Nonresponse issues when analysing business survey data

Abstract

Data issues due to nonresponse or missing data arises often in company surveys or in firm data. Missing data and nonresponse causes bias. Another problem that causes bias is omitted variables. Accordingly, it will lead to wrong conclusions. The idea behind this licentiate thesis is to address these problems. The aim is to develop an insight into how common problems can be solved by transforming the data and changing the statistical method. There is no claim that the method suggested in the papers is always optimal. Rather, the goal of the papers is to give an awareness of problems that occurs in quantitative business research.

List of Papers

This thesis is based on the following papers, which are referred to in the text by their roman numerals.

I Grek, Å., Hartwig, F. (2018). Growing profitable or growing from profits: A comment. *Under revise procedure for Journal of Business Venturing Insights as a research proposal to a special issue*.

II Grek, Å., Hartwig, F., Dougherty, M. (2018). Auxiliary variables for nonresponse adjustment in business surveys.

III Grek, Å., Hartwig, F., Dougherty, M. (2018). Determinants of debt leverage ratios in Swedish listed companies.

My contributions to the papers were as follows:

Paper I – Data processing, data analysis, writing and revising the manuscript

Paper II – Data processing, data analysis, writing and revising the manuscript

Paper III – Data processing, data analysis, writing and revising the manuscript

Additional work:

Daunfeldt, S.-O., Grek, Å., Hartwig, F. & Rudholm, N., 2017. Betydelsen av internt genererat kapital för en långsiktigt hållbar företagstillväxt. i: J. Lithander, red. *Perspektiv på kapitalförsörjning - En antologi om företagens finansiering och statens roll.* Östersund: Tillväxtanalys, pp. 33-51

Daunfeldt, S.-O., Grek, Å., Hartwig, F. & Rudholm, N., 2017. *Betydelsen av internt uppbyggt kapital för företagens tillväxt och överlevnad*, Östersund: Tillväxtanalys.

Contents

Introduction

Microdata analysis, "*is a multidisciplinary field of knowledge dealing with the collection modelling, compilation and interpretation of large data sets, together with underlying algorithms, methods and techniques*" (Högskolan Dalarna, 2012). The aim is to obtain skills in data collection, data assessment and transformation, data storage, analysis of data and decision making based on the analysis. The emphasis is on the whole data process. This thesis is written in Microdata analysis. Microdata analysis contain five parts. The first part is data collection; the second part and third part are data- capture, processing and storage. The fourth part is analysis of the data, by using statistical modelling, simulation techniques, etc. The fifth and final part is decision-making and actions. This thesis focus on data processing and analysis.

Data issues, methodological problems and model misspecification appear in various research areas; one of them is research on business data, mainly concerning papers in business and administration and economics. In three well-cited papers (Brounen, et al., 2004; Davidsson, et al., 2009; Graham & Harvey, 2001) in international journals, in the area of business and administration, problems were found concerning the lack of data and neglecting to handle nonresponse. Consequently, it leads to bias estimates and incorrect conclusions when the absence of data and methods is not addressed properly. The idea behind this thesis is to address these problems. The aim is to develop an insight into how common problems can be solved by transforming the data and changing the statistical method. There is no claim that the method suggested in the papers is always optimal. Rather, the goal of the papers is to give an awareness of problems that occurs in quantitative business research.

When data is not enough

In order to verify that results are consistent, it is essential that the results must satisfy the qualities of reproducibility and replicability (Leek & Peng, 2015). Reproducibility denotes the ability to reproduce results given the same data set. Replicability implies given the same research question, the results will be consistent with previous results (Peng, Reproducible research in computational science, 2011). This is an increasing problem, where the public is starting to doubt the results of research, especially after some research scandals as the one in Karolinska Institutet, where a researcher and his six colleges have been found fabricating results in highly esteemed journals as the Lancet, Biomaterials, Journal of Biomedical Materials Research and Thoracic Surgery Clinics (Karolinska Institutet , 2018). The lack of

replication and replicability has not only affect medical research, it has affected other disciplines, such as business and administration (Aguinis, Cascio, & Ramani, 2017) and psychology (Pashler & Wagenmakers, 2012).

What are the reasons for this replicability crisis? It can depend on several issues: omitted variables, poor study design, missing data. Some even go as far as claiming that it is the lack of proper statistical education which is one of the foundations of the reproducibility crisis (Peng, The reproducibility crisis in science: A statistical counterattack, 2015). However "*it does not change the fact that problematic research is conducted in the first place*" (Leek & Peng, 2015).

There are three ways a study can be replicated. The first concern is the sample, which can vary over another period, another country, or different companies. The second way to replicate a study is to vary the measures or methods. The third and final way is to use an extended replication. An extended replication uses one of the version mentioned previously and it is also uses an improved and increased specification (van Witteloostuijn, Dejardin, & Pollack, 2018). In paper I, an extended replication study on the previous, work by Davidsson et al. (2009).

When data is missing

Often when a survey is a part of the data collection method, nonresponse occurs. If the nonresponse is not missing completely at random, (MCAR)[1] there is no problem with bias in the estimates. However, if the nonresponse is missing at random (MAR)[2] the estimates will suffer of nonresponse bias (Rubin, 1976). In some studies, such as Brounen et al. (2004) and Graham and Harvey (2001), they do different analysis to try to estimate if the observations are MCAR. The problem is that there is no way to test if data are MAR or MCAR (Thoemmes & Rose, 2014). The nonresponse bias can be reduced to some extent by high response rates. Even if the response rate is high, there is no way to completely exclude the bias, especially if the expected response rate is highly correlated with the studied variables (Groves & Peytcheva, 2008). However, Bethlehem (2002) argued that response propensity is a random process, and not decided by the properties of the surveyed objects. Even if the response propensity will be missing, the consequence will still be that the data suffers of bias.

---

[1] MCAR is expressed as $P(R|Y)=P(R|Y_{obs}, Y_{mis})=P(R)$, where $P(R)$ is the unconditional probability distribution of missingness and Y is the conditional probability distribution of missingness given the unobserved values of the variable (Y_{mis}) and the observed values of the variable (Y_{obs}). See Rubin (1976) and Thoemmes and Rose (2014) for a thorough description.

[2] MAR is defined as $P(R|Y)=P(R|Y_{obs},Y_{mis}) = P(R|Y_{obs})$, for a more detailed explanation, see Rubin (1976) and Thoemmes and Rose (2014).

Consequently, something has to be done to solve the problem. There are many different methods, but we decided to use the method suggested by Lundström and Särndal (1999) is built on the method by Deville and Särndal (1992) and later the work was extended by Deville et al. (1993), the calibration method. The calibration method is highly dependent on auxiliary information. In paper II, an examination of the efficiency of different auxiliary variables are performed. Several variables were considered not appropriate to use in the calibration estimator.

When there are omitted variables

It is common in questionnaires to use ordinal scales to rank the importance of a certain variable. The surveys themselves can then contain possible explanatory variables or the possible explanatory variable can come from another source (data triangulation).

The omitted problem appears when possible explanatory variables is left out from the regression specification. According to Clarke (2005) the bias can be increased by including as many explanatory variables as possible. Therefore, a model needs to be selected carefully. Model selection through AIC and BIC is not feasible when the possible covariates are large. The option is then to use penalized models, to select the best explanatory variables. While the model selection of the Lasso is not consistent, the elastic net can be under certain restrictions (Jia & Yu, 2010).

Another problem that appears is due to the variable of interest follows an ordinal distribution. For the Lasso and the elastic net, there are limited options to modulate ordinal data. In paper III, an investigation of the feasible models is performed. The cumulative generalized monotone incremental forward stagewise (GMIFS) method and the parallel element link multinomial-ordinal (ELMO) model were considered to select the best models.

Summary of the papers

Paper I

In paper I, we do a replication of the study reported in the paper by Davidsson et al. (2009). Davidsson et al. (2009) investigate if firms with higher profitability are more likely to reach a state of both high profitability and high growth compared to firms, which in the starting phase had higher growth and lower profitability, on Swedish and Australian data. However, the paper extends the replication (van Witteloostuijn, Dejardin, & Pollack, 2018) of the work by Davidsson et al. (2009) in several ways. First by expanding the Swedish data to include the whole population, not only a sample. Second, by including micro firms[3]. Third, by expanding the period. By expanding the period, we can also be examining if the relationships reported in Davidsson et al. (2009) hold for a longer period. Fourth, by having one more growth measure. Accordingly, Davidsson et al. (2009) study a sample of 1 482 private limited companies with 10-250 employees under the period of 1997 – 2000. The replication study uses 294 319 private limited companies, and 2 323 142 observations over the years of 1997-2010.

The method used in the paper is state transition matrices. First all companies were classified in five ordinal categories, *Star*, *Profit*, *Growth*, *Middle* and *Poor*, depending on their quantile values for growth and profit. Profit is measured as the companies return on assets. In Davidsson et al. (2009) they use

$$Growth_{it}=(Sales_{it}-Sales_{i1})/Sales_{i1} \qquad 1$$

which was found to be problematic since the growth measure depends on the sales for year one to the coming years. A company registered early in the data set will have more time to grow and therefore a higher probability to be able to achieve a higher growth, compared to companies registered in the end of the period. Another growth indicator was suggested

$$Growth_{it}= ln\left(\frac{Sales_{it}}{Sales_{it-1}}\right) = ln\, Sales_{it} - ln\, Sales_{it-1} \qquad 2$$

When we used the growth measure suggested by Davidsson et al. (2009) (equation 1), the paper concludes the same relationship for short time period, up to two years, companies with a higher return on assets, have higher probability to achieve both high growth, and high profitability. In a longer

[3] Here we denote micro companies, as private limited companies with 0-9 employees, see European Commission (2003).

perspective, four years or more, companies that first grow before they build a high return on assets have a better chance of achieving long-term growth, and high profit, compared to the companies that build up high return to assets before they grow. Companies with high return on assets are more likely to end up with both low growth, and low profitability, compared to companies with a high growth in the initial state. The paper concludes that the hypothesis of (Davidsson, Steffens, & Fitzsimmons, 2009) holds but the relationship of the original state and the final state is weaker the longer the investigated period is. For the alternative growth measure, the findings indicate that companies, which first built a high return on assets before they grow, have a better chance of achieving long-term growth, compared to companies that grow before they build up capital.

Paper II

In paper II we used survey data on all Swedish listed companies for 2005 and 2008 (see Hartwig (2012) and Daunfeldt and Hartwig (2014)). The survey was a replicate of the study originally performed by Graham and Harvey (2001) in US and by Brounen et al. (2004) in UK, Netherlands, France and Germany. The main problem with the study by Graham and Harvey (2001) and Brounen et al. (2004) was the level of nonresponse recorded in the data 91, and 95 percent respectively, which will cause biased results if the observations are not missing completely at random (MCAR).

The calibration method by Lundström and Särndal (1999) and Särndal and Lundström (2005) was used. Monte Carlo simulation was performed and the results showed that even for large simulated nonresponse set, with the use of appropriate auxiliary variables, the simulation bias was maximum two percent. With one auxiliary variable, which was not so impropriate, the simulation biased and variance increased. For larger simulated nonresponse, the calibration method could not be estimated, for the impropriate auxiliary variables.

Paper III

Paper III is based on the same data as in paper II. In the survey by Graham and Harvey (2001), one of their research questions regards target debt. "*We ask directly whether firms have an optimal or 'target' debt-equity ratio. Nineteen percent of the firms do not have a target debt ratio or target range. Another 37% have a flexible target, and 34% have a somewhat tight target or range. The remaining 10% have a strict target debt ratio*" (Graham & Harvey, 2001, s. 211). In the survey by Hartwig (2012), later used by Daunfeldt and Hartwig (2014), the same question is used with the same answer option (see appendix 1, question nine in Daunfeldt and Hartwig (2014, s. 111)).

The aim of the paper was to examine; which company characteristics affect the chosen level of target debt. One opportunity arose when the previous literature in the field was insufficient. Therefore, another

method was needed to compensate for the lack of previous research. Therefore, the choice was to use a penalizing model.

The most common penalizing models are the lasso by Tibshirani (1996), and the elastic net by Zou and Hastie (2005). The problem is to fit an ordinal logit penalizing model. For the lasso model, there is not today any way to fit an ordinal lasso, but a multinomial lasso model can be fitted on the ordinal dependent variable, since the ordinal data is a special version of the multinomial data (Wurm, Rathouz, & Hanlon, 2017). The multinomial lasso model was fitted using two versions: the grouped and the ungrouped (Hastie, Tibshirani, & Wainwright, 2015).

Wurm et al. (2017) propose a class of models called the element link multinomial-ordinal (ELMO). The ELMO is a subset of the vector generalized linear models, and is usually fitted with a coordinate descent algorithm, to ordinal and multinomial regression models with an elastic net penalty. There are three version of the ELMO class model: the parallel-, nonparallel- and the semi-parallel model. The difference between the models is the shrinking procedure and parameters.

The final model used estimate the use of target debt, was the cumulative generalized monotone incremental forward stagewise (GMIFS) method by Archer et al. (2014). GMIFS fits a penalized method on an ordinal data by incremental forward stagewise (IFS) method.

The different penalizing models yielded different results. Wurm et al. (2017) did a simulation study, where they examined these models and they argued that cumulative GMIFS was the superior model, then the parallel and semi-parallel ELMO, came in a second place. In our case, the parallel ELMO and cumulative GMIFS both only selected quick asset ratio as the only possible explanatory variable. According to Friedman et al. (2001, p. 91), an unpenalized regression model can be fitted after a penalized model. From the unpenalized order logit models, we found that the only significant variable at five percent or lower was quick asset ratio.

Suggestions for future research

At the same time as working on the three papers, a knowledge gap was discovered. There is a lack of knowledge in company response behaviour to surveys. If the nonresponse in a survey is missing completely at random, (MCAR) there is no problem with bias. However, if the nonresponse is not MCAR, it will cause bias (Rubin, 1976). There is no way to test if the nonresponse is MCAR (Thoemmes & Rose, 2014). In Japec et al. (2000) they list some of the reasons why companies neglect to answer surveys, it may depend on more modest factors as the timing and wrong address, but also more company specific features. Other reasons may well be subject to staff change, ownership exchange, rearrangement, lack of money, lack of knowledge in the questioned topic, difficult and/ or demanding survey, poor or non-existent basis, as well as low priority. In surveys aimed towards individuals, it is known that certain personal characteristics lower the probability of obtaining an answer. Most difficult to reach are young, metropolitan, low skilled and foreign-born. If the study object to one or several of these categories, there probability of attaining an answer are lower than other categories (Höjer, 2017). The primary hypothesis we have is that company nonresponse may also be subject to other company characteristics, which can be detected in the accounting information.

The aim for the continuation is to investigate if the response rate depends on firm characteristics. A survey will be send to companies, with three reminders. The reminders will act as deadlines, and the respondents will be viewed as a respondent-group. The respondents in each respondent-group will be matched with their accounting information. Analysis will be carried out on each of the respondent-groups. The hope is to detect patterns in the accounting information of the companies responding without reminders, after the first reminder, after the second reminder and for those companies, which do not respond until the third reminder. The accounting data comes from the Swedish Companies Registration Office (Bolagsverket), where all limited companies has to send their annual report. The database contains over 170 variables for all limited companies.

Overall, the research so far has discussed the problems appearing with missing data, lack of data, and with omitted variables. My findings are expected, given the used methods. The recommendations I can draw from the research I conducted so far, is that more considerations should be done, when researchers on company data face data issues.

References

Aguinis, H., Cascio, W. F., & Ramani, R. S. (2017). Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies , 48*, 653-663.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., & Gentry, A. E. (2014). ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer informatics, 13*, 187-195.

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. Little (Eds.), *Survey Nonresponse* (pp. 275–288). New York: Wiley.

Brounen, D., de Jong, A., & Koedijk, K. (2004). Corporate finance in Europe: Confronting theory with practice. *Financial Management , 33*(4), 71-101.

Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science, 22*(4), 341-352.

Daunfeldt, S.-O., & Hartwig, F. (2014). What Determines the Use of Capital Budgeting Methods? Evidence from Swedish Listed Companies. *Journal of Finance and Economics, 2*(4), 101-112.

Daunfeldt, S.-O., Grek, Å., Hartwig, F., & Rudholm, N. (2017). Betydelsen av internt genererat kapital för en långsiktigt hållbar företagstillväxt. In J. Lithander (Ed.), *Perspektiv på kapitalförsörjning - En antologi om företagens finansiering och statens roll* (pp. 33-51). Östersund: Tillväxtanalys.

Daunfeldt, S.-O., Grek, Å., Hartwig, F., & Rudholm, N. (2017). *Betydelsen av internt uppbyggt kapital för företagens tillväxt och överlevnad.* Östersund: Tillväxtanalys.

Davidsson, P., Steffens, P., & Fitzsimmons, J. (2009). Growing profitable or growing from profits: Putting the horse in front of the cart? *Journal of Business Venturing, 24*(4), 388-406.

Deville, J.-C., & Särndal, C.-E. (1992, June). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association , 87*(418), 376-382.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *American statistical Association, 88*(423), 1013-1020.

European Commission. (2003). *Definition of micro, small and medium-sized enterprises.* Luxembourg: Office for Official Publications of the European Communities.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (10 ed., Vol. 1). New York, NY, USA: Springer series in statistics.

Graham, R. J., & Harvey, R. C. (2001). The theory and practice of corporate finance: evidence from the field . *Journal of Financial Economics, 60*, 187-243.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly, 72*(2), 167-189.

Högskolan Dalarna. (2012). *Curriculum for doctoral studies in Microdata Analysis* . Falun : Högskolan Dalarna.

Höjer, H. (2017, May 19). Vad tycker de som inte svarar? *Forskning & Framsteg*.

Hartwig, F. (2012). The Use Of Capital Budgeting And Cost Of Capital Estimation Methods In Swedish-Listed Companies . *The Journal of Applied Business Research, 28*(6), 1451-1476.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity.* Boca Raton, FL: CRC Press.

Japec, L., Ahtiainen, A., Hörngren, J., Lindén, H., Lyberg, L., & Nilsson, P. (2000). *Minska bortfallet.* Örebro : Statistiska centralbyrån .

Jia, J., & Yu, B. (2010). ON MODEL SELECTION CONSISTENCY OF THE ELASTIC NET WHEN p ≫ n. *Statistica Sinica, 20*(2), 595-611.

Karolinska Institutet . (2018, June 29). *Karolinska Institutet.* Retrieved September 20, 2018, from https://ki.se/en/news/seven-researchers-responsible-for-scientific-misconduct-in-macchiarini-case

Leek, J. T., & Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences, 112*(6), 1645-1646.

Lundström, S., & Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Offcial Statistics, 15*(2), 305-327.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226-1227.

Peng, R. D. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30-32.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* West Sussex, England: John Wiley &Sons Ltd.

Thoemmes, F., & Rose, N. (2014). A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems,. *Multivariate Behavioral Research, 49*, 443-459.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

van Witteloostuijn, A., Dejardin, M., & Pollack, J. M. (2018). *Journal of Business Venturing Insights*. Retrieved September 20, 2018, from https://www.journals.elsevier.com/journal-of-business-venturing-insights/call-for-papers/large-scale-replication-initiative-entrepreneurship

Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2017, June 19). Regularized Ordinal Regression and the ordinalNet R Package. *ArXiv e-prints*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series, 67*(2), 301–320.