Prediction in Human Recourse Managment

A gab between rhetoric and reality

Abhinna Shah Ariel Jeremi Wowor Bianca Ricci
Marcel Dieti

2025-07-10

Table of contents

List of Figures

List of Tables

1 List of Abbreviations

Abstract

Write your abstract here, summarizing your objectives, methods, results, and conclusion.

Abbreviation	Meaning
RSME	Root Mean Squared Error
MAE	Mean Absolute Error

2 Introduction to Prediction Modeling in HRM

Statistical modeling in the social sciences serves two distinct yet complementary purposes: explanation and prediction. Explanatory modeling, the dominant paradigm in Human Resource Management (HRM), focuses on testing theoretical propositions by estimating associations among constructs and evaluating in-sample model fit through metrics such as R^2, F-tests, and other standard indices (Shmueli, 2010). In contrast, predictive modeling assesses a model's capacity to generate accurate forecasts for new, unseen observations, typically using out-of-sample performance indicators such as the root mean squared error (RMSE)(Shmueli & Koppius, 2011).

This distinction carries important methodological consequences. While explanatory analysis emphasizes statistical significance and causal interpretation, predictive modeling prioritizes generalizability and accuracy, often employing techniques such as train–test splits or k-fold cross-validation to evaluate performance on new data (Hastie et al., 2013). Scholars have long warned against conflating explanation with prediction, as models optimized for in-sample performance may overfit the data and yield misleading insights when applied to new contexts (Forster, 2002; Forster & Sober, 1994).

This issue is particularly relevant in applied fields such as HRM, where empirical research frequently informs managerial decision-making. Prescriptive statements, such as the assumption that implementing a specific HR practice will enhance employee outcomes, implicitly rely on predictive validity. Yet, as Sarstedt and Danks (2021) highlight, a significant gap exists between these predictive ambitions and the methodological tools used in the field. In their review of the HRM literature, they find that while nearly all studies put forward practical recommendations, they rely exclusively on explanatory metrics for validation. This disconnect raises critical concerns about the real-world utility and robustness of the findings (Sarstedt & Danks, 2021).

2.1 Problem statement

The present thesis addresses this prediction–explanation gap by empirically assessing the extent to which explanatory performance translates into predictive accuracy. Specifically, we replicate and extend the job satisfaction model originally proposed by Drabe et al. (2015) using data from the International Social Survey Programme (ISSP) 2015. Our analysis focuses on three national contexts, namely Germany, Japan, and the United States, which represent leading industrial economies in Europe, Asia, and North America respectively, to examine both within-sample fit and cross-contextual generalizability.

We compare explanatory and predictive model performance using in-sample R^2 and both in-sample and out-of-sample RMSE. In doing so, we investigate (1) whether a model that fits its training data well also performs well in predicting new data from contextually similar samples, and (2) whether a model with strong explanatory power can generalize effectively across different national contexts.

In addition to this comparative evaluation, we integrate Necessary Condition Analysis (NCA) to identify structural prerequisites for achieving high job satisfaction. Unlike conventional regression approaches, which estimate average effects, NCA allows us to explore whether certain conditions are essential for the outcome to occur at all. By combining these perspectives, we aim to present a more comprehensive understanding of the methodological requirements for building robust and generalizable models in HRM research.

3 Modeling Job Satisfaction in Aging Workforces

Job satisfaction is a central concept in organizational research, linked to outcomes such as turnover intentions, absenteeism, and job performance (Drabe et al, 2015). While previous studies have explored the relationship between age and overall satisfaction levels, often suggesting a curvilinear or weakly positive association, relatively little is known about the specific determinants of job satisfaction across different age groups.

Addressing this gap, Drabe et al. (2015) investigate how the relevance of situational job characteristics varies with age. Based on prior empirical research and theoretical models of work design, they focus on seven key job facets that are consistently associated with job satisfaction across national contexts: income, job security, advancement opportunities, interesting work, autonomy, and the quality of relationships with supervisors and colleagues. These predictors span intrinsic, extrinsic, and social dimensions and have been validated in cross-national studies, such as those by Sousa-Poza and Sousa-Poza (2000).

To explain age-related differences, the authors draw on established frameworks in adult development and motivational psychology, including socio-emotional selectivity theory (Carstensen, 1995), the selection–optimization–compensation model (Baltes et al., 1999), and the lifespan theory of control (Heckhausen & Schulz, 1995). These perspectives suggest that older employees shift their focus from growth-oriented to maintenance- and loss-avoidance goals, prioritizing emotionally meaningful work and stable relationships over extrinsic rewards. Drabe et al.'s (2015) empirical findings confirm that older employees place greater value on collegial relationships and less emphasis on income, advancement, or job security, although these effects vary by country.

Overall, their model offers a theory-driven and empirically grounded framework for understanding

age-differentiated job satisfaction, particularly in the context of demographic change and international human resource management.

3.1 Evaluating Predictive Models using Regression Analysis

In assessing the predictive utility of a model, it is crucial to select metrics that capture different dimensions of performance. Among the various available metrics, R-squared (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are particularly valuable because they provide complementary insights into goodness-of-fit, error magnitude, and interpretability.

3.1.1 R-squared (R^2)

R-squared quantifies the proportion of variance in the dependent variable that is explained by the independent variables. It serves as a central indicator of in-sample goodness-of-fit and helps assess the strength of association within the estimation dataset (Sarstedt & Danks, 2021). R^2 is widely used due to its intuitive, scale-free interpretation as a percentage (ranging from 0 to 100), making it accessible for communicating how much of the outcome variance is accounted for by the model. However, a high in-sample R^2 is not necessarily indicative of strong predictive power. Models can exhibit high R^2 values by overfitting to sample-specific noise, which undermines generalizability to new data (Sarstedt & Danks, 2021).

To assess the explanatory power of a model we rely primarily on the R-squared (R^2) metric. R^2 measures the proportion of variance in the dependent variable that is explained by the independent variables included in the model (Sarstedt & Danks, 2021). As such, it offers a concise summary of how well the model fits the data used for estimation.

We use R^2 because it provides an intuitive and standardized measure of model fit, expressed as a percentage ranging from 0 to 100. A higher R^2 suggests that a larger share of outcome variation is captured by the model, making it particularly useful for communicating model performance within the training dataset. Its scale-free nature makes it suitable for comparing models across different samples or contexts.

3.1.2 Root Mean Square Error (RMSE)

RMSE is a commonly used metric for assessing predictive accuracy and reflects the standard deviation of residuals, i.e., the differences between predicted and observed values (Sarstedt & Danks, 2021). As a quadratic scoring rule, RMSE disproportionately penalizes larger errors, making it especially sensitive to extreme deviations. One of RMSE's primary advantages is that it is expressed in the same units as the dependent variable, which enhances its practical interpretability. For example, in a model predicting job satisfaction on a 1–7 scale, an out-of-sample RMSE of 0.9 conveys the typical magnitude of prediction error in real terms (Sarstedt & Danks, 2021). This makes RMSE a particularly informative metric for evaluating predictive utility on holdout data.

3.2 Understanding the Decline in Predictive Accuracy: Statistical and Contextual Perspectives

A frequent observation in predictive modeling, particularly within applied fields like Human Resource Management (HRM), is that model performance often degrades when applied to new data or shifted contexts. This reduction in predictive accuracy can be attributed to a combination of statistical factors and real-world contextual changes that affect the generalizability and stability of models.

One primary statistical explanation is overfitting. Overfitting occurs when a model is excessively tailored to its training data, capturing random noise in addition to meaningful patterns. While such models may display high explanatory power—such as a strong in-sample R^2—they tend to generalize poorly when applied to unseen data. This is because the patterns the model has learned may be idiosyncratic to the training set and not reflective of broader underlying relationships (Sarstedt & Danks, 2021).

Closely related to this is the issue of sample-specific variance. Each dataset represents a random sample drawn from a larger population and is thus influenced by its own unique idiosyncrasies. As a result, a model optimized on one sample may fail to replicate its performance on another due to sampling variability alone. This phenomenon is well-documented in studies where models with similar in-sample fit exhibit wide variability in out-of-sample performance metrics, such as RMSE, across different resampled datasets.

Beyond these statistical issues, there are substantive, real-world factors that contribute to the decline in predictive performance, particularly in the domain of HRM. One such factor is limited generalizability across contexts. Predictive models developed in one national, organizational, or cultural environment may not transfer well to others. For example, the drivers of employee satisfaction or turnover may differ significantly between countries such as Japan and the United States due to differences in institutional norms, social structures, and cultural expectations (Sarstedt & Danks, 2021). This context-specificity can severely limit the external validity of predictive models.

Another important contextual factor is temporal instability. The predictors of key HRM outcomes—such as job satisfaction or employee engagement—are not static over time. Economic shifts, labor market changes, technological advancements, and evolving workforce demographics can all alter the relevance or strength of previously identified predictors. Empirical evidence shows that models calibrated on data from one time period (e.g., 2005) often fail to maintain their predictive accuracy in subsequent years (e.g., 2015), highlighting how temporal changes can undermine model robustness (Sarstedt & Danks, 2021).

Together, these statistical and contextual explanations underscore the complexity of building predictive models that are both accurate and generalizable. Recognizing and addressing these challenges is essential for the responsible use of predictive analytics, especially in dynamic and context-sensitive fields such as HRM.

3.3 Assumptions Underlying Multiple Regression Analysis

In order to draw valid inferences from multiple regression analysis, several key statistical assumptions must be satisfied. Although it can be reasonably assumed that Drabe et al. (2015) tested the necessary assumptions in their original cross-national analysis of job satisfaction in the U.S., Japan, and Germany, this section underscores their theoretical relevance in the context of multivariate analysis. Meeting these assumptions is essential for obtaining unbiased, consistent, and efficient parameter estimates. Violations can lead to distorted coefficients, underestimated standard errors, inflated Type I error rates, and ultimately, erroneous conclusions (Hair et al., 2018, pp. 287–292). Accordingly, rigorous diagnostic testing forms an integral part of any regression-based empirical investigation.

The following subsections provide a theoretical overview of the four key assumptions underpinning

multiple regression analysis: (1) linearity of the phenomenon, (2) constant variance of the error terms, (3) normality of the error term distribution, and (4) absence of multicollinearity among predictors.

3.3.1 Linearity of the Phenomenon

The assumption of linearity posits that there is a straight-line relationship between each independent variable and the dependent variable. This implies that changes in the dependent variable are proportional to changes in the independent variables, holding other factors constant. If the true relationship is nonlinear and a linear model is applied, the resulting estimates will be biased and may misrepresent the strength or direction of effects (Hair et al., 2018, pp. 288–290).

Linearity is typically evaluated through a graphical diagnostic in which residuals are plotted against predicted values of the dependent variable. A random scatter of residuals supports the linearity assumption, whereas visible curves or systematic patterns suggest potential misspecification of the model. Such deviations may indicate the need for data transformation or the inclusion of nonlinear terms to adequately capture the relationship (Hair et al., 2018, pp. 288–290).

3.3.2 Constant Variance of the Error Terms

The assumption of homoscedasticity requires that the variance of residuals remains constant across all levels of the independent variables. When this condition is violated, known as heteroscedasticity, the residual variance changes systematically, which can result in biased standard errors and, consequently, unreliable hypothesis testing (Hair et al., 2018, p. 290).

To detect heteroscedasticity, researchers commonly inspect scatterplots of standardized residuals versus predicted values. A funnel-shaped or fan-like dispersion of residuals indicates potential violations. If heteroscedasticity is detected, several remedial strategies may be employed: transforming the offending variables using variance-stabilizing transformations, applying weighted least squares to adjust for varying variance, or computing heteroscedasticity-consistent standard errors to ensure robust inference (Hair et al., 2018, p. 290).

3.3.3 Normality of the Error Term Distribution

The assumption of normality holds that the residuals of the regression model should be normally distributed. While the estimation of regression coefficients does not require normality, hypothesis testing and the construction of confidence intervals depend on this assumption (Hair et al., 2018, p. 291).

Visual diagnostics such as histograms and Q–Q plots (i.e., normal probability plots) are commonly used to assess the normality of residuals. In a Q–Q plot, normally distributed residuals align closely along a 45-degree diagonal line, while systematic deviations, particularly in the tails, may indicate skewness or kurtosis. Although multiple regression is generally robust to moderate violations of normality, substantial departures may require data transformation or alternative estimation approaches (Hair et al., 2018, p. 291).

3.3.4 Absence of Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated, leading to redundancy in the model. It inflates standard errors, complicates the interpretation of coefficients, and can make estimates unstable. Variance Inflation Factors (VIFs) are commonly used to diagnose this issue, where VIF is the inverse of the tolerance value (VIF = 1/TOL). A VIF value above 10 is considered critical, although more conservative thresholds (e.g., VIF > 5 or even > 2.5) are often applied in practice depending on the sample size and context (Hair et al., 2018, pp. 312–316). Lower tolerance values indicate that a higher proportion of variance in a given predictor is shared with other variables, signaling potential multicollinearity.

3.4 Integrating Regression Analysis with Necessary Condition Analysis

Following the validation of linear regression assumptions outlined in the previous chapter, this research aims to extend the analysis of the replicated model through a comprehensive examination of covariates to determine the most significant determinants. When comparing variables measured in different units, establishing a common scale becomes essential. Consequently, standardized coefficients are employed to evaluate critical determinants within the job satisfaction model, as they

allow for assessment of each independent variable's relative magnitude and enable comparison of their predictive capabilities [Hair2018]. Therefore, this approach answers the question of which independent variable has the most influence on job satisfaction.

However, while standardized coefficients reveal factors are influential, they do not indicate whether specific conditions are prerequisite for particular outcomes. Relying solely on a traditional coefficient analysis with standardized coefficients would mean classifying the variable with the highest coefficient as the most important determinant and risking the neglect of other variables with lower values. This limitation is addressed through NCA, a methodology introduced by (Dul, 2016) that focuses on identifying necessary conditions within datasets. More specifically, these conditions represent fundamental prerequisites that must be present for particular outcomes to occur.

In 2020, Richter et al. extended NCA by integrating it with regression-based analysis using PLS-SEM (Richter & Sarstedt, 2020). In this study, instead of applying PLS-SEM, NCA is combined with a simple multiple regression model following the guidelines proposed by Richter et al. Details about the NCA methodology will be discussed in Chapter 3.

4 Methodological Approach for Replication and Extension

4.1 Assumption Testing of Underlying Multiple Regression Analysis

Prior to model estimation, we conducted diagnostic checks to evaluate whether the fundamental assumptions of multiple linear regression were met. The analysis was performed in JASP (Version 0.19.3) using a pre-processed version of the ISSP 2015 Work Orientations dataset, which had undergone reverse coding and case-wise deletion of missing values.

To ensure comparability across national contexts, the assumption testing was conducted on the pooled dataset, encompassing observations from the United States, Germany, and Japan. This unified diagnostic approach reflects the focus of our regression analysis on the combined sample. For the sake of parsimony and interpretability, we did not repeat the assumption checks separately for each country.

4.2 Replication Across Three Analytical Programs

Through multiple linear regression we estimated linear models that describe the relationship between job satisfaction and a set of predictors separately for each country, namely USA, Germany, and Japan, and the combined dataset. In order to ensure the comparability of the coefficients of the models it is necessary to standardize all continuous variables (e.g. education years, income) before estimating the linear models. According to Hair et al. (2018), standardization involves rescaling the variables to have a mean to zero and a standard deviation of one. This choice allows us to be able to

directly compare the relative strengths of each predictors' effect on job satisfaction, by quantifying the change in the outcome (measure in standard deviations) for a one standard deviation increase in the predictor.

Secondly, we evaluated each model using 10-fold cross-validation. For each country-specific dataset, the data was split into 10 subsets: in each of the 10 iterations, the model was trained on 9 of these folds and tested on the remaining one. This process was repeated so that each fold served once as the test set. After the all 10 rounds, we compute the average performance metrics (RMSE and R^2) across the 10 folds to get an estimate of the model's generalizability. This method allows to evaluate in a robust way the predictive performance and minimizing overfitting by repeatedly training the model on different subsets of the data and testing it on data it hasn't seen. This help to evaluate if the model captures specific noises of the training data, rather than the true underlying patterns.

To investigate whether better in-sample model fit (R^2) translates into higher predictive accuracy, we assessed both the explanatory and predictive power of models trained on contextually similar datasets. Specifically, we repeatedly drew random samples of 500 observations (1,000 replications) from a country-specific dataset. Each sample was split evenly into a training set and a test set, with 250 observations each. We then estimated a linear model on the training set and computed two metrics: the in-sample RMSE, measuring the error between predicted and actual job satisfaction within the training set, and the R^2, indicating the proportion of variance explained. Next, we generated predictions on the test set and calculated the out-of-sample RMSE to assess how the model performs on unseen data. Finally, we plotted the relationships between R^2 and out-of-sample RMSE, and between in-sample and out-of-sample RMSE, to examine how model fit relates to predictive performance.

To better understand how predictive performance behaves under similar levels of explanatory power, we also zoomed in on a narrow window around the median R^2 value (± 0.01). Within this range, we explored the spread of out-of-sample RMSE values to assess how much prediction accuracy can vary even when models show comparable in-sample fit.

Finally, to evaluate whether a model developed in one specific context can generalize well if applied to other contexts, we used each country-specific model to predict job satisfaction in the other countries' datasets, by computing RMSE on unseen data. Since predictions were on standardized scale, they were unstandardized using the training model's scale and mean to make predictions errors (RMSE) comparable to actual job satisfaction levels in the test country.

4.3 Data Source and Preprocessing Steps

4.3.1 ISSP 2015 Work Orientation Dataset

The ISSP 2015 Work Orientation module provides harmonized, publicly available survey data from multiple countries. We selected three national subsamples United States (USA; n = 915), Germany (GER; n = 899), and Japan (JPN; n = 635). The dependent variable, job satisfaction (job_sat), was measured on a 1–7 Likert scale. Independent variables matched those used by Drabe et al. These are as follows: (Drabe et al., 2015).

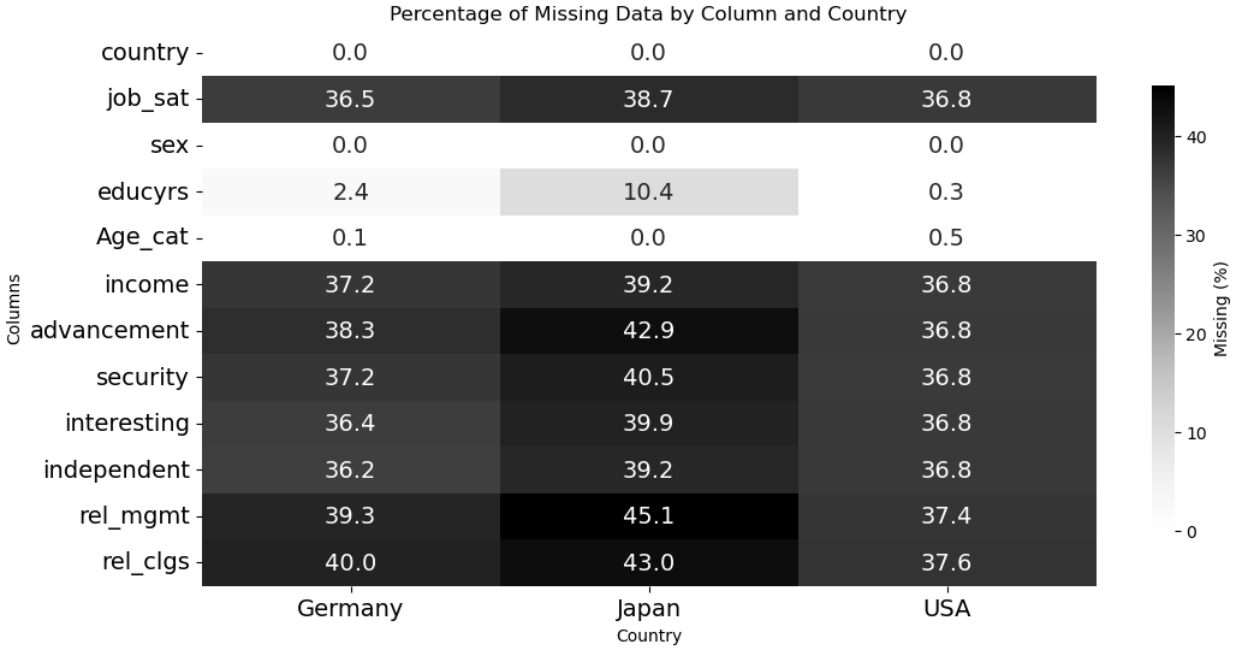
Table 4.1: Variables defining Job satisfaction

Variable	Intrinsic/Extrinsic
Gender (sex)	Intrinsic demographic
Age Category (Age_cat: ≤35, 36–49, ≥50)	Intrinsic demographic
Education (educyrs)	Intrinsic human-capital
Income	Extrinsic reward
Job security (security)	Extrinsic reward
Job interest (interesting)	Intrinsic job characteristic
Autonomy (independent)	Intrinsic job characteristic
Relationship with management (rel_mgmt)	Social work context
Relationship with colleagues (rel_clgs)	Social work context
Advancement opportunities (advancement)	Extrinsic reward
Job satisfaction (job_sat)	Outcome Variable

4.3.2 Handling Missing Data and Case-Wise Deletion

Survey items exhibited uneven missingness across countries (up to 35% for some items). In line with Sarstedt and Danks (2021), we applied case-wise deletion to maintain consistent sample composition for comparative modeling, accepting potential reductions in statistical power to preserve internal validity.

Figure 4.1: Heatmap of percentage Missing Data by column.



4.3.3 Reverse Coding

In the ISSP survey, higher numerical responses sometimes indicate less favorable conditions (e.g., "1 = Very satisfied" to "7 = Very dissatisfied"). To ensure that all predictors align directionally with job satisfaction, i.e., higher values lead to a higher outcome, we reversed scales so that greater values consistently denote more positive levels. This simplifies interpretation: a positive regression coefficient uniformly implies that increases in the predictor raise satisfaction.

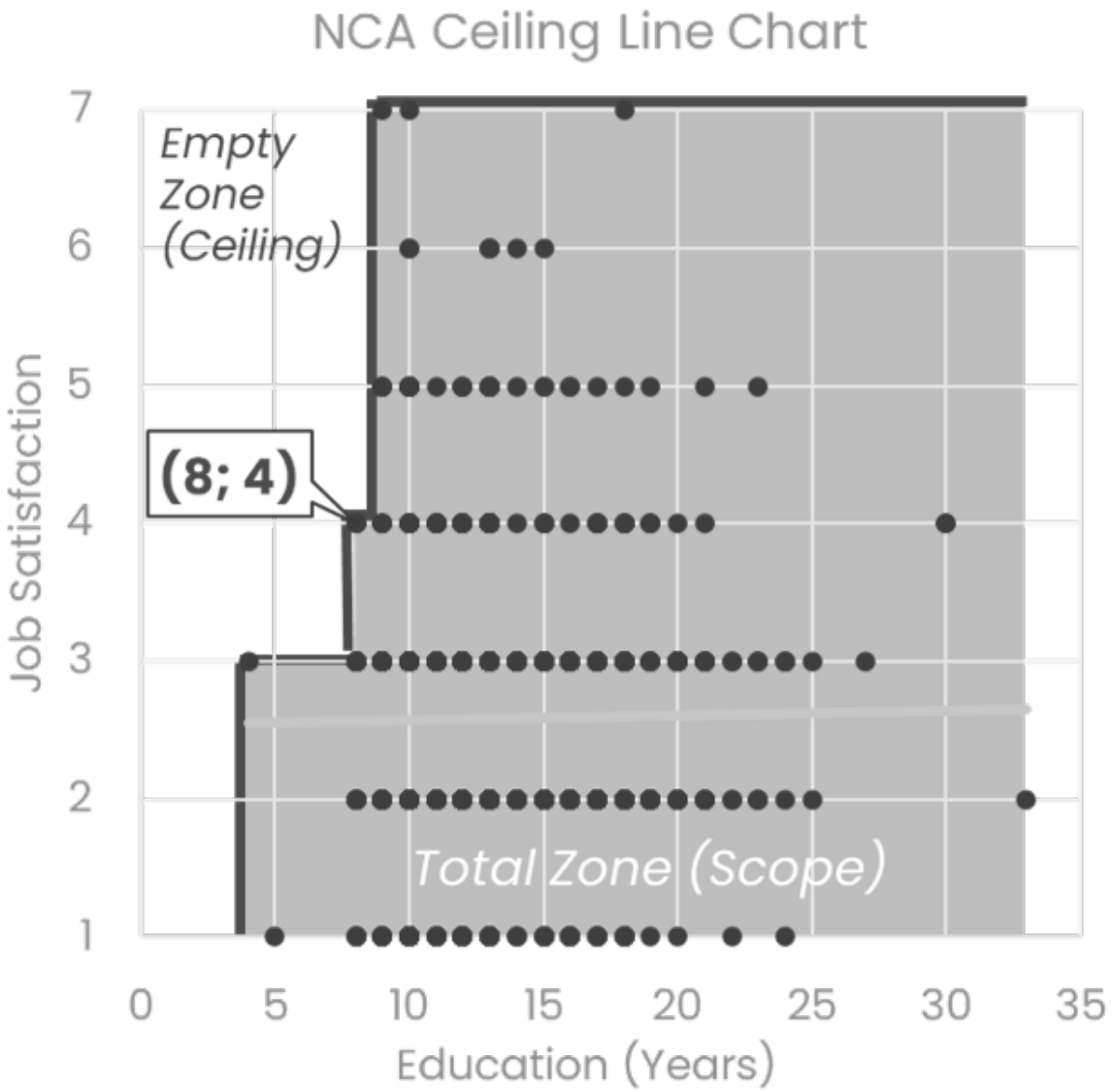
4.4 Extended Analysis of Determinants

To determine the most influential independent variable for job satisfaction, standardized coefficients are used for comparison. Following the examination of relationships between dependent and independent variables using these coefficients, NCA proceeds to identify areas in scatter plots that indicate the presence of necessary conditions (Richter & Sarstedt, 2020). In our model, the independent variables (X) presented in *Table Variables defining Job satisfaction* are treated as potential necessary conditions for Job Satisfaction (Y). This implies that if a necessary condition is not met, failure to achieve the desired outcome is guaranteed. However, NCA examines each variable individually, treating the necessary condition as operating in isolation and independently of context

(Richter & Sarstedt, 2020). Therefore, the absence of a necessary condition cannot be compensated for by other conditions or determinants.

NCA uses scatter plots to visualize the necessity relationship between an independent variable and an outcome, dividing the plot into two distinct areas, as shown in Figure: "Ceiling Line Chart for Education (Years) as Predictor of Job Satisfaction". The area where observations occur is called the total zone (scope), while the area without observations is referred to as the empty zone (ceiling).

Figure 4.2: Ceiling Line Chart for Education (Years) as Predictor of Job Satisfaction



An empty zone indicates a necessary condition, with its size providing a way to assess the strength of that requirement. To quantify this strength, (Richter & Sarstedt, 2020) describe two key NCA parameters, which are the ceiling accuracy and the necessity effect size d. The necessity effect size d measures how much of the outcome space is constrained by a necessary condition, ranging

from 0 to 1. Values of d are interpreted as small (0–0.1), medium (0.1–0.3), large (0.3–0.5), and very large (≥0.5) effects (Dul, 2016). While $d \geq 0.1$ is often used to support necessity hypotheses, its value indicates the substantive importance of the condition. Ceiling accuracy reflects the percentage of observations on or below the ceiling line. The CE-FDH line always achieves 100% accuracy, while lines like CR-FDH may have lower accuracy. Although no strict threshold exists, benchmarks (e.g., 95%) can help evaluate solution quality (Dul, 2016). However, because the data in this model are discrete rather than continuous, only the CE-FDH line is relevant, with 100% accuracy expected. Therefore, in this research, the necessity effect size measurement is the primary focus of the NCA.

5 Results: Findings/Analysis/Interpretation

5.1 Assumption Diagnostics of the Multiple Regression Model

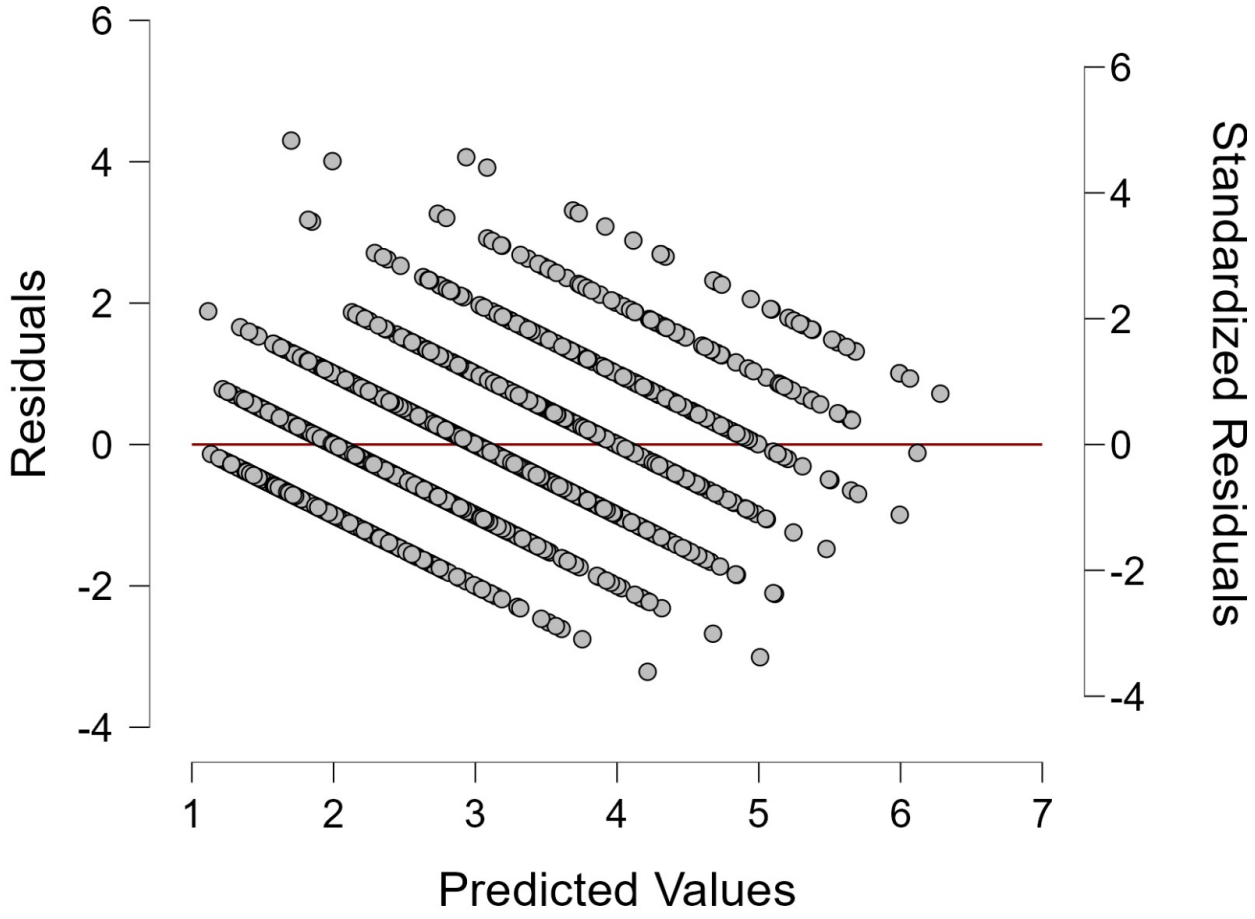
Prior to interpreting the regression results, we assessed whether the key assumptions underlying the ordinary least squares estimation were satisfied. This diagnostic step ensures the robustness and validity of subsequent inferences (Hair, 2018). The analysis was conducted for the pooled dataset, and all tests were performed using JASP (Version 0.19.3).

5.1.1 Linearity of the Phenomenon

To assess linearity, we examined the scatterplot of standardized residuals against predicted values (see Figure Standardized residuals against predicted values). The plot shows distinct parallel bands, which reflect the discrete 7-point Likert scale used to measure job satisfaction. Such patterns are expected in ordinal survey data and do not inherently violate the ass

Importantly, there is no systematic curvature or visible trend in the residuals. Therefore, the assumption of linearity appears to be reasonably satisfied.

Figure 5.1: Standardized residuals against predicted values



5.1.2 Constant Variance of the Error Terms

The same residual plot (Figure Standardized residuals against predicted values) was inspected to evaluate homoscedasticity. The spread of residuals is approximately uniform across the range of predicted values. No fan-shaped or funnel-like patterns are apparent, which would indicate heteroscedasticity.

Although the banding pattern persists due to the ordinal measurement scale, the residuals exhibit consistent variance, suggesting that the homoscedasticity assumption holds. Consequently, standard errors and significance tests derived from the model can be considered reliable.

5.1.3 Normality of the Error Term Distribution

Normality of residuals was assessed using a Q–Q plot and a histogram with density overlay (see Figures below). The Q–Q plot indicates that most standardized residuals fall closely along the diag-

onal reference line. A slight deviation at the upper tail is observed, but it remain within acceptable limits.

The histogram further supports this assessment by displaying a roughly symmetric, bell-shaped distribution. Taken together, both diagnostic plots suggest that the residuals are approximately normally distributed, thereby supporting the validity of significance tests and confidence intervals that depend on this assumption.

Figure 5.2: Histogram with density overlay

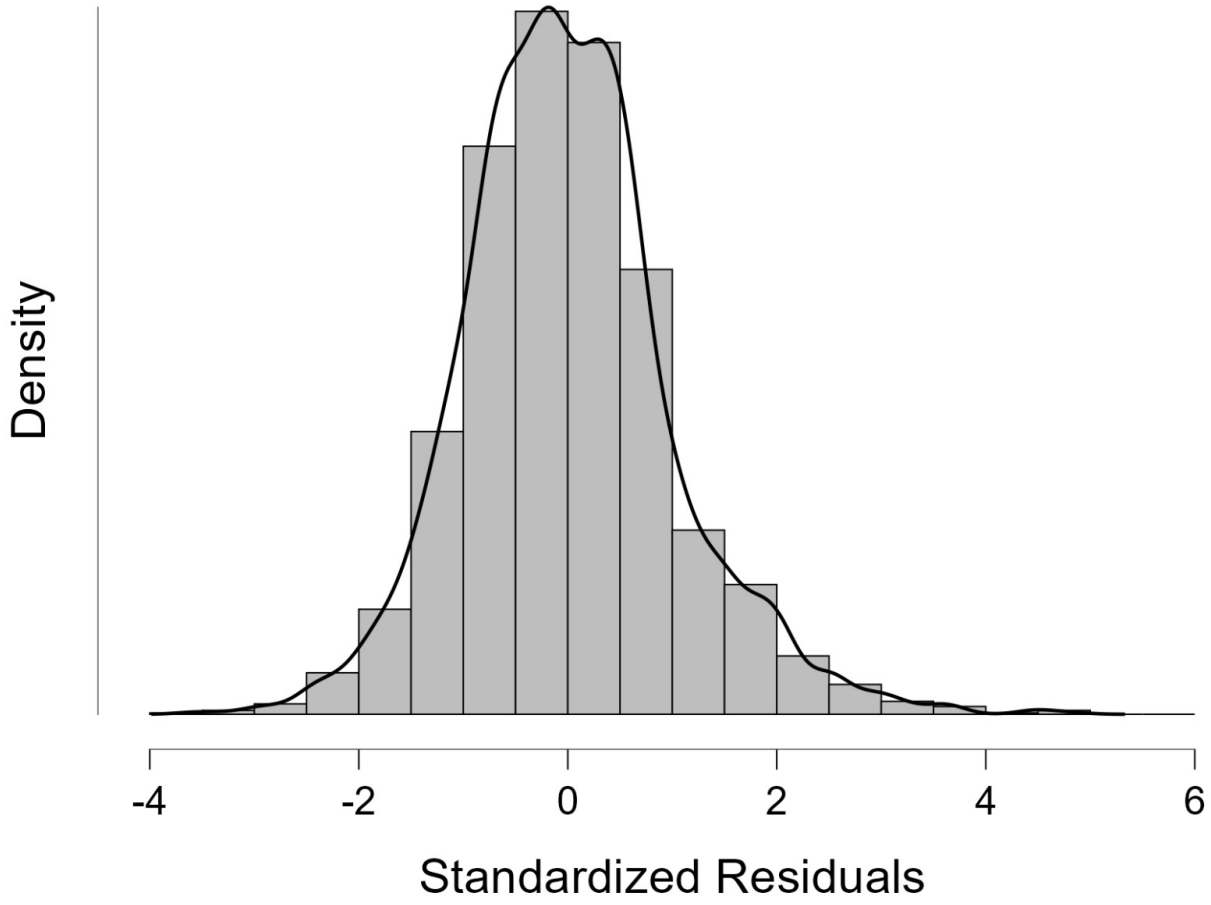
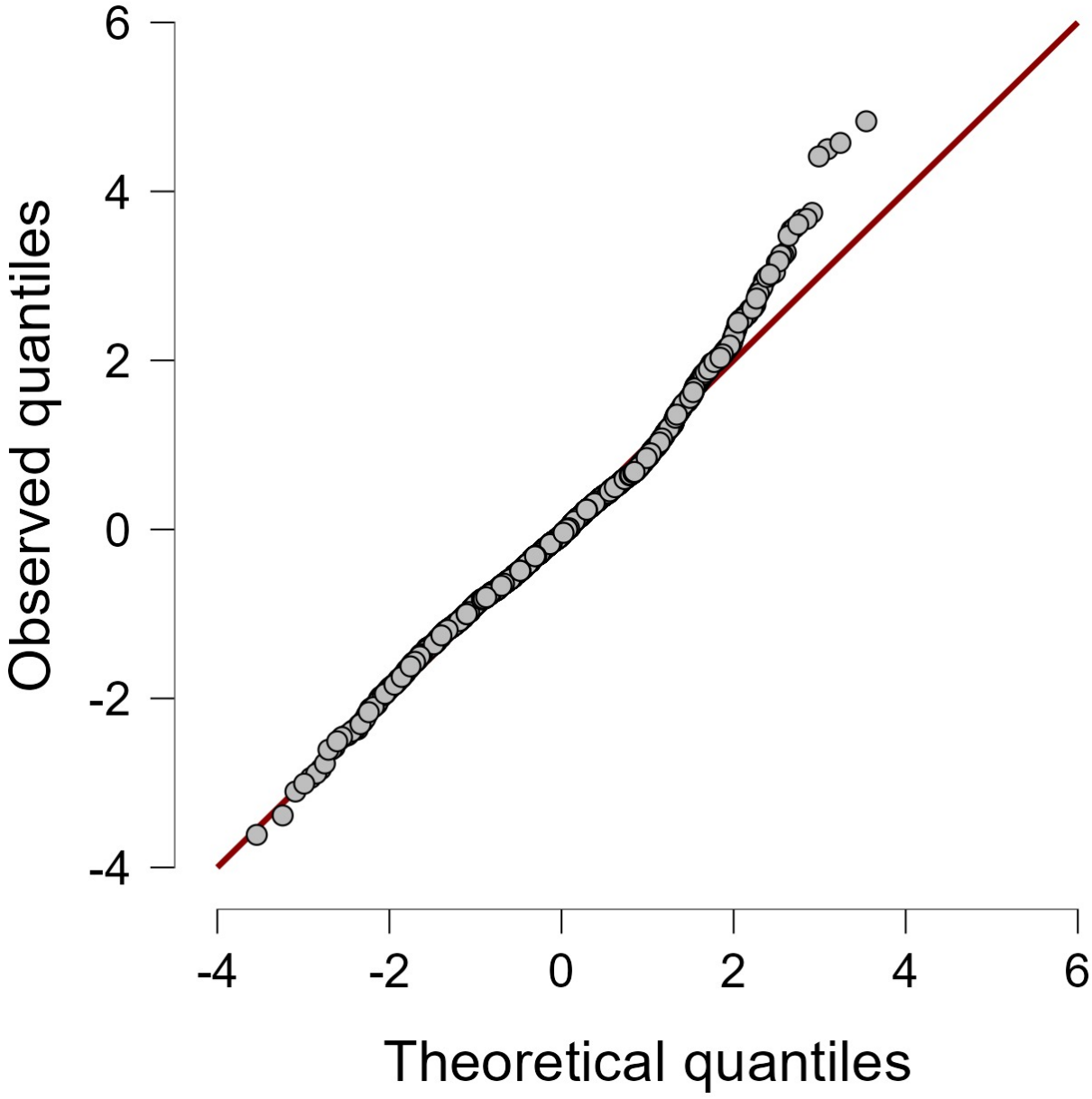


Figure 5.3: Q–Q Plot



5.1.4 Absence of Multicollinearity

To assess whether multicollinearity posed a threat to the interpretability of the regression coefficients, we examined the Variance Inflation Factor (VIF) and tolerance statistics for each predictor (see Figure 4). According to Hair et al. (2018), VIF values below 5 are generally considered acceptable, with more conservative thresholds set at 2.5 depending on context.

In the present analysis, all VIF values are well below the critical value of 5, ranging from 1.012 (Gender) to 1.30 (Good relationship with management). The corresponding tolerance values range

from 0.769 to 0.988, indicating that no predictor shares more than 23% of its variance with other independent variables. This suggests a low degree of redundancy among predictors.

The highest VIF observed (1.30 for "Good relationship with management") does not raise concern, as it still indicates a high level of predictor independence. Thus, multicollinearity is not present at problematic levels in the model. These findings support the statistical reliability and stability of the estimated regression coefficients and allow for valid interpretation of individual predictor effects (Hair et al., 2018, pp. 312–316).

5.1.5 Collinearity Statistics Table

The table below summarizes the collinearity statistics for the predictors in the multiple regression model:

Table 5.1: A Standardized coefficients can only be computed for continuous predictors.

Predictor	Coefficient (Unstandardized)	Standard Error	Coefficient (Standardized)a	t	p	Tolerance	VIF
Gender (reference: female)	-0.005	0.036		-0.128	0.898	0.988	1.012
Age (reference <=35)							
36-49	0.041	0.046		0.883	0.377	0.979	1.022
>=50	0.153	0.045		3.399	< .001		
Education	-0.02	0.006	-0.05	-3.482	< .001	0.98	1.021
Income	0.105	0.018	0.095	5.699	< .001	0.85	1.176
Advancement opportunities	0.044	0.019	0.041	2.296	0.022	0.786	1.272

Predictor	Coefficient (Unstandardized)	Standard Error	Coefficient (Standardized)a	t	p	Tolerance	VIF
Job security	0.096	0.018	0.088	5.442	< .001	0.877	1.14
Interesting job	0.448	0.02	0.381	22.3	< .001	0.83	1.205
Independent work	0.037	0.015	0.039	2.423	0.015	0.87	1.149
Good relationship with management	0.356	0.025	0.263	14.278	< .001	0.769	1.3
Good relationship with colleagues	0.162	0.028	0.105	5.819	< .001	0.788	1.269

5.2 Replication Results

In this chapter, we build on the models developed in the previous sections using the ISSP dataset to explore two key ideas.

First, we examine whether a model that fits its training data well (i.e., with high explanatory power) also performs well in predicting new data. Second, we assess whether a model trained in one specific context (such as one country) can generalize to different contexts.

All analyses were conducted in R and later replicated in Python to ensure robustness. Since results were consistent across both environments, in this chapter we report only the R-based plots for clarity. In the Appendix A are reported the results using Phyton.

5.2.1 Model Variability within Context

To address the first question, we focused on the German dataset and built 1000 models using repeated random sampling. In each replication, we drew a sample of 500 observations, then split it into two equally sized subsets: one for training and one as a holdout set.

Each model was trained on its training set, and we computed two in-sample metrics to evaluate explanatory power: R-squared and root mean square error (RMSE). We then used the same model to generate predictions on the holdout set and calculated the out-of-sample RMSE as a measure of predictive power.

Figures 1 and 2 illustrate how these two types of performance relate. In Figure 1, we plot R-squared (x-axis) against out-of-sample RMSE (y-axis); in Figure 2, the x-axis is the in-sample RMSE.

In both plots, we observe that models with similar in-sample performance can vary widely in their out-of-sample predictive accuracy. This means that high explanatory power does not guarantee high predictive power.

Interestingly, Figure 1 shows that higher R-squared values tend to be associated with higher prediction error, suggesting possible overfitting. Similarly, Figure 2 shows a negative trend between in-sample and out-of-sample RMSE: models that fit the training data very well often perform worse on new data.

To explore this further, we zoomed in on the models with an R-squared close to that of the original German model (R^2 between 0.366 and 0.386). Although these models all explain roughly the same proportion of variance, their prediction errors still vary greatly, from 0.75 to 0.94, meaning a 25% increase in error between the best and worst cases.

Figure 5.4: [1a] Relationship between (in-sample) R^2 and out-of-sample root mean square error (RMSE) [1.b] Relationship between in-sample RMSE and out-of-sample RMSE

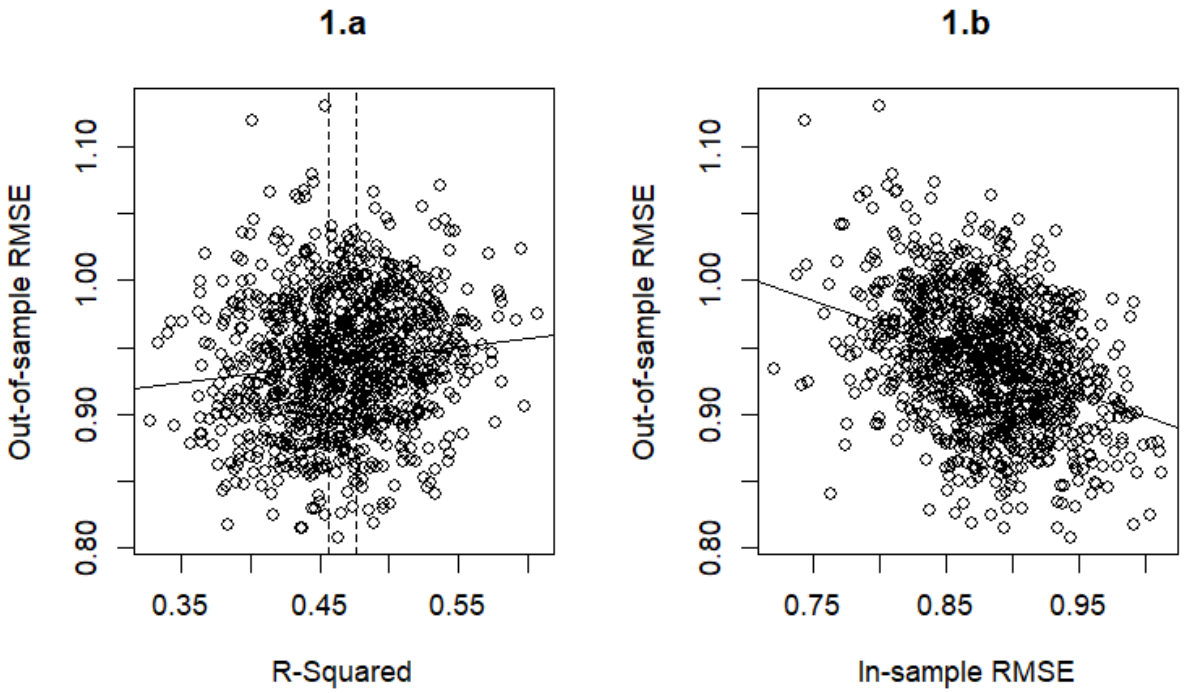
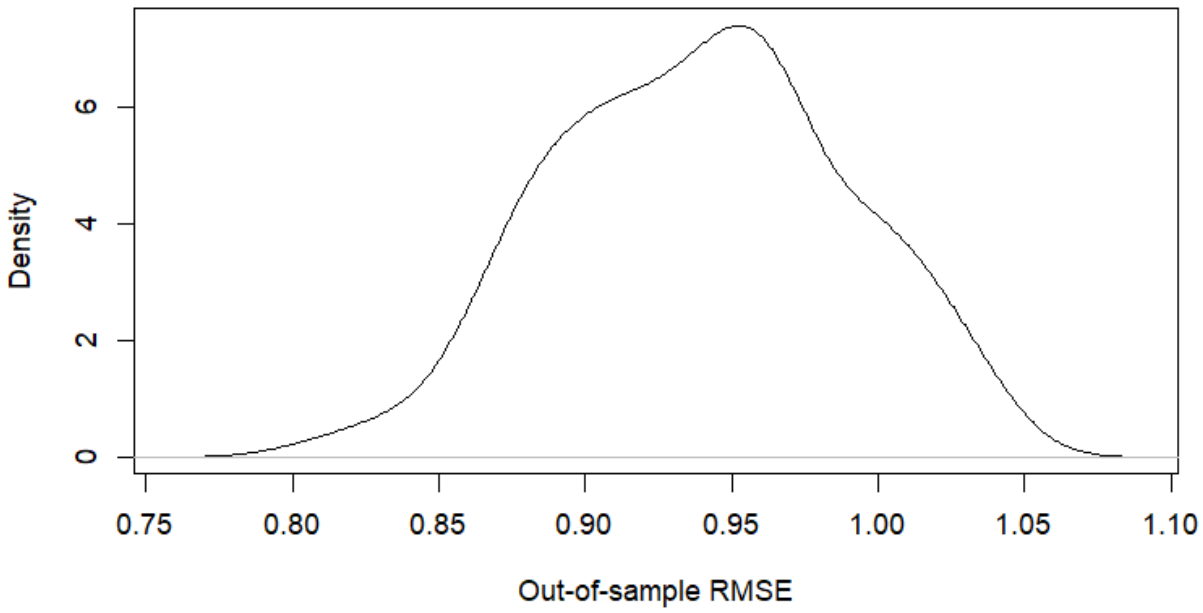


Figure 5.5: Density plot of predictive power for subsamples with $0.366 < R^2 < 0.386$. RMSE, root mean square error



5.2.2 Cross Country Generalizability of Predictive Power

In the second part of our analysis, we explore whether a model trained in one country can predict outcomes in other countries.

We used the full ISSP data for Germany, the USA, and Japan, building one model per country. The bar chart in Figure 4 shows the R-squared of each model on its own country's data, indicating how well it explains variation within its own context.

The heatmap in Figure 5 shows the out-of-sample RMSE when each country-specific model is used to predict data from the other countries. This allows us to compare generalizability across contexts.

From the results, the Japanese model shows to have the highest R-squared (0.508), meaning it fits its own data well. However, it performs poorly when predicting the USA data, suggesting it may be overfitting to context-specific patterns that don't transfer.

Furthermore, no other model predicts the Japanese data well, indicating that job satisfaction in Japan may follow idiosyncratic patterns not captured by models trained elsewhere.

The German model has the lowest explanatory power ($R^2 = 0.384$), but its data are reasonably well predicted by both the USA and Japan models. This may mean that the German model omits some relevant predictors or suffers from multicollinearity. At the same time, the predictors that matter in the USA and Japan might partially capture variation in the German context.

Figure 5.6: Ability of each country-specific model to fit its own country data

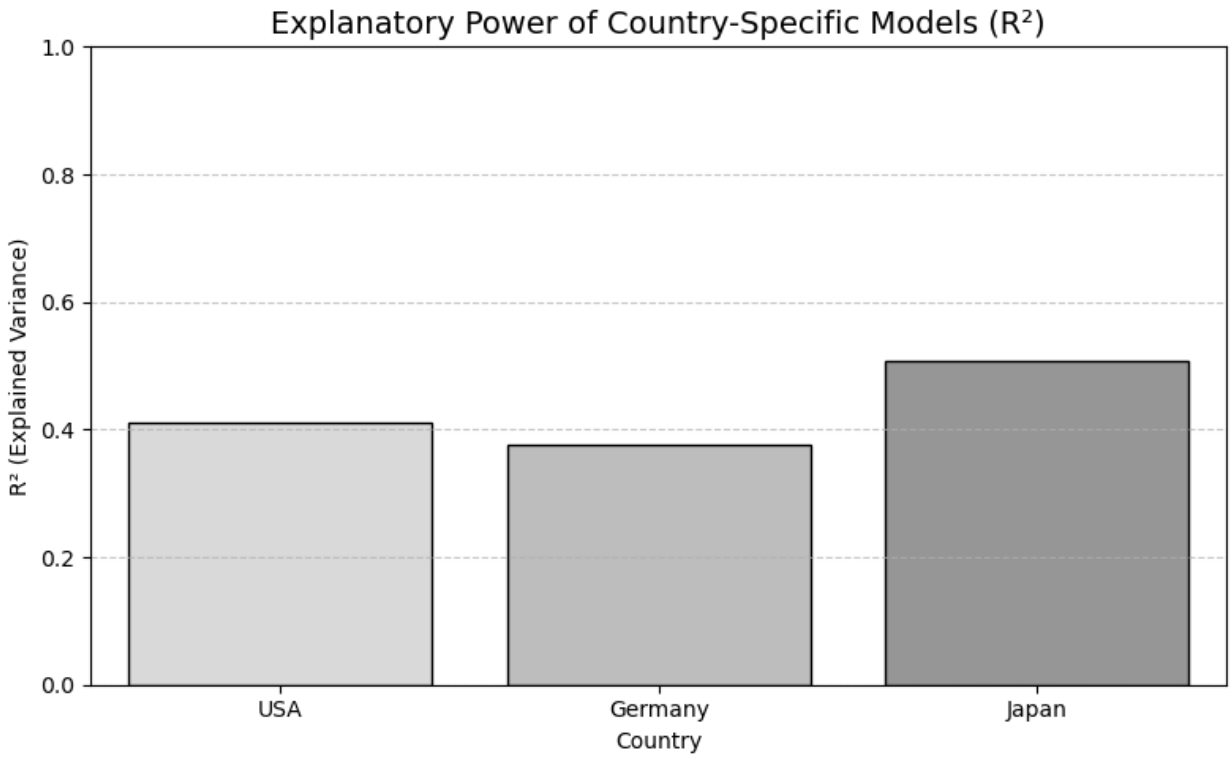
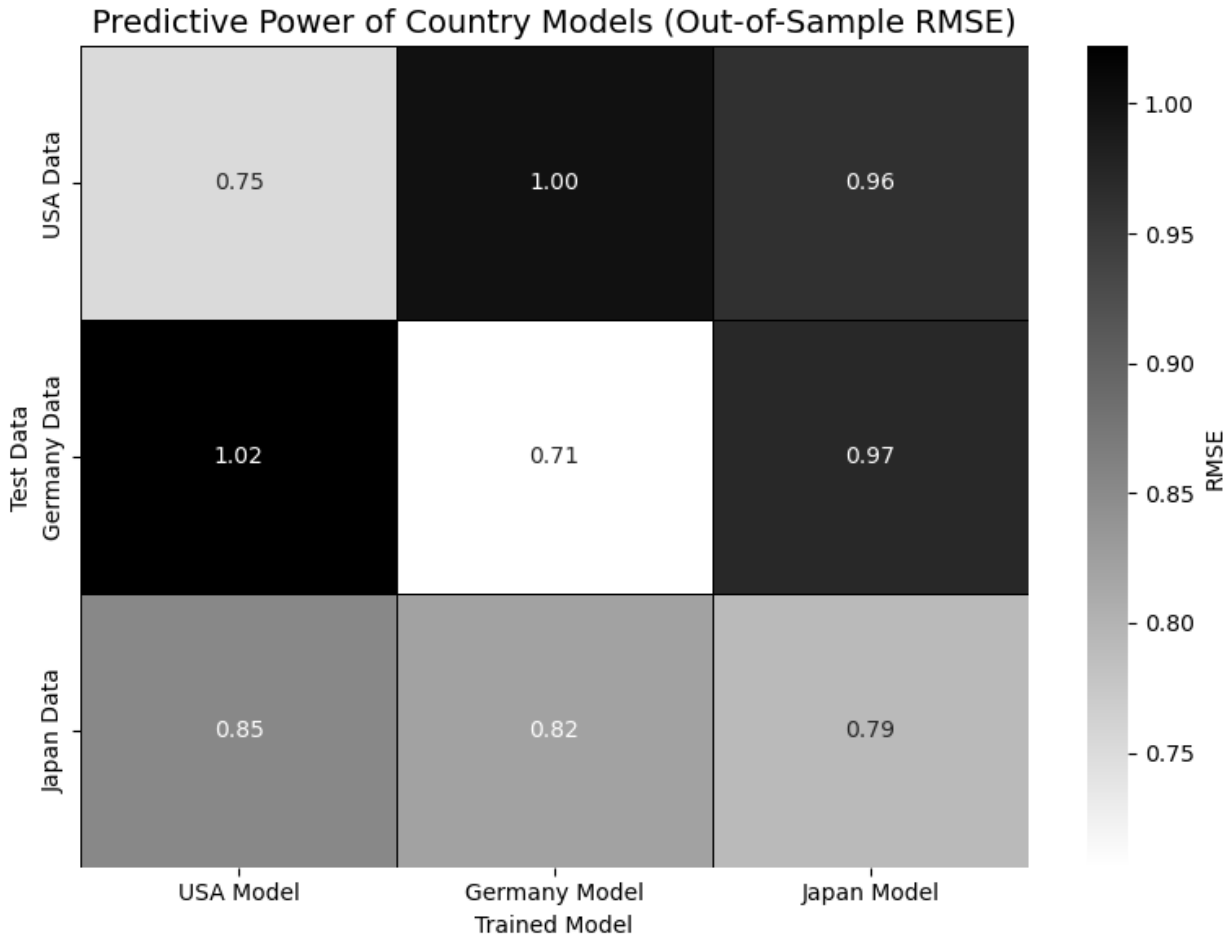


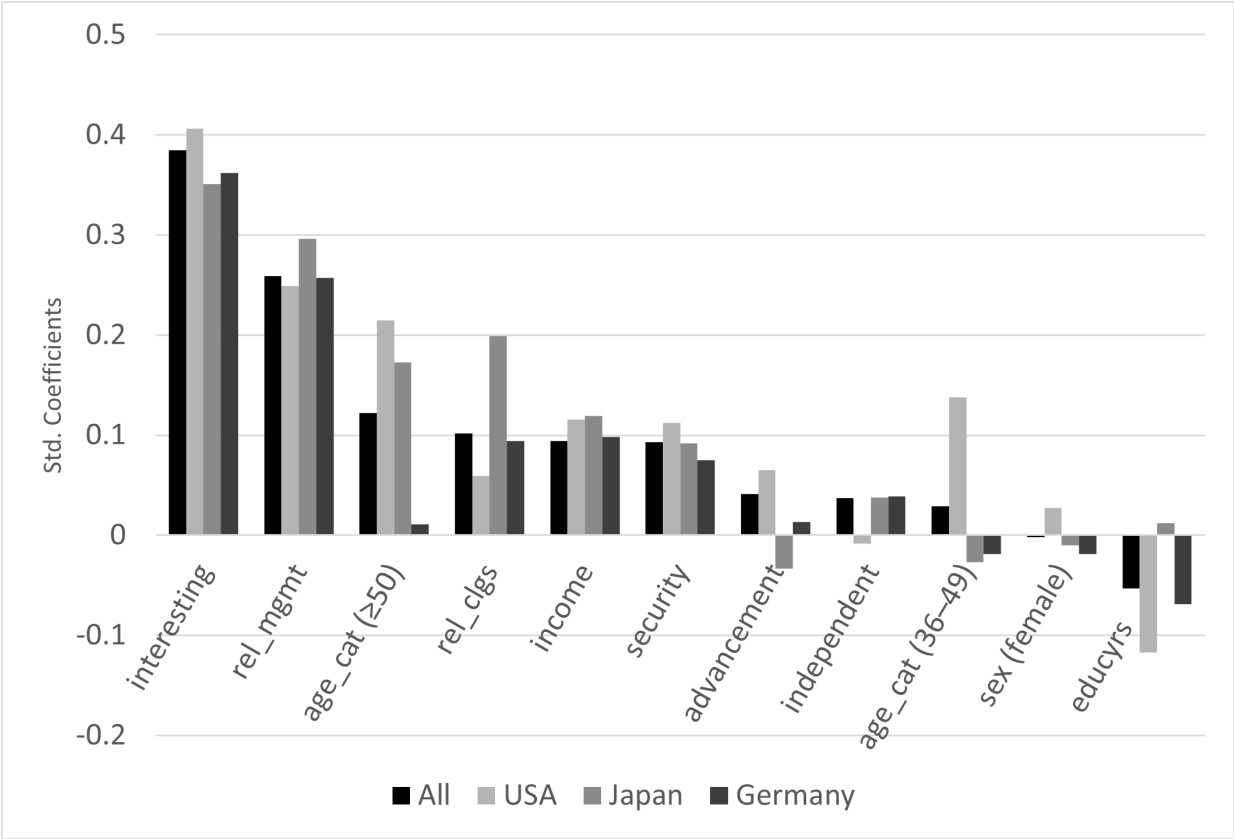
Figure 5.7: Predictive power of country-specific models



5.3 Extension Results

Figure xx illustrates the main determinants of job satisfaction, ranked from highest to lowest. Having an interesting job and maintaining a good relationship with management emerge as the strongest determinants, while education ranks among the weaker predictors. The overall ordering of predictors aligns closely with the original model from [paper xyz], showing only minor differences in the coefficient values. Although the order and direction of some predictors may vary across countries, these patterns remain consistent with those observed in the original model from [paper xyz]. The full set of coefficients across different languages (Python, R, and SmartPLS) is provided in **Table xy** in the appendix.

Figure 5.8: Cross-National Comparison of Standardized Regression Coefficients



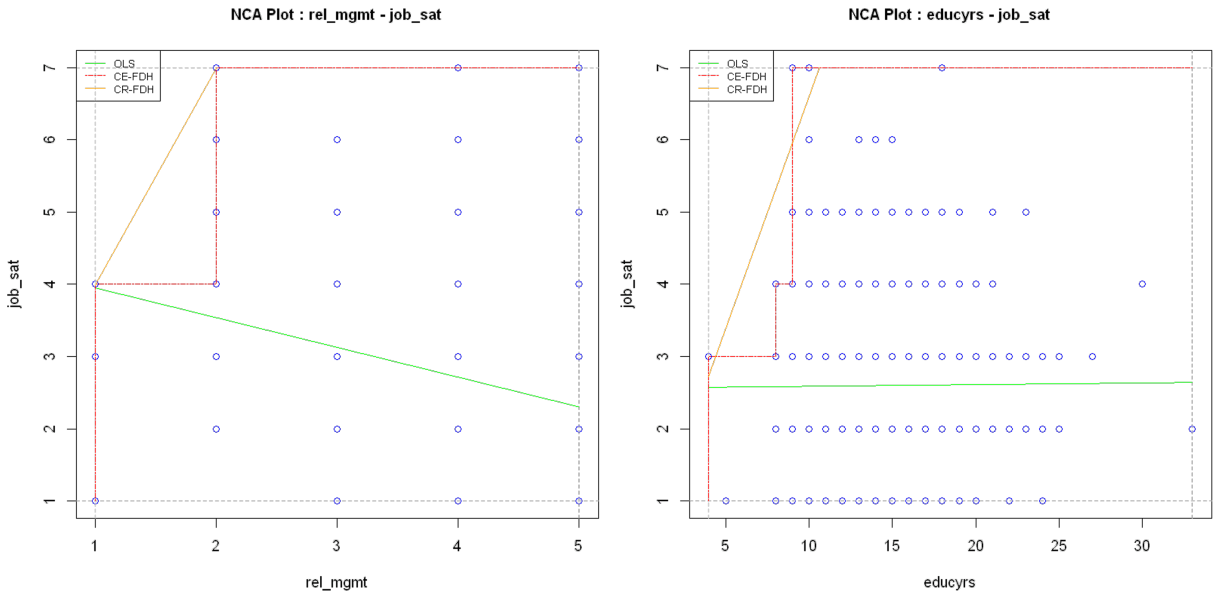
For the NCA analysis, control variables such as age group and gender are excluded because they are categorical. The variables interesting work, income, job security, working independently and opportunities for advancement show no empty zone in the scatter plots, with observations spread across all attribute levels. This indicates that there is no necessary condition for these independent variables. A summary table of all results is provided in the appendix.

In contrast, good relationship with management, good relationship with colleagues, and years of education*emerge as variables with necessary conditions. All of these variables have 100% ceiling accuracy and effect sizes greater than zero. However, the effect size for relationship with colleagues is 0.042, which is considered too small to be meaningful and can therefore be ignored. Relationship with management shows the highest effect size at 0.125 and is classified as the strongest determinant of job satisfaction. Years of education has an effect size of 0.109, ranking second in NCA.

These findings are further illustrated in Figure zz, where the step function ceiling line (CE-FDH) demonstrates specific threshold requirements for each variable. The left plot shows that achieving maximum job satisfaction requires a minimum relationship with management level of 2. When the relationship with management remains at the lowest level (1), job satisfaction cannot exceed level

4. The right plot reveals similar threshold patterns for education: reaching a job satisfaction level of 4 necessitates at least 8 years of education, while achieving maximum job satisfaction requires a minimum of 9 years of education. Additionally, the data indicates that a minimum threshold of 4 years of education is necessary to have any meaningful effect on job satisfaction.

Figure 5.9: NCA Plots



Comparing these NCA results with standard regression analysis reveals important distinctions. While both relationship with management and years of education demonstrate necessary conditions in NCA, they are classified differently based on their standardized coefficients in traditional analysis. Relationship with management is supported by its high standardized coefficient and is classified as a strong determinant in both approaches. Conversely, years of education is considered an insignificant determinant in standard regression analysis due to its low standardized coefficient, yet NCA reveals it as a necessary condition with meaningful predictive implications.

This extension enables us not only to identify the significance of determinants based on sufficiency but also to explore whether a condition must be present at all through necessity. This approach highlights that significant determinants can have a necessary condition or not, and the same applies to nonsignificant determinants.

Comparison of OLS regression results and Necessary Condition Analysis (NCA) findings.

Table 5.2: Comparison of OLS Regression Results and NCA Findings

Variable	OLS Results	NCA Results	Std. Coefficient	Effect Size
rel_mgmt	significant determinant	necessary condition with medium effect	0.343	0.125
educyrs	nonsignificant determinant	necessary condition with medium effect	-0.018	0.109
rel_clgs	slightly significant determinant	necessary condition with small effect	0.143	0.042
interesting	significant determinant	no necessary condition	0.433	0
income	nonsignificant determinant	no necessary condition	0.089	0
security	slightly significant determinant	no necessary condition	0.075	0
independent	nonsignificant determinant	no necessary condition	0.070	0
advancement	nonsignificant determinant	no necessary condition	0.021	0

6 Implications and Conclusions

6.1 Discussion of Key Findings

The analysis on models developed in similar contexts highlights the importance of jointly evaluating both explanatory and predictive power when assessing a model's performance. Our goal is to caution researchers against drawing prescriptive conclusions based solely on a model's in-sample fit, such as R-squared or in-sample RMSE, without verifying how well the model performs on unseen data.

As we have shown, models that explain a large proportion of variance in the training data do not necessarily perform well in predicting new observations. In fact, we observed that some models with high explanatory power exhibited relatively poor predictive performance, indicating a risk of overfitting.

A well-designed model should not only provide a good fit to its own data but also be able to generalize to new samples. However, there is no fixed relationship between in-sample metrics (e.g., R-squared, RMSE) and out-of-sample performance, which makes the predictive evaluation an essential step.

While explanatory modelling focuses on understanding relationships between variables, through the significance, direction, and size of coefficients, predictive modelling aims to accurately forecast new outcomes. These two goals are distinct, but not mutually exclusive.

By adopting an EP (Explanatory and Predictive) approach, researchers can strike a balance: developing models that both explain the data and support reliable prescriptive insights. Ultimately, when the objective is to derive actionable conclusions, predictive power must be explicitly assessed, explanatory power alone is not enough.

Secondly, by analyzing models developed on datasets contextually different, we show that models trained in one country do not necessarily generalize well to other countries. Patterns that hold in one cultural or institutional context may not apply elsewhere. We emphasize the importance of validating predictive performance across time, populations, or geographies, especially when a model is intended for broader use. Testing models on external data is key to understanding their generalizability and reliability.

6.2 Conclusion and Directions for Future Research

To conclude, while we were able to replicate the main results of the original study using R, the outcomes were not identical to those reported in the paper. This highlights a key challenge in replicating research: when data manipulation steps are not fully disclosed, even publicly available datasets and studies become difficult to reproduce with complete accuracy.

Secondly, our comparative analysis using different platforms such as R and Python showed that though these tools provide broadly consistent insights, the results can vary slightly depending on the platform used. This underlines the importance of transparency and standardization in research workflows.

Finally, our extension using NCA offered deeper insights into the model by identifying variables that are essential for achieving certain levels of job satisfaction. Through NCA we observed that certain variables which appeared insignificant in traditional analysis can nevertheless have a necessity effect.

This finding has practical implications for business practitioners. In our case, without NCA there is a risk of only focusing on good relationship with management if firms want to maintain job satisfaction while ignoring other variables. However, results of this study show that variables with low standardized coefficients can be important by providing necessity levels for certain outcomes. This approach can result not only in better predictive performance but also improved resource allocation decisions.

Future research could extend this model framework in several directions. Our work establishes the foundation for determining job satisfaction, which represents an important variable for predicting worker turnover, though this connection was not explored here. A logical next step would be

extending this model to worker turnover and using NCA to examine job satisfaction as a necessary condition for turnover outcomes. Researchers applying regression models should follow the framework of checking assumptions of linearity first. Future studies might identify other variables besides job satisfaction, such as organizational commitment, that influence worker turnover. Such results would reveal not only which variables are important but also what necessary conditions are required for specific outcomes.

7 References

8 Appendix

8.1 Comparison of Regression Results Across Statistical Software and Countries

The comparative analysis presented in the table validates the findings of the original study by replicating its regression models for job satisfaction across the USA, Japan, and Germany. The replication, conducted using both JASP and Python, shows a high degree of consistency in the results, confirming that the findings are robust and not dependent on the statistical software used.

Across all three countries, the model's predictive power varies, with Japan exhibiting the highest R-squared value (≈0.51), followed by the USA (≈0.44) and Germany (≈0.38). Despite these differences in overall model fit, certain predictors remain consistently significant. Specifically, having an "interesting job" and maintaining a "good relationship with management" emerge as the most influential factors for job satisfaction in all nations analyzed.

Table 8.1: Estimation of country specific models and predictive power and comparison of the results of the original paper with the one generated by us with JASP and Phyton.

Paper	JASP			Python			Coefficients		
	USA	JP	DE	USA	JP	DE	USA	JP	DE
Gender (ref: female)	0.027	-0.01	-0.019				0.027	-0.020	-0.014
Age (ref: <35)									

Paper	JASP			Python			Coefficients		
36–49	0.138	-0.027	-0.019				-0.138	0.037	-0.003
≥50	0.215	0.173	0.011				-0.215	-0.169	-0.023
Education	0.117	0.012	-0.069	-0.114	0.019	-0.069	0.116	-0.020	0.068
Income	0.116	0.119	0.098	0.117	0.123	0.098	-0.118	-0.124	-0.095
Advancement oppor-tuni-ties	0.065	-0.033	0.013	0.064	-0.042	0.023	-0.065	0.043	-0.024
Job se-curity	0.112	0.092	0.075	0.112	0.082	0.073	-0.112	-0.085	-0.076
Interesting job	0.406	0.351	0.362	0.405	0.35	0.349	-0.405	-0.366	-0.355
Independent work	0.008	0.038	0.039	-0.005	0.036	0.053	0.005	-0.039	-0.055
Good rel. w/ man-age-ment	0.249	0.296	0.257	0.247	0.297	0.265	-0.248	-0.304	-0.262
Good rel. w/ col-leagues	0.059	0.199	0.094	0.06	0.208	0.091	-0.060	-0.214	-0.090
R²	0.444	0.519	0.384	0.444	0.515	0.387	0.444	0.515	0.386
Adj. R²	0.437	0.511	0.377	0.437	0.507	0.38	0.437	0.507	0.378
n	915	635	899	913	668	933	913	666	929
CV R²	0.435	0.5	0.374				0.421	0.482	0.359
CV RMSE	0.912	0.949	0.876				0.917	0.955	0.792

8.2 Python Results

To illustrate that the results described in the thought experiment are robust for different environments, we report the results obtained by replicating the analysis of the research using Phyton. The two graphs below, showing the relationships between in-sample metrics (R-squared and in-sample RMSE) and the out-of-sample RMSE, confirm the absence of a fixed relationship between a model's explanatory power and its ability to predict unseen data, consistent with the results obtained using R.

Figure 8.1: OS RMSE vs R-squared using python

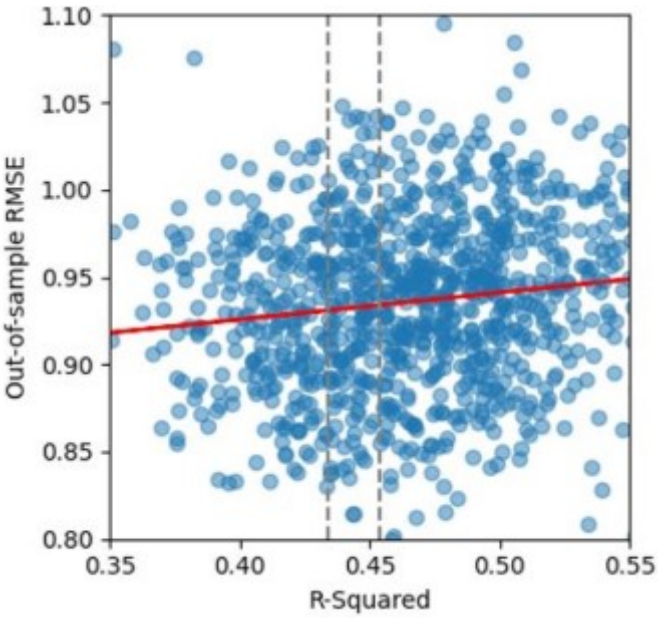
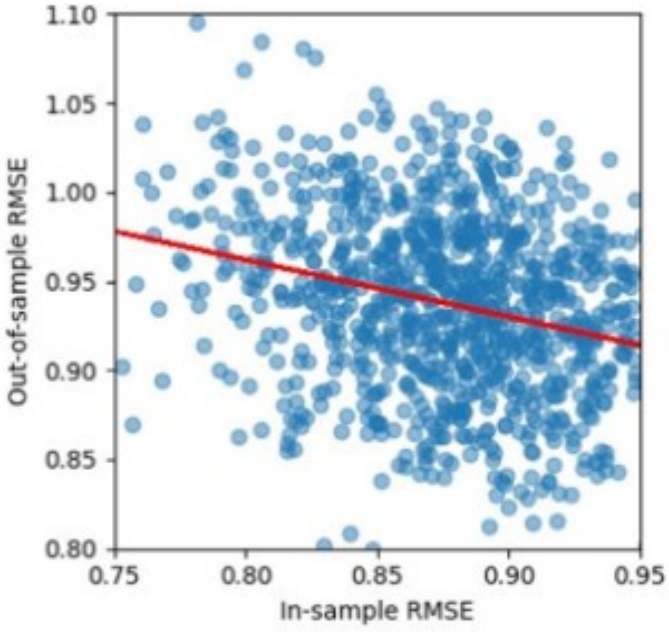


Figure 8.2: OS-IS RMSE VS IS RMSE using python



Moreover, the red regression lines highlight that models that fit the training data very well can still perform poorly when predicting unseen observations, which is a typical indication of overfitting.

The density plot of RMSE values for models with similar explanatory power still shows that, despite having similar R^2, their out-of-sample predictive performance varies considerably, once again confirming the results obtained with R.

By replicating in Python the second analysis evaluating how well country-specific models predict data from other countries, we obtained results consistent with those from R. While the specific behaviour of each model, such as Japan's strong in-sample fit and weak out-of-sample performance, or Germany's low in-sample fit, was already discussed in the previous chapter, the Python replication confirms the same cross-country patterns. As shown in the bar chart and heat map below, these results support the idea that models trained in one national context do not automatically transfer well in terms of predictive accuracy when applied to other contexts.

Figure 8.3: Explanatory power of country specific models (R^2)

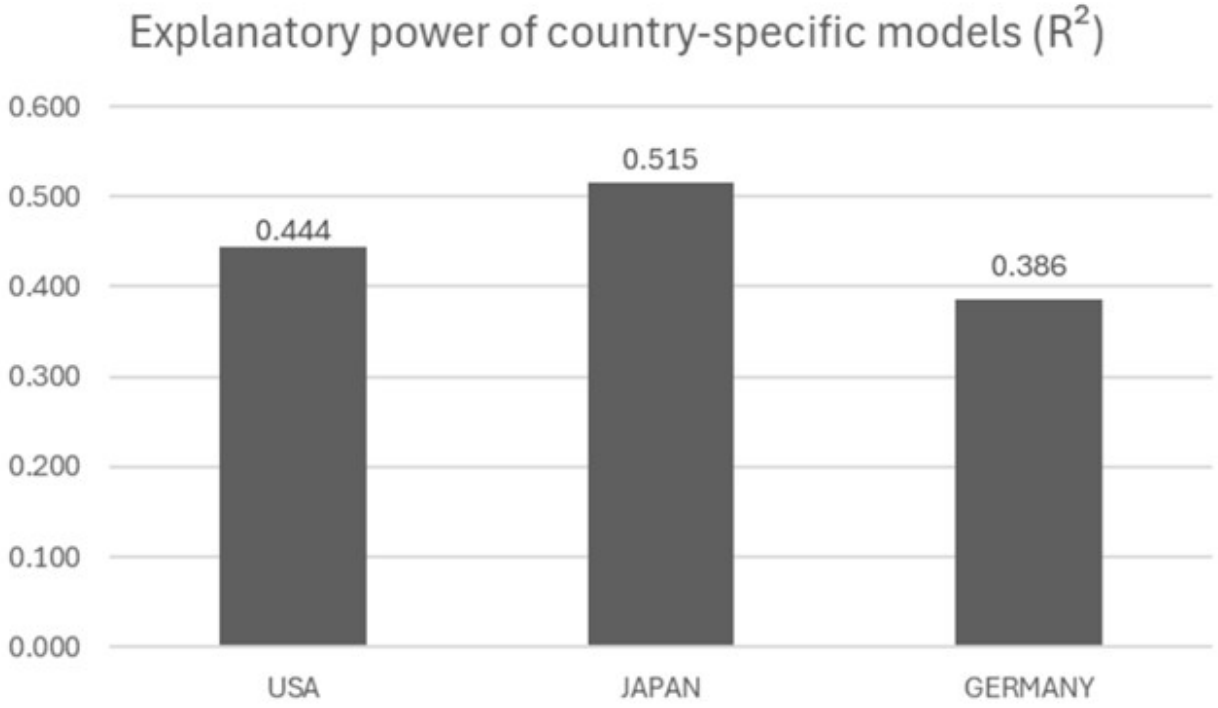


Table 8.2: Cross-country model performance showing RMSE values when models trained on one country are applied to predict data from other countries.

Model	USA	JAPAN	GERMANY
Model USA	N/A	1.351	0.886
Model JAPAN	1.183	N/A	1.193
Model GERMANY	0.812	1.317	N/A

We conclude that the findings observed in R are fully replicated in Python. The patterns remain consistent, and the results support the same core thesis of our analysis.

9 Use of AI Tools

In our research process, we utilized several AI tools to support different aspects of our work:

GitHub Copilot: Assisted us with coding tasks in both R and Python by providing intelligent code suggestions and automating repetitive coding patterns.

ChatGPT: Helped us refine text into more academically appropriate language.

Claude 4.0 and Gemini 2.5: These models were primarily used to restructure and improve the logical flow of our document, ensuring coherence and readability throughout.

Drabe, D., Hauff, S., & Richter, N. F. (2015). Job satisfaction in aging workforces: An analysis of the USA, japan and germany. *International Journal of Human Resource Management*, *26*(6), 783–805.

Dul, J. (2016). Necessary condition analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Sage Journals*, *19*(1), 10–52. https://doi.org/10.1177/1094428115584005

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*(3), S124–S134.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*(1), 1–35.

Hair, B., J. F. (2018). *Multivariate data analysis*. 259–370.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Richter, S., N. F., & Sarstedt, M. (2020). When predictors of outcomes are necessary: Guidelines for the combined use of PLS-SEM and NCA. *Industrial Management & Data Systems*, *80*(12), 2243–2267. https://doi.org/10.1108/IMDS-11-2019-0638

Sarstedt, M., & Danks, N. P. (2021). Prediction in HRM research—a gap between rhetoric and reality. *Human Resource Management Journal*, *32*(2), 485–513. https://doi.org/10.1111/1748-8583.12400

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.