Analyzing The Relationship Between Job Turnover And Gender Using NLS79 Data

Anisha Grover

January 30, 2023

1 Objective

In the private sector, it is common for employees to switch companies or employers for newer opportunities as well as for an increase in pay. If there is a difference in the job switching behaviour between men and women, then it will likely impact their income trajectory and may contribute to the gender wage gap that has been documented in the data. One may conjecture that women while raising a young family might value familiarity with the employer more than men of similar age and be more averse to switching jobs. At the same time many other factors also affect a change in employment. Not every change in employer is voluntary as people are fired or made to leave when a company fails. Ability may also play a role. Less able people might switch jobs more frequently as they get lower quality jobs that evict employees frequently or high ability people may switch jobs more frequently as employers compete for them by offering attractive pay packages. The objective of this analysis is to look for patterns in job switching behaviour of working age population and investigate if there are any differences by gender.

2 Data

I use the 1979 cohort of the National Longitudinal Survey of Youth (NLSY79) that tracks 12686 men and women who born in the United States during the years 1957-1964. Based on 28 rounds of survey (starting in 1979 and the latest one being in 2018), a forty year long panel data has been constructed and captures various aspects of employment and demographics of these men and women. The deidentified data is publically available here. Out of the 55000 variables available in the survey, relevant variables were selected for the analysis and transformed from the wide to long format. The long form data files are stored in the `data/interim` folder. The primary data were further transformed to create the required features using the script `code/gen_features.py`. The two data files containing features used in this analysis are stored in the `data/processed` and described below.

2.1 jobs_unique_df_data.csv

Variable	Type	Description
`Person_Id`	Integer	**Key**: Unique ID for each respondent.
`Total_Jobs`	Integer	Number of jobs held during the observed working life.
`Years_Job_History`	Integer	Number of years observed in employement data.
`Female`	Integer	1 if female, 0 if male.

Variable	Type	Description
`Gender`	Object	2 Categories: `Female` and `Male`
`Education_Ctg`	Object	4 Categories: `Less_School`, `Complete_School`,`College` and `Graduate`.
`Frac_Years_Pvt`	Float	Percentage of years employed in the private sector.
`Frac_Years_Gvt`	Float	Percentage of years employed in the government sector.
`Frac_Years_Self`	Float	Percentage of years employed in the self employment sector.

2.2 jobs_switch_df_data.csv

Variable	Type	Description
`Person_Id`	Integer	**Key**: Unique ID for each respondent.
`Calendar_Year`	Integer	**Key**: 4 digit calendar years from 1979 to 2018.
`Age`	Integer	Age of the respondent.
`Switch_Job_1`	Integer	1 if a new job is started and any job ends in the calendar year, 0 otherwise.
`Switch_Job_2`	Integer	1 if a new job is started in the calendar year, 0 otherwise.

3 Analysis

From the survey data we get the job history of each respondent which includes a list of all employers the respondent worked for as well as the starting and stopping dates for each employer. From this i have constructed two variables *Total_Jobs* and *Switch_Job_*, that will be analyzed in detail below. *Total_Jobs* counts all the employers a person worked at between the years 1979 to 2018. Greater the number of employers switched, higher will be the value of this variable. *Switch_Job_1* and *Switch_Job_2* are binary variables that are calculated for each calendar year a person is employed. The other key features are *Gender* and *Female* whose relationship with the two dependent variables will also be explored.

It is important to keep in mind that in this analysis switching a job is defined as switching an employer. Tracking occupational or position changes with the same employer is much harder in the NLS79 data due to inconsistency in the way this information is collected over different rounds.

[1]:
```
# Import modules
from pathlib import Path
import os

import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

import statsmodels.api as sm
import statsmodels.formula.api as smf
```

[2]:
```
# Find the path of this notebook and change the working directory to its parent
# folder i.e. set working directory to be 'python_sample' folder
os.chdir(Path(os.getcwd()).parents[0])
```

[3]:
```
# Run the file 'gen_features.py' to create dataframes needed for this analysis
%run ./scripts/gen_features.py
```

```
Processed data files saved.
```

[4]:
```
# Read data
jobs_unique_df = pd.read_csv('./data/processed/jobs_unique_data.csv')
jobs_switch_df = pd.read_csv('./data/processed/jobs_switch_data.csv')
```

Summarizing the *Female* variable shows that we have information on the work history of 12336 surveyed individuals out of which 49% are women. In this data women are as likely to have worked in their lifetime as men since approximately half the respondents who have ever worked are women. Note that the survey started with 12686 individuals out of which 6403 (50.5%) were categorized as men and remainiing as women.

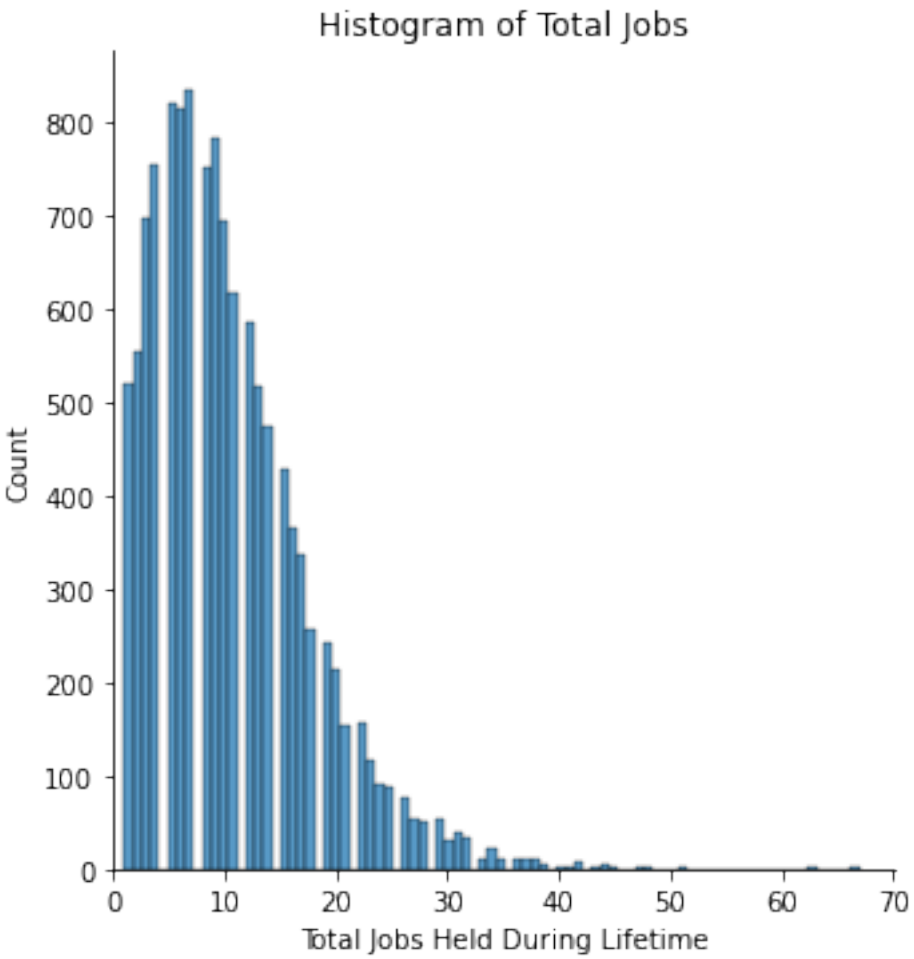
[5]:
```
# Describe sample information
jobs_unique_df['Female'].describe()
```

[5]:
```
count    12336.000000
mean         0.492218
std          0.499960
min          0.000000
25%          0.000000
50%          0.000000
75%          1.000000
max          1.000000
Name: Female, dtype: float64
```

3.1 Exploring `Total_Jobs` and its relationship to `Gender`

We start by looking at the distribution of *Total_ Jobs* that measures the total number of jobs a respondent held over their observed working lives between the years 1979 to 2018.

[6]:
```
# Historgam of Total_Jobs variable
plot01 = sns.displot(jobs_unique_df, x = "Total_Jobs")
plot01.set(xlabel = 'Total Jobs Held During Lifetime', title = 'Histogram of␣
 ↪Total Jobs')
plt.xlim(0, None)
plt.show()
```



[7]:
```
# Stats for Total_Jobs
jobs_unique_df['Total_Jobs'].describe()
```

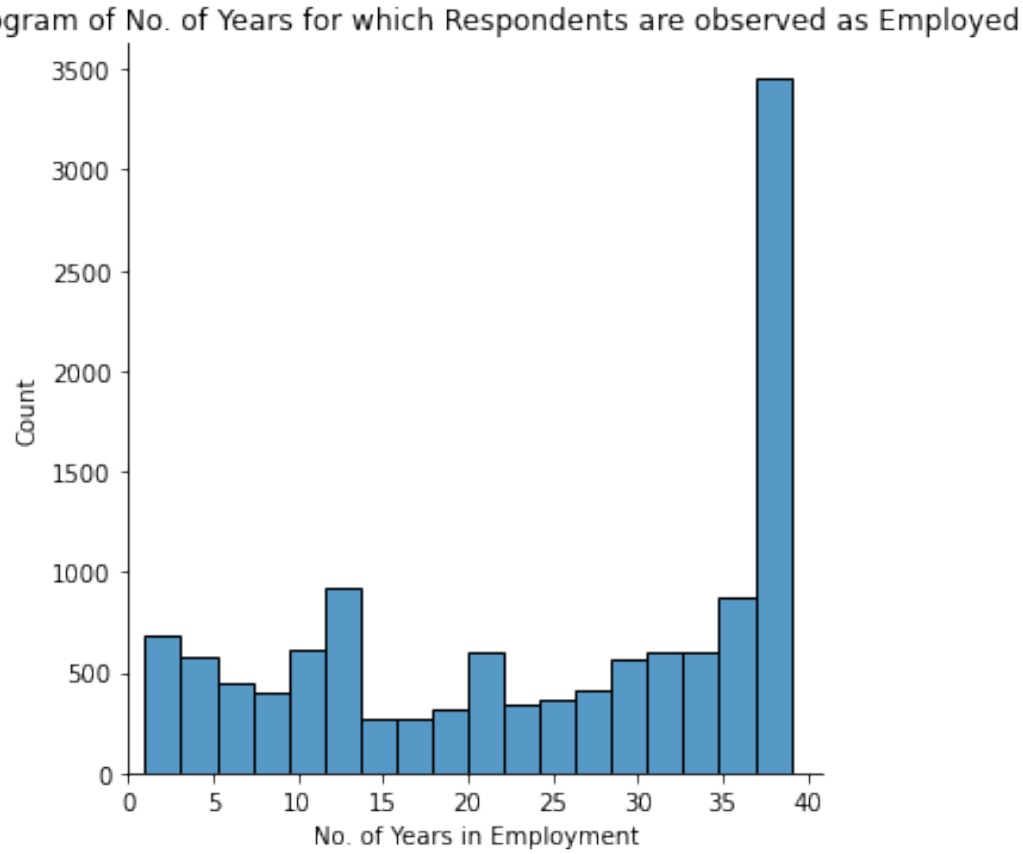
[7]:
```
count    12336.000000
mean        10.389348
std          7.000197
min          1.000000
25%          5.000000
50%          9.000000
75%         14.000000
max         67.000000
Name: Total_Jobs, dtype: float64
```

On average people hold **10** jobs during their working life and there is quite a bit of variation in jobs held. However, some of this variation in total jobs could be due to differences in the number of years a person is observed working. There are two kinds of reasons a respondent drop out of the employment history data. The first reason is that they became unemployed (could be voluntary or

involuntary) and the second reason is that they might have dropped out of the survey. A detail of the samples who became ineligible for the survey or who dropped can be found here. We must account for the differences in the amount of time observed in employment while looking at the behaviour of *Total_ Jobs* else our results will be biased. We ideally wish to compare the difference in the *Total_ Jobs* between men and women who have similar employment duration.

So i create the *Years_ Job_ History* variable that counts the number of years we see each respondent in the employment history data and look at its distribution.

[8]:
```
# Historgam of Years_Job_History variable
plot02 = sns.displot(jobs_unique_df, x = "Years_Job_History")
plot02.set(xlabel = 'No. of Years in Employment',
           title = 'Histogram of No. of Years for which Respondents are observed␣
 ↪as Employed')
plt.xlim(0, None)
plt.show()
```



Only around 28% of the respondents work throughout their life between the age of 20 till age 60. Others work for fewer years and this could be either due to breaks in their employment or due to them dropping out of the survey. Ideally, after ensuring that the reasons for dropping out are random and do not reduce the sample size of any gender significantly to affect validity of results,

we would like to remove from the sample people who drop out. But for this analysis we continue to keep them in the sample.

Even before we explore the differences in *Total_ Jobs* across gender, we should check if the behaviour of *Years_ Job_ History* varies by gender. In many countries it is observed that women drop out of the workforce after marriage or while raisin young children and the variation in the 'Years_Job_History' in this data might also be driven by differences in attachment to employment by gender. The linear regression below shows that in this data, gender does have a statistically signifcant correlation with *Years_ Job_ History*. However, women on average are observed for one year less than men in the employment data where men on average are observed for 25 years. This magnitude of difference is not very meaningful and i conclude that the behaviour of dropping out of the employment data is similar across genders.

[9]:
```
# Correlation between `Years_Job_History` and `Gender`
X01 = sm.add_constant(jobs_unique_df['Female'])
model01 = sm.OLS(jobs_unique_df['Years_Job_History'], X01).fit()
print(model01.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Years_Job_History   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     29.10
Date:                Mon, 30 Jan 2023   Prob (F-statistic):           7.01e-08
Time:                        15:51:35   Log-Likelihood:                -48944.
No. Observations:               12336   AIC:                         9.789e+04
Df Residuals:                   12334   BIC:                         9.791e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         25.2206      0.162    156.065      0.000      24.904      25.537
Female        -1.2425      0.230     -5.394      0.000      -1.694      -0.791
==============================================================================
Omnibus:                   247620.821   Durbin-Watson:                   1.231
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1249.355
Skew:                          -0.389   Prob(JB):                    5.08e-272
Kurtosis:                       1.649   Cond. No.                         2.60
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Next, we look at how much of the variation in *Total_ Jobs* is explained by *Years_ Job_ History*. The linear regression below shows that the number of years we see a person employed in this data explains a significant proprtion of variation in the total number of jobs held as the R-squared is fairly large.

[10]:
```
# Correlation between `Total_Jobs` and `Years_Job_History`
# Force no intercept in the model
model02 = sm.OLS(jobs_unique_df['Total_Jobs'],
                 jobs_unique_df['Years_Job_History']).fit()
print(model02.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:              Total_Jobs   R-squared (uncentered):                   0.719
Model:                             OLS   Adj. R-squared (uncentered):              0.719
Method:                  Least Squares   F-statistic:                          3.163e+04
Date:                 Mon, 30 Jan 2023   Prob (F-statistic):                        0.00
Time:                         15:51:35   Log-Likelihood:                         -40849.
No. Observations:                12336   AIC:                                     8.170e+04
Df Residuals:                    12335   BIC:                                     8.171e+04
Df Model:                            1
Covariance Type:            nonrobust
================================================================================
                         coef    std err          t      P>|t|      [0.025    0.975]
--------------------------------------------------------------------------------
Years_Job_History     0.3830      0.002    177.861      0.000       0.379     0.387
==============================================================================
Omnibus:                     1257.123   Durbin-Watson:                   1.748
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2494.778
Skew:                           0.667   Prob(JB):                         0.00
Kurtosis:                       4.753   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

The boxplot below indicates that much of the relationship between years observed in employment history data and total number of jobs held is driven by respondents who are observed working for less than 15 years. Hence, going forward in comparing *Total_Jobs* across genders, we will also control for *Years_Job_History*.

[11]:
```
# Boxplot of Total_Jobs by Years_Job_History
plot03 = sns.boxplot(data = jobs_unique_df, x = 'Years_Job_History', y =␣
 ↪'Total_Jobs')
plot03.set_xticks(range(0, max(jobs_unique_df['Years_Job_History']), 5))
plot03.set(xlabel = 'Years of Job History', ylabel = 'Total Jobs',
           title = 'Boxplot of Total Jobs by Years of Job History')
plt.show()
```

Job switching is affected by many different factors such as the sector or industry in which the person is employed, the role, education, etc. When trying to understand difference in job switching behaviour by gender, we need to check if the effect of gender is intermediated by these different factors. At the same time if a factor is not related to gender, we should control for it to ensure that we are comparing men and women who are as similar to each other on these other factors. In the following analysis i will evaluate the importance of some of these factors on *Total_Jobs* before doing a formal analysis of the effect of gender.

Job Sector

We can expect people to switch jobs with different frequency depending on their employment sctor (whether they are employed in the private sector or government), industry (whether they work in the automobile industry or the software industry) and their role (whether they work as a manager or as a specialist or a labourer).

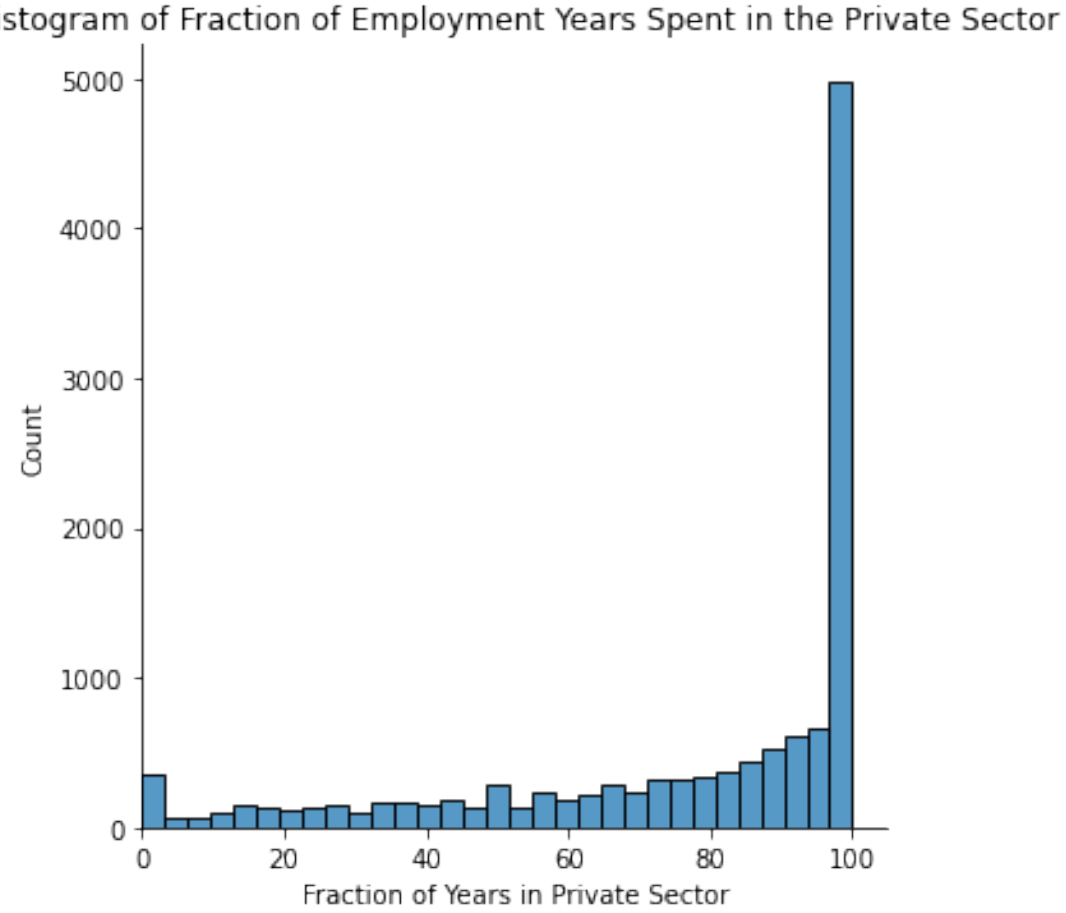
The survey records the industry, occupation(role) and class of worker (sector) for each employer of each respondent. I will not be including the industry and occupation as features in this analysis since these are categorical variables with large number of levels and need further work to reduce their dimensionality given the sample size. Also, the occupation and industry codes used in the survey changed over the years and need to be linked and corrected before adding them to the analysis. For now, i will only include the sector of the employer in this analysis.

Each employer/job can belong to any one of the three sectors: private, government or self employment.

Since people switch jobs between private, government and the self employment it is difficult to

cetagorize them into working in just one sector. However, we can still categorize people based on which sector they spend majority of their working life. For each respondent, we calculate the fraction of observed employment years spent in the private sector and plot its distribution below.

[12]:
```
# Histogram of fraction of employment years spent in the private sector
plot04 = sns.displot(jobs_unique_df, x = "Frac_Years_Pvt")
plot04.set(xlabel = 'Fraction of Years in Private Sector',
           title = 'Histogram of Fraction of Employment Years Spent in the␣
 ↪Private Sector')
plt.xlim(0, None)
plt.show()
```



[13]:
```
# Number of respondents that spend majority (more than 50%) of their employment
# years in the private sector.
jobs_unique_df.loc[jobs_unique_df['Frac_Years_Pvt'] > 50]['Person_Id'].count()
```

[13]: 9914

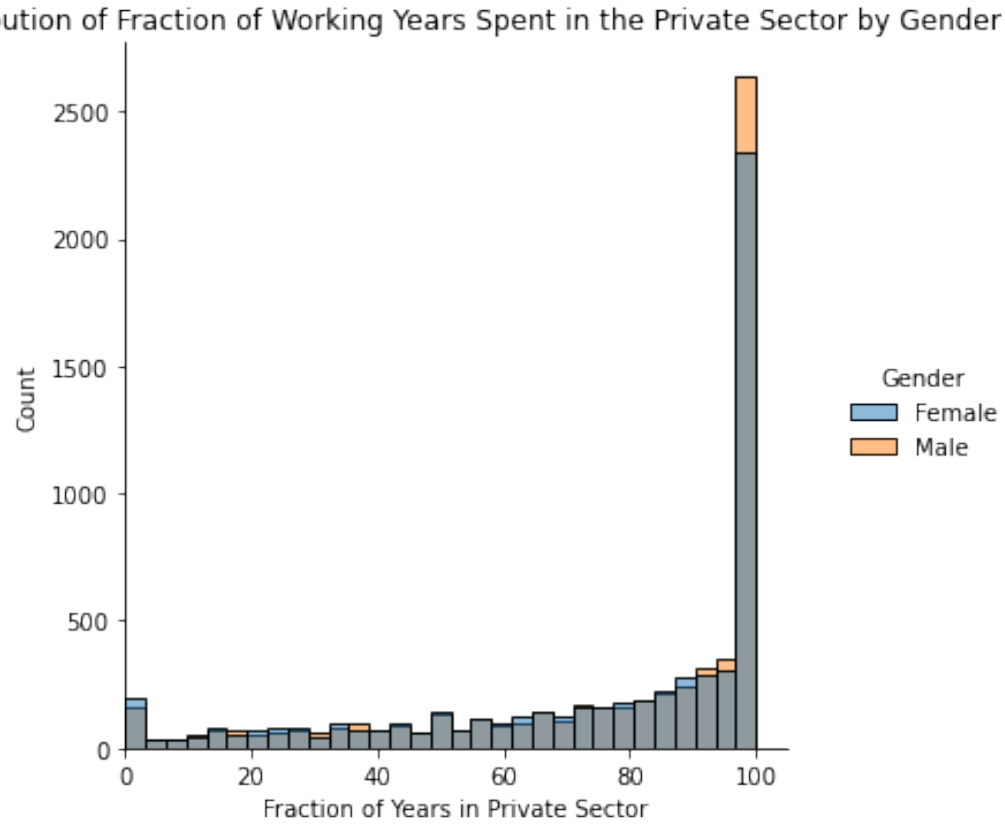
[14]:
```
# Correlation between Total_Jobs and Job_Types
(jobs_unique_df[['Total_Jobs','Frac_Years_Pvt','Frac_Years_Self','Frac_Years_Gvt']]
    .corr())
```

[14]:

	Total_Jobs	Frac_Years_Pvt	Frac_Years_Self	Frac_Years_Gvt
Total_Jobs	1.000000	0.110218	0.162477	-0.058394
Frac_Years_Pvt	0.110218	1.000000	-0.434187	-0.681356
Frac_Years_Self	0.162477	-0.434187	1.000000	-0.159055
Frac_Years_Gvt	-0.058394	-0.681356	-0.159055	1.000000

We can see that around 80% of the sample spends majority of their working life in the private sector. Also, people in the private sector are morelikely to switch jobs that people in the government or the self employed sector.

[15]:
```
# Histogram of the fraction of employment years spent in the private sector
# by gender

plot05 = sns.displot(jobs_unique_df, x = "Frac_Years_Pvt", hue = "Gender")
plot05.set(xlabel = 'Fraction of Years in Private Sector',
           title = 'Distribution of Fraction of Working Years Spent in the␣
 ↪Private Sector by Gender')
plt.xlim(0,None)
plt.show()
```



The distribution of the fraction of working years spent in the private sector does not seem very different for the two genders indicating there does not seem to be selection into the private sector by gender.

In the formal analysis of *Total_ Jobs* and *Gender*, i will control for the fraction of years spent in the private sector. I will also limit the sample to only those people who spend more than 60% of their working life in the private sector since our hypothesis seems to make sense only for people working in this sector.

Education

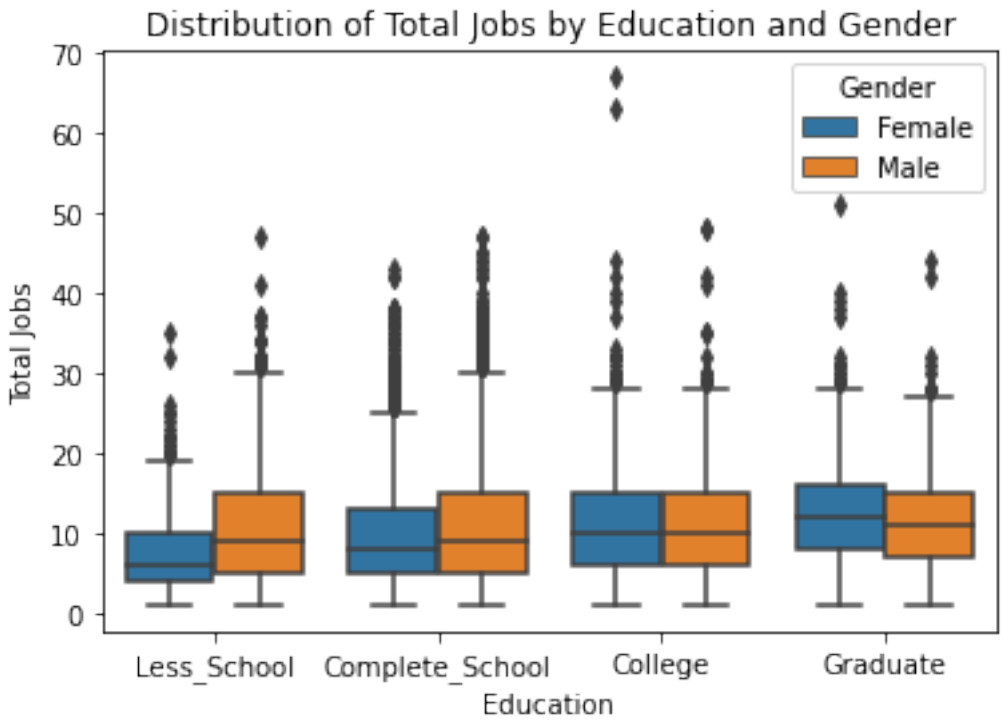
The survey records the highest degree completed by each respondent and this variable is an interger that indicates the number of years spent on one's education. I have factored this variables into four categories. *Less_ School* stands for not completing high school or less. *Complete_ School* indicates that the person completed high school but did not complete college. *College* indicates that the person has an undergraduate degree and *Graduate* indicates that a person also studies for a graduate degree.

[16]:
```
# Convert 'Education_Ctg' to categorical variable and order the levels for␣
↪plotting
jobs_unique_df['Education_Ctg'] = (jobs_unique_df['Education_Ctg'].
↪astype('category')
                                   .cat
                                   .reorder_categories(['Less_School',
                                                        'Complete_School',
                                                        'College',
                                                        'Graduate']))
```

[17]:
```
# Sample size of education categories by gender
pd.crosstab(jobs_unique_df['Education_Ctg'],jobs_unique_df['Gender'] )
```

[17]:
```
Gender           Female  Male
Education_Ctg
Less_School         598   864
Complete_School    3733  3976
College            1087   889
Graduate            654   535
```

[18]:
```
# Boxplot of Total_Jobs by education categories and gender
plot06 = sns.boxplot(data = jobs_unique_df, x = 'Education_Ctg', y =␣
↪'Total_Jobs',
                     hue = 'Gender')
plot06.set(xlabel = 'Education', ylabel = 'Total Jobs',
           title = 'Distribution of Total Jobs by Education and Gender')
plt.show()
```



A large fraction of the sample that has ever been employed seems to have completed schooling. More women seem to have college or graduate degrees than men. Looking at the boxplot, i do see some difference in *Total_ Jobs* by education that is not solely driven by gender differences. I will also include *Education_ Ctg* variable in the formal analysis.

Modeling `Total_Jobs` using Linear Regression

I now try to see the relationship between gender and total number of jobs held during a person's working life after controlling for the number of years they are observed working in the data, their education and the fraction of time they spend in the private sector.

[19]:
```
# Create dummy variables for education
jobs_unique_df = (jobs_unique_df
                      .merge(pd.
 ↪get_dummies(jobs_unique_df[['Person_Id','Education_Ctg']]
                                                .
 ↪set_index('Person_Id')['Education_Ctg']),
                             on = 'Person_Id', how = 'left'))
```

[20]:
```
# Regression model that only keeps respondents if they primarily work in the
# Pvt sector i.e. (Frac_Years_Pvt >= 60)

jobs_unique_df_pvt = jobs_unique_df.loc[(jobs_unique_df['Frac_Years_Pvt'] >= 60)]
X03 = sm.add_constant(jobs_unique_df_pvt[['Years_Job_History', 'Female',
```

```
                                        'Frac_Years_Pvt'    , 'College',
                                        'Complete_School'   , 'Graduate']])
model03 = sm.OLS(jobs_unique_df_pvt['Total_Jobs'], X03).fit(cov_type = 'HC3')
print(model03.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Total_Jobs   R-squared:                       0.230
Model:                            OLS   Adj. R-squared:                  0.229
Method:                 Least Squares   F-statistic:                     562.2
Date:                Mon, 30 Jan 2023   Prob (F-statistic):               0.00
Time:                        15:51:38   Log-Likelihood:                -30816.
No. Observations:                9433   AIC:                         6.165e+04
Df Residuals:                    9426   BIC:                         6.170e+04
Df Model:                           6
Covariance Type:                  HC3
==============================================================================
=====
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----
const                 16.0885      0.554     29.027      0.000      15.002
17.175
Years_Job_History      0.2500      0.005     51.576      0.000       0.240
0.259
Female                -0.6750      0.130     -5.205      0.000      -0.929
-0.421
Frac_Years_Pvt        -0.1132      0.006    -20.216      0.000      -0.124
-0.102
College               -0.6568      0.244     -2.690      0.007      -1.135
-0.178
Complete_School       -0.7161      0.191     -3.739      0.000      -1.091
-0.341
Graduate              -0.5981      0.290     -2.062      0.039      -1.167
-0.030
==============================================================================
Omnibus:                     1896.357   Durbin-Watson:                   1.833
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4350.204
Skew:                           1.141   Prob(JB):                         0.00
Kurtosis:                       5.421   Cond. No.                         858.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

According to the regression results, all variables are statistically significant. Females on average seem to work 0.68 years less than males. Economically this number implies a difference of less than

10% of total jobs held by men on average. Based on this data and analysis, men and women seem to be equally likely to switch jobs over their lifetime.

3.2 Analyzing Job Switching Behaviour Across Age

I would like to check for patterns in the probability of switching a job across age and by gender. I wish to check if women switch jobs less than men during the reproductive years but catch up to men in later years, given we have found that on average they hold as many jobs as men over their working life.

[21]:
```
# Check for correlation between the two job switching variables
jobs_switch_df[['Switch_Job_1', 'Switch_Job_2']].corr()
```

[21]:

	Switch_Job_1	Switch_Job_2
Switch_Job_1	1.000000	0.791292
Switch_Job_2	0.791292	1.000000

I have defined two variables to measure if a respondent switches job during a year they are observed as employed. *Switch_Job_1* takes value 1 if the respondent starts a new job and ends any job in a calendar year. *Switch_Job_2* takes value 1 if a respondent starts a new job in a calendar year. Since these variables are intended to measure the same thing, i look at their correlation above. They seem to be highly positively correlated and so in the further analysis i will only use *Switch_Job_1*.

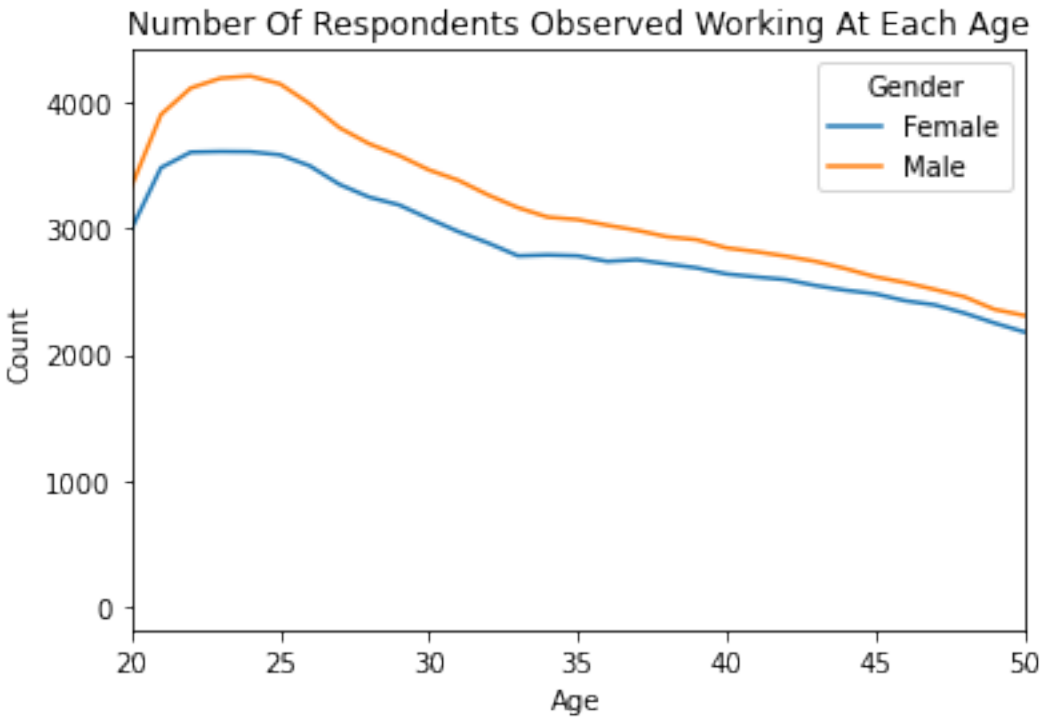
[22]:
```
# Merge jobs_switch_df with features from jobs_unique_df - where
# key is Person_Id

jobs_switch_df = (jobs_switch_df
                      .merge(jobs_unique_df[['Person_Id'         , 'Total_Jobs',
                                             'Years_Job_History', 'Female',
                                             'Gender'           ,␣
 ↪'Frac_Years_Pvt',
                                             'Less_School'      ,␣
 ↪'Complete_School',
                                             'College'          , 'Graduate']],
                             on = 'Person_Id', how = 'left'))
```

[23]:
```
# Keep only the respondents that have spent more than 60% of their emplyment
# years in the private sector

jobs_switch_df = jobs_switch_df.loc[jobs_switch_df['Frac_Years_Pvt'] >= 60]
```

[24]:
```
# Create a dataframe from jobs_switch_df that records for each age and gender,
# the number of respondents who are employed and the fraction of them that
# switches a job

age_switch_summary = (jobs_switch_df
                          .groupby(['Age', 'Gender'])
                          .agg(No_Resp = ('Person_Id', 'nunique'),
```

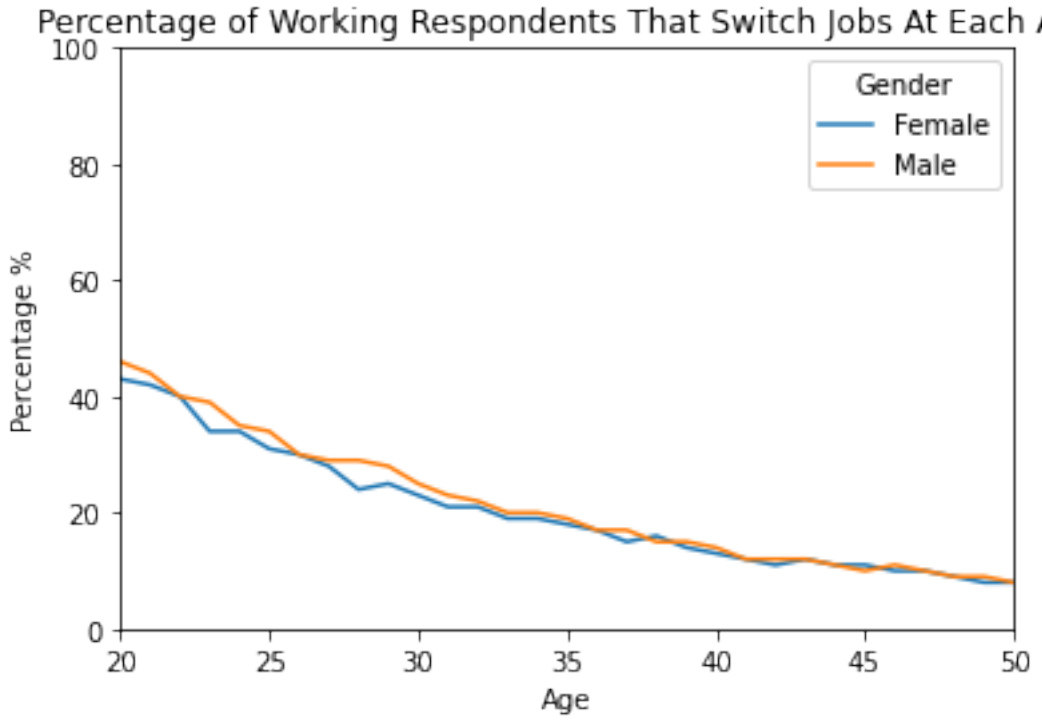
```
                                No_Switch_1 = ('Switch_Job_1', 'sum'),
                                No_Switch_2 = ('Switch_Job_2', 'sum'))
                            .reset_index())

age_switch_summary['Frac_Switch_1'] = round(age_switch_summary['No_Switch_1']
                                            * 100 /␣
 ↪age_switch_summary['No_Resp'])
```

[25]:
```
# Lineplot of number of respondents working at each age by gender

plot07 = sns.lineplot(x = 'Age', y = 'No_Resp', hue = 'Gender',
                      data = age_switch_summary)
plot07.set(title = 'Number Of Respondents Observed Working At Each Age',
           ylabel = 'Count')
plt.xlim(20, 50)
plt.show()
```



I plot the total number of respondents at each age to see if there are enough respondents working at each age and from both the genders. The sample seems large enough to be able to do this analysis by age and gender. The above graph again confirms that the number of working men in the sample is greater than women. Also, over the years there is a reduction in the number of both working men and women. Further analysis needs to be done to check if this patter is due to a reduction in the survey sample or due to people deciding to leave work.

[26]:
```
# Lineplot of the fraction of repondents that switch job at each age

plot08 = sns.lineplot(x = 'Age', y = 'Frac_Switch_1', hue = 'Gender',
                      data = age_switch_summary)
plot08.set(title = 'Percentage of Working Respondents That Switch Jobs At Each␣
 ↪Age',
           ylabel ='Percentage %')
plt.xlim(20, 50)
plt.ylim(0, 100)
plt.show()
```



This is an interesting graph and it shows that people switch jobs a lot more frequently in their 20s and 30s than later in their life. In fact the probability of switching jobs declines monotonically with age and for both genders. The high frequency of switching jobs in the early 20s could be explained by multiple part time and temporary jobs students do while in college. However, even during 30s this frequency remains at 20% implying that 1 in 5 people in their 30s and working tend to switch their job in a given year.

The probability of changing jobs does not seem to differ much between the two genders. Women do seem less likely to switch jobs from the age 25 till 35, but this difference looks small. We will confirm the patterns observed in this graph formally using a logistic regression with *Switch_Job_1* as the dependent variable.

[27]:
```
# Run a logistic regression of Switch_Job_1 on age, gender, their interaction
# and education

X04 = sm.add_constant(jobs_switch_df[['Female'         , 'Age'    ,␣
↪'Switch_Job_1',
                                      'Complete_School', 'College', 'Graduate']])
model04 = smf.logit(formula = 'Switch_Job_1 ~ Age + Female + Age:Female \
                                            + Complete_School + College \
                                            + Graduate',
                    data = X04).fit()
model04.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.492753
         Iterations 6
```

[27]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:           Switch_Job_1   No. Observations:               227828
Model:                          Logit   Df Residuals:                   227821
Method:                           MLE   Df Model:                            6
Date:                Mon, 30 Jan 2023   Pseudo R-squ.:                 0.09098
Time:                        15:51:41   Log-Likelihood:            -1.1226e+05
converged:                       True   LL-Null:                   -1.2350e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------

Intercept          1.3958      0.027     51.121      0.000       1.342       1.449
Age               -0.0757      0.001   -100.101      0.000      -0.077      -0.074
Female            -0.1453      0.034     -4.217      0.000      -0.213      -0.078
Age:Female         0.0031      0.001      2.801      0.005       0.001       0.005
Complete_School   -0.1581      0.017     -9.100      0.000      -0.192      -0.124
College           -0.1725      0.020     -8.436      0.000      -0.213      -0.132
Graduate          -0.1287      0.023     -5.501      0.000      -0.175      -0.083
==============================================================================

"""
```

[28]:
```
# Predict the probability of switching job for men and women who have
# completed school at age 30 and 45

jobs_switch_df_pred = pd.DataFrame(data = {'Female' : [0, 1, 0, 1],
                                           'Age' : [30, 30, 50, 50],
                                           'Complete_School' : [1, 1, 1, 1],
```

```
                                            'College' : [0, 0, 0, 0],
                                            'Graduate' : [0, 0, 0, 0]})

jobs_switch_df_pred['Prediction'] = round(model04.predict(jobs_switch_df_pred),␣
↪2)
print(jobs_switch_df_pred)
```

```
   Female  Age  Complete_School  College  Graduate  Prediction
0       0   30                1        0         0        0.26
1       1   30                1        0         0        0.25
2       0   50                1        0         0        0.07
3       1   50                1        0         0        0.07
```

The results from the logit model show that age, gender as well as education are all statistically significant at (1% level of significance), implying that all these factors are correlated to the frequency of switching jobs. The negative coefficients associated with *Age* indicates that people change jobs less with age. Similarly, women (indicated by a negative coefficient on *Female*) are less likely to switch jobs relative to men. As was hypothesized, the difference in job switching behaviour between men and women diminishes with age as the coefficient on *Age*x*Gender* is positive. According to the model, the probability that a man who has completed schooling and is 30 years old will switch a job is 26% and this probability for a women at the same education and schooling level is 25%. For both a man and a women who are 50 years old and who have completed high school the probability to switch a job is 7%. The model predictions imply that the small difference in job switching frequency between men and women in this data is mainly coming from differences in younger years. Greater years of education also lowers the likelihood of switching jobs which seems due to more stability associated with white collar jobs that more educated respondents are likely to be employed in.

Even though the factors included are correlated to frequency of switching jobs, they explain little variation in the dependent variable as the pseudo R-squared is 10%. This is understandable as i have not included in the model information about the industry and employment position or about geography, all features that are likely to be important.

3.3 Scope For Future Work

In future analysis, i can try to include important employment factors such as industry and position of employment as well as geography of the respondent to be able to conclude something more clearly about the causal effect of gender on job switching behaviour. Another extension of this work can be to look at how incomes change with a job switch and factors that affect the direction of the income change.

3.4 Acknowledgements

This analysis was initiated as a part of a group project requirement during the Correlation One Data Science For All Fellowship (July-Aug 2022). The idea for this analysis as well as the project code is my individual work and i am responsible for all the errors. I would like to thank my team members Danni Chen, Wenjing Dong, Yuan Du and Yifan Ma for feedback.