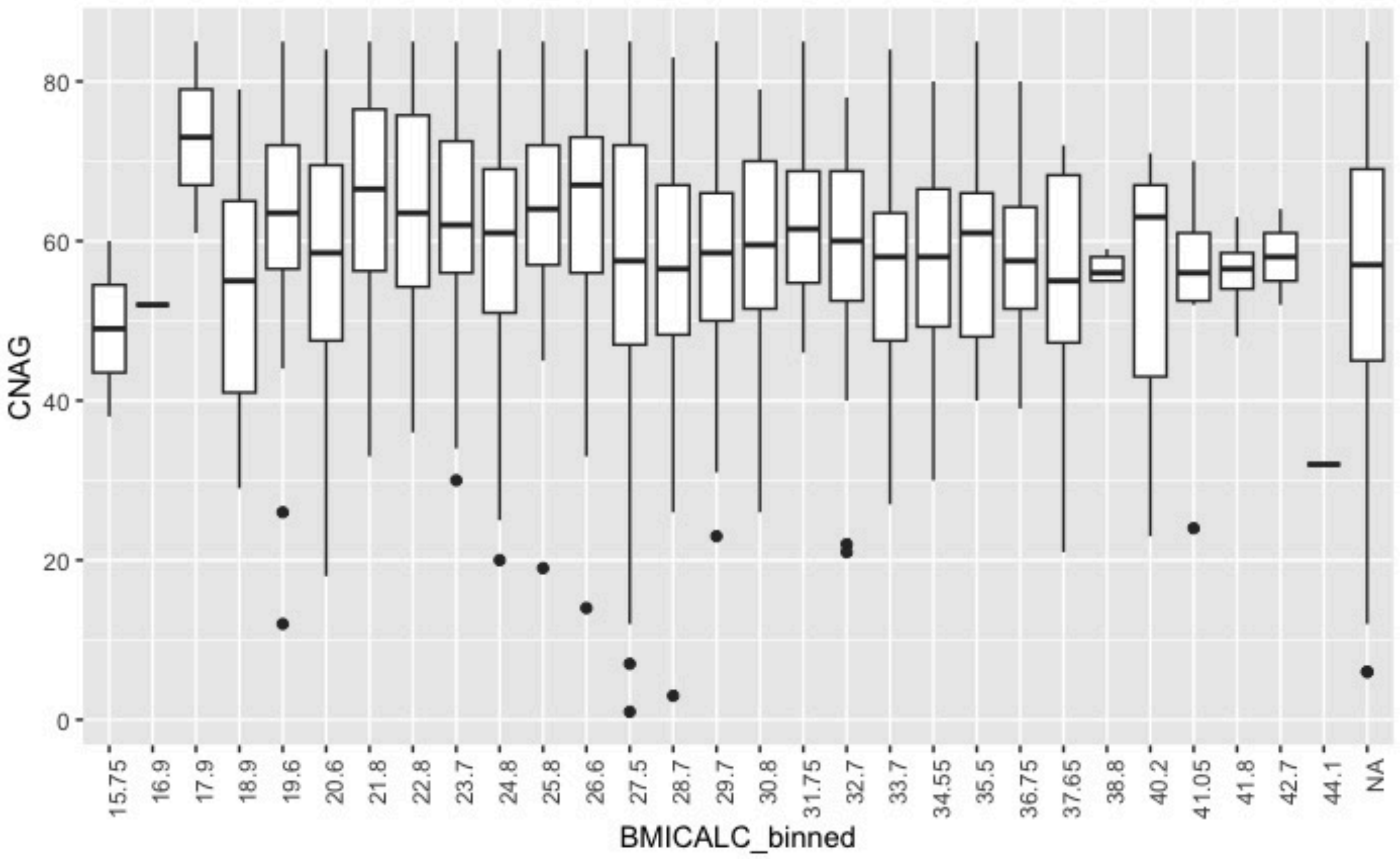
Health status self-reported by the person in question or evaluated by a family member



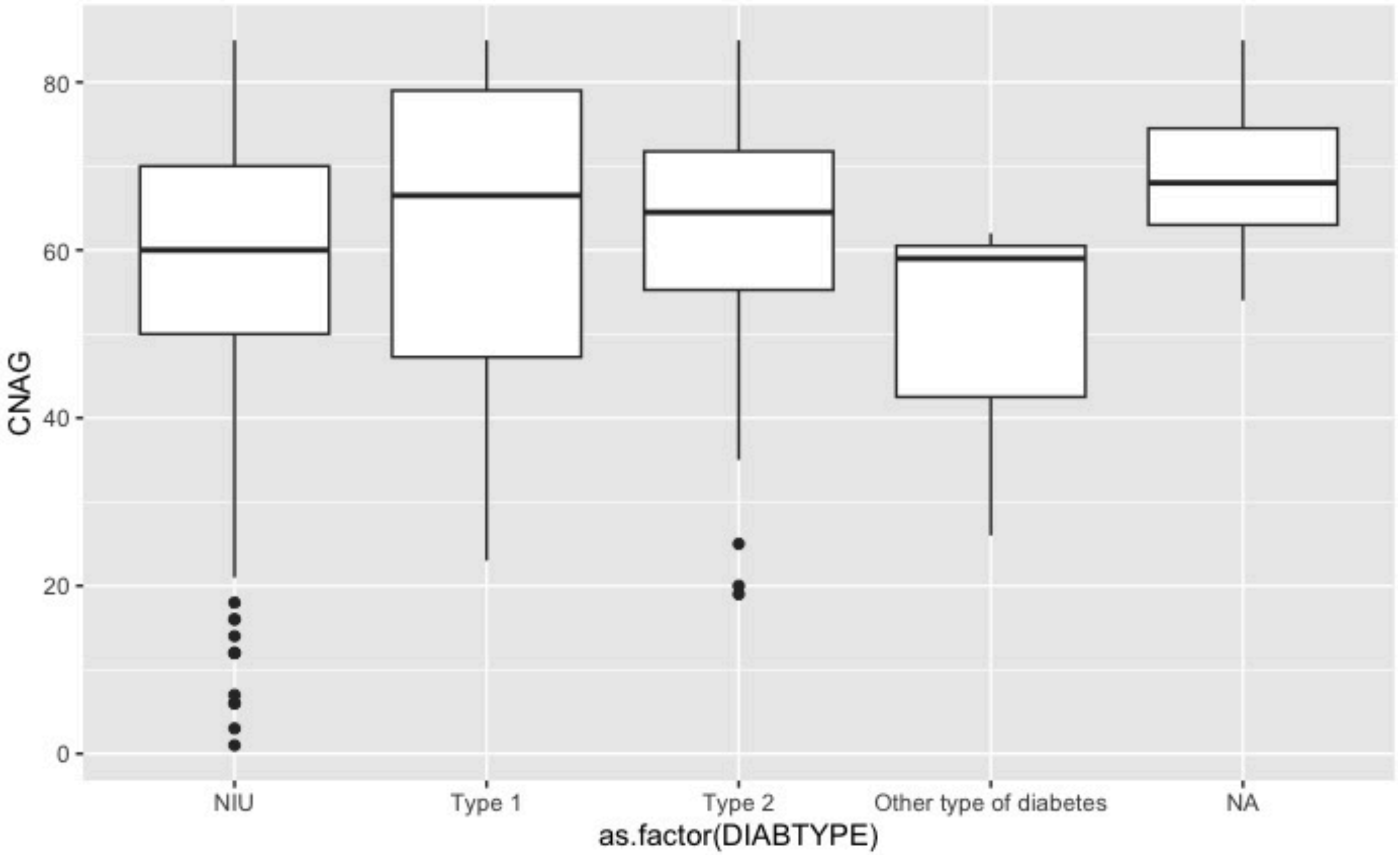
Weight (in pounds) of patients



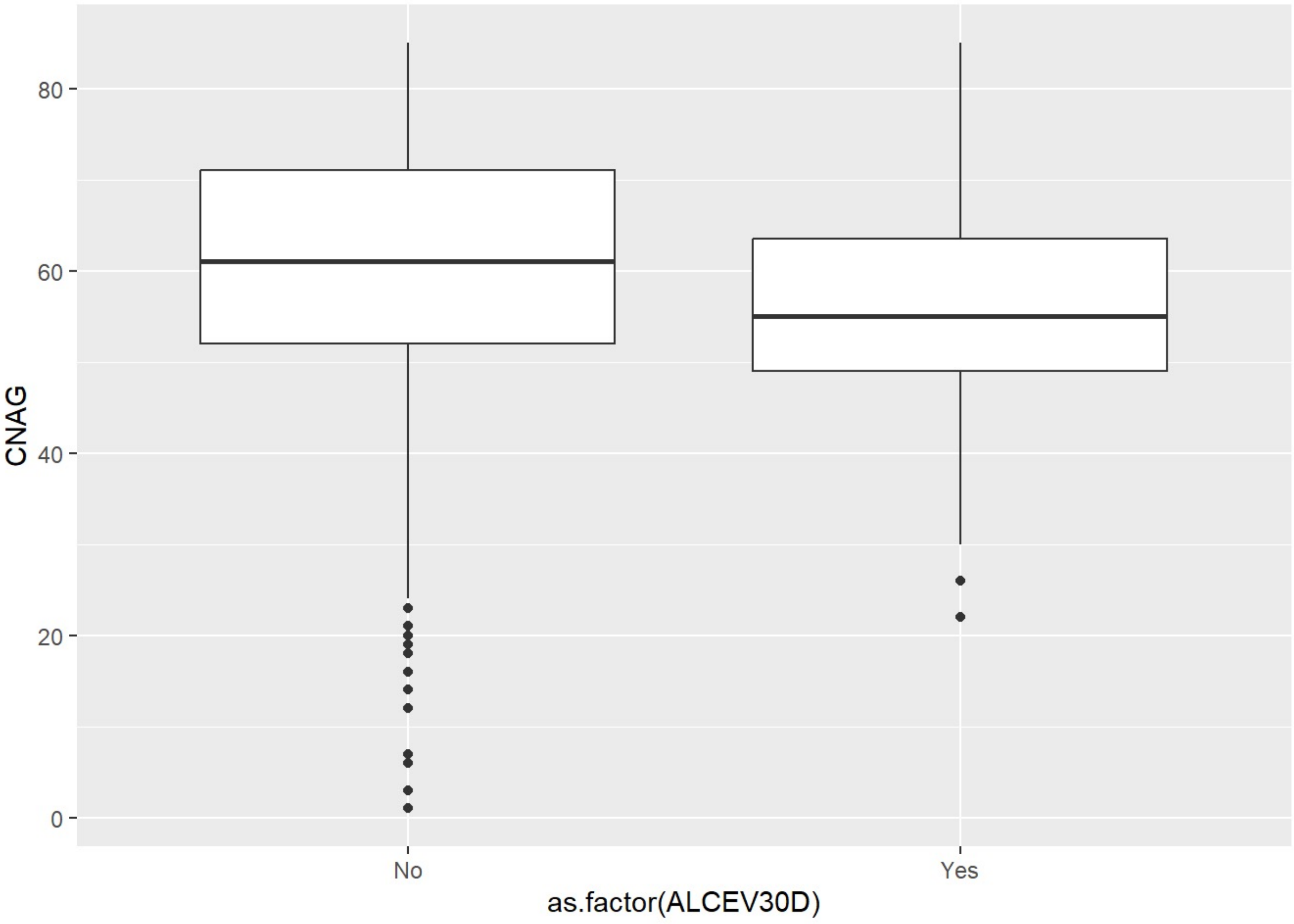
BMI of adults



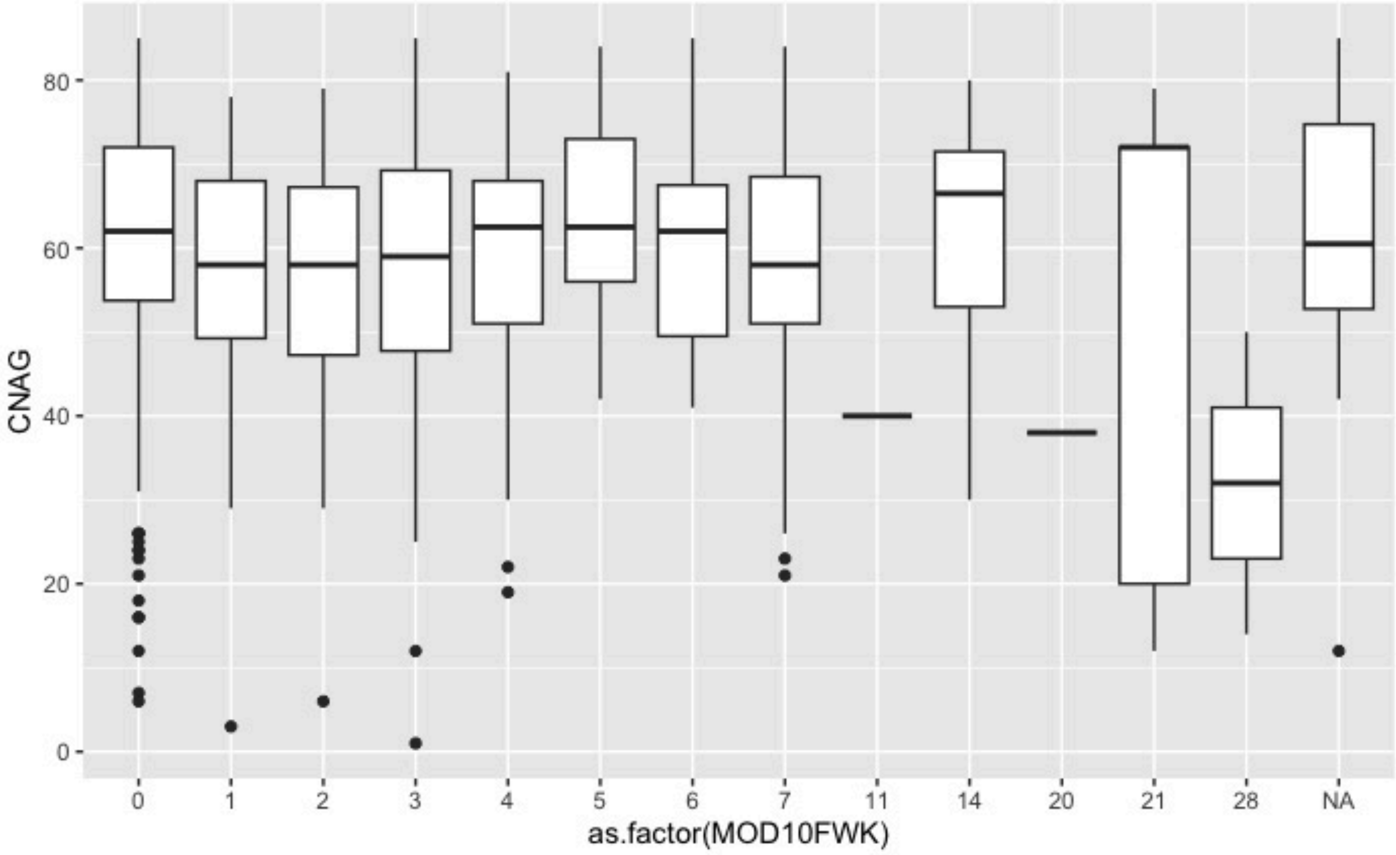
Type of diabetes



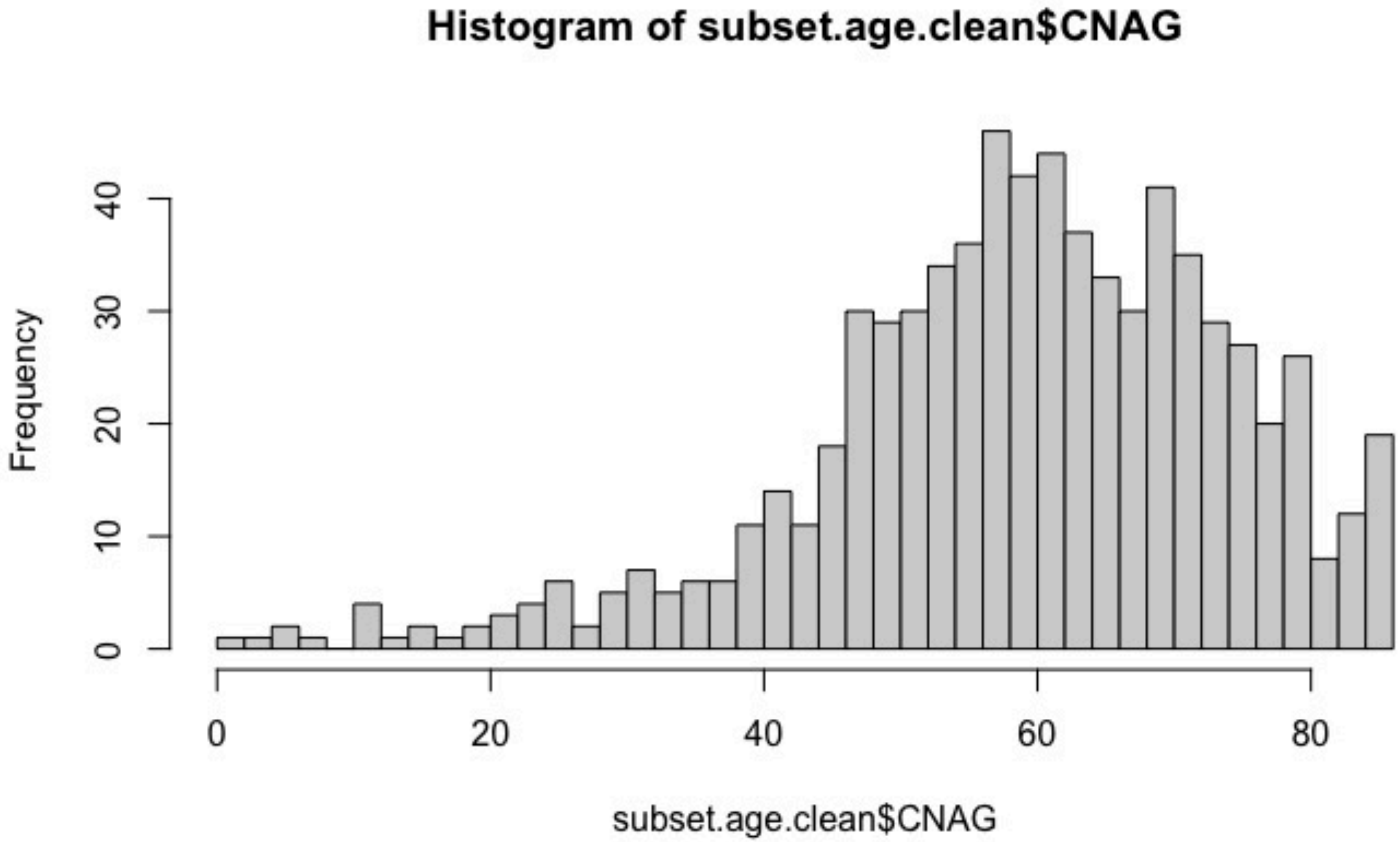
ALCEV30D reports whether, during the past 30 days, they ever had at least one drink.



MOD10FWK reports the frequency(within a week), in number of units, with which sample adults engaged in light or moderate leisure-time physical activities(at least 10 minutes).



Response variable(Age of first cancer diagnosis)



Fit the saturated model with covariates as adults' physical condition, educational background, and drinking habits and Poisson family with log link

```
fit0 = glm(CNAG ~ SEX + MAXEDUC + EMPSTAT + QTCINCFAM
           + HEALTH + HEIGHT + WEIGHT + BMICALC + DIABTYPE
           + ALCEV30D + MOD10FWK,
           data=subset.age.clean, family=poisson(link="log")
           na.action = na.omit)
```

Saturated Model performance

```
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
  (96 observations deleted due to missingness)
AIC: 5786.8
```

Reduced model using backward selection

```
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    
ALCEV30DYes                    -0.050852   0.017398  -2.923 0.003468 ** 
MOD10FWK                       -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
  (96 observations deleted due to missingness)
AIC: 5782.6
```

Performance compared to the full model

```
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
```

AIC saturated: 5786.8
AIC reduced: 5782.6

Fit the saturated model with Negative Binomial and log link

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

Negative Binomial model coefficient analysis

```
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
```

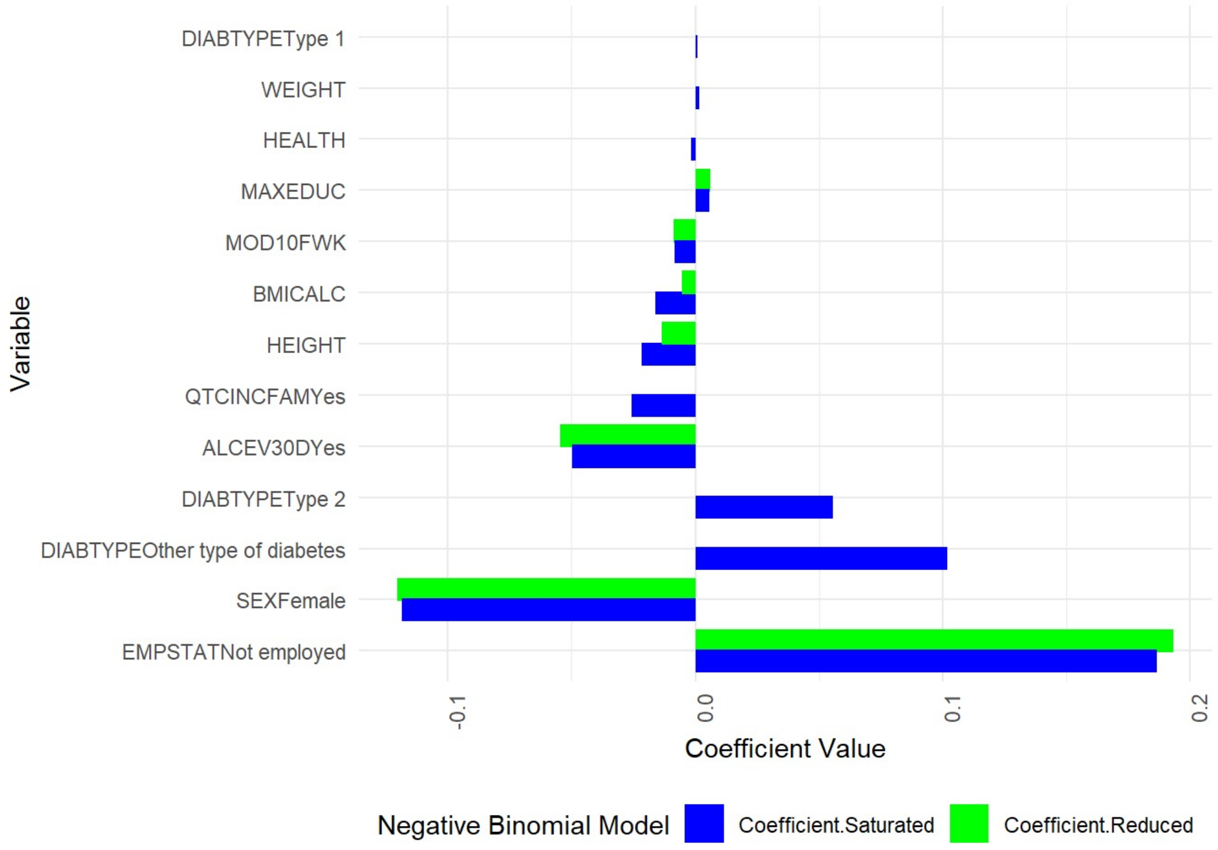
Reduced Negative Binomial model coefficient analysis

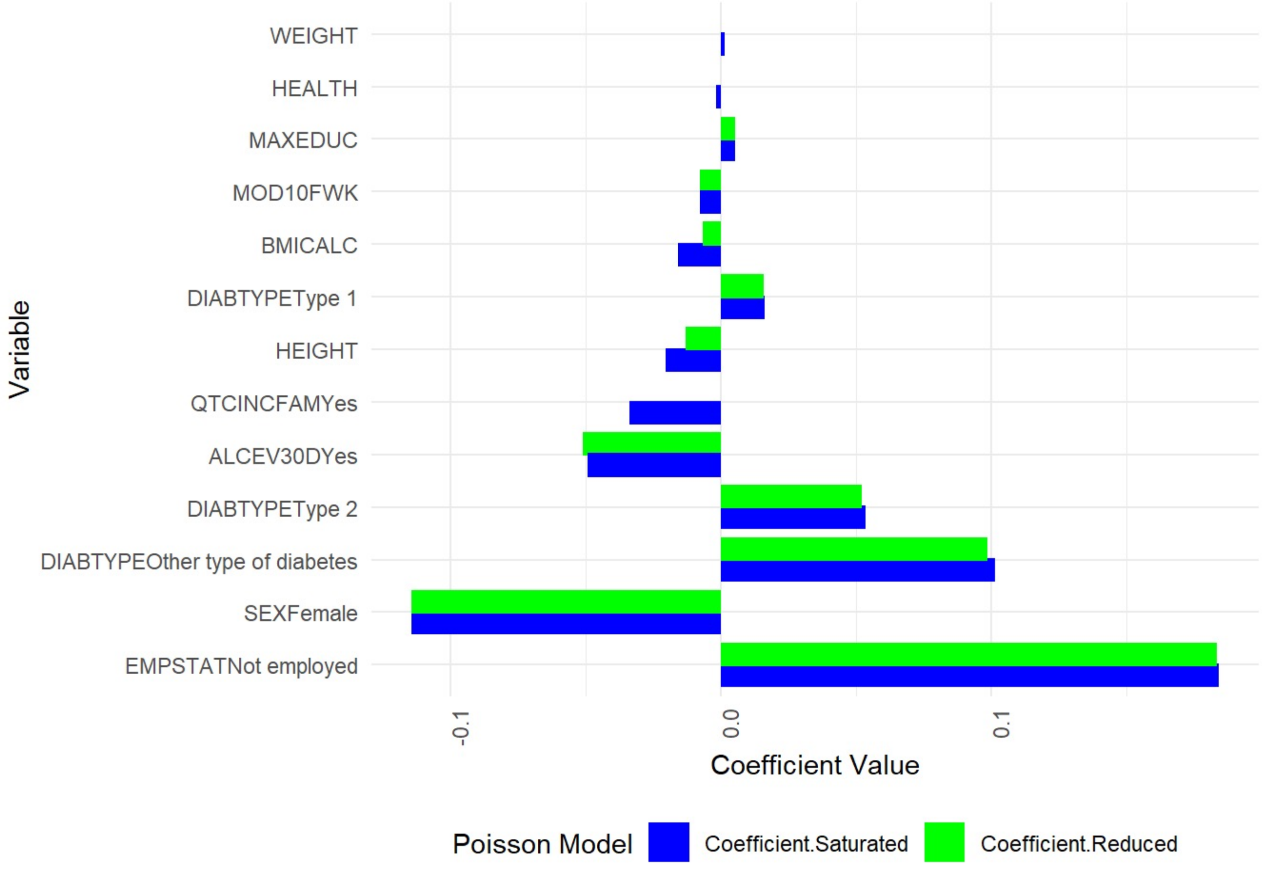
```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)          5.038414   0.274241  18.372  < 2e-16 ***
SEXFemale           -0.120446   0.029025  -4.150 3.33e-05 ***
MAXEDUC              0.005919   0.003831   1.545 0.122342    
EMPSTATNot employed  0.193127   0.023798   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
```

Parameter comparison between Poisson and Negative Binomial model





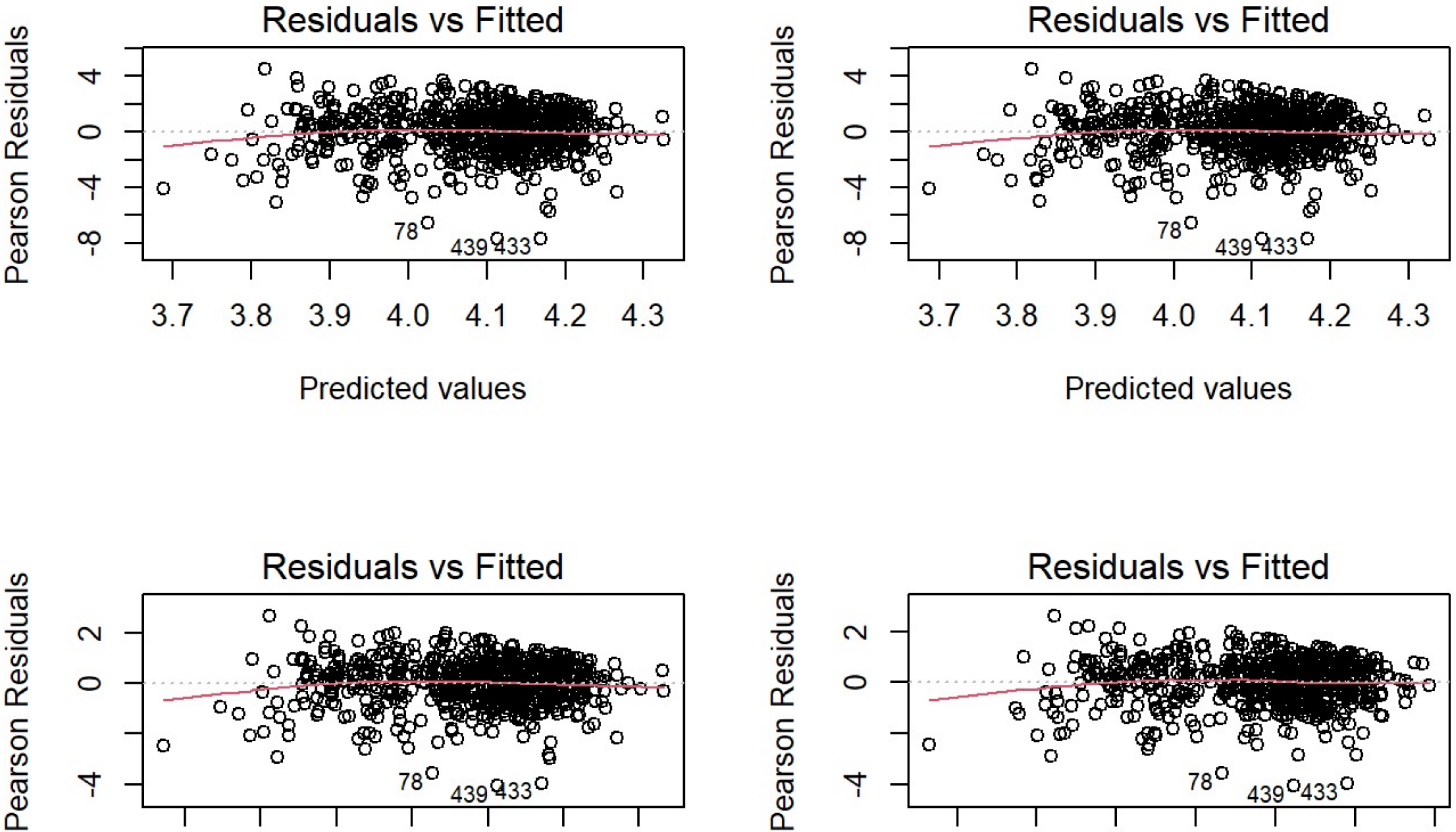
Residual Analysis between the two saturated model and two reduced model

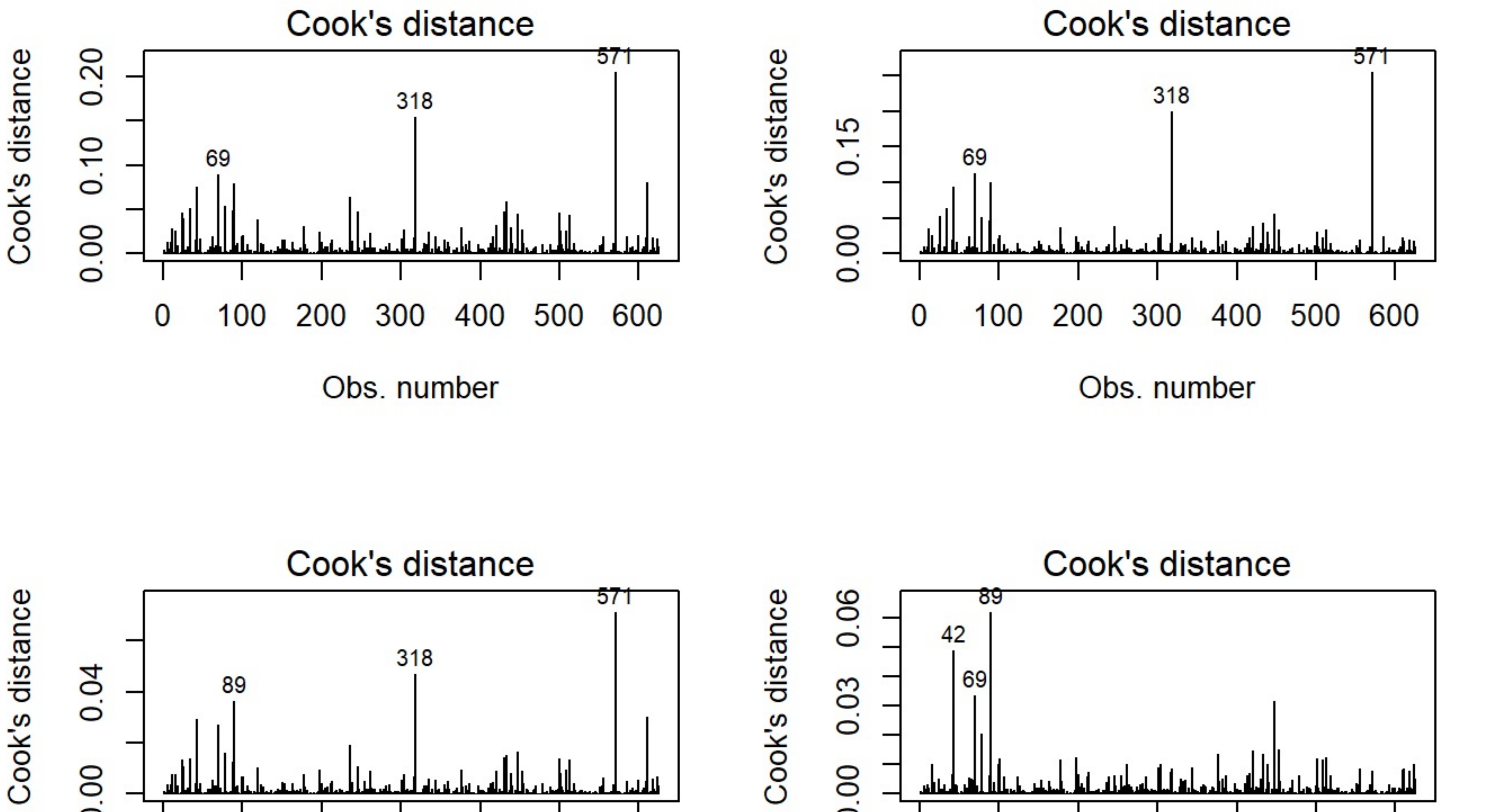
- Saturated Poisson model

Reduced Poisson model

- Saturated Negative Binomial model

Reduced Negative Binomial model


Residuals vs Fitted
Pearson Residuals
Predicted values
4
0
-4
-8
3.7
3.8
3.9
4.0
4.1
4.2
4.3
78
439
433
Residuals vs Fitted
Pearson Residuals
Predicted values
4
0
-4
-8
3.7
3.8
3.9
4.0
4.1
4.2
4.3
78
439
433
Residuals vs Fitted
Pearson Residuals
2
0
-4
78
439
433
Residuals vs Fitted
Pearson Residuals
2
0
-4
78
439
433



Cook's distance
Cook's distance
0.00
0.10
0.20
571
318
69
0
100
200
300
400
500
600
Obs. number
Cook's distance
Cook's distance
0.00
0.15
571
318
69
0
100
200
300
400
500
600
Obs. number
Cook's distance
Cook's distance
0.00
0.04
571
318
89
Cook's distance
Cook's distance
0.00
0.03
0.06
89
42
69

Conclusion

Being female is associated with a statistically significant decrease in the age of first cancer diagnosis compared to males

Having type 1 diabetes is associated with a significant decrease in the age of first cancer diagnosis

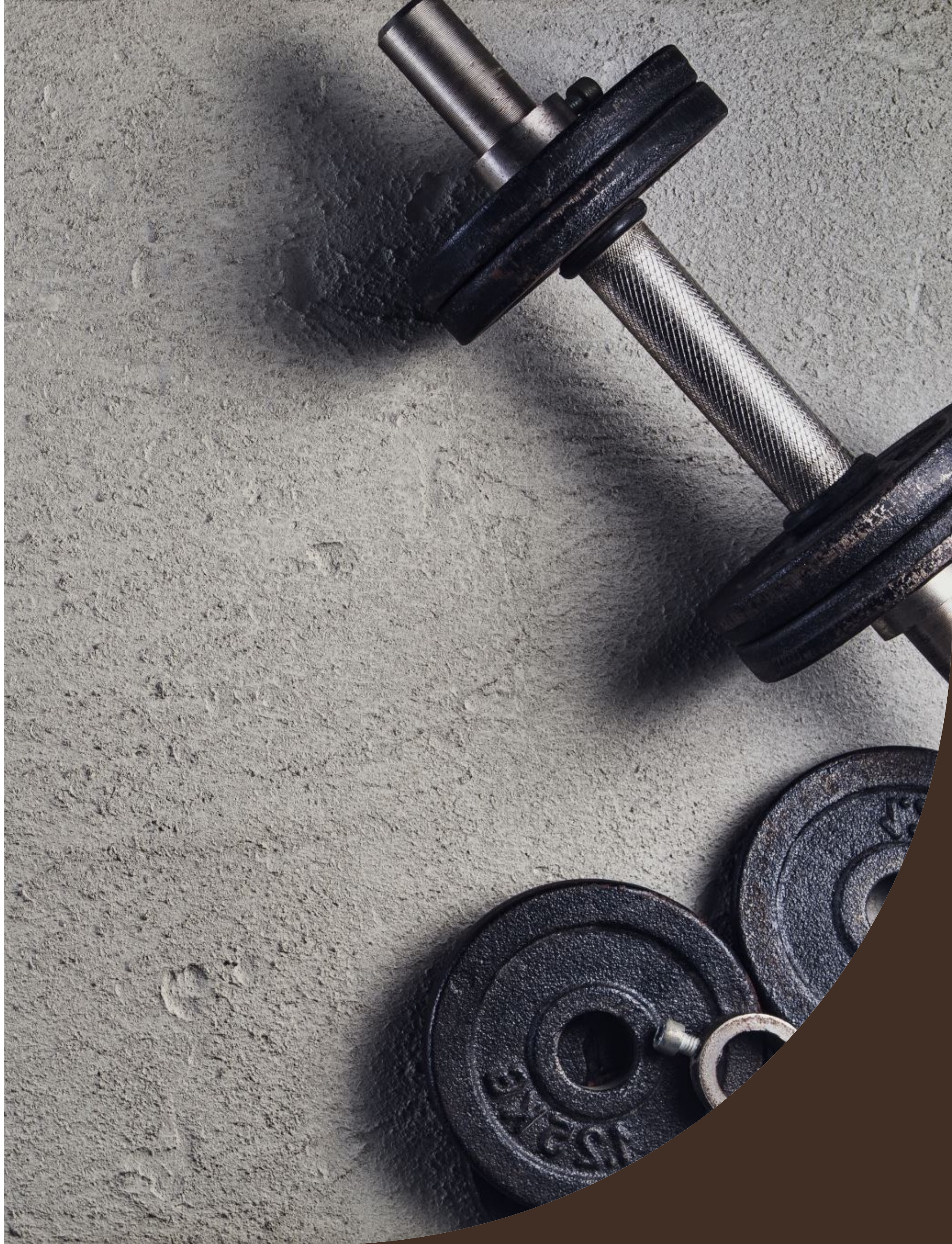
Having type 2 diabetes is associated with a significant increase in the age of first cancer diagnosis

Drinking at least once during a month can significantly reduce the age of first cancer diagnosis

Increasing the frequency of exercise also reduces the age of first cancer diagnosis

Limitations

- People who exercise regularly are diagnosed with cancer at an earlier age, and this is because our dataset is far from balanced, that is, most people who are not diagnosed with cancer are not in our dataset. Our dataset is dominated by older people who do little or no moderate exercise.



Thank you for listening!