Regression analysis
Final Assignment

Intention to Stay Project Report

Amir Omezzine
March 2024

Prof: Professor: Yashar Bashirzadeh

1) Objective

For this report, we will be using STATA and SPSS to analyze and test five hypotheses (Appendix A) using the data set of "Intention to stay" provided by Dr. Yashar Bashirzadeh. The dataset consists of a sample of 976 employees at a global firm. The set of variables are 1) Salary level measure on a scale of 1-25 2) Perceived recognition by management measured on 1-10 scale] 3) Attachment to the company measured on 1-10 scale 4) Seniority measured on 1-10 scale and 5) Intention to stay for 5 more years in the company measured on 1-10 scale. In this report each hypothesis result will be fully interpreted and justified with statistical evidence using STATA and SPSS.

2) Conceptual Model and Regression Equation

In our study, we aim to explore the dynamics influencing an employee's intention to stay within a company, focusing on the role of salary, perceived recognition by management, attachment to the company, and how these relationships are potentially moderated by the employee's seniority. The conceptual framework (Figure1) guiding our analysis is rooted in the hypothesis that while salary (independent variable X) directly impacts an employee's intention to remain with the firm (dependent variable Y), this effect is also mediated by how much recognition they perceive from management (Mediator M1) and their sense of attachment to the company (Mediator M2). Moreover, we posit that the strength and perhaps the nature of these relationships could vary across different levels of employee seniority(Moderator W), introducing a layer of complexity in how we understand the interplay between these variables.

To dissect these relationships methodically, our analysis is structured around a series of regression equations that collectively aim to capture the direct, mediated, and moderated effects within our conceptual model. First we begin with examining the direct impact of salary on an employee's intention to stay, formulated as:

Direct Effect Model Equation1: Intention to Stay = $\beta 0$ + $\beta 1$(Salary) + ε

This equation serves as the foundation, hypothesizing that salary alone can be a predictor of an employee's intention to stay.

Delving deeper, we explore the mediating roles of perceived recognition and attachment to the company through two separate sets of sub-models:

Mediation by Perceived Recognition:

Model 1 Equation 2a: Perceived Recognition = $\beta 0$ + $\beta 1$(Salary) + ε
Model 2: Equation 2b: Intention to Stay = $\beta 0$ + $\beta 1$(Perceived Recognition) + $\beta 2$(Salary) + ε

Mediation by Attachment to the Company:

Model 3: Equation 3a: Attachment to the Company = $\beta 0$ + $\beta 1$(Salary) + ε
Model 4: Equation 3b: Intention to Stay = $\beta 0$ + $\beta 1$(Attachment to the Company) + $\beta 2$(Salary) + ε

For the moderation models, we explore how seniority influences the mediated relationships. For the two moderators we have two equations:

- Moderation on Perceived Recognition's Effect:
  - **Equation 4: Intention to Stay = β0 + β1(Perceived Recognition) + β2(Seniority) + β3(Perceived Recognition × Seniority) + β4(Salary) + ε**
- Moderation on Attachment's Effect:
  - **Equation 5: Intention to Stay = β0 + β1(Attachment to the Company) + β2(Seniority) + β3(Attachment to the Company × Seniority) + β4(Salary) + ε**

A comprehensive model (Figure2), that integrates both mediators and the moderator, combines elements of Equations 4 and 5, accounting for both mediators and their interaction with seniority, alongside the direct effect of salary.

- **Equation 6: Intention to Stay = β0 + β1(Salary) + β2(Perceived Recognition) + β3(Attachment to the Company) + β4(Seniority) +β5(Perceived Recognition × Seniority) + β6(Attachment to the Company × Seniority) + ε**
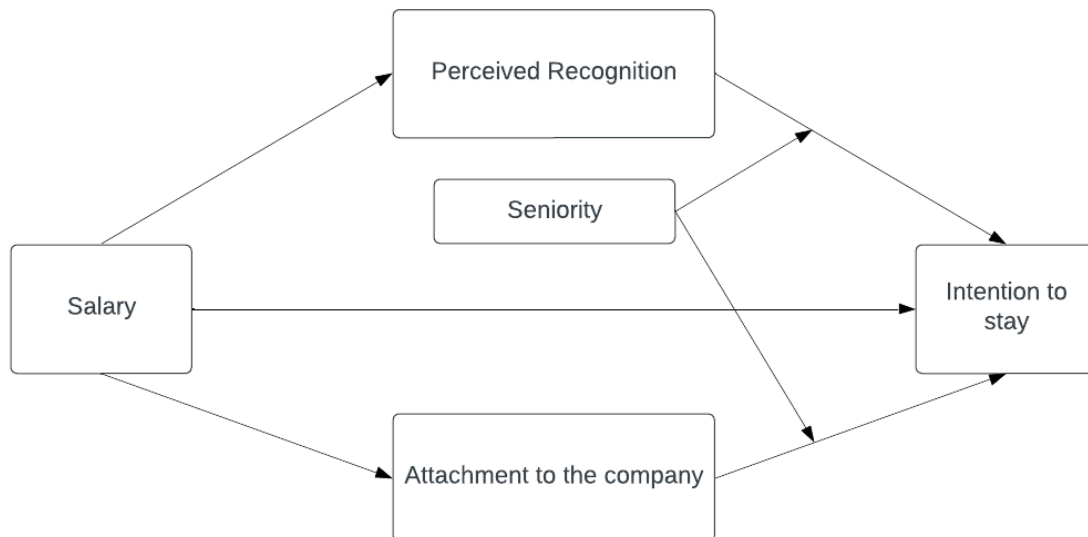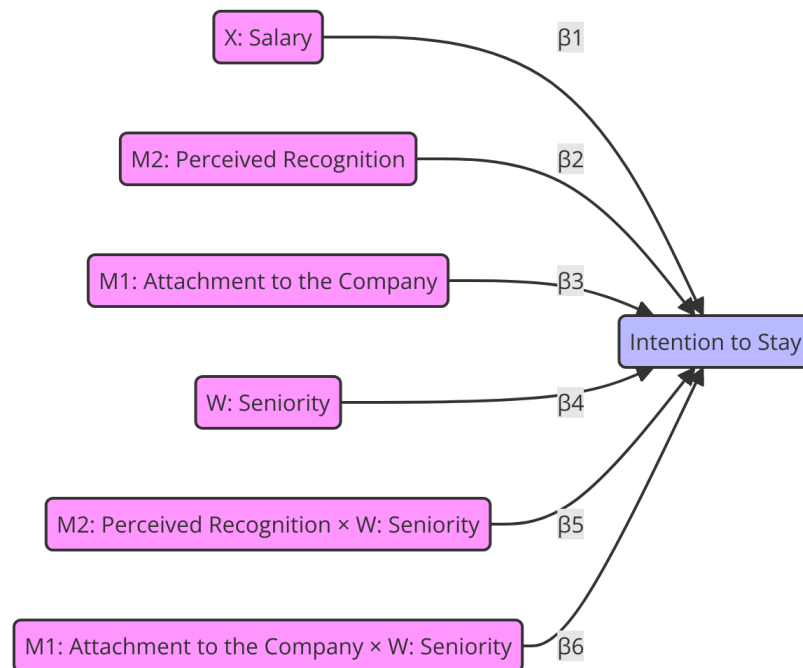
Figure1. Conceptual model.

Figure2. Comprehensive model (Equation 6)

3) Descriptive Statistics

Before we tested the hypothesis, we imported the dataset to Stata and had a descriptive analysis to be more familiar with its variables.

3.1) Missing Values

We examined the dataset for missing values, a crucial step to ensure the integrity and reliability of our findings. The dataset, comprising 976 observations from a global firm's employee sample, revealed a modest amount of missing data across several key variables (Table1). The percentages of missing data ranged from a negligible 0.20% for intention to stay up to 2.56% for both perceived recognition and attachment to the company.Given the relatively low percentages of missing data, we opted for listwise deletion as our method for handling missing values. This decision was underpinned by the assumption that the missingness was completely at random (MCAR), thereby minimizing the potential for bias in our subsequent analysis. Listwise deletion, while straightforward, was deemed appropriate due to the minimal impact it would have on our sample size and the overall integrity of our dataset. This approach allowed us to proceed with a clean dataset, ensuring that the analysis of the relationship between salary and intention to stay, our primary hypothesis, would be based on complete cases only. After addressing the missing values through listwise deletion, the refined dataset now consists of 920 observations.

Table1. Missing variables percentages.

Variable	Missing	Total	Percent Missing
salary	0	976	0.00%
recognit	25	976	2.56%
attachmt	25	976	2.56%
seniorit	16	976	1.64%
satisfac	9	976	0.92%
intentio	2	976	0.20%

3.2) Descriptive Variables

According to the descriptive of variables (Table2), the mean salary (M_{Salary} = 18.12) with a standard deviation (SD_{Salary}= 5.86), indicating a range of salary levels among the employees. Perceived recognition by management ($M_{recognit}$=7.31) with a standard deviation ($SD_{recognit}$= 1.59), suggesting varied perceptions of recognition. Attachment to the company ($M_{attachmt}$= 6.01) and a standard deviation ($SD_{attachmt}$=1.82), reflecting differing levels of employee attachment. Seniority ($M_{seniorit}$=5.77) with a standard deviation ($SD_{seniorit}$=1.92). The intention to stay for 5 more years ($M_{intentio}$= 6.84) and a standard deviation ($SD_{intentio}$ = 1.51).

Table2. Descriptive statistics

Variable	Obs	Mean	Std. dev.	Min	Max
salary	920	18.11848	5.861299	1	25
recognit	920	7.308696	1.590344	1	10
attachmt	920	6.005435	1.823845	1	10
seniorit	920	5.773913	1.919765	1	10
intentio	920	6.842391	1.508823	1	10
fitted_val~s	920	6.842391	.4710226	5.466725	7.395401
residuals	920	-6.48e-11	1.433417	-6.395401	4.533275

4) Data analysis

We conducted a regression analysis to examine the determinants of employees' intention to stay, using `intentio` as the dependent variable. The model included `salary`, `recognit` (Perceived Recognition), `attachmt` (Attachment to the Company), and `seniorit` (Seniority) as predictors. Additionally, we incorporated interaction terms `recog_sen` and `attach_sen` to assess how seniority moderates the effects of perceived recognition and attachment to the company on the intention to stay.

The regression analysis results (Table3) revealed that the model explains approximately 65.31% of the variance (R^2 = 0.6531, Adj R^2=0.6508,p<0.001), indicating a strong fit. Key findings include the positive and significant effects of salary, attachment to the company, and seniority on the intention to stay (p<0.05). Notably, perceived recognition did not show a statistically significant effect on its own (p = 0.69). The interactions between seniority with perceived recognition and attachment to the company were significant(p=0.00 < 0.05), suggesting that the influence of these factors on the intention to stay varies with employees' seniority levels. Specifically, the positive effect of perceived recognition on intention to stay increases with seniority, while the positive impact of attachment decreases as seniority rises.

Table3. Regression Equation 6.

Source	SS	df	MS			
				Number of obs	=	920
				F(6, 913)	=	286.51
Model	1366.43183	6	227.738639	Prob > F	=	0.0000
Residual	725.714905	913	.794868461	R-squared	=	0.6531
				Adj R-squared	=	0.6508
Total	2092.14674	919	2.27654705	Root MSE	=	.89155

intentio	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
salary	.0154854	.0055749	2.78	0.006	.0045442	.0264266
recognit	−.0212654	.0538693	−0.39	0.693	−.1269875	.0844566
attachmt	.5819574	.0532202	10.93	0.000	.4775093	.6864054
seniorit	.4088977	.0606727	6.74	0.000	.2898235	.5279719
recog_sen	.0287164	.0096797	2.97	0.003	.0097194	.0477135
attach_sen	−.0515247	.0088486	−5.82	0.000	−.0688907	−.0341588
_cons	1.504507	.311504	4.83	0.000	.8931596	2.115854

Before deciding to drop the variable perceived recognition (`recognit`) due to its lack of statistical significance in the initial regression analysis, it's prudent to first check the Ordinary Least Squares (OLS) assumptions.

5) Assumption of OLS estimation

5.1) Random sampling and no measurement error

In our analysis, the principles of random sampling and the absence of measurement error are critical yet challenging to fully verify. Random sampling ensures that every individual in the population has an equal chance of being selected, thereby making our sample representative of the broader population. This is essential for the generalizability of our findings. Similarly, the assumption of no measurement error is pivotal, as it ensures that the variables measured reflect their true values without bias or inaccuracies, which is crucial for the integrity of our regression results. However, in our specific case, directly testing for the fulfillment of these assumptions is not feasible.

5.2) Linearity: Multiple regression

In our case, linearity assumes that the change in the independent variable (y) (intention to stay) associated with a unit increase in an independent variable (x) (salary), while keeping other independent variables constant, is the same across all levels of salary. The assumption of linearity between the independent variables and the dependent variable was assessed by visually inspecting the scatter plot fitted values versus residuals (Figure3). Observing the distribution of residuals, which appear uniformly scattered without evident patterns or outliers (with few potential influential observations), supports the model's adherence to the linearity assumption, indicating its appropriateness for analyzing the intention to stay.

Figure3. Scatter plot of residuals
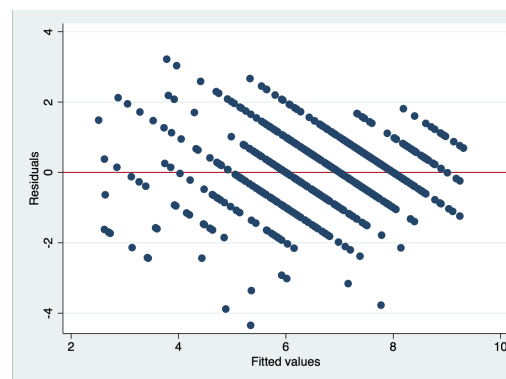
5.3) Variation in the Independent Variables

We ensured our independent variables, like salary and seniority, varied across our dataset. This variation is crucial; without it, regression analysis isn't possible. Our successful regression

analysis confirms that our variables have the necessary variation, indicating no violation of this assumption.

5.4) Zero conditional means

One critical assumption in OLS regression is the absence of endogeneity, ensuring that the error term exclusively captures the unexplained variation in the dependent variable not accounted for by the independent variables. Violating this assumption, where omitted variables correlate with both the dependent and independent variables, leads to biased and inconsistent OLS estimates. Directly detecting endogeneity from regression outputs is challenging, as it typically does not manifest in straightforward diagnostic tests. Addressing and identifying endogeneity often necessitates more sophisticated approaches, like employing instrumental variables, which can serve as proxies to isolate the independent variables' effect on the dependent variable, mitigating the bias introduced by endogeneity.

5.5) Homoscedasticity

Homoscedasticity is a key assumption in OLS regression analysis, which posits that the variance of the error terms is constant across all levels of the independent variables. This assumption ensures that the standard errors of the regression coefficients are accurately estimated, allowing for valid inference tests (e.g., t-tests and F-tests). To test for homoscedasticity in our regression model, we employ the Breusch-Pagan test. The test results (Table4) yielded a chi-square statistic of 179.21 with 1 degree of freedom and a p-value of less than 0.0001. Given the extremely low p-value, we reject the null hypothesis of constant variance among the error terms. This indicates the presence of heteroskedasticity within our regression model, suggesting that the variance of the error terms changes across different levels of the independent variable `intentio`.

Table4. Breusch- Pagan test for heteroscedasticity

Breusch−Pagan/Cook−Weisberg test for heteroskedasticity Assumption: Normal error terms Variable: Fitted values of `intentio`	
H0:	Constant variance
Chi2(1)	179.21
Prob> chi2	0.0000

5.6) *No Collinearity-Related Issue*

In our regression analysis, we addressed potential collinearity issues and assessed the influence of individual data points on our model. To ensure no single observation unduly influenced our results, we employed Cook's D, a diagnostic measure that identifies observations with a significant impact on the regression coefficients. A common benchmark is to scrutinize any data point with a Cook's D value greater than 1, as such points can disproportionately affect the model's estimates and conclusions. After calculating Cook's D for each observation in our dataset, we found that no observations exceeded this threshold (count = 0). This finding indicates that our model is free from undue influence by any single data point, enhancing the reliability and robustness of our analysis. Consequently, we can conclude that our regression results are stable and not distorted by outliers or highly influential observations

5.7) *Multicollinearity*

Our Variance Inflation Factor (VIF) test results (Table5) reveal significant multicollinearity among several predictors in our model, suggesting that the independent variables are not as distinct from each other as preferred for a robust regression analysis. Specifically, the interaction terms `recog_sen` and `attach_sen` exhibit exceptionally high VIF values (VIF>5) of 42.31 and 34.06, respectively, with corresponding Tolerance values well below the threshold of 0.1. This indicates a strong linear relationship between these variables and others in the model, potentially distorting the precision of our estimated coefficients. Additionally, `seniorit`, `attachmt`, and `recognit` show VIF values of 15.69, 10.89, and 8.49, respectively, further evidencing multicollinearity concerns. In contrast, `salary` demonstrates a VIF of 1.23, indicating no multicollinearity issue with this variable. With a mean VIF across variables at 18.78, our model suggests a need for corrective measures to address the significant multicollinearity observed, ensuring the reliability and validity of our regression analysis findings.

Table5. Collinearity diagnostics

```
Collinearity Diagnostics

                          SQRT                       R-
     Variable     VIF      VIF    Tolerance      Squared
------------------------------------------------------------
       salary     1.23     1.11     0.8100        0.1900
      recognit     8.49     2.91     0.1178        0.8822
      attachmt    10.89     3.30     0.0918        0.9082
      seniorit    15.69     3.96     0.0638        0.9362
      recog_sen    42.31     6.50     0.0236        0.9764
     attach_sen    34.06     5.84     0.0294        0.9706
------------------------------------------------------------
     Mean VIF     18.78


                          Cond
          Eigenval        Index
------------------------------------------
     1     6.6737        1.0000
     2     0.1890        5.9430
     3     0.0584       10.6939
     4     0.0474       11.8626
     5     0.0265       15.8636
     6     0.0034       44.2457
     7     0.0016       64.9348
------------------------------------------
     Condition Number        64.9348
     Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
     Det(correlation matrix)    0.0004
```

5.8) Normality of residuals

The residuals' normality was first examined through a kernel density plot (Figure 4), where the residuals' distribution seemed to look like a normal distribution. However, this visual similarity does not suffice to establish normality. The Shapiro-Wilk test (Table 6) provided a quantitative measure, giving a W statistic of 0.97789, which suggests a slight deviation from normality. But the p-value (Prob>z is 0.00000), indicating a statistically significant deviation from a normal distribution. The QQ plot (Figure 5) confirms these findings by displaying the residuals against a theoretically normal distribution. The QQ plot for these residuals shows a deviation from the line, particularly at the tails, which means there are more extreme values than what a normal distribution would predict. Combining the evidence from the kernel density estimate, the Shapiro-Wilk test, and the QQ plot, we can conclude that while the residuals display an approximation to normality, there is statistically significant evidence to suggest that they do not follow a normal distribution

Figure4. Kernel density estimate
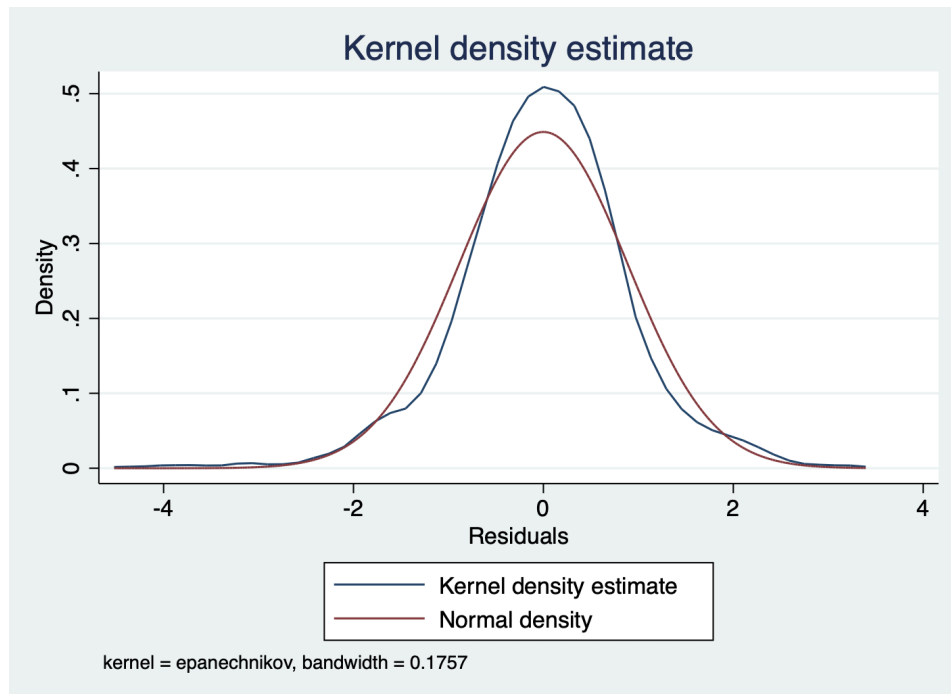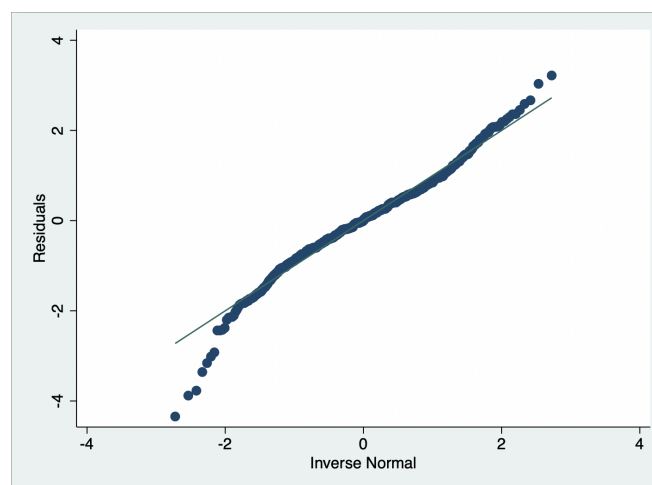

kernel = epanechnikov, bandwidth = 0.1757

Table6. Shapiro-Wilk W test for normal data

Variable	obs	W	V	Z	Prob>z
residuals	920	0.97789	12.923	6.315	0.0000

Figure5. Q-Q Plot.

6) The regression specification error: Omitted Variables

The Ramsey RESET test (Table7) for omitted variables, applied to our model with the intention to stay ($intentio$) as the dependent variable, yielded an F-statistic of 1.34 with a p-value of 0.2586. This test evaluates the null hypothesis that the model has no omitted variables by checking the significance of adding powers of the fitted values to the model. Given the p-value exceeds the conventional threshold of 0.05, we fail to reject the null hypothesis, suggesting our model does not suffer from omitted variable bias based on this test. This result indicates that, within the scope of the RESET test, our model adequately captures the essential predictors of the intention to stay without significant evidence of missing key explanatory variables.

Table7. Ramsey Reset test

Ramsey RESET test for omitted variables Omitted: Powers of fitted values of $intentio$	
H0	Model has no omitted variables
F(3, 910)	1.34
Prob > F	0.2586

7) Model Improvement

7.1) Robust standard error

In our regression analysis, we applied robust standard errors to address potential heteroscedasticity—variance in error terms not being constant across observations. Robust standard errors adjust the model's standard errors to be more reliable under such conditions, allowing for more accurate statistical inference. Despite this adjustment, the variable $recognit$ remained statistically insignificant, with a p-value of 0.824, as shown in the results (Table 8). This outcome suggests that, even after accounting for heteroscedasticity, $recognit$ does not have a statistically significant impact on the dependent variable, $intentio$. The application of robust standard errors was aimed at enhancing the reliability of our coefficient estimates under the assumption of heteroscedasticity. However, it does not inherently change the relationships between the independent variables and the dependent variable.

Table8. Linear regression after Robust standard errors

```
Linear regression                                    Number of obs    =         920
                                                     F(6, 913)        =      217.58
                                                     Prob > F         =      0.0000
                                                     R-squared        =      0.6531
                                                     Root MSE         =      .89155
```

intentio	Coefficient	Robust std. err.	t	P>|t|	[95% conf. interval]	
salary	.0154854	.0061876	2.50	0.012	.003342	.0276289
recognit	−.0212654	.0956596	−0.22	0.824	−.2090037	.1664728
attachmt	.5819574	.0839875	6.93	0.000	.4171264	.7467883
seniorit	.4088977	.0846489	4.83	0.000	.2427687	.5750267
recog_sen	.0287164	.0152278	1.89	0.060	−.0011691	.0586019
attach_sen	−.0515247	.0132839	−3.88	0.000	−.0775953	−.0254541
_cons	1.504507	.4951626	3.04	0.002	.5327175	2.476296

7.2) Mean-Centered

To address the issue of high Variance Inflation Factors (VIF) observed in our initial model, particularly with the variables recognit, attachmt, and seniorit, we implemented mean centering as a strategic solution. This technique involves subtracting the mean of each variable from its values, thereby centering the data around zero. The primary goal of mean centering in our context was to mitigate multicollinearity, especially important for models incorporating interaction terms, which can exacerbate the multicollinearity problem. Following mean centering, we generated two new interaction variables to explore the combined effects of recognition and seniority (Int1_sen_reco = recognit_c*seniorit_c), and attachment and seniority (Int2_sen_attach = attachmt_c*seniorit_c), on the intention to stay (intentio). These interaction terms allowed us to examine how the relationship between employee recognition (and attachment) and their intention to stay varies with their seniority level. The results, as detailed in Table 9, show a significant improvement in the model's performance. Notably, the coefficients for the mean-centered variables recognit_c, attachmt_c, and seniorit_c are statistically significant(p=0.000) , indicating a meaningful relationship with the dependent variable, intentio. This significance suggests that, after addressing multicollinearity through mean centering, the true effects of recognition, attachment, and seniority on the intention to stay become clearer.

Table10. Regression analysis after mean-centered applied.

Source	SS	df	MS			
Model	1366.43183	6	227.738639	Number of obs	=	920
Residual	725.714908	913	.794868464	F(6, 913)	=	286.51
				Prob > F	=	0.0000
				R-squared	=	0.6531
				Adj R-squared	=	0.6508
Total	2092.14674	919	2.27654705	Root MSE	=	.89155

intentio	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
salary	.0154854	.0055749	2.78	0.006	.0045442	.0264266
recognit_c	.1445407	.0251276	5.75	0.000	.0952261	.1938553
attachmt_c	.2844582	.0218739	13.00	0.000	.2415291	.3273872
seniorit_c	.3093491	.0202877	15.25	0.000	.2695331	.349165
Int1_sen_reco	.0287164	.0096797	2.97	0.003	.0097194	.0477135
Int2_sen_attach	-.0515247	.0088486	-5.82	0.000	-.0688906	-.0341588
_cons	6.630146	.1057818	62.68	0.000	6.422542	6.837749

Moreover, the VIF values post-mean centering and interaction term creation are substantially lower, with all variables showing VIFs well below the threshold of 5, indicating that multicollinearity is no longer a concern within our model (Table10). The mean VIF across variables has dropped to 1.69, further evidencing the effectiveness of mean centering in reducing multicollinearity.

Table11. VIF after mean-centered applied

Variable	VIF	1/VIF
recognit_c	1.85	0.541620
attachmt_c	1.84	0.543438
Int1_sen_r~o	1.77	0.564827
seniorit_c	1.75	0.570188
Int2_sen_a~h	1.69	0.593271
salary	1.23	0.810048
Mean VIF	1.69	

8) Final Regression Equation

the final regression equation for predicting the intention to stay (intentio) can be expressed as:

$$intentio = 1.5045 + 0.0155(salary) + 0.1445(recognit\_c) + 0.2845(attachmt\_c) + 0.3093(seniorit\_c) + 0.0287(Int1\_sen\_reco) - 0.0515(Int2\_sen\_attach)$$

The model's R-squared value of 0.6531 indicates that approximately 65.31% of the variability in employees' intention to stay is explained by the model. This high R-squared value suggests that the model does an effective job of capturing the factors that influence an employee's decision to stay with the company, with salary, recognition, attachment, seniority, and the specified interactions between these factors playing significant roles. With all predictors significantly impacting the dependent variable, we'll proceed to test our hypothesis on mediation paths using SPSS PROCESS Model 14 (Hayes 2022), directly addressing the roles these variables play in influencing employee retention decisions.

9) PROCESS Model 14

In our regression model we have a moderated mediation analysis with two mediators. To test the indirect effect, we used an SPSS PROCESS Model 14 (Hayes, 2022) with 5000 bootstrapped samples and 95% confidence interval with the "Intention" as the dependent variable, "Salary" as the independent variable, "Attachment" as the first and "Perceived recognition" as the second mediator and "seniority" as the moderator.

9.1) Direct effect of Salary on Intention to stay

The direct effect of salary on the intention to stay (intentio) is statistically significant (p = .0056), with a coefficient of .0155 and a standard error of .0056. This indicates a positive relationship between salary and employees' intention to stay, suggesting that as salary increases, so does the likelihood of employees intending to stay with the company. The confidence interval for this effect ranges from .0045 to .0264, reinforcing the significance of this finding. The statistically significant direct effect of salary on the intention to stay **supports our Hypothesis (H1)**.

9.2) Indirect effect of Salary on Intention to Stay

The analysis reveals that the effect of mean-centered attachment to the company (attach_c) on intention to stay is significant (β = .2845, SE = .0219, t(920) = 13.0044, p < .0001, 95% CI = [.2415, .3274]), **supporting H3**, which posits that attachment to the company mediates the relationship between salary and intention to stay. Similarly, the effect of mean-centered recognition by management (recog_c) on intention to stay is also significant (β = .1445, SE = .0251, t(920) = 5.7523, p < .0001, 95% CI = [.0952, .1939]), **supporting H2**, suggesting that recognition by management mediates the relationship between salary and intention to stay.

9.3) Moderated mediations

For the interaction between attachment to the company (attach_c) and seniority (seni_c), denoted as Int_1, the coefficient is -.0515 (SE = .0088, t = -5.8229, p < .0001, 95% CI = [-.0689, -.0342]). The Index of Moderated Mediation is -.0043 with a BootSE of .0013, and the bootstrap confidence interval ranges from -.0067 to -.0018, not including zero. The negative coefficient

and confidence interval indicate a significant moderated mediation effect where seniority negatively influences the mediation effect of attachment on the relationship between salary and intention to stay. Contrary to H5's expectation of a positive moderation, this result suggests that as seniority increases, the positive mediating effect of attachment decreases, thus **not supporting H5** in the anticipated direction.

The analysis of the interaction between recognition by management (recog_c) and seniority (seni_c), denoted as Int_2, reveals a coefficient of .0287 (SE = .0097, t = 2.9667, p = .0031, 95% CI = [.0097, .0477]). This indicates a statistically significant interaction effect, suggesting that the relationship between recognition by management and intention to stay varies with seniority. However, the Index of Moderated Mediation for this effect is .0034 with a BootSE of .0018, and the bootstrap confidence interval ranges from -.0002 to .0069.
The inclusion of zero in the confidence interval for the Index of Moderated Mediation means that **H4 is not supported**, which posited that the positive effect of recognition by management on the intention to stay is positively moderated by employees' seniority.

Appendix A

Hypothesis results

Hypothesis	Results
H1: Salary has a positive effect on the intention to stay	**Supported**
H2: The effect of salary on the intention to stay is mediated by the recognition by management	**Supported**
H3: The effect of salary on the intention to stay is mediated by the attachment to the company	**Supported**
H4: The positive effect of the recognition by management on the intention to stay is positively moderated by employees' seniority	**Not supported**
H5: The positive effect of the attachment to the company on the intention to stay is positively moderated by employees' seniority	**Not supported. The effect is negative**