Stay ahead of competition

Bibliographical Data and Classifications

Library of Congress Classification (LCC) http://lcweb.loc.gov/catdir/cpso/lcco/lcco_h.pdf	HF 5001-6182, HF 5410-5417.5, HF5438.8.C6
Journal of Economic Literature (JEL) http://www.aeaweb.org/journal/jel_class_system.html	C 44, M 31, C80
Gemeenschappelijke Onderwerpsontsluiting (GOO) Classification GOO	85.00, 85.03, 85.40
Keywords GOO	Bedrijfskunde / Bedrijfseconomie, marketing, methoden en technieken, operations research, econometrie, data, internet, concurrentie, redes (vorm)
Free keywords	econometrics, quantitative marketing, business data, big data, Bayesian inference, data mining, forecasting

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus Universiteit Rotterdam
Internet: www.erim.eur.nl

ERIM Electronic Series Portal:
http://hdl.handle.net/1765/1

Inaugural Addresses Research in Management Series
Reference number ERIM: EIA-2013-054-MKT
ISBN 978-90-5892-346-2
© 2013, Dennis Fok

Design and layout: B&T Ontwerp en advies (www.b-en-t.nl)
Print: Haveka (www.haveka.nl)

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

Stay ahead of competition

Address delivered in adjusted form at the occasion of accepting the appointment of
Endowed Professor of Applied Econometrics
at the Erasmus School of Economics, Erasmus University Rotterdam,
on behalf of Vereniging Trustfonds EUR, on Friday, October 4, 2013

Dennis Fok

Erasmus School of Economics
Erasmus Universiteit Rotterdam
Postbus 1738
3000 DR Rotterdam
E-mail: dfok@ese.eur.nl

Samenvatting

Het wijdverspreide gebruik van het internet en van computersystemen in het algemeen heeft er voor gezorgd dat er tegenwoordig data beschikbaar is over vrijwel alles. Deze data is van een grootte en detailniveau dat vroeger voor onmogelijk werd gehouden. Dit type data is vaak niet direct vergelijkbaar met de data die historisch gezien in de econometrie wordt bestudeerd. Dit alles brengt mogelijkheden en uitdagingen voor academische onderzoekers, bedrijven, en zelfs voor econometrie als wetenschapsgebied en opleiding. Al deze partijen kunnen, of zelfs moeten deze data gebruiken om hun concurrentie voor te blijven.

Abstract

The widespread use of the Internet and computer systems has led to a situation where data are available on almost everything. The volume and the level of detail of these data is something we considered to be impossible until a few years ago. Researchers in economics and business now have access to a new variety of data. Such data are often not directly comparable to the data that have historically been considered in econometrics. This brings opportunities and challenges for academic researchers, companies, and even econometrics as a field and as an educational program. All parties involved can, or even have to use these data to stay ahead of the competition.

Content

1. Introduction

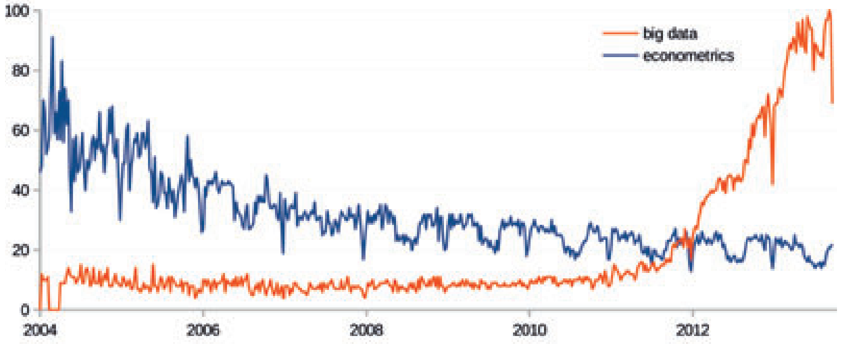
Dear Rector Magnificus,
dear board members of the Vereniging Trustfonds,
dear colleagues, friends, and family,
dear distinguished guests.

Competition is pervasive. Companies have to fight off their competitors, academics compete with each other on new findings and over limited space in the top academic journals, and even the field of econometrics is in competition with other fields. Today, I would like to discuss how all three of these can stay ahead of the competition. I will advocate one way to achieve this, namely through an econometric view on currently available data. I will specifically focus on online data and micro data available within firms. In many ways, these data differ from the data that have historically been studied in econometrics and econometrics as a field seems to ignore some important developments in this area. In the long run, the field faces a threat of being overtaken by computer science and machine learning.

We are currently often confronted with the term Big Data in the news and in the popular press. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process them using standard database management tools or traditional modeling approaches. Big data dominates popular discussions on data analysis and there seems to be more interest in this term than in econometrics. Figure 1 substantiates this claim by showing the relative interest in "Econometrics" and in "Big Data" as measured by the number of Google search requests containing these words. In itself this figure is an example of a new type of data that is currently available. The figure gives time, from 2004 to 2013, on the horizontal axis and a relative measure of search requests on the vertical axis. The maximum number of search requests in a month is scaled to 100. The figure clearly shows that interest in big data exploded in 2011 and 2012, and has increased by a factor 10 over the past three years. Overall the trend for econometrics is declining, and since 2012 there has been more interest in big data than in econometrics.

Big data seems to be everywhere and the media claim that it is something that will change our lives dramatically. I do not entirely agree with this statement. On the one hand, a large part of this is true, but on the other hand, big data is a hype. Many things that are discussed have already been around for a long time. Only the label has changed, perhaps under the influence of a smart marketeer. Many questions under this label have been studied in the domains of applied econometrics, quantitative marketing, and applied statistics. Econometricians have a long history of studying various phenomena using large quantities of data, and most of these problems are closely related to economics and business. In the recent past, and before the big data hype, people for example studied on-line behavior of individuals, and consumer-level preferences based on revealed choices.[1] So, in this sense big data is not really new. However, this does not mean that we can simply ignore the trends I have shown.

Figure 1. Development of Google searches on the terms "Big Data" (red) and "Econometrics" (blue) (Source: Google trends, September 2013, see http://www.google.com/trends)



I will not deny that the quantity, depth, and breadth of available data have grown exponentially. In fact, the amount of data may have grown beyond what can be modeled using standard econometric techniques. Especially available online data have increased exponentially. Let me give you some examples of the amount of data that we generate: in 2012 every minute of the day, Google received 2 million queries, Facebook users shared 600 thousand pieces of content, 204 million emails were sent, and 571 new websites were created.[2] Together with the increase in business data, we now have information at a very detailed level. As a result, many aspects of human behavior can be quantitatively studied.

This data availability is a goldmine for econometricians. Data are available on almost everything! However, making good use of such data is not easy, and some challenges still have to be resolved. In the next half an hour, I will take you through some of the essential requirements to really obtain information from available data. At the same time, I will identify a number of challenges that require further research. As my personal focus is on business related questions, I will mainly focus on studies that are relevant for consumers or businesses.

[1] See for example, Johnson et al. (Management Science, 2004), and Chintagunta (Marketing Science, 1992).

[2] Source: http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/

2. Studying business processes

In business, managers are now well aware that the data they collect, or could collect, about their customers and competitors can be very valuable. The idea that information is the key to good business processes, and subsequently to sound financial performance is now well known. Currently available data are extremely detailed. As I have said before, we have information about almost everything. In many business contexts, we can summarize available data by a number of W's:

Who contacts Whom, with What, When, Why and with What outcome?

Let me give you an example. Telecom companies regularly contact their customers with offers. These offers are usually aimed at retaining customers or persuading them to renew their contract. The company will at least store information on 1) the customer they contacted; 2) the current status of the customer; 3) the customer's current calling pattern; 4) the offer made to the customer; 5) why the customer was contacted; and 6) what happened after the contact. After a number of customers have been contacted, the company can use the data to assess the added value of the offer. In some cases, such analysis may give unexpected results. For example, in one of our studies we found that the contact had a negative value for a particular group of individuals – those who were not very satisfied with the telecom company. When contacted, these people were reminded of the fact that they could terminate their contract, and they immediately decided to do so. The company had not only wasted its efforts, it has also lost a customer earlier in time.

In order to conduct studies on business processes like in this example, we need three key ingredients: 1) data; 2) models; and 3) techniques. I will next discuss these ingredients one by one.

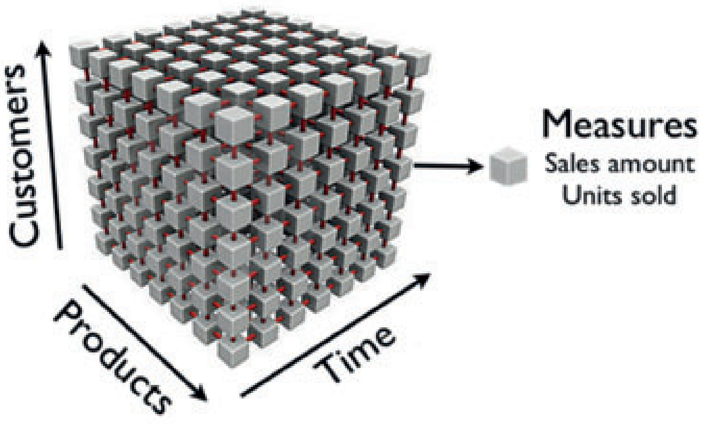
3. Data

Types of data

Datasets can be classified according to their dimensions. *Cross-sectional* data give information on the behavior of a number of individuals at one point in time. *Time-series data* represent the development of a single variable, for example, the output of one firm, over time. *Two-dimensional panel data* are a combination of both of these and give information on, for example, the behavior of 1000 individuals over time. These three options form the classical datasets that have been considered in theoretical econometric work.

However, current questions and data do not always fit this one-dimensional or two-dimensional setup. Current data are often multi-dimensional. For example, we may have data on individuals visiting different websites over a period of time and their actions on these sites, or data on individuals making purchases in various product categories over time. Dependent on how we view such data, we have at least three dimensions: the individuals, the websites or product categories, and time.

Figure 2. Example of data cube for three dimensions (customers, products, and time)



One way to visualize a three-dimensional data set is using a so-called data cube. Figure 2 illustrates a data cube for the three dimensions: customers, products, and time. With this cube, I illustrate a case where data are available for individual customers on their purchases over time across a number of different products. Each measurement in this example is represented by a box and gives the sales or expenditures by a customer of a product in a particular period. Each connection between two boxes symbolizes a potential relationship between products, customers, or time periods. Complicating things further, we may even have more than three dimensions, for example, when we follow customers over different geographical regions, or individuals within a group, such as, a firm or a household. In general, we have multiple measures. Besides sales data, we also have information about, for example, price and promotional activities. Some data may even not be numeric, for example, descriptions of, or reviews on available products.

Current econometric theory focuses on numeric data of one or two dimensions. My simple examples above highlight the need for more work on more complex types of data.

Data collection

Although a lot of data are readily available online or from a company's administrative records, data collection remains important. All company actions lead to data. In everyday business, we should think about all the information that is generated in this way. In some cases, this means that actions could be tailored so that we obtain the most informative data. We should not only think about the direct benefits, but also about how to maximize the information value of the to-be-collected data.

Let me give you another example. Suppose a company is interested in selling an upgrade to one of their services. Although customers can just buy this upgrade, the company decides to contact some customers to inform them about this upgrade. During a brainstorming session, an employee comes up with the idea that they should only contact their best customers as these are the most likely to buy the upgrade. This would ensure that contacting them would actually be worth the effort. Although this sounds sensible at first, there are four main issues with this idea:

1. The claim may not be true:
   The best customers may not be the most likely to buy the upgrade.
2. It is suboptimal:
   Although the best customers may be likely to buy the upgrade, they may buy it anyway.
3. It complicates later analysis:
   A basic econometric analysis to evaluate the impact of the campaign would compare the upgrade frequency of those who are contacted versus those who are not. The difference between these two groups should represent the impact of the campaign. However, the target selection will lead to a distorted view of the true effect. The most important difference between the two groups is that the contacted group contains the best customers by definition. In econometrics, this problem is known as *endogeneity*; the company's action is based on the expected outcome. Although solutions to this problem are available, they all require knowledge of the selection process. The company should therefore at least formalize the decision rules of the applied target selection and keep these for further reference.
4. It could make later analysis impossible:
   Suppose that all good customers, according to some definition, are contacted. It is now literally impossible to measure the impact of the campaign. Being contacted is now equivalent to being a good customer. There is no way to disentangle the impact of, being contacted and being a good customer, on the probability of buying the upgrade.

Although contacting all good customers may be optimal from a short-term business perspective, it prevents or at least complicates the possibility of econometrically investigating its impact. In other words, we cannot confirm or refute that the selection was indeed a good idea. This may be good for the person who came up with this idea, but it is not optimal for the firm. We should consider the option value of analyzing and learning about customers, and develop a smart campaign that later allows for a sound econometric analysis.

In line with the above, failures are not as bad as they may seem at first. A failed marketing campaign is never a wasted effort if the campaign is designed well. It will always generate information that can be used to further optimize later actions.

Sharing data

Broadly speaking, academic research in this domain has two potential targets. First, it may aim to develop new methodologies, that is, create new models or new techniques. Second, its objective may be to develop and test theories on individual behavior. We need access to real-life data for both.

The big data development seems to imply that unlimited data are available to everyone. This is certainly not true. Data are available, but not to everyone. Business data is usually collected and kept by companies themselves. Companies are well aware of the value of this data, and their first natural reaction is to protect them. I firmly believe that the academic and corporate community can both benefit from information sharing and collaboration. For researchers, this means that academic progress can be made and in return, companies are informed about the latest methods and academic developments, and they get to be the first to implement new techniques. This information sharing can result in academic papers with large contributions, even without revealing sensitive company information. This seems to be the optimal combination where both parties can stay ahead of their competition.

Fortunately, several companies make their data available for academic use. For example, IRI has made very detailed data available on the sales records of many supermarkets in the United States. This data is accessible to everyone. Other companies choose to make their data available only to a selected group of researchers. For example, The Wharton Customer Analytics Initiative of the University of Pennsylvania acts as a "matchmaker" for companies who want to do this. Companies select groups of researchers through a competitive process based on research proposals. Recently, Peter Verhoef and Tammo Bijmolt from the University of Groningen, Matilda Dorotic from BI Norwegian Business School, and I were granted access to detailed data on loyalty card usage. Direct cooperation between a company and a team of researchers is also possible. Recently, Bas Donkers, our PhD student, Bruno Jacobs, and I reached an agreement with a large online store. We will be allowed access to part of their massive database for our research. All these collaborations create opportunities for companies and researchers to learn and develop new skills.

4. Models

The second key ingredient for analyzing business processes is models. To an outsider, the concept of a model may be unclear. Let me, therefore, first try to define what an econometrician considers to be a model:

A model is a mathematical/statistical representation of reality and always involves simplifications.

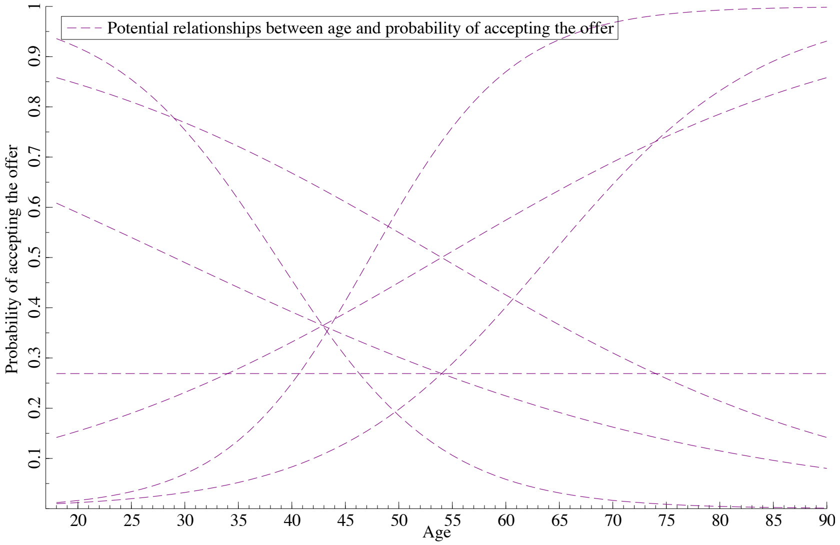
I am sure that this definition will still not be clear to everyone. For example, you may ask: "What is a statistical representation of reality?" Let me explain. A statistical representation assigns probabilities to possible outcomes and describes the potential relationship between these probabilities and external factors. Parts of this relationship will be unknown, but these parts can be learned or estimated when data are available.

As an example, let us consider a customer's response to an offer made by a salesperson. There are two potential outcomes: accept or reject the offer. A model should therefore assign probabilities to these outcomes for every individual. This probability may depend on a number of characteristics of the customer, for example, age. An econometric model specifies a particular form for the relationship between age and probability. In general, the exact shape of the relationship is unknown, but the econometric model gives the shapes that we want to consider. Figure 3 shows a number of potential relationships for a particular model. Each line represents a potential shape, and each shape can be characterized by two numbers.[3] These numbers are referred to as parameters. If data are available, econometric methods allow us to identify the line that best fits the data. In other words, the parameters can be estimated. This estimated relationship can next be used to identify the best targets for a salesperson. Of course, in reality more characteristics play a role than just age. Including more characteristics leads to more parameters and makes it more difficult to visualize the main idea of the econometric model.

[3] The model that is considered here is a logit model, the two parameters are the intercept and the impact of age.

A model needs to capture two special features in the context that we are discussing today. These two features are *heterogeneity* and *sparseness*.

Figure 3: Examples of potential relationships between an individual's characteristic and a probability as implied by an econometric model



Heterogeneity

Individuals are different! Even if two individuals have exactly the same observed characteristics, they may behave differently. All individuals have their own peculiarities which translate into specific behavior. The technical word that is used to refer to these differences is heterogeneity. It is important that models take heterogeneity into account. One of the most popular techniques to do so is to allow each individual to have its own specific parameters in the model. If we have adequate data, we can control for and even estimate these individual-level parameters. In the end, this allows us to identify and develop tailor-made actions for specific individuals.

Sparseness

Although we currently often have too much rather than too little data, in some important aspects, the information in the data can still be limited. In other words, a large data set may be quite small in terms of the information it provides. Suppose we collected information on purchases made by a large number of individuals over a large set of products over time. We would know who buys what products on each day. Such a dataset potentially gives us a lot of information on the preferences of these individuals. Combined with external information, we could make a detailed study of the short and long-term impact of advertising including competitive cross-effects. However, such data is limited in an important aspect. There will be many days on which a particular individual does not buy a specific product, and there will even be products that he never buys. This especially holds true if we consider large assortments such as all the products available at Albert Heijn or at Amazon.com. In other words, there are many zero-observations. Datasets like this are called sparse datasets.

The challenge with sparse data is that all the zeros may make it difficult to see the information that is in the data. In this case, the challenge is to construct the model that "explains" the zeros, but also finds information on the actual purchases. Using the structure of an econometric model is helpful in making a good analysis, especially in these cases.

Why do we need models?

The media often states that we no longer need models in this time because big data should give all the necessary information by itself. I do not agree with this. First of all, many big data techniques actually do rely on models. Secondly, an approach based on a probabilistic model has many advantages as it allows us to quantify the uncertainty of the effects we find, and to make optimal decisions even when there is such uncertainty. Finally, a model-based approach helps to control the risk of finding spurious relationships, in other words, "discovering" results which are not really there. I will discuss two of these arguments.

Let us consider whether current big data approaches actually do not use any models. As mentioned earlier, a model is a statistical representation of reality. A model can also be seen as a way to describe the assumptions made in an analysis. In the model example I gave before (see Figure 3), the implicit assumption is that there is a smooth relationship between age and probability. Suppose that we need to predict the probability that a customer who is 36 years old plus 185 days accepts or rejects the offer made by a salesperson. Someone like this may not have been observed before, and without assumptions it would be impossible to predict his behavior. Using the smoothness assumption, we conclude that this individual must be very similar to people in the age group 36 to 37. This allows us to predict behavior. Conversely, as soon as a comparison with "similar" people is made in any technique, a smoothness assumption is used. In this sense, all analyses are based on a set of assumptions. However, in an econometric model we make these assumptions explicit.

The models that are currently used in the big-data context are often very simple. In fact, they are sometimes so simple that we may not recognize them as models anymore. The main reason is not that these simple models are the best, but that it is very difficult to apply other, more complicated models on a large scale. Here is another opportunity for academic researchers to contribute to the field of econometrics. We need to find a more efficient way to use complex models on a large scale. I will come back to this issue in the next part of this lecture.

Another need for models is related to the earlier mentioned concept of endogeneity. In principle, we often aim to uncover causal relationships between variables. We may want to predict the choices individuals make when their income rises. Merely looking for patterns in a dataset does not always yield trustworthy information. Even if many individuals with a high income buy a particular product, this does not mean that there is a causal relationship between income and the product. An increase in income does not necessarily lead to an increase in purchase probability. For example, the product may be particularly popular in certain affluent regions. Econometric models and techniques can be used to control for this. Merely searching for patterns in a data file is more vulnerable to finding false relationships.

5. Techniques

The third key ingredient for analyzing business processes is techniques. Almost all models contain unknowns: things that have to be estimated. An econometrician uses data to estimate these unknowns. In general, the unknowns are parameters, numbers that specify a particular relationship like the numbers that specify the shape of the relationship in Figure 3. Various techniques are available to perform this estimation task. The general principle is that we should set the parameters such that available data is best explained. In some situations, this task is straightforward. However, as the models and datasets become larger and more complex, the estimation task also becomes more challenging. Just calculating the “fit” of a model may become non-trivial.

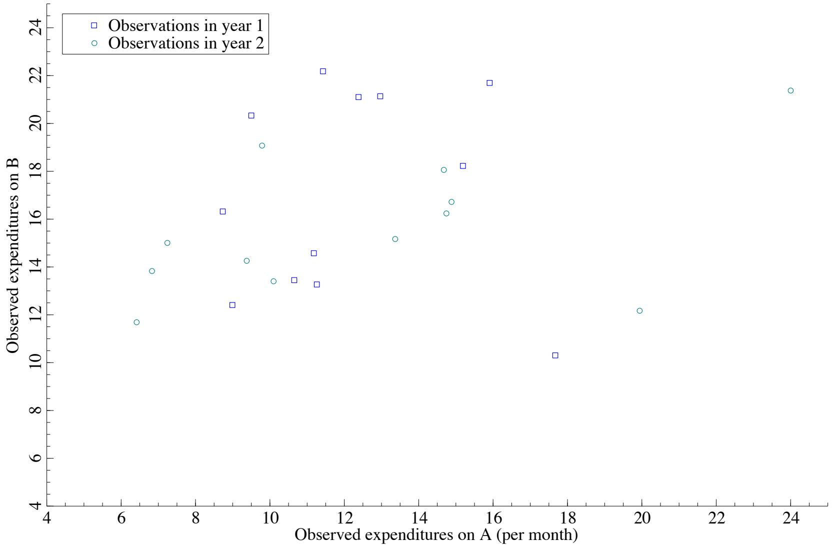
The fit of a model is often based on the implied likelihood of the observed behavior of individuals. According to the model, how likely is it that individuals behave the way they did? Many estimation procedures are directly or indirectly based on this idea. If the model contains heterogeneity, it is usually not easy to calculate this likelihood, and computer simulation methods are necessary. The recent past has shown a substantial development of such simulation methods. Estimation methods can roughly be classified in two groups: frequentist and Bayesian estimation. The difference is in the fundamental view on model parameters. In the frequentist view, model parameters are unknown, but fixed quantities. The Bayesian view focuses on uncertainty. Here it is acknowledged that we will never be sure. A whole range of different relations between variables may be true, though some relations will be more likely than others. Technically speaking, parameters are considered to be random variables, and the uncertainty in the parameters can be represented by probability density functions.

The Bayesian paradigm has become very popular and promising in recent years. It allows us to make individual-level inferences even if we only have limited data for a particular individual and we cannot obtain very precise estimates at the individual level. This technique allows us to personalize the results even if little information is available. Let me give you an example.

An example of Bayesian inference

Suppose we want to model the purchases of individuals over time in two related product categories. Assume that we have observations on the total monthly expenditure for each individual over a two-year period. In our earlier terminology, we could call this a three-dimensional panel, with individuals, time, and product categories as dimensions. Individuals are heterogeneous: they differ in their regular spending levels. Some individuals tend to buy more than others. For one specific individual, the data may look like the squares and circles in Figure 4. Each month in the two-year period is represented by a square (year 1) or circle (year 2). For example, the lower left circle indicates that in a particular month this individual spent about €6 on category A and almost €12 on category B. Of course, the expenditure differs across all months. How could we calculate the future value of this individual to the company? To answer this question, we would need to predict future spending.

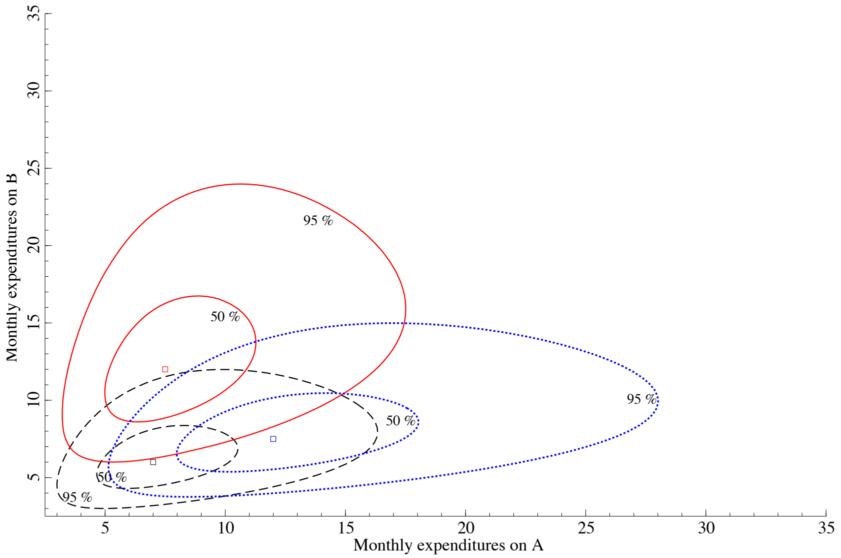
Figure 4: Observed expenditures of an individual on two product categories



With 24 observations, we could forecast future spending by calculating the average sales in both categories. This may be a reasonably accurate estimate. But how accurate is it and what would we do if we had fewer observations, or more categories to study? In other words, what would we do if the data were sparse? In this case, it would not be feasible to use simple averages. This is where Bayesian econometrics comes in. Even after observing the data, we are not sure about this individual. The observed expenditures give us some information about this individual, but they also contain noise. For example, how likely is it that the relatively extreme observation in the top right corner of Figure 4 will occur again in the future? To answer questions like these, we need to retrieve the underlying structure from the observations.

The first step is to specify a probabilistic model that describes how likely certain spending levels of an individual are. Technically speaking, we need to specify the distribution of the expenditures, where this distribution may depend on unknown quantities. A relatively simple model is obtained by assuming that the expenditures follow a joint log-normal distribution. In simple terms, this means that expenditures in a particular month are based on a constant (but unknown) baseline level which is multiplied by a noise component in each month. For example, in some months the individual buys 5% more than the baseline for A and 3% more for B, while in another month 7% less is spent on A and B. Finding the signal in the observations now becomes equivalent to trying to figure out the unknown baseline levels.

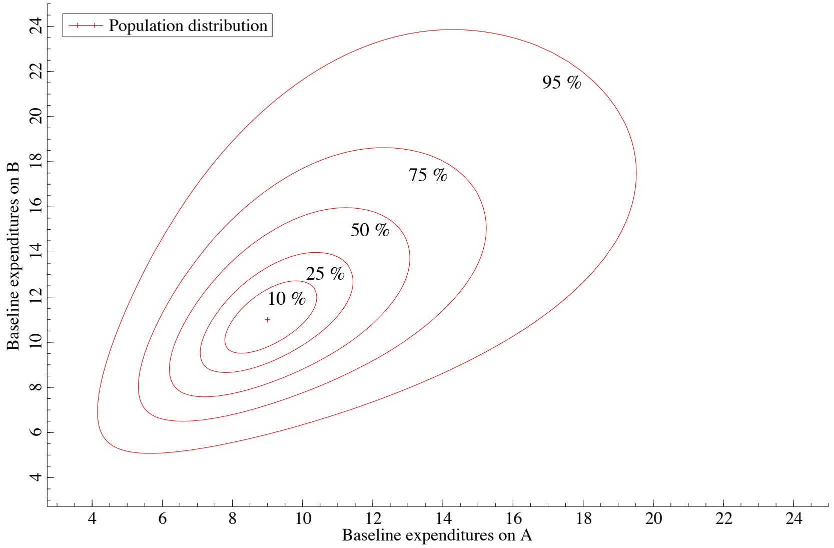
Figure 5: Modeling monthly expenditures: three potential distributions



This model is illustrated in Figure 5 where I represent three potential spending distributions for this individual using three different colors. Each distribution is represented by what is known as the 50% and 95% highest density regions. These regions should be interpreted as follows: for each distribution 50% of all monthly expenditure combinations are in the innermost region and 95% of the observations are in the larger region. For the black distribution, this means that 50% of the expenditures on A are roughly in the €5 to €10 interval. Under this distribution, spending more than €16 on A is unlikely. In this case, each distribution can be completely characterized by its center point, which represents the baseline expenditures. In the figure, these are indicated by squares. Estimating the baseline expenditures now becomes equivalent to finding the best distribution.

We should also take into account that we do not only have information on this individual, but also on many other individuals. Even if we do not have a lot of data on one individual, people are likely to be somewhat similar in their behavior. This information could also be used. Our knowledge on all other individuals can be summarized by the joint distribution of the baseline expenditures in the population. Such a distribution is graphically depicted in Figure 6, again using highest density regions.

Figure 6: Population distribution on baseline expenditures

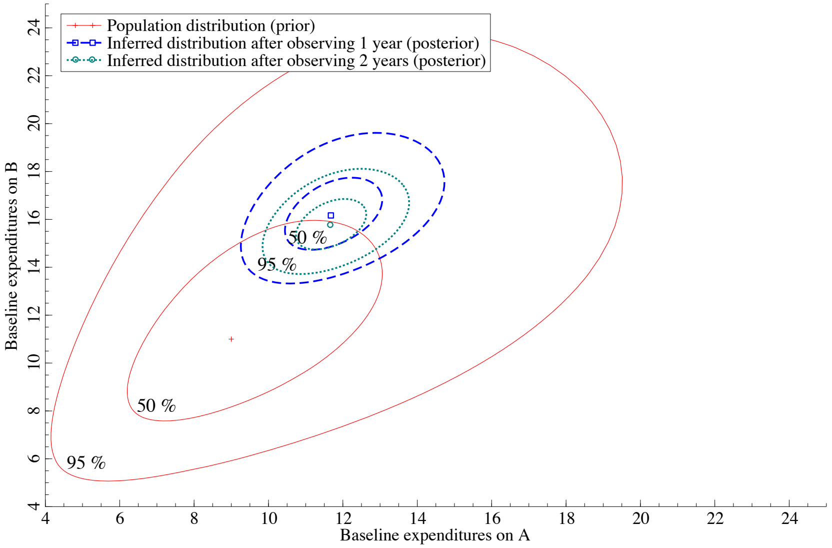


Note that this is a distribution on the baseline expenditures, that is, the possible center points in the previous figure. This summarizes everything that we know about the baseline expenditure of an individual before having seen any purchases by that individual. Note that this information is again specified in terms of probabilities. We do not know the baseline sales exactly, but we do have an idea about the likelihood of different possibilities.

Now assume that we have observed the purchases of the individual over the first year. These are represented by the green squares in Figure 4. The population distribution now no longer accurately reflects our knowledge on the baseline expenditures of the individual. For example, we know that, relative to the population, this individual has a high expenditure pattern. The baseline expenditures are therefore more likely to be in the upper right area of the distribution than in the lower left area in Figure 6.

In the Bayesian terminology, we should use the observed expenditures (data) to update the population distribution (prior) to reflect our current knowledge. Bayesian statistics tells us exactly how to do this. In the interest of time I will not go into details. The updated knowledge can again be represented in the form of a distribution. This distribution is known as the posterior distribution. Figure 7 gives the posterior distribution after incorporating the first year in long-dashed lines. For clarity, I only show the 50% and 95% highest density regions in this figure.

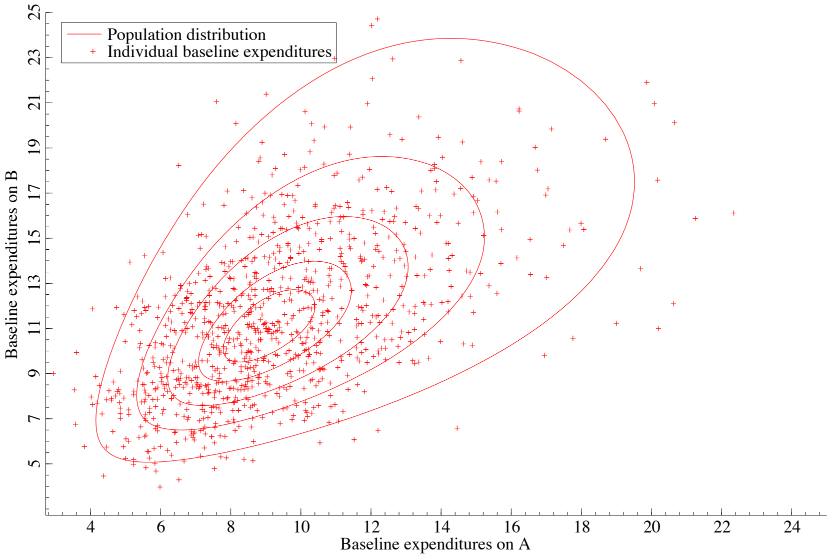
Figure 7: Inferred distribution on baseline expenditures using 1 or 2 years of observations



From the graph, it is clear that we have already learned a lot using just twelve observations. The size of the 95% region has shrunk considerably. Also note that the location of the distribution has moved. We are now relatively confident that this individual is a high, but not extreme spender. The regions in the dotted lines indicate the posterior distribution after having observed two years. The additional year of data enhances our knowledge; the size of the highest density regions again shrink and the distribution moves a bit as well. As more and more information comes in, we will obtain a more precise view on this individual.

I am sure that some of you will have noticed that I cheated a bit in this example. If the baseline expenditures of an individual are unknown, how can we ever know the population distribution? The answer is that we have to estimate this distribution as well. Let us briefly take a look at this problem. Suppose that we now know the baseline expenditures of a set of individuals. Based on this set, we can estimate the population distribution relatively easily. This is illustrated in Figure 8. Each small plus-symbol represents the baseline expenditures of a single individual. In total, 1000 individuals are represented in the figure. The distribution should now be chosen such that it matches these individuals as well as possible. For example, roughly 50 individuals should be outside of the 95% highest density region. This is a relatively simple estimation task.

Figure 8: Estimating the population distribution from individual baseline expenditures



However, this step is not feasible either as we do not know the individual expenditures for sure. In practice, we have to combine both steps and perform both sides of the analysis simultaneously. This methodology allows us to learn about the population distribution and about the individual behavior at the same time.

Real life applications

The example I have just discussed was intentionally kept simple. In practice, many more things need to be modeled and estimated. In the model we only considered the baseline sales to be unknown. In general, we also need to estimate the magnitude of the noise. Moreover, we may want to include more complicated dependencies between the product categories, and between the expenditures and marketing instruments. Moreover, it is quite likely that the expenditures in one month have an impact on the expenditures in the next month. Finally, demographic variables may be a part of our prior knowledge.

By extending the model, the necessary computations also become more complicated. For the above example, all calculations are relatively easy. However, even when the model is extended only slightly, we have to rely on computer simulation methods. These methods are well developed, but come at the expense of computation time. Although the speed of computers has increased at an exponential rate, we have been able to keep up with this trend in the complexity of the problems that we study. This literally means that days or weeks of computer time are sometimes needed to perform a particular estimation task. With the increasing size of datasets and more complicated models in mind, the current methods may soon prove to be infeasible. I see two main solutions here which I will discuss in turn.

Computer science solutions

First, we need to make more efficient use of current computing power. The developments in computer chips are no longer in faster chips, but in more chips that operate simultaneously. In the recent past, considerable progress has been made by programming languages that allow for parallel computing. This means that different computations can be done at the same time by multiple processors within one computer, but also across a number of computers in a network. More research is needed to make full use of these possibilities in econometrics, as it often requires an approach that is fundamentally different from what is currently the standard.[4] Next to the development of techniques, making progress in this area also requires access to the adequate hardware.

Approximations to the solution or to the problem

Another insight that may change our treatment of estimation problems is that some problems are just too large to solve to optimality. We need to accept that in these cases we can only obtain an approximation to the solution. However, the challenge of how to arrive at a good approximation remains. This route requires more insight into the properties of approximate solutions to be

[4] Recent contributions are made by Durham and Geweke (working paper, 2012).

sure that they are good enough.[5] In a recent paper with former PhD student Tim Salimans, we have suggested such an approximation method for a particular class of models.[6]

As an alternative to approximating the solution to the actual problem, we could consider approximating the problem and finding the exact solution to the approximate problem. This perhaps uses the best of both worlds. We specify a complete econometric model so that we have an accurate representation of behavior, but ignore some details in the estimation that are not relevant for obtaining a good solution. Often dependencies between certain variables in a model can be ignored. We should study such approximations in detail and assess which are good and which are not to be used. Developing further insights on finding good, usable solutions is a promising area for future research. These computational challenges are too important for econometrics to leave to computer science and machine learning.

[5] See for example Rue et al (Journal of the Royal Statistical Society B, 2009).
[6] Salimans and Fok (working paper, 2013).

6. Education

Next to research, education is another important part of our academic life. The Erasmus University Rotterdam can be seen as the birthplace of econometrics, and we can be proud of having the largest educational program in econometrics in the world. I see econometrics as the ideal choice for students who wish to learn to extract value from data that is available in a wide variety of companies. This view is confirmed each and every year by the fact that almost all of our students manage to find a job, or have at least received a job offer close to the date of their graduation. Already during their thesis work, many students make a significant contribution to improving business processes. This again highlights that econometric skills are very much needed in practice. In fact, they seem to become more and more wanted in business. Some have even named “data analyst”[7] or “data scientist”[8] as the sexiest job of the 21st century.[9]

However, we need to think about the future of our program. We also need to stay ahead of competition. We compete with other econometrics programs in the Netherlands and other countries, but in the future competition from other domains will increase. As I discussed already, additional skills are needed to contribute to current issues. We need to ensure that our students remain up to date with the state-of-the-art techniques, and can deal with the massive datasets that become available. They will require specific skills to work with less well-structured micro data, such as, multi-dimensional panels, clickstream data, data in the form of written text, and choices out of very large choice sets. This requires a somewhat different view on econometrics. We need to go beyond the classical methods and incorporate recent developments in this area into our program.

All of this is very possible in Rotterdam, we have always focused on applying econometrics to practically relevant problem in various domains. Venturing into a new domain fits perfectly into this mindset. All we need to do is to keep an open mind and, I am sure that we will continue to stay ahead of the competition.

[7] CNBC, June 2013 (http://www.cnbc.com/id/100792215)
[8] Davenport and Patil (Harvard Business Review, 2012)
[9] However, the word econometrics is not used in either publication; the words big data are.

7. Summary

Although Big Data has features of a hype, it creates many opportunities and challenges for the development and application of econometric techniques. It creates opportunities for companies to learn more about their customers' needs and their competitors' behavior. It also provides opportunities for academics as it requires new models and techniques that can deal with large quantities and new types of data. One of the opportunities is to increase the collaboration and information sharing between the academic and corporate communities. As a result, companies will have access to the latest methods and academic developments, and researchers will have access to business data to test their developed methods. Finally, this development creates opportunities to enhance our educational program. If we integrate these techniques and applications into our program, we will further increase the contribution that our students can make in business.

Dankwoord

As we approach the end of this lecture it is time for me to thank a number of people. I will do this in Dutch. Aan het einde van deze rede wil ik een aantal mensen bedanken.

Geachte leden van het College van Bestuur van de Erasmus Universiteit, geachte Decaan, De Vereniging Trustfonds,

Ik vind het nog steeds een voorrecht om aan de Erasmus Universiteit te mogen werken. Ik heb deze universiteit in veel verschillende rollen mogen meemaken: als student Econometrie, als promovendus, en als wetenschappelijk medewerker. In elk van deze rollen ben ik trots geweest op mijn universiteit. Ik ben dankbaar voor het vertrouwen dat de universiteit in mij uitspreekt met mijn benoeming tot hoogleraar.

Hooggeleerde Franses, beste Philip Hans,

Hoewel ik je net onder je titel van decaan al heb bedankt, verdien je zeker een aparte vermelding. In alle fasen van mijn wetenschappelijke ontwikkeling heb jij een belangrijke invloed gehad. In mijn tijd als student heb je een grote rol gespeeld in mijn beslissing om het promotietraject in te gaan. In de vier jaar erna heb je als promotor onmisbare sturing gegeven aan mijn onderzoek. Door jouw enthousiaste begeleiding was er nooit een gebrek aan nieuwe ideeën. We hebben samen de toepassing van econometrie in de marketing verder ontwikkeld. Ik ben je dankbaar voor het vertrouwen dat je in mij stelt, en ik hoop in de toekomst gezamenlijk nog aan vele onderzoeken te werken.

Hooggeleerde Paap, beste Richard,

Als co-promotor heb je een onmisbare invloed gehad op mijn promotie-onderzoek en die invloed is merkbaar in bijna al mijn latere onderzoeken. Ik heb veel van je geleerd. Ik herinner mij vooral onze soms verhitte discussies over econometrische vraagstukken. Ik geloof dat er wel eens collega's dachten dat we ruzie hadden. Uiteindelijk denk ik dat we allebei veel baat hebben gehad bij deze discussies. Ik hoop in de toekomst verder met je samen te werken aan de ontwikkeling van nieuwe technieken.

Beste co-auteurs,

Ieder van jullie heeft een unieke inbreng in de projecten die we gezamenlijk doen. Sommigen brengen specifieke technische kennis in, of kennis van de marketing theorie. Anderen leveren zeer belangrijke informatie over de context van een vraagstuk. Sommigen hebben zelfs als extra taak de uitdaging om het project bovenop mijn stapel met nog-te-doen te krijgen. Ik waardeer de samenwerking met jullie zeer. Zonder jullie zou ik het niet allemaal kunnen.

Beste promovendi en ex-promovendi,
Beste Carlos, Yuri, Wei, Tim, Bruno en Aiste,

Het is een groot plezier om getalenteerde mensen te mogen begeleiden. Jullie zorgen er mede voor dat ik nieuwe dingen blijf leren. Laten we proberen om samen de concurrentie voor te blijven door vernieuwend onderzoek te blijven doen.

Beste vertegenwoordigers van het bedrijfsleven,

Zoals ik in mijn rede al heb genoemd is het voor wetenschappers zoals ik van groot belang om te kunnen samenwerken met het bedrijfsleven. Ik ben dankbaar voor de mogelijkheden die jullie mij, onze promovendi en onze studenten bieden en ik hoop dat nog vele projecten zullen volgen die voor beide partijen toegevoegde waarde hebben.

Beste collega's van de capaciteitsgroep Econometrie,

Ik dank jullie allen voor de prettige werkomgeving. Het sociale aspect van onze groep maakt het een plezier om naar de universiteit te komen.

Lieve ouders,

Ik wil jullie bedanken voor de mogelijkheden die jullie mij hebben geboden. Jullie hebben mij altijd gesteund in de keuzes die ik heb gemaakt. Daarnaast hebben jullie mij geleerd om kritisch te zijn. Dit is een eigenschap waar ik als onderzoeker veel aan heb.

Sofie en Julia, mijn lieve dochters,

Mede door jullie is het elke werkdag een plezier om weer naar huis te gaan. Het is geweldig om jullie te zien opgroeien. Jullie zijn gewoon onmisbaar.

Mijn allerliefste Sonja,

Jij bent het belangrijkste in mijn leven, jij zorgt ervoor dat alles de moeite waard is. Ik hoop samen met jou oud te worden.

Ik heb gezegd, I have said.

Erasmus Research Institute of Management - ERIM

Inaugural Addresses Research in Management Series
ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

Balk, B.M., *The residual: On monitoring and Benchmarking Firms, Industries and Economies with respect to Productivity,* 9 November 2001, EIA-07-MKT, ISBN 90-5892-018-6, http://hdl.handle.net/1765/300

Benink, H.A., *Financial Regulation; Emerging from the Shadows,* 15 June 2001, EIA-02-ORG, ISBN 90-5892-007-0, http://hdl.handle.net/1765/339

Bleichrodt, H., *The Value of Health*, 19 September 2008, EIA-2008-36-MKT, ISBN/EAN 978-90-5892-196-3, http://hdl.handle.net/1765/13282

Boons, A.N.A.M., *Nieuwe Ronde, Nieuwe Kansen: Ontwikkeling in Management Accounting & Control*, 29 September 2006, EIA-2006-029-F&A, ISBN 90-5892-126-3, http://hdl.handle.net/1765/8057

Brounen, D., *The Boom and Gloom of Real Estate Markets*, 12 December 2008, EIA-2008-035-F&A, ISBN/EAN 978-90-5892-194-9, http://hdl.handle.net/1765/14001

Bruggen, G.H. van, *Marketing Informatie en besluitvorming: een inter-organisationeel perspectief*, 12 October 2001, EIA-06-MKT, ISBN 90-5892-016-X, http://hdl.handle.net/1765/341

Commandeur, H.R., *De betekenis van marktstructuren voor de scope van de onderneming*. 05 June 2003, EIA-022-MKT, ISBN 90-5892-046-1, http://hdl.handle.net/1765/427

Dale, B.G., *Quality Management Research: Standing the Test of Time; Richardson, R., Performance Related Pay – Another Management Fad?*; Wright, D.M., *From Downsize to Enterprise: Management Buyouts and Restructuring Industry*. Triple inaugural address for the Rotating Chair for Research in Organisation and Management. March 28, 2001, EIA-01-ORG, ISBN 90-5892-006-2, http://hdl.handle.net/1765/338

De Cremer, D., *On Understanding the Human Nature of Good and Bad Behavior in Business: A Behavioral Ethics Approach*, 23 October 2009, ISBN 978-90-5892-223-6, http://hdl.handle.net/1765/17694

Dekimpe, M.G., *Veranderende datasets binnen de marketing: puur zegen of bron van frustratie?*, 7 March 2003, EIA-17-MKT, ISBN 90-5892-038-0, http://hdl.handle.net/1765/342

Dijk, D.J.C. van, *"Goed nieuws is geen nieuws"*, 15 November 2007, EIA-2007-031-F&A, ISBN 90-5892-157-4, http://hdl.handle.net/1765/10857

Dissel, H.G. van, *"Nut en nog eens nut" Over retoriek, mythes en rituelen in informatiesysteemonderzoek*, 15 February 2002, EIA-08-LIS, ISBN 90-5892-018-6, http://hdl.handle.net/1765/301

Donkers, A.C.D., *"The Customer Cannot Choose"*, Apruil 12, 2013, ISBN 978-90-5892-334-9, http://hdl.handle.net/1765/39716

Dul, J., *"De mens is de maat van alle dingen" Over mensgericht ontwerpen van producten en processen.*, 23 May 2003, EIA-19-LIS, ISBN 90-5892-038-X, http://hdl.handle.net/1765/348

Ende, J. van den, *Organising Innovation*, 18 September 2008, EIA-2008-034-ORG, ISBN 978-90-5892-189-5, http://hdl.handle.net/1765/13898

Groenen, P.J.F., *Dynamische Meerdimensionele Schaling: Statistiek Op De Kaart*, 31 March 2003, EIA-15-MKT, ISBN 90-5892-035-6, http://hdl.handle.net/1765/304

Hartog, D.N. den, *Leadership as a source of inspiration,* 5 October 2001, EIA-05-ORG, ISBN 90-5892-015-1, http://hdl.handle.net/1765/285

Heck, E. van, *Waarde en Winnaar; over het ontwerpen van electronische veilingen*, 28 June 2002, EIA-10-LIS, ISBN 90-5892-027-5, http://hdl.handle.net/1765/346

Heugens, Pursey P.M.A.R., *Organization Theory: Bright Prospects for a Permanently Failing Field*, 12 September 2008, EIA-2007-032 ORG, ISBN/EAN 978-90-5892-175-8, http://hdl.handle.net/1765/13129

Jansen, J.J.P., *Corporate Entrepreneurship: Sensing and Seizing Opportunities for a Prosperous Research Agenda*, April 14, 2011, ISBN 978-90-5892-276-2, http://hdl.handle.net/1765/22999

Jong, A. de, *De Ratio van Corporate Governance*, 6 October 2006, EIA-2006-028-F&A, ISBN 90-5892-128-X, http://hdl.handle.net/1765/8046

Jong, M. de, *New Survey Methods: Tools to Dig for Gold*, May 31, 2013, ISBN 978-90-5892-337-7, http://hdl.handle.net/1765/40379

Kaptein, M., *De Open Onderneming, Een bedrijfsethisch vraagstuk*, and Wempe, J., Een maatschappelijk vraagstuk, Double inaugural address, 31 March 2003, EIA-16-ORG, ISBN 90-5892-037-2, http://hdl.handle.net/1765/305

Knippenberg, D.L. van, *Understanding Diversity*, 12 October 2007, EIA-2007-030-ORG, ISBN 90-5892-149-9, http://hdl.handle.net/1765/10595

Kroon, L.G., *Opsporen van sneller en beter. Modelling through*, 21 September 2001, EIA-03-LIS, ISBN 90-5892-010-0, http://hdl.handle.net/1765/340

Maas, Victor S., *De controller als choice architect,* October 5, 2012, ISBN 90-5892-314-1, http://hdl.handle.net/1765/37373

Magala, S.J., *East, West, Best: Cross cultural encounters and measures,* 28 September 2001, EIA-04-ORG, ISBN 90-5892-013-5, http://hdl.handle.net/1765/284

Meijs, L.C.P.M., *The resilient society: On volunteering, civil society and corporate community involvement in transition,* 17 September 2004, EIA-2004-024-ORG, ISBN 90-5892-000-3, http://hdl.handle.net/1765/1908

Meijs, L.C.P.M., *Reinventing Strategic Philanthropy: the sustainable organization of voluntary action for impact,* February 19, 2010, ISBN 90-5892-230-4, http://hdl.handle.net/1765/17833

Oosterhout, J., *Het disciplineringsmodel voorbij; over autoriteit en legitimiteit in Corporate Governance,* 12 September 2008, EIA-2007-033-ORG, ISBN/EAN 978-90-5892-183-3, http://hdl.handle.net/1765/13229

Osselaer, S.M.J. van, *Of Rats and Brands: A Learning-and-Memory Perspective on Consumer Decisions,* 29 October 2004, EIA-2003-023-MKT, ISBN 90-5892-074-7, http://hdl.handle.net/1765/1794

Pau, L-F., *The Business Challenges in Communicating, Mobile or Otherwise*, 31 March 2003, EIA-14-LIS, ISBN 90-5892-034-8, http://hdl.handle.net/1765/303

Peccei, R., *Human Resource Management And The Search For The Happy Workplace*. January 15, 2004, EIA-021-ORG, ISBN 90-5892-059-3, http://hdl.handle.net/1765/1108

Peek, E., *The Value of Accounting*, October 21, 2011, ISBN 978-90-5892-301-1, http://hdl.handle.net/1765/32937

Pelsser, A.A.J., *Risico en rendement in balans voor verzekeraars*. May 2, 2003, EIA-18-F&A, ISBN 90-5892-041-0, http://hdl.handle.net/1765/872

Pennings, E., *Does contract complexity limit oppoortunities? Vertical organization and flexibility.,* September 17, 2010, ISBN 978-90-5892-255-7, http://hdl.handle.net/1765/20457

Pronk, M., *Financial Accounting, te praktisch voor theorie en te theoretisch voor de praktijk?, June 29,* 2012, ISBN 978-90-5892-312-7, http://hdl.handle.net/1765/1

Rodrigues, Suzana B., *Towards a New Agenda for the Study of Business Internationalization: Integrating Markets, Institutions and Politics, June 17,* 2010, ISBN 978-90-5892-246-5, http://hdl.handle.net/1765/20068

Roosenboom, P.G.J., *On the real effects of private equity*, 4 September 2009, ISBN 90-5892-221-2, http://hdl.handle.net/1765/16710

Rotmans, J., *Societal Innovation: between dream and reality lies complexity*, June 3, 2005, EIA-2005-026-ORG, ISBN 90-5892-105-0, http://hdl.handle.net/1765/7293

Smidts, A., *Kijken in het brein, Over de mogelijkheden van neuromarketing*, 25 October 2002, EIA-12-MKT, ISBN 90-5892-036-4, http://hdl.handle.net/1765/308

Smit, H.T.J., *The Economics of Private Equity*, 31 March 2003, EIA-13-LIS, ISBN 90-5892-033-X, http://hdl.handle.net/1765/302

Stremersch, S., *Op zoek naar een publiek....*, April 15, 2005, EIA-2005-025-MKT, ISBN 90-5892-084-4, http://hdl.handle.net/1765/1945

Verbeek, M., *Onweerlegbaar bewijs? Over het belang en de waarde van empirisch onderzoek voor financierings- en beleggingsvraagstukken,* 21 June 2002, EIA-09-F&A, ISBN 90-5892-026-7, http://hdl.handle.net/1765/343

Waarts, E., *Competition: an inspirational marketing tool,* 12 March 2004, EIA-2003-022-MKT, ISBN 90-5892-068-2, http://ep.eur.nl/handle/1765/1519

Wagelmans, A.P.M., *Moeilijk Doen Als Het Ook Makkelijk Kan, Over het nut van grondige wiskundige analyse van beslissingsproblemen,* 20 September 2002, EIA-11-LIS, ISBN 90-5892-032-1, http://hdl.handle.net/1765/309

Whiteman, G., *Making Sense of Climate Change: How to Avoid the Next Big Flood.* April 1, 2011, ISBN 90-5892-275-5, http://hdl.handle.net/1765/1

Wynstra, J.Y.F., *Inkoop, Leveranciers en Innovatie: van VOC tot Space Shuttle,* February 17 2006, EIA-2006-027-LIS, ISBN 90-5892-109-3, http://hdl.handle.net/1765/7439

Yip, G.S., *Managing Global Customers,* 19 June 2009, EIA-2009-038-STR, ISBN 90-5892-213-7, http://hdl.handle.net/1765/15827