Intro to Modeling

Asem Berkalieva

11/23/2019

Libraries

```r
library(tidyverse)
```

Data

```r
# all years (cleaned data)
hrs <- read.csv('../../data/hrs_finalized.csv')
```

Binomial Simple GLM

The simplest form: treating every observation as independent. Not what we want, but we will build up. Our response, depression, is a binary variable, so we run a binomial regression. Our covariates are income and time (years).

```r
simple_bin <- glm(depression ~ income + years + income*years,
                  data = hrs, family = "binomial")

summary(simple_bin)
```

```
##
## Call:
## glm(formula = depression ~ income + years + income * years, family = "binomial",
##     data = hrs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5737  -0.5730  -0.5722  -0.5715   2.1985
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.500e+00  2.229e+01  -0.202    0.840
## income        2.473e-07  7.852e-07   0.315    0.753
## years         1.378e-03  1.107e-02   0.124    0.901
## income:years -1.238e-10  3.902e-10  -0.317    0.751
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11461  on 13603  degrees of freedom
## Residual deviance: 11455  on 13600  degrees of freedom
## AIC: 11463
##
## Number of Fisher Scoring iterations: 4
```

GEE: Independence

Accounting for repeated measures per person, we fit a binomial GEE models. First, we use 'independence' correlation structure, which reduces the model back to a simple regression, as it does not account for repeated measures within individuals:

```
library(gee)
gee_ind <- gee(depression ~ income + years + income*years, person_id,
               data = hrs, family = binomial,
               corstr = "independence")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##   (Intercept)        income         years  income:years
## -4.499815e+00  2.472698e-07  1.378388e-03 -1.237587e-10
```

```
# summary
ss_ind <- data.frame(summary(gee_ind)$coefficients)
ss_ind <- data.frame(ss_ind, pvalue = 2 * (1 - pnorm(abs(ss_ind[,5]))))
round(ss_ind, 4)
```

```
##              Estimate Naive.S.E. Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)   -4.4998    22.2956 -0.2018     22.6547  -0.1986 0.8426
## income         0.0000     0.0000  0.3149      0.0000   0.3018 0.7628
## years          0.0014     0.0111  0.1245      0.0113   0.1225 0.9025
## income:years   0.0000     0.0000 -0.3171      0.0000  -0.3040 0.7612
```

GEE: Exchangeable

Now, we account for repeated measures within individuals. Results are identical to those using an independence correlation matrix.

```
gee_exch <- gee(depression ~ income + years + income*years, person_id,
                data = hrs, family = binomial,
                corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##   (Intercept)        income         years  income:years
## -4.499815e+00  2.472698e-07  1.378388e-03 -1.237587e-10
```

```
# summary
ss_exch <- data.frame(summary(gee_exch)$coefficients)
ss_exch <- data.frame(ss_exch, pvalue = 2 * (1 - pnorm(abs(ss_exch[,5]))))
round(ss_exch, 4)
```

```
##              Estimate Naive.S.E. Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)   -4.4998    22.2956 -0.2018     22.6547  -0.1986 0.8426
## income         0.0000     0.0000  0.3149      0.0000   0.3018 0.7628
## years          0.0014     0.0111  0.1245      0.0113   0.1225 0.9025
## income:years   0.0000     0.0000 -0.3171      0.0000  -0.3040 0.7612
```

Mixed Model

```r
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.5.2
```

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

```
## Warning: package 'Matrix' was built under R version 3.5.2
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```r
glmer <- lme4::glmer(depression ~ income + years + income*years + (1|person_id),
               data = hrs, family = binomial)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

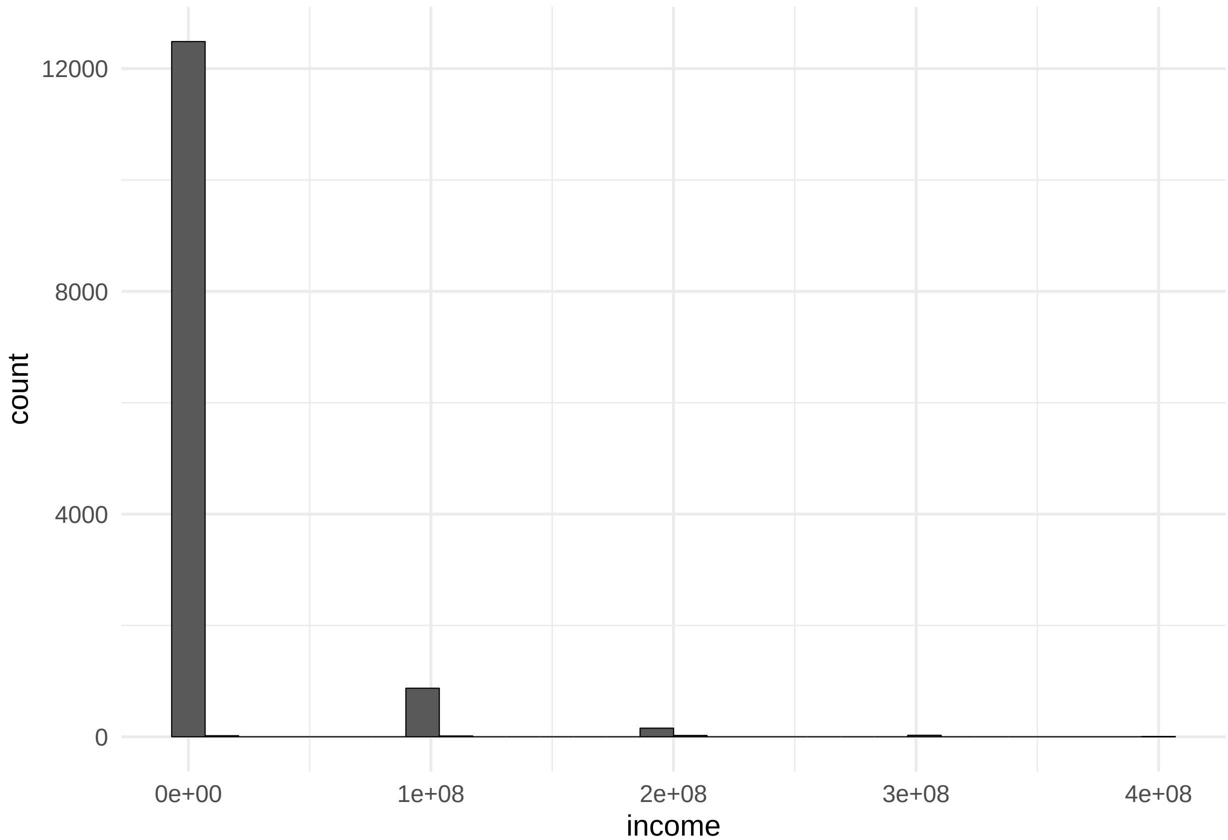
```
## Error in length(value <- as.numeric(value)) == 1L: (maxstephalfit) PIRLS step-halvings failed to redu
```

```r
summary(glmer)
```

```
## Error in object[[i]]: object of type 'closure' is not subsettable
```

Rescale

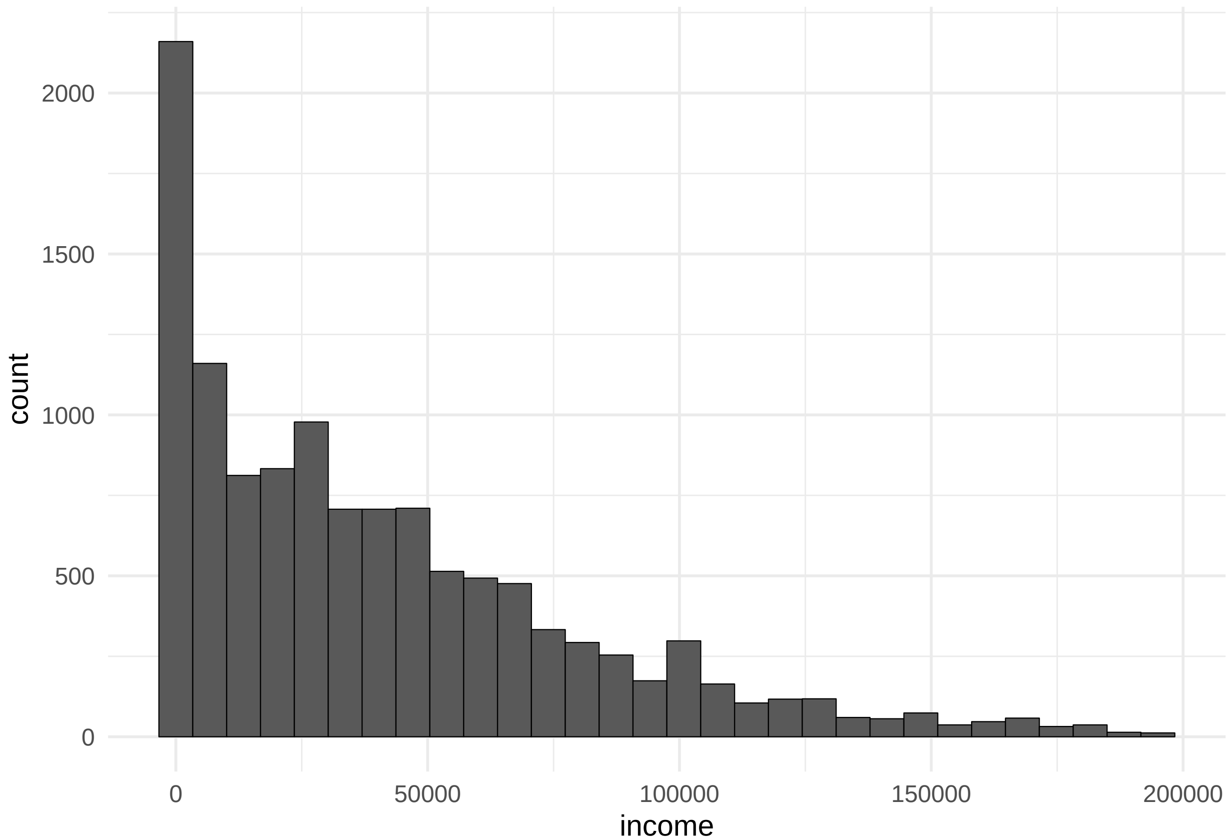
```r
library(ggplot2)
hrs %>%
  ggplot(aes(x=income)) +
    geom_histogram(col='black', lwd=0.2, bins=30) +
    theme_minimal()
```




```r
# REMOVE OUTLIERS
five_num <- summary(hrs$income)
for_out  <- (five_num[5] - five_num[2])*1.5
```

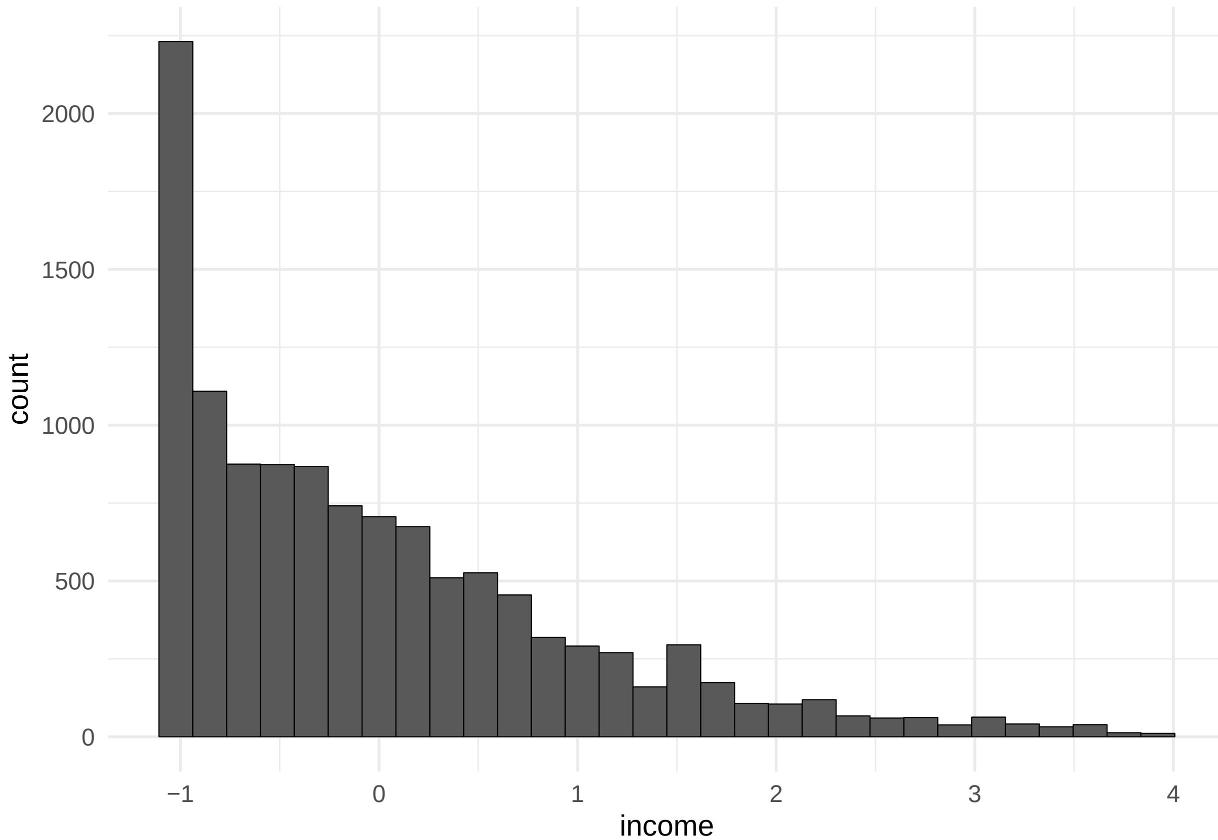
```r
# CREATE SUBSET
hrs_sub <- hrs %>%
  filter(income>0) %>%
  filter(income<five_num[5]+for_out)

hrs_sub$years[which(hrs_sub$years==2010)] <- 1
hrs_sub$years[which(hrs_sub$years==2012)] <- 2
hrs_sub$years[which(hrs_sub$years==2014)] <- 3
hrs_sub$years[which(hrs_sub$years==2015)] <- 4
```

```r
hrs_sub %>%
  ggplot(aes(x=income)) +
    geom_histogram(col='black', lwd=0.2, bins=30) +
    theme_minimal()
```




```r
# RESCALE VALUES
hrs_sub <- hrs_sub %>%
  mutate(income=scale(income))
```

```r
hrs_sub %>%
  ggplot(aes(x=income)) +
    geom_histogram(col='black', lwd=0.2, bins=30) +
    theme_minimal()
```




```
head(hrs_sub %>% select(person_id, depression, income, years))
```

```
##      person_id depression     income years
## 1 h010001pn010          0 -1.0280197     1
## 2 h010003pn030          0 -1.0163293     1
## 3 h010063pn010          1 -1.0232523     1
## 4 h010397pn010          0 -0.5284499     1
## 5 h010773pn020          0 -0.7313209     1
## 6 h010893pn010          0  1.1959536     1
```

Simple Logistic Regression

```
simple_bin <- glm(depression ~ income + years + income*years,
                  data = hrs_sub, family = "binomial")

summary(simple_bin)
```

```
##
## Call:
## glm(formula = depression ~ income + years + income * years, family = "binomial",
##     data = hrs_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6987  -0.6279  -0.5627  -0.4519   2.5494
##
```

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.770e+00  3.073e-02 -57.587   <2e-16 ***
## income       -3.292e-01  3.565e-02  -9.235   <2e-16 ***
## years         4.031e-05  3.020e-05   1.335    0.182
## income:years -2.975e-05  3.383e-05  -0.879    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10128.0  on 11832  degrees of freedom
## Residual deviance:  9977.9  on 11829  degrees of freedom
## AIC: 9985.9
##
## Number of Fisher Scoring iterations: 4
```

```r
simple_bin2 <- glm(depression ~ income + years,
                  data = hrs_sub, family = "binomial")

summary(simple_bin2)
```

```
##
## Call:
## glm(formula = depression ~ income + years, family = "binomial",
##     data = hrs_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6878  -0.6289  -0.5619  -0.4574   2.5176
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.774e+00  3.055e-02 -58.050   <2e-16 ***
## income      -3.460e-01  3.035e-02 -11.401   <2e-16 ***
## years        4.676e-05  2.920e-05   1.601    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10128.0  on 11832  degrees of freedom
## Residual deviance:  9978.7  on 11830  degrees of freedom
## AIC: 9984.7
##
## Number of Fisher Scoring iterations: 4
```

GEE: Independence

```r
library(gee)
gee_ind <- gee(depression ~ income + years + income*years, person_id,
               data = hrs_sub, family = binomial,
               corstr = "independence")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
##   (Intercept)        income         years  income:years
## -1.769907e+00 -3.292130e-01  4.030755e-05 -2.975132e-05
```

```
# summary
ss_ind <- data.frame(summary(gee_ind)$coefficients)
ss_ind <- data.frame(ss_ind, pvalue = 2 * (1 - pnorm(abs(ss_ind[,5]))))
round(ss_ind, 4)
```

```
##              Estimate Naive.S.E.  Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)   -1.7699     0.0309 -57.2427      0.0311 -56.9121 0.0000
## income        -0.3292     0.0359  -9.1794      0.0404  -8.1541 0.0000
## years          0.0000     0.0000   1.3265      0.0000   1.3228 0.1859
## income:years   0.0000     0.0000  -0.8740      0.0000  -0.8060 0.4203
```

GEE: Exchangeable

```
gee_exch <- gee(depression ~ income + years + income*years, person_id,
                data = hrs_sub, family = binomial,
                corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
##   (Intercept)        income         years  income:years
## -1.769907e+00 -3.292130e-01  4.030755e-05 -2.975132e-05
```

```
# summary
ss_exch <- data.frame(summary(gee_exch)$coefficients)
ss_exch <- data.frame(ss_exch, pvalue = 2 * (1 - pnorm(abs(ss_exch[,5]))))
round(ss_exch, 4)
```

```
##              Estimate Naive.S.E.  Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)   -1.7699     0.0309 -57.2427      0.0311 -56.9121 0.0000
## income        -0.3292     0.0359  -9.1794      0.0404  -8.1541 0.0000
## years          0.0000     0.0000   1.3265      0.0000   1.3228 0.1859
## income:years   0.0000     0.0000  -0.8740      0.0000  -0.8060 0.4203
```

Mixed Model

```
library(lme4)
glmer <- lme4::glmer(depression ~ income + years + income*years + (1|person_id),
               data = hrs_sub, family = binomial)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0692327
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unider
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```r
summary(glmer)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: depression ~ income + years + income * years + (1 | person_id)
##    Data: hrs_sub
##
##      AIC      BIC   logLik deviance df.resid
##   8755.4   8792.3  -4372.7   8745.4    11828
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6467 -0.1756 -0.1567 -0.1314  3.2955
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  person_id (Intercept) 5.964    2.442
## Number of obs: 11833, groups:  person_id, 3273
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.259e+00  1.156e-01 -28.200  < 2e-16 ***
## income       -3.361e-01  5.772e-02  -5.823 5.78e-09 ***
## years         5.819e-05  3.799e-05   1.532    0.126
## income:years -3.607e-05  4.162e-05  -0.867    0.386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) income years
## income       0.044
## years       -0.189 -0.077
## income:yers -0.021 -0.425  0.164
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code: 0
## Model failed to converge with max|grad| = 0.0692327 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```