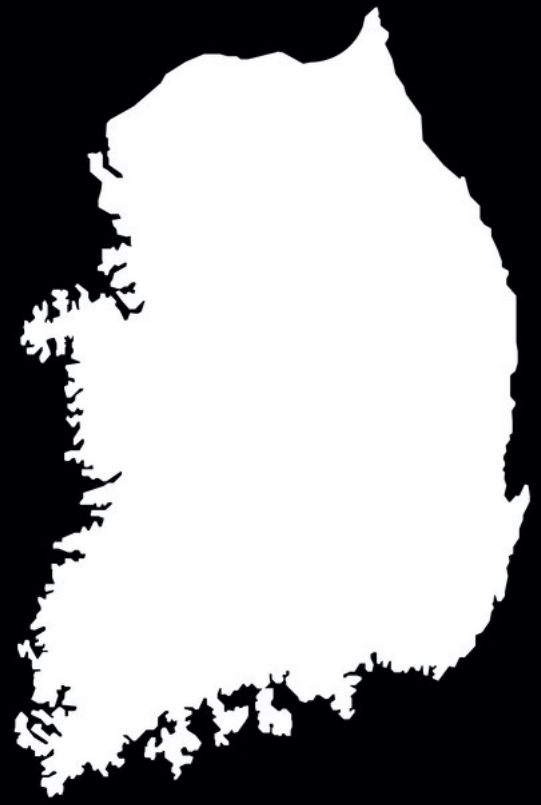
# Research Paper:

"Longitudinal small and medium enterprise (SME) data on survival, research and development (R&D) investment, and patent applications in [South] Korea's innovation clusters from 2008 to 2014"

- Researcher: Byung-Keun Kim, Phd
- Researcher: JungTae Hwang, Phd
- Study Review:  Sam Dreyfuss
- Johns Hopkins University – Summer Term 2023

South Korea



# Outline of Presentation:

- Research Paper Summary
  - Background Information on Study
  - Researcher Motivation
  - Literature Review Finding
- Empirical Results Summary
  - Estimation Method
  - Model Overview
  - Data Sample Overview
  - Findings and Discussion of Results
- Extension and Analysis
  - Complementary Research

# Research Paper Summary – Background Information

- Recessions impact small and medium sized enterprises (SMEs) most significantly
- When a recession occurs, a firm can decide to either increase or decrease spending on research and development (R&D).
  - **Increasing R&D expenditure**: more expensive in the short-term, but can put the firm in a competitive position vs. peers post recession
  - **Decreasing R&D expenditure**: less expensive in the short-term, but can put the firm in a less competitive position vs. peers post recession
- Smaller firms have historically tended to invest more heavily, while larger firms have typically chosen to refrain from additional investment

# Research Paper Summary – Research Motivation

- There have been few studies looking at the effects of innovation (R&D expenditure) on SMEs during recessions

- Previous studies have been focused R&D investment during expansionary periods of the business cycle and not accounted for the risk of additional investment

- There are even fewer studies which specifically looked the R&D's impact on firm survival within South Korea

# Research Paper Summary – Literature Review Findings

- The main findings of the study included:
  - The effect of expenditure on R&D for SMEs is not identical for all firms
  - Investment in R&D during a recession is generally a poor choice for SMEs
    - For select innovative firms investment in R&D can be a superior strategy
  - Firm size and can lower the probability of market exit (bankruptcy)

- Link to original research paper and data: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6557752/
- Link to SPSS code: https://docs.google.com/document/d/e/2PACX-1vRrRtfmdTtEnZaVBdeM6p3kb3x2uzNb2NP4KTXwB_qQWzMkpxwu71lVnAmxt1u3fI1JC7df5d6HAvc9/pub
- Original Data: https://docs.google.com/spreadsheets/d/e/2PACX-1vQOwjyWcp4kZ6N8WQKUVcK4Xgu43ch_mSfZghZ34KJnTBAuQwnME5doQ62_cpDCAyuKfcdlOMzU-1Jj/pub?gid=2145868861&single=true&output=csv
- Screened Data: https://docs.google.com/spreadsheets/d/e/2PACX-1vSohBhgT15-0i8yb1g7Ng8PYZxVr81tTe9-jlW-s859kzuK-OdvNFGhyOJizVCnG4NlXDHa05L4FtuO/pub?gid=607866883&single=true&output=csv

# Empirical Results Summary – Estimation Method

- To conduct this specific type of analysis typically referred to as a survival or hazard analysis, multiple factors hypotheses were considered:
- Hazard models are discussed within the course textbook (Chapter 22)
- **Firm Age:**
  - Hypothesis 1: Young firms are more likely to exit than old firms during a recession
- **Firm Size (log(# of employees)):**
  - Hypothesis 2: Small firms are more likely to exit more frequently than large firms during a recession
- **Investment in R&D (Number of Research Personnel as proxy):**
  - Hypothesis 3: During a recession, SMEs that invest in R&D have higher probabilities of survival than those of non-investing firms

# Empirical Results Summary– Model Overview

- A **time-dependent Cox regression model** based on panel data was chosen due to its effectiveness for survival analyses
- A Cox regression model is a type of statistical analysis used in survival analysis to examine the relationship between covariates (independent variables) and the time to an event of interest, such as death, failure, relapse, or other types of occurrences
- Due to the longitude dataset structure, censoring and time-varying explanatory variables with event-history data were implemented
  - As the survival rates were not monitored once the experiment was completed and firms were permitted to leave the dataset (ex: disappearing for unknown reason) **right censoring** was specifically used within the model
- **Backwards elimination** was implemented to assist in identifying which factors had the most influence on the survival rate.
  - Backwards elimination begins with a full variable model that includes all potential covariates and iteratively removing variables that are found to be less significant or less relevant to the outcome.
  - Its also useful in time-dependent Cox regression framework due to its ability to improve model specification, avoid overfitting, and help reduce multicollinearity

# Empirical Results Summary – Cox Regression Explained

- A time-dependent Cox regression model is an extension of the **Cox regression model**
- The Cox regression model computes what are called “Hazard Ratios”
  - log(Hazard Ratios) = Model Coefficient
- The hazard ratio (HR) represents the change in the hazard rate (the risk of experiencing the event of interest) associated with a one-unit change in the independent variable, while holding other variables constant
- Interpretation of Coefficient and Hazard Ratio (HR) via examples:

| **Example 1: Clinical Trial for Drug Efficacy** | |
|---|---|
| Independent Variable: | Treatment (1 = T, 0 = C) |
| Coefficient ($\beta$): | -0.45 |
| Conversion: | $e^{-0.45}$ = 0.64 (HR) |
| HR Interpretation: | T has (1-0.64) = 36% lower Hazard Risk than C* |

T = Treatment, C = Control

| **Example 2: Survival & Tumor Size** | |
|---|---|
| Independent Variable: | Tumor Size (cont. var) |
| Coefficient ($\beta$): | 0.06 |
| Conversion: | $e^{0.06}$ = 1.06 (HR) |
| HR Interpretation: | Each unit size (cm) of tumor increases fatality risk by 6% |

# Empirical Results Summary – Cox Regression Explained

- A key assumption of the cox regression model is the **proportional hazards assumption** which states that the hazard ratio (the relative risk of experiencing the event) between two groups of individuals remains constant over time
- The time component in the panel dataset invalidates the original Cox regression's assumptions that the hazard ratio remains for (potentially) multiple variables remain constant over time, using a time-dependent Cox regression model allows you to model situations where the hazard ratios change over time, providing a more flexible and accurate analysis of survival data

# Empirical Results Summary – Data Sample

- Firms within the dataset were selected based on industry sector and employment size
- Manufacturing firms were selected for the study including, but excluded cigarette companies
- Only firms with less than 500 full-time employees were included in the study (including parent company in the case the studied company was a subsidiary)
- Firms aged 3 years or less were excluded from the dataset due to the possibility of measurement error
- Final dataset included 588 firms – including 58 firms which failed to survive during the research period
- Due to the possibility of multicollinearity within the data, the study employed a backward elimination technique to identify variables

# Empirical Results Summary – Results

- I recreated the study's results in SPSS using a time-dependent Cox regression model (w backwards elimination)
- Recreated model coefficients and p-values matched the original research results perfectly

| Original Study Results: | | |
|---|---|---|
| Ind. Var. | $\beta$ | Significance |
| Age | -0.030 | 0.304 |
| Size – ln(# of employees) | -0.476 | 0.003 |
| # of R&D Employees | -0.060 | 0.150 |

| Recreated Study Results: | | |
|---|---|---|
| Ind. Var. | $\beta$ | Significance |
| Age | -0.030 | 0.304 |
| Size – ln(# of employees) | -0.476 | 0.003 |
| # of R&D Employees | -0.060 | 0.150 |

# Empirical Results Summary – Results Discussion

| Study Results: | | | |
|---|---|---|---|
| Ind. Var. | $\beta$ | Significance | Notes |
| Age | -0.030 | 0.304 | Removed in backward stepwise, low significance |
| Size - log(# of employees) | -0.476 | 0.003 | Most significant variable, negative coefficient implies minus impact on hazard function (positive for survival) |
| # of R&D Employees | -0.060 | 0.150 | Positive for survival, not significant |

# Empirical Results Summary – Results Interpretation

- I was surprised to find only a single variable to be statistically significant at higher confidence levels, but all results seem intuitively valid

| **Independent Variable:** | **Size log(# of employees)** |
|---|---|
| Coefficient ($\beta$): | -0.476 |
| Conversion: | $e^{-0.476}$ = 0.62 (HR) |
| HR Interpretation: | Each unit increase in size of firm decreases firm bankruptcy risk by (1-0.62) = 38% |

# Empirical Results Summary – Real World Significance

- In addition to the typical use cases for this model (ex: healthcare) there are many other applicable analyses this type of model might prove useful
  - Understanding employee ”churn” within a company (what factors contribute to employees staying and leaving?)
  - Understanding what causes clients to cease doing business with another firm
  - Understanding what factors cause a borrower to be a higher quality credit risk (less likely to default on a loan)
- I aim to use a similar framework (time-dependent Cox regression model) to help me within the financial technology (fintec) space upon graduation
  - Effective in answering: what type of firms/individuals might we want to avoid loaning money to?

# Empirical Results Summary – Research Limitations

- The research could have been expanded to include more variables and combinations of variables from the original dataset (ex: annual R&D expenditure proportional to annual Sales)
- The dataset included some variables recorded in months (months to hazard) and others in years (firm founding year) which ultimately added inaccuracy into the model as unit value should be consistent
- Testing other model such as an accelerated failure time (AFT) model, or machine learning models (random forest, gradient boost, etc. adapted for a survival analysis) would have been an interesting complement to the study to further validate parameter estimations
- Stepwise methodology within a cox-regression has also been shown to be less effective at identifying statistically significant variables when the study involves larger numbers of variables

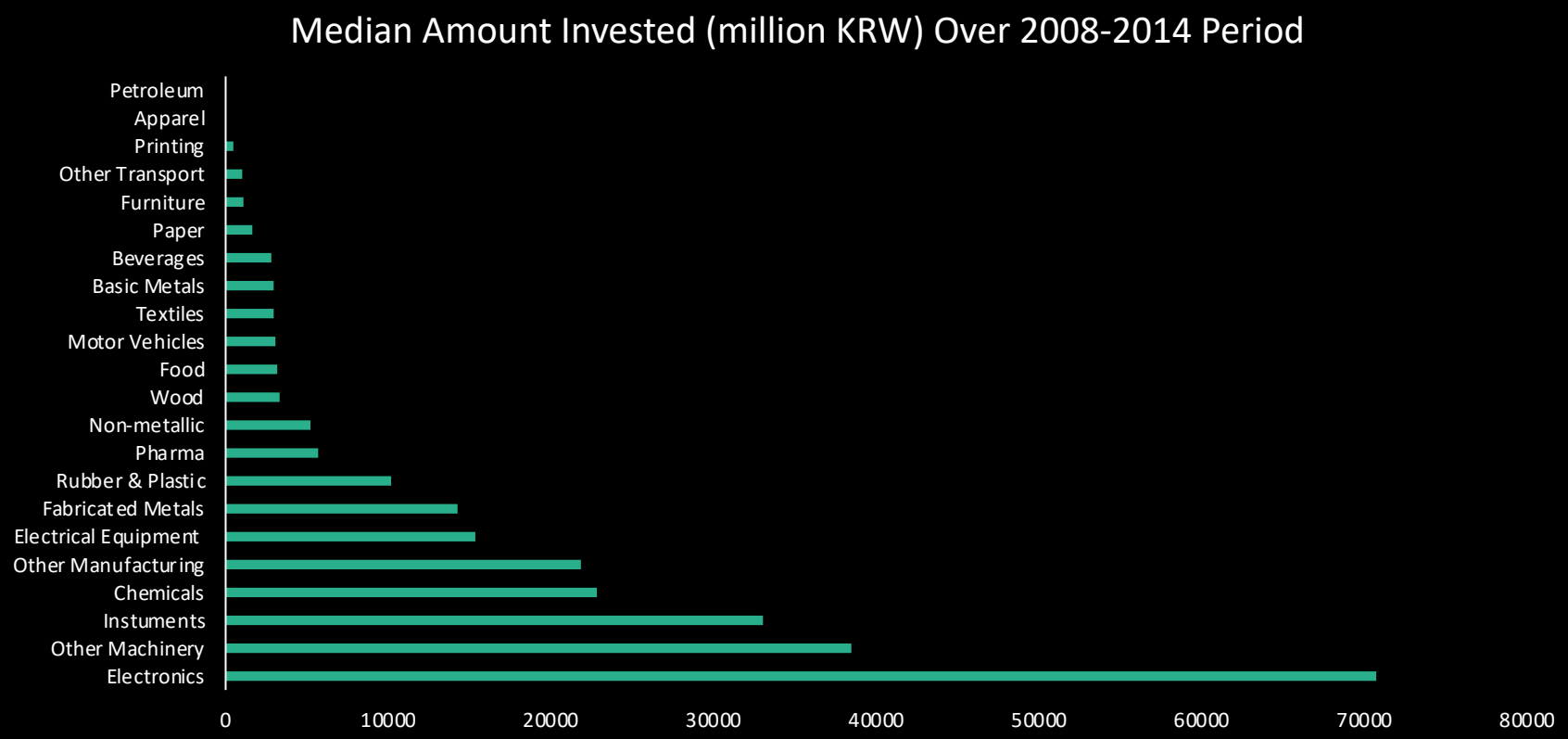
# Empirical Results Summary – Research Limitations

- The researchers chose to use a manufacturing sector index and not an alternative index such as the “information technologies & communications” offered by the **KSIC (Korean Standard Industry Classification** – or at least combined the indices to create a composite index
- R&D within each sector is likely to differ dramatically
- Sectors that made it to the final dataset include:

| Food | Beverages | Textiles | Apparel | Wood | Paper | Printing | Petroleum | Chemicals | Pharma | Furniture |
|---|---|---|---|---|---|---|---|---|---|---|
| Rubber & Plastic | Non-Metallic | Basic Metals | Fabricated Metals | Electronics | Instruments | Electrical Equipment | Other Machinery | Motor Vehicles | Other Transport | Other Manufacturing |

# Empirical Results Summary – Research Limitations

- Many of the sectors found in the dataset invest close to nothing in R&D (during both economic contractions and expansion periods) – this begs the question if firms in these sectors even engage in these types of capital outlays as part of a corporate strategy




# Extension & Analysis – Complementary Research

- Exploring how additional variables excluded from the original study (both variables found in the data file and variables I self generated) was an interest of mine
- **Firm Business Sector:**
  - Hypothesis 1: Specific business sectors naturally have a better survival likelihood during a recession
- **Research Costs As A Percentage of Total Sales:**
  - Hypothesis 2: Firms that invest a smaller *proportion* of revenue in R&D will have a higher likelihood for survival
- **Region:**
  - Hypothesis 3: Geographic region may play a role in determining a firms survival rate
- **Dataset Restriction:**
  - Hypothesis 4: By restricting the original dataset to sector which using increasing and decreasing R&D expenditures as a corporate strategy – we can improve upon the original model

# Extension & Analysis – Complementary Research

Assumptions made:

- **Firm Business Sector:**
  - The manufacturing index defined by KSIC would be broad enough to have both traditionally cyclical and non-cyclical sectors included
- **Region:**
  - The 3 regions selected in the study (Daejeon, Gwangju, and Daegu) would have unique enough properties to show up in a statistical study
- **Dataset Restriction:**
  - I wouldn't have to restrict the dataset/sector count too materially which would lessen the observation count too drastically and potentially introduce higher standard errors

# Extension & Analysis – Complementary Research

- Excluding additional variables pertaining to sector of firm, research cost as a proportion of sales, and region might be problematic for several reasons:
- **Bias in Hazard Ratios:**
  - This can result in misleading conclusions about the effects of variables on survival outcomes. The relative risk of an event occurring at any given time can be over or understated.
- **Confounding:**
  - Omitting relevant variables can introduce confounding into the analysis. Confounding occurs when a variable is related to both the independent variable and the dependent variable. This can lead to incorrect p-value estimates of the independent variable on the dependent
- **Misspecification of Model:**
  - The model may fail to capture the complex relationship between variables and their changing effects over time, resulting in a poor fit to the data and potentially inaccurate predictions

# Extension & Analysis – Complementary Research

- My findings from analyzing variables sector (KSIC), R&D as % of Sales (RnDSales), and region (Region) when analyzed together and not including variables size, age, and RnDP

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.024 | 0.027 | 0.376 |
| RnDSales | 0.000 | 0.007 | 0.971 |
| Region | -0.358 | 0.281 | 0.201 |

- No variables were statistically significant at the 95% confidence level, with region having the lowest p-value (closest to significant)

# Extension & Analysis – Complementary Research

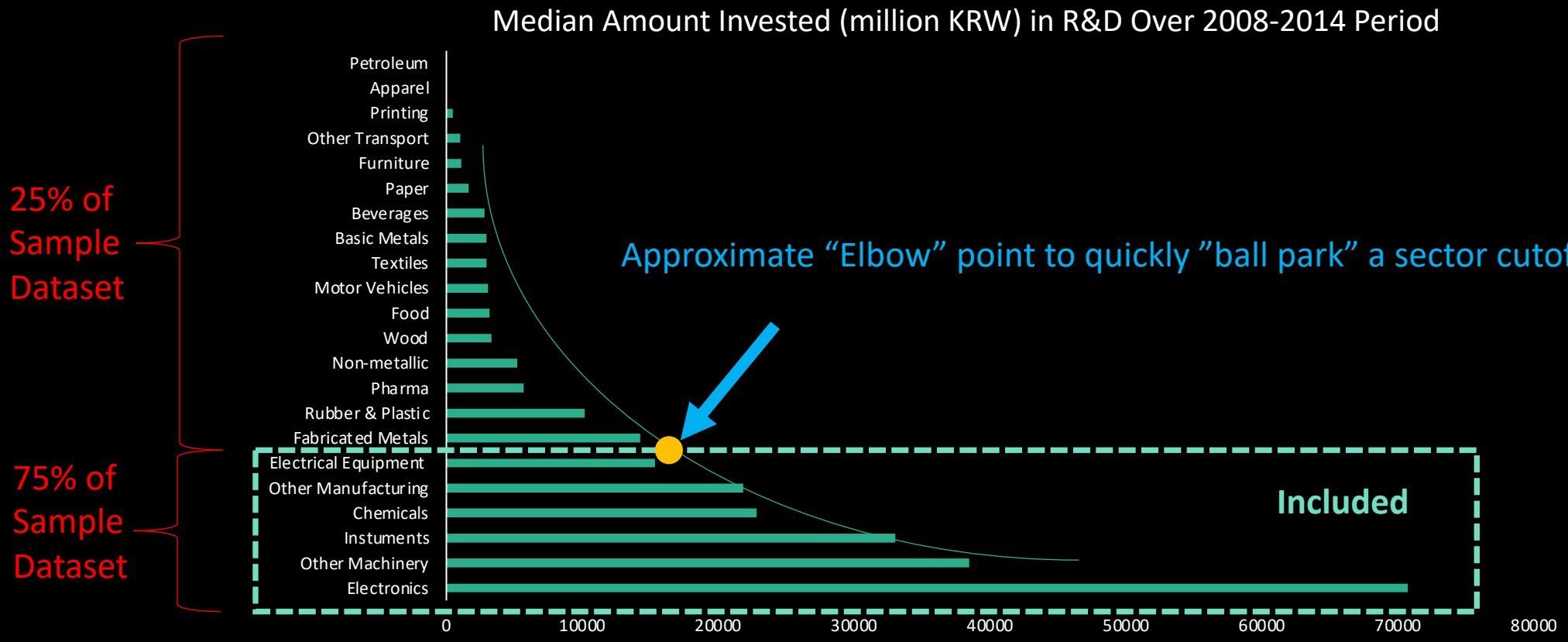
- My findings from analyzing variables sector (KSIC), R&D as % of Sales (RnDSales), and region (Region) when added to the previously analyzed variables in the model (size, age, and RnDP)

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.027 | 0.026 | 0.304 |
| RnDSales | -0.001 | 0.006 | 0.935 |
| Region | -0.312 | 0.285 | 0.275 |

- All variables continued to be insignificant
- Original test statistics and coefficients for size, agev, and RnDP remained unchanged at the second decimal place of precision

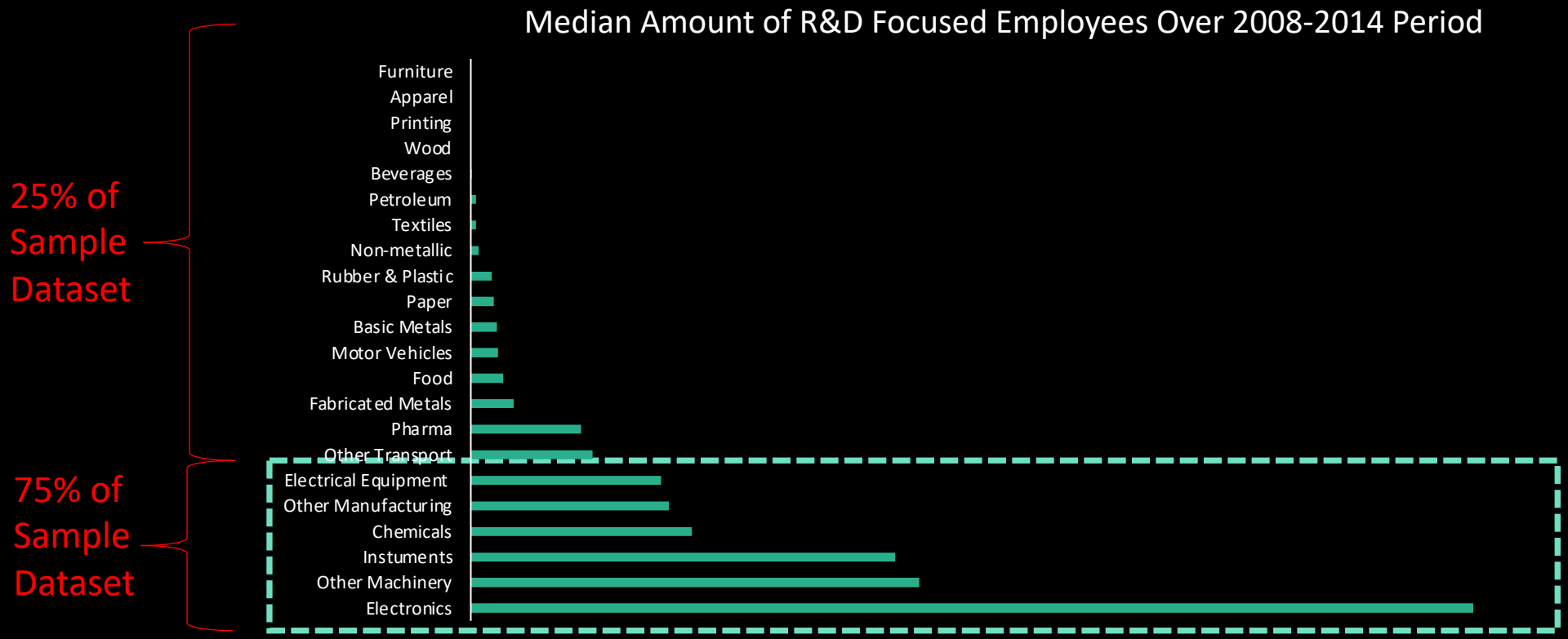
# Extension & Analysis – Complementary Research

- What happens if we screen the dataset for sectors which makes more sense to be in the dataset in the first place?



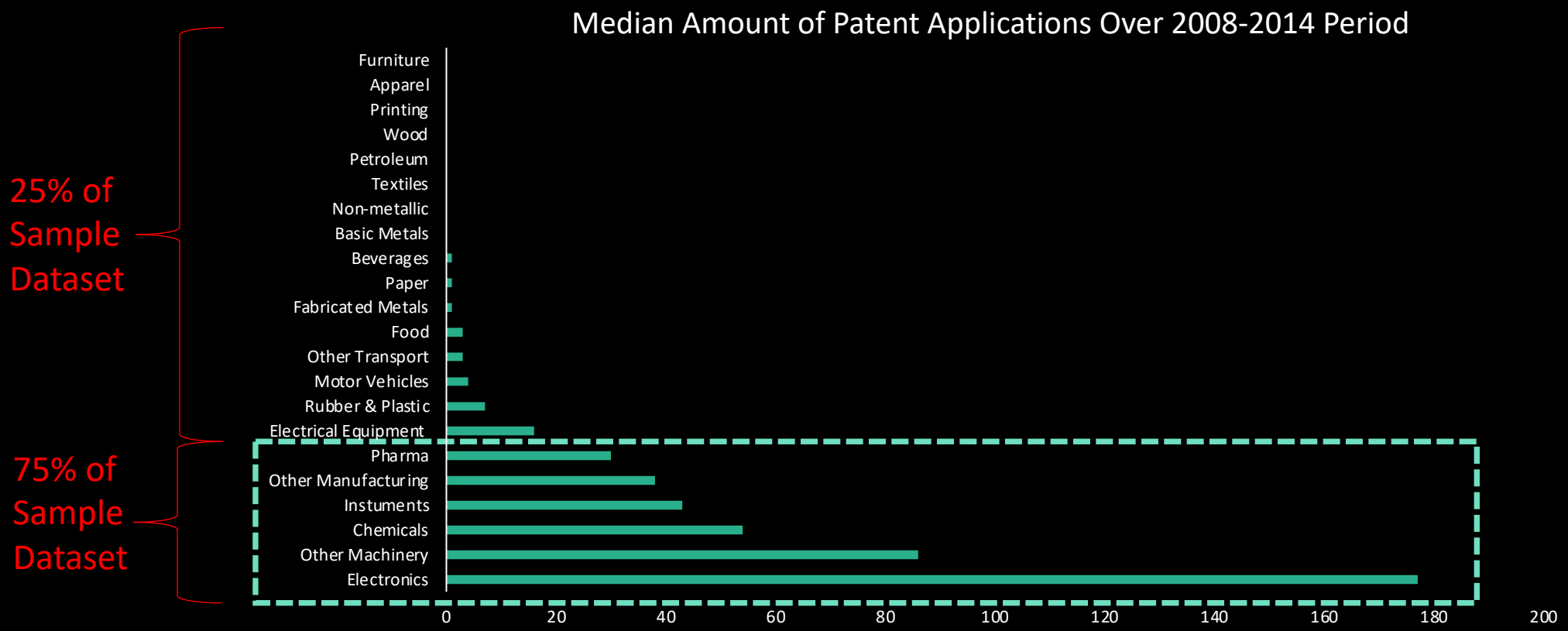

# Extension & Analysis – Complementary Research

- When looking at the same 6 selected sectors from the previous slide the same 6 sectors fall into the exact same bucketing



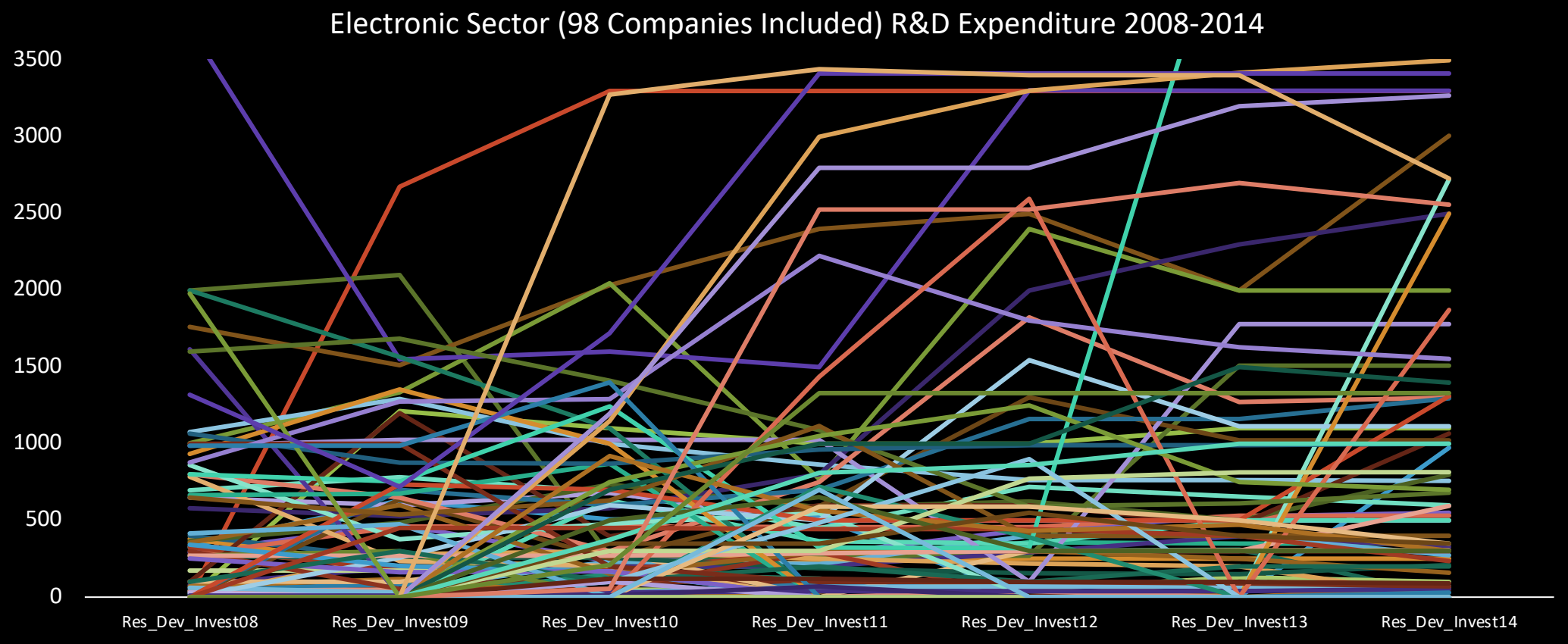

# Extension & Analysis – Complementary Research

- Almost identical bucketing when patents from each sector are analyzed (Pharma and Electrical Equipment almost change spots)




# Extension & Analysis – Complementary Research

- I also spot checked a couple of the largest sectors to confirm there were no obvious trends in company level R&D expenditure values from 2008-2014




# Extension & Analysis – Complementary Research

- My findings from analyzing variables sector (KSIC), R&D as % of Sales (RnDSales), and region (Region) when analyzed together and not including variables size, age, RnDP

All Sectors:

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.024 | 0.027 | 0.376 |
| RnDSales | 0.000 | 0.007 | 0.971 |
| Region | -0.358 | 0.281 | 0.201 |

Material Drop (KSIC) – Still Insignificant

Screened for Sectors Who Participate in R&D Investment:

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.187 | 0.146 | 0.199 |
| RnDSales | 0.000 | 0.006 | 0.968 |
| Region | -0.358 | 0.281 | 0.174 |

# Extension & Analysis – Complementary Research

- My findings from analyzing variables sector (KSIC), R&D as % of Sales (RnDSales), and region (Region) when added to the previously analyzed variables in the model (size, agev, RnDP)

All Sectors:

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.027 | 0.026 | 0.304 |
| RnDSales | -0.001 | 0.006 | 0.935 |
| Region | -0.312 | 0.285 | 0.275 |

Screened for Sectors Who Participate in R&D Investment:

| Ind. Var. | $\beta$ | Standard Error | Significance |
|---|---|---|---|
| KSIC | -0.151 | 0.195 | 0.436 |
| RnDSales | -0.001 | 0.006 | 0.923 |
| Region | -0.421 | 0.281 | 0.254 |

# Extension & Analysis – Conclusion

- While none of my added variables turned out to be statistically significant for both the original dataset and screened dataset – adding my new variables to the original variables and testing the model on the screened sectors which actually engage in research activities resulted in the following
- # of R&D Employees becomes statistically significant at 90% level, p-value for age is cut by >50%

| Original Study Results (from slide 11): | | |
|---|---|---|
| Ind. Var. | $\beta$ | Significance |
| Age | -0.030 | 0.304 |
| Size – ln(# of employees) | -0.476 | 0.003 |
| # of R&D Employees | -0.060 | 0.150 |

| Screened/Added Variables Results: | | |
|---|---|---|
| Ind. Var. | $\beta$ | Significance |
| Age | -0.062 | 0.136 |
| Size – ln(# of employees) | -0.429 | 0.037 |
| # of R&D Employees | -0.092 | 0.086 |
| Region | -0.438 | 0.242 |

# Extension & Analysis – Conclusion Summary

- By adding in previously unaccounted for variables in combination with screening the original dataset to better represent the researchers own original hypotheses, one of the previously insignificant variables (# of R&D employees) became significant at a >90% confidence level

| **Independent Variable:** | **# of R&D Employees** |
|---|---|
| Coefficient ($\beta$): | -0.092 |
| Conversion: | $e^{-0.092}$ = 0.912 (HR) |
| HR Interpretation: | Each R&D employee hired by firms specifically within the chemicals, electronics, instruments, electrical equipment, other machinery, or other manufacturing sectors decreases firm bankruptcy risk during a recession by approximately (1-0.912) = 8.7% |

- This finding adds sector precision to the researcher's statements that "R&D investment positively affects the probability of survival only when firms are innovative"

# If You Made It This Far - Thank You For Your Time

Sources:


Econometric Analysis of Cross Section and Panel Data, Jeffrey M. Woolridge

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6

https://stats.oarc.ucla.edu/sas/seminars/sas-survival/

https://www.hsph.harvard.edu/wp-content/uploads/sites/148/2012/10/fewell_stataj04.pdf