Case Study 3 R Supplement, Stat 230

Helen Moses

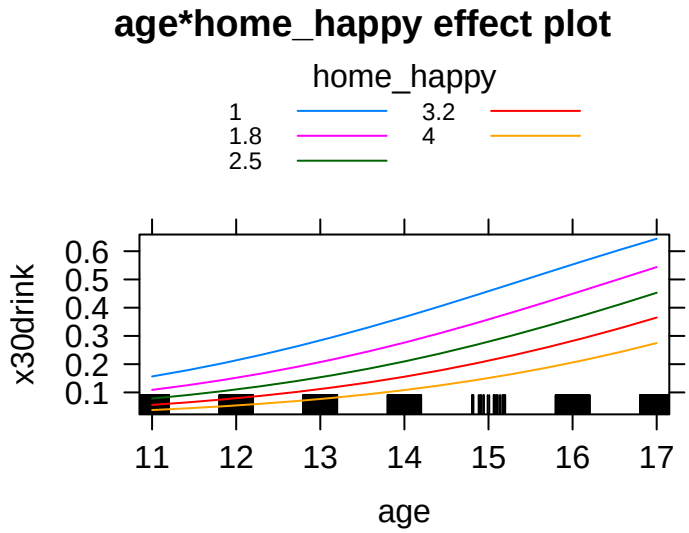
1(Visualize the Data).

```
#Change the data set to binary success/failure (any response greater than zero days is now a success, n
iys$x30drink <- ifelse(iys$x30drink == "0 days", 0, 1)

#Change the data set so that the categorical survey responses become quantitative
iys<- iys %>%
  mutate(home_happy = as.integer(recode_factor(home_happy, "Strongly disagree" = '1', "Disagree" = '2',

iys<- iys %>%
  mutate(home_al_drug = as.integer(recode_factor(home_al_drug, "Strongly disagree" = '1', "Disagree" =

iys<- iys %>%
  mutate(time_no_adult = as.integer(recode_factor(time_no_adult, "0 hours" = '1', "1-2 hours" = '2', "3
```

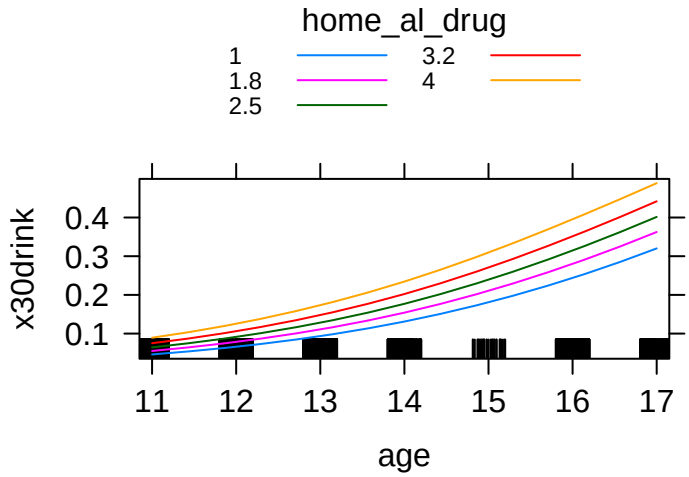
```
#Fitting the initial model
iys_glm <- glm(x30drink ~ age + home_happy + home_al_drug + time_no_adult, data = iys, family ="binomial
summary(iys_glm)
```

```
##
## Call:
## glm(formula = x30drink ~ age + home_happy + home_al_drug + time_no_adult,
##     family = "binomial", data = iys)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8861  -0.6074  -0.3546  -0.2139   2.7200
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.88838    0.45467 -15.150  < 2e-16 ***
## age            0.37951    0.02678  14.172  < 2e-16 ***
## home_happy    -0.52105    0.05831  -8.936  < 2e-16 ***
## home_al_drug   0.23602    0.05055   4.669 3.03e-06 ***
## time_no_adult  0.41589    0.04161   9.994  < 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: 3524.3  on 3554  degrees of freedom
## Residual deviance: 2770.7  on 3550  degrees of freedom
##   (184 observations deleted due to missingness)
## AIC: 2780.7
##
## Number of Fisher Scoring iterations: 5
```

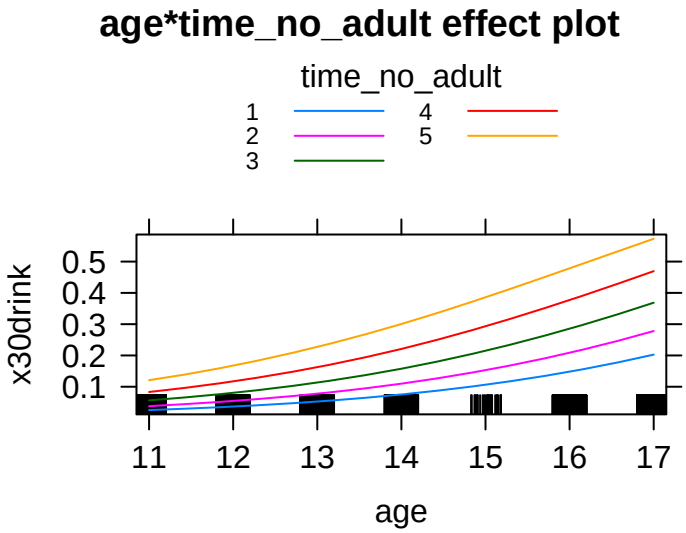
```
#Visualising the data through effects plots
iys_eff <- Effect(c("age", "home_happy"), iys_glm)
plot(iys_eff, multiline = TRUE, type = "response")
```




```
iys_eff1 <- Effect(c("age", "home_al_drug"), iys_glm)
plot(iys_eff1, multiline = TRUE, type = "response")
```




```
iys_eff2 <- Effect(c("age", "time_no_adult"), iys_glm)
plot(iys_eff2, multiline = TRUE, type = "response")
```



Based on the output of the model and the effects plot, there appears to be a clear association between alcohol abuse and the three explanatory variables of interest. However, before confirming this claim, we need to check the model assumptions.

2 (Are Model Assumptions Satisfied).

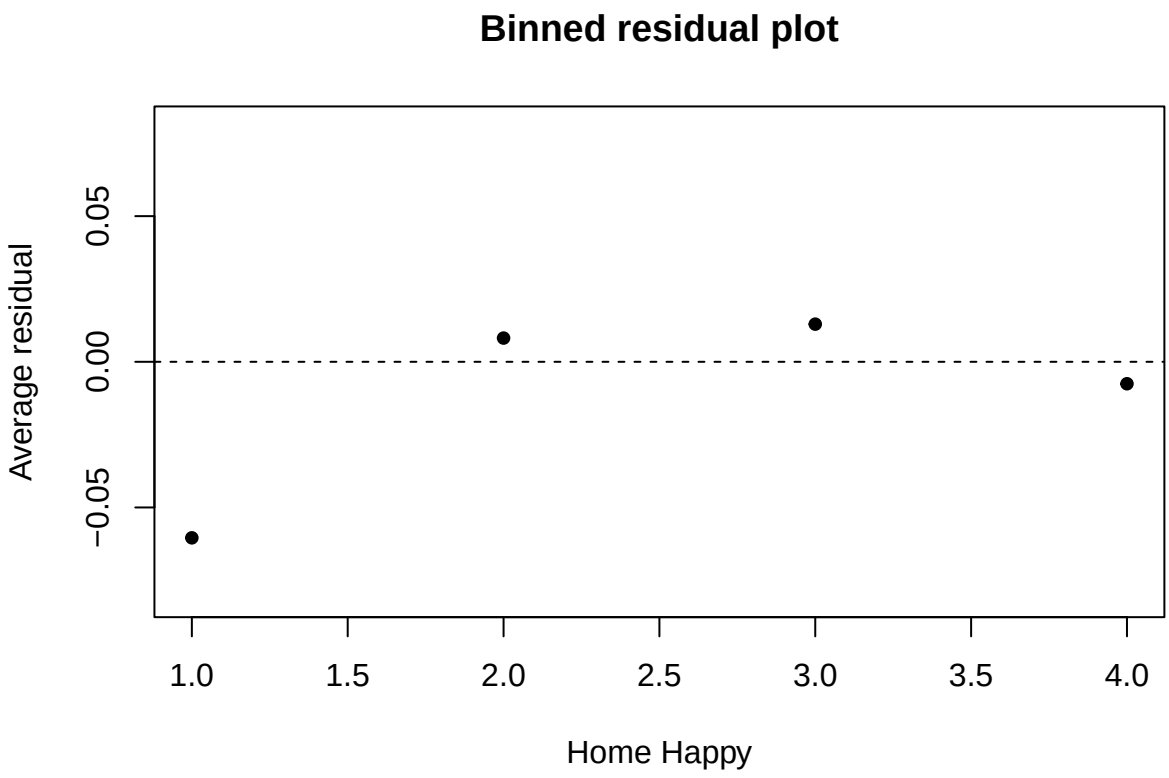
Based on how the data was collected, and what the data is, we know that the assumptions of independence and randomness are met, and we do not have any tools to confirm this in R.

```
# Testing for multicolinearity in the data
vif(iys_glm)
```

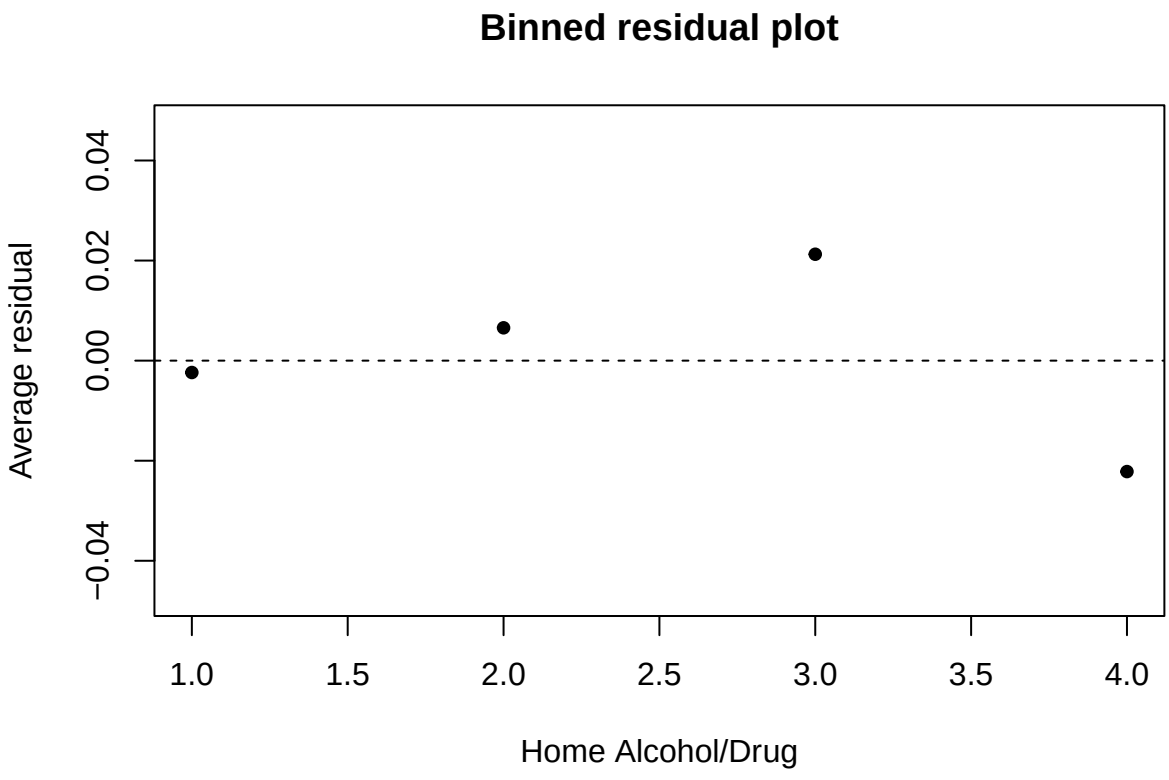
```
##           age    home_happy  home_al_drug time_no_adult
##      1.083302      1.034503      1.031827      1.086635
```

All of the VIF values are low, so there are no concerns of multicollinearity!

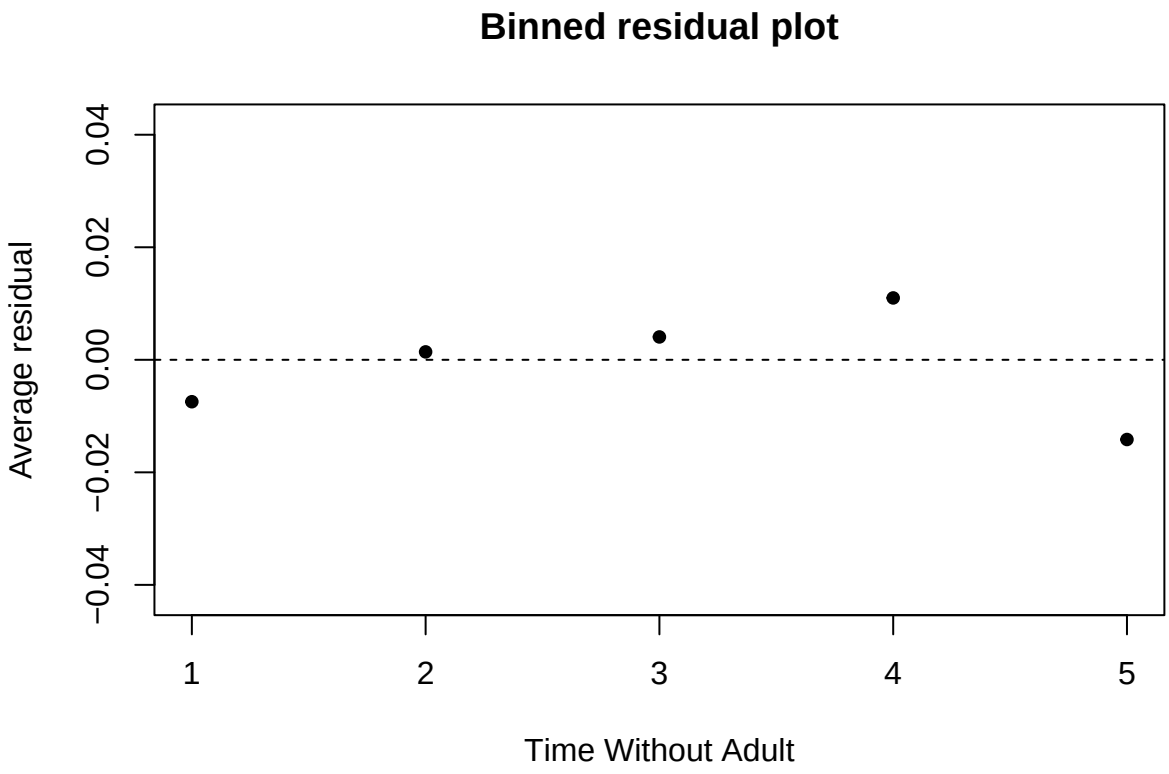
```
# Creating a binned residual plot
iys_glmaug <- augment(iys_glm) |>
  mutate(.resp.resid = resid(iys_glm, type = "response"))

arm::binnedplot(iys_glmaug$home_happy, iys_glmaug$.resp.resid, xlab = "Home Happy", col.int = NULL)
```



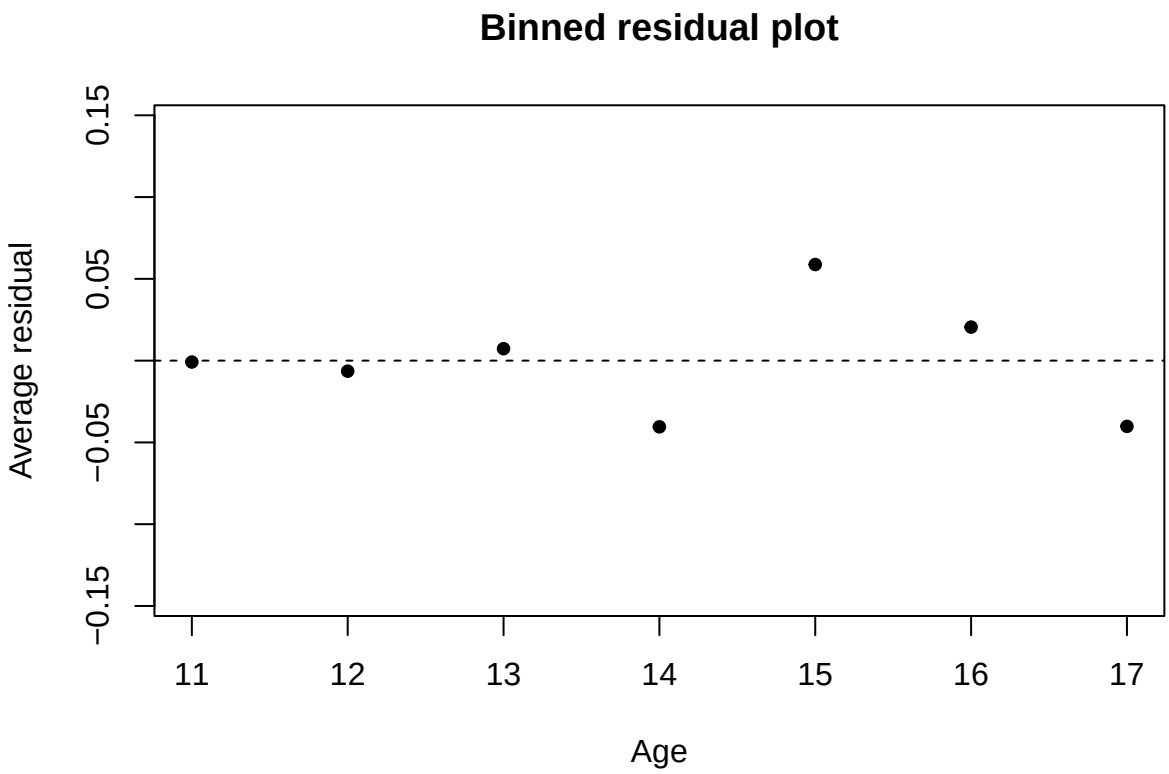

```
arm::binnedplot(iys_glmaug$home_al_drug, iys_glmaug$.resp.resid, xlab = "Home Alcohol/Drug", col.int = 
```




```
arm::binnedplot(iys_glmaug$time_no_adult, iys_glmaug$.resp.resid, xlab = "Time Without Adult", col.int =
```




```
arm::binnedplot(iys_glmaug$age, iys_glmaug$.resp.resid, xlab = "Age", col.int = NULL)
```



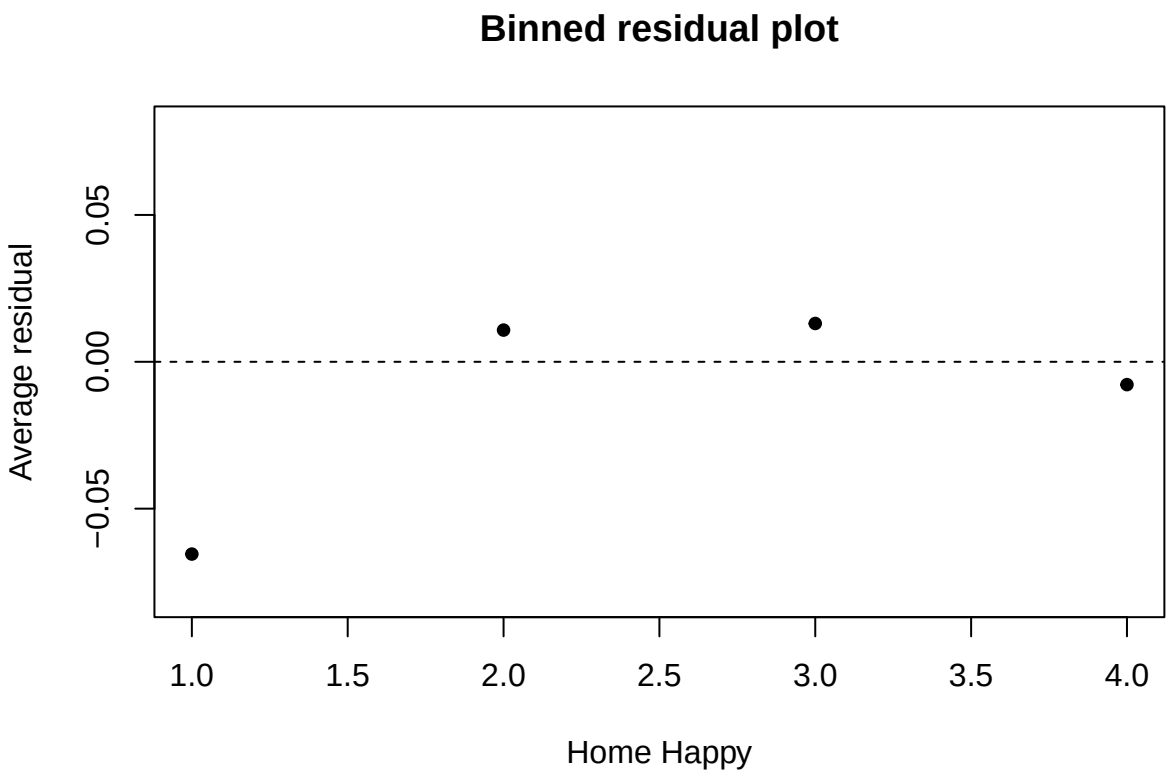
Based on the binned residual plots, we can see that there is a slight concern of non-linearity as a clear curve pattern can be seen in all of them. However, in the age plot, we can see evidence of a bit of a tail, suggesting that interaction terms might be necessary.

3(Transforming the Model).

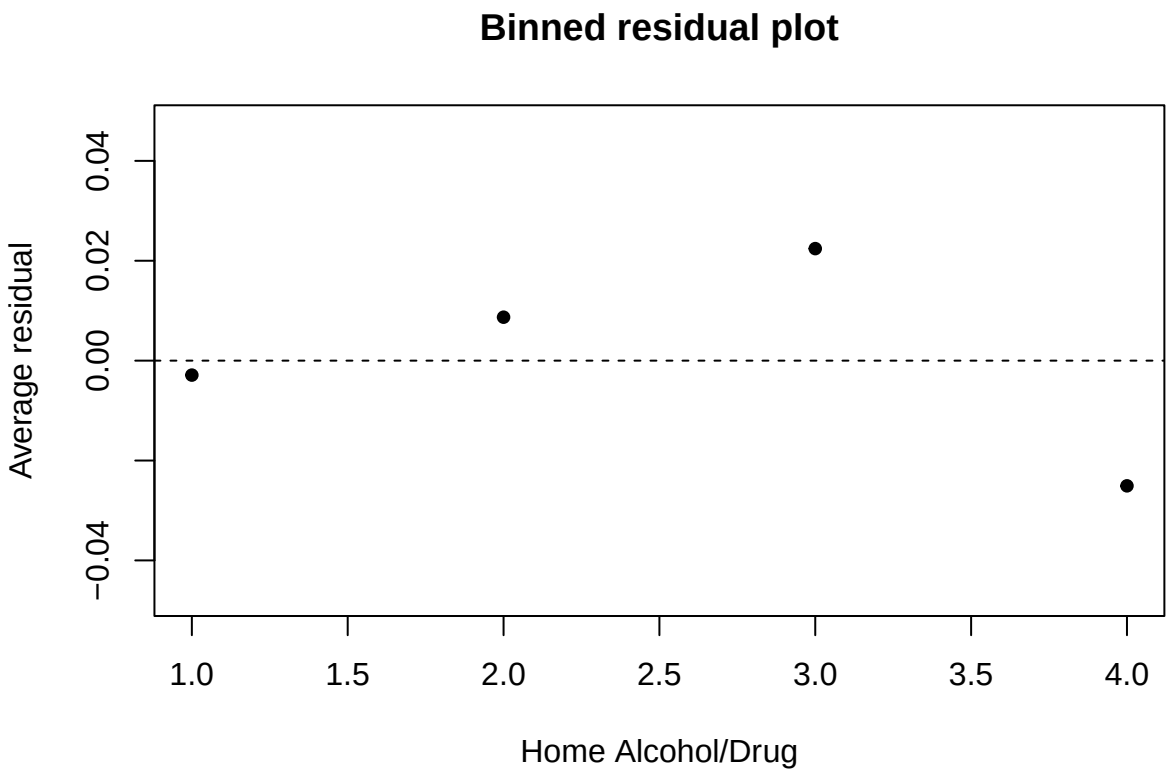
```
iys_glm1 <- glm(x30drink ~ age + home_happy + home_al_drug + time_no_adult + age*home_happy + age*home_a
summary(iys_glm1)
```

```
##
## Call:
## glm(formula = x30drink ~ age + home_happy + home_al_drug + time_no_adult +
##     age * home_happy + age * home_al_drug + age * time_no_adult,
##     family = "binomial", data = iys)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7677  -0.6208  -0.3608  -0.1921   2.7288
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -5.556718   2.019561  -2.751  0.00593 **
## age                0.289307   0.137360   2.106  0.03519 *
## home_happy        -1.233046   0.451639  -2.730  0.00633 **
## home_al_drug       0.707486   0.377028   1.876  0.06059 .
## time_no_adult      0.429724   0.314110   1.368  0.17129
## age:home_happy     0.048548   0.030571   1.588  0.11227
## age:home_al_drug  -0.032396   0.025699  -1.261  0.20745
## age:time_no_adult -0.001045   0.021255  -0.049  0.96078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3524.3  on 3554  degrees of freedom
## Residual deviance: 2766.0  on 3547  degrees of freedom
##   (184 observations deleted due to missingness)
## AIC: 2782
##
## Number of Fisher Scoring iterations: 6
```

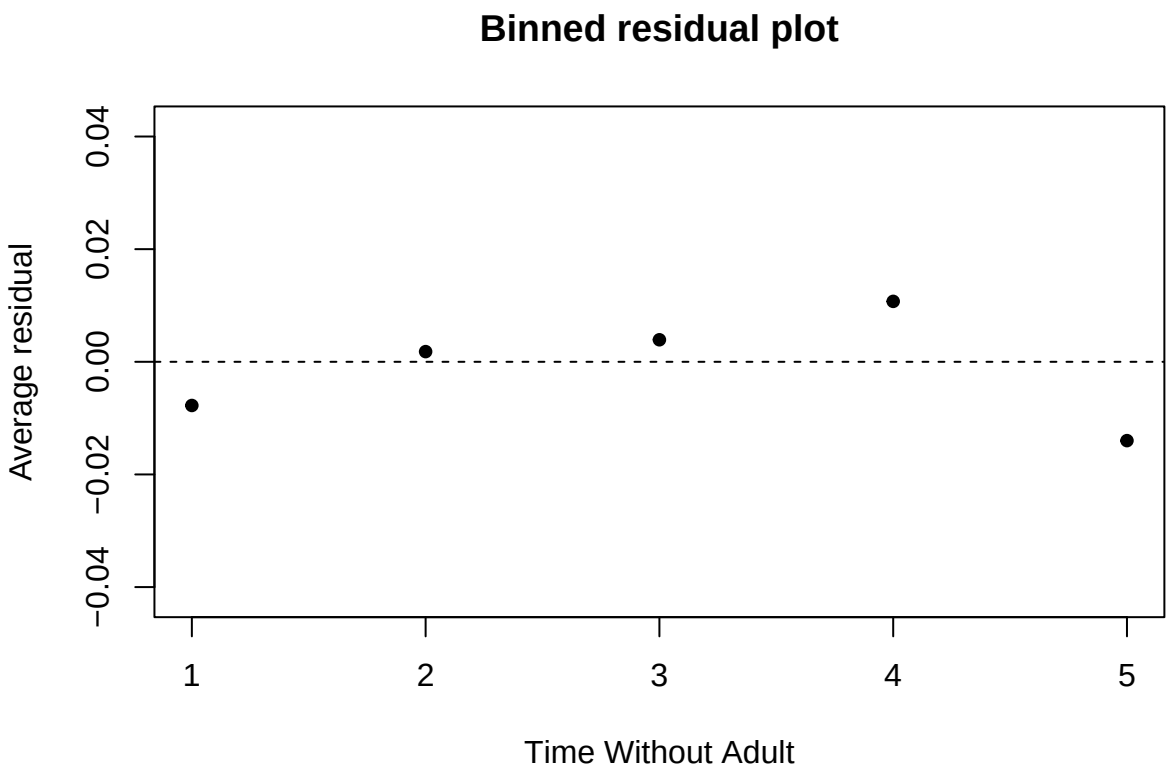
```
iys_glmaug1 <- augment(iys_glm1) |>
  mutate(.resp.resid1 = resid(iys_glm1, type = "response"))

arm::binnedplot(iys_glmaug1$home_happy, iys_glmaug1$.resp.resid1, xlab = "Home Happy", col.int = NULL)
```



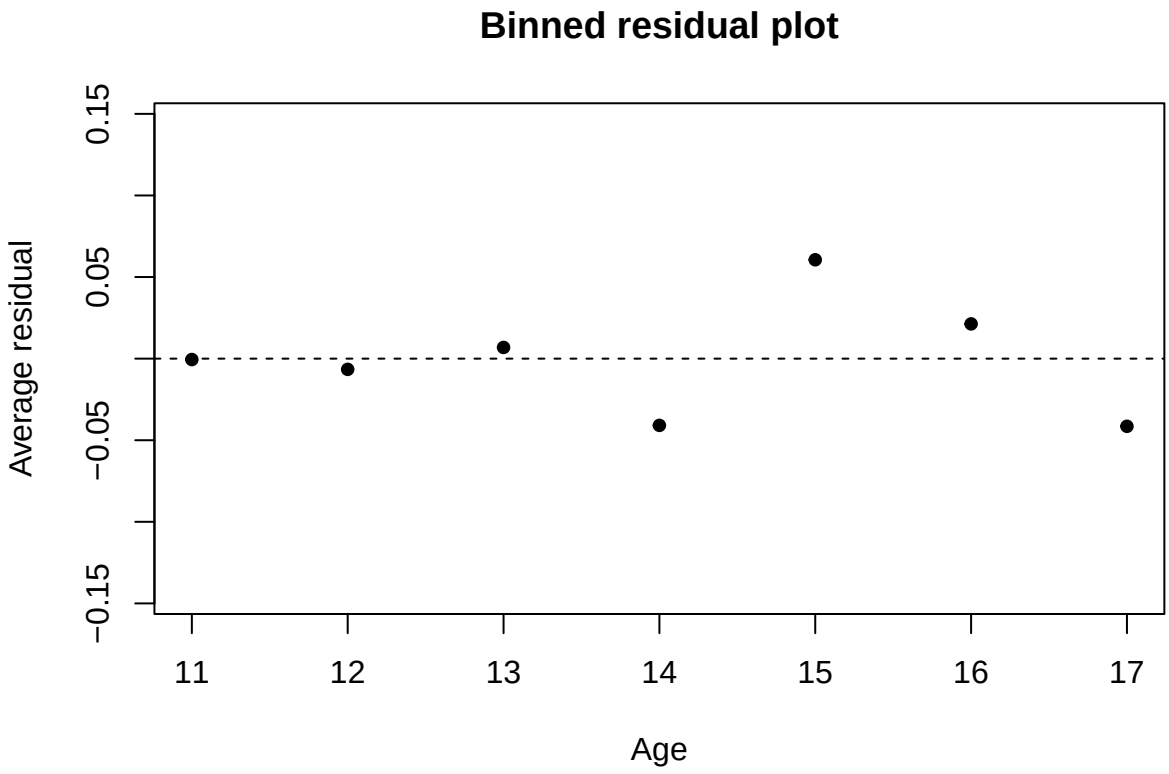

```
arm::binnedplot(iys_glmaug1$home_al_drug, iys_glmaug1$.resp.resid1, xlab = "Home Alcohol/Drug", col.int
```




```
arm::binnedplot(iys_glmaug1$time_no_adult, iys_glmaug1$.resp.resid1, xlab = "Time Without Adult", col.i
```



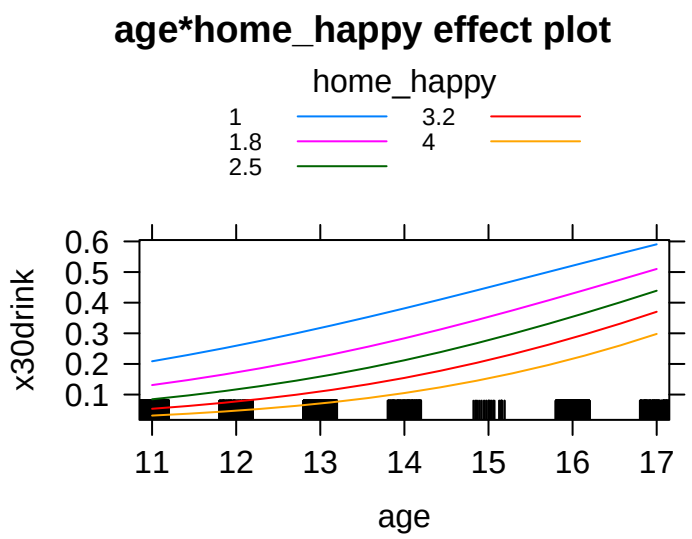

```
arm::binnedplot(iys_glmaug1$age, iys_glmaug1$.resp.resid1, xlab = "Age", col.int = NULL)
```



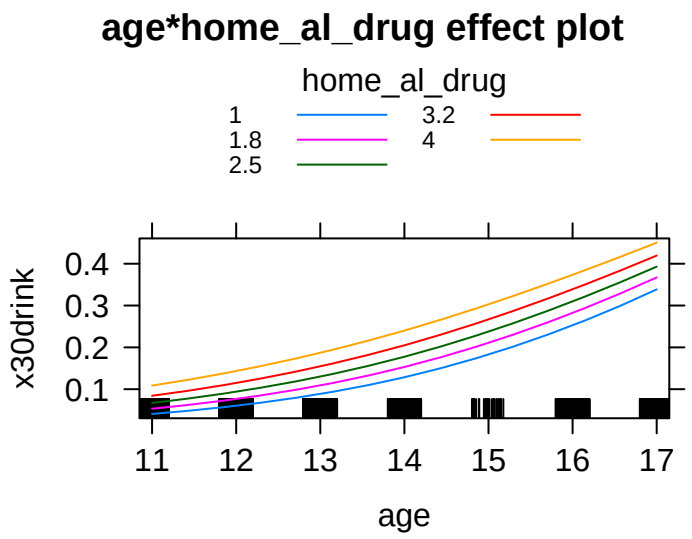
The binned residual plots look the exact same, suggesting there are still problems of non linearity.

```
#Visualising the new data through effects plots
iys_eff <- Effect(c("age", "home_happy"), iys_glm1)
```

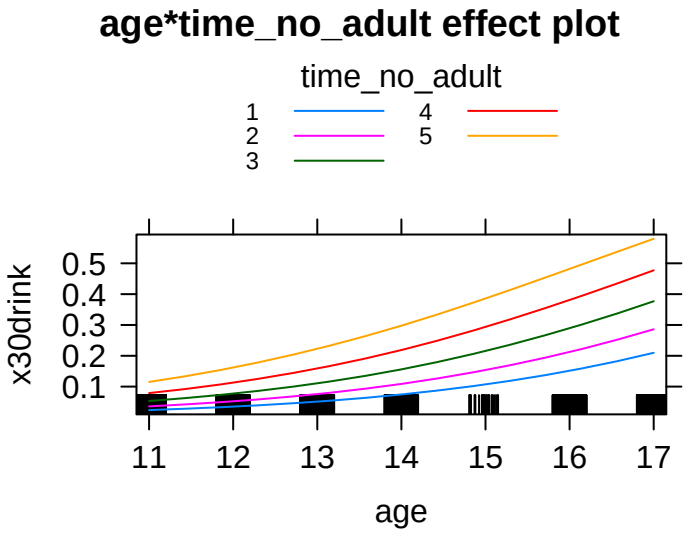
```
plot(iys_eff, multiline = TRUE, type = "response")
```




```
iys_eff1 <- Effect(c("age", "home_al_drug"), iys_glm1)
plot(iys_eff1, multiline = TRUE, type = "response")
```




```
iys_eff2 <- Effect(c("age", "time_no_adult"), iys_glm1)
plot(iys_eff2, multiline = TRUE, type = "response")
```



The new effects plots show the same association as before, but the lines are slightly less curved in each of the plots.

Even though the data still has concerns of linearity, I will still test to see if the new model has more predictive power.

```
#Original model Somers' D and Goodman-Kruskal gamma
lrm(x30drink ~ age + home_happy + home_al_drug + time_no_adult, data = iys)
```

```
## Frequencies of Missing Values Due to Each Variable
##      x30drink           age    home_happy  home_al_drug time_no_adult
##             0             0            33            51           111
##
## Logistic Regression Model
##
##  lrm(formula = x30drink ~ age + home_happy + home_al_drug + time_no_adult,
##      data = iys)
##
##
##                         Model Likelihood       Discrimination    Rank Discrim.
##                               Ratio Test              Indexes          Indexes
##  Obs          3555    LR chi2     753.62       R2       0.304    C       0.815
##   0           2856    d.f.             4      R2(4,3555)0.190    Dxy     0.629
##   1            699    Pr(> chi2) <0.0001    R2(4,1684.7)0.359    gamma   0.634
##  max |deriv| 1e-07                            Brier    0.124    tau-a   0.199
##
##                Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept     -6.8884 0.4547 -15.15 <0.0001
##  age            0.3795 0.0268  14.17 <0.0001
##  home_happy    -0.5211 0.0583  -8.94 <0.0001
##  home_al_drug   0.2360 0.0506   4.67 <0.0001
##  time_no_adult  0.4159 0.0416   9.99 <0.0001
##
```

```
#New model Somers' D and Goodman-Kruskal gamma
lrm(x30drink ~ age + home_happy + home_al_drug + time_no_adult + age*home_happy + age*home_al_drug + ag
```

```
## Frequencies of Missing Values Due to Each Variable
##      x30drink           age    home_happy  home_al_drug time_no_adult
```

```
##             0             0            33            51           111
##
## Logistic Regression Model
##
##  lrm(formula = x30drink ~ age + home_happy + home_al_drug + time_no_adult +
##      age * home_happy + age * home_al_drug + age * time_no_adult,
##      data = iys)
##
##
##                        Model Likelihood       Discrimination    Rank Discrim.
##                              Ratio Test              Indexes          Indexes
##  Obs          3555    LR chi2     758.31       R2       0.305    C       0.815
##   0           2856    d.f.             7      R2(7,3555)0.191    Dxy     0.629
##   1            699    Pr(> chi2) <0.0001    R2(7,1684.7)0.360    gamma   0.633
##  max |deriv| 1e-11                            Brier    0.124    tau-a   0.199
##
##                      Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept           -5.5567 2.0196 -2.75  0.0059
##  age                  0.2893 0.1374  2.11  0.0352
##  home_happy          -1.2330 0.4516 -2.73  0.0063
##  home_al_drug         0.7075 0.3770  1.88  0.0606
##  time_no_adult        0.4297 0.3141  1.37  0.1713
##  age * home_happy     0.0485 0.0306  1.59  0.1123
##  age * home_al_drug  -0.0324 0.0257 -1.26  0.2074
##  age * time_no_adult -0.0010 0.0213 -0.05  0.9608
##
```

The Somers' D and Goodman-Kruskal gamma are the same for both models, suggesting the interaction terms do not change the model's predictive powers. To confirm this, I will conduct a drop in deviance test, as shown below.

G: 2770.7 - 2766.0 = 4.6

p-value: 1-pchisq(4.6, df = 3) = .204

*This confirms the intercation terms do not add any predictive powers.

Because the addition of those interaction terms did not improve the model or resolve the concern of non linearity, I need to try a new transformation.

4 (New Transformation).

```
# Log transforming the explanatory variables
iys <- iys |>
  mutate(loghome_happy = log(home_happy))

iys <- iys |>
  mutate(loghome_al_drug = log(home_al_drug))

iys <- iys |>
  mutate(logtime_no_adult = log(time_no_adult))
```

```
#Fitting the new log transformed model
iys_glm2 <- glm(x30drink ~ age + loghome_happy + loghome_al_drug + logtime_no_adult, data = iys, family
summary(iys_glm2)
```

```
##
```

```
## Call:
## glm(formula = x30drink ~ age + loghome_happy + loghome_al_drug +
##     logtime_no_adult, family = "binomial", data = iys)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9359  -0.5955  -0.3529  -0.1825   2.7941
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.04010    0.42671 -16.498  < 2e-16 ***
## age               0.38127    0.02676  14.245  < 2e-16 ***
## loghome_happy    -1.14634    0.14030  -8.170 3.07e-16 ***
## loghome_al_drug   0.50271    0.09809   5.125 2.97e-07 ***
## logtime_no_adult  1.19130    0.12361   9.638  < 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: 3524.3  on 3554  degrees of freedom
## Residual deviance: 2781.4  on 3550  degrees of freedom
##   (184 observations deleted due to missingness)
## AIC: 2791.4
##
## Number of Fisher Scoring iterations: 5
```

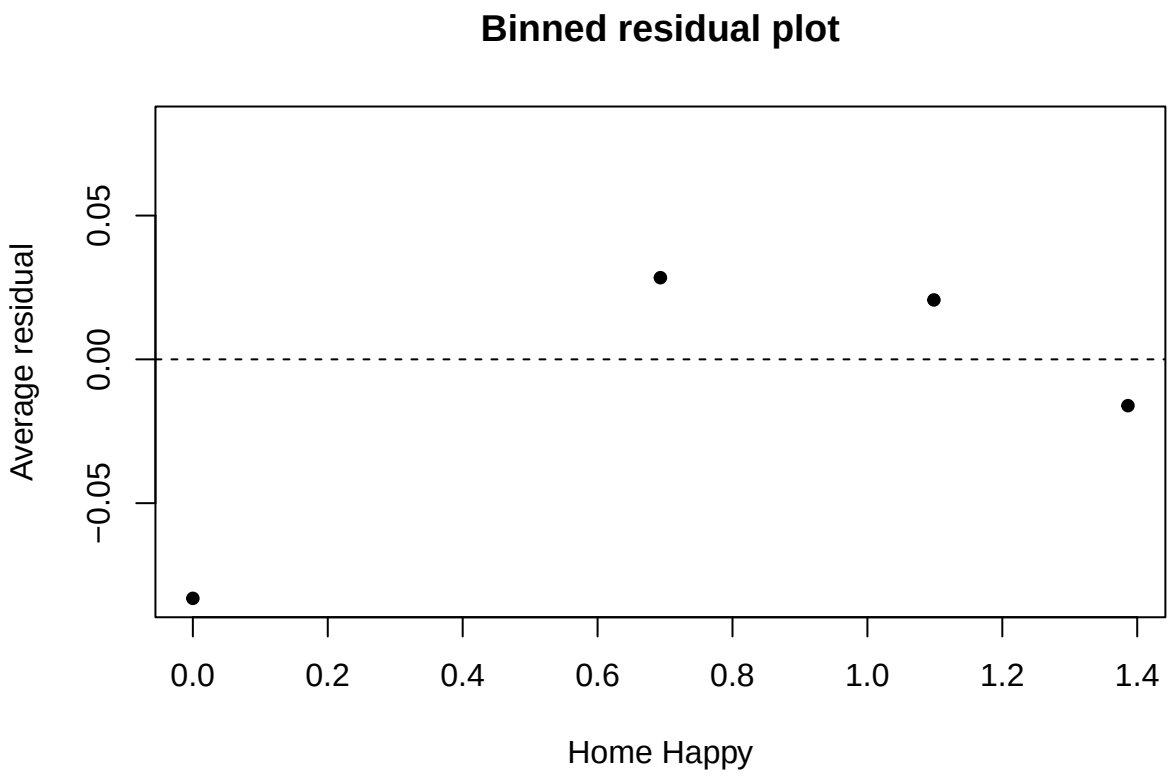
```
#New model Somers' D and Goodman-Kruskal gamma
lrm(x30drink ~ age + loghome_happy + loghome_al_drug + logtime_no_adult, data = iys)
```

```
## Frequencies of Missing Values Due to Each Variable
##         x30drink              age    loghome_happy  loghome_al_drug
##                0                0               33               51
## logtime_no_adult
##              111
##
## Logistic Regression Model
##
##  lrm(formula = x30drink ~ age + loghome_happy + loghome_al_drug +
##      logtime_no_adult, data = iys)
##
##
##                        Model Likelihood       Discrimination    Rank Discrim.
##                              Ratio Test              Indexes          Indexes
##  Obs          3555    LR chi2     742.90       R2       0.300    C       0.811
##   0           2856    d.f.             4      R2(4,3555)0.188    Dxy     0.622
##   1            699    Pr(> chi2) <0.0001    R2(4,1684.7)0.355    gamma   0.626
##  max |deriv| 9e-07                           Brier    0.125    tau-a   0.197
##
##                   Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept        -7.0401 0.4267 -16.50 <0.0001
##  age               0.3813 0.0268  14.25 <0.0001
##  loghome_happy    -1.1463 0.1403  -8.17 <0.0001
##  loghome_al_drug   0.5027 0.0981   5.13 <0.0001
##  logtime_no_adult  1.1913 0.1236   9.64 <0.0001
```

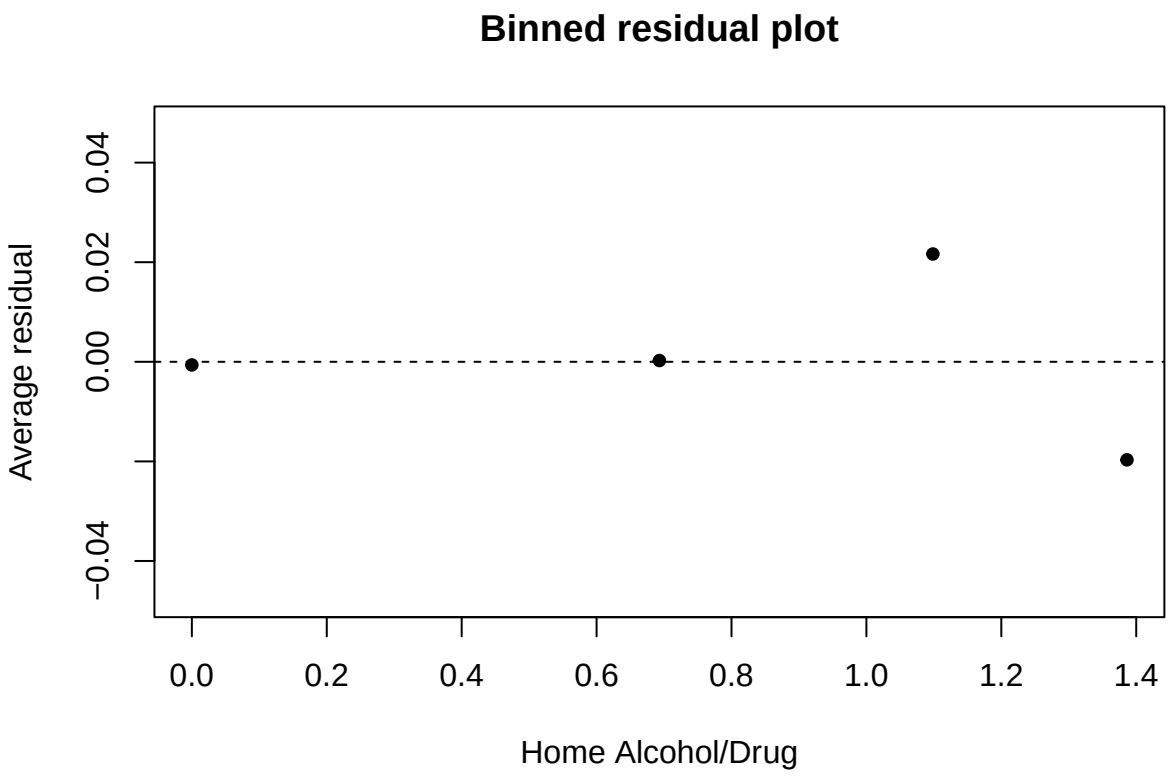
##

These values are slightly lower than they were in the original model, but they are still large enough to suggest the model has strong predictive power.

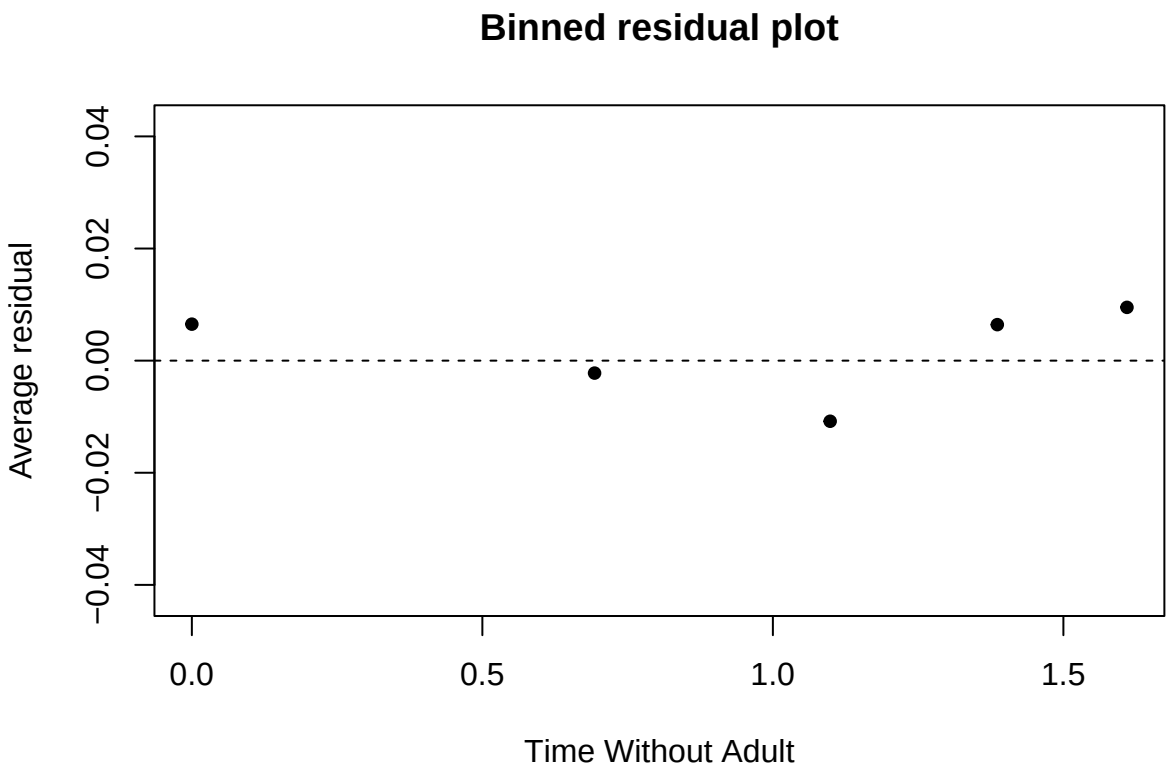
```
#Plotting the binned residuals of the new log transformed data
iys_glmaug2 <- augment(iys_glm2) |>
  mutate(.resp.resid2 = resid(iys_glm2, type = "response"))

arm::binnedplot(iys_glmaug2$loghome_happy, iys_glmaug2$.resp.resid2, xlab = "Home Happy", col.int = NULL
```



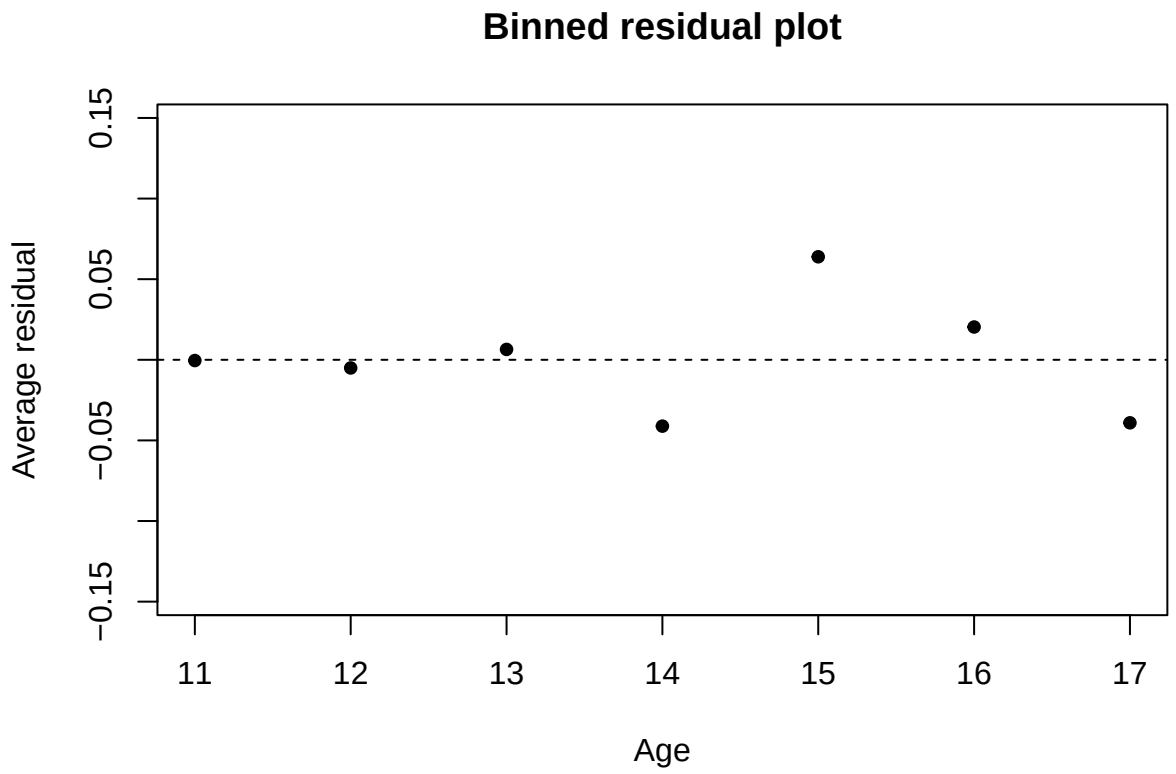

```
arm::binnedplot(iys_glmaug2$loghome_al_drug, iys_glmaug2$.resp.resid2, xlab = "Home Alcohol/Drug", col.
```




```
arm::binnedplot(iys_glmaug2$logtime_no_adult, iys_glmaug2$.resp.resid2, xlab = "Time Without Adult", co
```



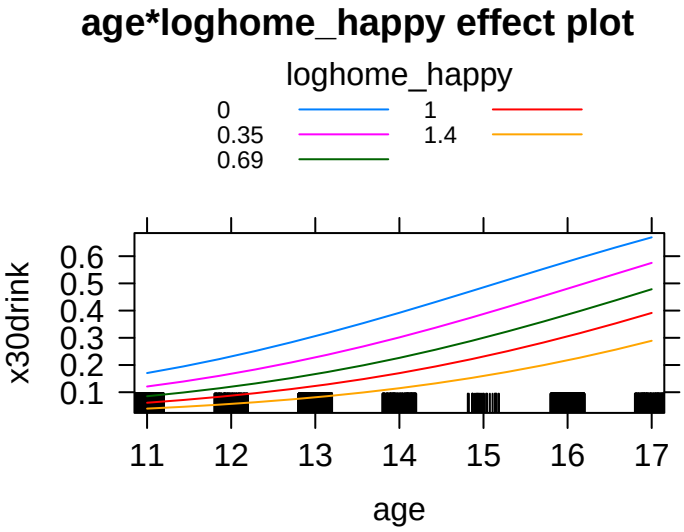

```
arm::binnedplot(iys_glmaug2$age, iys_glmaug2$.resp.resid2, xlab = "Age", col.int = NULL)
```



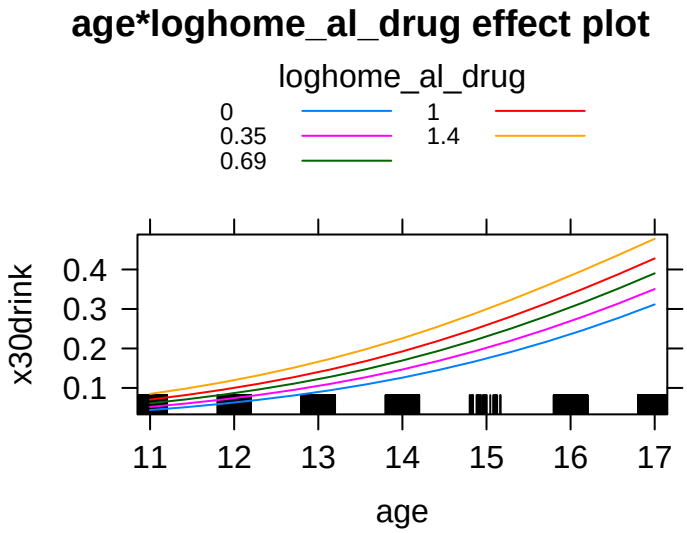
The home_happy binned residual plot still seems to suggest some issues of non linearity, but the other plots look much better*. Because the issue of linearity does not seem fixed for the home_happy variable, I noted it in my report and suggested that the emphasis be put on the conclusions found by analyzing time_no_adult and home_al_drug.

*They still have a slight curvature, but is pretty good, especially considering this is an observational study conducted in an uncontrolled environment.

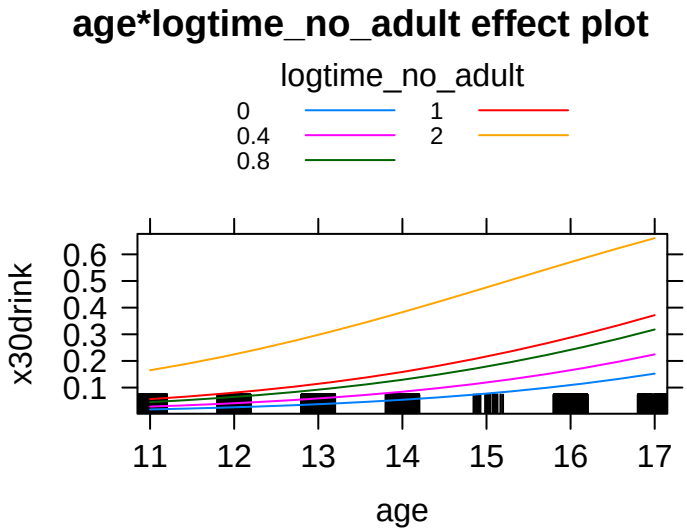
```
#Visualising the new data through effects plots
iys_eff <- Effect(c("age", "loghome_happy"), iys_glm2)
plot(iys_eff, multiline = TRUE, type = "response")
```




```
iys_eff1 <- Effect(c("age", "loghome_al_drug"), iys_glm2)
plot(iys_eff1, multiline = TRUE, type = "response")
```




```
iys_eff2 <- Effect(c("age", "logtime_no_adult"), iys_glm2)
plot(iys_eff2, multiline = TRUE, type = "response")
```



These new effect plots still seem to show an association between alcohol abuse and the three variables of interest, after accounting for each other and age.

6(Testing An Association)

```
#Constructing a confidence interval for all of the explanatory variables
confint.default(iys_glm2, level = .95)
```

```
##                       2.5 %     97.5 %
## (Intercept)      -7.8764365 -6.2037541
## age               0.3288131  0.4337278
## loghome_happy    -1.4213306 -0.8713484
## loghome_al_drug   0.3104644  0.6949515
## logtime_no_adult  0.9490281  1.4335657
```

Home_happy Using Wald's test, z: -8.170 p-value: <.001 –> significant

CI:(-1.421, -.087)

Home_al_drug Using Wald's test, z: 5.125 p-value: <.001 -> significant

CI: (.31, .69)

Time_no_adult Using Wald's test, z: 9.638 p-value: <.001 -> significant

CI: (.949, 1.43)

*Alcohol abuse is associated with all of the explanatory variables of interest after accounting for each other and age.