Statistical analysis of company dissolution in Denmark

1st Lasse Buschmann Alsbirk
ITU, FIU Denmark
Copenhagen, Denmark
laal@itu.dk

2nd Julius Tind Westmann
ITU
Copenhagen, Denmark
jutw@itu.dk

3rd Katalin Literati-Dobos
ITU
Copenhagen, Denmark
klit@itu.dk

4th Alexandru Clefos
ITU
Copenhagen, Denmark
alecl@itu.dk

***Abstract*—This study investigates corporate dissolution dynamics in Denmark using publicly available data from the CVR register. By examining differences between voluntary and forced dissolutions and factors influencing company lifespans, the analysis highlights the central role of financial health, operational metrics, and executive characteristics in corporate outcomes. Employing hypothesis testing, survival analysis, and Generalized Linear Models, the study reveals systematic differences between dissolution types and identifies key predictors such as equity variability and financial statement submissions. Despite the dataset's static nature, missing data, and high co-linearity, the findings provide valuable insights for understanding corporate dissolution dynamics.**

I. Introduction

Corporate survival and dissolution are critical aspects of the economic landscape, reflecting the health and stability of industries, regions, and financial systems. Understanding why companies terminate—whether voluntarily or involuntarily—offers valuable insights into the factors that influence corporate lifespans and termination risks. These insights are vital for policymakers, regulators, and researchers who aim to design interventions that support business sustainability and economic resilience.

The dynamics of corporate dissolution are influenced by a range of factors, including financial performance, structural characteristics, industry affiliation, and ownership composition. Forced dissolutions, such as bankruptcies, represent one category of corporate termination that has drawn significant attention. For instance, corporate structures have been shown to play a critical role in facilitating illicit activities, such as money laundering schemes, which exploit corporate entities to create a veil of legitimacy. The Danish Financial Intelligence Unit (FIU Denmark) estimates that 68 billion DKK is laundered annually, much of it through corporate entities [1]. Shell corporations, strawman directors, and other mechanisms have been highlighted in investigations like the Panama Papers [2], underscoring the connection between corporate dissolution and financial crime.

Therefore, this study aims to understand general temporal differences between companies and the factors that affect their termination risks. It is also motivated by the Danish Financial Intelligence Unit's interest in better understanding the dynamics of corporate entities undergoing forced dissolution. Using data from the publicly accessible Danish business register (CVR), this paper aims to answer three key research questions:

1) Can we predict the lifespan of companies across different industry types and identify the covariates that influence their termination risks?
2) To what extent do companies that undergo forced dissolution differ from those that dissolve voluntarily?
3) Can we predict the lifespan and dissolution category (forced vs. voluntary) of companies using publicly available CVR data?

To address these questions, we conduct survival analysis to estimate lifespans and survival probabilities of companies, using Kaplan-Meier survival curves and Cox proportional hazards models to identify key covariates influencing termination risks. Hypothesis tests are employed to compare characteristics of companies dissolved voluntarily and by force providing insights into structural, financial, and industry-specific dynamics. To answer the third research question, we use Generalized Linear Models (GLM) to model company lifespan (regression) and dissolution type (classification) respectively. The research offers broader insights into survival probabilities across company types and termination risks and highlights how forced dissolution disproportionately affects certain industries and financial profiles.

By integrating descriptive and predictive analyses, this study provides insights into the statistical characteristics, survival dynamics, and predictive factors associated with corporate dissolution and survival in Denmark, contributing to practical efforts to combat financial crime.

II. Data

The Danish Business Authority maintains the official Danish business register (CVR, "Det Centrale Virksomhedsregister") and makes it available in full to the public through an API[1]. The database serves as a comprehensive source of information on businesses operating within Denmark and includes all legally registered entities, such as private companies, public companies, sole proprietorships, associations, and nonprofits. In addition to basic company information (including addresses and financial statements for certain company types),

[1]https://datacvr.virk.dk/artikel/system-til-system-adgang-til-cvr-data

the CVR register describes company-company and person-company relationships related to incorporation, directorships, and ownership. Most of the information in the CVR register is self-reported either by company representatives or professional advisors such as bookkeepers and auditors.

A. Company selection and calculated variables

To ensure consistency in reporting standards and data availability, this analysis focuses on limited companies (A/S) and limited liability companies (ApS) incorporated on or after 01-01-2000, with at least one registered owner. The set of *dissolved* companies meeting these criteria comprises 109,430 records in total and divided into two categories: forced dissolutions (e.g. bankruptcy) and voluntarily dissolution (e.g. dissolution by declaration). Additionally, 50,000 active companies are randomly sampled to enable comparative analysis for specific questions, as detailed below.
The full set of variables describing each company record can be found in appendix table VII. Some of these variables, such as company type, incorporation/termination dates and industry codes, are represented directly as-is from the source data. In addition, we compute a number of variables of particular interest to the hypothesis. These variables are mostly count-based and include the number of industry codes reported by a company, the number of addresses registered to the company, as well as variables computed from a company's relationships, e.g. the number of dissolved companies owned by a company's UBO(s). Finally, under the assumption that certain financial figures are indicators of forced/voluntary dissolution, we compute the mean and standard deviation of reported equity value, financial result and reported staff expenses for all available financial statements submitted by a company.

B. Missing Data

Reporting standards can vary based on the type, size and activity of a company[2]. As a result, certain variables contain missing values. As shown in appendix table VIII, the proportion of missing values is highest for variables based on financial statements, as some companies are either exempt from filing financial statements or terminate before doing so. The handling of missing data is detailed in the sections below.

III. Analyzing Company Lifespan and Termination Factors Using Survival Analysis and Hazard Models

Katalin Literati-Dobos

This part of the project examines factors influencing company termination using the above-described dataset. The objective was to estimate survival probabilities, identify critical covariates, and quantify their impact on company survival, providing insights into industrial and financial vulnerabilities. The research questions focused on identifying survival probabilities across industries and company types, along with determining the key structural, financial, and activity-related factors that influence termination risks.

The analysis comprised three stages: Kaplan-Meier visualization of survival probabilities, Cox proportional hazards models to quantify covariate impacts, and comparative Log-Rank Tests for statistical comparisons.

First, **Kaplan-Meier survival analysis** was conducted yearly to visualize survival probabilities over time for different industry types. In the dataset, survival time and censoring were calculated using `incorporation_date`, `termination_date`, and `company_status`. A binary variable - True or False - was created for terminated and active companies. Right censored data accounts for 31.36% of the dataset. The analysis is yearly since monthly analysis can lead to sparse data points, particularly for industries with fewer observations or longer lifespans. Yearly aggregation ensures sufficient data within each time point for robust estimates. Aggregating 674 industry codes into 20 main categories was practical for deriving meaningful insights, and plotting confidence intervals accounted for sample size differences and quantified uncertainty in survival probabilities using Greenwood's formula. Kaplan-Meier curves revealed high uncertainty for industries with fewer observations, making it particularly insightful to examine industry types with the highest number of observations.

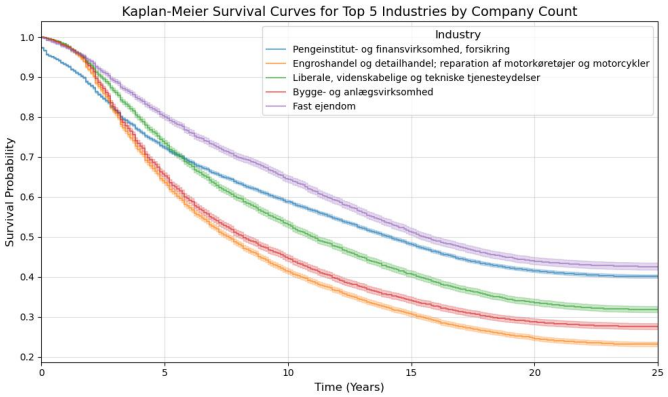

Fig. 1. Kaplan-Meier Survival Curves for industry types with the highest number of observations

In Fig. 1, steep declines were observed for scientific and technical services, retail, and construction industries with high termination proportions (68.2-76.9%). In contrast, slower declines were noted for financial institutions and real estate businesses, which included more active companies (57.4-59.9%). Also, the lower start of the financial industry's survival curve reflects a higher early termination rate. However, as time progresses, surviving companies in this sector appear to stabilize and outperform others, leading to the crossing of curves. This pattern suggests an industry dynamic where early risk is high, but long-term survival is stronger for financial companies.

Second, the **Cox Proportional Hazards Model** quantified covariates' effects through three variations. Structural,

[2]https://erhvervsstyrelsen.dk/hvilke-virksomheder-skal-indsende-aarsrapporter

financial, and activity metrics are analyzed separately in the Cox Proportional Hazards model to ensure methodological clarity and statistical robustness. Continuous variables (e.g., `number_of_owners`) and categorical variables (e.g., `industry_type`) behave differently in the model: categorical predictors require dummy encoding, which introduces baseline categories, while continuous variables are interpreted as unit changes in hazard. Investigating these metrics separately reduces the risk of multicollinearity, allows for clearer identification of each group's independent effects on survival, and improves model interpretability and numerical stability, particularly when working with high-dimensional data.

Model 1 included structural factors like `industry_main_category`, `company_type`, `number_of_names` and `number_of_owners`. Categorical variables were converted into dummy variables, continuous predictors were centered to improve model stability, and a moderate penalizer (L2 regularization with penalizer=1) was applied to address potential collinearity. The model shows some predictive value (concordance of 0.58), and the covariates collectively have a statistically significant effect on survival times (log-likelihood ratio test with extremely small p-value, effectively 0).

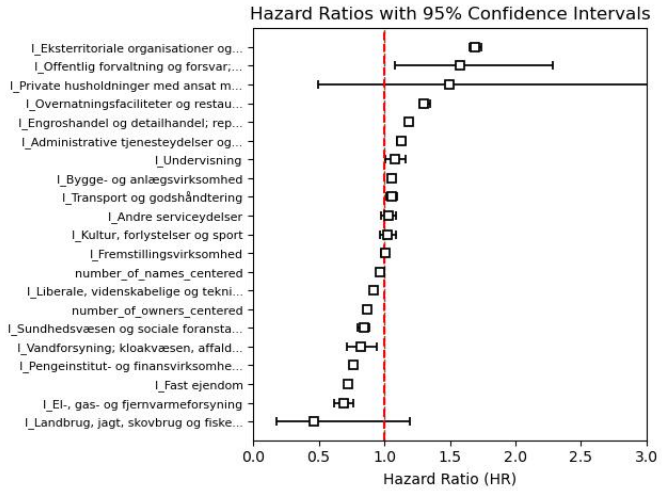

Fig. 2. Hazard Ratios for Structural Factors

The Cox Proportional Hazards plot (Fig. 2) confirms the same order of the five industries seen in Kaplan-Meier (Fig. 1) but reveals important additional information: the two industries with the lowest Kaplan-Meier curves (*Construction* and *Wholesale and retail trade*) now show HRs above 1, indicating higher termination risk. In contrast, the three industries with steeper Kaplan-Meier declines (*Real estate, Finance*, and *Professional, scientific, and technical activities*) have HRs below 1, suggesting stabilization. An interesting example is *Human health and social work activities*, not plotted in Kaplan-Meier due to its smaller sample size, which shows an HR below 1, indicating lower termination risk. Also, industries with fewer than 10 observations had hazard ratios crossing the HR = 1 line, indicating statistical insignificance. Additionally, the plot suggests that a smaller number of owners may increase the risk of termination, potentially linked to fraud due to concentrated power and lack of oversight, whereas a higher number of owners reduces this risk by fostering shared responsibility, better governance, and enhanced accountability.

Model 2 investigated financial metrics individually to assess their isolated effects on termination risk, ensuring interpretability and avoiding multicollinearity. Each financial metric was tested in a separate Cox Proportional Hazards model, producing hazard ratios with 95% confidence intervals (Fig. 3). With 43,206 observations and 30,014 events, the analysis revealed significant predictive value for most of the metrics; they reduced the risk of termination, as indicated by hazard ratios below 1. However, `mean_staff_expenses` was inconclusive, as its confidence interval included the null value (HR = 1), suggesting insufficient statistical evidence to confirm its effects. Additionally, the larger confidence intervals observed for all the metrics, reflect the impact of a smaller effective sample size, reducing precision and increasing uncertainty in the estimates. The plot suggests that both mean and standard equity, as well as mean and standard result, reduce the hazard ratio as they grow. This indicates that strong average financial performance and adaptable practices, whether in equity or operational results, contribute to survival. Stability (low variability) may not be more critical than the average values; instead, variability appears to have a protective effect when coupled with strong mean performance in both equity and results. High variability may reflect active financial management, such as capital injections or reinvestments, which can enhance a company's ability to adapt and survive.

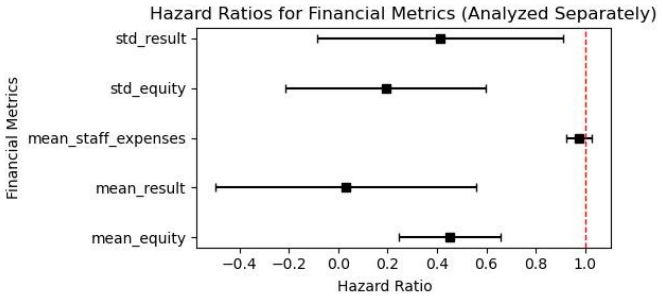

Fig. 3. Hazard Ratios for Financial Metrics

Model 3 incorporated activity metrics (`num_statements` and `num_statement_submitters`). Achieving a concordance of 0.75, it demonstrated that submitted financial statements significantly reduce termination risks, highlighting operational activity as a protective factor as seen in Fig. 4.

Finally, **Log-Rank Tests** compared survival distributions across industries and company types. A global test confirmed significant differences ($p < 0.005$), prompting pairwise comparisons. To balance false positives (Type I errors) and false negatives (Type II errors) across 190 comparisons, the Benjamini-Hochberg adjustment was applied, controlling the False Discovery Rate while retaining statistical power. Results are shown in a Heatmap (Fig. 8) of P-values for Survival

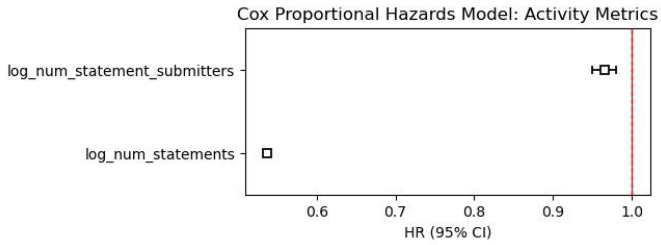

Fig. 4. Hazard Ratios for Activity Metrics

Differences. Each cell represents the significance level of survival differences between two groups.

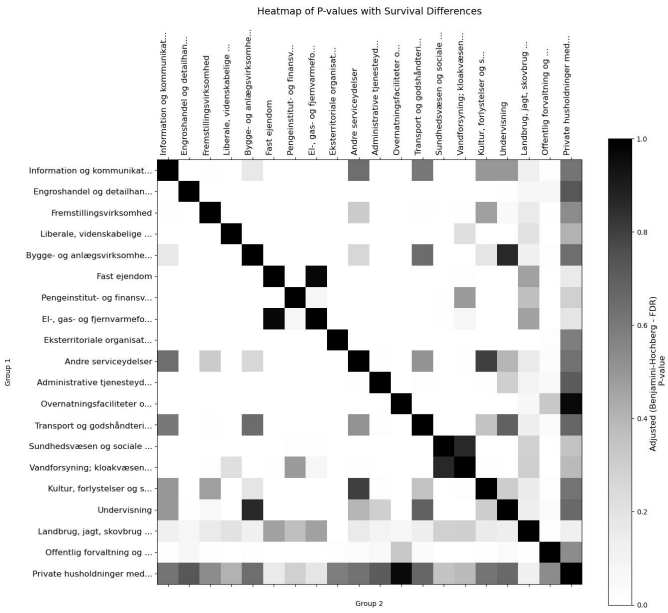

Fig. 5. Heatmap of P-values with Survival Differences

The heatmap [3] reveals that most groups differ significantly from each other, as evidenced by the high concentration of light-colored cells, particularly in rows like *"Information and communication"* and *"Public administration and defense; social security"*. However, some groups show similar survival behavior, evident in darker regions. Using a matrix format instead of a table allows for quick identification of significant pairwise differences and clustering patterns across groups, which would be challenging to interpret in a tabular format.

IV. Hypothesis Testing

Lasse Buschmann Alsbirk

We conduct hypothesis testing to compare two distinct populations of companies: those that were dissolved **voluntarily** and those dissolved **by force**. Our aim is to determine whether significant differences exist between these groups across multiple variables in the dataset, including count-based, binary, and categorical features.

For each variable X, we test the null hypothesis (H_0) that the distributions of X are identical for companies dissolved voluntarily and those dissolved by force. The alternative hypothesis (H_A) posits that a significant difference exists between the two distributions.

The nature of the variables dictates the choice of statistical tests. For count-based variables (e.g., `num_statements`, `num_addresses`, `number_of_primary_industry_codes`), which are non-negative and skewed, we use the **two-sided Mann-Whitney U Test**. This test is non-parametric and does not require the data to follow a normal distribution, making it appropriate for skewed or zero-inflated data. The Mann-Whitney U test assesses whether the ranks of values differ between two independent groups, providing a robust alternative to parametric tests when differences in the central tendency (e.g., medians) are expected. While the test does not strictly test for median differences, it is sensitive to shifts in location, as illustrated in Figure 6, where the number of financial statements submitted by companies shows notable differences between the two populations.

For binary variables, including original features such as `one_person_has_all_roles` and one-hot-encoded versions of categorical variables like `root_industry` and `company type`, we use the **chi-squared test of independence**. This test assesses whether the proportions of the binary outcome differ significantly between the two dissolution types. To facilitate comparisons across variables of different types, we also encode certain continuous variables, such as financial performance metrics, as binary indicators. For instance, the variable describing the mean financial result (in DKK) reported in financial statements is transformed into the binary variable `mean_result_negative`, which indicates whether the mean financial result is negative. This transformation ensures that the chi-squared test can be consistently applied while enabling straightforward comparisons across binary and one-hot-encoded variables.

A. Addressing multiple comparisons

Performing hypothesis tests on a large number of variables introduces the risk of Type I errors (false positives), where some null hypotheses are incorrectly rejected purely by chance. To address this, we apply two multiple testing correction methods: Holm-Bonferroni correction and Benjamini-Hochberg (FDR) correction.

The Holm-Bonferroni correction controls the family-wise error rate (FWER), ensuring that the probability of making even one false rejection remains below a specified threshold (α=0.05). This method adjusts the p-values in a sequential, step-down manner, starting with the smallest p-value and progressively applying a stricter threshold for subsequent tests. Compared to the classic Bonferroni correction, Holm's method is less conservative while maintaining strong control over FWER. Holm's correction is particularly appropriate here,

[3] A more detailed heatmap, included in the appendix, features annotations on cells with significant p-values, highlighting the direction and magnitude of survival differences.

where the focus is on minimizing false positives to ensure robust conclusions in a hypothesis-driven analysis.

For comparison, we also apply the Benjamini-Hochberg (BH) procedure, which controls the false discovery rate (FDR) - the expected proportion of false rejections among all rejected hypotheses. The BH correction is more liberal and better suited for exploratory analyses where identifying as many significant results as possible is the priority. However, because our analysis targets rigorous validation of group differences, and because the number of tests performed is in the range 40-50, we prioritize Holm-Bonferroni correction over BH to ensure stricter control of Type I error.

Both adjusted p-values are reported to allow for transparency and to demonstrate the effect of different correction methods.

B. Effect Size and Practical Significance

While extremely small p-values provide strong statistical evidence for rejecting the null hypothesis, they do not convey the magnitude or practical importance of the observed differences. To address this, we compute effect size measures tailored to the type of test:

- For count-based variables tested with the Mann-Whitney U Test, we report the Rank-Biserial Correlation (RBC), which quantifies the strength and direction of the difference between the two distributions.
- For binary variables tested with the chi-squared test, we compute the Phi coefficient (Φ), which measures the association strength in 2x2 contingency tables.

Effect size values provide context for interpreting the results and identifying variables with meaningful differences, even when statistical significance is achieved.

C. Hypethesis testing results

The hypothesis testing results are reported in appendix table VIII, sorted in order of increasing p-value after Holm-Bonferroni correction. Due to the large sample size (over 100,000 companies in most tests), nearly all p-values are extremely small, often approaching machine precision. As expected, we observe that the Holm correction generally produces higher adjusted p-values than the BH-adjusted p-values; however, in this analysis, no adjusted p-values exceed the $\alpha = 0.05$ threshold for the null hypothesis. These results provide strong statistical evidence to reject the null hypothesis across all variables, indicating systematic differences between companies dissolved voluntarily and those dissolved by force.

Nonetheless, the large sample size must be carefully considered when interpreting the results. In such large datasets, even minor differences between groups can achieve statistical significance. To address this, we rely on effect size measures to evaluate the practical significance of the observed differences, ensuring that our findings are both statistically robust and meaningful in practice.

Key findings include the `num_statements` variable, where the Holm-adjusted p-value is effectively zero, and the Rank-Biserial Correlation (RBC) indicates a large effect size (0.2515). This result highlights a substantial difference in the number of financial statements submitted between companies dissolved voluntarily and those dissolved by force. As illustrated in Figure 6, companies dissolved voluntarily are skewed toward submitting a higher number of financial statements. In a similar vein, the `executive_num_prev_bankruptcies` variable also shows an adjusted p-value of 0.0 and a large negative RBC value (-0.2437). This result suggests that the presence of executives with prior bankruptcies significantly increases the likelihood of forced dissolution. Interestingly, `executive_num_prev_companies_dissolved` variable, though still statistically significant, shows a low effect size (RBC) of -0.022, suggesting that the presence of executives with prior company dissolutions is not to the same extent a strong indicator of dissolution type. Furthermore, the chi-squared test for the binary variable `mean_equity_negative` demonstrates extreme statistical significance alongside a meaningful effect size (0.2446). The corresponding contingency table, shown in Table I, confirms that companies reporting negative equity values are disproportionately more likely to undergo forced dissolution, reflecting differences in financial performance between the two groups.

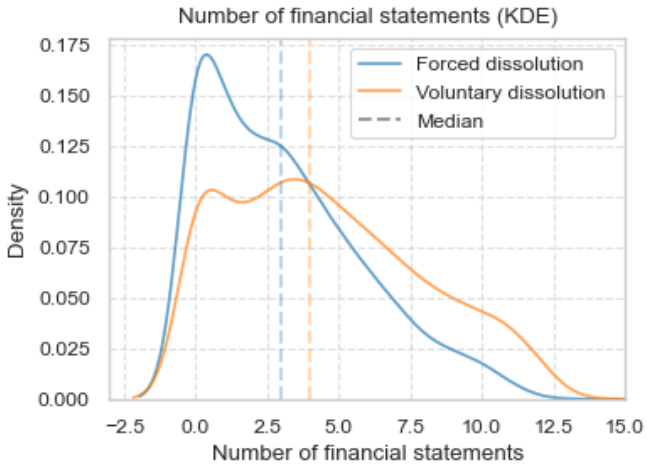

Fig. 6. KDE plot of the number of financial statements submitted. Adjusted p-value (Holm): 0.0000. Rank-Biserial Correlation: 0.2515

TABLE I
2X2 CONTINGENCY TABLE FOR TYPE OF DISSOLUTION AND THE BINARY VARIABLE `MEAN_EQUITY_NEGATIVE`. CHI-SQUARED ADJUSTED P-VALUE (HOLM): 0.0, ϕ-COEFFICIENT: 0.2446

Dissolution Type	Mean equity negative = 0	Mean equity negative = 1
Voluntary Dissolution	47,915	11,086
Forced Dissolution	19,190	13,575

Other variables, such as `number_of_names`, `num_addresses`, and `root_industry_F`, the latter indicating membership of the construction industry, exhibit

highly significant p-values as well, accompanied by moderate to large effect sizes. These results collectively indicate that companies dissolved by force systematically differ from those dissolved voluntarily across a wide range of operational, financial, and structural characteristics.

V. Modeling Company Lifespan

Julius Tind Westmann

Companies dissolve for many different reasons and how long they last can vary greatly. We want to see what parameters affect the lifespan of dissolved companies, both those who were forcefully dissolved and those who closed for other reasons.

We made 2 Generalized linear models, one Poisson and one Gamma, that tried to predict lifespan based on the other parameters. We chose those because the lifespan is in discrete months and is never below 0, though there is a big enough range that it can be thought of as continuous for the purposes of the Gamma distribution. Figure 7 shows the distribution of lifespans for companies that have closed. We can see that lifespan has a clear peak and a heavy tail. Non-forced companies have a smaller peak than the other 2. All of them peak at around the same place and have similar maximums. The minimum lifespan for all companies is 13, max is 298, and the mean is ~ 97.2 with a standard deviation of ~ 58.7, thereby a variance of around 3400. A simple, 1 parameter Poisson model with no features, would have a variance equal to the mean (expected value), so it is clear that something more complicated is required.

We removed rows with NaN variables, which are the ones with missing information. The data still has over 20,000 datapoints afterwards, 13020 who were forced to dissolve and 12546 who weren't. We chose not to have num_statements as a parameter, as longer lifespan just gives more time for financial statements, so it doesn't tell us anything interesting. For categorical data, we one-hot encoded it and chose "Bygge- og anlægsvirksomhed" (Building and construction) as the default, by dropping it from the data such that the coefficients of the other values represent the difference compared to "Bygge- og anlægsvirksomhed". We normalized each input parameter to go from 0 to 1, so their coefficients could be more easily compared.

We split the data in to 5000 test dataset and the rest train. The test set had 2585 companies who were forced to dissolve and 2415 who didn't. We judge the quality of the models by their mean squared error of the test set. The models were optimized by minimizing the deviance between predicted result and true result, as this is equivalent to MLE for GLM's. The first model performed comically poorly, getting MSE above 10^{24} in some cases.

When looking at the coefficients for the parameters, mean_staff_expenses and std_staff_expenses stand out as clear outliers having parameters of ± 60-70 for all dissolutions and for no forced dissolution, where most other parameters were close to 0 and no other was above 10. Looking closer, it's likely due to extreme outliers. The mean is in the 100,000's while the largest is in the 10 Billions. If these outliers also had a particularly high lifespan, and the feature isn't important otherwise, it could be overfitting to very little data. We have other measurements of finances, so instead of removing the outliers and keeping staff_expenses, we removed the staff_expenses features. This significantly improved the models.

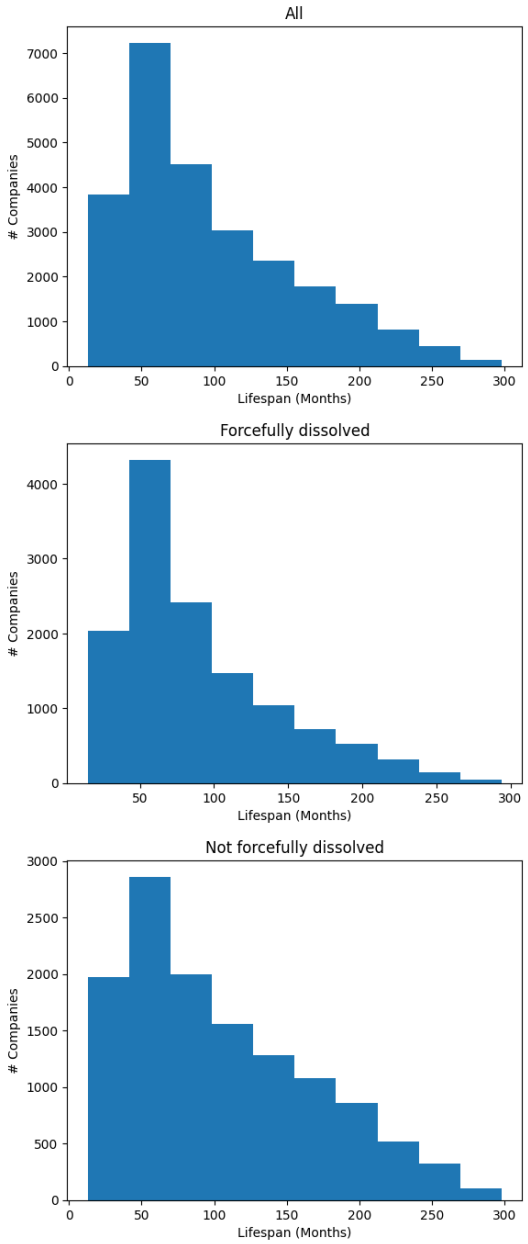

Fig. 7. Histograms of the Lifespan of companies.

Table II shows the deviance and MSE of Both models and if it was for forced dissolution companies. All means that all companies were used. The Gamma Distribution performed had about a quarter of the deviance of the Poisson, but the Poisson performed better on the test set for all and no forced dissolution. The deviance of Poisson and Gamma are very similar, so it is suprising that they are so far apart. The

gamma distribution performed extremely poorly when looking at companies that weren't forced to dissolve. It still had the best performance in one of the categories, but the other too were very bad.

The MSE is still very large for all instances. The variance for lifespan is around 3400 and none of the model are close, having at least 6000 higher MSE. A glance at the predictions on the test set shows that most prediction were below 20 while 1 or 2 were above 100 or above 1000. Any conclusion made from these models should therefore be taken with a grain of salt, as it could likely be very wrong.

model	Forced Dissolution	Deviance	MSE
Poisson	All	229233	11,136
Poisson	True	103932	35,812
Poisson	False	118327	14,483
Gamma	All	2345	199,339
Gamma	True	1100	9,798
Gamma	False	1155	315,462

TABLE II
DEVIANCE OF THE DIFFERENT MODELS.

Table III shows the value of the coefficient of the most important parameters overall, judge by their distance from 0, as well as their standard error. Most of the coefficients in the models were close to 0. We can see that financial information is important for the models with both equity and result, results being revenue minus expenses.

Forced Dissolution	Parameter	Value P	std err P	Value G	std err G
All	std_equity	2.993	0.246	-0.503	0.622
True	std_equity	-1.292	0.323	-0.987	1.352
False	std_equity	-3.240	0.108	-2.227	0.350
All	mean_equity	-0.207	0.075	-1.2704	0.406
True	mean_equity	0.560	0.343	-0.617	1.516
False	mean_equity	0.271	0.058	-0.680	0.318
All	n_owners	-0.717	0.022	-0.583	0.096
True	n_owners	1.230	0.031	-0.924	0.124
False	n_owners	-0.064	0.030	-0.052	0.149
All	mean_result	6.668	0.015	6.788	0.082
True	mean_result	5.001	0.053	5.122	0.234
False	mean_result	7.059	0.012	7.119	0.065
All	std_result	-1.992	0.056	3.843	0.273
True	std_result	3.261	0.090	3.999	0.433
False	std_result	0.422	0.093	3.731	0.362

TABLE III
KEY PARAMETERS AND THEIR STANDARD ERROR FOR THE MODELS. P IS POISSON AND G IS GAMMA.

Mean_result had the highest coefficient for all models. Some of it might be residue from from the outliers in expenses, we do see that it's valued higher when we have no forced dissolution, but it's still high overall. This shouldn't come as a big surprise, since ideally having a solid financial situation should result in longer lifespan. The lower value for forced dissolution, could indicate that even with financial success, you life might be cut short due to other factors.

The standard deviation for results in the Gamma model, seem to indicate that having more varying results correlates to longer lifespan. This might just be because high mean results will allow for bigger standard error, since it isn't normalized beforehand. For Poisson, it is only companies that were forced to dissolve that had a high coefficient. This could indicate that having a high deviation might give you more wiggle room to argue for not being dissolved by saying losses are temporary and there are potential future gains, while companies that dissolved without force don't benefit from that. For all companies, it shows a high negative effect, indicating that volatile results are bad for the overall health of the company.

Most of the coefficients are very small and change a lot between model, and data. The number of owners has a very high coefficient, specifically in the Poisson distribution for forced dissolution. This could indicate that switching owners might obfuscate how well you're doing financially. Number of owners is generally negatively correlated, which indicates that a healthy company doesn't change owner that often.

Interestingly, the coefficient for standard deviation of equity varies a lot both in the Poisson and Gamma distributions, though not in the same way. Both Poisson and Gamma have a high negative value for non-forced companies, but only Poisson has a high positive value for all companies, while Gamma has only has negative values. The negative value could indicate that a volatile equity is a killer. It means dissolves to this before it dissolves for any other reason.

While the coefficients vary a lot across the models, their standard error is pretty low across the board, except equity in the Gamma model for forced dissolution, which has standard errors higher than their values, meaning they could flip sign within 1 standard deviation.

The different industry types did not appear to have any significant influence on any of the models. It could be that "Bygge- og anlægsvirksomhed" is in the middle, so when comparing other industries they don't seem that much different, though most of the number were very close to zero, so the effect is definitely minor. The most significant of the binary data, was if the company was incorporated or not, having a coefficient of around 0.15 for all models.

With the models poor performances, it might have been better to not drop all NaN values. There is also a possibility that a company having a NaN value is significant. We could have left out some of the financial data, as they have the most NaN values, and therefore are the reason for most of the removed data. The financial data is however also the most significant portion of the models, so it might just have made it less useful.

The data might also have been better model by a normal distribution, with a cut-off point at 0, so that it doesn't go below.

It could also simply be the case that the data can't predict company lifespan. There are many outside influences that can make a company that appears fine suddenly tank in value. There can also be internal conflicts that can't be represented in the data, but that has a significant influence on the company.

VI. Analyzing Factors Contributing to Forced Company Dissolution Using Generalized Linear Models

Alexandru Clefos

The dataset comprises several columns of interest, notably `cause_of_termination` and `forced_dissolution`. A rigorous analysis of the data revealed intriguing patterns. Given the complexity and high dimensionality inherent in the dataset, manually modeling the relationships between the numerous features and the target variable poses substantial challenges. To address this, we propose the utilization of a **Generalized Linear Model (GLM)**. This approach enables the interpretation of model weights to identify which variables exert the most significant influence on the target.

Prior to model training, the dataset undergoes a comprehensive cleaning and processing phase employing established techniques. Numerical columns are subjected to normalization, standardization, and binning, while categorical variables are transformed using one-hot or binary encoding. This prepares the dataset for robust feature engineering.

However, the feature engineering phase significantly augments dataset dimensionality. This expansion introduces redundant noise, dilutes predictive power due to weak correlations with the target variable, and risks overfitting. **Lasso regularization** was implied but when features are highly correlated, Lasso can pick one feature at random to keep and shrink others to zero. The dataset's columns after processing, using a correlation matrix show a high correlation between columns resulting in zero weights.

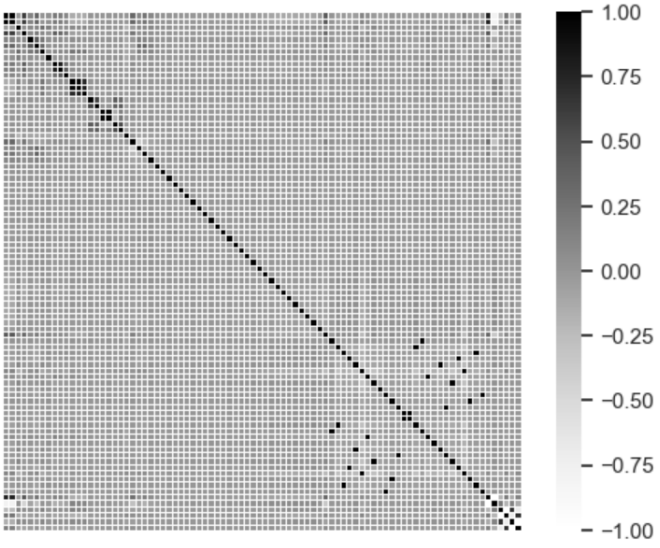

Fig. 8. Correlation Matrix of the transformed dataset

Thus, **Ridge** is more effective in contexts with multicollinearity since it does not zero out coefficients but rather balances them. Moreover, because **Ridge** doesn't enforce sparsity, it doesn't completely discard weak features but rather adjusts them based on their contribution.

Column Name	Coefficient
num_statements	-0.6984
num_addresses	-1.1192
num_employees	3.296
mean_staff_expenses	0.6184
owner_num_prev_companies	-0.0045
owner_num_prev_companies_dissolved	0.0020

TABLE IV
RIDGE REGULARIZATION RESULTS SAMPLE

A negative coefficient suggests that as the "number of financial statements" increases, the likelihood of forced dissolution decreases. Companies with more financial statements may be more organized, transparent, or have a longer operational history, which might correlate with greater financial stability and a lower risk of forced dissolution. A positive coefficient means that as the number of employees increases, so does the likelihood of forced dissolution. Larger staff sizes might increase operational costs, and if not matched by revenue, could lead to financial difficulties. Alternatively, larger firms often face greater scrutiny or complex challenges which can increase the risk of forced dissolution. The coefficients near to zero indicate that the impact is relatively minor in changing the predicted value, indicating that while there's a positive association, the feature might not be a strong predictor by itself in the context of the model.

After the selection and validation of regularization, the **Generalized Linear Model** (GLM) is defined. Considering the binary nature of the target variable, the logistic (sigmoid) activation function is utilized. By assuming a Bernoulli distribution for the outcome, this function effectively connects the expected mean with the weighted sum through the logistic transformation.

For robust model evaluation, **cross-validation** techniques were applied. Specifically, the dataset is bifurcated into five subsets, with training conducted over five iterations using permutations of four subsets, and evaluation performed on the remaining subset. This **Stratified Cross-Validation** method ensures model results are not exaggerated or erroneous, accounting for potential pitfalls during the train-test split phase.

Iteration	Score
1	0.705427
2	0.731751
3	0.690317
4	0.882428
5	0.850229

TABLE V
STRATIFIED CROSS-VALIDATION RESULTS

Upon completing the training phase and validation against chosen metrics, the model's weights are analyzed to discern which variables most substantially impact predictions. In the context of the sigmoid function applied, model weight interpretation is achieved through understanding how each feature's coefficient influences the log-odds of the response variable, which in turn translates into changes in probability.

$$Intercept(log-odds) : -0.60556 \quad (1)$$

$$Intercept\ odds\ ratio : 0.54576 \quad (2)$$

The intercept in the log-odds scale indicates the log-odds of "forced dissolution" when all predictor variables are held at zero. A negative log-odds suggests that, at this baseline level, the odds of "forced dissolution" are less than 1, meaning the event is less likely than not. An odds ratio of 0.5458 indicates that, at the baseline level (all features equal zero), the odds of "forced dissolution" are about 54.58% of the odds of not having a "forced dissolution". This suggests that in the absence of other information, `forced_dissolution` is less likely than not.

Index	Feature	Coefficient	Odds Ratio
8	mean_equity	-3.529889	0.029308
9	std_equity	0.733813	2.083009
14	company_type_Anpartsselskab	0.770638	2.161145
9	std_equity	0.733813	2.083009
18	number_of_primary_industry_codes_4.0	0.545246	1.725032
3	number_of_owners	1.710295	5.530592
6	num_addresses	3.428811	30.839967
2	number_of_names	3.915264	50.162391

TABLE VI
SAMPLE OF COEFFICIENTS AND ODDS RATIOS FOR FEATURES AFFECTING FORCED DISSOLUTION

Companies with higher mean equity are significantly less likely to face dissolution, underscoring the protective value of financial stability. In contrast, increased equity variability, a higher number of owners, and structured as an 'Anpartsselskab' appear to increase the risks of dissolution, possibly due to operational complexities or governance challenges. Furthermore, firms with multiple addresses or frequent name changes exhibit a markedly higher likelihood of dissolution, suggesting instability or identity issues as key risk factors. Increased financial variability, as indicated by a higher standard deviation of equity, is associated with a significantly higher likelihood of the target outcome, which potentially indicates a greater risk due to financial instability. Furthermore, a greater number of specific primary industry codes suggest increased odds of the outcome, possibly highlighting complexity or industry-specific risks. Together, these findings underscore the potential impact of financial and operational factors on the likelihood of the modeled outcome.

Collectively, these findings highlight the importance of maintaining financial robustness and operational consistency to mitigate the risk of forced dissolution, providing valuable guidance for corporate management and policy making.

VII. DISCUSSION

This study examined corporate dissolution dynamics in Denmark, leveraging data from the CVR register to explore differences between voluntary and forced dissolutions, factors influencing company lifespans, and the predictability of dissolution categories. The dataset reflected the complexities of real-world data: skewed distributions, co-linearity, and substantial missing data. These characteristics necessitated diverse statistical approaches tailored to address specific challenges and derive meaningful insights.

A number of factors emerged as important indicators of both company lifespan and dissolution type. Financial metrics, such as equity values and their variability, were strongly associated with corporate outcomes. Persistently negative equity was linked to higher likelihoods of forced dissolution, while positive equity and financial stability correlated with longer lifespans. Beyond financial performance, operational metrics such as the number of financial statements submitted (`num_statements`) also played a significant role. Companies submitting more financial statements tended to have longer lifespans and were more likely to dissolve voluntarily. While this may initially suggest that financial transparency is a marker of corporate health, alternative interpretations are plausible. The number of financial statements submitted could reflect company size or age, as larger or older companies are subject to stricter reporting obligations. This finding highlights how the dataset's structure and external regulatory factors influence variable interpretation.

The analysis also underscored the high degree of co-linearity among variables, as illustrated by the correlation matrix in Figure 7. Financial and operational metrics are inherently interrelated, complicating efforts to isolate individual effects. Regularization techniques, such as ridge regression applied in our survival analysis and GLM classification model, proved effective in addressing these dependencies by penalizing redundant predictors and focusing on the most influential variables. This approach provided a more holistic view of company characteristics compared to hypothesis testing, which evaluates variables independently and does not account for such overlaps.

The dataset's complexity extended beyond co-linearity. Survival analysis revealed how company lifespan varies significantly across industries, with construction and retail sectors experiencing shorter lifespans compared to finance and real estate. Stratifying the data to account for such differences allowed for more nuanced insights but also introduced challenges like data sparsity and the need for domain-specific knowledge. Defining key attributes like "company size" is not straightforward in this context, as holding companies with minimal operational activity may report large financial results, while operational firms with hundreds of employees may report relatively modest revenues. Such ambiguities demonstrate the limitations of relying solely on raw data without incorporating domain expertise.

Additionally, the heavy-tailed nature of the corporate landscape introduced extreme values that posed challenges for predictive modeling. GLM experiments showed sensitivity to these outliers, reinforcing the importance of robust preprocessing and careful variable selection in real-world analyses.

VIII. Conclusion

This study demonstrates the potential of using publicly available data to analyze corporate dissolution dynamics in Denmark, addressing key questions about company lifespan, dissolution categories, and differences between forced and voluntary dissolutions. The results reveal that company lifespan can be partially predicted using survival analysis, with significant variations across industries. For example, construction and retail companies exhibited shorter lifespans compared to finance and real estate, and key covariates such as equity values and financial statement submissions emerged as important predictors of termination risks. However, the static nature of the dataset limited the ability to fully capture dynamic processes such as financial decline or operational disruptions, which likely play a crucial role in lifespan predictions.

Dissolution category (forced vs. voluntary) was also predictable to some extent, with financial metrics like negative equity and variability in financial performance being the strongest indicators of forced dissolution. Operational metrics, including the number of financial statements submitted, further distinguished voluntary dissolutions, though this variable's interpretation remains nuanced. While it might reflect financial transparency, it is also influenced by company size, lifespan, and regulatory thresholds, highlighting the complexities inherent in analyzing real-world data.

The study also revealed systematic differences between companies dissolved voluntarily and by force. Forced dissolutions were associated with persistent financial instability and executive histories of prior bankruptcies, while voluntary dissolutions were linked to longer lifespans and greater financial stability. However, these differences were shaped by the interdependencies among variables, and required careful interpretation to avoid overestimating the importance of individual predictors.

While the findings provide valuable insights, the dataset's static nature and missing data limited the scope of the analysis. Temporal changes, such as shifts in financial health or operational strategies, remain unexplored, and external factors like macroeconomic conditions and inter-company relationships were not captured. Future research should incorporate temporal methods to capture the evolution of risk factors and expand datasets to include external variables like macroeconomic conditions. These advancements could deepen our understanding of corporate dissolution and inform strategies to mitigate its risks.

References

[1] "The national risk assessment of money laundering 2022," Hvidvasksekretariatet, Financial Intelligence Unit Denmark, January 2023.

[2] H.-C. H. Chang, B. Harrington, F. Fu, and D. N. Rockmore, "Complex systems of secrecy: The offshore networks of oligarchs," *PNAS Nexus*, vol. 2, no. 3, p. pgad051, Mar. 2023.

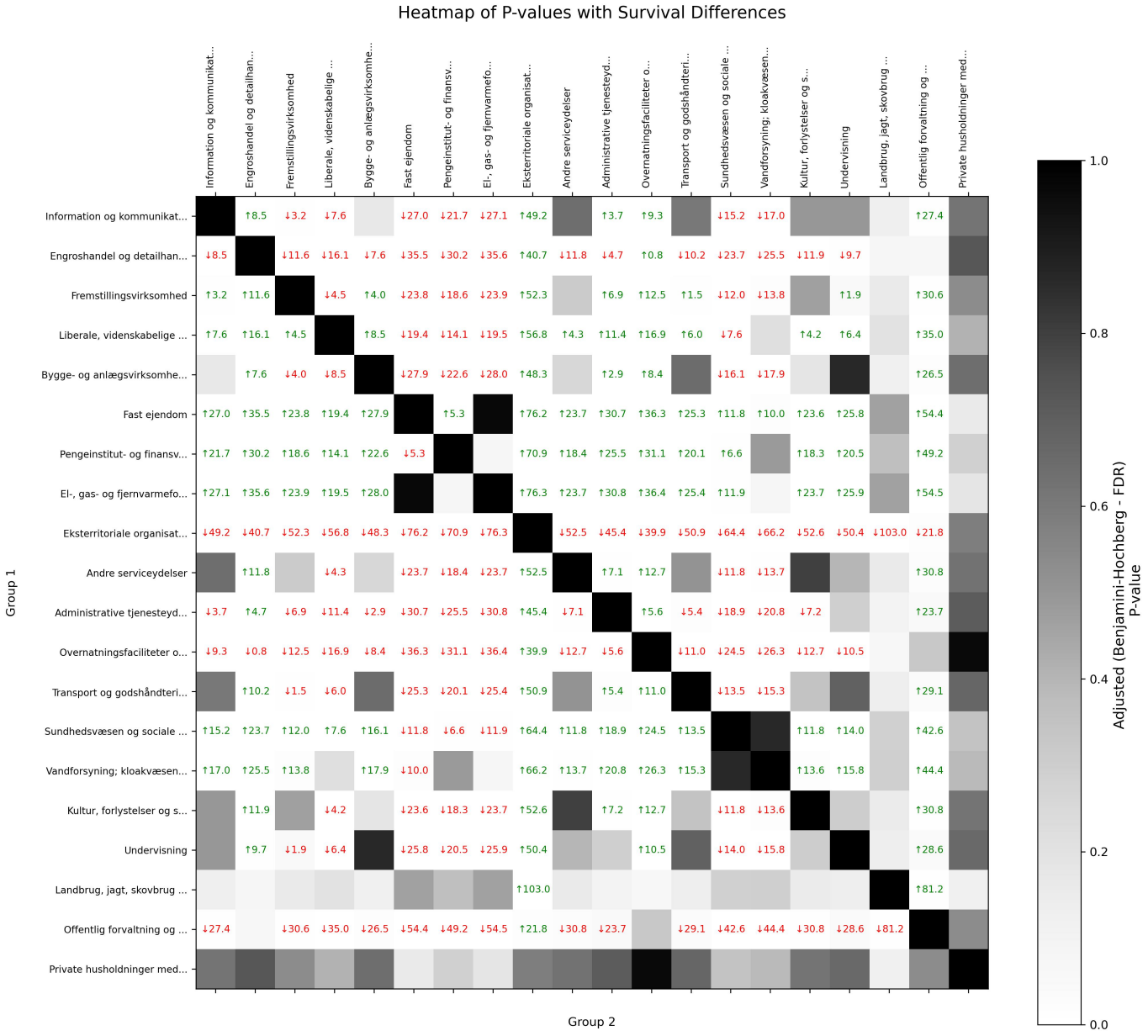

Fig. 9. Heatmap of P-values with Survival Differences
A significant difference is represented by a p-value < 0.05, indicating a statistically meaningful disparity between the groups being compared.

TABLE VII
DATA COLUMN SPECIFICATION

Column name	Description	Data type	Example record
cvr_nummer	Incorporation number	int	31470080
company_status	Status of company	string (category)	KONKURS
industry_code	Lowest-level industry code	string (category)	412000
company_type	Company type	string (category)	Anpartsselskab
incorporation_date	Date of incorporation	date	2008-05-16
termination__date	Date of dissolution.	date	2016-10-25
lifespan_months	Number of months between incorporation date and termination date.	int	101
number_of_names	Total number of names registered to the company	int	1
number_of_primary_industry_codes	Total number of industry codes registered by the company. It can only have one at a time	int	2
number_of_owners	Total number of UBOs	int	1
num_statement_submitters	Total number of external companies which have submitted financial statements on behalf of the company	int	1
num_statements	Total number of financial statements submitted by the company	int	2
mean_equity	Mean equity value in the company's financial statements (can be null)	float	209416.5
std_equity	Standard dev. equity value in the company's financial statements (0 if mean is null)	float	6623.469219
mean_staff_expenses	Mean value reported for personnel/staff expenses (can be null)	float	548042
std_staff_expenses	Standard dev. value reported for personnel/staff expenses (0 if mean is null)	float	71259.39298
mean_result	Mean value reported for annual result, which is revenue subtracted by all expenses (can be null)	float	7903
std_result	Standard dev. value reported for result (0 if mean is null)	float	24423.46822
num_addresses	Number of addresses registered to the company	int	1
num_employees_year	Year of the latest registration of employee count.	int	2014
num_employees	latest registration of employee count.	float	1
owner_num_prev_companies	Number of companies owned by UBO	int	0
owner_num_prev_companies_dissolved	Number of dissolved companies owned by UBO	int	0
owner_num_prev_bancruptcies	Number of bancrupt companies owned by UBO	int	0
one_person_has_all_roles	True if a single person has all roles in the company (executive, founder, owner)	Binary	1
coowned_by_company	True if the company is fully or partly owned by another company	binary	1
coowned_by_holding_company	True if the company is fully or partly owned by a holding company	binary	0
executive_num_prev_companies	Number of companies executives have had executive roles in	int	2
executive_num_prev_companies_dissolved	Number of dissolved companies executives have had executive roles in	int	0
executive_num_prev_bancruptcies	Number of bancrupt companies executives have had executive roles in	int	0
root_industry_code	Top-level industry code	string (category)	F
root_industry_name	Top-level industry name	string (category)	Bygge- og ...
forced_dissolution	True if company was forced to dissolve, False if company was voluntarily dissolved	binary	TRUE
dissolved	True if company is dissolved, False if company is active	binary	TRUE

TABLE VIII
RESULTS OF MULTIPLE HYPOTHESIS TESTS WITH P-VALUE ADJUSTMENTS

Variable name	Number of not-null values	Statistical test	Test statistic	p-value	Effect type	Effect value	Adjusted p-value (BH)	Adjusted p-value (Holm)
num_statements	109430	Mann-Whitney	1.05616e+09	0	RBC	0.251557	0	0
mean_equity_negative	91766	chi-squared	5494.4	0	Phi	0.244692	0	0
executive_num_prev_bancruptcies	109430	Mann-Whitney	1.75504e+09	0	RBC	-0.243707	0	0
mean_result_negative	90969	chi-squared	3258.75	0	Phi	0.189269	0	0
num_statement_submitters	109430	Mann-Whitney	1.15664e+09	0	RBC	0.180348	0	0
root_industry_F	109430	chi-squared	2560.6	0	Phi	0.152969	0	0
root_industry_K	109430	chi-squared	2557.99	0	Phi	0.152891	0	0
owner_num_prev_bancruptcies	109430	Mann-Whitney	1.6242e+09	0	RBC	-0.150987	0	0
root_industry_I	109430	chi-squared	1692.89	0	Phi	0.124379	0	0
number_of_names	109430	Mann-Whitney	1.57543e+09	0	RBC	-0.116426	0	0
coowned_by_company	109430	chi-squared	1176.76	6.86133e-258	Phi	0.103699	2.3079e-257	1.85256e-256
owner_num_prev_companies	109430	Mann-Whitney	1.25061e+09	5.29399e-240	RBC	0.113756	1.63231e-239	1.37644e-238
root_industry_L	109430	chi-squared	1044.97	3.02108e-229	Phi	0.0977199	8.59845e-229	7.55269e-228
one_person_has_all_roles	109430	chi-squared	977.972	1.1026e-214	Phi	0.0945355	2.91403e-214	2.64625e-213
root_industry_G	109430	chi-squared	732.861	2.13823e-161	Phi	0.0818357	5.27429e-161	4.91792e-160
company_type_Aktieselskab	109430	chi-squared	662.498	4.27628e-146	Phi	0.077808	9.30719e-146	9.40781e-145
company_type_Anpartsselskab	109430	chi-squared	662.498	4.27628e-146	Phi	0.077808	9.30719e-146	9.40781e-145
num_addresses	109428	Mann-Whitney	1.5327e+09	3.27922e-145	RBC	-0.0861747	6.74062e-145	6.55844e-144
executive_num_prev_companies	109430	Mann-Whitney	1.2909e+09	1.36915e-125	RBC	0.0852098	2.66623e-125	2.60138e-124
number_of_owners	109430	Mann-Whitney	1.30236e+09	8.45317e-116	RBC	0.0770836	1.56384e-115	1.52157e-114
root_industry_N	109430	chi-squared	511.611	2.8297e-113	Phi	0.0683757	4.98566e-113	4.81049e-112
root_industry_H	109430	chi-squared	491.272	7.53556e-109	Phi	0.0670028	1.26734e-108	1.20569e-107
coowned_by_holding_company	109430	chi-squared	380.337	1.05119e-84	Phi	0.0589543	1.69105e-84	1.57679e-83
num_employees	44700	Mann-Whitney	2.71848e+08	7.3279e-73	RBC	-0.08843	1.12972e-72	1.02591e-71
lifespan_months	109430	Mann-Whitney	1.3388e+09	3.90045e-46	RBC	0.0512639	5.77266e-46	5.07058e-45
number_of_primary_industry_codes	109430	Mann-Whitney	1.34962e+09	1.75142e-42	RBC	0.0435969	2.49241e-42	2.10171e-41
mean_staff_expenses_zero	30160	chi-squared	185.125	3.68594e-42	Phi	0.078346	5.05111e-42	4.05454e-41
num_employees_year	44700	Mann-Whitney	2.67035e+08	4.93824e-37	RBC	-0.0691608	6.52553e-37	4.93824e-36
root_industry_M	109430	chi-squared	150.294	1.49528e-34	Phi	0.0370597	1.90777e-34	1.34575e-33
root_industry_S	109430	chi-squared	142.926	6.10128e-33	Phi	0.0361399	7.52492e-33	4.88103e-32
owner_num_prev_companies_diss.	109430	Mann-Whitney	1.45691e+09	1.80059e-31	RBC	-0.032436	2.14909e-31	1.26041e-30
root_industry_J	109430	chi-squared	130.724	2.84602e-30	Phi	0.0345628	3.29071e-30	1.70761e-29
root_industry_C	109430	chi-squared	109.39	1.33317e-25	Phi	0.031617	1.49477e-25	6.66585e-25
mean_staff_expenses_positive	30160	chi-squared	88.4602	5.18686e-21	Phi	0.0541575	5.64453e-21	2.07474e-20
root_industry_Q	109430	chi-squared	55.024	1.19068e-13	Phi	0.0224237	1.25872e-13	3.57204e-13
executive_num_prev_companies_diss.	109430	Mann-Whitney	1.44242e+09	1.7014e-10	RBC	-0.0221695	1.74867e-10	3.40281e-10
mean_staff_expenses_negative	30160	chi-squared	5.57759	0.0181917	Phi	0.013599	0.0181917	0.0181917