# Retirement Pension and Mortality Disparities in Andalusia

*Mathias Voigt*

This document serves as summary for the research project on health inequalities in the retired population. It contains a few graphs, tables, code chunks, and figures that are not (yet) displayed in the paper/article version which will be send to the journal. As it can be updated easily, I hope it will help us stay up-to-date with the most current developments of the analysis on the effect on individual and partner level pension and come to quick decisions about how to represent the results. It can also help us to find out which variables and what way to code them might bring the best results. I will keep this document as up-to-date as possible and save changes in the dropbox immediately. May it help us to come to a result that everybody can live with.

## Part 1 - Sample

To answer how health differences in the retired population of Spain are determined by social inequalities, we took a sample from the full population sample from the BDLPA which contains 734682 cases which qualified through following characteristics:

1. Individuals registered in Andalusia in 2002 who are still alive at 2011
2. Everybody who is eligible for receiving a public pension between 2011-2016 (including disability)
3. Everybody between the age 65 and 95 - this includes everybody who will turn 65 during the observation period
4. Everybody who receives a public retirement pension at some point during the observation period

### Who do we exclude and how this affects validity?

As our data covers public pensions, there are a few population groups that might be systematically excluded. Due to the setup of the analysis (focus on the working life course), it was chosen to include only individuals with active working life histories indicated through contribution to social security. Possible selection biases occur for the following groups.

1. Very rich individuals with private pensions
2. Very poor individuals who never contributed to the public pension funds
3. Most importantly, women who took over the role as care giver or stayed home for other reasons

### Entering and Re-entering

The BDLPA follow-up study of a cohort of individuals based on the census 2001. To my knowledge, the data we have does not capture individuals who immigrated or registered after 2002, which would for example exclude more recent retirement migrants. Even if this is a large number, selection based on this measure should not introduce to many problems to our analysis as we are interested in the ones who contributed to social security in Andalusia.

## The partner data

Given the meager results from the individual level analysis (a Null finding with regard to income), it was proposed to use a sub population of the aforementioned sample which would include 202929 individuals who were married and cohabitating between 2001 and 2011 and where both partners receive some form of public pension. From a theoretical point of view, we would be able to use an approximation of the household income (both partners combined pension). The practical advantages would be that we find expected income effects for this sub-population. **I would like to use exclusively this subsample of the population for the following analysis despite the flaws (for instance the selecitivity for women) and the extra work of changing the paper.**

# Further decisions on selection of the population

## Left Truncation

This is not problematic but to mention it, the data is left-truncated. Starting the observation at 65 introduces a survivor bias. The choice to that can however be justified by a) our interest in the retired population and b) that we do not have income information for earlier ages (only for widows and disabled individuals).

## Time Lag

The time lag between two observations might be a problem for the census based variables which are additionally used to measure the sociology-economic position of individuals. This would include the ownership of cars and housing. Given the age of the observed individuals, it is in my opinion unlikely that there were large-scale changes with regard to these variables, however, a ten year time lag in a period of economic fluctuations can add some bias i.e. there might be individuals who owned a car in 2001 cannot drive anymore in 2011 or could not replace it. It will also be a strong assumption, as you have mentioned before, to say that there were no residential changes at older ages which could affect the house ownership status. As these are indirect measures for wealth, I assume (and this is written in the paper right now) that the distributions have not changed that much over for the population of interest.

## Disability

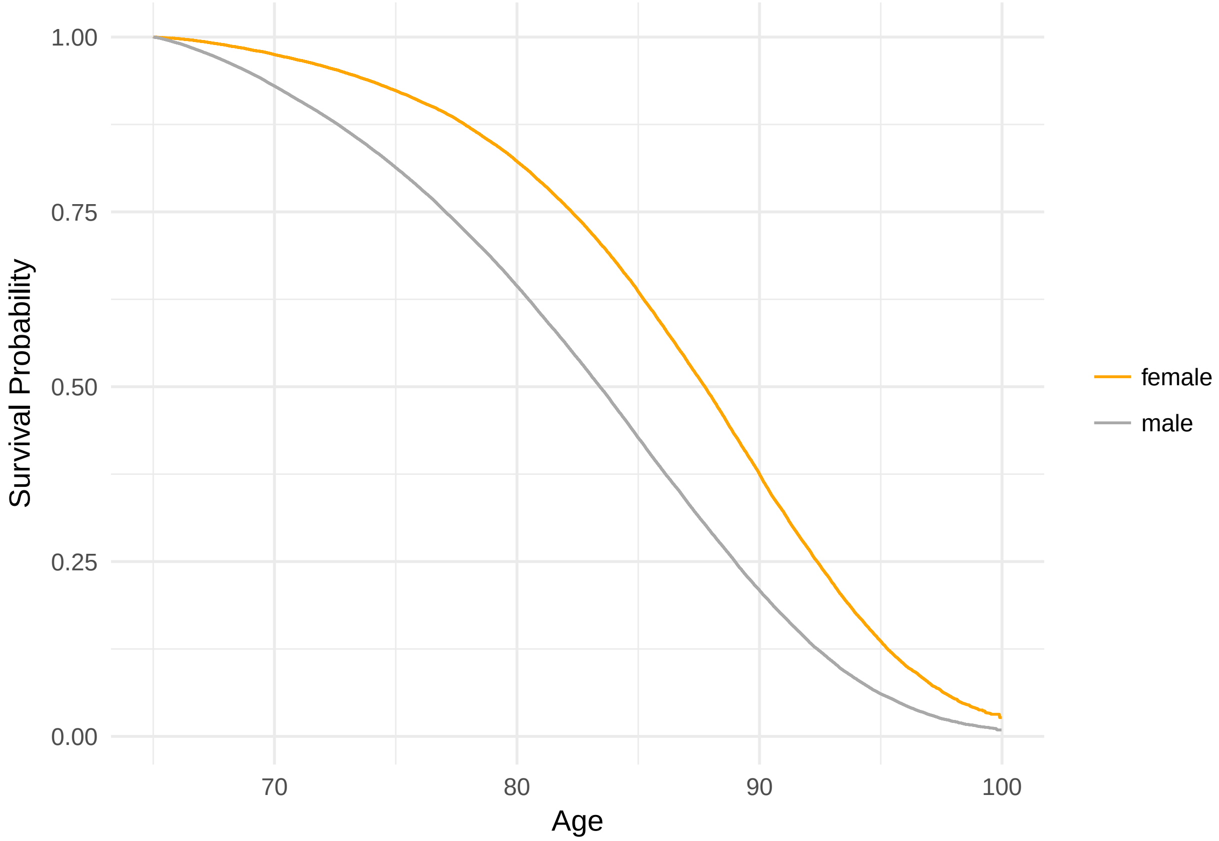
The difficulties to distinguish the disabled early retirees from the “normal” retirees gives me the most headaches. The way the data (and the law) is structured, it is to my knowledge impossible to disentangle what part of the retirement pension of a former disabled after age 65 is based on his/her contribution to social security and what part is because of his/her disability. The solution for now is to include only individuals age 65 and older (including probably severely disabled) and control if they have received a disability pension before their 65-th birthday. This leads to the following problem:

*italics*The assumption that higher pension income indicates a higher socioeconomic status throughout life as indicator for life time contribution to social security will not hold for the disabled population where higher “income” is probably related to more severity of the disability. This will lead to the paradox that the highest paid individuals in this substantial sub-population have the highest chances to die first.

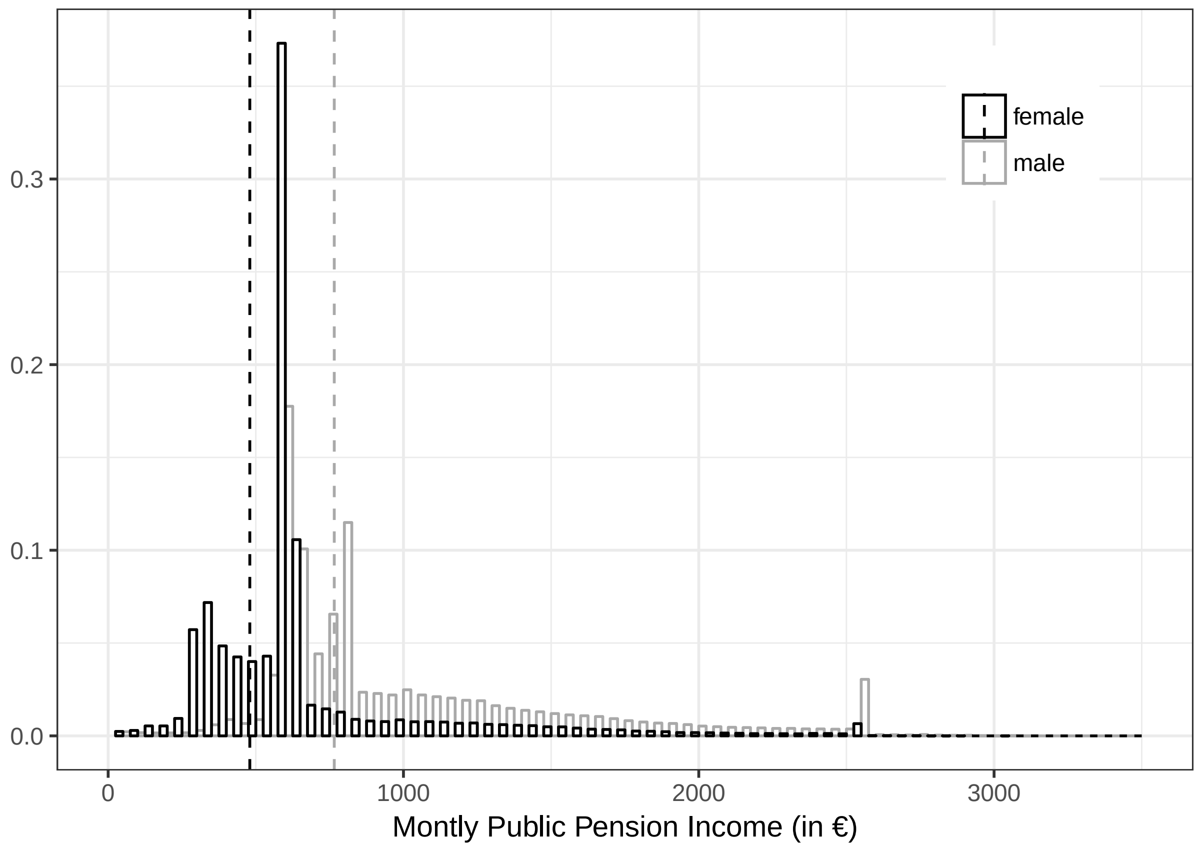
Fran suggested to run a stratified model (disabled/not disabled) which would be one possibility to account for the issue. In this case we might have to be careful in how we explain the decision to treat all disabled individuals as different group. Another possibility is to use a frailty model. As we have theoretically no idea about the severness of the disability, we could use a random effect to account for this unobserved heterogeneity. **I think that this is a major road block and I would like to discuss again how to deal with it the best way**.

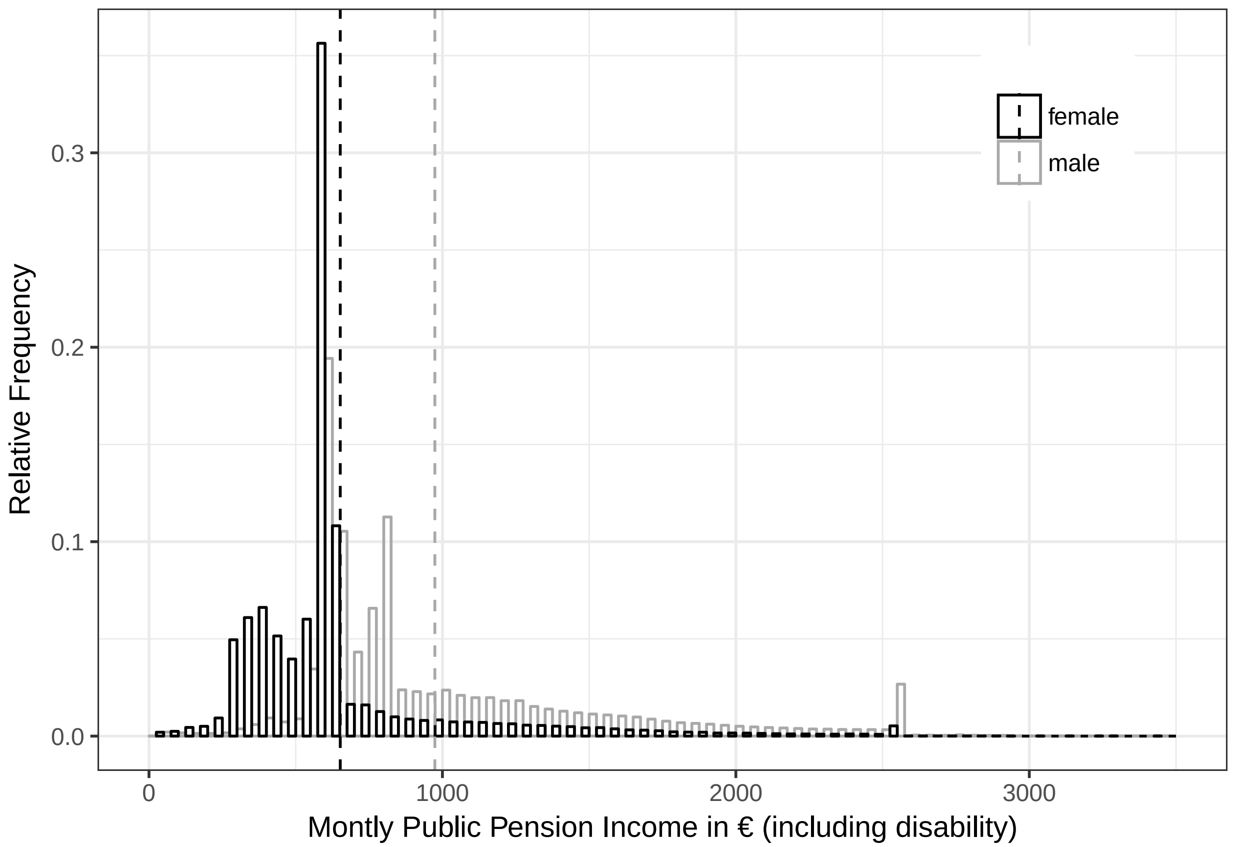
### Male/Female survival and income differences

As we know from the paper presented at EPC 2016, Andalusian women have a survival advantage over men, at least in the generations covered by the BDLPA. This results in more "rectangular" survival curve (here for the full population) and less variation in the ages at death. These differences in LE and life span disparity are probably part of the explanation for the results we get for the individual level relationship between income and survival. There are less female deaths during the observation period and they also seem to be less spread.



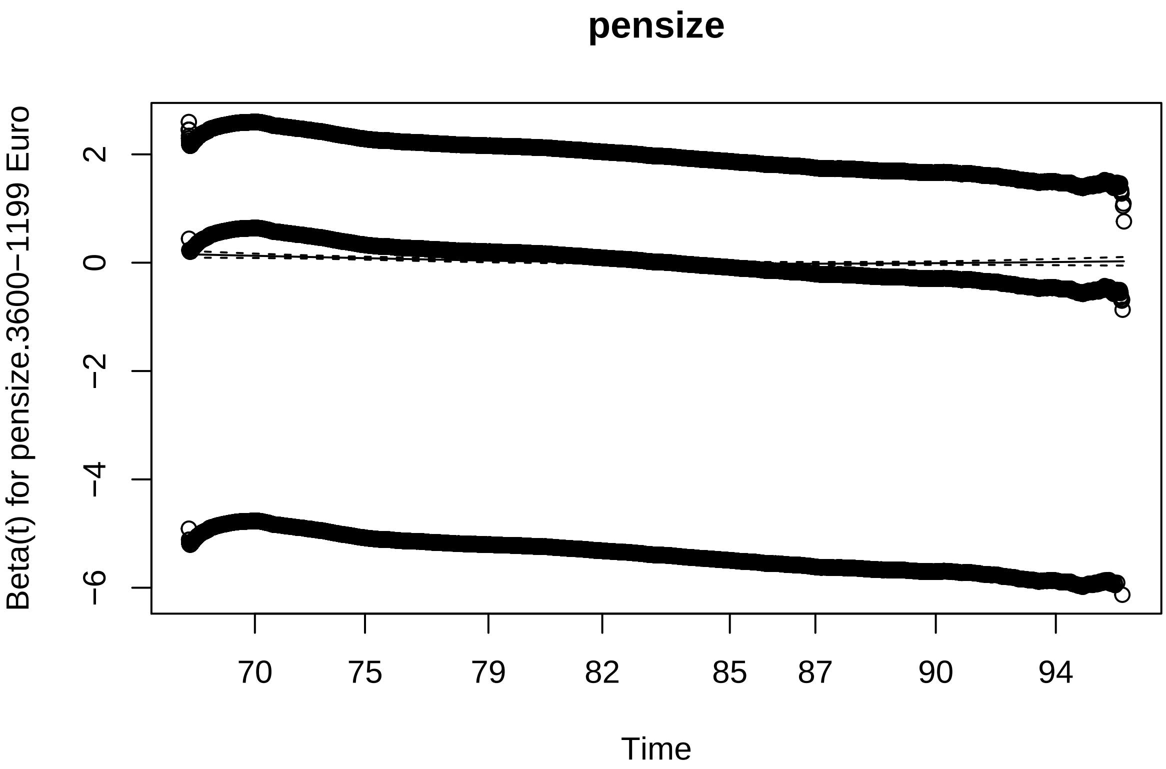
We also know that there are substantial differences in the labor force participation, average salaries and therefore contribution to social security between men and women. The driving force behind our findings on female survival, in my opinion, lies in the distributions of incomes and how the variable enters the model. Even if we select only women who have payed into social security, there incomes are skewed to the lower end of the income distribution and more compressed than mens incomes. The plots show the relative income distribution for both sexes, the first plot just for retirement pensions, the second also including disability allowance.


female
male
0.0
0.1
0.2
0.3
0
1000
2000
3000
Montly Public Pension Income (in €)

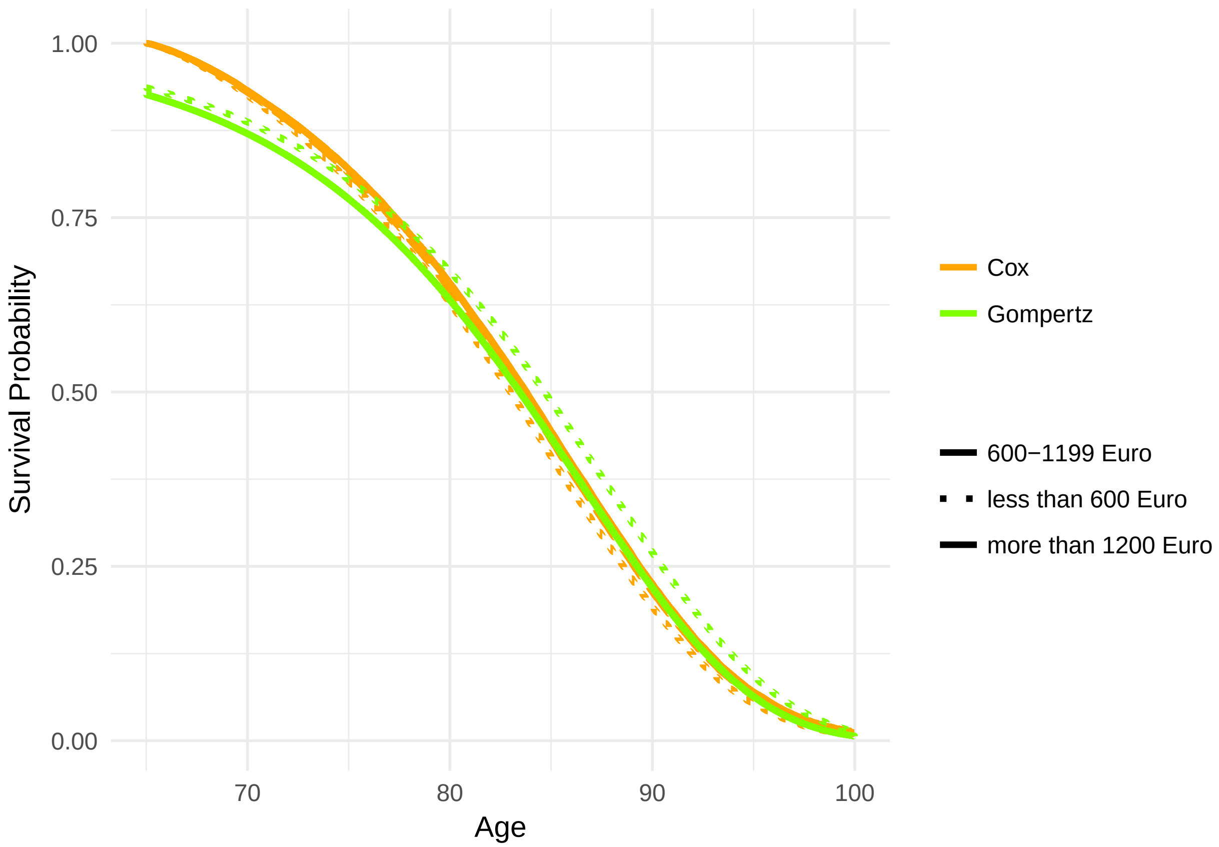


## Part 2 - Model and variable choices

So far we were using Cox models to estimate the survival differences between different income groups and other variables. The comments from the LONGPOP master class made me think if it might be more accurate to use a parametric survival model with a Gompertz baseline distribution. It was repeatedly shown that human adult mortality follows a Gompertz distribution and we would not need to make assumptions about proportional hazards or else. I ran a few tests and found that indeed fit the data better than the Cox model. It would not be an extensive amount of work to change it. Indeed it would not be more than an afternoon of work. I just show a few results for males comparing models with a categorical income variable, the same income variable we have used so far.

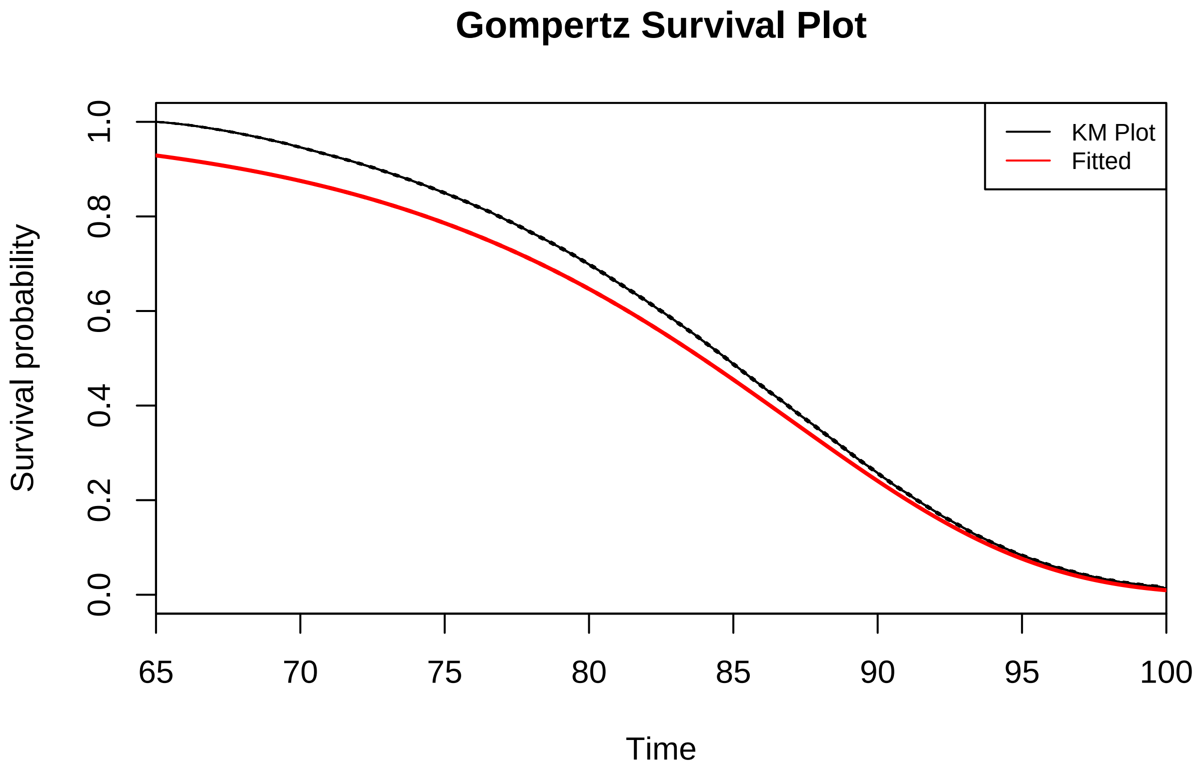


```
##   Distribution     AIC
## 2     Gompertz 1030952
## 1          Cox 2215633
```



I ran a few tests with other parametric distributions which confirmed that the Gompertz curve fits the data best (for more model comparison see code file: 032_ModelTests.R).

```
##   Distribution     AIC
## 3     gompertz 1031670
## 2        lnorm 1038268
## 1          exp 1135792
```



## Measuring income

Another change I would like to propose is the way pension income is measured. Until now, we use a categorical income variable based on a mix of pension and disability income after age 65 (which mentioned earlier cannot be really told apart). We tried to different categorical variables and for both the highest income is relatively small and spread (values in percent, 0-1 corresponds to the proportions who experienced the event).

```
##                          0    1
## more than 2000 Euro  4.6  0.6
## 1000-1999 Euro      16.0  3.4
## 650-999 Euro        19.7  4.4
## less than 650 Euro  40.7 10.5
```

```
##                          0   1
## more than 1200 Euro 15.4 2.9
## 600-1199 Euro       34.4 8.2
## less than 600 Euro  31.3 7.8
```

Inspired by George Alters works on the Pensilvanian train company, I played around with a few different income measures. It turned out that using the log-pension income instead of the categories improves the fit of the models substantially. This choice, apparently often used in economics (see for instance the discussion here) comes with at least two further advantages. The higher the pension income the less will add an additional Euro a month, so there will not be so much weight on the few individuals with higher incomes. Secondly, we avoid grouping individuals by arbitrary thresholds. We can discuss the interpretation, but first statistical test also confirm the improvement in fit (here with the full Cox model for the individual analysis).

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ pensize + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ pensize.3 + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##     loglik  Chisq Df P(>|Chi|)
## 1 -1105415
## 2 -1105392 44.686  1 2.313e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ pensize.3 + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ INCOME + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##     loglik  Chisq Df P(>|Chi|)
## 1 -1105392
## 2 -1105434 83.117  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ INCOME + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ log(INCOME) + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##     loglik  Chisq Df P(>|Chi|)
## 1 -1105434
## 2 -1105434 0.7254  0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ pensize + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ pensize.3 + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##    loglik  Chisq Df P(>|Chi|)
## 1 -364628
## 2 -364648 40.024  1 2.508e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ pensize.3 + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ INCOME + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##    loglik  Chisq Df P(>|Chi|)
## 1 -364648
## 2 -364629 38.817  1 4.655e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##  Cox model: response is  Surv(time = entry.age.r, time2 = exit.age, event = event)
##  Model 1: ~ INCOME + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##  Model 2: ~ log(INCOME) + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status + hh
##    loglik Chisq Df P(>|Chi|)
## 1 -364629
## 2 -364635 11.99  0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Call:
## coxph(formula = Surv(time = entry.age.r, time2 = exit.age, event = event) ~
##     log(INCOME) + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status +
##         hh, data = subset(retire, SEXO == "male"))
##
##   n= 478921, number of events= 102033
##
##                              coef exp(coef)  se(coef)       z Pr(>|z|)
## log(INCOME)              0.018696  1.018872  0.007582   2.466  0.01367 *
## ESREAL5Tertiary Educ.   -0.055474  0.946037  0.016915  -3.280  0.00104 **
## ESREAL5Secondary Educ.  -0.028458  0.971943  0.009678  -2.941  0.00328 **
## ESREAL5Primary Educ.    -0.018280  0.981886  0.007676  -2.381  0.01725 *
## mobilno car available    0.199027  1.220215  0.006938  28.687  < 2e-16 ***
## HousRegowned            -0.135397  0.873369  0.010402 -13.016  < 2e-16 ***
## FNAC                     0.012296  1.012372  0.002210   5.563 2.65e-08 ***
## DIS                      0.347618  1.415691  0.006778  51.285  < 2e-16 ***
## civil.statusother forms  0.182733  1.200493  0.007605  24.027  < 2e-16 ***
## civil.statuswidowed      0.159867  1.173355  0.011656  13.716  < 2e-16 ***
## hhwith partner only     -0.055174  0.946321  0.006846  -8.059 7.77e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## log(INCOME)                1.0189     0.9815    1.0038    1.0341
## ESREAL5Tertiary Educ.      0.9460     1.0570    0.9152    0.9779
## ESREAL5Secondary Educ.     0.9719     1.0289    0.9537    0.9906
## ESREAL5Primary Educ.       0.9819     1.0184    0.9672    0.9968
## mobilno car available      1.2202     0.8195    1.2037    1.2369
## HousRegowned               0.8734     1.1450    0.8557    0.8914
## FNAC                       1.0124     0.9878    1.0080    1.0168
## DIS                        1.4157     0.7064    1.3970    1.4346
## civil.statusother forms    1.2005     0.8330    1.1827    1.2185
## civil.statuswidowed        1.1734     0.8523    1.1469    1.2005
## hhwith partner only        0.9463     1.0567    0.9337    0.9591
##
## Concordance= 0.574  (se = 0.001 )
## Rsquare= 0.01   (max possible= 0.99 )
## Likelihood ratio test= 4895  on 11 df,   p=0
## Wald test            = 5082  on 11 df,   p=0
## Score (logrank) test = 5123  on 11 df,   p=0
```

```
## Call:
## coxph(formula = Surv(time = entry.age.r, time2 = exit.age, event = event) ~
##     log(INCOME) + ESREAL5 + mobil + HousReg + FNAC + DIS + civil.status +
##         hh, data = subset(retire, SEXO == "female"))
##
##   n= 255761, number of events= 36747
##
##                              coef exp(coef)  se(coef)      z Pr(>|z|)
## log(INCOME)              0.089890  1.094054  0.013195  6.812 9.60e-12 ***
## ESREAL5Tertiary Educ.   -0.192380  0.824993  0.036295 -5.300 1.16e-07 ***
## ESREAL5Secondary Educ.  -0.129737  0.878326  0.018845 -6.885 5.80e-12 ***
## ESREAL5Primary Educ.    -0.078871  0.924159  0.012944 -6.093 1.11e-09 ***
## mobilno car available    0.059622  1.061435  0.011663  5.112 3.19e-07 ***
## HousRegowned            -0.078012  0.924953  0.016635 -4.690 2.74e-06 ***
## FNAC                     0.020995  1.021217  0.003693  5.685 1.31e-08 ***
## DIS                      0.331670  1.393293  0.010662 31.109  < 2e-16 ***
## civil.statusother forms  0.161487  1.175257  0.014647 11.025  < 2e-16 ***
## civil.statuswidowed      0.141835  1.152386  0.016161  8.776  < 2e-16 ***
## hhwith partner only     -0.004631  0.995380  0.011311 -0.409    0.682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## log(INCOME)                1.0941     0.9140    1.0661    1.1227
## ESREAL5Tertiary Educ.      0.8250     1.2121    0.7683    0.8858
## ESREAL5Secondary Educ.     0.8783     1.1385    0.8465    0.9114
## ESREAL5Primary Educ.       0.9242     1.0821    0.9010    0.9479
## mobilno car available      1.0614     0.9421    1.0374    1.0860
## HousRegowned               0.9250     1.0811    0.8953    0.9556
## FNAC                       1.0212     0.9792    1.0139    1.0286
## DIS                        1.3933     0.7177    1.3645    1.4227
## civil.statusother forms    1.1753     0.8509    1.1420    1.2095
## civil.statuswidowed        1.1524     0.8678    1.1165    1.1895
## hhwith partner only        0.9954     1.0046    0.9736    1.0177
##
## Concordance= 0.572  (se = 0.002 )
## Rsquare= 0.005   (max possible= 0.943 )
## Likelihood ratio test= 1354  on 11 df,   p=0
## Wald test            = 1370  on 11 df,   p=0
## Score (logrank) test = 1380  on 11 df,   p=0
```

These are just the results for the Log-likelihood ratio test for one model (Cox model with all covariates), but so far it can be confirmed that applying the logarithm of the income leads to a better model fit. I would be happy to discuss that! Although the fit is improved, the results are still not as expected (at least not by the theory). For both sub-populations there seem to be a negative effect of income on survival once all covariates are included. ## Number of children in the household instead of household size This is just a suggestion to change this covariate in the full model or make a mix between the variables because it will help me to explain the situation a little bit better.

# Part 3 - The final model how I would use it

This is just to show how I would imagine the final model we present in the paper for now. It just touches on the partner data and therefore has a few more variables than the individual table and substantially less individuals. I have to admit that the search for the right model might not be over but in my opinion, we are getting closer. What do you think?

```
## Call:
## flexsurvreg(formula = Surv(time = entry.age.r, time2 = exit.age,
##     event = event) ~ log(hhincome) + SEXO + ESREAL5 + mobil +
##     HousReg + p.surv + DIS + log(FNAC) + DIS_p + ESREAL5_p +
##     hijo + bw, data = pen.coupl, dist = "gompertz")
##
## Estimates:
##                                 data mean  est        L95%       U95%
## shape                                  NA   1.99e-01   1.88e-01   2.10e-01
## rate                                   NA   1.96e-01   8.94e-73   4.28e+70
## log(hhincome)                    7.18e+00  -1.82e+00  -1.86e+00  -1.78e+00
## SEXOmale                         5.00e-01   1.36e+00   1.30e+00   1.41e+00
## ESREAL5No or Incomplete Educ.    5.47e-01   3.74e-02  -8.33e-02   1.58e-01
## ESREAL5Secondary Educ.           1.50e-01   4.67e-01   3.54e-01   5.81e-01
## ESREAL5Primary Educ.             2.73e-01   3.77e-02  -8.04e-02   1.56e-01
## mobilno car available            2.68e-01  -1.14e-01  -1.49e-01  -7.85e-02
## HousRegNot owned                 6.67e-02   1.13e-01   4.79e-02   1.79e-01
## p.survwidowed                    1.18e-01  -2.17e+00  -2.24e+00  -2.09e+00
## DIS                              2.59e-01   1.03e+00   1.00e+00   1.07e+00
## log(FNAC)                        7.57e+00  -8.79e-01  -2.25e+01   2.07e+01
## DIS_p                            2.59e-01   9.35e-02   5.95e-02   1.28e-01
## ESREAL5_pTertiary Educ.          2.99e-02   1.59e+00   1.50e+00   1.69e+00
## ESREAL5_pSecondary Educ.         1.50e-01  -1.01e+00  -1.11e+00  -9.14e-01
## ESREAL5_pPrimary Educ.           2.73e-01  -4.57e-01  -5.15e-01  -3.98e-01
## hijoOnly Partner                 4.14e-01   2.13e-02  -1.66e-02   5.92e-02
## hijoTwo or more children in hh   2.81e-01  -3.14e-01  -3.75e-01  -2.54e-01
## bwbreadwinner                    1.61e-01   1.33e+00   1.29e+00   1.36e+00
##                                 se         exp(est)   L95%       U95%
## shape                            5.70e-03         NA         NA         NA
## rate                             1.64e+01         NA         NA         NA
## log(hhincome)                    2.19e-02   1.62e-01   1.55e-01   1.69e-01
## SEXOmale                         2.69e-02   3.88e+00   3.68e+00   4.09e+00
## ESREAL5No or Incomplete Educ.    6.16e-02   1.04e+00   9.20e-01   1.17e+00
## ESREAL5Secondary Educ.           5.79e-02   1.60e+00   1.42e+00   1.79e+00
## ESREAL5Primary Educ.             6.02e-02   1.04e+00   9.23e-01   1.17e+00
## mobilno car available            1.81e-02   8.92e-01   8.61e-01   9.24e-01
## HousRegNot owned                 3.33e-02   1.12e+00   1.05e+00   1.20e+00
## p.survwidowed                    3.79e-02   1.14e-01   1.06e-01   1.23e-01
## DIS                              1.68e-02   2.81e+00   2.72e+00   2.91e+00
## log(FNAC)                        1.10e+01   4.15e-01   1.73e-10   9.93e+08
## DIS_p                            1.73e-02   1.10e+00   1.06e+00   1.14e+00
## ESREAL5_pTertiary Educ.          4.82e-02   4.92e+00   4.48e+00   5.41e+00
## ESREAL5_pSecondary Educ.         4.89e-02   3.64e-01   3.31e-01   4.01e-01
## ESREAL5_pPrimary Educ.           2.98e-02   6.33e-01   5.97e-01   6.71e-01
## hijoOnly Partner                 1.94e-02   1.02e+00   9.83e-01   1.06e+00
## hijoTwo or more children in hh   3.08e-02   7.30e-01   6.88e-01   7.76e-01
## bwbreadwinner                    1.93e-02   3.77e+00   3.63e+00   3.91e+00
##
## N = 202929,  Events: 26663,  Censored: 176266
## Total time at risk: 878891.7
## Log-likelihood = -118103, df = 19
## AIC = 236244
```