22 Effect of wages on remaining employed

## Aims

The aims of this notebook are to understand the effect of wages on remaining employed.

The variables to look at are:

* employees\_w12\_paygl: gross pay at last payment
* employees\_w12\_paynl: take home pay at last payment
* seearngrs\_dv: self-employment earnings - gross
* seearnnet\_dv: self-employment earnings - net

We are predominantly interested in the extent to which different levels of the above affect probability of those employed remaining employed at the next wave.

## Preparation

We are not yet sure whether the derived wages variables includes second jobs.

Let’s now look at the range of take-home pay by wave for those employed

# A tibble: 11 × 4  
 wave lower\_q median\_q upper\_q  
 <chr> <dbl> <dbl> <dbl>  
 1 a 882. 1508 2481.  
 2 b 917. 1583 2500   
 3 c 932 1600 2500   
 4 d 953. 1625 2583.  
 5 e 990 1667. 2607   
 6 f 1000 1700 2667.  
 7 g 1018. 1733. 2708.  
 8 h 1080 1772. 2800   
 9 i 1090. 1800 2833.  
10 j 1167. 1900 2968   
11 k 1233. 2000 3035.

Because the median and distribution of wages changed with each wave, in order to make use of all waves worth of data as predictors of remaining employment, we should probably look at the effect of wage-specific quantiles on remaining employed

Now let’s start with the simplest model, just predicting whether next\_wave is still employed

Call:  
glm(formula = not\_employed ~ z\_payg\_dv, family = binomial(link = "logit"),   
 data = df\_ind\_wages\_normalised)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-0.7429 -0.4677 -0.3053 -0.2005 2.9981   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -1.14626 0.01511 -75.87 <2e-16 \*\*\*  
z\_payg\_dv -3.33670 0.03963 -84.19 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 90359 on 163615 degrees of freedom  
Residual deviance: 81514 on 163614 degrees of freedom  
AIC: 81518  
  
Number of Fisher Scoring iterations: 6

The more relative pay people in employment get, the lower the probability of moving from employment becomes, and this term is statistically significant.

What does this mean substantively?

1 2 3 4 5   
0.14020359 0.07720525 0.05653982 0.04115915 0.02154949

For someone earning less than 80% of earners, the probability of no longer being employed in the next wave is 14%. For someone earning more than 80% of earners, the probability of no longer being employed in the next wave is 2%. For someone earning the median amount, the probability of not being employed in the next wave is 5.6%.

Now let’s add the standard controls

Call:  
glm(formula = not\_employed ~ sex + splines::bs(age, 5) + z\_payg\_dv,   
 family = binomial(link = "logit"), data = df\_ind\_wages\_normalised)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.5968 -0.3699 -0.2712 -0.1980 3.0013   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.92840 0.04069 22.815 <2e-16 \*\*\*  
sexmale 0.01957 0.02161 0.906 0.365   
splines::bs(age, 5)1 -3.06056 0.09541 -32.078 <2e-16 \*\*\*  
splines::bs(age, 5)2 -3.19548 0.07231 -44.194 <2e-16 \*\*\*  
splines::bs(age, 5)3 -3.72907 0.09070 -41.115 <2e-16 \*\*\*  
splines::bs(age, 5)4 -2.47387 0.07098 -34.854 <2e-16 \*\*\*  
splines::bs(age, 5)5 -1.07421 0.06181 -17.378 <2e-16 \*\*\*  
z\_payg\_dv -1.98243 0.04503 -44.027 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 90359 on 163615 degrees of freedom  
Residual deviance: 73228 on 163608 degrees of freedom  
AIC: 73244  
  
Number of Fisher Scoring iterations: 6

Once age is controlled for in the standard way, the effect of sex becomes non-significant.

Let’s compare the models

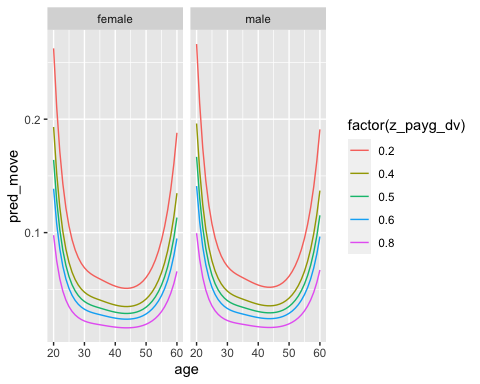
df AIC  
mod\_00 2 81518.31  
mod\_01 8 73244.14

df BIC  
mod\_00 2 81538.32  
mod\_01 8 73324.18

The more complex model is still preferred

Let’s work out what mod\_01 is implying

Let’s see what this shows:

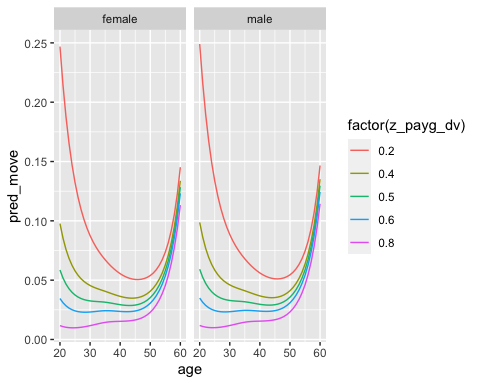


The model predicts a strongly U-shaped relationship with age. There is no interaction between age and quantile. We could look at that next

df AIC  
mod\_00 2 81518.31  
mod\_01 8 73244.14  
mod\_02 13 72006.08

df BIC  
mod\_00 2 81538.32  
mod\_01 8 73324.18  
mod\_02 13 72136.15

The model with interactions between age and relative wage is preferred over the next most complex model. As before, let’s look at what the model predicts



Comparatively wages have a very strong influence on probability of remaining employed at younger adult ages. These influences diminish with age. From the early to mid 50s the probability of leaving employment increases regardless of comparative wages.

There are a couple more modifications to consider:

* Include an interaction between sex and comparative wages (for example, are women more likely to exit employment if they receive low wages than men, or vice versa?)
* Move to multinomial logistic regression, to model where people move if they move from employment

Let’s start with the sex interaction term

df AIC  
mod\_00 2 81518.31  
mod\_01 8 73244.14  
mod\_02 13 72006.08  
mod\_03 14 72008.03

df BIC  
mod\_00 2 81538.32  
mod\_01 8 73324.18  
mod\_02 13 72136.15  
mod\_03 14 72148.10

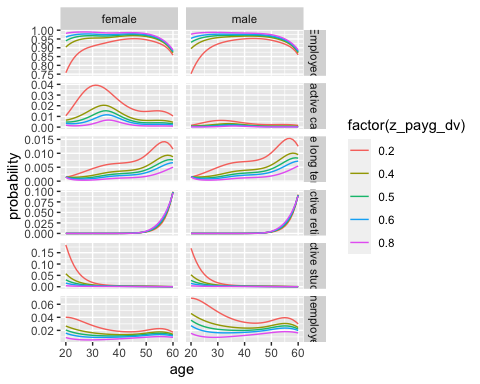
The model with a sex:wage interaction is *not* preferred to the model without such terms.

Now let’s run the multinomial logit model:

# weights: 98 (78 variable)  
initial value 318382.034948   
iter 10 value 60337.114215  
iter 20 value 53099.169011  
iter 30 value 49280.455396  
iter 40 value 48155.231702  
iter 50 value 47801.646126  
iter 60 value 47358.005159  
iter 70 value 47306.211706  
iter 80 value 47260.578908  
iter 90 value 47073.166846  
iter 100 value 47035.764377  
final value 47035.764377   
stopped after 100 iterations

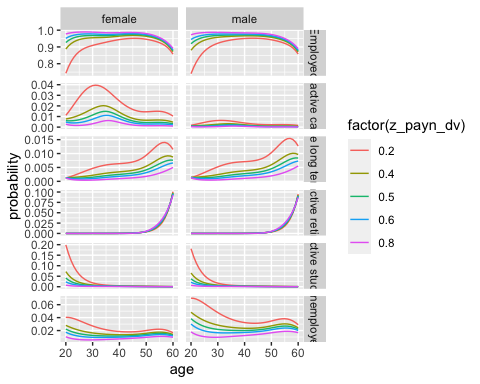
Let’s now see what it predicts:

Now to visualise it, focusing on Unemployment, Inactive care, Inactive long term sick, Inactive student, and Inactive retired



Let’s look at the equivalent for net

# weights: 98 (78 variable)  
initial value 318382.034948   
iter 10 value 60451.180405  
iter 20 value 53017.520478  
iter 30 value 49595.894512  
iter 40 value 48231.978397  
iter 50 value 47910.906606  
iter 60 value 47429.744570  
iter 70 value 47397.698092  
iter 80 value 47350.414975  
iter 90 value 47174.938954  
iter 100 value 47120.210532  
final value 47120.210532   
stopped after 100 iterations



I don’t think there’s any substantive difference whether using gross or net pay. For consistency I’m going to keep using gross pay for the counterfactual

## Counterfactual Simulation

Let’s model the following counterfactual:

* If someone is earning at at least the 40th percentile, their pay stays as is;
* If someone is earning below the 40th percentile, their pay is moved up to the 40th percentile

base counterfactual  
Employed 10882.48087 11230.64693  
Inactive care 89.46711 52.04180  
Inactive long term sick 49.18924 38.07907  
Inactive other 26.78770 22.78726  
Inactive retired 229.76486 227.66949  
Inactive student 352.69113 95.13879  
Unemployed 234.61909 198.63665

base counterfactual  
Employed 100 103.19933  
Inactive care 100 58.16864  
Inactive long term sick 100 77.41343  
Inactive other 100 85.06615  
Inactive retired 100 99.08804  
Inactive student 100 26.97510  
Unemployed 100 84.66346

In this scenario, the proportion of those employed who remain employed is expected to increase by around 3%, the proportion transitioning from employment to full-time care to fall by around 40%, from employment to long-term sick to fall by around 22%, and to unemployment by around 15%. Many of these transitions from employment are quite unusual; instead such transitions may occur over multiple waves, often via unemployment through an intermediate state. The following table shows the proportion of the population projected to move into each state in both the baseline and counterfactual scenario:

base counterfactual  
Employed 91.7191814 94.6535772  
Inactive care 0.7540422 0.4386161  
Inactive long term sick 0.4145743 0.3209361  
Inactive other 0.2257707 0.1920545  
Inactive retired 1.9364927 1.9188326  
Inactive student 2.9725337 0.8018440  
Unemployed 1.9774049 1.6741395

So, in the baseline scenario around 92% of those employed in wave T are predicted to remain employed at wave T+1, around 3% to become a student at wave T+1, around 2% to become retired, or to become unemployed. In the counterfactual scenario the proportion remaining employed is projected to increase retention of employment by 2.9%.

## Notes from Martin

The function ecdf converts the real wages into a value between 0 and 1, where e.g. 0.2 is 20% up the wages distribution, with 80% of waged employees earning more than them.

This allows us to compare relative position across all the waves without adjusting wages

Model 1 - probability of moving from employment to other categories, based on age sex and wage position

We included an interaction between wages and age

But we don’t think at the broadest level sex makes a difference

This is all standard logistic regression

We don’t think we can use the Minimum Income Standard because this operates at a household income

Could a wage based intervention increase inequalities?

By having people who moved into education remaining in better paid work?

Should we give a different floor? By age  group?