Labour Economics and Policy Evaluation. Investigating employee attrition: causal inference using Regression Discontinuity Design, Instrumental Variables and Random Control Trials

Lorenzo Rossi, University of Milan

17/01/2022

Abstract

The main factors that contribute to employee attrition were analysed using causal inference techniques such as Regression Discontinuity Design (RDD), Instrumental Variables (IV) and Random Control Trials (RCT), on an anonymous IBM dataset. The research found that having to work overtime widely causes employees to quit their job, but sufficient training helps preventing this issue. Income and age play an important role as well. Low remunerative positions present higher rates of attrition, especially in younger age groups. At the same time, promotions don't reduce the probability of attrition, except in higher-income employees. Morevoer, employees who have already worked in many companies are more likely to quit.

Introduction

This paper will perform some of the most important causal inference techniques in order to understand the main factors that influence employee attrition. The data on which this research is based come from an anonymized dataset from IBM.

Employee attrition occurs when the size of the workforce of a company diminishes over time due to unavoidable factors such as employee resignation for personal or professional reasons. Contrarily to turnover, when the company makes efforts to replace the lost employee, in attrition cases the vacancy remains unfilled, or the employer completely eliminates that position. Usually, employers look to reach a low attrition rate since it means that their employees are satisfied and they don't have to invest in hiring and training new people. Attrition is an inevitable part of any business and, as said, some forms are unavoidable, like if an employee is retiring or is moving to another city. But attrition may be caused by several reasons, and it is of utmost importance to understand its processes in order to consciously address or prevent it, and not let it become a cause of concern for the company, the employer and the other workers.

While sometimes a company needs to eliminate positions to stay financially afloat, assign new duties to particular employees or implement new technologies that can replace the labor force, having high levels of attrition may lead to a lack of continuity, training gaps, and a lack of institutional knowledge. It can take a long time to fill positions (especially specialized roles), and leaving these positions empty, it can become difficult to fill these positions later. These holes can cause burnout for the remaining employees and lower overall productivity, leading to unhappy customers and struggles for the company.

The following analysis will try to address the issue of attrition. The first part will focus on data visualisation and exploration, while the second one will be dedicated to the implementation of causal inference models. Fort this paper, the following were choses:

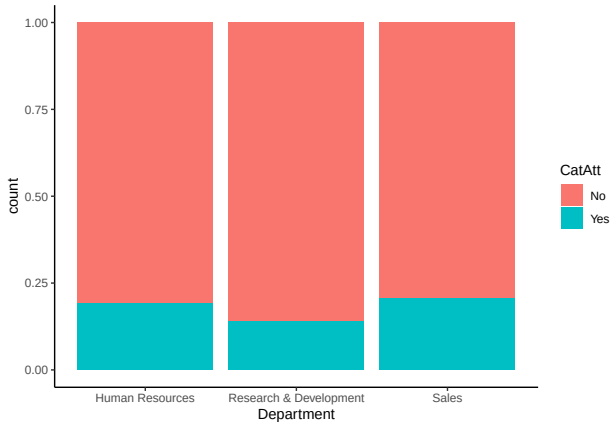
1. Regression Discontinuity Design

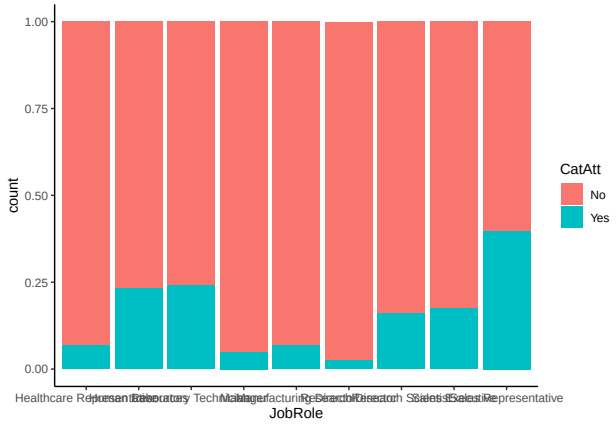
2. Instrumental Variables
3. Random Control Trials

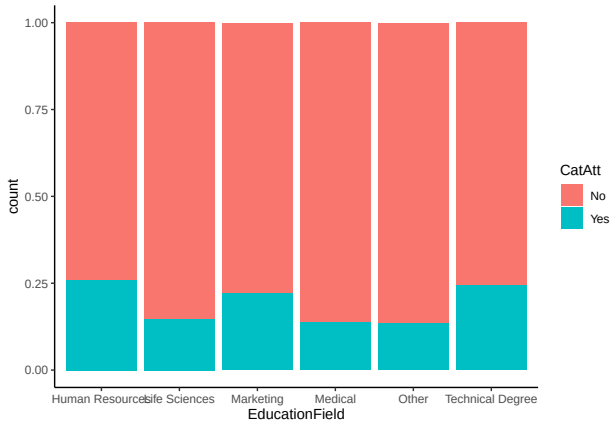
Data analysis and visualisation

The dataset consists in 28 variables, both categorical and numerical. The main variable is Attrition, indicated with 1 or 0, depending on the status of the employee. Each employee is affiliated with its job role, its department and its education background. Along with some primary variable like age, monthly income, gender and marital status, there are several others concerning the professional profile of the related employee, such as the years in its current role, in its current company or with its current manager, the number of companies in which the person worked, the years since its last promotion, if the subject works overtime and many more.

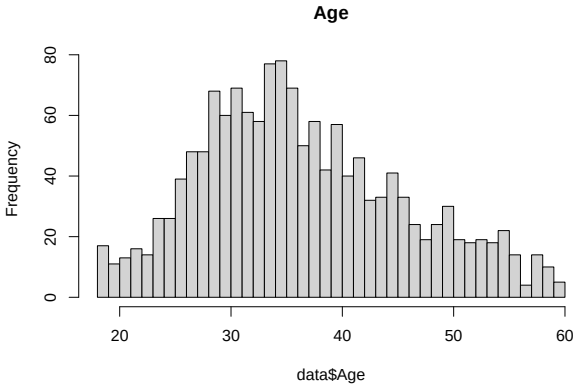
The first part of the data analysis consists in having a visual perspective of attrition in relation with certain background variables. The following charts show the attrition rate in each department, for each job role and education background. As it is clear, the department of **R&D** presents a lower attrition rate than HR and Sales. The job role who suffers more from attrition among its employees is the **Sales Representative**, followed by **Laboratory Technicians** and HR personnel. Finally, the degrees that are associated with higher rates of attrition are the ones in **Human Resources** and the **Technical Degree**

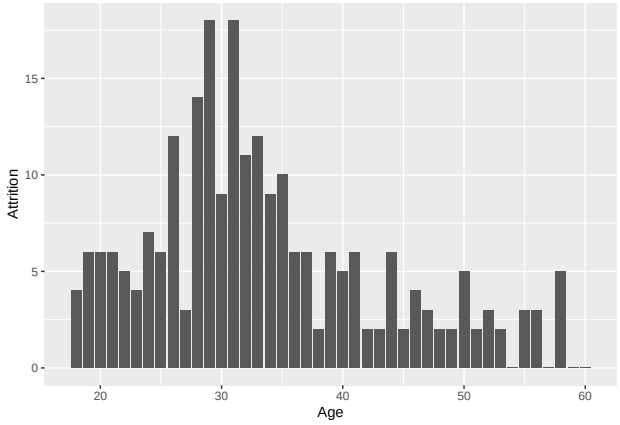


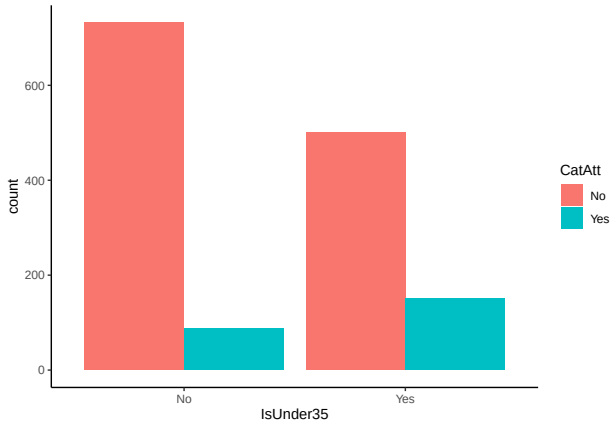


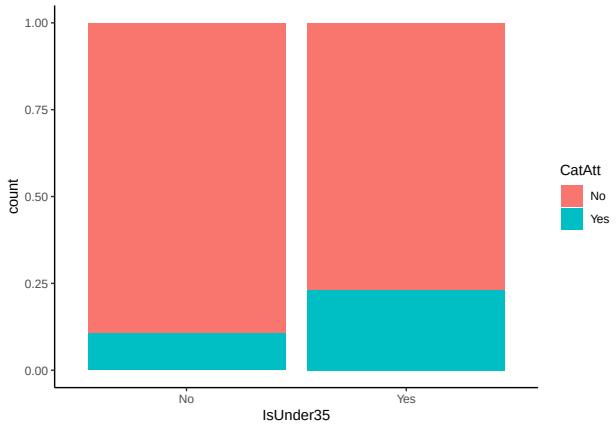


Taking a look at the *Age* variable, despite its distribution being not particularly skewed, the frequency of attrition is clearly concentrated in the younger age groups. Therefore, a new categorical variable was created, in order to identify the individuals under 35 years old. The consequent chart confirms an higher attrition rate for this age group.

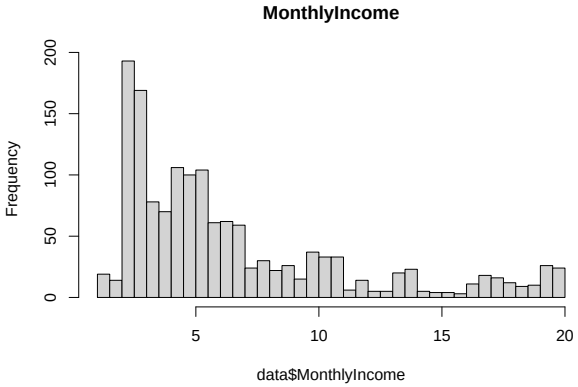




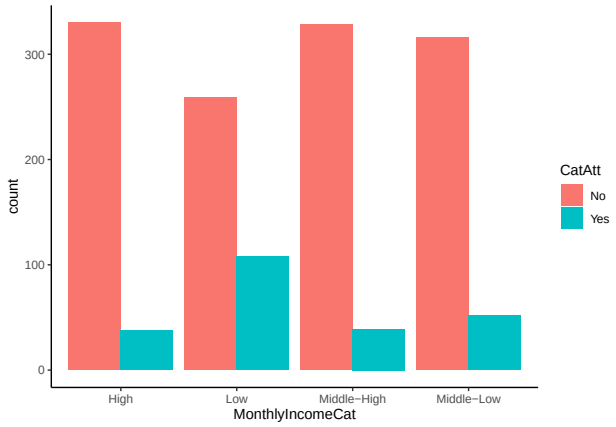




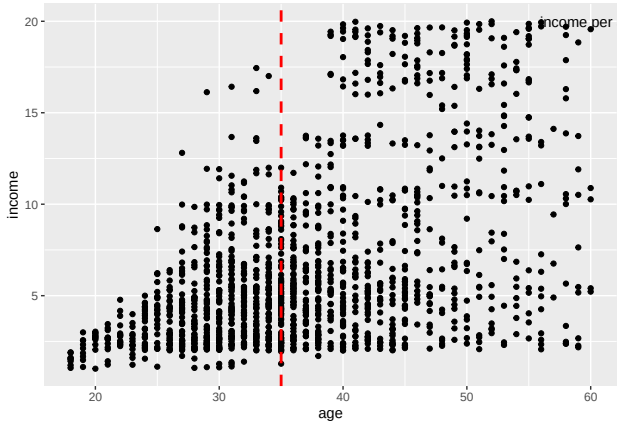
A similar procedure was iterated when visualizing the values of monthly income. However, in this case the skeweness of the distribution is much more clear, revealing that the majority of subjects earns around 5.000 dollars per month or less. The distribution was divided into a four-categories variable and it's clear that lower incomes present an higher level of attrition




```
##      0%     25%     50%     75%    100%
##   1.009   2.911   4.919   8.379 19.999
```



Finally, one last analysis consists in visualizing, through a scatter plot, the relation between age and income. As one may imagine, despite some outliers, younger employees are associated with lower incomes than the older ones.



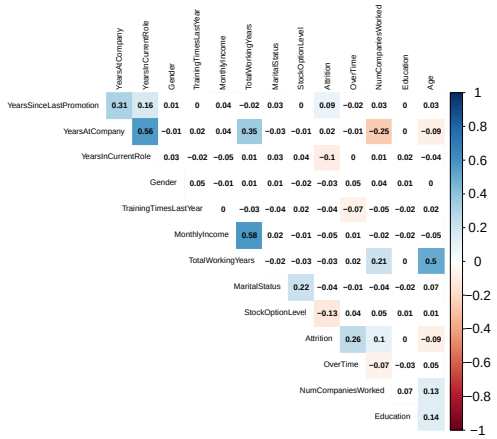
Regression models

Despite recognizing the value of certain variables like job involvement or job satisfaction, that certainly reflect the the opinion of the employee towards the workplace, the choice for the variables to insert into the models was done by selecting just the quantitative and "objective" ones, meaning the variables that reflect and the status of the worker and exactly quantify specific aspects of its professional profile (such as age, income, total working years, years in the current company or last promotion), in order to perform the most impartial analysis. The selection is the following:

- Attrition
- Overtime work
- MaritalStatus
- Education
- Gender
- Age

- NumCompaniesWorked
- YearsInCurrentRole
- YearsSinceLastPromotion
- YearsAtCompany
- StockOptionLevel
- TrainingTimesLastYear
- JobRole

To have an initial overview of the relations among the variables, the first step was the computation of the following *correlation matrix* between the numerical variables. Positive and negative correlations are indicated with the blue and red color, respectively. White spaces are associated with a p-value greater than 0.05 and thus, non significant.



Logistic Regression

The first model consist in a simple logistic regression in which all the selected variables have been regressed on the dependent one, *Attrition*, in order to have a first look at what mainly influences the variable of interest. It's possible to notice that many of the selected variables influence, either negatively or positively the value of Attrition. The exception are *Gender*, *MaritalStatus*, *TrainingTimeLastYear* and *YearsAtCompany*, because of the high p-value. An impressive coefficient is the one related to *OverTime* (1.49), also associated with a p-value even lower than 0.01, making this variable a particularly relelvant factor.

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

However, before reaching conclusions, it's important to control many things in our model of regression; one of the most important is the variance inflation factor (VIF) which gives us insights into the variables which can be "dangerous" to what concern multicollinearity. According to Hair et al. (1995) the maximum acceptable level of VIF is 10, whereas according to Ringle et al. (2015) the maximum acceptable level of VIF is 5.

$\chi^2(12)$	227.61
Pseudo-R^2 (Cragg-Uhler)	0.24
Pseudo-R^2 (McFadden)	0.18
AIC	1096.97
BIC	1165.78

	Est.	S.E.	z val.	p
(Intercept)	0.56	0.41	1.38	0.17
OverTime	1.49	0.16	9.32	0.00
MaritalStatus	-0.24	0.17	-1.45	0.15
Education	0.02	0.08	0.29	0.77
Gender	-0.28	0.16	-1.73	0.08
Age	-0.04	0.01	-4.03	0.00
MonthlyIncome	-0.10	0.03	-3.66	0.00
NumCompaniesWorked	0.13	0.03	3.89	0.00
YearsInCurrentRole	-0.15	0.04	-3.99	0.00
YearsSinceLastPromotion	0.14	0.04	3.89	0.00
YearsAtCompany	0.01	0.03	0.19	0.85
StockOptionLevel	-0.47	0.11	-4.40	0.00
TrainingTimesLastYear	-0.11	0.06	-1.70	0.09

Standard errors: MLE

```
##              OverTime         MaritalStatus             Education
##              1.043656              1.096983              1.070617
##                Gender                   Age         MonthlyIncome
##              1.017322              1.471349              1.592220
##   NumCompaniesWorked   YearsInCurrentRole YearsSinceLastPromotion
##              1.205095              2.315268              2.163991
##       YearsAtCompany     StockOptionLevel TrainingTimesLastYear
##              3.600546              1.101064              1.010738
```

None of the VIF values stands outside the aforementioned critical values. There is no risk of multicollinearity and thus, no need to drop any variable.

There are several other variables that present a significant explanatory power. The value of *YearsInCurrentRole* suggests that the longer an employee keeps its position, the lower are the possibilities for him to quit. *MonthlyIncome* and *Age* present a negative coefficient, meaning that higher income and age result in a reduction of probability for the employee to abandon the company and as seen previously, higher income is often correlated with older age. Moreover, higher *StockOptionLevel* (which is the opportunity for an employee to purchase, or assign to others, previously issued shares of his company) values, the lower is the attrition rate: remaining in the same company gives the opportunity to benefit from the profit of the shares that the employee owns. There are also positive coefficients among the variables. The values assigned to the coefficients of *NumCompaniesWorked* and *YearsSinceLastPromotion* suggest that if the number of companies in which the employee previously worked is high, or different years have passed since his last promotion, then the possibility for the individual to resign from the current position is higher. The second logistic regression was performed on *JobRole*, confirming the results previously obtained in the charts: **Sales Representative**, **HR professionel** and **Laboratory Technicians** are the jobs the see the highest ratio of attrition.

```
## # weights:  27 (16 variable)
## initial  value 3229.920129
```

```
## iter  10 value 2953.960997
## iter  20 value 2938.110744
## final  value 2938.107461
## converged
```

```
## Call:
## multinom(formula = JobRole ~ Attrition, data = data2)
##
## Coefficients:
##                         (Intercept)    Attrition
## Human Resources          -1.1150982  1.402810594
## Laboratory Technician     0.4791922  1.450711468
## Manager                  -0.2293014 -0.358516361
## Manufacturing Director    0.1012421  0.004142155
## Research Director        -0.4472740 -1.056789202
## Research Scientist        0.6972407  0.955693153
## Sales Executive           0.7906974  1.055150912
## Sales Representative     -0.8920115  2.191332492
##
## Std. Errors:
##                         (Intercept) Attrition
## Human Resources           0.1821973 0.4771167
## Laboratory Technician     0.1152081 0.3748530
## Manager                   0.1360369 0.5741319
## Manufacturing Director    0.1249172 0.4761468
## Research Director         0.1449720 0.7950653
## Research Scientist        0.1108080 0.3803520
## Sales Executive           0.1091524 0.3749273
## Sales Representative      0.1679201 0.4118398
##
## Residual Deviance: 5876.215
## AIC: 5908.215
```

Regression Discontinuity Design

RDD is performed by applying a standard OLS but splitting the observations according to a threshold, above (or below) which the treatment Di is assigned.

$Di = 1(Ri > t)$ or $Di = 1(Ri < t)$

With Ri as the running variable and t its threshold.

$Yi = \beta 0 + \beta 1Ri + \beta 2Di(Ri > t)$ or $\beta 2Di(Ri < t)$

Depending on the aim of the analysis. By comparing the observations lying close to the threshold, it is possible to estimate the average treatment effect (ATE).

In this case the RDD was performed by looking at two important variables, with high level of significance, from the previous regression: *MonthlyIncome* and *Age*. Starting from the income as running variable Ri, the threshold was chosen by looking at the quantile previously computed in the data visualization, finding in 4.919 the 50% value of the four quartiles. Dummmy Di values 1 and 0 were assigned to observation above and below the threshold, respectively. As expected, values above the threshold decrease the probability of attrition by 6%. The equation for the RDD was built in the as follows, similarly to the previous theoretical example:

$ATTi = \beta 0 + \beta 1(MIi - t) + \beta 2Di(MIi > t)$

Where $ATTi$ is the *Attrition* variable, t is the threshold (4.919 in this case), $MIi-t$ is the difference between the value of the monthly income of i and $Di(MIi > t)$ the dummy assigned to i whether it falls above or below the threshold.

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(2)$	47.10
Pseudo-R^2 (Cragg-Uhler)	0.05
Pseudo-R^2 (McFadden)	0.04
AIC	1257.48
BIC	1273.36

	Est.	S.E.	z val.	p
(Intercept)	-1.45	0.11	-13.82	0.00
threshold	-0.28	0.22	-1.26	0.21
I(MonthlyIncome - 4.919)	-0.10	0.03	-3.20	0.00

Standard errors: MLE

The same experiment was repeated by setting 2.911, the limit value of the first quartile (or the **Low Income** category), as threshold. The second model presents higher coefficients. Above the threshold, the probability of attrition decreases by 14%.

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(2)$	65.36
Pseudo-R^2 (Cragg-Uhler)	0.07
Pseudo-R^2 (McFadden)	0.05
AIC	1239.23
BIC	1255.10

	Est.	S.E.	z val.	p
(Intercept)	-0.92	0.12	-7.97	0.00
threshold	-0.81	0.18	-4.44	0.00
I(MonthlyIncome - 2.911)	-0.06	0.02	-2.73	0.01

Standard errors: MLE

The last application of RDD was done though the age variable. In this case, the threshold was chosen from the division previously computed in the data visualization, selecting 35 as the threshold value. This time the coefficients present lower values, but above the threshold the probability of attrition decreases as well (-2%).

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(2)$	44.78
Pseudo-R^2 (Cragg-Uhler)	0.05
Pseudo-R^2 (McFadden)	0.03
AIC	1259.80
BIC	1275.68

	Est.	S.E.	z val.	p
(Intercept)	-1.37	0.13	-10.61	0.00
threshold	-0.55	0.24	-2.28	0.02
I(Age - 35)	-0.03	0.01	-1.86	0.06

Standard errors: MLE

Instrumental Variables

IVs are an efficient tool that is used to study the indirect effect of certain exogenous variables that would have no direct effect on the dependent one or help preventing the effects of endogeneity within other variables in the system. Because of their effects on the regression model, these exogenous variables are called, therefore, instruments.

Consider a linear regression structure:

$Yi = a + Xi\beta 1 + \varepsilon i$

If an exogenous variable has a direct effect on Xi but not on Yi, then Xi may be decomposed in its own regression with the exogenous variable:

$Xi = \pi 0 + Zi\pi 1 + \nu i$ and $\pi 1 \neq 0$

Where Zi is the exogenous variable which "shows" its effect only through Xi, and therefore the main regression may be rewritten as follows.

$Yi = a + (\pi 0 + Zi\pi 1 + \nu i)\beta 1 + \varepsilon i$

In which the variable Xi was substituted by its own regression, in a process called Two-Stages Least Squares (2SLS). In this way, it's possible to measure the impact of the exogenous variable on Yi.

The first issue is to identify the potential instruments. To be one, the exogenous variable must be correlated with Xi once the other exogenous variables have been netted out, thus having $\pi 1 \neq 0$. This is tested through the 2SLS. Another way to prove the instrumentality of a variable comes from knowledge of the theory related to the experiment or proofs found in other empirical studies.

IV 1: training and overtime work

In the current analysis, the variable *TrainingTimesLastYear* will be tested as an instrument for overtime work. From the previous correlation matrix, it was noticeable a negative correlation between the two variables, therefore a 2SLS will be performed in order to find an actual dependence. Also, one could argue that giving more training to an employee may increase its skills and productivity, and consequently reducing the probability of having to work overtime. To confirm the previous hypothesis, the following regression reveals

an highly significant negative coefficient of training times on overtime work, meaning that the variable helps to explain the dependent one. Therefore, *TrainingTimesLastYear* can be considered a valid instrument.

Observations	1470
Dependent variable	OverTime
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(1)$	9.37
Pseudo-R^2 (Cragg-Uhler)	0.01
Pseudo-R^2 (McFadden)	0.01
AIC	1746.16
BIC	1756.75

	Est.	S.E.	z val.	p
(Intercept)	-0.54	0.14	-3.94	0.00
TrainingTimesLastYear	-0.14	0.05	-3.02	0.00

Standard errors: MLE

Proceeding with the 2SLS, the previous regression was fitted into a second regression with the attrition variable. In this way, it will be possible to see the effect of training on *Attrition*. In R, this is possible by extracting the value of the coefficient from the previous model and set is as the regressor of the second equation. The coefficient has been named **IVttls** (Instrumental Variable *TrainingTimesLastYear*).

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(1)$	5.35
Pseudo-R^2 (Cragg-Uhler)	0.01
Pseudo-R^2 (McFadden)	0.00
AIC	1297.24
BIC	1307.82

	Est.	S.E.	z val.	p
(Intercept)	-2.98	0.59	-5.07	0.00
IVttls	4.66	2.03	2.29	0.02

Standard errors: MLE

Interestingly, in the second stage it's possible to notice an higher coefficient (4.65) than the one assigned to overtime work in the first logistic model. An explanation may be: lower number of training session cause the increment of overtime work and thus, the probability for the employee to quit the current job. The 2SLS proved *TrainingTimesLastYear* to have a relevant explanatory role through its effect on one of the main variables that affect attrition the most.

The equation of the whole procedure, taking into account the previous theoretical example, is the following:

$ATTi = a + OTi\beta 1 + \varepsilon i$

Where $ATTi$ is *Attrition* and OTi is *OverTime*

$OTi = \pi 0 + TTLSi\pi 1 + \nu i$

Where $TTLSi$ is *TrainingTimesLastYear*. Therefore

$ATTi = a + (\pi 0 + TTLSi\pi 1 + \nu i)\beta 1 + \varepsilon i$

IV 2: number of companies and years in the current company

From the initial logistic regression, the variable *YearsAtCompany* didn't have a significant explanatory power on attrition rate. However, taking a look again at the previous matrix, the variable shows a negative correlation with *NumCompaniesWorked*. Therefore, a second 2SLS regression will be performed, in order to verifiy if the variable may be considered a valid instrument too (in this case affecting *NumCompaniesWorked*). The hypothesis is that spending longer periods working for one company consequently reduces the total number of companies in which the employee worked in, and therefore this reduces the possibility of a short term abandon of the current workplace.

As expected, the first stage regression shows a negative and highly significant coefficient for *YearsAtCompany*, making it a relevant explanatory variable.

Observations	1470
Dependent variable	NumCompaniesWorked
Type	OLS linear regression

F(1,1468)	20.88
R^2	0.01
Adj. R^2	0.01

	Est.	S.E.	t val.	p
(Intercept)	3.03	0.10	30.82	0.00
YearsAtCompany	-0.05	0.01	-4.57	0.00

Standard errors: OLS

Iterating in the same way of the previous IV (following the identical equation structure), the coefficient of the first regression named **IVyac** (Instrumental Variable *YearsAtCompany*) is fitted in the regression with the main variable of interest: attrition.

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(1)$	32.08
Pseudo-R^2 (Cragg-Uhler)	0.04
Pseudo-R^2 (McFadden)	0.02
AIC	1270.50
BIC	1281.09

	Est.	S.E.	z val.	p
(Intercept)	-6.23	0.92	-6.78	0.00
IVyac	1.67	0.33	5.07	0.00

Standard errors: MLE

The results are similar to the ones obtained through the previous model. The coefficient of the instrumental variable is higher (1.67) than the value of *NumCompaniesWorked* in the initial logistic regression. An employee that has been working for little time in the current company but has worked in several previous companies has an higher probability to resign compared to others who spent more time in the same workplace.

However, when both IVs are included in the general logistic regression, it's possible to see that only the only related to overtime work maintains its significance. Therefore it's possible to conclude that, even if it explains attrition through *NumCompaniesWorked*, the variable of years in the current company doesn't generate as much influence as the other variables previously included in the regression.

Observations	1470
Dependent variable	Attrition
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(10)$	126.98
Pseudo-R^2 (Cragg-Uhler)	0.14
Pseudo-R^2 (McFadden)	0.10
AIC	1193.60
BIC	1251.82

	Est.	S.E.	z val.	p
(Intercept)	-1.89	1.86	-1.01	0.31
IVttls	5.33	2.09	2.55	0.01
IVyac	0.30	0.57	0.52	0.60
MaritalStatus	-0.26	0.16	-1.61	0.11
Education	0.02	0.07	0.32	0.75
Gender	-0.15	0.15	-0.95	0.34
Age	-0.03	0.01	-2.89	0.00
MonthlyIncome	-0.08	0.03	-3.10	0.00
YearsInCurrentRole	-0.15	0.04	-3.92	0.00
YearsSinceLastPromotion	0.14	0.04	3.87	0.00
StockOptionLevel	-0.43	0.10	-4.13	0.00

Standard errors: MLE

Random Control Trial

RCTs are based on a basic regression structure:

$Yi = \alpha + \beta Di + \varepsilon i$

Where Di is a dummy variable associated with assigning the subject to a particular treatment or not. It must not be confused with the RDD, since in the RCT there's no threshold, but just the random assignment to the treatment. The estimates of the treatment effect (ATE) will be as follows:

$\beta[OLS = E[Yi|Di = 1] - E[Yi|Di = 0]$

Starting from the previous IV analysis related to the effect of training on overtime work, it would be advisable to go deeper in this relation and look which level of training helps preventing overtime work. Since the maximum value if training session present in the dataset is 6 (with the minimum of 0), the variable was split in 7 different treatment dummies, each one related to the level of training (TT0 … TT6, with "TT" meaning "training times"). A value equal or higher than 1, corresponding to variables *YearsInCurrentRole* and *YearsAtCompany*, was selected in the creation of the treatment dummies as well.

The equation of the regression is as follows:

$ATTi = \alpha + \beta Dij + \varepsilon i$

Where Dij is the sum of the dummies for the treatment, denoted by i, the individual and j, the treatment.

$\beta[OLS = E[Yi|Dij = 1] - E[Yi|Dij = 0]$

Will be the estimates.

The result of the regression shows that treatments corresponding to more sessions of training (**TT5**,**TT6**), reduce the probability of working overtime by over 80%.

Observations	1470
Dependent variable	OverTime
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(7)$	20.95
Pseudo-R^2 (Cragg-Uhler)	0.02
Pseudo-R^2 (McFadden)	0.01
AIC	1746.58
BIC	1788.93

	Est.	S.E.	z val.	p
(Intercept)	-0.79	0.14	-5.74	0.00
TT0	-0.11	0.36	-0.30	0.76
TT1	-0.50	0.33	-1.51	0.13
TT2	0.06	0.17	0.33	0.74
TT3	-0.22	0.18	-1.26	0.21
TT4	0.07	0.25	0.28	0.78
TT5	-0.82	0.31	-2.66	0.01
TT6	-1.05	0.43	-2.43	0.02

Standard errors: MLE; Continuous predictors are mean-centered and scaled by 1 s.d.

Conclusion

The paper had the aim to investigate the main causes of employee attrition. In order to do this, three models of causal inference were used: Regression Discontinuity Design (RDD), Instrumental Variables (IV) and Random Control Trials (RCT).

The results has showed that attrition is highly influenced by several factors. Through the RDD it was possible to define that workers with lower income have higher probability of suffering from attrition. At the same time, younger individuals (under 35) present the same behavior. Moreover, the analysis lead to the conclusions that these two variables are correlated: younger workers are paid less than older colleagues. One of the most important feature that presents an high level of explanatory power towards attrition rate attrition is overtime work. Employees suffering from overtime work are more likely to quit their current positions but through the analysis done by Instrumental Variables and Random Control Trials, it was faound that enough training may prevent the risk of working overtime. Another factor analysed through the research, with the help of IVs, was that employees who worked in several companies are more likely to quit their current workplace in the short term.

It is important for employers to address the issue of employee attrition. This phenomena isn't always a bad sign for the business, because it may be the result of reaching financial stability, shifting of resources or implementation of new production processes. However, being able to understand the causes of attrition in order to prevent high rates that would lead to critical situations within the company is a core mission of any business.