finalproject

2025-06-17

LOaD packages

```
require(haven)
```

```
## Lade nötiges Paket: haven
```

```
require(foreign)  # for loading spss data
```

```
## Lade nötiges Paket: foreign
```

```
require(sandwich)  # for robust standard errors
```

```
## Lade nötiges Paket: sandwich
```

```
## Warning: Paket 'sandwich' wurde unter R Version 4.4.3 erstellt
```

```
require(lme4)      # for multilevel analysis library(lmerTest)  # for posttests
```

```
## Lade nötiges Paket: lme4
```

```
## Warning: Paket 'lme4' wurde unter R Version 4.4.3 erstellt
```

```
## Lade nötiges Paket: Matrix
```

```
require(lmtest)    # for robust standard errors and likelihood ratio tests
```

```
## Lade nötiges Paket: lmtest
```

```
## Warning: Paket 'lmtest' wurde unter R Version 4.4.3 erstellt
```

```
## Lade nötiges Paket: zoo
```

```
## Warning: Paket 'zoo' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'zoo'
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(dplyr)
```

```
##
## Attache Paket: 'dplyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych) # for descriptives
```

```
## Warning: Paket 'psych' wurde unter R Version 4.4.3 erstellt
```

```
library(quest) # for descriptives of multi-level data
```

```
## Warning: Paket 'quest' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'quest'
```

```
## Das folgende Objekt ist maskiert 'package:psych':
##
##     winsor
```

```
## Das folgende Objekt ist maskiert 'package:dplyr':
##
##     changes
```

```
## Das folgende Objekt ist maskiert 'package:stats':
##
##     decompose
```

```
library(mlmhelpr) # for multi-level applications
```

```
## Warning: Paket 'mlmhelpr' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'mlmhelpr'
```

```
## Das folgende Objekt ist maskiert 'package:quest':
##
##     center
```

```
## Das folgende Objekt ist maskiert 'package:psych':
##
##     reliability
```

```
library(ggplot2) # for plotting
```

```
##
## Attache Paket: 'ggplot2'
```

```
## Die folgenden Objekte sind maskiert von 'package:psych':
##
##     %+%, alpha
```

```
require(stargazer) # for making nice tables
```

```
## Lade nötiges Paket: stargazer
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(lmerTest) # for conducting post estimation test on ml models
```

```
## Warning: Paket 'lmerTest' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'lmerTest'
```

```
## Das folgende Objekt ist maskiert 'package:lme4':
##
##     lmer
```

```
## Das folgende Objekt ist maskiert 'package:stats':
##
##     step
```

```
library(performance) # for R2 calculation
```

```
## Warning: Paket 'performance' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'performance'
```

```
## Das folgende Objekt ist maskiert 'package:mlmhelpr':
##
##     icc
```

```
library(lattice) # for plotting
library(plm) # for fixed effect models
```

```
## Warning: Paket 'plm' wurde unter R Version 4.4.3 erstellt
```

```
##
## Attache Paket: 'plm'
```

```
## Die folgenden Objekte sind maskiert von 'package:dplyr':
##
##     between, lag, lead
```

```
library(tidyr) # tranforming data from wide to long and vice versa
```

```
##
## Attache Paket: 'tidyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:Matrix':
##
##     expand, pack, unpack
```

```
library(broom.mixed) # for tidy() on mixed models
```

```
## Warning: Paket 'broom.mixed' wurde unter R Version 4.4.3 erstellt
```

```
library(sjPlot)
```

```
## Warning: Paket 'sjPlot' wurde unter R Version 4.4.3 erstellt
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
pequiv <- readRDS("C:/Users/noaha/AppData/Local/Temp/fe870aea-e78e-46ec-9dd1-8757510678f3_SOEP-CORE.v39eu_R_EN.zip.8f3/R_EN/soepdata/pequiv.rds")
```

Create Subset with relevant variables

```
#pequiv$happiness <- pequiv$p11101
#pequiv$workhrs <- pequiv$e11101
#pequiv$employmentstatus <- pequiv$e11102
#pequiv$netincome <- pequiv$i11102
#pequiv$race <- pequiv$d11112ll
#pequiv$was_in_hospital <- pequiv$m11101
#pequiv$age <-pequiv$d11101
#d11109 years of education
#occupation e11105_v1
#m11126 Current Self-Rated Health Status
#x11101ll   Person Identification Number
#pid    Never Changing Person ID
#syear  Survey Year
#d11101 Age of Individual
#d11102ll   Gender of Individual
#cid    Original HH Number
#hid    Current Wave HH Number
#d11106 Number of Persons in HH
#d11107 Number of Children in HH
#d11104 Marital Status of Individual
#l11101 State of Residence
#e11103 employment level

vars <- c(
  # Original core variables
  "d11109",        # Years of education
  "m11126",        # Self-rated health
  "pid",           # Person ID (stable)
  "syear",         # Year
  "d11101",        # Age
  "d11102ll",      # Gender
  "cid",           # Original household ID
  "hid",           # Current wave household ID
  "d11106",        # Household size
  "d11107",        # Number of children
  "d11104",        # Marital status
  "e11103",        #employment level
  "i11110",        #individual labor earnings
  "p11101",        # Happiness
  "e11102",        # Employment status
  "i11102",        # HH Net income
  "m11101"         # Was in hospital
)

#create subset
pequiv_subset <- pequiv[, vars]

#rename variables
pequiv_subset <- pequiv_subset %>%
  rename(
    education = d11109,
```

```
  health = m11126,
  pid = pid,
  year = syear,
  age = d11101,
  gender = d11102ll,
  hh_orig = cid,
  hh_wave = hid,
  hh_size = d11106,
  hh_kids = d11107,
  marital_status = d11104,
  happiness = p11101,
  employmentstatus = e11102,
  hhnetincome = i11102,
  was_in_hospital = m11101,
  employmentlvl = e11103,        #employment level
 indiv_wage =i11110       #individual labor earnings
)
```

Inspect Dataset

```
str(pequiv_subset)
```

```
## tibble [1,148,926 × 17] (S3: tbl_df/tbl/data.frame)
##  $ education       : dbl+lbl [1:1148926] 15.0,  9.0, 12.0, 10.5, -2.0, -2.0,  9.0,  9.0,
-1...
##    ..@ label       : chr "Number of Years of Education"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ health          : dbl+lbl [1:1148926] -2, -2, -2, -2, -2, -2, -2, -2, -2, -2, -2, -2,
-2...
##    ..@ label       : chr "Current Self-Rated Health Status"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:13] -8 -7 -6 -5 -4 -3 -2 -1 1 2 ...
##    .. ..- attr(*, "names")= chr [1:13] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ pid             : dbl+lbl [1:1148926]  101,  102,  103,  201,  202,  203,  301,  302,
3...
##    ..@ label       : chr "Never Changing Person ID"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ year            : dbl+lbl [1:1148926] 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1
9...
##    ..@ label       : chr "Survey Year"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ age             : dbl+lbl [1:1148926] 54, 44, 21, 58, 28, 24, 24, 23,  0, 64, 73, 30, 2
6...
##    ..@ label       : chr "Age of Individual"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ gender          : dbl+lbl [1:1148926] 1, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 1, 2,
2,...
##    ..@ label       : chr "Gender of Individual"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:10] -8 -7 -6 -5 -4 -3 -2 -1 1 2
##    .. ..- attr(*, "names")= chr [1:10] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ hh_orig         : dbl+lbl [1:1148926]  19,  19,  19,  27,  27,  27,  35,  35,  35,  43,
...
```

```
##    ..@ label       : chr "Original HH Number"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ hh_wave         : dbl+lbl [1:1148926]  19,  19,  19,  27,  27,  27,  35,  35,  35,  43,
...
##    ..@ label       : chr "Current Wave HH Number"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ hh_size         : dbl+lbl [1:1148926] 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 2, 2, 1, 1, 1,
1,...
##    ..@ label       : chr "Number of Persons in HH"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ hh_kids         : dbl+lbl [1:1148926] 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,
0,...
##    ..@ label       : chr "Number of Children in HH"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ marital_status  : dbl+lbl [1:1148926]  1,  1,  2,  4, -1, -1,  2,  2, -1,  3,  3,  2,
2...
##    ..@ label       : chr "Marital Status of Individual"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:15] -8 -7 -6 -5 -4 -3 -2 -1 1 2 ...
##    .. ..- attr(*, "names")= chr [1:15] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ employmentlvl   : dbl+lbl [1:1148926]  1,  3,  1,  3,  3,  3,  1,  1, -1,  1,  3,  2,
2...
##    ..@ label       : chr "Employment Level of Individual"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:11] -8 -7 -6 -5 -4 -3 -2 -1 1 2 ...
##    .. ..- attr(*, "names")= chr [1:11] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ indiv_wage      : dbl+lbl [1:1148926]  4481,     0,  6187,     0,  5245,     0, 12169,
1...
##    ..@ label       : chr "Individual Labor Earnings"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
```

```
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ happiness       : dbl+lbl [1:1148926]  8,  8,  8, 10, -2, -2, 10, 10, -2,  0,  6, 10,
4...
##    ..@ label       : chr "Overall life satisfaction"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:10] -8 -7 -6 -5 -4 -3 -2 -1 0 10
##    .. ..- attr(*, "names")= chr [1:10] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ employmentstatus: dbl+lbl [1:1148926]  1,  0,  1,  0,  0,  0,  1,  1, -1,  1,  0,  1,
1...
##    ..@ label       : chr "Employment Status of Individual"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:10] -8 -7 -6 -5 -4 -3 -2 -1 0 1
##    .. ..- attr(*, "names")= chr [1:10] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
##  $ hhnetincome     : dbl+lbl [1:1148926]  9286,  9286,  9286, 15988, 15988, 15988, 18029,
1...
##    ..@ label       : chr "HH Post-Government Income"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:8] -8 -7 -6 -5 -4 -3 -2 -1
##    .. ..- attr(*, "names")= chr [1:8] "[-8] Question this year not part of survey" "[-7] O
nly available in less restricted edition" "[-6] Version of questionnaire with modified filter
ing" "[-5] Not included in this version of the questionnaire" ...
##  $ was_in_hospital : dbl+lbl [1:1148926]  0,  0,  0,  0, -2, -2,  0,  1, -2,  1,  1,  0,
0...
##    ..@ label       : chr "Overnight hosp stay"
##    ..@ format.stata: chr "%54.0g"
##    ..@ labels      : Named num [1:10] -8 -7 -6 -5 -4 -3 -2 -1 0 1
##    .. ..- attr(*, "names")= chr [1:10] "[-8] Question this year not part of survey" "[-7]
Only available in less restricted edition" "[-6] Version of questionnaire with modified filte
ring" "[-5] Not included in this version of the questionnaire" ...
```

```
summary(pequiv_subset$happiness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -5.00   -2.00    6.00    4.24    8.00   10.00
```

```
table(pequiv_subset$happiness)
```

```
##
##     -5     -2     -1      0      1      2      3      4      5      6      7
##   3773 359319   2482   3534   3178   8805  18190  24293  84065  78875 163327
##      8      9     10
## 241930 104585  52570
```

```
summary(pequiv_subset$employmentstatus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.000   0.000   0.000   0.187   1.000   1.000
```

```
table(pequiv_subset$gender, useNA = "ifany")
```

```
##
##     -1      1      2
##    846 565398 582682
```

```
table(pequiv_subset$occupation, useNA = "ifany")
```

```
## Warning: Unknown or uninitialised column: `occupation`.
```

```
## < table of extent 0 >
```

```
attr(pequiv_subset$gender, "labels")
```

```
##             [-8] Question this year not part of survey
##                                                     -8
##         [-7] Only available in less restricted edition
##                                                     -7
##  [-6] Version of questionnaire with modified filtering
##                                                     -6
## [-5] Not included in this version of the questionnaire
##                                                     -5
##                     [-4] Inadmissable multiple response
##                                                     -4
##                                  [-3] Implausible value
##                                                     -3
##                                     [-2] Does not apply
##                                                     -2
##                                          [-1] No answer
##                                                     -1
##                                                [1] Male
##                                                      1
##                                              [2] Female
##                                                      2
```

```
attr(pequiv_subset$marital_status, "labels")
```

```
##            [-8] Question this year not part of survey
##                                                    -8
##        [-7] Only available in less restricted edition
##                                                    -7
##  [-6] Version of questionnaire with modified filtering
##                                                    -6
## [-5] Not included in this version of the questionnaire
##                                                    -5
##                   [-4] Inadmissable multiple response
##                                                    -4
##                                [-3] Implausible value
##                                                    -3
##                                   [-2] Does not apply
##                                                    -2
##                                        [-1] No answer
##                                                    -1
##                                           [1] Married
##                                                     1
##                                            [2] Single
##                                                     2
##                                           [3] Widowed
##                                                     3
##                                          [4] Divorced
##                                                     4
##                                        [5] Separated
##                                                    5
##                           [6] Over 18 and NotW,Partnr6
##                                                    6
##                          [7] Under 18 And NotW,Partnr7
##                                                    7
```

Data handling

Prepare data for analysis

```
#all non-answers recoded to NA
pequiv_subset[] <- lapply(pequiv_subset, function(x) {
  if (is.numeric(x) || is.integer(x)) {
    x[x %in% -1:-8] <- NA
  }
  return(x)
})


#check
table(pequiv_subset$gender, useNA = "ifany")
```

```
##
##      1      2   <NA>
## 565398 582682    846
```

```
#recode gender to new female variable with 0 male and 1 for female
pequiv_subset <- pequiv_subset %>%
  mutate(female = case_when(
    gender == 1 ~ 0,  # Male
    gender == 2 ~ 1,  # Female
    TRUE ~ NA_real_
  ))
table(pequiv_subset$female, useNA = "ifany")
```

```
##
##      0      1   <NA>
## 565398 582682    846
```

```
pequiv_subset <- pequiv_subset %>%
  mutate(
    married = ifelse(marital_status == 1, 1, 0)
  )

#more preprocessing
#only core working age group
pequiv_subset <- pequiv_subset %>%
  filter(age >= 30, age <= 49)



#only beyond 2002 (less attrition more consistent survey questions)
pequiv_subset <- pequiv_subset %>%
  filter(year >= 2002, year <= 2020)
table(pequiv_subset$year, useNA = "ifany")
```

```
##
##  2002  2003  2004  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014
##  9972  9236  8818  8249  8550  7804  7202  7499 12820 13292 12492 14421 12068
##  2015  2016  2017  2018  2019  2020
## 11953 13213 14197 13016 12546 13601
```

##more pre-processing

```
#create income log
pequiv_subset$log_hhincome <- log1p(pequiv_subset$hhnetincome)




#create OECD equivalence scale for income
pequiv_subset <- pequiv_subset %>%
  mutate(eq_income = hhnetincome / sqrt(hh_size),
         log_eq_income = log1p(eq_income))  # if skewed
```

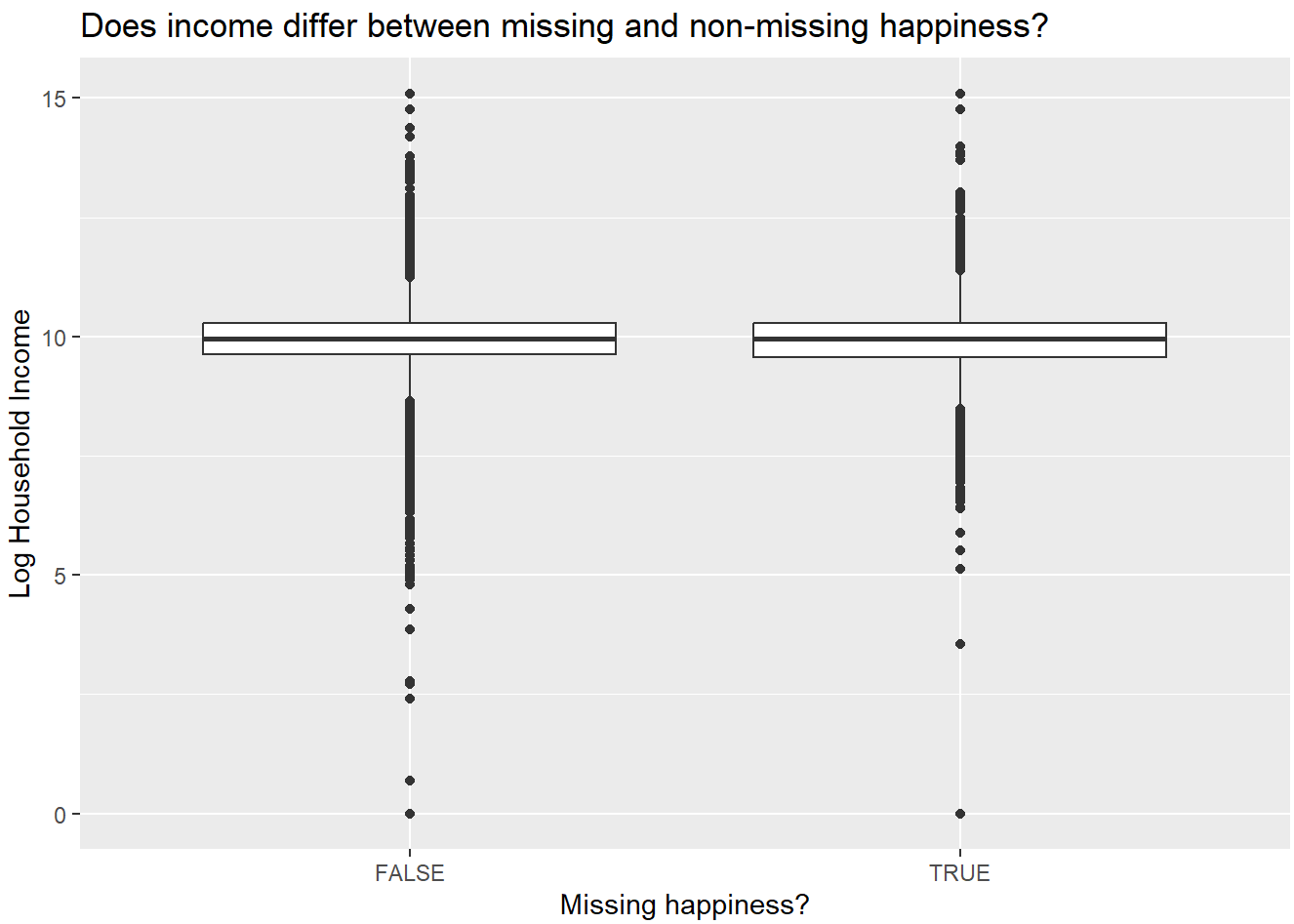
##check and handle missings

```
colSums(is.na(pequiv_subset))
```

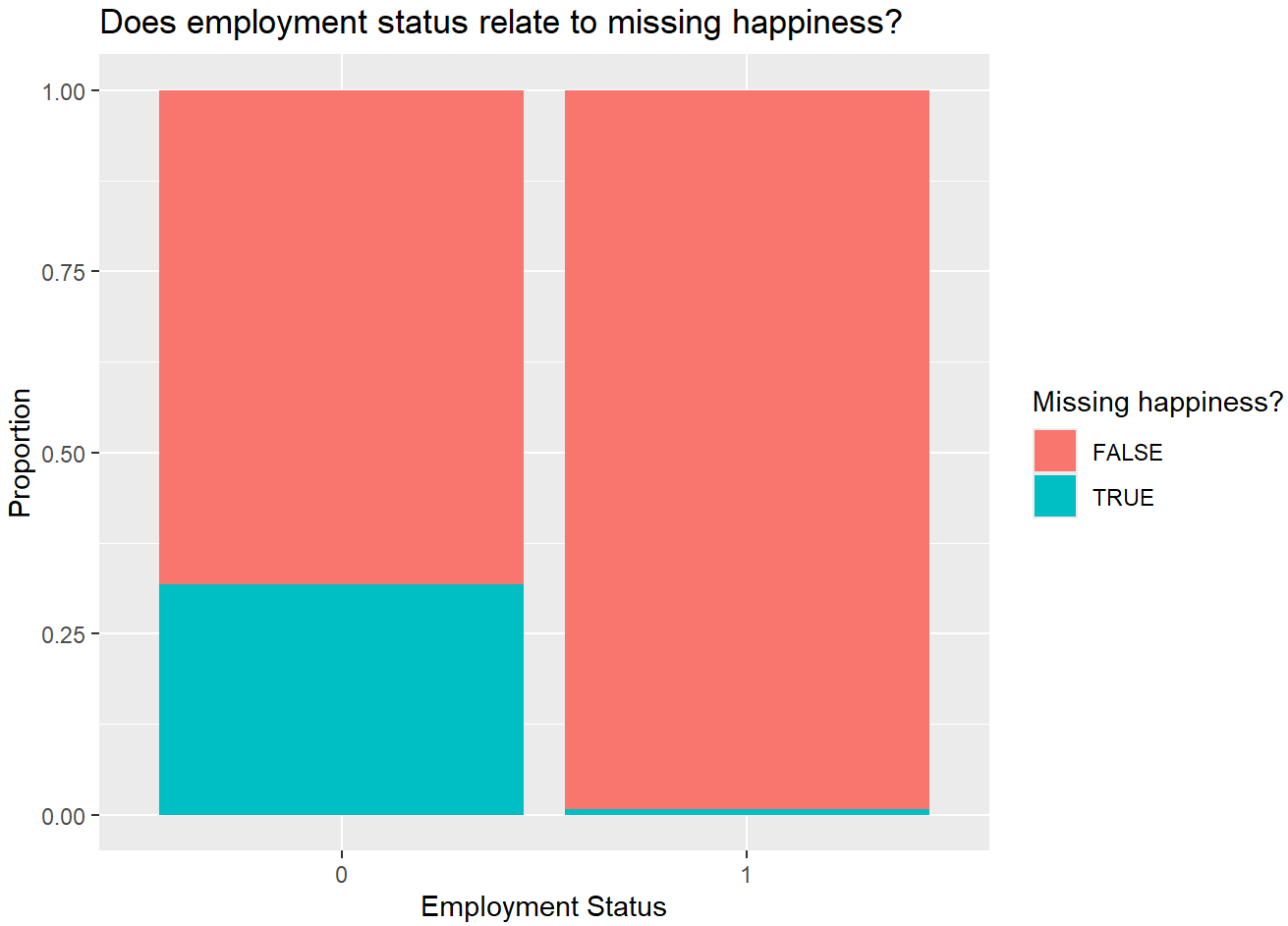
```
##        education           health              pid             year
##            21771            18117                0                0
##              age           gender          hh_orig          hh_wave
##                0                5                0                0
##          hh_size          hh_kids   marital_status    employmentlvl
##                0                0            18806                0
##       indiv_wage        happiness employmentstatus      hhnetincome
##             5721            19397                0             5721
##  was_in_hospital           female          married     log_hhincome
##            32196                5            18806             5721
##        eq_income    log_eq_income
##             5721             5721
```

```
#significant missings:  health 18,000 ; married 18806 ; happiness 19397  ; hhnetincome 5721



# Visual check: does X differ by missingness in happiness?

# Income
ggplot(pequiv_subset, aes(x = is.na(happiness), y = log_eq_income)) +
  geom_boxplot() +
  labs(x = "Missing happiness?", y = "Log Household Income",
       title = "Does income differ between missing and non-missing happiness?")
```

```
## Warning: Removed 5721 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

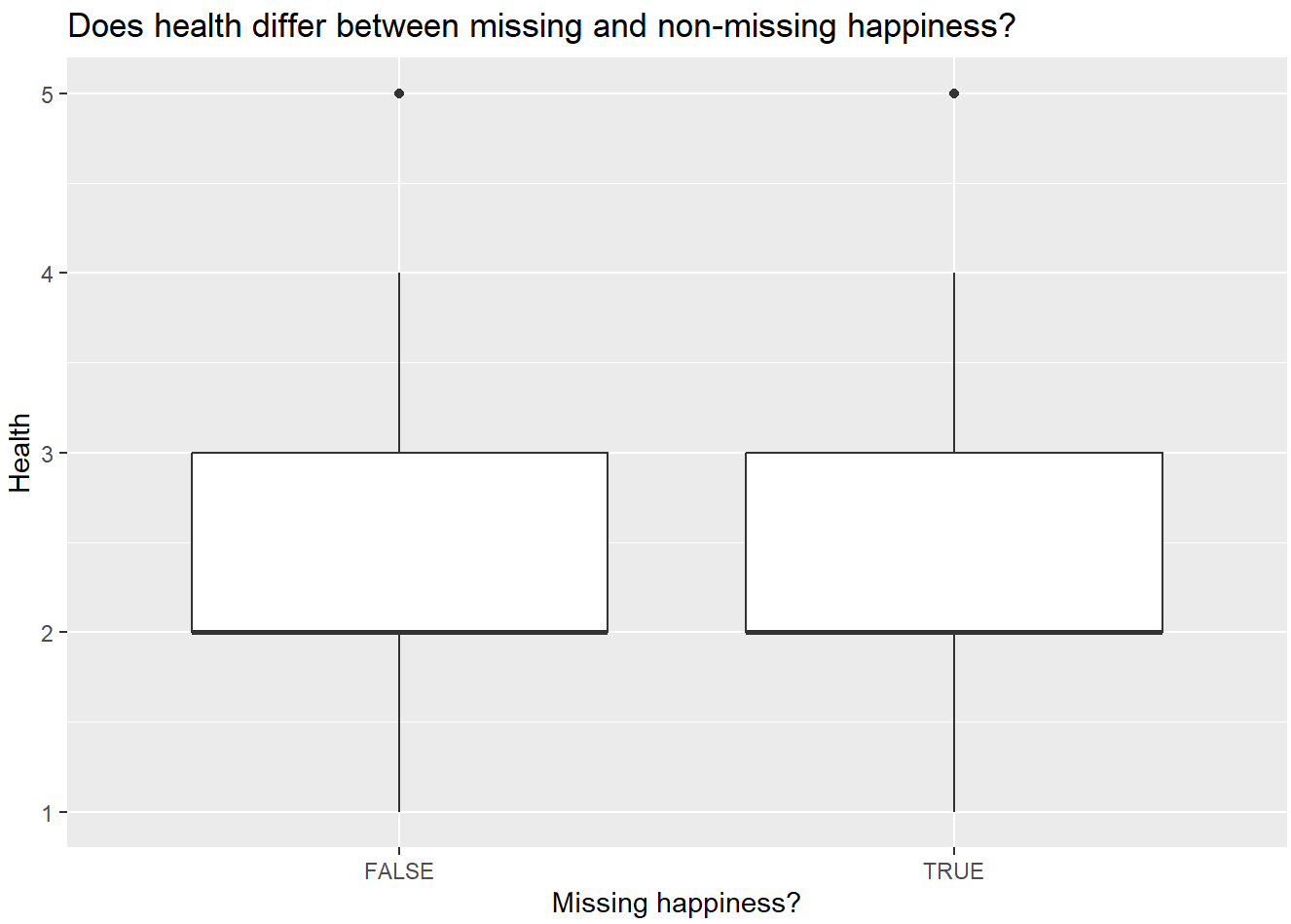



```
# Employment status
ggplot(pequiv_subset, aes(x = factor(employmentstatus), fill = is.na(happiness))) +
  geom_bar(position = "fill") +
  labs(x = "Employment Status", y = "Proportion",
       fill = "Missing happiness?",
       title = "Does employment status relate to missing happiness?")
```

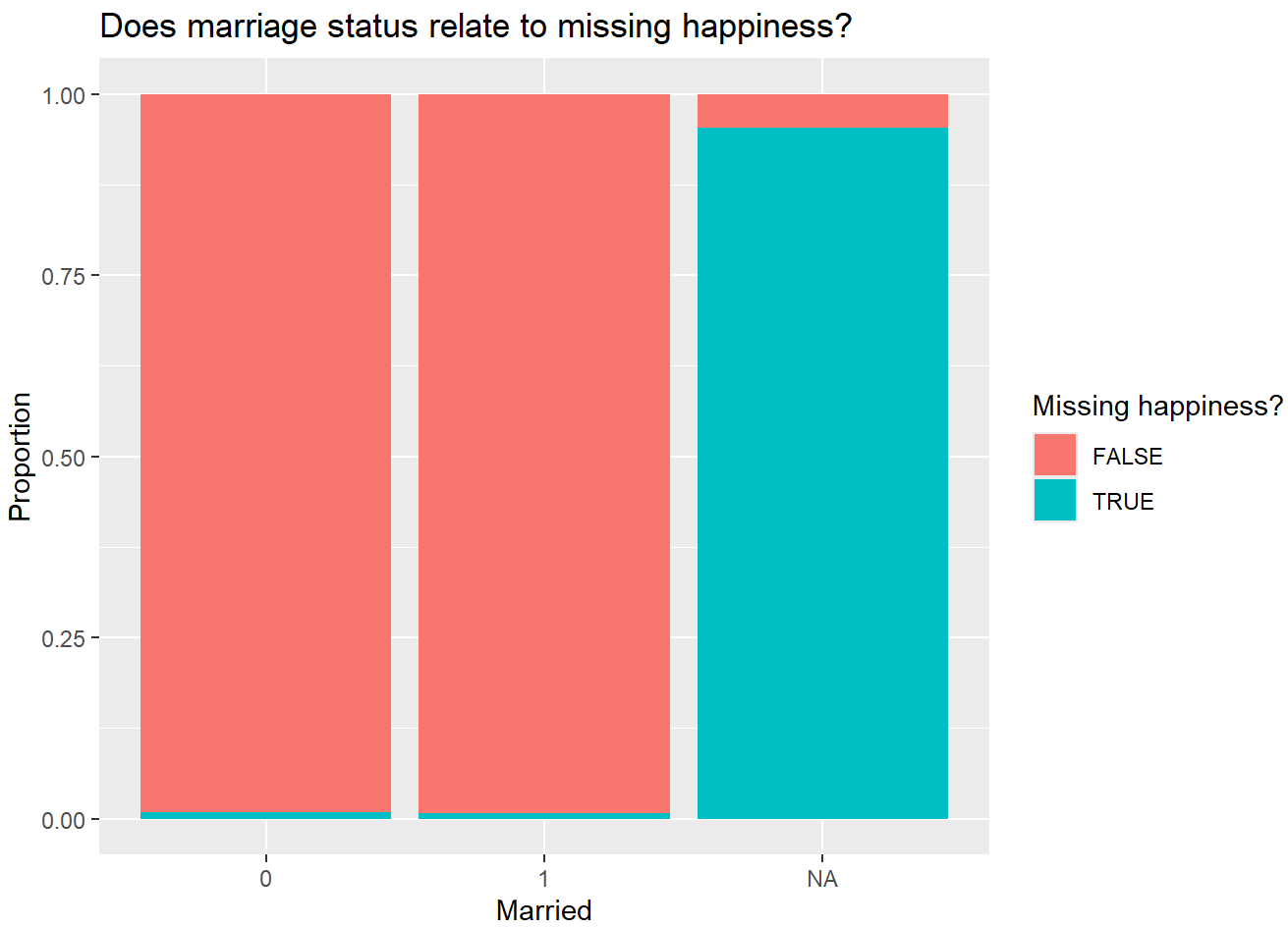



```
#health
ggplot(pequiv_subset, aes(x = is.na(happiness), y = health)) +
  geom_boxplot() +
  labs(x = "Missing happiness?", y = "Health",
       title = "Does health differ between missing and non-missing happiness?")
```

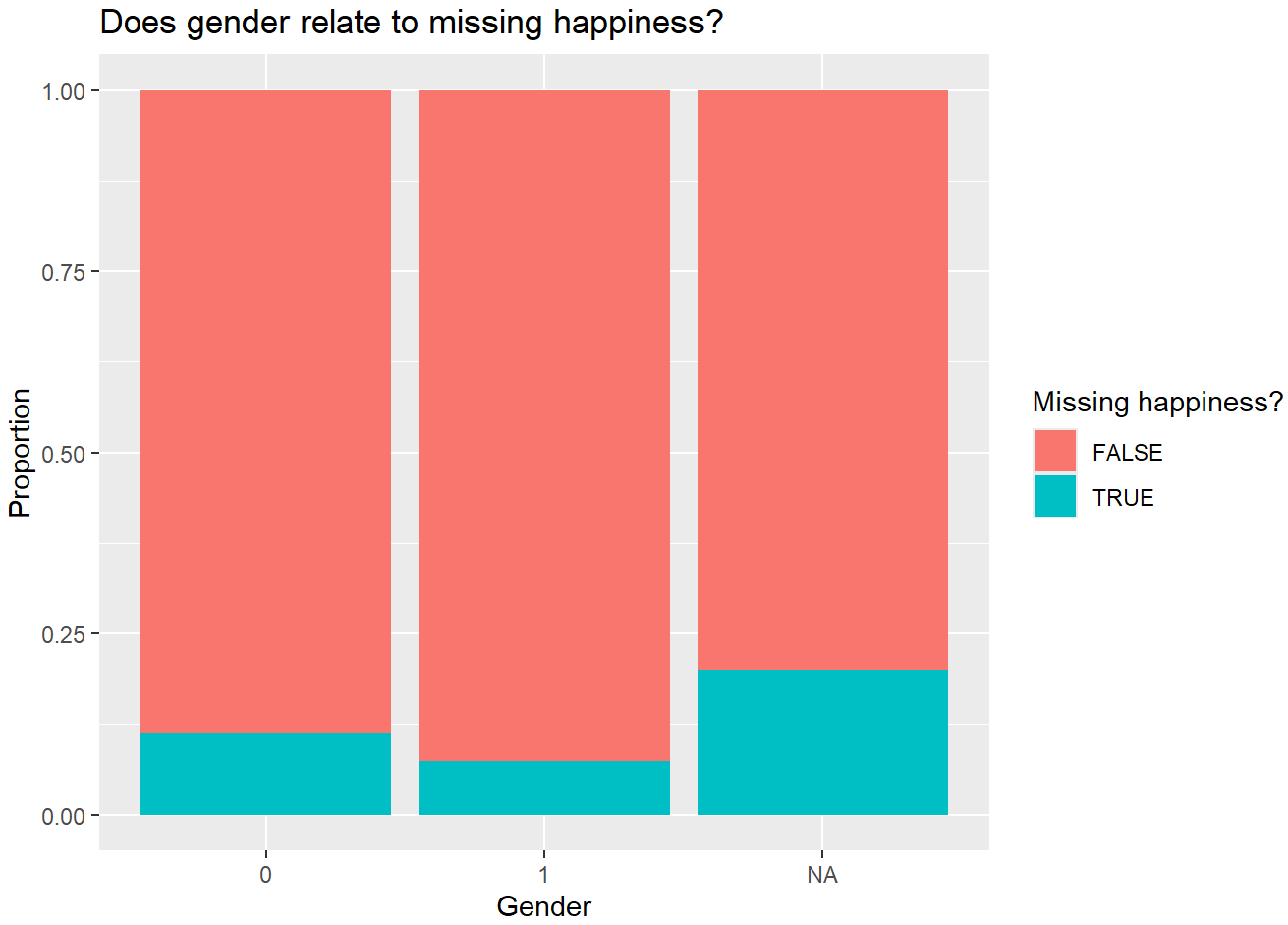
```
## Warning: Removed 18117 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
# Married
ggplot(pequiv_subset, aes(x = factor(married), fill = is.na(happiness))) +
  geom_bar(position = "fill") +
  labs(x = "Married", y = "Proportion",
       fill = "Missing happiness?",
       title = "Does marriage status relate to missing happiness?")
```




```
# Gender
ggplot(pequiv_subset, aes(x = factor(female), fill = is.na(happiness))) +
  geom_bar(position = "fill") +
  labs(x = "Gender", y = "Proportion",
       fill = "Missing happiness?",
       title = "Does gender relate to missing happiness?")
```




```
#remove incomplete observations on key variables
pequiv_subset <- pequiv_subset %>%
  filter(!is.na(health), !is.na(employmentstatus), !is.na(female), !is.na(happiness), !is.na
(hhnetincome), !is.na(married))
```

##create income brackets and check

```
#create income per hh member brackets and calc mean happiness
pequiv_subset <- pequiv_subset %>%
  filter(!is.na(happiness), !is.na(hhnetincome)) %>%
  mutate(income_bracket2 = cut(
    eq_income,
    breaks = c(0, 15000, 30000, 45000, 70000, 120000, Inf),
    labels = c("1", "2", "3", "4", "5", "6"),
    right = FALSE  # bracket includes lower bound, excludes upper
  ))
mean_happiness_by_bracket <- pequiv_subset %>%
  group_by(income_bracket2) %>%
  summarise(
    n = n(),
    mean_happiness = mean(happiness, na.rm = TRUE)
  )

print(mean_happiness_by_bracket)
```

```
## # A tibble: 6 × 3
##   income_bracket2     n mean_happiness
##   <fct>           <int>          <dbl>
## 1 1               45636           6.74
## 2 2               97899           7.30
## 3 3               30638           7.65
## 4 4                9382           7.77
## 5 5                1820           8.01
## 6 6                 581           8.10
```

##check descriptives of variables of interest

```
mean(pequiv_subset$happiness)
```

```
## [1] 7.254356
```

```
sd((pequiv_subset$happiness))
```

```
## [1] 1.718516
```

```
psych::describe(pequiv_subset$happiness)
```

```
##    vars      n mean   sd median trimmed  mad min max range  skew kurtosis se
## X1    1 185956 7.25 1.72      8    7.25 1.48   0  10    10 -1.05      1.5  0
```

```
table(pequiv_subset$employmentstatus)
```

```
##
##      0      1
##  34106 151850
```

More data preprocessing

```
#check number of ppl in dataset
length(unique(pequiv_subset$pid))
```

```
## [1] 39161
```

```
#check some correlations
cor(pequiv_subset$health, log1p(pequiv_subset$hhnetincome), use = "complete.obs")
```

```
## [1] -0.08579306
```

```
cor(pequiv_subset$health, pequiv_subset$happiness, use = "complete.obs")
```

```
## [1] -0.4245338
```

```
attr(pequiv_subset$health, "labels")
```

```
##            [-8] Question this year not part of survey
##                                                    -8
##        [-7] Only available in less restricted edition
##                                                    -7
##  [-6] Version of questionnaire with modified filtering
##                                                    -6
## [-5] Not included in this version of the questionnaire
##                                                    -5
##                   [-4] Inadmissable multiple response
##                                                    -4
##                                [-3] Implausible value
##                                                    -3
##                                   [-2] Does not apply
##                                                    -2
##                                        [-1] No answer
##                                                    -1
##                                         [1] Very good
##                                                     1
##                                              [2] Good
##                                                     2
##                                        [3] Satisfactory
##                                                     3
##                                              [4] Poor
##                                                     4
##                                               [5] Bad
##                                                     5
```

```
cor(pequiv_subset$health, pequiv_subset$employmentstatus, use = "complete.obs")
```

```
## [1] -0.07301336
```

```
cor(pequiv_subset$happiness, pequiv_subset$employmentstatus, use = "complete.obs")
```

```
## [1] 0.08450555
```

```
cor(pequiv_subset$happiness, pequiv_subset$hhnetincome, use = "complete.obs")
```

```
## [1] 0.1452998
```

```
cor(pequiv_subset$health, pequiv_subset$age, use = "complete.obs")
```

```
## [1] 0.1389251
```

```
cor(pequiv_subset$gender, pequiv_subset$hhnetincome, use = "complete.obs")
```

```
## [1] -0.004837729
```

```
cor(pequiv_subset$happiness, pequiv_subset$year, use = "complete.obs")
```

```
## [1] 0.1398737
```

```
cor(pequiv_subset$happiness, pequiv_subset$marital_status, use = "complete.obs")
```

```
## [1] -0.1523129
```

```
attr(pequiv_subset$marital_status, "labels")
```

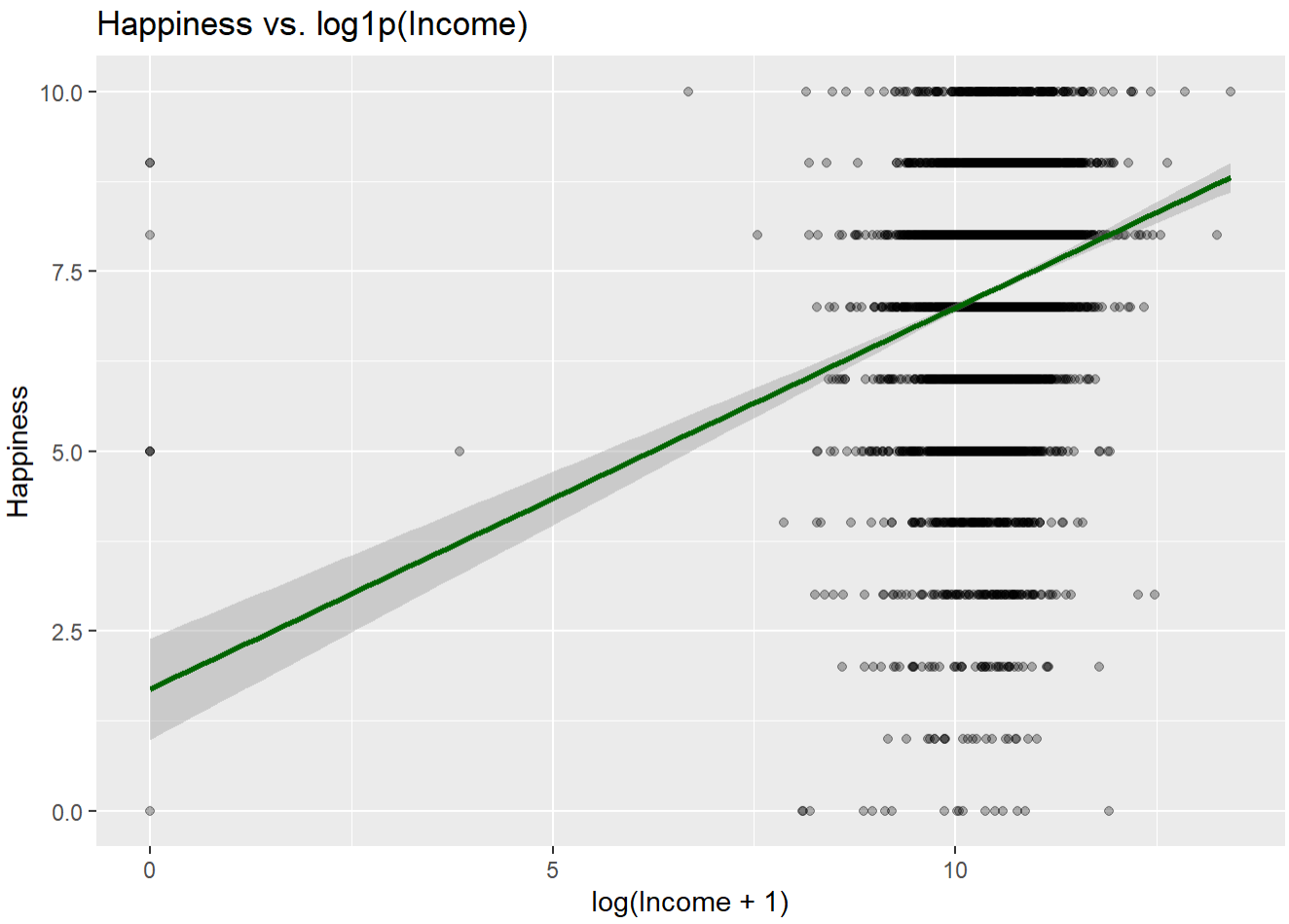
```
##             [-8] Question this year not part of survey
##                                                     -8
##         [-7] Only available in less restricted edition
##                                                     -7
##  [-6] Version of questionnaire with modified filtering
##                                                     -6
## [-5] Not included in this version of the questionnaire
##                                                     -5
##                    [-4] Inadmissable multiple response
##                                                     -4
##                                 [-3] Implausible value
##                                                     -3
##                                    [-2] Does not apply
##                                                     -2
##                                         [-1] No answer
##                                                     -1
##                                            [1] Married
##                                                      1
##                                             [2] Single
##                                                      2
##                                            [3] Widowed
##                                                      3
##                                           [4] Divorced
##                                                      4
##                                         [5] Separated
##                                                      5
##                           [6] Over 18 and NotW,Partnr6
##                                                      6
##                          [7] Under 18 And NotW,Partnr7
##                                                      7
```

```
cor(pequiv_subset$happiness, log1p(pequiv_subset$hhnetincome), use = "complete.obs")
```

```
## [1] 0.208591
```

```
#try to plot log income over happiness
set.seed(123)  # for reproducibility

plot_data <- pequiv_subset %>%
  filter(!is.na(happiness), !is.na(hhnetincome)) %>%
  sample_n(5000)  # adjust sample size as needed

ggplot(plot_data, aes(x = log1p(hhnetincome), y = happiness)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", color = "darkgreen") +
  labs(
    title = "Happiness vs. log1p(Income)",
    x = "log(Income + 1)",
    y = "Happiness"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



#center and standardize some variables

```
pequiv_subset <- pequiv_subset %>%
  mutate(
    # Center age around median
    age_c = age - median(age, na.rm = TRUE),

    # Time variable: years since 2002
    baseyear = year - 2002,

    # Standardized log household income
    log_eq_income_z = as.numeric(scale(log_eq_income)),

    # Number of children in household, centered around median
    hh_kids_c = hh_kids - median(hh_kids, na.rm = TRUE),

    # Household size centered around median
    hh_size_c = hh_size - median(hh_size, na.rm = TRUE),

    # Happiness centered around median
    happiness_c = happiness - median(happiness, na.rm = TRUE),

    # Health centered around median
    health_c = health - median(health, na.rm = TRUE)


  )
```

create health lag

```
# Create lagged health variable where the first observation per pid is retained
pequiv_subset <- pequiv_subset %>%
  arrange(pid, year) %>%
  group_by(pid) %>%
  mutate(
    health_lag = dplyr::lag(health),
    health_lag = if_else(row_number() == 1, health, health_lag)
  ) %>%
  ungroup()

# View side-by-side for a sample of individuals
pequiv_subset %>%
  filter(pid %in% sample(unique(pid), 10)) %>%
  dplyr::select(pid, year, health, health_lag) %>%
  arrange(pid, year)
```

```
## # A tibble: 58 × 4
##    pid       year      health               health_lag
##    <dbl+lbl> <dbl+lbl> <dbl+lbl>            <dbl+lbl>
##  1 2144802   2002      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  2 2144802   2003      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  3 2144802   2004      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  4 2144802   2005      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  5 2144802   2006      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  6 2144802   2007      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  7 2144802   2008      3 [[3] Satisfactory] 3 [[3] Satisfactory]
##  8 2144802   2009      2 [[2] Good]         3 [[3] Satisfactory]
##  9 2144802   2010      3 [[3] Satisfactory] 2 [[2] Good]
## 10 2144802   2011      3 [[3] Satisfactory] 3 [[3] Satisfactory]
## # ℹ 48 more rows
```

###create descriptive table of key variables

```
# key variables
desc_vars <- pequiv_subset %>%
  select(
    happiness,
    eq_income,
    log_eq_income,
    employmentstatus,
    female,
    married,
    baseyear,
    health_lag
  )

# Create descriptive statistics table
descriptives <- psych::describe(desc_vars)

# Print nicely for copy-paste
round(descriptives[, c("mean", "sd", "min", "max", "n")], 2)
```

```
##                     mean       sd min       max      n
## happiness           7.25     1.72   0      10.0 185956
## eq_income        24191.31 21171.77   0 3605147.5 185956
## log_eq_income        9.93     0.66   0      15.1 185956
## employmentstatus     0.82     0.39   0       1.0 185956
## female               0.54     0.50   0       1.0 185956
## married              0.67     0.47   0       1.0 185956
## baseyear             9.53     5.39   0      18.0 185956
## health_lag           2.38     0.90   1       5.0 185956
```

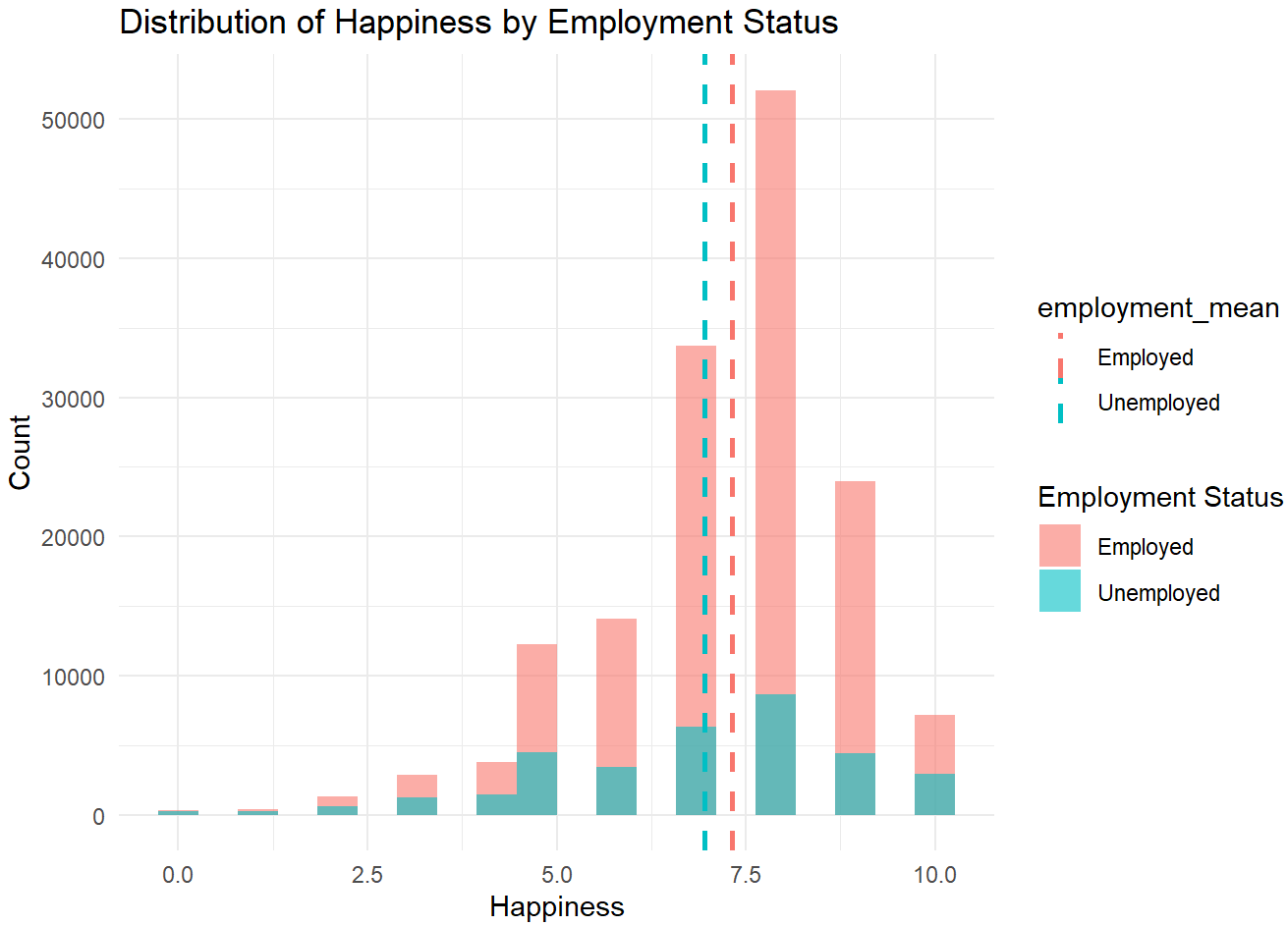
#plot happiness by employmentstatus

```
pequiv_subset %>%
  group_by(employmentstatus) %>%
  summarise(mean_happiness = mean(happiness, na.rm = TRUE),
            n = n())
```

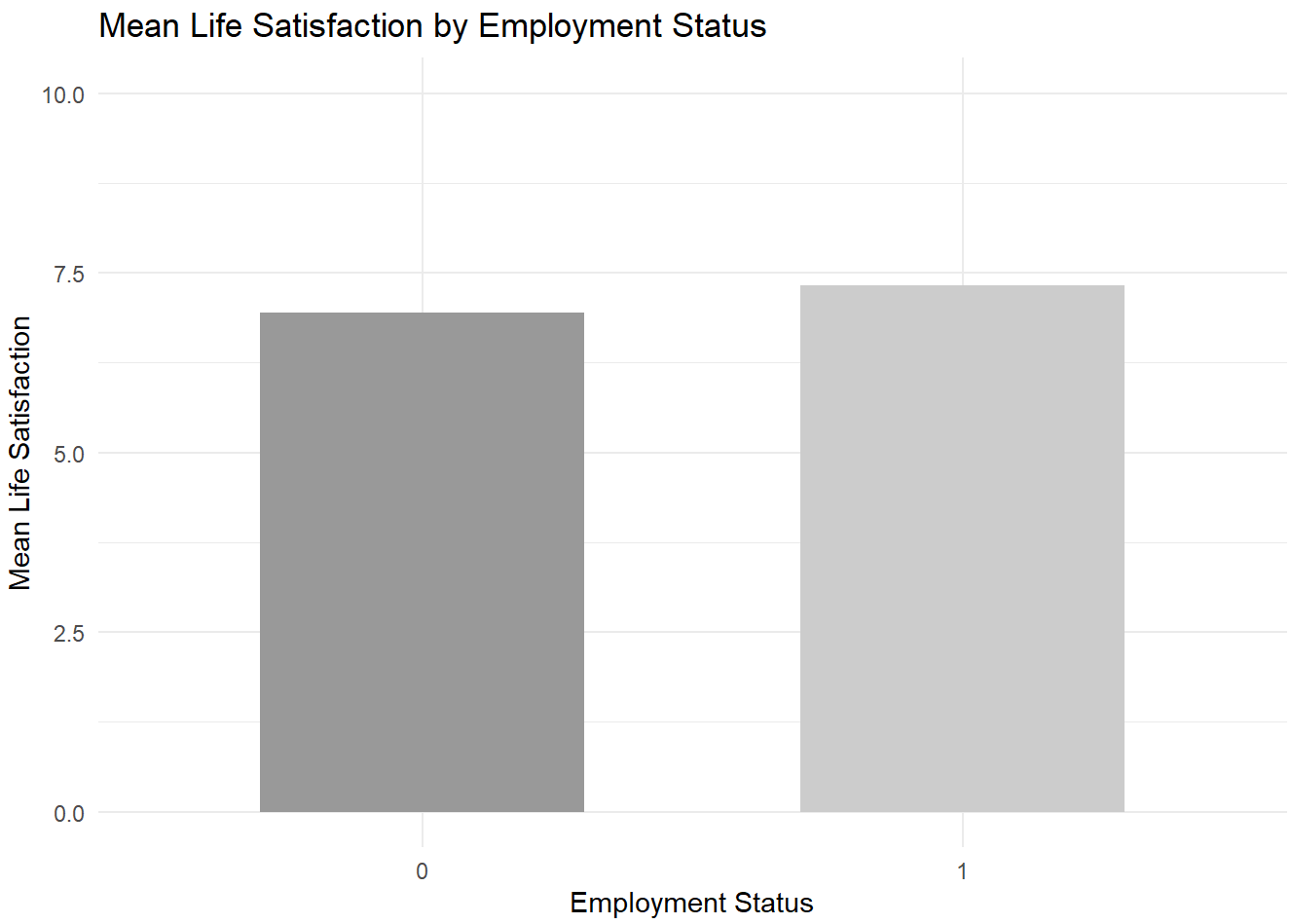
```
## # A tibble: 2 × 3
##   employmentstatus     mean_happiness      n
##   <dbl+lbl>                     <dbl>  <int>
## 1 0 [[0] Not Employed]           6.95  34106
## 2 1 [[1] Employed]               7.32 151850
```

```
happiness_summary <- pequiv_subset %>%
  filter(employmentstatus %in% c(0, 1)) %>%  # exclude NAs or other values
  group_by(employmentstatus) %>%
  summarise(avg_happiness = mean(happiness, na.rm = TRUE)) %>%
  mutate(
    employment_mean = if_else(employmentstatus == 1, "Employed", "Unemployed")
  )

pequiv_subset %>%
  filter(employmentstatus %in% c(0, 1)) %>%
  mutate(employment_mean = if_else(employmentstatus == 1, "Employed", "Unemployed")) %>%
  ggplot(aes(x = happiness, fill = employment_mean)) +
  geom_histogram(position = "identity", alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Happiness by Employment Status",
       x = "Happiness",
       y = "Count",
       fill = "Employment Status") +
  theme_minimal() +
  geom_vline(data = happiness_summary,
             aes(xintercept = avg_happiness, color = employment_mean),
             linetype = "dashed", size = 1)
```

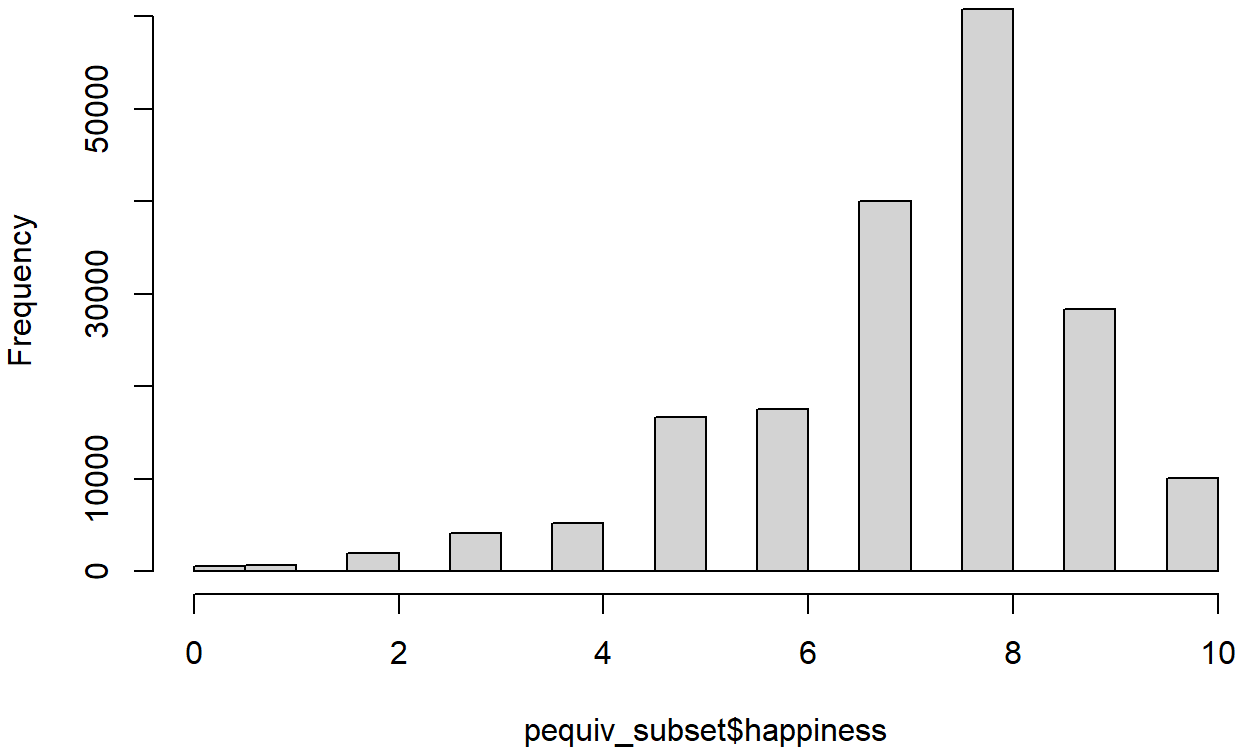
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Calculate mean happiness by employment status
mean_happiness <- pequiv_subset %>%
  group_by(employmentstatus) %>%
  summarise(mean_happiness = mean(happiness, na.rm = TRUE))
mean_happiness$employmentstatus <- as.factor(mean_happiness$employmentstatus)

# Create the bar plot
ggplot(mean_happiness, aes(x = employmentstatus, y = mean_happiness, fill = employmentstatu
s)) +
  geom_bar(stat = "identity", width = 0.6) +
  scale_fill_manual(values = c("gray60", "gray80")) +
  labs(title = "Mean Life Satisfaction by Employment Status",
       x = "Employment Status",
       y = "Mean Life Satisfaction") +
  ylim(0, 10) +
  theme_minimal() +
  theme(
    legend.position = "none",
    text = element_text(size = 11)
  )
```



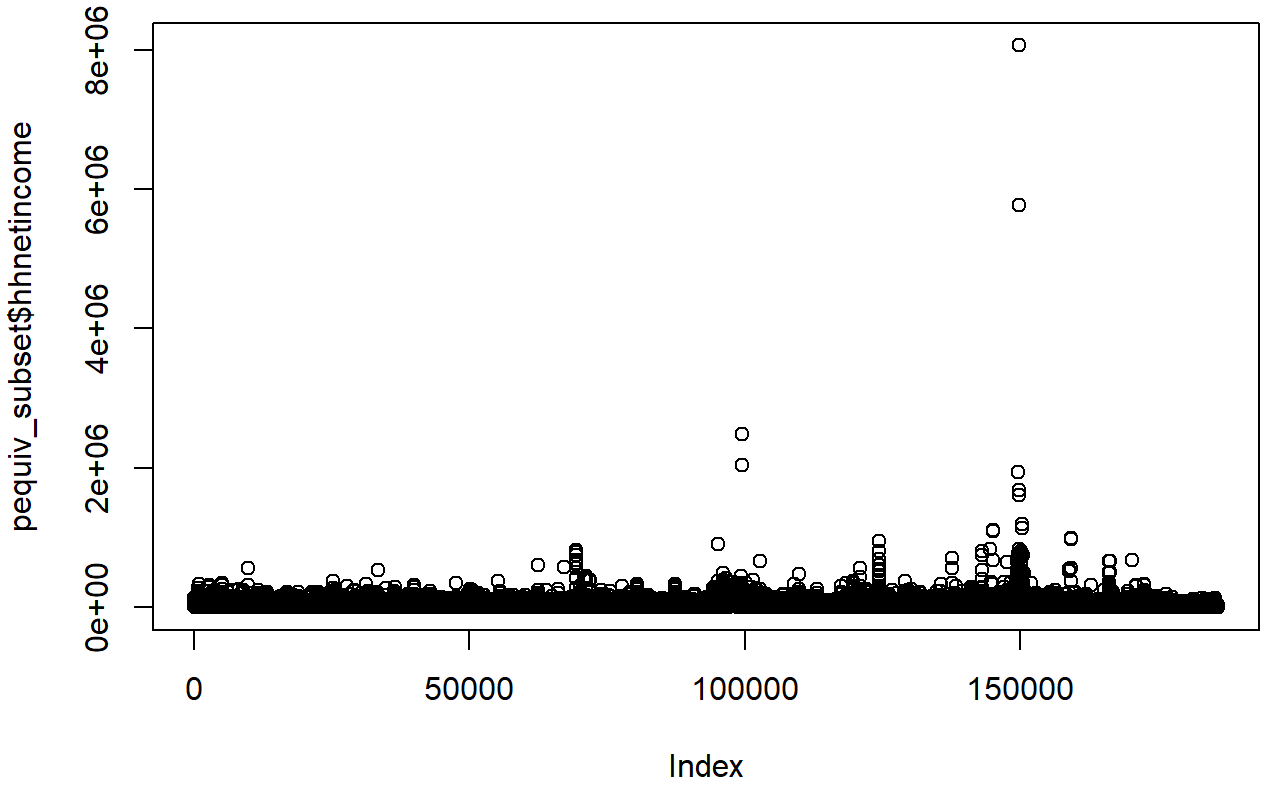

```
hist(pequiv_subset$happiness)
```



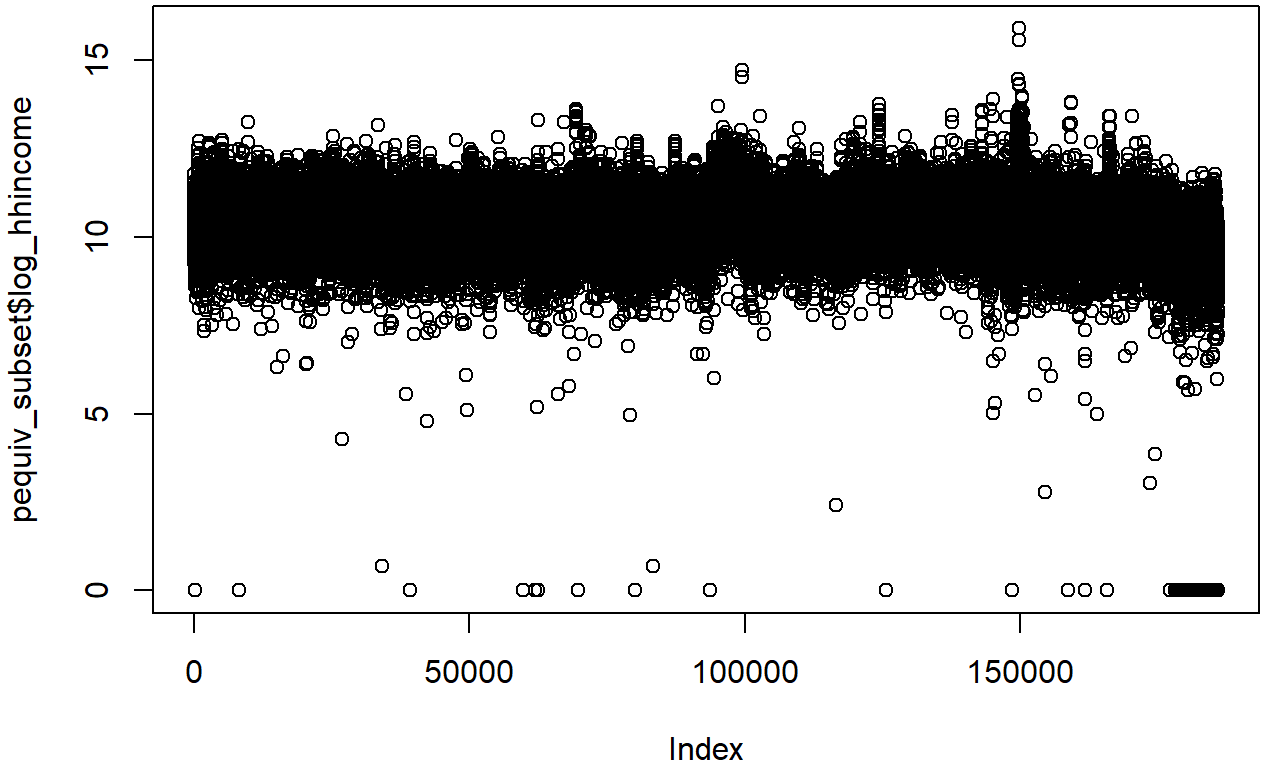

```
psych::describe(pequiv_subset$happiness)
```

```
##    vars      n mean   sd median trimmed  mad min max range  skew kurtosis se
## X1    1 185956 7.25 1.72      8    7.25 1.48   0  10    10 -1.05      1.5  0
```

```
plot(pequiv_subset$hhnetincome)
```




```
plot(pequiv_subset$log_hhincome)
```



#simple regression and robust standard errors

#simple ols modes to compare

```
ols1 <- lm(happiness ~ scale(log_eq_income) +  employmentstatus, data = pequiv_subset)
summary(ols1)
```

```
##
## Call:
## lm(formula = happiness ~ scale(log_eq_income) + employmentstatus,
##     data = pequiv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7770 -0.7495  0.4376  0.9422  7.0970
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.146421   0.009620  742.90   <2e-16 ***
## scale(log_eq_income) 0.281367   0.004163   67.59   <2e-16 ***
## employmentstatus     0.132177   0.010756   12.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.692 on 185953 degrees of freedom
## Multiple R-squared:  0.03095,    Adjusted R-squared:  0.03094
## F-statistic:  2970 on 2 and 185953 DF,  p-value: < 2.2e-16
```

```
ols2 <- lm(happiness ~ scale(log_eq_income) + female*employmentstatus, data = pequiv_subset)
summary(ols2)
```

```
##
## Call:
## lm(formula = happiness ~ scale(log_eq_income) + female * employmentstatus,
##     data = pequiv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7703 -0.7119  0.4489  0.9179  7.2119
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.895940   0.017745  388.61   <2e-16 ***
## scale(log_eq_income)      0.272374   0.004194   64.94   <2e-16 ***
## female                    0.342557   0.020397   16.80   <2e-16 ***
## employmentstatus          0.376086   0.018980   19.82   <2e-16 ***
## female:employmentstatus -0.326785   0.022173  -14.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.69 on 185951 degrees of freedom
## Multiple R-squared:  0.03244,    Adjusted R-squared:  0.03242
## F-statistic:  1558 on 4 and 185951 DF,  p-value: < 2.2e-16
```

```
ols4 <- lm(happiness ~ scale(log_eq_income) + female*employmentstatus + scale(health_lag) + s
cale(baseyear) + married , data = pequiv_subset)
summary(ols4)
```

```
##
## Call:
## lm(formula = happiness ~ scale(log_eq_income) + female * employmentstatus +
##     scale(health_lag) + scale(baseyear) + married, data = pequiv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5202 -0.7696  0.2056  1.0011  7.8305
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.497412   0.017247  376.73   <2e-16 ***
## scale(log_eq_income)      0.194370   0.003888   50.00   <2e-16 ***
## female                    0.436832   0.018753   23.29   <2e-16 ***
## employmentstatus          0.433924   0.017534   24.75   <2e-16 ***
## scale(health_lag)        -0.581575   0.003619 -160.68   <2e-16 ***
## scale(baseyear)           0.244034   0.003625   67.32   <2e-16 ***
## married                   0.471284   0.007738   60.90   <2e-16 ***
## female:employmentstatus -0.364210   0.020410  -17.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.549 on 185948 degrees of freedom
## Multiple R-squared:  0.1872, Adjusted R-squared:  0.1872
## F-statistic:  6119 on 7 and 185948 DF,  p-value: < 2.2e-16
```

```
#ols2 <- lm(happiness ~ scale(log_hhincome) + relative_year +  relative_year2 +  female  + sc
ale(health) + scale(baseyear) , data = pequiv_trans)
#summary(ols2)

psych::describe(pequiv_subset$hhnetincome)
```

```
##    vars      n     mean       sd  median  trimmed      mad min     max   range
## X1    1 185956 43260.06 39730.12 37964.5 43260.06 18674.09   0 8061355 8061355
##     skew kurtosis    se
## X1 67.56 11480.82 92.13
```

```
quantile(pequiv_subset$hhnetincome, probs = seq(0, 1, 0.1), na.rm = TRUE)
```

```
##        0%       10%       20%       30%       40%       50%       60%       70%
##       0.0   17855.0   24144.0   29041.5   33525.0   37964.5   42895.0   48871.0
##       80%       90%      100%
##   57042.0   71114.0 8061355.0
```

```
pequiv_subset <- pequiv_subset %>%
  mutate(income_bracket_num = as.integer(income_bracket2) - 1)
table(pequiv_subset$income_bracket_num)
```

```
##
##     0     1     2     3     4     5
## 45636 97899 30638  9382  1820   581
```

```
ols12 <- lm(happiness ~ scale(income_bracket_num) , data = pequiv_subset)
summary(ols12)
```

```
##
## Call:
## lm(formula = happiness ~ scale(income_bracket_num), data = pequiv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9744 -0.8597  0.3972  1.1403  3.1403
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.254356   0.003915 1852.93   <2e-16 ***
## scale(income_bracket_num) 0.320955   0.003915   81.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.688 on 185954 degrees of freedom
## Multiple R-squared:  0.03488,    Adjusted R-squared:  0.03488
## F-statistic:  6721 on 1 and 185954 DF,  p-value: < 2.2e-16
```

```
ols10 <- lm(happiness ~ scale(income_bracket_num) +  female*employmentstatus + scale(health_l
ag) + scale(baseyear) + married , data = pequiv_subset)
summary(ols10)
```

```
##
## Call:
## lm(formula = happiness ~ scale(income_bracket_num) + female *
##     employmentstatus + scale(health_lag) + scale(baseyear) +
##     married, data = pequiv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4589 -0.7763  0.2093  1.0068  5.3315
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.434279   0.016856  381.72   <2e-16 ***
## scale(income_bracket_num)  0.200653   0.003798   52.83   <2e-16 ***
## female                     0.496216   0.018623   26.64   <2e-16 ***
## employmentstatus           0.496541   0.017159   28.94   <2e-16 ***
## scale(health_lag)         -0.575280   0.003624 -158.73   <2e-16 ***
## scale(baseyear)            0.236392   0.003629   65.14   <2e-16 ***
## married                    0.477168   0.007712   61.87   <2e-16 ***
## female:employmentstatus   -0.422868   0.020289  -20.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.548 on 185948 degrees of freedom
## Multiple R-squared:  0.1885, Adjusted R-squared:  0.1885
## F-statistic:  6170 on 7 and 185948 DF,  p-value: < 2.2e-16
```

##create table with ols models and clustered se

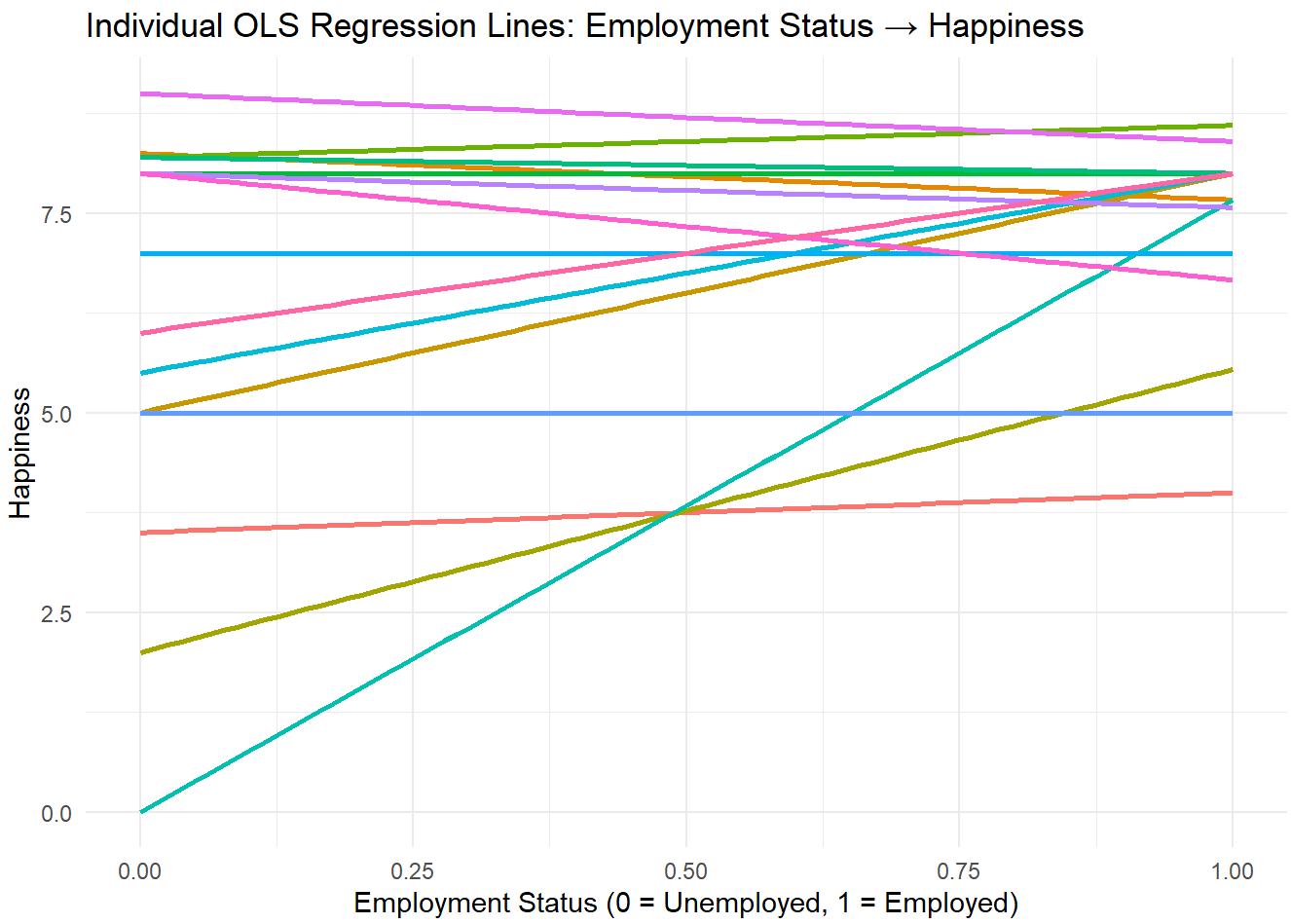
```
library(lmtest)
library(sandwich)
library(stargazer)

# Clustered standard errors by pid
cluster_se <- function(model, cluster_var) {
  vcovCL(model, cluster = cluster_var, type = "HC1")
}

# Compute clustered SEs
cl_ols1 <- cluster_se(ols1, ~pequiv_subset$pid)
cl_ols2 <- cluster_se(ols2, ~pequiv_subset$pid)
cl_ols4 <- cluster_se(ols4, ~pequiv_subset$pid)

# Stargazer with clustered SEs
stargazer(ols1, ols2, ols4,
          se = list(sqrt(diag(cl_ols1)), sqrt(diag(cl_ols2)), sqrt(diag(cl_ols4))),
          type = "text",  # or "text/ latex" / "html"
          title = "OLS Models Predicting Happiness with Clustered SEs",
          column.labels = c("OLS1", "OLS2", "OLS4"),
          dep.var.labels = "Happiness",
          no.space = TRUE,
          digits = 3,
          keep.stat = c("n", "rsq", "adj.rsq"))
```

```
##
## OLS Models Predicting Happiness with Clustered SEs
## ================================================
##                              Dependent variable:
##                          ---------------------------
##                                    Happiness
##                            OLS1      OLS2      OLS4
##                             (1)       (2)       (3)
## ------------------------------------------------
## scale(log_eq_income)    0.281***  0.272***  0.194***
##                          (0.009)   (0.009)   (0.007)
## female                             0.343***  0.437***
##                                    (0.041)   (0.033)
## employmentstatus        0.132***  0.376***  0.434***
##                          (0.020)   (0.036)   (0.029)
## scale(health_lag)                            -0.582***
##                                               (0.006)
## scale(baseyear)                              0.244***
##                                               (0.006)
## married                                      0.471***
##                                               (0.014)
## female:employmentstatus            -0.327*** -0.364***
##                                    (0.042)   (0.034)
## Constant                7.146***  6.896***  6.497***
##                          (0.018)   (0.035)   (0.030)
## ------------------------------------------------
## Observations             185,956   185,956   185,956
## R2                        0.031     0.032     0.187
## Adjusted R2               0.031     0.032     0.187
## ================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

##plot some random slopes

```
library(dplyr)
library(lme4)

# Step 1: Sample 70 individuals
set.seed(123)
sampled_pids <- sample(unique(pequiv_subset$pid), 70)

# Step 2: Subset and clean data
d <- pequiv_subset %>%
  filter(pid %in% sampled_pids) %>%
  filter(!is.na(employmentstatus) & !is.na(happiness)) %>%
  mutate(employmentstatus = as.numeric(employmentstatus))

##  plot
ggplot(d, aes(x = employmentstatus, y = happiness, group = pid)) +
  geom_smooth(method = "lm", se = FALSE, aes(color = factor(pid))) +
  labs(x = "Employment Status (0 = Unemployed, 1 = Employed)",
       y = "Happiness",
       color = "Individual (pid)",
       title = "Individual OLS Regression Lines: Employment Status → Happiness") +
  theme_minimal() +
  theme(legend.position = "none")
```

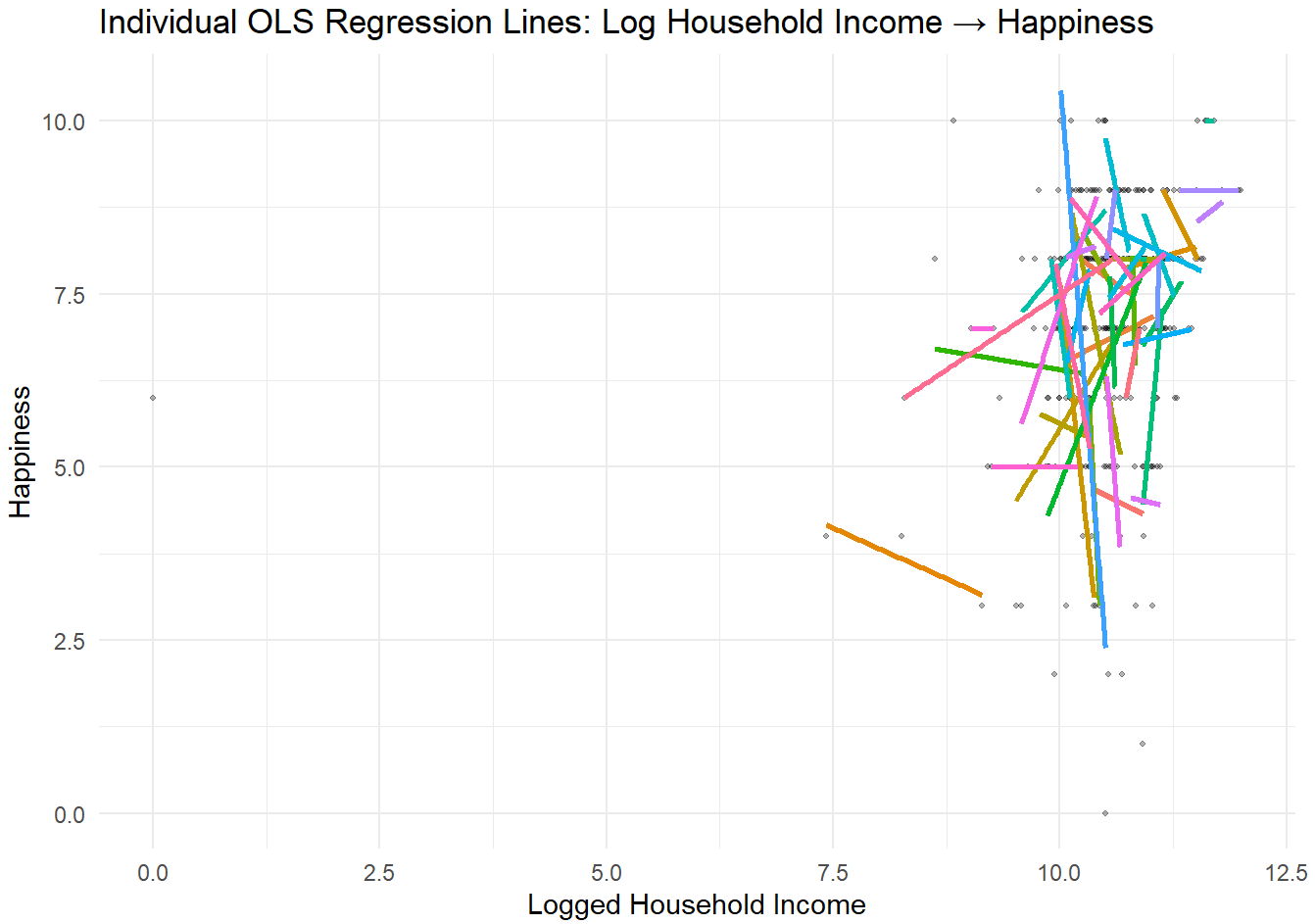
```
## `geom_smooth()` using formula = 'y ~ x'
```



#plot individual regression lines to see whether the slopes vary across participants

```
# Step 3: Plot individual OLS regression lines
ggplot(d, aes(x = log_hhincome, y = happiness, group = pid)) +
  geom_point(alpha = 0.3, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(color = factor(pid)), alpha = 0.5) +
  labs(x = "Logged Household Income",
       y = "Happiness",
       color = "Individual (pid)",
       title = "Individual OLS Regression Lines: Log Household Income → Happiness") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



##1 MUlti-level model: NUll model, Level 1, Level 2, random slopes, cross level interaction

```
#null model
mnull <- lmer(happiness ~ 1 + (1 | pid), data = pequiv_subset, REML = F)
summary(mnull)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: happiness ~ 1 + (1 | pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  657247.3  657277.7 -328620.7  657241.3    185953
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.3431 -0.4251  0.1052  0.5370  6.0823
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 1.613    1.270
##  Residual             1.426    1.194
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 7.300e+00  7.380e-03 3.685e+04   989.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#compute ICC
icc_null <- icc(mnull)
icc_null
```

```
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.531
##   Unadjusted ICC: 0.531
```

```
#about 53 percent of the variance is at the pid level

#add key level 1 predictors
ml1a <- lmer(happiness ~  scale(log_eq_income)
                   + employmentstatus + (1 | pid),
                 data = pequiv_subset, , REML = F)
summary(ml1a)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: happiness ~ scale(log_eq_income) + employmentstatus + (1 | pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  656046.2  656096.9 -328018.1  656036.2    185951
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.3315 -0.4248  0.0937  0.5347  6.1076
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 1.547    1.244
##  Residual             1.427    1.194
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                       Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)          7.221e+00  1.109e-02 1.066e+05  651.27   <2e-16 ***
## scale(log_eq_income) 1.313e-01  4.450e-03 1.598e+05   29.50   <2e-16 ***
## employmentstatus     1.197e-01  1.079e-02 1.849e+05   11.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) sc(__)
## scl(lg_q_n)  0.222
## emplymntstt -0.755 -0.243
```

```
#Add lvl1 predictors
ml1 <- lmer(happiness ~ scale(health_lag) + scale(log_eq_income) +  scale(baseyear) + married
                   + employmentstatus + (1 | pid),
                 data = pequiv_subset, , REML = F)
summary(ml1)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula:
## happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + (1 | pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  648883.6  648964.6 -324433.8  648867.6    185948
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.4424 -0.4360  0.0884  0.5451  5.8905
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 1.134    1.065
##  Residual             1.441    1.200
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           6.941e+00  1.286e-02  1.124e+05  539.60   <2e-16 ***
## scale(health_lag)    -3.159e-01  3.949e-03  1.775e+05  -79.98   <2e-16 ***
## scale(log_eq_income)  1.161e-01  4.314e-03  1.456e+05   26.91   <2e-16 ***
## scale(baseyear)       1.076e-01  4.275e-03  1.427e+05   25.16   <2e-16 ***
## married               3.737e-01  1.058e-02  1.068e+05   35.33   <2e-16 ***
## employmentstatus      1.282e-01  1.055e-02  1.808e+05   12.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(_) sc(__) scl(b) marrid
## scl(hlth_l) -0.013
## scl(lg_q_n)  0.255  0.008
## scale(bsyr) -0.042 -0.044 -0.084
## married     -0.581  0.017 -0.104 -0.010
## emplymntstt -0.669  0.024 -0.261  0.026  0.054
```

```
#add lvl2 predictors
ml2 <- lmer(happiness ~ scale(health_lag) + scale(log_eq_income) +  scale(baseyear) + married
                  + employmentstatus + female + (1 | pid),
                data = pequiv_subset, REML = F)
summary(ml2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula:
## happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 | pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  648810.0  648901.2 -324396.0  648792.0    185947
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.4476 -0.4356  0.0885  0.5453  5.8917
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 1.130    1.063
##  Residual             1.441    1.201
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           6.872e+00  1.507e-02  9.095e+04  455.93   <2e-16 ***
## scale(health_lag)    -3.175e-01  3.950e-03  1.775e+05  -80.37   <2e-16 ***
## scale(log_eq_income)  1.149e-01  4.314e-03  1.458e+05   26.63   <2e-16 ***
## scale(baseyear)       1.082e-01  4.273e-03  1.424e+05   25.33   <2e-16 ***
## married               3.762e-01  1.057e-02  1.066e+05   35.58   <2e-16 ***
## employmentstatus      1.373e-01  1.059e-02  1.818e+05   12.96   <2e-16 ***
## female                1.131e-01  1.300e-02  3.306e+04    8.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(_) sc(__) scl(b) marrid emplym
## scl(hlth_l)  0.008
## scl(lg_q_n)  0.236  0.009
## scale(bsyr) -0.042 -0.045 -0.084
## married     -0.508  0.016 -0.105 -0.009
## emplymntstt -0.619  0.020 -0.263  0.027  0.057
## female      -0.522 -0.036 -0.035  0.011  0.024  0.098
```

```
#add random slopes
mrs <- lmer(happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) + married
                  +employmentstatus + female + (1 +employmentstatus | pid),
                data = pequiv_subset, REML = F)
summary(mrs)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula:
## happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 + employmentstatus |      pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  647563.3  647674.8 -323770.6  647541.3    185945
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.5224 -0.4320  0.0901  0.5422  6.0131
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  pid      (Intercept)      1.7324   1.3162
##           employmentstatus 0.5458   0.7388   -0.66
##  Residual                  1.4070   1.1862
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           6.908e+00  1.638e-02  3.567e+04 421.888  < 2e-16 ***
## scale(health_lag)    -3.127e-01  3.960e-03  1.781e+05 -78.975  < 2e-16 ***
## scale(log_eq_income)  1.216e-01  4.454e-03  1.141e+05  27.309  < 2e-16 ***
## scale(baseyear)       1.111e-01  4.265e-03  1.366e+05  26.055  < 2e-16 ***
## married               3.632e-01  1.049e-02  1.013e+05  34.632  < 2e-16 ***
## employmentstatus      1.192e-01  1.259e-02  1.314e+04   9.474  < 2e-16 ***
## female                9.551e-02  1.276e-02  3.180e+04   7.486  7.3e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(_) sc(__) scl(b) marrid emplym
## scl(hlth_l)  0.013
## scl(lg_q_n)  0.225  0.013
## scale(bsyr) -0.047 -0.042 -0.101
## married     -0.467  0.012 -0.103 -0.006
## emplymntstt -0.706  0.011 -0.246  0.044  0.049
## female      -0.467 -0.036 -0.026  0.004  0.030  0.074
```

```
#add cross-level interactions
m_interaction_rs <- lmer(happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyea
r) + married
                              + employmentstatus * female +
                              (1 + employmentstatus | pid),
                            data = pequiv_subset,REML = F)

#add cross-level interaction 2
m_interaction2_rs <- lmer(happiness ~ scale(health_lag) + scale(log_eq_income)*female + scale
(baseyear) + married
                              + employmentstatus * female +
                              (1 + employmentstatus | pid),
                            data = pequiv_subset,REML = F)
summary(m_interaction2_rs)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: happiness ~ scale(health_lag) + scale(log_eq_income) * female +
##     scale(baseyear) + married + employmentstatus * female + (1 +
##     employmentstatus | pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  647481.6  647613.3 -323727.8  647455.6    185943
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.5154 -0.4322  0.0900  0.5417  6.2660
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  pid      (Intercept)      1.7128   1.3088
##           employmentstatus 0.5363   0.7323   -0.65
##  Residual                  1.4074   1.1863
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)                  6.778e+00  2.255e-02  2.173e+04 300.606  < 2e-16
## scale(health_lag)           -3.128e-01  3.959e-03  1.780e+05 -79.015  < 2e-16
## scale(log_eq_income)         9.674e-02  5.930e-03  1.027e+05  16.315  < 2e-16
## female                       2.928e-01  2.629e-02  1.694e+04  11.136  < 2e-16
## scale(baseyear)              1.125e-01  4.274e-03  1.366e+05  26.334  < 2e-16
## married                      3.573e-01  1.051e-02  1.019e+05  33.981  < 2e-16
## employmentstatus             2.698e-01  2.177e-02  1.731e+04  12.390  < 2e-16
## scale(log_eq_income):female  5.035e-02  8.925e-03  1.238e+05   5.641 1.69e-08
## female:employmentstatus     -2.273e-01  2.665e-02  1.526e+04  -8.531  < 2e-16
##
## (Intercept)                 ***
## scale(health_lag)           ***
## scale(log_eq_income)        ***
## female                      ***
## scale(baseyear)             ***
## married                     ***
## employmentstatus            ***
## scale(log_eq_income):female ***
## female:employmentstatus     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(_) sc(__) female scl(b) marrid emplym s(__):
## scl(hlth_l)  0.009
## scl(lg_q_n)  0.256  0.000
## female      -0.766 -0.016 -0.211
## scale(bsyr) -0.073 -0.042 -0.059  0.049
```

```
## married     -0.312  0.011 -0.025 -0.022 -0.005
## emplymntstt -0.857  0.008 -0.276  0.735  0.071 -0.004
## scl(lg_q_): -0.135  0.016 -0.658  0.222 -0.031 -0.076  0.179
## fml:mplymnt  0.688 -0.002  0.224 -0.875 -0.054  0.041 -0.817 -0.245
```

```
###this imodel is not in the paper since changes in effects are negligeble, and the additiona
l explained variance so small that it would only make the table less readable


#check out final model
summary(m_interaction_rs)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula:
## happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus * female + (1 + employmentstatus |      pid)
##    Data: pequiv_subset
##
##       AIC       BIC    logLik  deviance  df.resid
##  647511.4  647633.0 -323743.7  647487.4    185944
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.5198 -0.4322  0.0902  0.5417  5.9772
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  pid      (Intercept)      1.7183   1.3108
##           employmentstatus 0.5383   0.7337   -0.65
##  Residual                  1.4073   1.1863
## Number of obs: 185956, groups:  pid, 39161
##
## Fixed effects:
##                            Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)              6.795e+00  2.236e-02  2.140e+04 303.951  < 2e-16 ***
## scale(health_lag)       -3.130e-01  3.959e-03  1.780e+05 -79.071  < 2e-16 ***
## scale(log_eq_income)     1.188e-01  4.467e-03  1.158e+05  26.582  < 2e-16 ***
## scale(baseyear)          1.132e-01  4.272e-03  1.368e+05  26.502  < 2e-16 ***
## married                  3.617e-01  1.049e-02  1.012e+05  34.490  < 2e-16 ***
## employmentstatus         2.475e-01  2.144e-02  1.704e+04  11.546  < 2e-16 ***
## female                   2.597e-01  2.566e-02  1.590e+04  10.120  < 2e-16 ***
## employmentstatus:female -1.903e-01  2.585e-02  1.474e+04  -7.361 1.93e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(_) sc(__) scl(b) marrid emplym female
## scl(hlth_l)  0.011
## scl(lg_q_n)  0.223  0.013
## scale(bsyr) -0.078 -0.042 -0.106
## married     -0.326  0.012 -0.100 -0.007
## emplymntstt -0.855  0.005 -0.214  0.077  0.010
## female      -0.762 -0.019 -0.088  0.057 -0.006  0.725
## emplymntst:  0.682  0.002  0.087 -0.063  0.024 -0.810 -0.868
```

```
#compute LRtests
lrtest(ml1, mnull)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + (1 | pid)
## Model 2: happiness ~ 1 + (1 | pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -324434
## 2   3 -328621 -5 8373.8  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ml1a, mnull)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(log_eq_income) + employmentstatus + (1 | pid)
## Model 2: happiness ~ 1 + (1 | pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -328018
## 2   3 -328621 -2 1205.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ml1, ml1a)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + (1 | pid)
## Model 2: happiness ~ scale(log_eq_income) + employmentstatus + (1 | pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -324434
## 2   5 -328018 -3 7168.7  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ml2, ml1)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 | pid)
## Model 2: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + (1 | pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -324396
## 2   8 -324434 -1 75.541  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(mrs, ml2)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 + employmentstatus |
##     pid)
## Model 2: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 | pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -323771
## 2   9 -324396 -2 1250.7  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(m_interaction_rs, mrs)
```

```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus * female + (1 + employmentstatus |
##     pid)
## Model 2: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus + female + (1 + employmentstatus |
##     pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -323744
## 2  11 -323771 -1 53.933  2.075e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(m_interaction2_rs,m_interaction_rs)
```

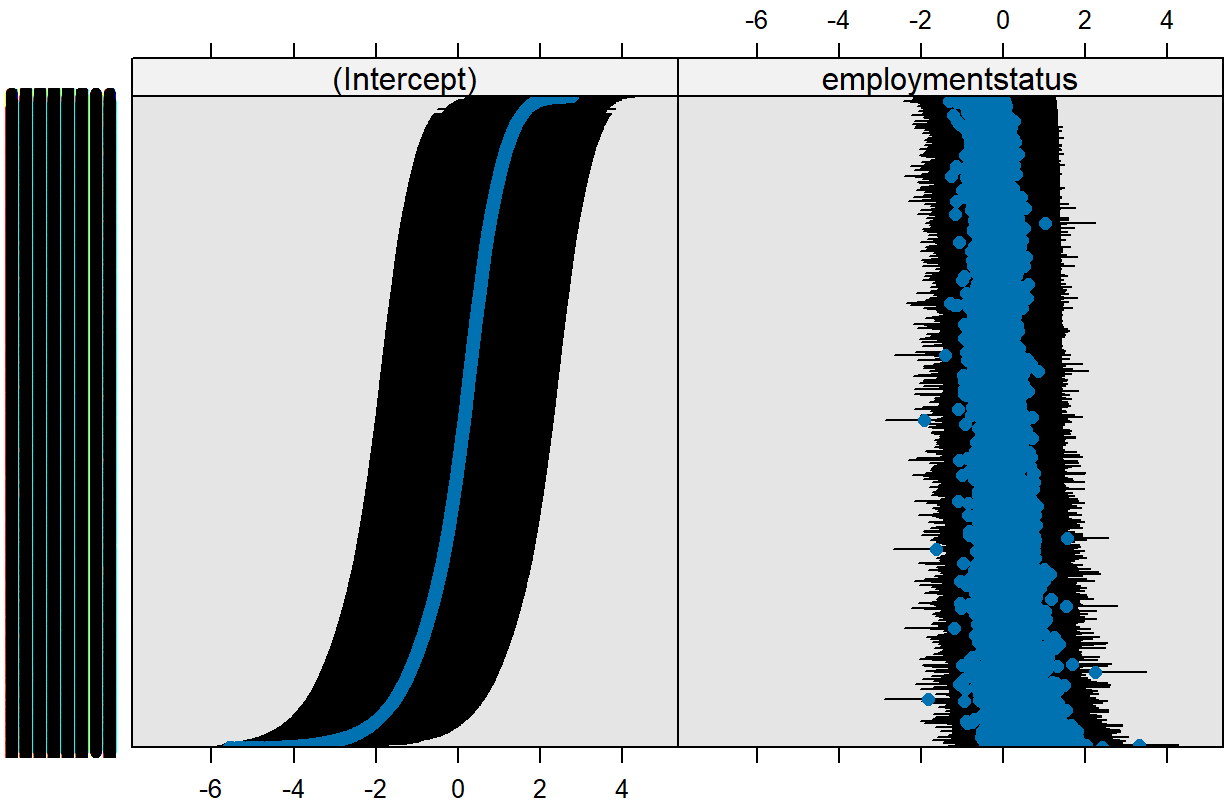
```
## Likelihood ratio test
##
## Model 1: happiness ~ scale(health_lag) + scale(log_eq_income) * female +
##     scale(baseyear) + married + employmentstatus * female + (1 +
##     employmentstatus | pid)
## Model 2: happiness ~ scale(health_lag) + scale(log_eq_income) + scale(baseyear) +
##     married + employmentstatus * female + (1 + employmentstatus |
##     pid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  13 -323728
## 2  12 -323744 -1 31.786  1.722e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The likelihood ratio tests are statistically significant, implying that the
#more complex model fits the data statistically significantly better than
#the simpler model


#see whether random slope makes sense
#yes: statistically significant
#ranova(m_interaction_rs)
```

#dotplot

```
library(lattice)
dotplot_model<- dotplot(ranef(m_interaction_rs))
dotplot_model
```

```
## $pid
```



interpreting income effects

```
# Mean and SD of log-income
mean_log_inc <- mean(pequiv_subset$log_eq_income, na.rm = TRUE)
sd_log_inc   <- sd(pequiv_subset$log_eq_income, na.rm = TRUE)

# A 1 SD increase means multiplying income by exp(sd_log_inc)

income_multiplier <- exp(sd_log_inc)
income_multiplier
```

```
## [1] 1.931625
```

```
#1SD increase in log income corresponds with 93% increase in income

#what happens when income doubles:

delta_log = log(2)
delta_std = delta_log / sd_log_inc
delta_std
```

```
## [1] 1.052836
```

```
#Then multiply by your model coefficient:
delta_happiness = 0.12 * delta_std
delta_happiness
```

```
## [1] 0.1263403
```

```
#doubling income means 0.126 higher expected happiness
```

##variance explained by interaction (Proportional Reduction in Random Slope Variance (PRV))

```
# Full variance-covariance matrix
vc_mrs <- VarCorr(mrs)$pid
vc_int <- VarCorr(m_interaction_rs)$pid

# Extract random slope variances
var_slope_A <- as.numeric(vc_mrs[2, 2])
var_slope_B <- as.numeric(vc_int[2, 2])

# Calculate proportional reduction
prv_slope <- (var_slope_A - var_slope_B) / var_slope_A
prv_slope
```

```
## [1] 0.01377208
```

```
((as.data.frame(VarCorr(mrs))[2,4])-(as.data.frame(VarCorr(m_interaction_rs))[2,4])) /
(as.data.frame(VarCorr(mrs))[2,4])
```

```
## [1] 0.01377208
```

```
#1.4% of in between-person variance in the effect of employment on life satisfaction is expla
ined by gender differences
```

##Visualize

```
library(emmeans)
```

```
## Warning: Paket 'emmeans' wurde unter R Version 4.4.3 erstellt
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```
# Step 1: Get estimated marginal means
emm1 <- emmeans(m_interaction_rs, ~ employmentstatus * female)
```

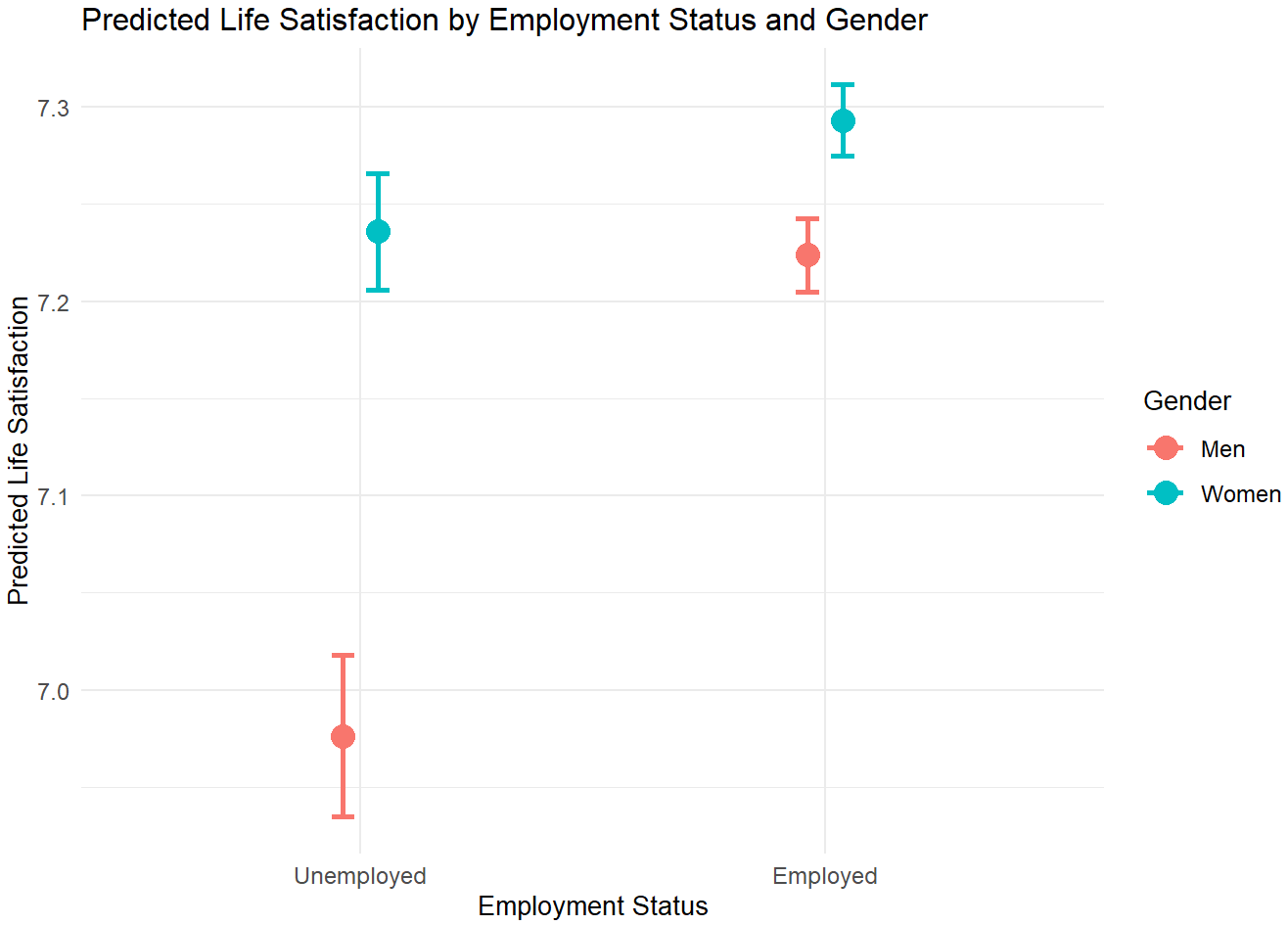
```
## Note: D.f. calculations have been disabled because the number of observations exceeds 300
0.
## To enable adjustments, add the argument 'pbkrtest.limit = 185956' (or larger)
## [or, globally, 'set emm_options(pbkrtest.limit = 185956)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 300
0.
## To enable adjustments, add the argument 'lmerTest.limit = 185956' (or larger)
## [or, globally, 'set emm_options(lmerTest.limit = 185956)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
# Step 2: Convert to data frame
emm_df <- as.data.frame(emm1)
emm_df
```

```
##  employmentstatus female   emmean          SE  df asymp.LCL asymp.UCL
##                 0      0 6.976322 0.021235224 Inf  6.934702  7.017942
##                 1      0 7.223815 0.009665879 Inf  7.204870  7.242759
##                 0      1 7.235977 0.015282351 Inf  7.206024  7.265930
##                 1      1 7.293170 0.009343678 Inf  7.274857  7.311483
##
## Results are averaged over the levels of: married
## Degrees-of-freedom method: asymptotic
## Confidence level used: 0.95
```

```
# Step 3: Optional – relabel for clarity
emm_df <- emm_df %>%
  mutate(
    employment = factor(employmentstatus, labels = c("Unemployed", "Employed")),
    gender = factor(female, labels = c("Men", "Women"))
  )

# Step 4: Plot
employmentplot <- ggplot(emm_df, aes(x = employment, y = emmean, color = gender)) +
  geom_point(position = position_dodge(width = 0.15), size = 4) +  # thicker points
  geom_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL),
                position = position_dodge(width = 0.15),
                width = 0.1, size = 1) +  # thicker CI bars
  labs(
    x = "Employment Status",
    y = "Predicted Life Satisfaction",
    color = "Gender",
    title = "Predicted Life Satisfaction by Employment Status and Gender"
  ) +
  theme_minimal(base_size = 10) +  # smaller overall sizing
  theme(
    plot.title = element_text(size = 12),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 9),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 9)
  )

employmentplot
```

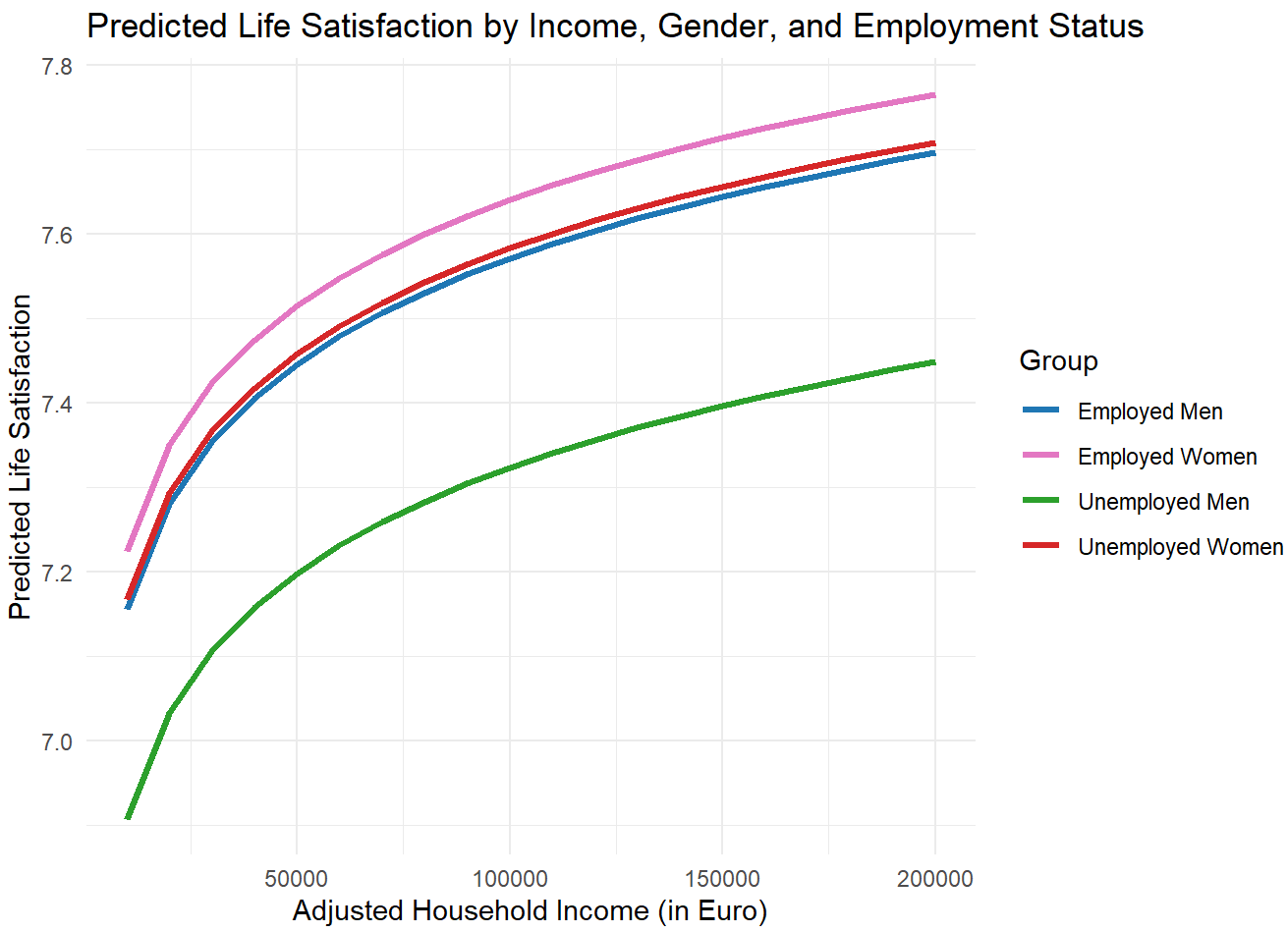


##visualize life satisfaction for differing incomes

```
##create a new grid with reasonable income values and predict life satisfaction for them
###try with four different gender/employment cmbinations

# Step 1: Create raw income range
income_vals <- seq(10000, 200000, by = 10000)
base_df <- expand.grid(
  raw_income = income_vals,
  employmentstatus = c(0, 1),
  female = c(0, 1)
)

# Step 2: Log1p-transform income
base_df$log_eq_income <- log1p(base_df$raw_income)

# Step 3: Standardize log_hhincome
mu <- mean(pequiv_subset$log_eq_income, na.rm = TRUE)
sigma <- sd(pequiv_subset$log_eq_income, na.rm = TRUE)
base_df$log_hhincome_scaled <- (base_df$log_eq_income - mu) / sigma

# Step 4: Fill in average values for other covariates
base_df <- base_df %>%
  mutate(
    health_lag = mean(pequiv_subset$health_lag, na.rm = TRUE),
    baseyear = mean(pequiv_subset$baseyear, na.rm = TRUE),
   married = mean(pequiv_subset$married, na.rm = TRUE)
  )
head(base_df)
```

```
##   raw_income employmentstatus female log_eq_income log_hhincome_scaled
## 1      10000                0      0      9.210440         -1.09160026
## 2      20000                0      0      9.903538         -0.03884001
## 3      30000                0      0     10.308986          0.57700437
## 4      40000                0      0     10.596660          1.01395821
## 5      50000                0      0     10.819798          1.35288816
## 6      60000                0      0     11.002117          1.62981524
##   health_lag baseyear   married
## 1    2.38344 9.530626 0.6693142
## 2    2.38344 9.530626 0.6693142
## 3    2.38344 9.530626 0.6693142
## 4    2.38344 9.530626 0.6693142
## 5    2.38344 9.530626 0.6693142
## 6    2.38344 9.530626 0.6693142
```

```
# Step 5: Predict happiness (fixed effects only)
base_df$predicted_happiness <- predict(m_interaction_rs, newdata = base_df, re.form = NA)

# Step 6: Label groups for plotting
base_df <- base_df %>%
  mutate(
    group = case_when(
      employmentstatus == 1 & female == 0 ~ "Employed Men",
      employmentstatus == 1 & female == 1 ~ "Employed Women",
      employmentstatus == 0 & female == 0 ~ "Unemployed Men",
      employmentstatus == 0 & female == 1 ~ "Unemployed Women"
    )
  )

# Step 7: Plot
incomeplot <- ggplot(base_df, aes(x = raw_income, y = predicted_happiness, color = group)) +
  geom_line(size = 1.2) +
  labs(
    x = "Adjusted Household Income (in Euro)",
    y = "Predicted Life Satisfaction",
    color = "Group",
    title = "Predicted Life Satisfaction by Income, Gender, and Employment Status"
  ) +
  theme_minimal() +
  scale_color_manual(values = c(
    "Employed Men" = "#1f77b4",
    "Employed Women" = "#e377c2",
    "Unemployed Men" = "#2ca02c",
    "Unemployed Women" = "#d62728"
  ))
incomeplot
```



##checking explained variances

```
# ml1a vs the null model
#explained var at lvl 1 by key predictors
1-((sigma(ml1a)^2/ (sigma(mnull)^2)))
```

```
## [1] -0.0001656869
```

```
#explained var at lvl 2
1-(as.numeric(VarCorr(ml1a)) / as.numeric(VarCorr(mnull)))
```

```
## [1] 0.04090421
```

```
# ml1 vs the null model
#explained var at lvl 1
1-((sigma(ml1)^2/ (sigma(mnull)^2)))
```

```
## [1] -0.01034937
```

```
#explained var at lvl 2
1-(as.numeric(VarCorr(ml1)) / as.numeric(VarCorr(mnull)))
```

```
## [1] 0.2966687
```

```
#model_2 vs null model
#explained var at lvl 1
1-((sigma(ml2)^2/ (sigma(mnull)^2)))
```

```
## [1] -0.01057826
```

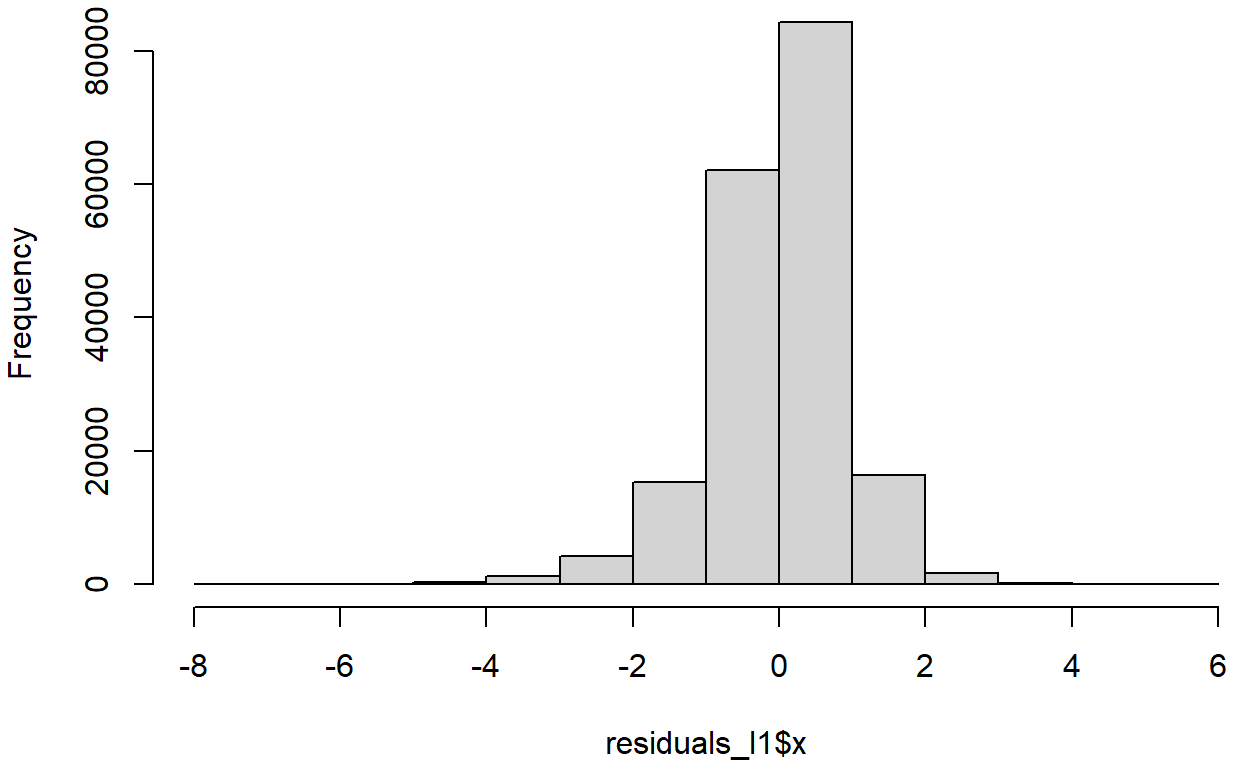
```
#explained var at lvl 2
1-(as.numeric(VarCorr(ml2)) / as.numeric(VarCorr(mnull)))
```

```
## [1] 0.2995537
```

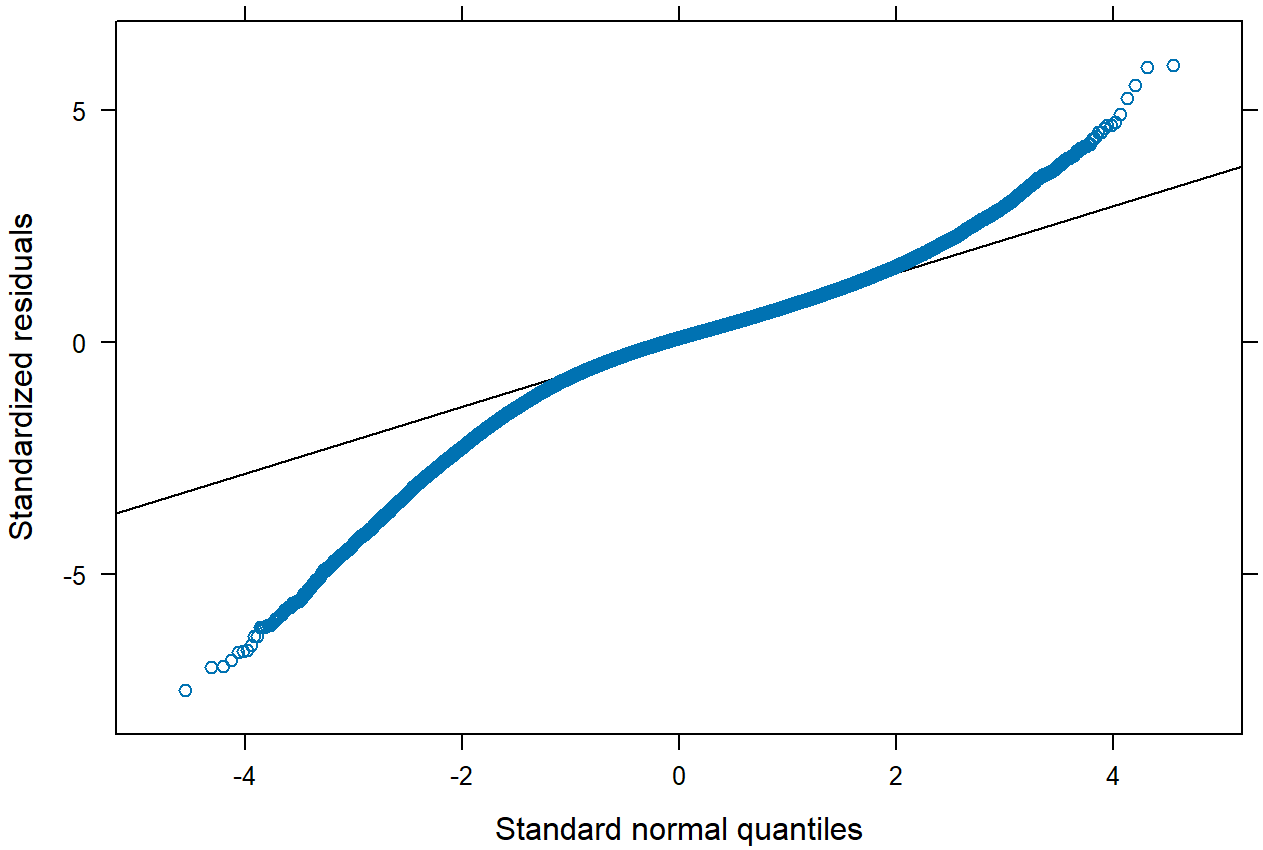
```
#different table like in hox book
#tab_model(
#  mnull, ml1, ml2, mrs, m_interaction_rs,
#  show.re.var = TRUE,       # Show random effects variances
#  show.icc = TRUE,          # Show ICC
#  show.aic = TRUE,          # Show AIC
#  show.bic = TRUE,          # Show BIC
#  dv.labels = c("M1: Null Model", "M2: L1 Predictors", "M3: +Gender", "M4: Random Slope", "M
5: Interaction RS"),
#  title = "Multilevel Models Predicting Happiness"
#)
```

#MLM model assumptions

```
model_interaction_rs <- m_interaction_rs

#residuals level 1
z_residuals_l1 <- resid(model_interaction_rs, type = "pearson", scale = T)

#save as DF
residuals_l1 <- data.frame(x = z_residuals_l1 )

hist(residuals_l1$x)
```




```
#not perectly normal distributed

qqmath(model_interaction_rs)
```




```
##deviations at the top and bottom
```

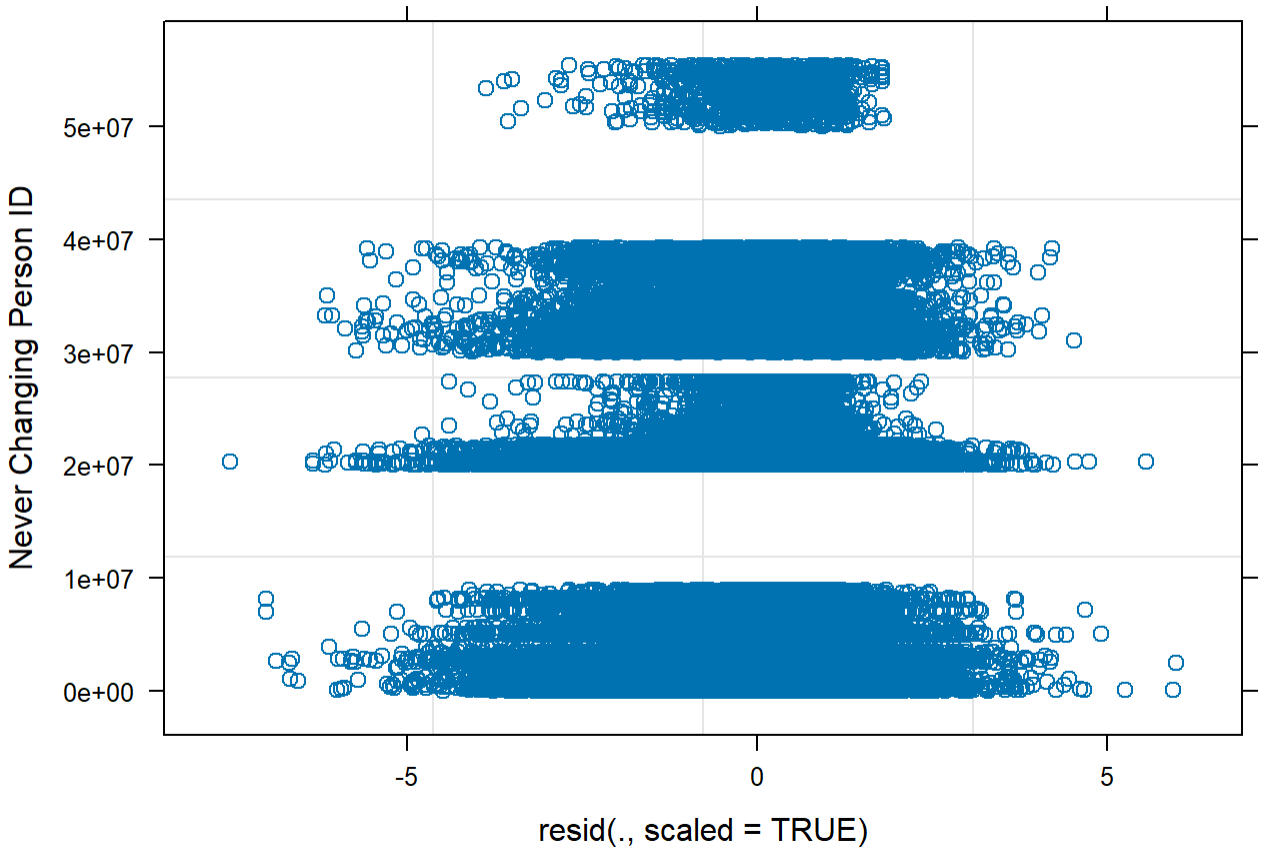
#Check whether the variance of the L1 residual errors is the same in all groups

```
# Save the L1 residuals. let's add these as a variable to the dta
pequiv_subset$l1resid<-residuals(model_interaction_rs)

pequiv_subset$abs.l1resid<-abs(pequiv_subset$l1resid)
# absolute value of residuals
pequiv_subset$abs.l1resid2<-pequiv_subset$abs.l1resid^2
# squared absolute value of the residuals
Levene.Model.F <- lm(abs.l1resid2 ~ pid, data=pequiv_subset)
#ANOVA of the squared residuals
anova(Levene.Model.F) #displays the results
```

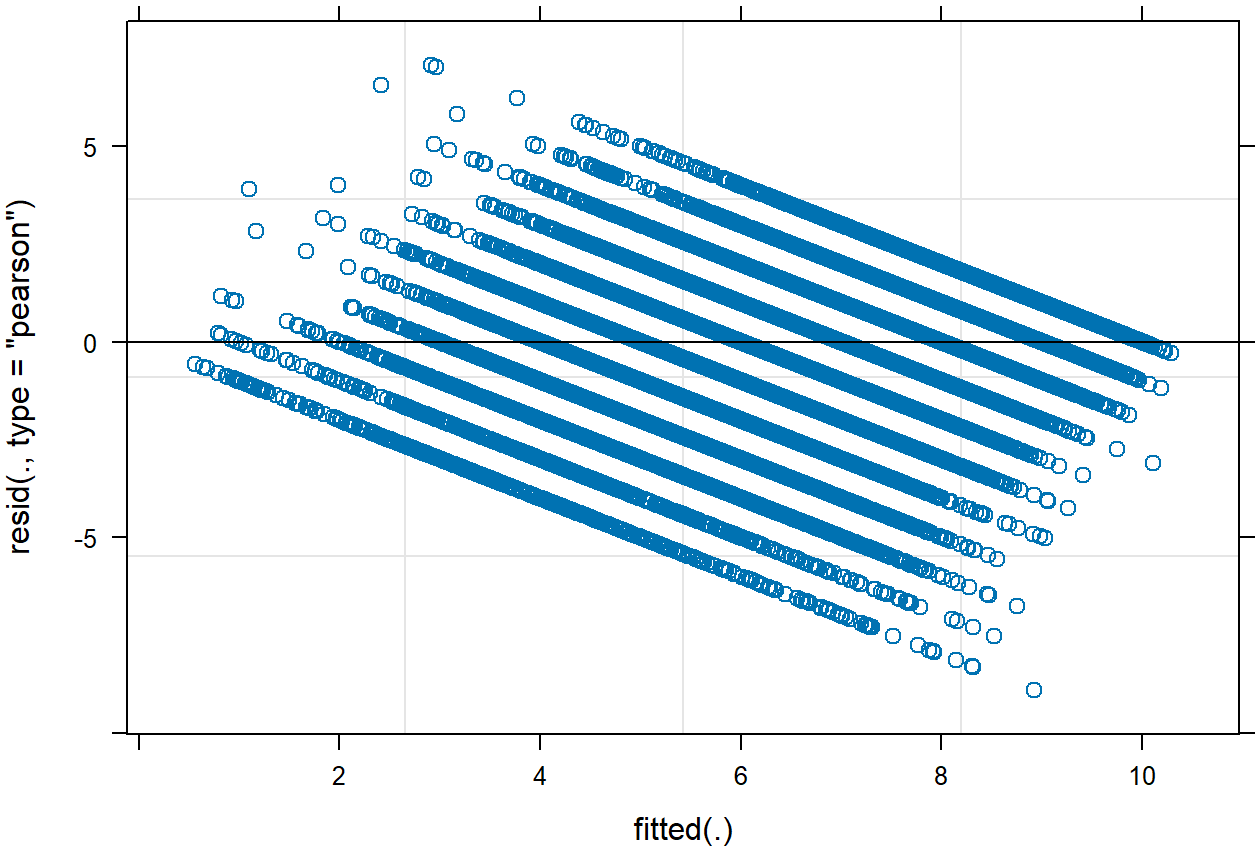
```
## Analysis of Variance Table
##
## Response: abs.l1resid2
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## pid            1      28  28.035   3.997 0.04558 *
## Residuals 185954 1304275   7.014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##significantly different at 5% f-test
```

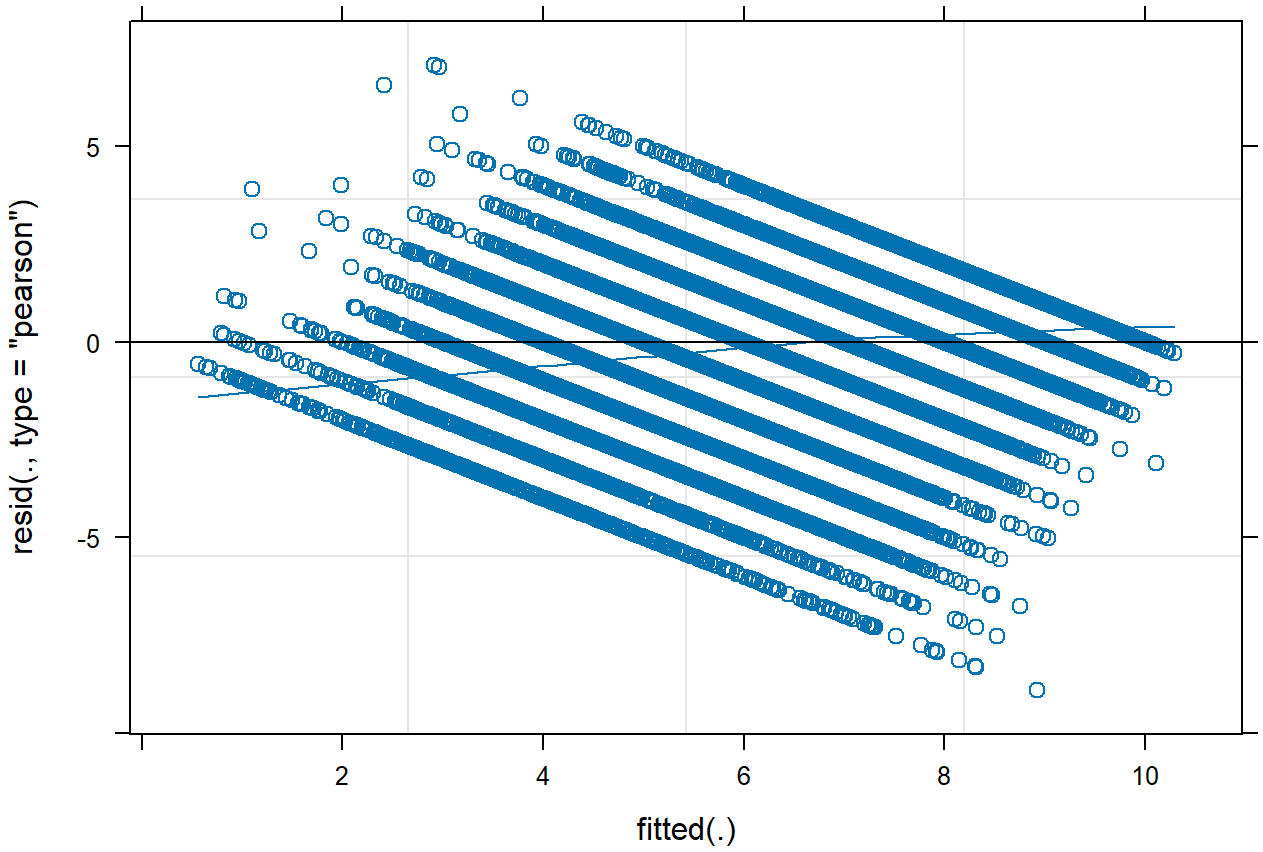
```
plot(model_interaction_rs, pid ~ resid(., scaled=TRUE))
```




```
##slide deviation to the left

plot(model_interaction_rs,
     form = resid(., type = "pearson") ~ fitted(.), abline=0)
```




```
##o check for failures of normality, nonlinearity, and heteroscedasticity. If these assumptio
ns are met, the plotted points should be evenly divided above and below their mean value of z
ero. That is, there should be no relation between the variability around the zero-line and th
e fitted value (variability around the line should not be a function of the fitted values).
plot(model_interaction_rs, type = c("p", "smooth"))
```

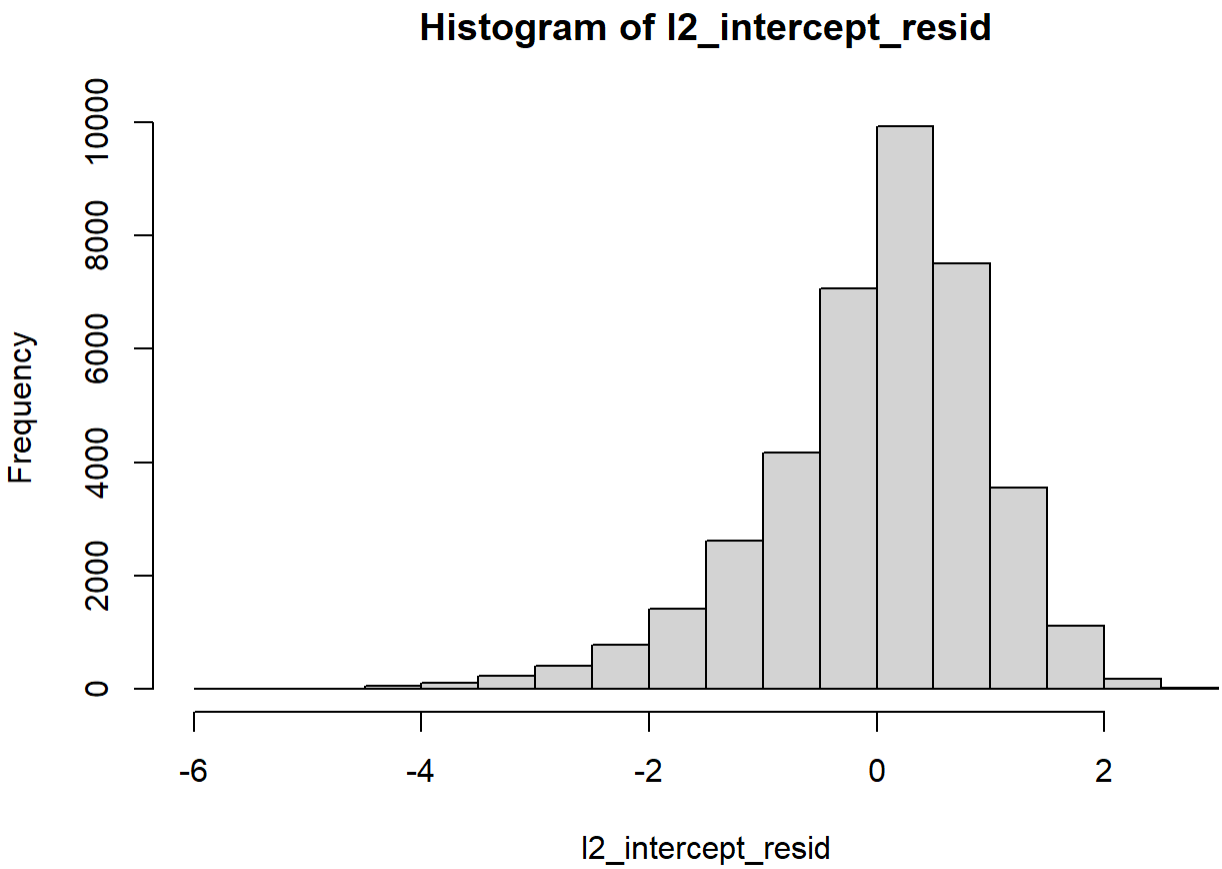



```
##smooth line is not horizontal but close enough
```

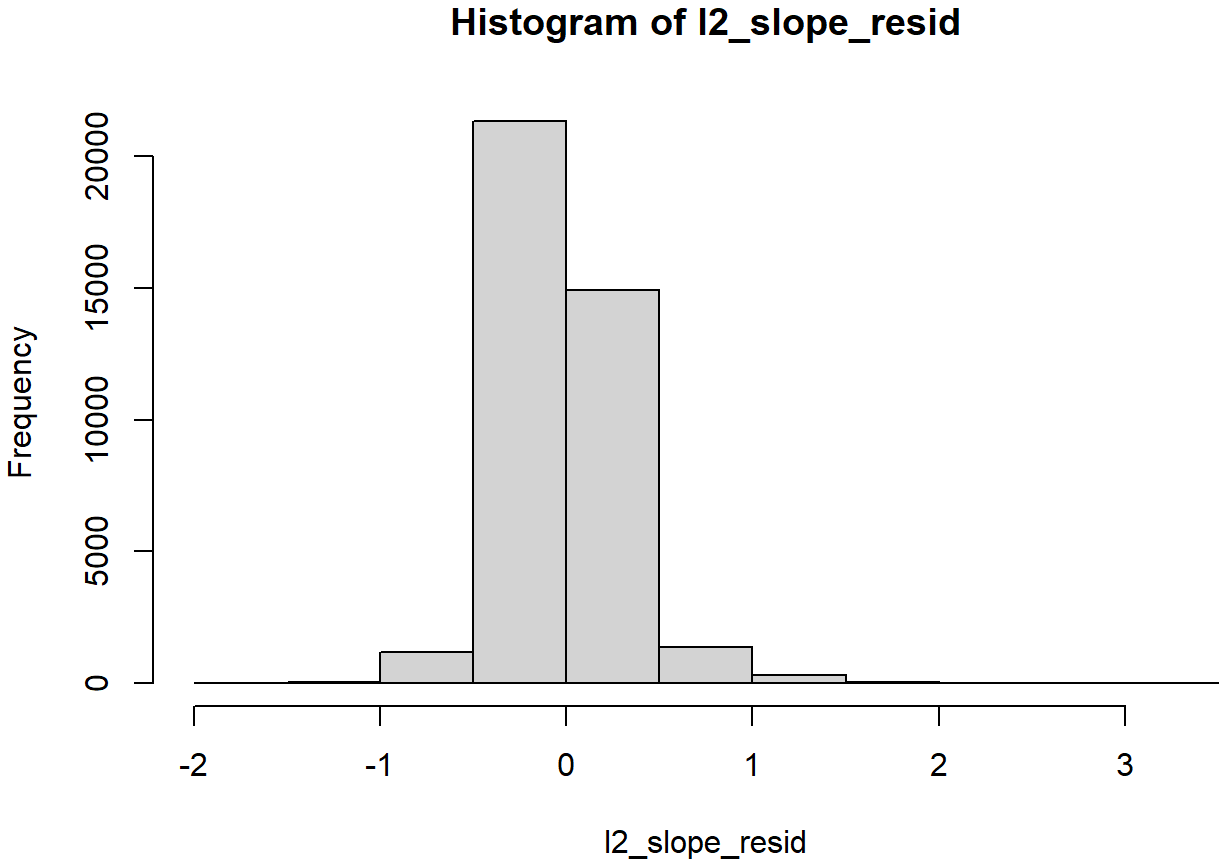
```
#l2 data
l2_data <- pequiv_subset %>%
  group_by(pid) %>% # group data by clustering variable, student
  mutate(happiness_mean = mean(happiness)) %>% # create mean gpa per student
  select(pid, happiness_mean, female) %>%
  unique()


head(l2_data)
```

```
## # A tibble: 6 × 3
## # Groups:   pid [6]
##   pid       happiness_mean female
##   <dbl+lbl>          <dbl>  <dbl>
## 1  203                7.14      0
## 2  601                5         0
## 3  602                5.17      1
## 4 1501                6.83      0
## 5 1602                6.38      1
## 6 1704                9.75      1
```

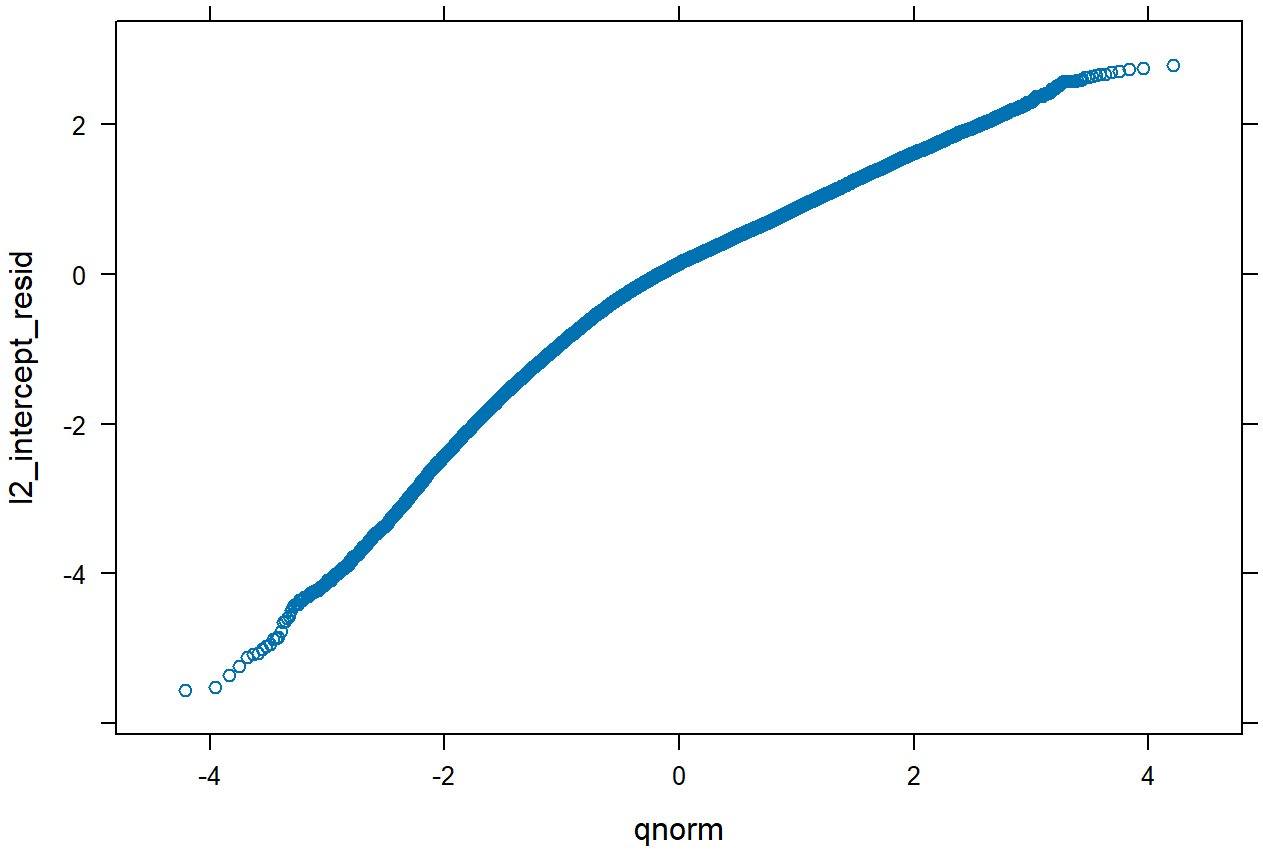
```
#residuals level 2
l2_intercept_resid = ranef(model_interaction_rs)$pid [, 1]
l2_slope_resid = ranef(model_interaction_rs)$pid[, 2]

#plot histograms
hist(l2_intercept_resid)
```



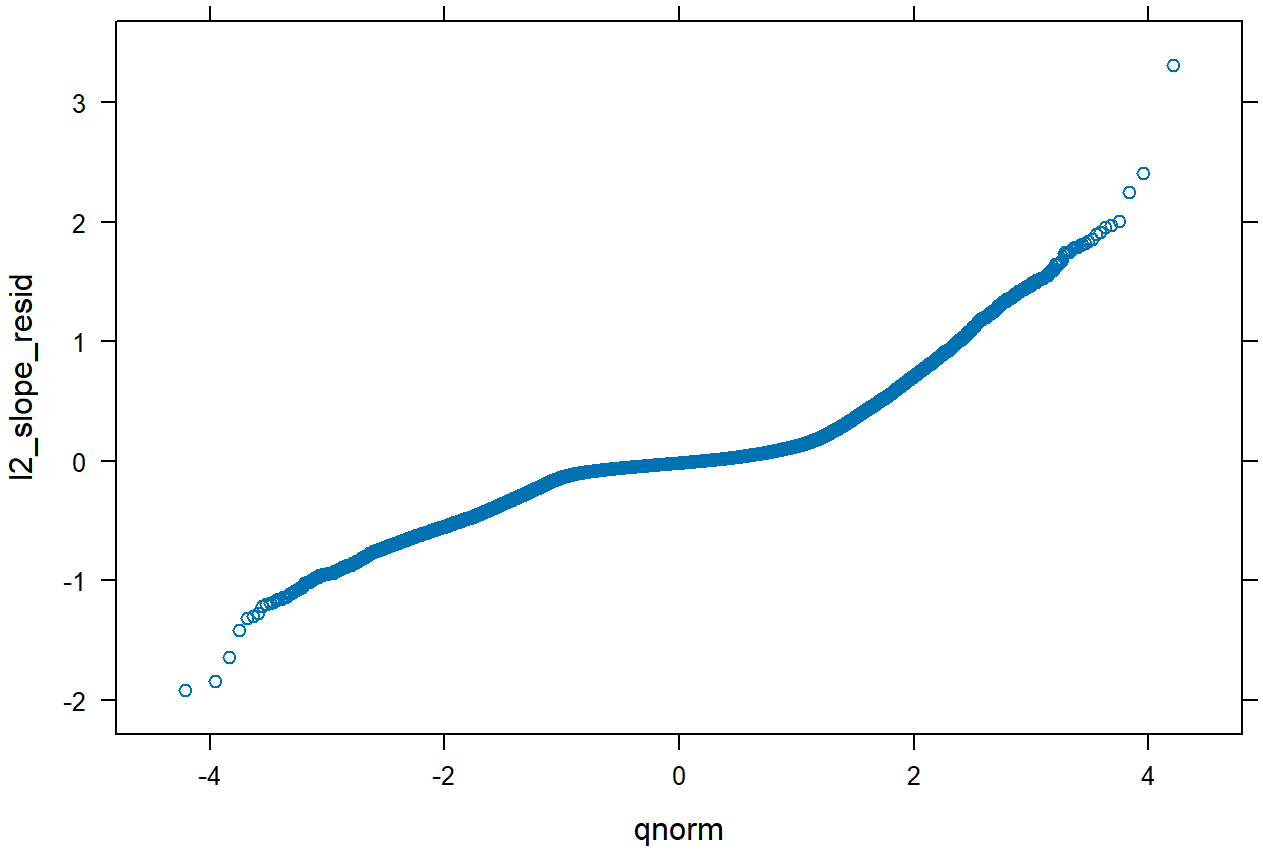

```
hist(l2_slope_resid)
```




```
#plot qq plots
qqmath(l2_intercept_resid)
```



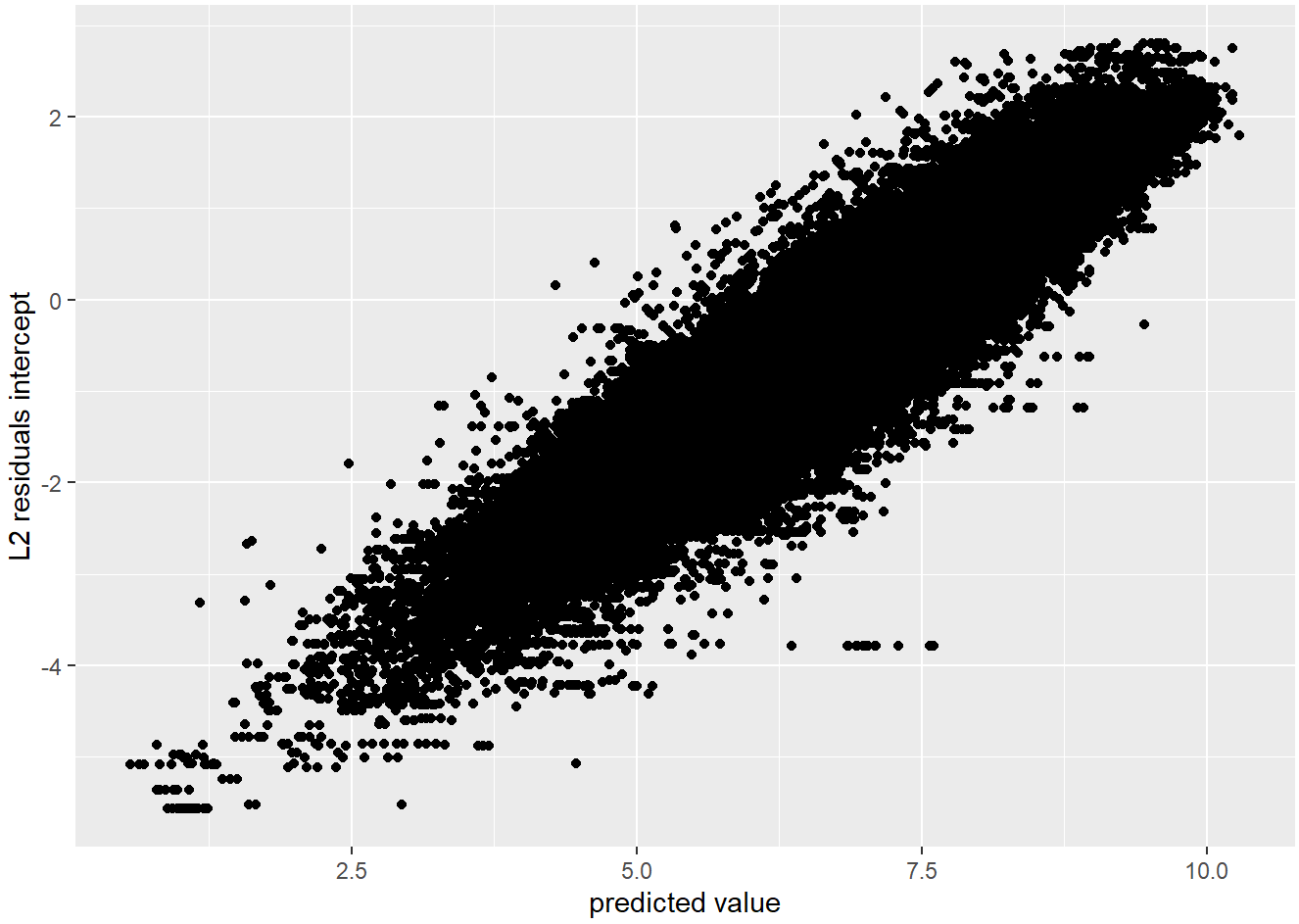

```
qqmath(l2_slope_resid)
```



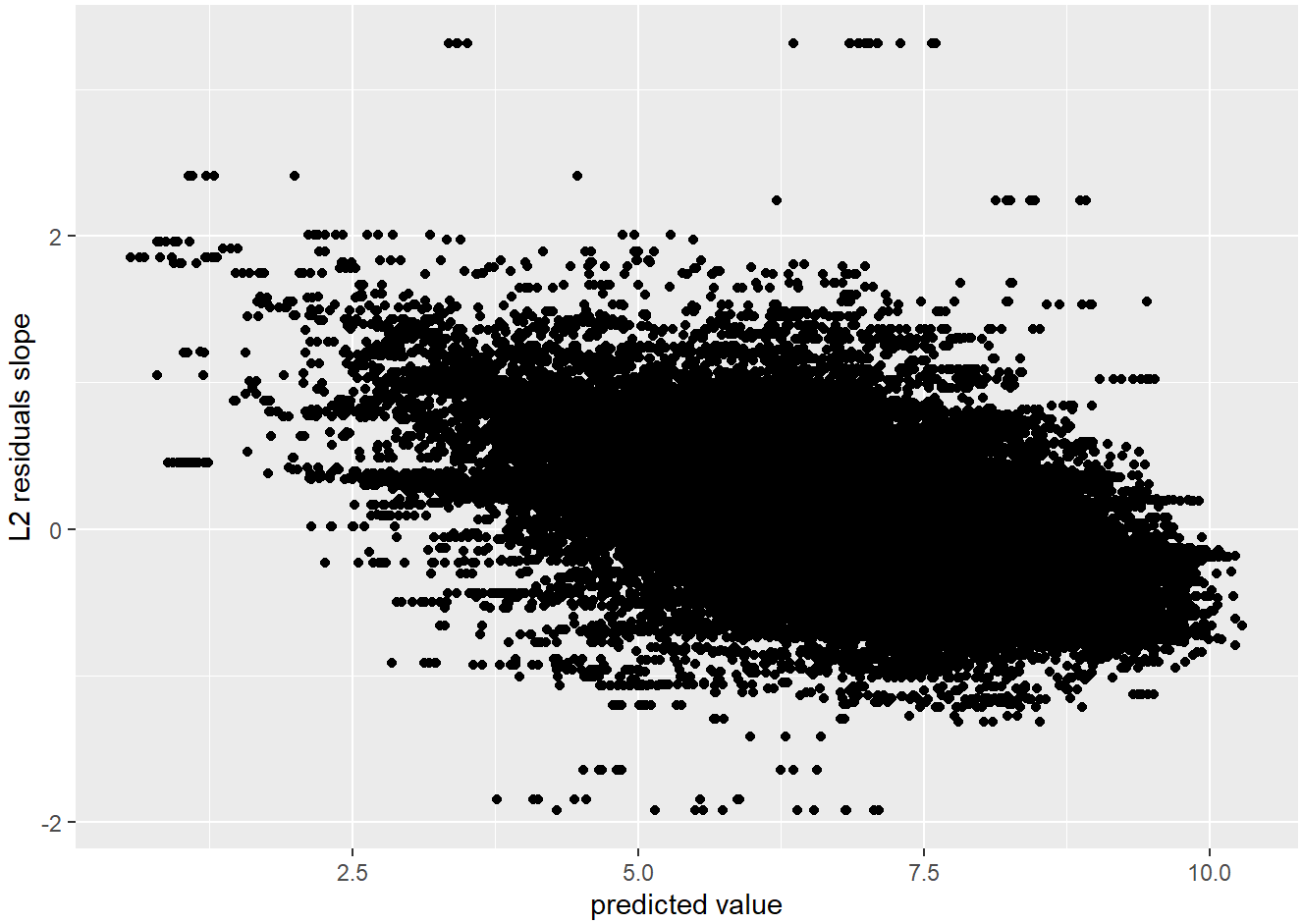

```
###level 2 residuals also not perfectly normal distributed but closer to normal distribution.
affects standard errors


#add the lvl 2 residuals to lvl 2 data

l2_data$intercept_resid <- l2_intercept_resid
l2_data$slope_resid <- l2_slope_resid


# merge level-2 data to predicted scores (these predicted scores are at the
# l1-level)
pequiv_subset_pred <-pequiv_subset %>% inner_join(l2_data[,], by = "pid" ,
                                                  relationship="many-to-one")
pequiv_subset_pred
```

```
## # A tibble: 185,956 × 39
##    education health    pid   year  age   gender  hh_orig hh_wave hh_size hh_kids
##    <dbl+lbl> <dbl+lbl> <dbl> <dbl> <dbl> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+l>
##  1 18        1 [[1] V… 203   2002  42    1 [[1]… 27       60313  1       0
##  2 18        2 [[2] G… 203   2003  43    1 [[1]… 27       60313  1       0
##  3 18        1 [[1] V… 203   2004  44    1 [[1]… 27       60313  1       0
##  4 18        1 [[1] V… 203   2005  45    1 [[1]… 27       60313  1       0
##  5 18        2 [[2] G… 203   2006  46    1 [[1]… 27       60313  1       0
##  6 18        2 [[2] G… 203   2007  47    1 [[1]… 27       60313  1       0
##  7 18        2 [[2] G… 203   2008  48    1 [[1]… 27       60313  1       0
##  8 18        2 [[2] G… 601   2002  48    1 [[1]… 60      108898  2       0
##  9 18        2 [[2] G… 601   2003  49    1 [[1]… 60      108898  2       0
## 10 18        3 [[3] S… 602   2002  44    2 [[2]… 60          60  4       2
## # ℹ 185,946 more rows
## # ℹ 29 more variables: marital_status <dbl+lbl>, employmentlvl <dbl+lbl>,
## #   indiv_wage <dbl+lbl>, happiness <dbl+lbl>, employmentstatus <dbl+lbl>,
## #   hhnetincome <dbl+lbl>, was_in_hospital <dbl+lbl>, female.x <dbl>,
## #   married <dbl>, log_hhincome <dbl>, eq_income <dbl>, log_eq_income <dbl>,
## #   income_bracket2 <fct>, age_c <dbl>, baseyear <dbl>, log_eq_income_z <dbl>,
## #   hh_kids_c <dbl>, hh_size_c <dbl>, happiness_c <dbl>, health_c <dbl>, …
```

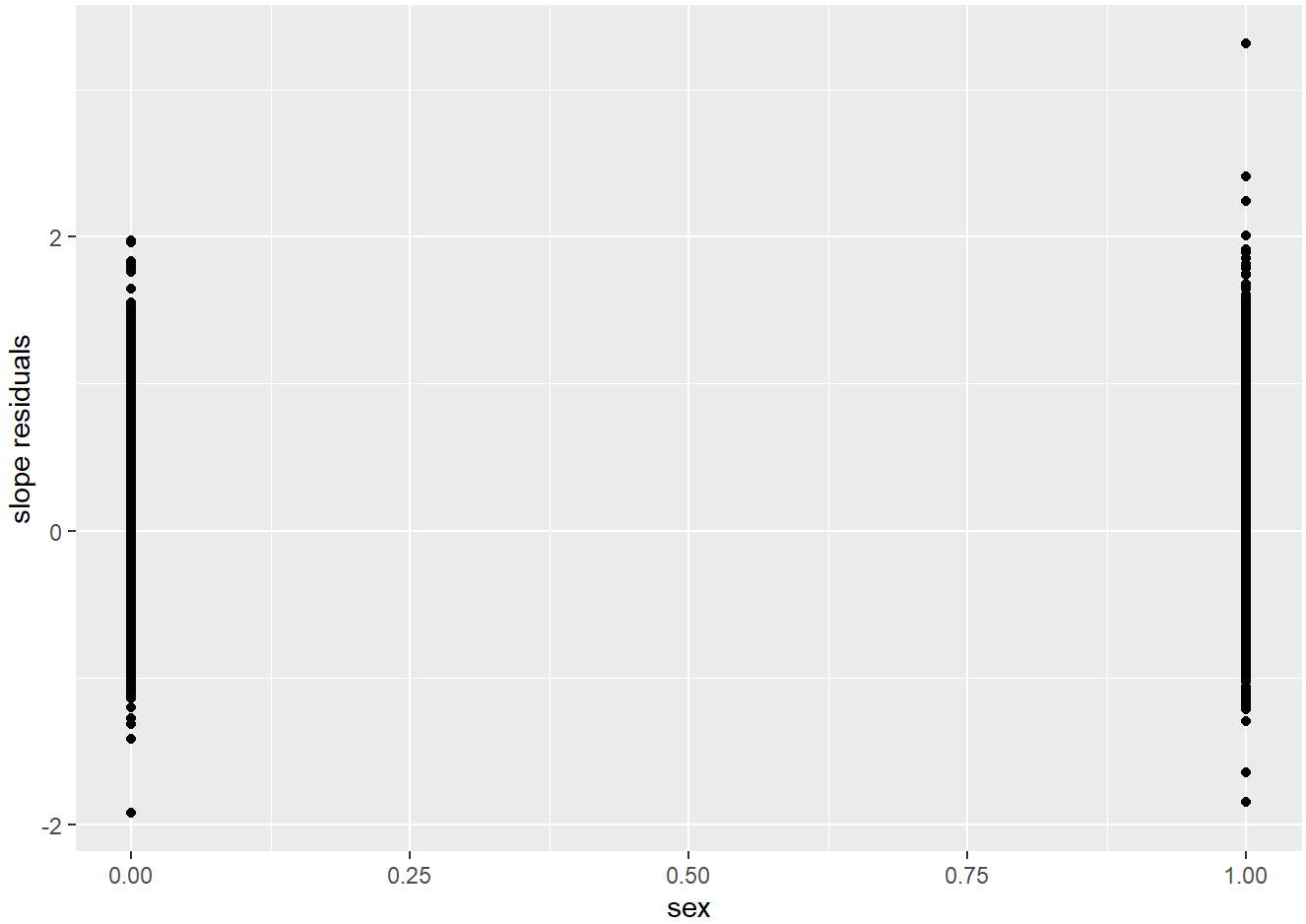
```
#add predicted scores
pequiv_subset_pred$predlmer = predict(model_interaction_rs)


pequiv_subset_pred %>%
  ggplot(mapping = aes(x = predlmer, y = intercept_resid)) +
  geom_point() +
  labs(x = "predicted value", y = "L2 residuals intercept")
```




```
pequiv_subset_pred %>%
  ggplot(mapping = aes(x = predlmer, y = slope_resid)) +
  geom_point() +
  labs(x = "predicted value", y = "L2 residuals slope")
```



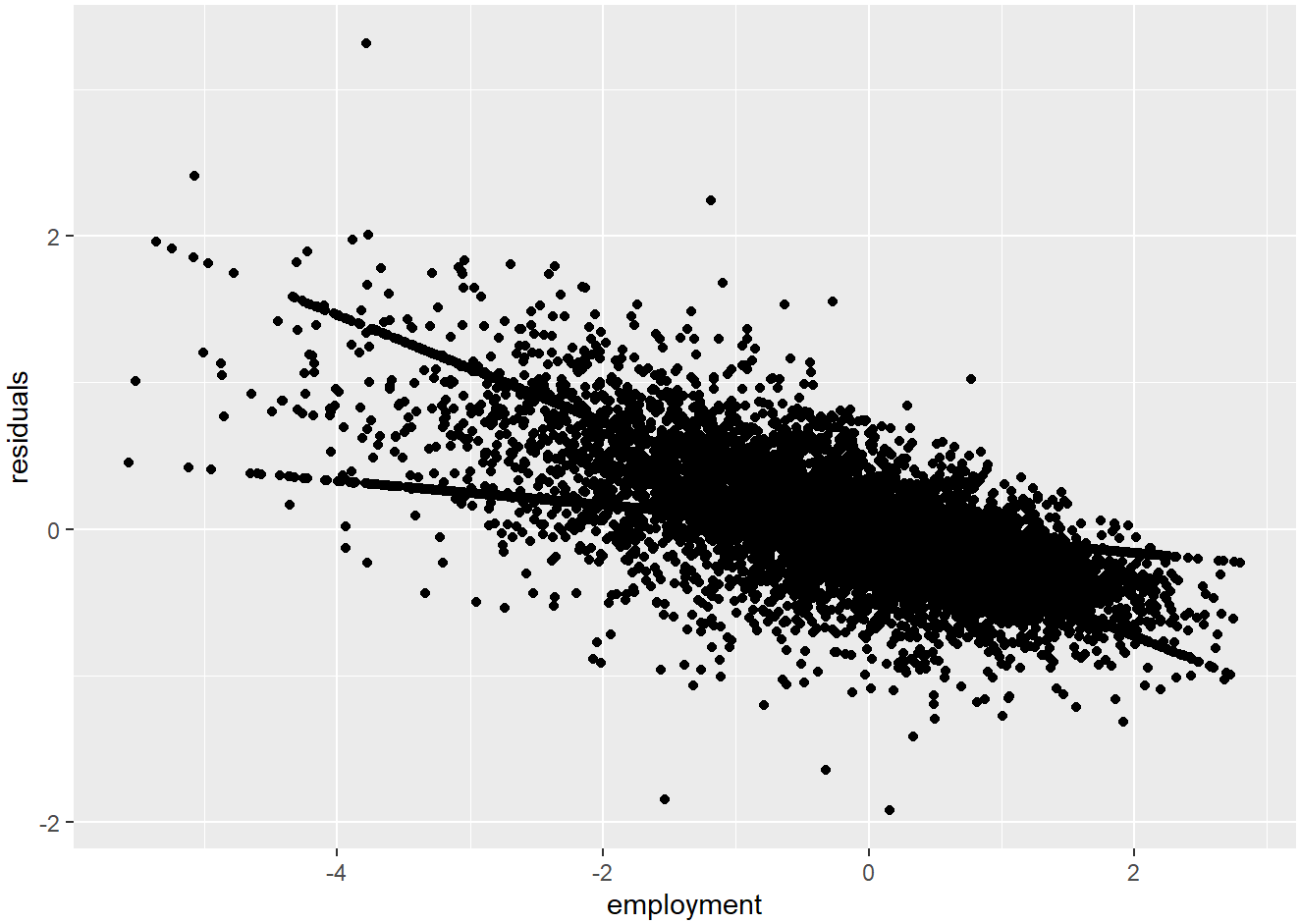

```
# look at to figure out whether to standardize residuals
# https://link.springer.com/article/10.3758/s13428-021-01709-z#Sec34
#by Sara


l2_data %>%
  ggplot(mapping = aes(x = female, y = slope_resid)) +
  geom_point() +
  labs(x = "sex", y = "slope residuals")
```




```
##gender and slope residuals not strongly correlated
cor.test(l2_data$slope_resid, as.numeric(l2_data$female))
```

```
##
##  Pearson's product-moment correlation
##
## data:  l2_data$slope_resid and as.numeric(l2_data$female)
## t = -1.198e-09, df = 39159, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.009904297  0.009904297
## sample estimates:
##           cor
## -6.053856e-12
```

```
# gender not correlated to random slope (employmentstatus)


l2_data %>%
  ggplot(mapping = aes(x = intercept_resid, y = slope_resid)) +
  geom_point() +
  labs(x = "employment", y = "residuals")
```




```
#there seems to be a negative correlation between intercept and slope residuals
cor.test(l2_data$slope_resid, as.numeric(l2_data$intercept_resid))
```

```
##
##  Pearson's product-moment correlation
##
## data:  l2_data$slope_resid and as.numeric(l2_data$intercept_resid)
## t = -210.92, df = 39159, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7338867 -0.7246129
## sample estimates:
##        cor
## -0.7292833
```

```
#strong negative correlation. bad?

# by including a covariance between intercept and slope variance we
# account for a correlation between the intercept variance and slope variance! -> maybe need
to do this?
```

check for influential cases using leverage measures. use the function hlm_influence

```
#doesnt work- too much CPU use
#infl <- hlm_influence(model_interaction_rs, level = "pid")
#print(infl, width = Inf)
```

##create variables to mark transitions to (un)-employment

```
table(pequiv_subset$employmentstatus)
```

```
##
##      0      1
##  34106 151850
```

```
attr(pequiv_subset$employmentstatus, "labels")
```

```
##             [-8] Question this year not part of survey
##                                                     -8
##         [-7] Only available in less restricted edition
##                                                     -7
##  [-6] Version of questionnaire with modified filtering
##                                                     -6
## [-5] Not included in this version of the questionnaire
##                                                     -5
##                    [-4] Inadmissable multiple response
##                                                     -4
##                                 [-3] Implausible value
##                                                     -3
##                                    [-2] Does not apply
##                                                     -2
##                                         [-1] No answer
##                                                     -1
##                                       [0] Not Employed
##                                                      0
##                                          [1] Employed
##                                                      1
```

```
pequiv_subset <- pequiv_subset %>%
  arrange(pid, year) %>%
  group_by(pid) %>%
  mutate(
    employmentstatus_lag = dplyr::lag(employmentstatus),
    emp_to_unemp = if_else(employmentstatus_lag == 1 & employmentstatus == 0, 1, 0, missing =
NA_integer_),
    unemp_to_emp = if_else(employmentstatus_lag == 0 & employmentstatus == 1, 1, 0, missing =
NA_integer_)
  ) %>%
  ungroup()

# Summary tables
table(pequiv_subset$emp_to_unemp, useNA = "always")
```

```
##
##      0      1   <NA>
## 169365   4382  12209
```

```
table(pequiv_subset$unemp_to_emp, useNA = "always")
```

```
##
##      0      1   <NA>
## 151036   7968  26952
```

#filter dataset for those who become unemployed

```
ids_emp_to_unemp <- pequiv_subset %>%
  filter(emp_to_unemp == 1) %>%
  distinct(pid)

employment_df <- pequiv_subset %>%
  filter(pid %in% ids_emp_to_unemp$pid)

length(unique(employment_df$pid))
```

```
## [1] 3890
```

##plot happiness time to transitioning from emp to unemp

```
# Step 1: Identify transitions from employment to unemployment
transition_years <- pequiv_subset %>%
  filter(emp_to_unemp == 1) %>%
  select(pid, transition_year = year)

# Step 2: Identify control group (never transitioned)
control_ids <- pequiv_subset %>%
  anti_join(transition_years, by = "pid") %>%
  distinct(pid)

# Step 3: Assign "fake" transition year to control group (median year per person)
control_fake_transition <- pequiv_subset %>%
  filter(pid %in% control_ids$pid) %>%
  group_by(pid) %>%
  summarise(transition_year = floor(median(year, na.rm = TRUE)))

# Step 4: Create treated group with relative year
treated_traj <- pequiv_subset %>%
  inner_join(transition_years, by = "pid") %>%
  mutate(
    relative_year = year - transition_year,
    group = "Transitioned"
  ) %>%
  filter(relative_year %in% -5:5)
```

```
## Warning in inner_join(., transition_years, by = "pid"): Detected an unexpected many-to-man
y relationship between `x` and `y`.
## ℹ Row 206 of `x` matches multiple rows in `y`.
## ℹ Row 1 of `y` matches multiple rows in `x`.
## ℹ If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
# Step 5: Create control group with relative year
control_traj <- pequiv_subset %>%
  inner_join(control_fake_transition, by = "pid") %>%
  mutate(
    relative_year = year - transition_year,
    group = "Control"
  ) %>%
  filter(relative_year %in% -5:5)

# Step 6: Combine and summarize with confidence intervals
happiness_compare <- bind_rows(treated_traj, control_traj) %>%
  group_by(group, relative_year) %>%
  summarise(
    mean_happiness = mean(happiness, na.rm = TRUE),
    sd = sd(happiness, na.rm = TRUE),
    n = sum(!is.na(happiness)),
    se = sd / sqrt(n),
    ci_lower = mean_happiness - 1.96 * se,
    ci_upper = mean_happiness + 1.96 * se,
    .groups = "drop"
  )

# Step 7: Plot
ggplot(happiness_compare, aes(x = relative_year, y = mean_happiness, color = group, fill = gr
oup)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), alpha = 0.2, color = NA) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  labs(
    title = "Average Happiness Around Employment → Unemployment Transition",
    x = "Years Relative to Transition",
    y = "Average Happiness",
    color = "Group",
    fill = "Group"
  ) +
  theme_minimal()
```


Average Happiness Around Employment → Unemployment Transition
7.4
7.2
7.0
6.8
Average Happiness
-5.0
-2.5
0.0
2.5
5.0
Years Relative to Transition
Group
Control
Transitioned

```
# Summarize transition group happiness trajectory with CIs
happiness_transitions_only <- treated_traj %>%
  group_by(relative_year) %>%
  summarise(
    mean_happiness = mean(happiness, na.rm = TRUE),
    sd = sd(happiness, na.rm = TRUE),
    n = sum(!is.na(happiness)),
    se = sd / sqrt(n),
    ci_lower = mean_happiness - 1.96 * se,
    ci_upper = mean_happiness + 1.96 * se,
    .groups = "drop"
  )

# Plot
ggplot(happiness_transitions_only, aes(x = relative_year, y = mean_happiness)) +
  geom_line(color = "steelblue", size = 1.2) +
  geom_point(color = "steelblue", size = 2) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), alpha = 0.2, fill = "steelblue") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  labs(
    title = "Average Happiness Around Employment → Unemployment Transition",
    x = "Years Relative to Transition",
    y = "Average Happiness"
  ) +
  theme_minimal()
```

##separately for men and women

```
# Step 1: Label gender using `female` variable
treated_traj <- treated_traj %>%
  mutate(
    gender_label = if_else(female == 1, "Women", "Men")
  )

# Step 2: Summarize happiness by gender and relative year
happiness_by_gender <- treated_traj %>%
  group_by(gender_label, relative_year) %>%
  summarise(
    mean_happiness = mean(happiness, na.rm = TRUE),
    sd = sd(happiness, na.rm = TRUE),
    n = sum(!is.na(happiness)),
    se = sd / sqrt(n),
    ci_lower = mean_happiness - 1.96 * se,
    ci_upper = mean_happiness + 1.96 * se,
    .groups = "drop"
  )

# Step 3: Plot
ggplot(happiness_by_gender, aes(x = relative_year, y = mean_happiness, color = gender_label,
fill = gender_label)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), alpha = 0.2, color = NA) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  labs(
    title = "Happiness Around Employment → Unemployment Transition, by Gender",
    x = "Years Relative to Transition",
    y = "Average Happiness",
    color = "Gender",
    fill = "Gender"
  ) +
  theme_minimal()
```


Happiness Around Employment → Unemployment Transition, by Gender
Average Happiness
7.0
6.5
6.0
-5.0
-2.5
0.0
2.5
5.0
Years Relative to Transition
Gender
Men
Women