Statistical modeling and business expertise, or where is the truth?

Igor Mandel
Advanced Marketing Models, 37 Overlook Terrace, Suite 3F
New York, NY 10033, igor.mandel@AMModelsInc.com

Abstract

Statistical model results may strongly contradict domain knowledge or expectation, which generates many problems considered in this article. In a nutshell, who is right: a statistician stating that advertising does not work (based on the model), or a marketing officer stating the opposite (based on experience or "gut feeling")? The answer is not as obvious as it seems. All statisticians face this dilemma almost every time they present results to a "non-statistical" audience. The combinations of subjective and objective, derived and desirable, free and forced aspects of modeling could be complicated, often posing difficult scientific and moral challenges. The problem is considered from practical and theoretical points of view. A brief review of important statistical concepts is given to position a statistician's role in dialog with a client. Different types of statistical and non-statistical uncertainties, which are usually not holistically considered in a modeling, are systemized.

Key words: statistical modeling, marketing mix modeling, regression, causality analysis, yield analysis, business decisions.

Introduction

Statistics is the science of mass phenomena under the condition of uncertainty (it is one out of many possible definitions, but it bears most of the important features). Business is the art of "getting things done", presumably pursuing one's profit or the "general interests of society" (if the term "business" is applied to governmental structures, which is often the case). Statistical analysis or modeling is performed by qualified specialists and is based on the developed, always-changing methodology. Decision makers who execute business solutions are not specially qualified; rather they gain status through right of birth, talent, hard work, or luck. Even this simple difference sets the grounds for many possible (and real) odds between two ways of thinking and acting. More and more business decisions in recent years have been made based on some statistical knowledge, just as they always have been (maybe unconsciously) throughout history. This knowledge is not necessarily "scientific"; it's just "business intuition" based on some key parameters. As a hero in a novel (Murukami 1999), a successful owner of a restaurant chain explains: before opening a new restaurant he goes to some preliminarily picked corner in Tokyo and observes people running there day by day. He doesn't take any notes; he just tries to imagine what will be changed at this corner if he opens a restaurant there. In couple of weeks of such observations he makes a decision. He chooses to either purchase a spot for a new restaurant or to go to another corner. Is it a "statistical" decision? Yes indeed, but not the only one.

Imagine that an educated statistician always follows this businessman and collects data about people running around, their demographics and so on. At the end of the task he proposes the recommendations to his client, and it happens that they are the opposite of the boss's intention. What is the expected outcome? Most likely, a businessman will follow his gut rather than any scientific evidence. The first question is why? And the next question is – what is the statistician's reaction to that decision? This article is about possible answers to questions like these. Its *subject is the relationship between statistician and client*. A client is determined either an employing company for a statistician or an outsider's company (further a chain of stores is considered as a client and an object of modeling). Typically a statistician believes in domain knowledge of a client, whereas the client believes in statistical techniques – otherwise a contact (and contract) would be impossible. They both are right and wrong. The statistician is right to say that domain knowledge is lying somewhere on the client side, but he may be wrong in locating it there. A businessman is right trusting statistical techniques because there is nothing else he can do (to say that "science does not work" is taking a step back and cannot be defended on an upper level). But he is wrong

because, with rare exceptions, statistics cannot guarantee his desired outcome. Specification of these statements is what this article is about.

Out of many business problems and statistical settings for their solutions I have selected, maybe, the most important and popular one, which is the finding dependence between business outcome (sales, profit, and so on) and affecting variables, based on observational data. And while conclusions of the article may be applied to other types of problems and data (say, experimental design), this assumption shows up clearly in some of the further considerations.

The article is structured as follows. In the first part I consider different business ambiguities and uncertainties, which directly or indirectly affect modeling. In the second part I try to summarize statistical uncertainties in such a way that it goes beyond just a list of typical problems and also touches quite fundamental issues about the nature of statistical and other models. The conclusion synthesizes the two parts and draws the general picture of the real environment in which any business modeling takes place; it also proposes some further steps.

1. Business uncertainties

A statistician reasonably expects formulation of the modeling problem from the client, but surprisingly often discovers that it is either non-existent, vague, unmeasured, or a combination of the three. Typical business uncertainties in relation to modeling could be classified as follows.

1.1. Model's goal

The goal should be measured in a precise and indisputable way in the form of a statistical variable, but even this "simplest" part of the modeling process is not always clear. For example, the original impulse of the client was to model sales as a function of many possible influencing factors; at least this is what he enthusiastically said in the first statistician-client ("kick-off") meeting. Then, on a more concrete level, the following questions are raised:

a) weekly, daily, monthly or quarterly sales are of real interest;

b) sales of the entire chain, regions, or individual stores;

c) what is the sales measure? Total revenue for the given period of time as registered at checkout or as finally booked (could be just once a week)? With or without returns? With or without coupon sales?

d) is it really sales the client is interested in, or maybe profit? Or maybe ultimately, the client's share price on the stock market is the goal?

e) does the client care about short run (1-3 years) or long run benefits from the model?

Questions like these are very typical. While working with one large financial institution, I waited about three weeks for their definition of revenue - in fact, they didn't have a definition ready at their fingertips. In the end, of course, a definition was provided (no one can start without that). However, the client's internal vagueness remains and may easily bear its fruits in a later stage of modeling and ultimately undermine the best model. It cannot be avoided or predicted, since it has an intrinsically personal nature (see also 1.4). The only way to reduce uncertainty is through the ultimate decision maker choosing modeling as a course of action (like modeling a privately owned company under direct supervision of an owner). But even in this case one still should account for the variability of the owner's intentions over time. As a result of this uncertainty, a modeler is always at risk of being accused of building a poor model simply because the client's goal has consciously or unconsciously changed during the course of modeling. I once had a client panic when sales started to drop below the respective weekly sales level of the previous year (relying on the model "to fix" this trend). But we modeled weekly sales, not their year old increments!

It is important to note that we are not talking about well known problems of multi-criteria optimization in decision making, which tries to solve problems of conflicting interests when they are measured. We are talking about internal uncertainty - the client's inability to formulate the goal in precise terms. It causes the necessity for statisticians to work with surrogates of the real goals, which might have vital consequences for the model's implementation.

1.2. Data for modeling

When a goal is determined, the statistician wants data for modeling. Usually, a general intention is to collect as much information as possible, both granular (stores are more preferable than regions, weeks than years, and so on) and comprehensive (as much "undependable" variables as possible). The apotheosis of that intention is data mining practice, where up to terabytes of information are gathered to answer basically the same question about "dependence" as in the "normal" model. Everything said about goal uncertainty is to be applied here, but quadrupled. Usually the level of client's knowledge and assurance about "factors" is comparable to that of a statistician, i.e., close to zero (except very general things, which are true because they are trivial).

When he deals with a technical person who is usually responsible for data collection they at best may provide him with perfect data, but without a sense of the data's meaning. Some indicators were defined three years ago but their meanings were lost. A very good guy, who quit the company just last December, was responsible for those variables and didn't leave any notes. For some variables the method of calculation was changed three times during the year and values were not readjusted. The client is not one intentional entity, but a complicated system of different departments with thousands of personal interrelations. When a client's goal is formulated and supported by data, it should always be understood within a narrow frame of a concrete authorized person (department) working with a statistician. There is always a chance that replacing the person or restructuring the company would change the whole model (see also 1.4).

For example, what may happen if a model of sales is needed, but a statistician works with the marketing department? He may naively think that price is an important factor of sales in natural units. Well, you are right, but price is not registered in our department. Can we get it from another one? Unfortunately, no, because the two computer accounting systems are incompatible, or the two department bosses do not talk to each other, or data collection is too complicated. So, do we model sales without price data? Yes. Or should we use a substitute for price which is "sales divided by number of buyers", even realizing it's a mixture of real price change and structure of purchase? It's a good idea, too. The main principle here is that a statistician usually collects not what he wants or what the theory tells him, but what is easily available. That is why hundreds of factors (not 3-5) are gathered in pure empirical style with the hope of finding a pearl in that pile of ore. So ultimately, domain knowledge in that stage of modeling is transformed into a huge data set – i.e., a statistician was wrong in his hope of finding something deeper from the domain specialists.

All that has nothing to do with well known problems of data quality itself – data errors, omissions, sparseness, incorrect indicators, like those who may inspire one to use instrumental variables in econometrics (Echambadi et al. 2006), measurable biases, etc. It's pure business and organizational uncertainty, which is inevitable in modeling practice. The scale of that uncertainty is sometimes very impressive. In one project, for example, the client really cannot give an explanation why two ways of measurement of the most important indicator completely contradicted each other in many instances – and the model was ultimately built with that discrepancy in the data. In another example a client has realized just after the modeling that a dependent variable's values are about three times (!) bigger than the correct ones (it's after "final data checking" before the modeling). This last example is especially interesting, because it shows, first, that often the "client" does not have enough of a feel for its own data to catch such a rough own error on the fly; and second, that the model may work perfectly with the wrong data, still "explaining" them very well.

1.3. Domain knowledge quality

What kind of domain knowledge does a statistician hope to get from his client? Any phenomenon he is asked to model has a very complex nature. Any person he received an order or data from is responsible only for part of it (often a very small part). The entire picture is always out of sight, and this is a fundamental problem (exactly as with goal definition). For example, the goal of a client in a marketing department is to model the effectiveness of different marketing activities. However, one cannot do it in an isolated fashion because there is no variable as "result of adverting" to be explained. A statistician has an aggregated variable such as sales instead which doesn't allow him to immediately answer the limited advertising problem. From this point forward, his direct client is not his guide anymore (for example, the huge area of price, distribution and logistics factors is no longer the marketing department's responsibility). Is the

solution to go directly to the CEO? This is not a real option considering that the CEO does not usually work with models because his personal vision of the needs and factors could be completely different from that of the marketing department.

Domain knowledge, as any other, is in somebody's head, or very rarely, in documents. There are many "domains" within a company and they usually just weakly interact. There was a meeting that I attended with accounting and marketing department specialists of one of the largest retail chains in the country. The goal of the meeting was to figure out a way to calculate "current profits" by matching reported sales with accounted expenses. A modeling company initiated the meeting in an attempt to model the "correct" indicator (the profit), rather than sales. Unfortunately, however, it became clear that neither side was willing to consider the possibility for mutual conciliation – i.e., marketing people, who wanted a model, couldn't comply with accountants, who were completely indifferent to the model. The result of two hours of talking between the two departments was a complete failure – in spite of the obvious importance of that indicator for the company, not for just the modeling sake. Ultimately, the sales were modeled, which in fact was not what everybody wanted.

This kind of scenario is not atypical for large companies, and even if the problem is very local and, supposedly, available for understanding by only one person, the domain knowledge still remains a mystery. A while back, I analyzed some medical data and discussed the results with a very well qualified doctor (the client). I first asked him what kind of correlations he expected for some variables. Even though his answers were definitive, the correlation analysis showed that the doctor was wrong for most of his hypothesis. Who is correct? What if paired correlations were indeed misleading, because "real" relations depended on other predictors that were not considered? Should the statistician insist that he or she is right (see part 2) and ignore the doctor's experience? Or should the statistician reconsider a model if it does not fit from an expert's view?

1.4. Interests of peoples involved

Not only are the visions and knowledge of the people involved in modeling from the business side drastically different, but so are their interests, which I already had a chance to stress a couple of times earlier. There is no such thing as an "objective model of the company". Each manager is not only responsible for some part of the business but is also remunerated for specific indicators within that part. Thus, it is in his best interest to have a model that confirms those indicators and thus his achievements. In our example, the marketing manager would love to have a model confirming that marketing efforts do play a big role in sales. Let's say that if the model shows that 15% of sales are due to advertising, and these results are trusted by upper management, he or she has great leverage to ask for additional funding; the manager can experiment with different types of media to further elevate sales for the brand. However, if the model states that only 1% of the sales are due to advertising, what will the expected reaction from the management be? It would be a lot harder to ask not only for modeling continuation, but also for additional funding. The conclusions for both the modeler and the client's manager from these facts are more or less predictable…

We should add to the client's interests the same type of ***interests of the statisticians***, which has nothing to do with science. The statisticians are living people, working under all the same conflicts and controversies as businessmen do. If modeling is done ***internally within a company that depends on government,*** like the pharmaceutical industry, the statisticians are tied to the company's business objectives. These objectives (for example, to get approval from the FDA) may be far away from "scientific truth" (as a result, studies like notorious the "grounding" of smoking harmlessness for lung cancer may appear – see discussion in Pearl 2000 and remark of Joseph Kadane in Pearl 2003). If the statisticians work ***internally, but solve direct business problems*** (like data mining specialists in banks), they seem less biased, but are typically faced with the routine production of recommendations. These circumstances often do not leave room for implementation of better techniques or development. Plus, these statisticians report directly to the marketing (or other) department – where all the above listed problems persist. If the statisticians are in ***consulting company***, there are strong incentives to please the client rather than to create the best possible model. If the simplest model, which is not checked by the client, pays off, what is the motivation for the statistician (and the company) to use more complicated and expensive, though better techniques? Few modeling companies try to use competitive advantages by modernization of their tools, though it should be a main road in a future. And finally, if the statisticians are in ***academia***, they are absolutely free, but not free from grants (see everything listed above). For this or any other reason, about 1-2% of academic studies estimated as

intentionally fraudulent (Hand 2007) – what is the percent in business area than? But even if the freedom is real – another problem came up. A good friend of mine, a professor of statistics and author of several books, said to me once: "You know, I really envy you – you have access to real problems and real data every day, I can just dream about it." Well, his dream came true for too many, but there was a certain price. To summarize, a statistician who is equipped with perfect knowledge and freely selecting the best way to solve a properly formulated practical problem exists only in an ideal world, in a pure Platonic sense.

A fresh example of the interests involved in modeling is the ongoing story of ***global warming,*** which culminated with A. Gore receiving the Nobel Prize for his timely alarm to mankind. While billions of dollars are already spent for GW studies and propaganda, and the question about human activity as the originator of GW is "clear" for millions around the globe, other models tell us a completely different story, stressing that current warming is a part of a cyclical process in the Earth's history, which is caused by solar and cosmic ray activity and so on http://video.google.fr/videoplay?docid=-4123082535546754758. GW defenders (a big part of the respectable scientific community) are accused by much smaller, but also very respectful part in being paid for predetermined conclusions. These "deniers" are accused of being paid by the "oil industry" though http://www.durangobill.com/Swindle_Swindle.html. The truth has been lost in debates and has accrued pure political attributes, while in fact it's just enough to make a simple (in my uneducated view of climatology) model to answer at least one key question: is warming causing the rise in atmospheric CO2 or vice versa? However, any honest correct and independent answer to that simple question in fact doesn't matter – it will be buried again under mutual political accusations. So, modeling (and truth) takes a back seat when interests reign…

All these types of **business uncertainty** – ***goals, data, knowledge and interests*** – are deeply embodied in the modeling process, mutually interconnected and heavily influencing the results. However, they are not typically considered in a holistic and practical way. So how can statisticians deal with this situation?

2. Statistical uncertainties

2.1. Ideal statistical problem and modeling reality

Let's assume that all the mentioned obstacles do not exist; the statistician has transparently determined the goal; has access to data of any type, supported by extremely deep domain knowledge; and has no more interests than to make an ideal model. Is he/she able to do it? The answer is, for the majority of cases, no. There are many reasons for this, but before we go into that, let's consider examples when the answer is yes.

If the goal is to make an automated system recognizing handwriting of digits (i.e., zip codes on letters), a model could be extremely successful, with a stable error rate less than 1% (Hastie, Tibshirani, and Friedman 2001, p.362-366). There are several factors contributing to it:

a) The volume of data used for training is big and easily updated.

b) The data collected are complete in a sense that practically all types of handwriting are presented with sufficient quantities.

c) The data is not dynamically sensitive: humans do not have a tendency to change their handwriting over the course of time (at least in the short term).

d) The predictors for correct recognition are limited and clear: black or white pixels within a predetermined two-dimensional space.

e) Predictors should not be interpreted in any deep substantial way; whatever combination of pixels works better is enough for the model.

f) There is a way to detect the errors because each case of recognition could be manually checked. The checkers, in turn, may have some error rate, but then another check could be performed, and this checking process is rapidly converging. It allows determining the algorithm's precision with high accuracy.

All 6 conditions, in my opinion, are key for a robust and successful statistical model (with the most questionable point e) – see below). Violations of any of them may not only result in the model's failure but even worse, make the character and size of failure unpredictable. If we go back to our marketing example, what similarities exist to those "ideal model" conditions?

a) *Volume of data and chances for permanent updating* are usually limited;

b) *Completeness* of data assumes ***full*** variability of predictors: variable values should run from a theoretical minimum to a theoretical maximum if it is quantitative or have all possible grades if it is

qualitative. This is a condition that is hard to achieve; practically it never holds in socio-economic applications – thus, surprises are always possible.

c) *The dynamic aspect* of socio-economic data is obvious, even if the concrete data set does not have a dynamic component. For that reason, any model, with or without a time scale, will not be applicable to some future period without going into the gray zone of "forecasting". Forecasting, however, as experience shows, is the most ungrateful occupation. This is, in fact, sufficient enough argument to constrain any modeling.

d) *Set of potential predictors* is actually infinite. It does not mean that for statistical modeling, more and more independent variables are required or that hundreds of predictors in data mining are not enough. It rather means that, unlike in digit recognition, the very character of the predictors, keeping aside their measurement, is usually unknown. In our example, the real reason why people would buy an advertised brand could be a function of such transient and immeasurable factors as type of content in ad, competitor's discounts, fads and so on. Try, for instance, to rationalize the recent boom around Google's shares (well, many will do it with big wisdom – a posteriori), and then use those found "factors" for another company or for the same Google, but years later. By the same token, the "Russian Google" Jandex.ru did not spend a dollar for advertising, yet got around 70% of the huge growing search engine market. Why? Word of mouth? Right, but why does it sometimes flourish and sometimes not? All those key questions are usually completely left out of modeling (see also 2.2).

e) *Interpretability*. Usually socio-economic models need a strong interpretation, but there can also be a strong contradiction with statistics and economics. There are many instances where (statistically) good models are corrected or rejected just because they are not interpreted in the proper manner. However, "proper" usually just means a way that is favorable for the client. For instance, the modeler will never include an advertising variable in a regression model if there is a strong negative coefficient (which occurs very often). Why? Because he or she does not believe that any ad works to hurt the product in any way. Could it be, theoretically, that a bad ad repels customers and, indeed, produces negative effect on sales? Yes, but it is very unlikely. So, statisticians here play the role of a censor who anticipates the client's expectations and avoids possible negative feedback. There are many examples like this; the interpretation stage of the model always introduces a bias. Thus it itself prevents the model from being "truthful".

f) *Errors* of statistical models are practically unavailable for testing. Different criteria of goodness of fit, either traditional (R squared like) or more comprehensive (with penalty functions like in SVM models), derived from bootstrapping or Monte Carlo – they all provide meaningful results only if applied to samples drawn from the same general population. But this is exactly the point: no one knows for sure about that – see b) and c). For example, any dynamic model could be tested on the basis of holdout samples. But those samples are just some parts of a time series. Could the same results be applied to the near future? To answer the question, one should know that the holdout belongs to the same "general population" which contains all given data. The same, but in less direct manner, is applied to cross-sectional models where procedures of error checking are much better developed. If one makes a model based on random sample from a database (like in data mining), and then applies it to a testing sample and finds similar results, it is considered evidence that the model is correct (Han and Kamber 2006). It is right to a certain extent. Mainly, when the model is applied to the outside world and does not work, the same old problem of "sample and population" will be raised again. All that creates a situation when the quality of the model could be as high as possible, but any expansions of these findings to other areas in time or space, which is ultimately a goal of any model, are untested and thus not convincing.

For those reasons, a model like the described handwritten digit recognition may be considered perfect and therefore very rare. Respectively, all others models, where violation of mentioned principles is obvious, are far from perfect but very frequent.

2.2. Fundamental statistical uncertainties

According to commonly accepted views, any modeling has two goals: first, to "imitate nature", i.e., capture some (presumably causal) relations, and second, "predict the future". Without going into lengthy discussions about these two aspects, let's just say that if the first goal is achieved (i.e., causal relations are captured well enough), the second should follow, but not vice versa. Despite the very common belief that the best criterion for the model is its ability to predict the future, it is not necessarily correct. As one of the

founders of modern sociology bitterly remarked half a century ago, criticizing "*numerology*" (the mildest term he used dubbing the formalistic use of mathematics in social sciences, together with, say, *quantophrenia*), the best predictions are usually done without any models, while a very good model fails to predict anything if circumstances change (Sorokin 1956). All that is very much true for statistical modeling now. Let's briefly consider here the quality of the ***predictive models*** having dependent and independent variables.

Technically, estimation of parameters in statistical models is always based on the solution of an optimization problem of a certain type - least squares, penalized criteria, maximum likelihood or something else. In fact, the criterion of optimization is oriented to some kind of approximation of the data observed and predicted by the model. If so, there is no accepted way within this paradigm to distinguish the causal and coincidental "factors". In the letter recognition example, the question of causality is unimportant and does not have any meaning; the purely pragmatic solution is acceptable (any pixels may be good or bad predictors depending on location). But the same pragmatism applied to socio-economical problems without appealing to real causes of a process, is at least inappropriate and may fight back at any moment. I cannot consider all the problems related to quality of statistical models but will briefly describe just two of the most popular and important classes of predictive models, *regression* and *causal analysis*. They represent **descriptive** and **causal** concepts within statistics and as such generate **fundamental uncertainty**, which any statistician may face, as will be shown below.

2.2.1. Regression analysis

Regression analysis and its multiple variations are by far the most widely utilized modeling approach. It is a strategic tool utilized by many of the world's top corporations for marketing mix models, data mining, and volume forecasting. However, as regression analysis's prominence grew, it was not without a variety of significant problems periodically highlighted in the literature (see recent accounts in Echambadi, Campbell and Agarwal 2006, who also touched problems with SEM and partial least squares approaches; Bemmaor and Franses 2005; Naik, Schultz and Srinivasan 2007); many of them were summarized in a monograph (Berk 2003). I'll focus here on some other important issues which have a controversial nature.

2.2.1.1. ***Multicollinearity (MC)***, on the one hand, is usually considered as one of the big problems of regression analysis (see early accounts in Theil 1971), but on the other hand, its interpretation and explanation leaves the field open for much confusion and many misconceptions. For example, some authors state that the signs of the regression coefficients are not altered due to multicollinearity (Franses and Heij 2002). Others write: "The only cure for near-multicollinearity is to reduce the number of explanatory variables by imposing restrictions on the parameters…If economic theory does not provide guidelines for parameter restrictions, the only other option is to delete the variables from the model that cause the problem." (Bierens 2005). Yet others say: "NEVER NEVER NEVER try to solve multicollinearity by 'throwing out' one (or more) of the intercorrelated independent variables" (Losh 2004). Many textbooks just do not consider that at all. I may assume that the confusion takes place because several aspects of the problem are mixed together. Without deep discussion, I will try just to separate them into parts.

Formally, whatever the reasons for MC are, it affects regression estimation in direct proportion to the level of correlation. There are concrete conditions (Lipovetsky and Conklin 2006a), under which two types of distortions – inflation of coefficients and alternation of regression and correlation coefficient signs - take place. But the presence of MC doesn't itself tell what remedy to use. It depends on the real causes of MC – and this point seems to be usually underestimated in the literature (at least, the above cited and other statements make one think that). One may distinguish five types of *causes of multicollinearity*.

a) Different variables reflect different aspects of the same process and in that sense are redundant (like many economical variables describing growth, etc.).

b) Different variables change in a synchronized way, but reflect different processes (like advertising, going simultaneously by many channels within one campaign).

c) Different variables influence each other (like education affects salary level, salary affects income, income affects education and purchasing power, and they all are highly correlated while the dependent variable is sales).

d) Different variables are correlated due to functional relations between them when they are measured (like price may have negative relation with sales in units not because people react to price change

by buying more, but because price is a ratio of sales in dollars and sales in units; or two out of three shares which summed up to 100% will be always correlated, and so on.)

e) Different variables are correlated just randomly (like birth rate and the number of storks on trees, once reported)

Scenario a) is known and investigated the most. The logical way to struggle with this type of MC is either to leave just one variable out of a bunch of highly correlated ones or to replace them with some kind of common factor or principal component. Both approaches are actively used, and both somehow get rid of redundancy in the data. There is a lot of literature about it, and I wouldn't go further. The problem, though, is that this approach sometimes claims to be a panacea for all cases of MC, as, for example, in Bierens 2005, cited above, or in numerous practical applications.

Scenario b) assumes that one can't get rid of other predictors, since they are not redundant. One has no choice but to estimate coefficients under conditions of high MC. Technically it means that matrix inversion become unstable due to the closeness of the determinant to zero; it results in arbitrary distribution of regression coefficient values between correlated variables. Without pretending to complete awareness, I may distinguish the following methodologically different ways of solving this problem.

1) The historically first solution was proposed in a general frame of *ill-posed problems* by A. Tikhonov in 1943 (Tikhonov and Arsenin 1977) and known as *regularization*. The idea was to add some regularization parameter to a diagonal of the matrix to be inverted (covariance matrix for regression). In the statistics community the term *ridge regression* for this approach is much more popular after its reinvention (in Hoerl and Kennard 1970). Ridge regression has one serious problem, noted in the very beginning: the value of the regularization parameter is in fact arbitrary and unknown; some additional considerations for its definition should be applied. Ridge regression looked so appealing to statisticians that in the first decade after its inception it was already possible to compare through numerical simulation ten different proposed analytical and iterative methods of parameter estimation (Gibbons 1981). As for now, one of the best seems to be the cross-validation approach (Golub, Heath, and Wahba 1979), where the optimal value is such that the leave-one-out validation error is minimal. Methodologically close to ridge regression is the *lasso* approach (Hastie, Friedman, and Tibshirani, 2001), where in fact the quadratic penalty function is replaced by the sum of absolute values. However, no convincing evidence exists about the uniqueness of any choice of regularization parameter, nor about typical size of bias, obtained with a given optimal value.

2) The problem was very early considered from a *Bayesian perspective* (Leamer 1973), where it was treated as a problem of selection of the correct priors. However, A. Leamer correctly noted that in that case one problem is in fact replaced by another – the difficulties in obtaining "correct" priors are of the same level as those for obtaining correct values of regression coefficients. It is especially true for multivariate situations, the only practically important cases, when any priors are very hard to justify. It could be shown though this that ridge regression estimates could be derived in Bayesian style, assuming that prior distribution of the coefficients is known (Hastie, Tibshirani and Friedman 2001, p.60; *I'm very grateful to I. Lipkovich for bringing that fact to my attention.*). This is a good illustration of how one type of uncertainty (about priors) can be transformed into another type (about the correct value of the ridge parameter).

3) *Equity estimator* (Krishnamurthi and Rangaswamy 1987) is a non-parametric procedure, unlike ridge regression. It makes an approximation of the original variables by a set of new orthogonal variables (in fact principal components, although the authors don't say that), then regress these components to the dependent variable, and ultimately recalculate the regression coefficients back to the original variables. Quite intensive numerical comparison (Wildt 1993; Rangaswamy and Krishnamurthi 1995) has shown that this approach outperforms both OLS and ridge regression (with one specifically selected regularization parameter) in some situations, but does not in others. However, analytically there is not much to be said about that approach.

4) A recently proposed *eventual ridge regression* (Lipovetsky 2006) obtained a unique solution, for which the classical functional of ridge regression was transformed into another one, with two parameters. The new term in penalty function minimizes the differences between pair and multivariate coefficients of regression. But the latter are equal to the former in a case of orthogonality. That is to say, eventual ridge regression may be considered as a combination of the two ideas: the original ridge regression and othogonalization, which is an essence of the equity estimator. However, the advantage here is that if in the equity estimator the orthogonality is interpreted in terms of somehow artificial principal components, here it has a very straight meaning. As shown, the solution not only became unique but also approximates the data much better than ridge regression,

keeps signs of regression coefficients the same as ones of correlations of predictors with the dependent variable (which is guaranteed neither by ridge nor by equity estimator) and bears other nice statistical features. In principle, it could be applied to any MC situation – however definitely with a significant loss of meaning if the real causes of MC are not determined.

The general feature of all considered approaches (except the Bayesian one) is that regression coefficients of strongly collinear variables (in standardized form) tend to be close to each other. This stems from the nature of penalty functions and algorithms used. However, in reality, strongly collinear variables may have completely different outcomes (like two correlated advertising channels may have very different efficiency). In that sense all approaches are just convenient approximation, admitting the sad fact that when two variables are "non-distinguishable" their outcomes are not distinguishable either. It is not a very promising message for models which are intended to reveal reality.

Scenario c) has only one appropriate solution – using the different techniques of simultaneous equations (SEM – see 2.2.2). However, the real implementation of it faces so many troubles, especially with dozens or even hundreds of potential predictors (which are not unique to economic problems), that the use of it is much more limited than regression modeling or even impossible. Moreover, even when applied, the MC still does not disappear and may affect estimation.

Scenario d) is quite complicated to handle. There is a big scope of works about modeling of constrained ratios, or compositional data (Aitchison 1986), but the question about separation of effects of functionally of related variables from "real correlation" is not solved in general way, to the best of my knowledge. As a result, many correlations in economic models may be in fact just consequences of normalization common to all variables (say, by population), not of "the real causes". However, this scenario has not been studied well enough and remains quite ambiguous, at least for the reason that there are many possible ways to introduce different "measurement biases", which may need a different treatment.

Scenario d) needs no comments – if one feels that the variable is random, it should be, of course, taken out of the model. However, is there a way to say that? Or more generally, can one distinguish between all of the above-listed causes of MC?

In real data sets, most likely, all combinations of scenarios a)-d) may appear. What to do if the modeler does not know the real causes of correlations? What method to prefer? In practice, mainly the choice is between reductionism and ridge regression approaches. I'm not aware of consistent studies comparing the relative losses and advantages if those techniques are applied to improper situations – so, generally, multicollinearity leaves a modeler with a high probability of being somehow mistaken.

2.2.1.2. ***Selecting the best predictors and correct estimation of regression coefficients.*** There are many options for selecting the best predictors, including stepwise, backward/forward techniques and new methods that continue to emerge (see Casella and Moreno 2006 and others). Yet all of these methods share a similar problem - "reduction of dimensionality". This was conceived a century ago in Spearmen's concept of factor analysis, and many approaches have been developing since that time. All of that can be classified as "numerical" implementation of Occam's razor paradigm, "Entities should not be multiplied unnecessarily" and intended to obtain a most compact and meaningful description of the problem and/or data (see also the discussion in 2.2.1.1). However, some issues arise when a compact model is not the desired outcome. For example, in the marketing mix models the goal is not to include as few variables as possible to get the best approximation but to include *as many as possible* in order to explain the impact from each marketing initiative. In this situation a reductionist approach loses its ground; the selection of variables problem is transforming into a problem of conscious specification of the model based on a predetermined design. For instance, advertising of one particular channel (say, TV) is often initially represented as a set of many variables, different from each other by decay rates and lags, when the so-called adstock concept of the prolonged advertising effect is applied (Broadbent 1997). A possible design in that case may be the following: each advertising channel should be represented in a model by just one variable. In this situation the best that statistics may recommend is a tool to estimate each variable's contribution into the general model's quality in some controllable way.

The promising way to solve this type of problem is *Shapley Value Regression* (Lipovetsky and Concline 2005). This method offers the ability to estimate the significance of each variable while accounting for all possible contributions from this variable in combination with others. It allows working with multicollinear variables (see above) and guarantees the same signs of the regression and correlation coefficients. Since each variable receives its "absolute value", a modeler may include as many variables as he wants, limited

only by common sense and the desired level of data approximation. The concept of data reduction may easily be combined with this idea as well. Characteristically, the eventual ridge regression and Shapley value regression often provide fairly similar results, though the complete testing of it is not done yet.

But even if the best selection in the sense of a model's fitting and other nice statistical features is obtained, the deeper questions still remain unanswered: do the selected variables really reflect the "drivers" of the outcome or they are just coincidental factors? Are there other drivers not included in the model? Ultimately, is that model causal or not? Before going into discussion (2.2.2), let's just dwell for a moment on one very important and often neglected aspect of the modeling – how the model's quality depends on correct specification of the model, or, in an even narrower sense, on *composition of the model*.

It is known that values of regression coefficients depend on the "environment", i.e., on other predictors in a model. For example, some variable *X* may enter the regression equation with a *positive significant* t-statistic, while with another set of predictors (sometimes just slightly different) it enters with a *negative significant* t-statistic. Theoretically, it may happen only if the variables affect each other (case c) in 2.2.1.1). But if this is true, regression analysis should be immediately replaced by something like simultaneous equations, which try to capture this kind of interaction. In practice it rarely happens for the following main reasons: hard interpretability of simultaneous equations models; much higher statistical level of modeler required for it; difficulties of implementation for many variables, and ultimately because they do not guarantee the ultimate correctness (see 2.2.2). As a result, regression analysis is still the main instrument for practical modeling. The interaction between variables is either ignored or utilized in a specific process of model building, which is called the "art of modeling", ninety percent of which is in fact the art of selecting the proper set of variables to make them meaningful.

This art is usually the result of some combination of predictors, selected by algorithms, "domain expertise", or by both as deemed satisfactory from the modeler's point of view. But paradoxically enough, this kind of logic is in fact exploiting that very multicollinearity in the data, which by the same commonly accepted logic is one of the main troubles. Indeed, if multicollinearity didn't exist, all coefficients of multivariate regression would be equal to those of pair regressions and would not change after addition or deletion of other predictors. In that case the selection of predictors would depend only on the desired level of their contribution (see above). But all that works as described only when correlations between predictors are indeed zeroes. Even a small departure from that (which is practically inevitable) may result in disproportionately large uncertainty in estimation of parameters as a function on total composition, in a quite unexpected way.

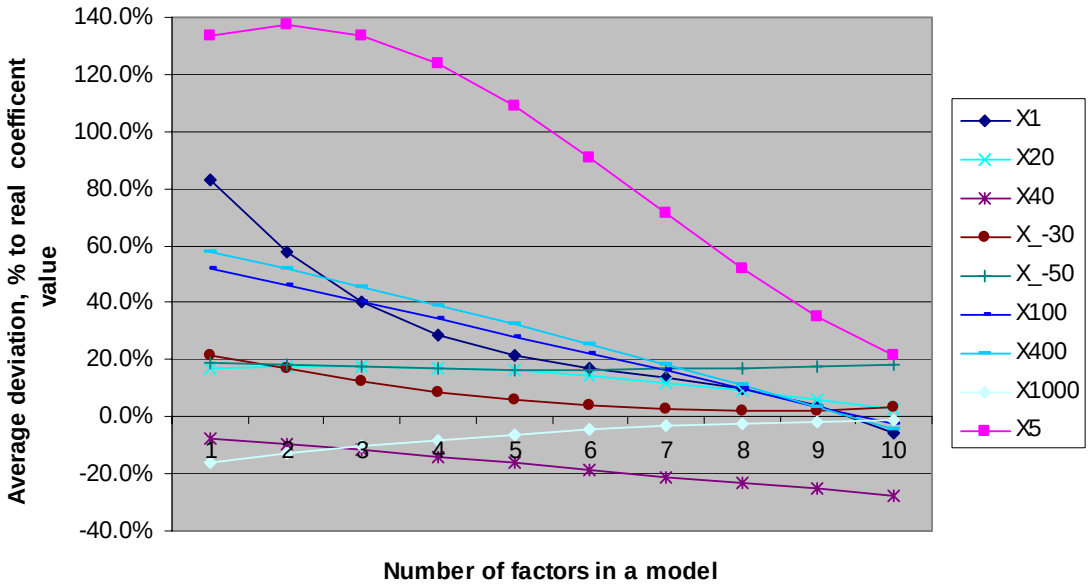
This problem, in spite of its obvious importance, does not find a definitive answer (see discussion in Franses and Heij 2002). Usually, textbooks consider an idealized situation when there is a postulated model with *K* variables and there is a mechanism of estimating the coefficients of these *K* variables. Practically, there are proposals to run as many regressions as possible (couples, triples, etc.) on a concrete data set and then to see the lower and upper boundaries for each coefficient. If the lower limit for positive and the upper limit for negative coefficients lie higher and lower than zero respectively, the given coefficient is considered stable – at least always positive or negative (Leamer 1985). However, the cases in which this occurs are extremely rare. Consequently a soft modification of this approach was considered (Sala-I-Martin 1997), where instead of a strong location of the boundary at above or below zero some 95% criterion of occurrences is proposed. These modifications allow for the inclusion of many more variables. Those experiments (for actual economic times series) have significant learning value focusing on the instability of estimations, even when t-stats and other conventional statistics tell the opposite. However, since the real values of these parameters are unknown, the final conclusion remains unclear. The problem ultimately can be reduced to the selection of the most stable for particular data set variables. In that sense it is matching somehow the ideas of Shapley regression, though without its theoretical appeal.

It makes sense to compare *known* values of the regressions' parameters versus the *estimated* parameters against different subsets of predictors. I have done such experiments on generated data with quite interesting results (Mandel 2007). Without going into the details of data generation, let's just say that the dependent variable was formed as a linear function of ten random variables with known regression coefficient plus noise. Then all possible regressions (with one predictor, two predictors, etc.) on that data set were calculated. One typical chart from this study is presented in Figure 1. The vertical axis shows an average percent deviation of the estimated coefficient from the real value; the horizontal axis, number of predictors

used. For example, the value 135% for variable X5 against 3 shows the average estimation of coefficient for X5 out of all triples possible was higher than the real value of 135%.

As expected, the majority of predictors are better estimated when the number of predictors in the model increases. However, estimation of variable x40, for instance, steadily worsens as the number of predictors is increased. Other variables have slight non-conversion effects as well. In general, the graph demonstrates a large variety of curves, all of which assume high nonlinearity and non-predictability of the process. Only one predictor, x1000, meets two strong conditions–independence from other predictors and monotonic improvement of estimation when the specification approaches the real value. All other variables have different problems (one variable is not even shown here because the deviation is huge, from 200% to 400%).

Figure 1 – Estimation of regression coefficients based on number of predictors



In general, the experiment shows that for the data *with and without multicollinearity*:

a) the estimated values may be far away from the real values of the coefficients;

b) the estimated values may diverge from the real values when the number of variables used is approaching the actual number of predictors originated in the outcome.

c) if variables which have not affected the generated outcome (coincidental, but correlated with outcome) are used in estimation, the divergence becomes much stronger.

These results indicate that even if a mechanism of data generation is exactly as it is supposed to be, in a regression model it cannot identify the real structure of the data generation even in simple cases. This fact alone undermines the blind use of regression as a universal tool.

2.2.2. Causality analysis

From a practical (and especially business) point of view, the problem of causality is seen very simply: if a model is not causal, it's not a model at all. How can a modeler explain to the client that "correlation is not causation", that "regression does not mean causal" and so on? The entire topic of the difference between statistical and causal models is out of the brackets in the relationship between statistician and client. But within statistics it is a never-ending battlefield.

Causal analysis started almost simultaneously with the inception of modern statistics dating back to S. Wright's first article on path analysis published in 1918 (see complete account in Wright 1921, under the characteristic title "Correlation and Causation"). Path analysis grew to become a very rich field of *structural (or simultaneous) equations modeling*, *SEM* (Kline 2005). Several new branches of causal analysis have developed in the last three decades, some of which have presented ideologies seemingly quite contrary to the traditional thought. These branches could be vaguely grouped as *Bayesian causal networks* (Spirtes, Glymour, and Scheines 2001), *potential outcomes concept* (Rubin 2006), *temporal statistical (Granger causality) concept* (Hoover 2001), and *interventionist concept* (Pearl 2000; Robins 2003). In some cases these approaches overlap between themselves and SEM, so their aforementioned names and groupings are

tentative. In an excellent review (Pearl 2003), a different stratification is given: a) graphical models; b) structural equations; c) counterfactual or potential outcomes concept, and d) unified concept, combining all of the above-mentioned in one paradigm. Indeed, it sounds as though seemingly different things started to converge into one theory (keeping Granger's concept outside of it though), which is definitely not finished yet. I can't consider here even briefly the huge scope of related works. But I may rather say with higher confidence what these works *are not*.

1. Not all of the concepts share the same **definition of causality**. Clive Granger calls variable X a cause of variable Y if the knowledge of X helps in predicting the future values of Y better than solely the knowledge of Y (in time series setting). It has often been noted (see Sorensen 2005 and the main line of thinking in this article) that this type of causality has a pure statistical sense with all associated problems of distinction between "real" and "false" predictors.

Donald Rubin defines the causal effect as the difference between real and potential outcomes on the same object. The last one, being impossible to observe in nature, should be approximated by comparing the object with other objects with the most similar values of other predictors (which is obtained by using propensity scores and matching). The idea here is completely different from Granger's and very similar to the concept of "what-if history". The latter theory, although tempting, is usually deservedly denied its place in historical studies, but seems more relevant for statistics. This is because not just one unique historical event, but many observations are available. However even in that interpretation a "what-if" assumption often seems quite irrational or absurd. A good example is in Berk 2003, when such an approach is applied to analyze job discrimination against women. The "potential outcome" should reflect a situation when a woman is supposedly "replaced" by a man. The author very reasonably concludes that such a major change will also cause an immediate change in all other variables as well (see similar arguments by S. Fienberg and Haviland in Pearl 2003). Thus the concept of isolated causal influence will be broken.

Maybe, the most cited work on the subject to date is the book Pearl 2000, which surprisingly does not have a strong definition of causality. It seems that the author accepts a wide range of definitions all throughout the text. It is not only my impression (see, for example, comments by D. Heckerman and R. Shachter that Pearl's "…definition of causal model was (and still is) unclear" (Pearl 2003, p.328). It's especially sad, because the normal style of Judea Pearl is quite straightforward (keeping apart his undoubtedly important results).

SEM or path analysis, technically, is an extension of regression analysis expanded to a system of interrelated equations. SEM generally uses the same definition: "cause" is what "affects" the result and is measured by a path coefficient (i.e., a specific type of regression). Let's do a simple mental experiment: there are two non-correlated variables, supposedly affecting an outcome. It is known that one affects it causally, another not at all. The path coefficient between these predictors is zero. I do not see a way that SEM may strictly detect the qualitative difference between these two variables. Moreover, regression, being a particular case of SEM without interrelation between predictors, may also be called causal – which is very questionable though. It is especially clear when all correlations (and thus indirect path coefficients) are zeroes – then SEM is reduced to regression with all its problems, without any causal aspect to it.

Maybe such discordance in definitions is not random. As Lofty Zadeh argues, the causality is un-definable in principle (Zadeh 2003). It's particularly interesting to hear from the creator of a fuzzy sets theory: causality seems to him fuzzier than almost everything around what fuzzy sets theory claims to express with its language. I can't now go into my own interpretation of causality or discuss many other existing definitions. To expand the picture, let's just bring one quote from an abstract of an article in a leading sociological journal: *"The last decade featured the emergence of a significant and growing literature concerning comparative-historical methods. This literature offers methodological tools for causal and descriptive inference that go beyond the techniques currently available in mainstream statistical analysis"* (Mahoney 2004, p.81). After such a statement one may expect to find further at least some of the works referred to above, but in vain. "Comparative-historical methods" are yet something different. I may just emphasize once again that something is definitely not too clear here, even among the most distinguished specialists and approaches.

2. As a result, different approaches do not solve the **same problems** and definitely do not solve **all problems**. A potential outcome approach is a method for a single, though very important type of data, when one needs to compare "treatment vs. no treatment". It could be considered a problem of estimating the effect of one qualitative (usually binary) variable to the quantitative outcome; it can't work with a qualitative

independent variable (although there are some attempts of generalization). Granger causality methods work with both quantitative and qualitative data, but only in a time series setting. SEM is the most universal in terms of a variety of possible data settings but also is the closest to a classical regression and potentially remote from "real" causal notions. Plus, SEM needs to formulate so many underlying assumptions that may not be manageable in a real situation with many variables, similar to Bayesian models. Even after recent demonstration in (Pearl 2003) that three key approaches (SEM, interventionist and potential outcomes) could be formally considered as identical under some conditions, many questions remain unanswered – some criticism concerning different approaches, presented above, may just be extended to others as well.

3. Causal approaches are almost exclusively concerned with the treatment of **variables, not the objects** by which these variables are measured. More precisely, the typical question asked is "*does variable A cause B?"* rather than *"How do different objects behave when/if A causes B?"* This difference looks subtle, but from a statistical point of view it may be very important, because many aspects of actual situations depend on the concrete types of data that are observed. A systematic implementation of SEM in homogeneous groups, preliminarily obtained by clustering (which is widely accepted as a normal practice for statistics in general) could prove to be a real nightmare due to its difficult interpretability, and thus is practically not used. On a different note, J. Pearl devoted a chapter in his book to Simpson's paradox, with the following key phrase: "*The explanation for Simpson's paradox should be clear to readers of this book, since we have taken great care in distinguishing **seeing** from **doing***" (Pearl 2000, p. 174). Then a paradox is explained in a causal interventionist style. However, a whole paradox could definitely be explained from a pure statistical point of view. If treatment works well for males and females separately and doesn't work for the "population" as a whole it is because the "population" was not homogeneous. Thus the "population" needed to be separated by sex, as any other heterogeneous population where less dramatic yet still important distortions are routinely found. The recent analysis of the "paradox" (Lipovetsky and Conclin 2006) shows it clearly without any appeal to the distinction between "seeing and doing", however important this distinction is per se.

Taking a further step from groups to individual objects, one should admit that in real life causal action takes place in each particular moment of time and for each particular entity. For example, today's advertising affects people with one rate, but in the next week the rate will be different. However, nearly all statistical *and causal* methods ignore that fact and make an estimation of one parameter for one predictor, as if advertising always worked with the same efficiency. The attempt to overcome that trouble was done in the approach called ***yield analysis*** (Mandel 2003; Demidenko and Mandel 2005). In this paradigm, estimations of influence are made in each data point and vary over time and space, analogously to some non-parametric regression models, but with such an important difference, that no smoothing technique is used. Experiments show that the algorithm is able to recover the individual coefficients used in data generation in many situations, i.e., really reveals the mechanism of the influence. Within that concept it's easy to demonstrate that variables producing positive contributions in each data point may still be negatively correlated to the outcome due to the rates of production (yields) changing accordingly. Hence the correlation analysis would become irrelevant to the real causes of events, if one believes in this type of data generation mechanism (which seems very close to reality). Accordingly, the *whole regression machinery will not work*, SEM included…

4. A deeper problem with statistical models is that the models may be farther from the real mechanism of data origination than assumed in most comprehensive statistical schemes. The classical statistical model (with or without a causal slant) postulates an existence of some "data generator," which is usually some kind of function of generic type. Estimation procedures are built based on the generation mechanism, with all possible problems mentioned and not mentioned above. However, all these models live their own life, practically in complete isolation from other powerful theories describing a world, possibly, in a much more convincing and logical way. Let's just name here the most important interrelated approaches of that type.

a) **Theory of complex systems** and especially **theory of self-organizing and evolving systems** (Hemelrijk 2005) may explain the core phenomena in social in economic life, and yet statistical models typically do not take it into account at all.

b) The latest developments in **sociophysics** (Ball 2006) and **econophysics** (McCauley 2004) convincingly show that many social and economic situations may be described by physical or physical-like laws. Interactions between individuals, collective and herding behavior of masses govern wealth and income distribution, company size, conflict severity, financial indicators, word of mouth, social networking and

many other phenomena. A self-organization concept is incorporated in physical models very naturally as well (through equilibrium and self-criticality phenomena and so on). However, only rare efforts may be found to bridge statistical and physical models - see discussion in Kuznetsov and Mandel 2007, where **mediaphysics** presents the possible integration of the two approaches.

c) **Agent based modeling (ABM)** uses a different paradigm than statistics and incorporates physical and self-organizing principles in a much more organic way than statistics. The impressive results in economics of this type of models (North and Macal 2007) may be a sign that *ABM could be a real challenge for current dominance of statistical models* in that area.

Let's just hope that ultimately the entire new paradigm of statistical inference will be developed, where *not goodness of fit or even predictability play a key role, but relevance of causal explanations to the observed facts. And causality will be understood not in the current limited and drastically simplified manner, but as a feature derived from the revealed mechanisms of the complex dynamic systems.*

2.3. Technical and probabilistic uncertainties

Technical uncertainty appears when the main approach is chosen, but there are many (uncertain) other problems to be solved. Partly it overlaps with already considered problems (like selection of the best predictors--on the one hand, it's a fundamental problem because it reflects the core of the causal process, but on the other hand, it's also technical in a more narrow data mining sense). Here I would touch just two but very important "technical" issues: a **lack of consistency** among numerous concrete procedures and a **lack of time** for making a good model.

Each method in statistics is in fact a group of algorithms which sometimes is very large. If in the 1980s one could count and describe about one hundred algorithms of cluster analysis, which was, of course, still an incomplete set (Mandel 1988), a task like that now looks impossible (at least I don't know such an exhaustive inventory for the most important groups of methods). Even for the described above "ideal" and comparatively narrow letter recognition problem, the authors applied 5 different layouts of neural networks. Each of them could win, and one did. But the total number of possible algorithms for this problem could be enormous and fill the hundreds of statistical magazines published in the world each month. Who would dare say that he or she knows even a comparatively narrow area? Each algorithm, if not purely heuristic, is based on some strong formal assumptions. Who would dare say that he or she really knows not only those hundreds of algorithms, but also all the assumptions behind them? Most importantly, who would dare say that he or she can check their validity in each particular case? Bayesian averaging, a fast developing branch of statistics, is actually about the art and science of comparing several models of one process (see the through survey in Lipkovich 2003). How does all that huge academic richness, which remains unused in practice 99.9% of the time, affect the day-to-day modeler's work with concrete data and specific problems? It is not a surprise that modeling has become like a magician's art. Each performer uses his or her tricks which are based mainly on personal habits, customs and passions and just lastly on their relevance to problems. Additionally, it is no wonder that if one wants to know exactly what tricks have been used within specific software by a specialized company, with all details, assumptions, constraints, and formulas, he will never get an answer. The modeling business, especially in its data mining or marketing mix modeling in industrialized forms, has progressed into a typical black-box occupation such as dentistry or medicine. However, unlike medicine, the modeling business is without a natural control and feed-back from the government and often even from the client. The rare appeals to disclose details of the techniques used (Naik, Schultz and Srinivasan 2007) are not heard by the community.

As a result, the modeling industry as a whole produces thousands of models, which are incomparable in their final results, by techniques, and by validity to the "reality." Modeling has an extremely transient form. Everything is going to be changed in the next week or month without a reference to the previously done and highly advertised work. Nothing is even intended to be constructed into strong, solid and lasting "regularities" even remotely reminiscent of the physical laws. The client often sees this inconsistency and either simply gives up or drops the model just after doing it (see the interesting admission of this fact by a top manager in Bemmaor and Franses 2005). Just recently I asked a very good, mathematically educated manager at a large financial institution, why instead of a powerful optimization procedure they use just rough estimation. The answer was that a) they don't have time for that and b) a model will be redone very soon anyway – so, why worry? This situation of rush and fuzziness is perfectly reflected in the following

words of a prominent marketing researcher: "...I have never found widely used procedures, for example Gaussian least-squares *regression or multiple factor analysis, to be of any use* – rather like blood-letting. One reason was that *these techniques have not led to a lasting scientific discovery over the last 100 years, or even to any claim for one*… In the 1990s … some 30+ leading modelers worldwide were invited to apply their own preferred analytic procedure to some simple repeat-buying data for new cars… We found … that the 20+ participating modeling experts had given 20+ different answers" (Ehrenberg 2005, *my italics–I.M*). But such sobering statements are rather exceptions; much more often one hears enthusiastic promises that model's use would "add value," "increase awareness," and "multiply ROI."

In summary, even if one has a lot of time and is equipped with all modern techniques, the modeler will still not be able defend his creation with the same zeal and assurance as that with which an engineer can defend a newly built machine. And although the President of the American Statistical Association may feel that a "…statistician is an engineer" (McNaulty 2006), he it is not. If it were true, this "engineer's products" would be used for only one customer and for only one moment, life in such an engineered world will be like life in a time before Watt. On the other hand, the "engineering view" of statistics, being in fact very popular among practitioners, is one of the main de-stimulating forces, preventing modelers from adopting the new approaches.

Besides all of these considerations, the most investigated is the classical concept of **probabilistic uncertainty**, which comes directly from the probabilistic nature of the processes. Its is associated with the fact that our limited estimations of anything from the "sample" are never equal to the "real" parameters of "all". In one of the most comprehensive accounts, "bias-variance trade-off" consists of 10 different elements like "truth," "realization," and "closest fit" (Hastie, Tibshirani, and Friedman 2001, p.198-200). These elements are the matter of main concerns for mathematical statistics (some of them, like misspecification error, go also beyond it – see 2.2.2). Thousands of works are dedicated to determining different confidence intervals. As many specialists think, the entire science of statistics is about these types of things. Progress there is unquestionable; however, how is it related to our original problem? Is it not enough to note that in 95% of practical works in business, confidence intervals are not even reported to the client and that if they were, the client wouldn't know what to do with them?

Conclusion

As the discussion above shows, **a general modeling uncertainty** consists of business and statistical components, which together make the modeling much fuzzier than one may expect. Although it is an inevitable and almost immeasurable part of the modeling process, it is a practically unanalyzed phenomenon in its entirety. The situation is schematically represented in Fig.2. Apparently, the only measurable probabilistic uncertainty is a small part of the real uncertainty, although, maybe 90% of all efforts in statistics historically have been so designated.

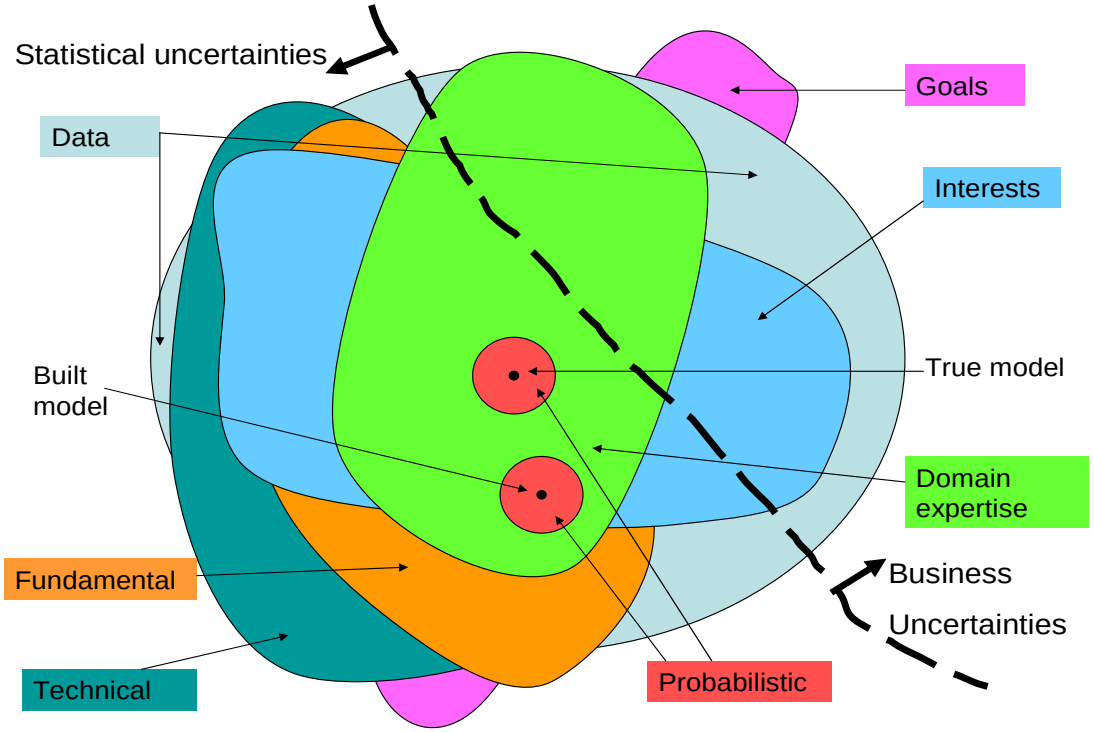
Business uncertainties directly interfere with statistical ones - all of them cross the demarcation line between the two classes. But business is completely unaware of pure statistical problems, and respectively statistical uncertainties do not cross the line. Two types of uncertainties – related with data and domain expertise – are created usually by both statistician and client. For different types of clients uncertainties could be quite different. For example, in genetics and engineering, the goals and interests uncertainties are usually minimal, but in the pharmaceutical industry they could be high. The intercrossing of business uncertainties into statistical areas also means that the statistician experiences not only the statistical, but also the typical internal business problems discussed in 1.4.

This chart implies that much wider than probabilistic yet immeasurable (now?) confidence intervals for any model are the right uncertainty indicators. Thus, it undermines the popular clichés about the "dumb deep-pocket client," who should be taught, and the "wise truth-bearer scientist," who would be a teacher. What conclusions could be drawn from it?

First, a wide confidence interval, being not formally determined, allows manipulating a model's conclusions in a freer style than is usually accepted. A modeler may feel comfortable, including in a model different constraints or limitations coming not from a theoretical point of view, but from very earthy, even transient, business considerations, if it looks more interpretable. In general, an important criterion of model interpretability should be, in my opinion, partly shifted from the statistical to the business domain. It means, particularly, there would be an increased role of such informal factors as general concordance of all model

components. For example, whatever the model says, if the advertising success rate in some regions is one hundred times higher than in other regions, it's more reasonable to cut it off and make the rates more even, without any "statistical justification" or choose another model.

Figure 2. General modeling uncertainty



Second, the wide confidence level mitigates possible moral conflicts in client-statistician relations (Lesser and Nordenhaug 2004) and eventually may tighten those relations. Morally, a statistician should not believe himself to be a guru until fundamental problems are solved. As a result, a statistician could be much more flexible in any direction, hoping that he is still in the wide but hidden confidence interval drawn in Figure 2. Practically, an honest disclosure to the client with all of the above-described circumstances could lead to relations between them in the only proper way. Instead of one model, many models should be done. Instead of having the statistician or client have no feedback to the model, access to the modeling implementation results should be assured in the beginning. Instead of having "all accessible" data, data collection should be organized specially for modeling purposes and so on. All these means of decreasing business uncertainties are beneficial, but they should first be determined. The same logic follows with statistical uncertainty. If the client understands that there is such an inevitable thing, he may be willing to give more time and possibly money for applying several techniques instead of one and will not rely anymore on different "automatic solutions," which are still offered by some companies. Further important analysis of ethical and business aspects in the pair "client-statistician" in situations of high uncertainty is possible based on the powerful mathematical-psychological theory developed in Lefebfvre 2002, which is a topic for special study.

Third, I would propose organizing a regular contest of socio-economic marketing models, analogous to ones running on UCI Depositary data sets (Asuncion and Newman 2007). This popular depositary does not have typical marketing data and is oriented only on machine learning specialists. Marketing mix models, for example, the most problematic type of modeling and one of the main topics in this article, affecting billions of dollars of advertising budget, practically are not tested in that way. Typical problems of **confidentiality** of business data, which complicate a results comparison, could be overcome by special coding of data or otherwise.

Fourth, the wider realization of the fact that applied modeling has many uncertainties other than those of a probabilistic nature may help in shifting the attention of experts from pure mathematical-statistical aspects to fundamental and business-related problems.

Acknowledgements

Together with Dr. D. Hauser, I presented on JSM 2006 materials, partly used in this article; I had extremely fruitful discussions with Dr. I. Lipkovich and Dr. S. Lipovetsky time and again; V. Kamensky made a program on SAS for an experiment with selection of predictors. I'm heartily thankful to all of them for their help and contribution. I was very lucky to get two extremely knowledgeable and serious anonymous reviewers, who definitely didn't count their time for series of very important remarks, recommendations, and new references. I took them all into account, what, I believe, significantly improved the quality of the article.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman & Hall

Asuncion, A and Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. University of California

Ball P. (2006) Critical mass. How one thing leads to another. Farrar, Straus and Giroux

Bemmaor A. C. and Franses P.H. (2005) The diffusion of marketing science in the practitioners' community: opening the black box. Applied stochastic models in business and industry, 2005; 21:289–301

Berk R. (2003) Regression Analysis: A Constructive Critique. Sage Publications

Bierens H. Multicollinearity. (2005) Pennsylvania State University

htpp://econ.la.psu.edu/~hbierens/MULTCOL.PDF

Broadbent S.(1997). Accountable Advertisement. UK, Admap Publications.

Casella G. and Moreno E. (2006) "Objective Bayesian Variable Selection." *Journal of the American Statistical Association, 101, 157-167*

Demidenko E., Mandel I. (2005) "Yield Analysis and Mixed Model." *Joint Statistical Meeting 2005, Proceedings, p.1636-1643*

Ehrenberg A. (2005) "My Research in Marketing." *Admap Magazine*, Issue 461

Echambadi R., Campbell B. and Agarwal R. (2006) Encouraging Best Practice in Quantitative Management Research: An Incomplete List of Opportunities. *Journal of Management Studies 43:8 December 2006 0022-2380*

Franses P.H. and Heij C. (2002) "Estimated parameters do not get the wrong sign due to collinearity across included variables." *ERIM report series research in management*, ERIM 2002-31-MKT, Erasmus research institute of management, March 2002

Han J. and Kamber M. (2006) Data Mining: Concepts and Techniques. Elsevier

Hand D. (2007) Deception and dishonesty with data: fraud in science. *Significance*, v.4, Issue 1, 22-25.

Hastie T., Tibshirani, R. and Friedman, J. (2001) The elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag

Hemelrijk C., editor (2005) Self-organization and Evolution of Biological and Social Systems. Cambridge University Press

Hoerl, A.E. and Kennard, R.W. (1970) "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12(3):55-67

Gibbons D.G. (1981) A simulation studies of some ridge estimators. *Journal of American Statistical association,* 76 (March)

Golub G., Heath M, and Wahba G. (1979) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, v. 21, 2

Granger, C. W. J. and Newbold, P. (1974) “Spurious regressions in econometrics.” *Journal of Econometrics* 2

Hoover K. (2001) Causality in Macroeconomics. Cambridge University Press

Keller-McNulty S. (2006) “Epiphany: I Am an Engineer.” Amstat News, Vol. 344, February, 2

Kline R. B. (2005) Principles and Practice of Structural Equation Modeling. The Guilford Press

Krishnamurthi L. and Rangaswamy A. (1987) The equity estimator for marketing research. *Marketing Science* v.6, 4.

Kuznetsov D. and Mandel I. (2007) “Statistical physics of media processes: Mediaphysics.” *Physica A*, Vol. 377

Leamer E. (1973) Multicollinearity:A Bayesian interpretation. *The review of economics and statistics, 1*.

Leamer E. (1985) “Sensitivity Analyses Would Help.” *American Economic Review*, 57(3)

Lefevbre V. (2002) Algebra of Conscience. Kluwer Academic Publishers

Lesser L. M. and Nordenhaug E. (2004) “Ethical Statistics and Statistical Ethics: Making an Interdisciplinary Module.” *Journal of Statistics Education*, Vol. 12, No. 3

Lipkovich, I. (2002) “Bayesian Model Averaging and Variable Selection in Multivariate.” *Ecological Models.* PhD Dissertation, Virginia Polytechnic Institute.

Lipovetsky S. and Conklin M. (2005) “Incremental net effects in multiple regression.” *International journal of mathematical education in science and technology*, Vol. 36, 4

Lipovetsky S. and Conklin M. (2006) “Data Aggregation and Simpson’s Paradox Gauged by Index Numbers.” *European Journal of Operational Research*, 172

Lipovetsky S., (2006) “Two-parameter ridge regression and its convergence to the eventual pairwise model.” *Mathematical and Computer Modelling*, 44

Lipovetsky S. and Conklin M. (2006a) “A Model for Considering Multicollinearity” International Journal of Mathematical Education in Science and Technology, 2003, 34

McCauley J. (2004) Dynamics of Markets: Econophysics and Finance, *Cambridge University Press*

Murukami H. (1999) The Wind-up Bird Chronicle, *Vintage*

Losh S.C. (2004) Guidance to Introductory Statistics, *Florida State University, http://edf5400-01.fa04.fsu.edu/Guide8.html*

Mahoney J. (2004) Comparative-historical methodology. *Annual Review of Sociology*. 30:81–101

Mandel I. (1988) “Cluster analysis.” *Finance and Statistics*, Moscow (Russian; English review by V. Kamensky, *Journal of Classification*, 1990, 7, 119-123)

Mandel I. (2003) “Multicollinearity problem in marketing studies.” *Joint Statistical Meeting Proceedings*

Mandel I. (2007) Statistical aspects of marketing mix models. *Joint Statistical Meeting Proceedings*

McCauley J. (2004) Dynamics of Markets: Econophysics and Finance. Cambridge University Press

Naik P. A., Schultz D.E. and Srinivasan S. (2007) Perils of Using OLS to Estimate Muitimedia Communications Effects. Journal of Advertising Research, September.

North M. and Macal C. (2007) Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press Pearl J. (2000) Causality. Cambridge University Press, Cambridge, UK

Pearl J. Statistics and Causal Inference: A Review (*together with following discussion and a rejoinder*) (2003). *Test* V. 12, No. 2, pp. 281–345

Rangaswamy A. and Krishnamurthi L. (1995) Equity Estimation and Assessing Market Response: A Rejoinder. *Journal of Marketing Research* Vol. XXXII

Robins J. M. (2003) General methodological considerations. *Journal of Econometrics*, 112

Rubin D. (2006) Matched Sampling for Causal Effects. Cambridge University Press

Sala-I-Martin X. X. (1997) I Just Ran Two Million Regressions. *The American Economic Review*, Vol. 87, No. 2

Sorensen B. (2005) Granger Causality. *Economics 7395*, March 1

Sorokin P.A. (1956) Fads and foibles in Modern Sociology and related sciences. Henry Regnery Company (Gateway edition, 1965)

Spirtes P., Glymour C., and Scheines R. (2001) Causation, Prediction, and Search (Adaptive Computation and Machine Learning). The MIT Press

Theil H. Principles of Econometrics (1971). John Wiley and Sons, Inc.

Tikhonov A.N. and Arsenin V.Y. (1977) Solutions of Ill-Posed Problems. Winston

Wildt A. R. (1993) Equity Estimation and Assessing Market Response, *Journal of Marketing Research*, 30, 4 (November)

Wright S. (1921) "Correlation and Causation." *Journal of Agricultural* Research, Vol. 20, No. 7

Zadeh L. (2003) Causality is indefinable. www.cs.berkeley.edu/~nikraves/zadeh/Zadeh2.doc.