Canadian Community Health Survey Analysis

Zafer Rabin Saba

28.10.2021

Section 1 - Data Cleaning

```
library(tidyverse)
cchs <- read_csv("cchs2015.csv") #2015/2016 Canadian Community Health Survey
                                 #Public Use Microdata File
```

```
cchs.clean <- cchs %>%
  filter(dhhgage==4 |dhhgage==5 | dhhgage==6 | dhhgage==7) %>%
  mutate(
    age=case_when(
         dhhgage == 4 ~ 22.5,
         dhhgage == 5 ~ 27.5,
         dhhgage == 6 ~ 32.5,
         dhhgage == 7 ~ 37.5),
    female=ifelse(dhh_sex==2,1,0),
    married=ifelse(dhhgms==1|dhhgms==2,1,0),
    hhsize=dhhdghsz,
    alcohol.weekly=alwdvwky
        )%>%
  select(alcohol.weekly,age,female,married,hhsize)
```

```
summary(cchs.clean)
```

```
##  alcohol.weekly         age            female         married
##  Min.   :  0.000   Min.   :22.50   Min.   :0.000   Min.   :0.0000
##  1st Qu.:  0.000   1st Qu.:27.50   1st Qu.:0.000   1st Qu.:0.0000
##  Median :  2.000   Median :32.50   Median :1.000   Median :1.0000
##  Mean   :  4.457   Mean   :30.78   Mean   :0.561   Mean   :0.5214
##  3rd Qu.:  6.000   3rd Qu.:37.50   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :100.000   Max.   :37.50   Max.   :1.000   Max.   :1.0000
##  NA's   :386                                       NA's   :5
##      hhsize
##  Min.   :1.00
##  1st Qu.:2.00
##  Median :3.00
##  Mean   :2.87
##  3rd Qu.:4.00
##  Max.   :5.00
##  NA's   :3
```

Survey respondents in my sample: - Consume 4.457 drinks per week on average - Have an average age of 30.78 - 56.1% of them are female - 52.14% of them are married - The average household size is 2.87

Section 2 - Analysis

To understand the potential demographic determinants of alcohol consumption, I estimate the following population model:

$$Alch_i = \beta_0 + \beta_1 age_i + \beta_2 female_i + \beta_3 married_i + \beta_4 hhsize_i + e_i$$

Where: -$Alch_i$ measures the number of drinks per week for person i, -age_i is the age of person i (of course it is not exact since we took the average of the age bracket) -$female_i$ represents an indicator where it is 1 if i is female and 0 otherwise -$married_i$ represents an indicator where it is 1 if i is married and 0 otherwise -$hhsize_i$ represents the household size of person i. However, it is maxed at 5 according to the documentation, so we consider values higher than 5 as 5 too. -e_i represents the error term

```
ols <- lm(data=cchs.clean, alcohol.weekly ~ age+female+married+hhsize)
summary(ols)
```

```
##
## Call:
## lm(formula = alcohol.weekly ~ age + female + married + hhsize,
##     data = cchs.clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.918 -3.542 -2.152  1.417 93.095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.295277   1.046338   6.972 4.19e-12 ***
## age          0.001884   0.032446   0.058  0.95369
## female      -2.950742   0.325246  -9.072  < 2e-16 ***
## married      0.003039   0.374200   0.008  0.99352
## hhsize      -0.451275   0.135599  -3.328  0.00089 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.319 on 2062 degrees of freedom
##   (391 observations deleted due to missingness)
## Multiple R-squared:  0.04703,    Adjusted R-squared:  0.04519
## F-statistic: 25.44 on 4 and 2062 DF,  p-value: < 2.2e-16
```

From the regression, the effects of age and married are not statistically significant. Therefore, interpreting their coefficients is not appropriate. There is no detection of statistically significant linear dependence of the mean of Alch with neither age nor married. Female is statistically significant. We can interpret that females have lower alcohol consumption by 2.95 drinks per week on average, when keeping other variables constant. Household size is also statistically significant. With each one increase in household size, (until 5 since that is the max) we expect the weekly alcohol consumption to decrease by 0.45 drinks on average, keeping other variables constant. Even though the affect of hhsize on alch is statistically significant, we cannot be sure about causality because linear regression alone is not enough to establish causality. For instance, we might have omitted variables. It is also possible that the affect is other way around: Maybe people who are alcoholics prefer to live alone. Overall, causality is possible but it is not certain. We need to do additional analysis for causality. I do not expect the hhsize to be uncorrelated with the population error. I believe that the size of the error would be affected by the hhsize. This is because I expect that our model would do a better job for explaining the alcohol consumption for smaller size of households, but a worse job for large households because I believe there would be more heterogenity in alcohol consumption for people living in larger households due to things we do not observe here.

I augment my model by including an additional regressor representing the interaction between age and hhsize since we may expect the effect of hhsize on alcohol consumption to vary by the level of age.

$$Alch_i = \beta_0 + \beta_1 age_i + \beta_2 female_i + \beta_3 married_i + \beta_4 hhsize_i + \beta_5 hhsize_i age_i + e_i$$

Each variable in this model is described above with the exception of the $hhsize_i * age_i$ interaction term which represents the product of household size and age for person i.

```
ols2 <- lm(data=cchs.clean, alcohol.weekly ~ age+female+married+hhsize+age*hhsize)
summary(ols2)
```

```
##
## Call:
## lm(formula = alcohol.weekly ~ age + female + married + hhsize +
##     age * hhsize, data = cchs.clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.911 -3.537 -2.155  1.427 93.097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.429405   2.356333   3.153  0.00164 **
## age         -0.002418   0.075097  -0.032  0.97432
## female      -2.951328   0.325455  -9.068  < 2e-16 ***
## married     -0.003061   0.386408  -0.008  0.99368
## hhsize      -0.498156   0.750262  -0.664  0.50678
## age:hhsize   0.001534   0.024146   0.064  0.94935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.321 on 2061 degrees of freedom
##   (391 observations deleted due to missingness)
## Multiple R-squared:  0.04704,    Adjusted R-squared:  0.04472
## F-statistic: 20.35 on 5 and 2061 DF,  p-value: < 2.2e-16
```

The interaction effect is not statistically significant, so we can interpret that as: We have no evidence that the effect of hhsize on alcohol consumption depends on the level of age. The inclusion of the intersection term is not justified in our model. Therefore, it does not make sense to interpret the coefficient in this case.

However, if it was statistically significant, we could interpret that as:

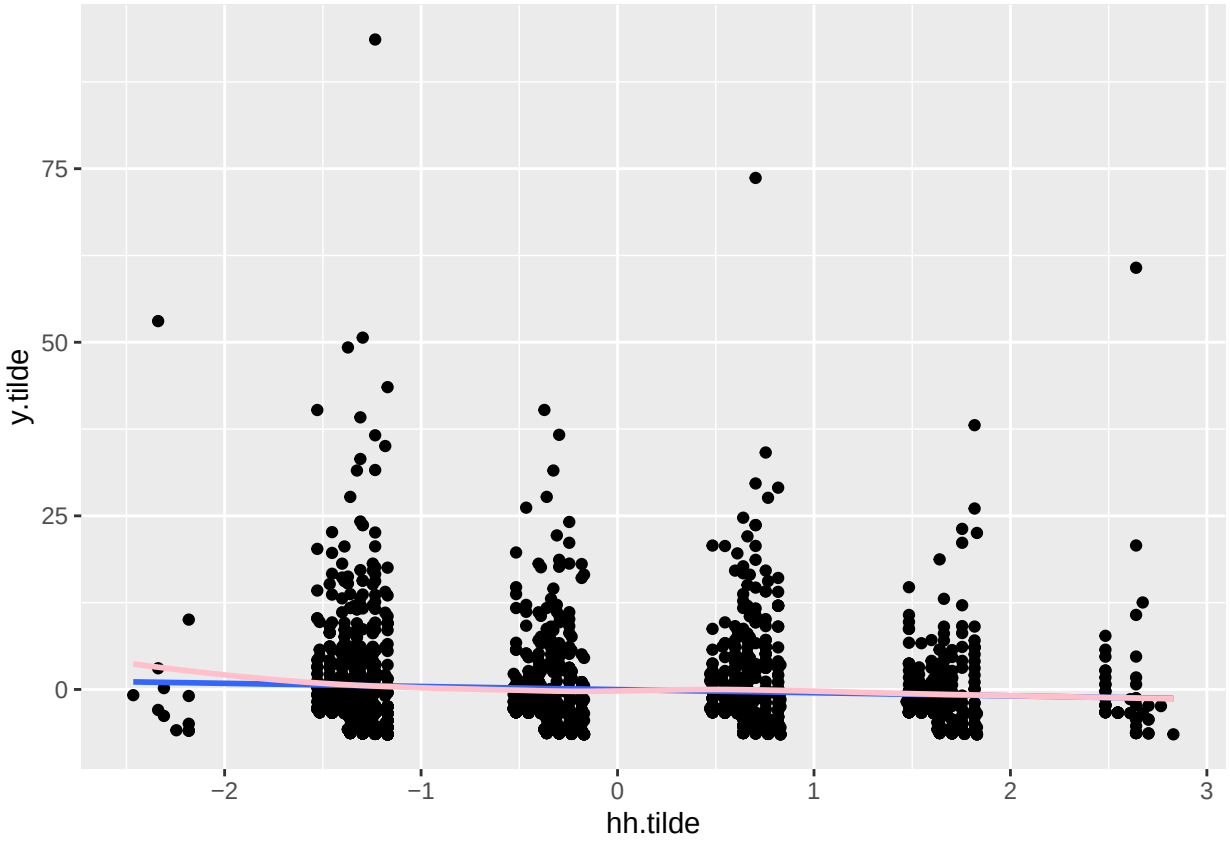
The effect of one unit increase in hhsize on alch is expected to increase by 0.00153 for every one year increase in age, holding other things constant.

```
library(modelr)
library(ggplot2)

ols3 <- lm(data=cchs.clean, alcohol.weekly ~ age+female+married)
cchs.clean2 <-  add_residuals(cchs.clean,ols3, var="y.tilde")

ols4 <- lm(data=cchs.clean, hhsize ~ age+female+married)
cchs.clean2 <-  add_residuals(cchs.clean2,ols4, var="hh.tilde")
```

```
ggplot(data = cchs.clean2) +
  geom_point(aes(y=y.tilde, x=hh.tilde)) +
  geom_smooth(method=lm, se=FALSE, aes(y=y.tilde, x=hh.tilde))+
  geom_smooth(method="loess", se=FALSE, aes(y=y.tilde, x=hh.tilde),color="pink")
```



From the graph, it seems that the smooth line(pink) and the linear line(blue) are nearly on top of each other, so the relationship does not appear to be non-linear.