14\_estimating\_pafs\_for\_education

Jon Minton

## Notes and Caveats

* the feend\_dv variable (in xwavedat.dta) looks inconsistent compared with hiqual\_dv
  + We need to look further at this variable and how it’s derived.

# Aim

This notebook will aim to estimate PAFs attributable to low/no education.

Variables to consider:

* feend - age left further education
* hiqual\_dv - highest qualification (will require some recoding)

## Discussion with Martin

* We are unsure where the ‘other qualification’ group fits within within a simple and meaningful hierarchy going from none, to gcse, to A-level to degree
  + We could look at making it equivalent both to gcse, or to A-level, and modelling under both assumptions?
* We could also look at age left further education?

## Data loading and prep

devtools::load\_all(here::here('R'))

ℹ Loading economic\_inactivity

base\_dir\_location <- "big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls"  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ readr::edition\_get() masks testthat::edition\_get()  
✖ dplyr::filter() masks stats::filter()  
✖ purrr::is\_null() masks testthat::is\_null()  
✖ dplyr::lag() masks stats::lag()  
✖ readr::local\_edition() masks testthat::local\_edition()  
✖ dplyr::matches() masks tidyr::matches(), testthat::matches()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(nnet)  
  
  
varnames <- c(  
 "jbstat", "dvage", "sex", 'hiqual\_dv'  
 )  
  
extract\_what <- c(  
 "labels", "values", "labels", "labels"  
 )  
  
ind\_data <- get\_ind\_level\_vars\_for\_selected\_waves(  
 varnames = varnames, vartypes = extract\_what  
)  
  
ind\_data\_highestqual <-   
 ind\_data |>   
 mutate(  
 qual\_group = case\_when(  
 hiqual\_dv %in% c("No qual", "No qualification") ~ "None",  
 hiqual\_dv %in% c("GCSE etc", "A level etc", "A-level etc", "Other qual", "Other qualification", "Other higher") ~ "Some",   
 hiqual\_dv %in% c("Degree", "Other higher degree") ~ "Degree",  
 TRUE ~ NA\_character\_  
 )  
 ) |>   
 mutate(  
 qual\_group = ordered(qual\_group, levels = c("None", "Some", "Degree"))  
 ) |>   
 mutate(  
 dvage = ifelse(dvage < 0, NA, dvage)  
 ) |>   
 rename(age = dvage) |>   
 filter(between(age, 25, 60)) %>% #As highest qualification starting at 25 not 16 years of age |>   
 filter(complete.cases(.))

Now we need to standardise the hiqual\_dv variable

ind\_data\_highestqual$hiqual\_dv |> unique()

[1] "Other higher" "GCSE etc" "A level etc"   
 [4] "Other qual" "Degree" "No qual"   
 [7] "No qualification" "Other qualification" "Other higher degree"  
[10] "A-level etc"

Now let’s model

mod\_null <- multinom(  
 next\_status ~ this\_status \* sex + splines::bs(age, 5),  
 data = ind\_data\_highestqual  
)

# weights: 238 (198 variable)  
initial value 448049.703642   
iter 10 value 178421.309710  
iter 20 value 148331.001206  
iter 30 value 134058.694236  
iter 40 value 121762.926198  
iter 50 value 113294.317291  
iter 60 value 105900.282500  
iter 70 value 101724.243604  
iter 80 value 97038.430735  
iter 90 value 94476.811055  
iter 100 value 93843.295952  
final value 93843.295952   
stopped after 100 iterations

mod\_highestqual <- multinom(  
 next\_status ~ this\_status \* sex + splines::bs(age, 5) + qual\_group,  
 data = ind\_data\_highestqual  
)

# weights: 252 (210 variable)  
initial value 448049.703642   
iter 10 value 201836.011318  
iter 20 value 176324.237601  
iter 30 value 151805.208749  
iter 40 value 137060.460419  
iter 50 value 121425.009912  
iter 60 value 111968.668436  
iter 70 value 103691.474057  
iter 80 value 99393.187434  
iter 90 value 95815.113158  
iter 100 value 93548.359467  
final value 93548.359467   
stopped after 100 iterations

Now to compare model fit

AIC(mod\_null, mod\_highestqual)

df AIC  
mod\_null 126 187938.6  
mod\_highestqual 138 187372.7

BIC(mod\_null, mod\_highestqual)

df BIC  
mod\_null 126 189242.3  
mod\_highestqual 138 188800.6

Having grouped the qualifications into three simple ordered categories, AIC and BIC now indicate the variable leads to improved fit.

## Indicative vignettes

Let’s start by thinking of some demographic groups, and assigning them either none, some or degree as their qualification level.

Let’s assume they start off either Employed or Unemployed

illustrative\_populations <-   
 expand\_grid(  
 age = seq(25, 55, by = 5),  
 sex = c("male", "female"),  
 this\_status = c("Employed", "Unemployed", "Inactive care"),  
 qual\_group = c("None", "Some", "Degree")  
 ) |>   
 mutate(  
 qual\_group = ordered(qual\_group, levels = c("None", "Some", "Degree"))  
 )

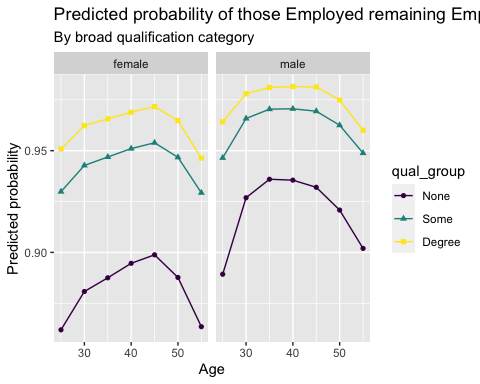
Now to get predictions for each of the above

predictions <- predict(mod\_highestqual, newdata = illustrative\_populations,   
 type = "probs")  
  
predictions\_predictors <- bind\_cols(illustrative\_populations, predictions) |>   
 pivot\_longer(cols = c("Employed":"Unemployed"), names\_to = "predicted\_next\_status", values\_to = "predicted\_probability")

First visualisation, for males and females who start off employed, what’s the estimated effect of different levels of qualification on remaining employed?

predictions\_predictors |>   
 filter(this\_status == "Employed") |>   
 filter(predicted\_next\_status == "Employed") |>   
 ggplot(aes(x = age, y = predicted\_probability, shape = qual\_group, colour = qual\_group)) +   
 facet\_wrap(~sex) +   
 geom\_point() +   
 geom\_line() +  
 labs(  
 x = "Age",   
 y = "Predicted probability",  
 title = "Predicted probability of those Employed remaining Employed",  
 subtitle = "By broad qualification category"  
 )

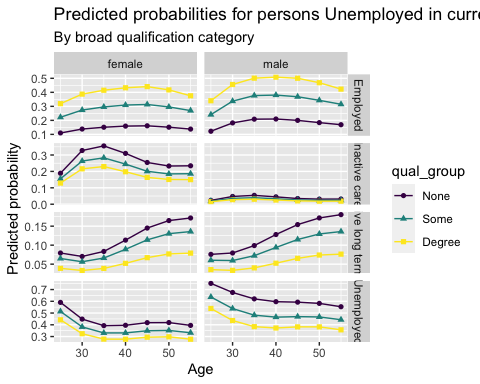
Warning: Using shapes for an ordinal variable is not advised



What about from unemployment to employment, and from unemployment to either long-term sick or carer?

predictions\_predictors |>   
 filter(this\_status == "Unemployed") |>   
 filter(predicted\_next\_status %in% c("Employed", "Unemployed", "Inactive care", "Inactive long term sick")) |>   
 ggplot(aes(x = age, y = predicted\_probability, shape = qual\_group, colour = qual\_group)) +   
 facet\_grid(predicted\_next\_status~sex, scales = "free\_y") +   
 geom\_point() +   
 geom\_line() +   
 labs(  
 x = "Age",   
 y = "Predicted probability",  
 title = "Predicted probabilities for persons Unemployed in current wave",  
 subtitle = "By broad qualification category"  
 )

Warning: Using shapes for an ordinal variable is not advised

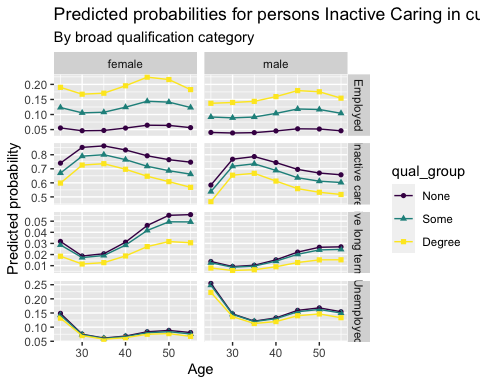


Another scenario:

Let’s consider the inactive care to employment, unemployment, long-term sick

predictions\_predictors |>   
 filter(this\_status == "Inactive care") |>   
 filter(predicted\_next\_status %in% c("Employed", "Unemployed", "Inactive care", "Inactive long term sick")) |>   
 ggplot(aes(x = age, y = predicted\_probability, shape = qual\_group, colour = qual\_group)) +   
 facet\_grid(predicted\_next\_status~sex, scales = "free\_y") +   
 geom\_point() +   
 geom\_line() +   
 labs(  
 x = "Age",   
 y = "Predicted probability",  
 title = "Predicted probabilities for persons Inactive Caring in current wave",  
 subtitle = "By broad qualification category"  
 )

Warning: Using shapes for an ordinal variable is not advised



## Estimated effects of changing no qualifications to some qualifications

Let’s take an illustrative wave, such as j, and modify anyone who has no qualification to the some qualifications category

scenario\_baseline <-   
 ind\_data\_highestqual |>   
 filter(wave == 'j')  
  
scenario\_counterfactual <-   
 scenario\_baseline |>   
 mutate(qual\_group = as.character(qual\_group)) |>   
 mutate(qual\_group = ifelse(qual\_group == "None", "Some", qual\_group)) |>   
 mutate(qual\_group = ordered(qual\_group, levels = c("None", "Some", "Degree")))

distribution\_results <-   
 calculate\_baseline\_counterfactual\_distribution(  
 m = mod\_highestqual,  
 d\_base = scenario\_baseline,   
 d\_counter = scenario\_counterfactual  
 )  
  
distribution\_results

# A tibble: 28 × 4  
 output scenario state value  
 <chr> <chr> <chr> <dbl>  
 1 absolute baseline Employed 13046.   
 2 absolute counterfactual Employed 13085.   
 3 absolute baseline Inactive care 947.   
 4 absolute counterfactual Inactive care 929.   
 5 absolute baseline Inactive long term sick 738.   
 6 absolute counterfactual Inactive long term sick 723.   
 7 absolute baseline Inactive other 92.9  
 8 absolute counterfactual Inactive other 91.4  
 9 absolute baseline Inactive retired 570.   
10 absolute counterfactual Inactive retired 579.   
# ℹ 18 more rows

distribution\_results |> make\_tabular\_summary()

# A tibble: 7 × 5  
 state baseline counterfactual abs\_label rel\_label  
 <ord> <dbl> <dbl> <glue> <glue>   
1 Employed 13046. 13085. 39 more 0.3% up   
2 Unemployed 631. 616. 14 less 2.3% down  
3 Inactive long term sick 738. 723. 15 less 2.1% down  
4 Inactive student 97.7 98.8 1 more 1.1% up   
5 Inactive care 947. 929. 17 less 1.8% down  
6 Inactive retired 570. 579. 9 more 1.5% up   
7 Inactive other 92.9 91.4 1 less 1.6% down

And with a formatted table

distribution\_results |>   
 make\_tabular\_summary(  
 outputType = 'kable',   
 title = "Estimated effect of ensuring everyone has some qualifications",  
 subtitle = "Data used: Wave J of UKHLS"  
 )

Estimated effect of ensuring everyone has some qualifications

|  | Scenarios | | Comparisons | |
| --- | --- | --- | --- | --- |
|  | Baseline | Counterfactual | Absolute | Relative |
| **Active** | | | | |
| Employed | 13046 | 13085 | 39 more | 0.3% up |
| Unemployed | 631 | 616 | 14 less | 2.3% down |
| **Inactive** | | | | |
| Long-term Sick | 738 | 723 | 15 less | 2.1% down |
| Student | 98 | 99 | 1 more | 1.1% up |
| Full-time Carer | 947 | 929 | 17 less | 1.8% down |
| Retired | 570 | 579 | 9 more | 1.5% up |
| Other | 93 | 91 | 1 less | 1.6% down |
| Data used: Wave J of UKHLS |  |  |  |  |

What if everyone had a degree!?

scenario\_baseline <-   
 ind\_data\_highestqual |>   
 filter(wave == 'j')  
  
scenario\_counterfactual\_degree <-   
 scenario\_baseline |>   
 mutate(qual\_group = as.character(qual\_group)) |>   
 mutate(qual\_group = "Degree") |>   
 mutate(qual\_group = ordered(qual\_group, levels = c("None", "Some", "Degree")))

calculate\_baseline\_counterfactual\_distribution(  
 m = mod\_highestqual,  
 d\_base = scenario\_baseline,   
 d\_counter = scenario\_counterfactual\_degree  
) |>   
 make\_tabular\_summary(  
 outputType = 'kable',   
 title = "Estimated effects if everyone had activity propensities of those with degrees",  
 subtitle = "Data used: Wave J of UKHLS"   
 )

Estimated effects if everyone had activity propensities of those with degrees

|  | Scenarios | | Comparisons | |
| --- | --- | --- | --- | --- |
|  | Baseline | Counterfactual | Absolute | Relative |
| **Active** | | | | |
| Employed | 13046 | 13291 | 245 more | 1.9% up |
| Unemployed | 631 | 538 | 92 less | 14.6% down |
| **Inactive** | | | | |
| Long-term Sick | 738 | 580 | 158 less | 21.4% down |
| Student | 98 | 111 | 14 more | 14% up |
| Full-time Carer | 947 | 834 | 112 less | 11.9% down |
| Retired | 570 | 659 | 89 more | 15.6% up |
| Other | 93 | 107 | 14 more | 15.5% up |
| Data used: Wave J of UKHLS |  |  |  |  |

This appears to have a more substantive effect, as expected.

Let’s now look at the effect of moving everyone up one level (except degree, as they’re already at the top level)

scenario\_counterfactual\_up\_one <-   
 scenario\_baseline |>   
 mutate(qual\_group = as.character(qual\_group)) |>   
 mutate(qual\_group = ifelse(qual\_group == "Some", "Degree", qual\_group)) |>  
 mutate(qual\_group = ifelse(qual\_group == "None", "Some", qual\_group)) |>   
 mutate(qual\_group = ordered(qual\_group, levels = c("None", "Some", "Degree")))

calculate\_baseline\_counterfactual\_distribution(  
 m = mod\_highestqual,  
 d\_base = scenario\_baseline,   
 d\_counter = scenario\_counterfactual\_up\_one  
) |> make\_tabular\_summary(  
 outputType = 'kable',   
 title = "Estimated effect if everyone's qualification class moved up one level (up to degree)",  
 subtitle = "Data used: Wave J of UKHLS"   
 )

Estimated effect if everyone's qualification class moved up one level (up to degree)

|  | Scenarios | | Comparisons | |
| --- | --- | --- | --- | --- |
|  | Baseline | Counterfactual | Absolute | Relative |
| **Active** | | | | |
| Employed | 13046 | 13265 | 219 more | 1.7% up |
| Unemployed | 631 | 545 | 85 less | 13.5% down |
| **Inactive** | | | | |
| Long-term Sick | 738 | 610 | 128 less | 17.4% down |
| Student | 98 | 110 | 13 more | 12.8% up |
| Full-time Carer | 947 | 848 | 99 less | 10.5% down |
| Retired | 570 | 640 | 70 more | 12.2% up |
| Other | 93 | 104 | 11 more | 12.4% up |
| Data used: Wave J of UKHLS |  |  |  |  |