CHAPTER 6

DISCUSSION

HISTORICAL PERSPECTIVE

Statistical concepts and methods have been strongly developed in the last century with major contributions of Francis Galton, Karl Pearson, Ronald Fisher and Jerzy Neyman at the end of the 19^{th} and the beginning of the 20^{th} century. For instance the concept of the correlation coefficient can be traced back to Karl Pearson,[1] whereas regression analysis goes back to Francis Galton[2] and George Yule.[3] The introduction of two other important regression methods can be contributed to David Cox: *logistic regression* analysis in 1958[4] and the *proportional hazards model* in 1972.[5] All these regression-based techniques are still extensively used in many research areas for the prediction and explanation of various phenomena.

When the main objective is to estimate a single treatment effect, as is common in medical and pharmaceutical research, the *randomized experiment* is the gold standard. Very influential on the design of such experiments was the work of Fisher, who is in general credited for the invention of randomized experiment in 1925.[6,7] The concept of random assignment of treatments goes back to Neyman,[8,9] but even in 1885 randomization was used by Charles Peirce.[10]

When an experiment is not possible or when subjects are not randomly assigned to treatments, it can still be the main objective to estimate a single treatment effect or exposure and to adjust for the *confounding* influence of other factors that are prognostic for the outcome. It is common practice to use the earlier mentioned conventional regression-based methods for this purpose, but alternative methods have specifically been developed to adjust for confounding. Two of these methods are investigated in more detail in this thesis: the method of *propensity scores* and *instrumental variables*. The first is a relatively recent method, developed by Rosenbaum & Rubin in 1983.[11] It has its fundament in matching on a continuous variable in the work of Rubin[12,13] and Cochran & Rubin.[14] The other main topic of this thesis is the method of *instrumental variables*. The related problem of *solving the identification problem* in simultaneous equation models goes back to Philip Wright in 1928,[15,16] whereas the term instrumental variable first appeared in work of Reiersøl in 1945.[17,18] The first appearance in medical research was probably in 1989 by Permutt and Hebel.[19]

The development and the improvement of methods to adjust for confounding are important, because results from observational studies can only be used to inform clinical practice if the effect estimates of treatment are valid. In the last decade an *increasing use* of propensity scores in medical studies can be noticed, but still there are questions on how these methods are best applied in different settings. Although instrumental variables as an adjustment method is less known and in general *less applicable*, the same is true for this method: how and when to apply this method. Therefore, the aim of this thesis was to assess the strengths and limitations of alternative adjustment methods (Chapter 2), to compare these methods to conventional regression-based methods (Chapter 3), to demonstrate less straightforward applications (Chapter 4) and to further develop these alternative methods (Chapter 5).

WHY TO USE ALTERNATIVE ADJUSTMENT METHODS?

To adjust for confounding in non-randomized studies the factor of interest and all possible confounders are usually included in a regression model. Alternatively one can use methods of propensity scores and instrumental variables to adjust for confounding. Contrary to linear, logistic or Cox proportional hazards regression, these methods have been developed with a randomized controlled trial in mind, which is the preferred research design to estimate intended treatment effects. Propensity score methods and instrumental variables have one variable of main interest (the treatment or exposure variable) and are primarily concerned with *similarity of treatment groups*. This is not true for regression-based methods in which all variables, confounders and treatment variable, have technically the same place in the outcome model. Although methods of propensity scores and instrumental variables can also use regression methods to finally estimate treatment effects, the philosophy is quite distinct from directly estimating adjusted treatment effects with a linear, logistic or Cox proportional hazards regression model.

PERFECT SIMILARITY

Important questions in medical research are whether a certain drug is effective to prevent or cure a specific disease and whether exposure to some environmental factor influences health. Such questions are primarily directed to find an *average causal effect* of a factor of interest, the treatment or exposure variable. One hypothetical way to answer such questions is to observe all individuals of a certain sample in two different states at the same moment. For instance, observe the cholesterol level when a patient is treated by a drug and compare it to its level when the *same patient* is not treated by this drug over the *same time period.* The direct causal effect of treatment for all individuals will then be exactly known because all other possible explanations except treatment did not change. Unfortunately it is physically impossible to measure the same person at the same time in two different states, treated and untreated, or exposed and not exposed. This problem could be solved by using two exactly identical individuals in order to observe this pair at the same time, one as treated and the other as untreated. For experiments with animals two identical (e.g. genetically) subjects could be used to be assigned to two different treatments. However, for humans this is virtually impossible, even when we restrict similarity to only those factors that are prognostic for the outcome.

AVERAGE SIMILARITY ON GROUP LEVEL

To make a valid assessment of an average treatment effect it is too restrictive to demand that *individuals* should be similar between both treatment groups. It should be sufficient for assessing an average treatment effect to reach on average similarity on group level. This can be achieved by a procedure known as *randomization*, a term contributed to Fisher.[6,20] Two groups of patients are on average similar on all characteristics if the assignment to these groups was completely at random, or in other words, if all individuals had the same probability to be in one of the treatment groups. Nowadays this procedure is the scientific standard for medical research of interventions. This procedure assumes that the researcher has *control over treatments* or exposures in order to let the toss of a coin decide which of the treatments are given to the individuals.

NO CONTROL OVER TREATMENTS OR EXPOSURES

Unfortunately it is not always possible for researchers to have control over treatments, exposures or, more general, the factor of interest. This means that when observing a certain population many other factors can be explanations for the average difference in outcome between treatment groups, because treatment groups do not differ only by treatment. To adjust for factors that differ between treatment groups and at the same time are prognostic for the outcome, a distinction can be made by three possible approaches. The first is a regression-based approach in which all confounders and the treatment are included in the same model. The second is to create similarity of treatment groups within subcategories based on the confounders (propensity score methods) and the third is to identify groups that are similar with regard to confounders, but differ with regard to treatment (method of instrumental variables). These approaches are further explained in the next paragraphs.

REGRESSION-BASED APPROACH

A first approach is based on the simple general principle *'if you can't beat them, join them'*. If you can not get rid of competing, alternative explanations beforehand by controlling treatments or exposures, then combine them in a joint model with the factor of interest in order to adjust for their confounding influence on the treatment effect estimate. Such statistical models are for instance linear regression analysis, logistic regression analysis and Cox proportional hazards regression. Although one is primarily interested in the treatment or exposure effect, such models *simultaneously* estimate an adjusted treatment effect as well as all individual effects of the potential confounders in the model.

CREATE SIMILARITY OF GROUPS

A second approach of handling the problem of confounding is based on the idea that the undesirable differences between treatment groups are due to *different probabilities* of belonging to one of the treatment groups. If all factors that both influence treatment and outcome are known, these probabilities (the true *propensity scores*) can be determined. For individuals or groups of individuals who have the same probability to be treated, it is as if the toss of a coin has decided who was actually treated and who was not. In practice, all true confounders are unknown and the propensity scores have to be estimated with only *observed confounders*. This implies that the propensity score does not replace a randomized experiment because adjustment for unobserved factors is not possible. Using the estimated propensity score to create *subgroups or pairs* of subjects, the original comparison between *all* treated and *all* untreated subjects is now replaced by a comparison within those subgroups or pairs. Although regression-based methods are also frequently used in propensity score methods, it is different from the first approach because in a propensity score model confounders are *not directly related to the outcome*.

USE RELATED 'TREATMENT GROUPS'

A third approach is based on the idea that if we can not directly relate treatment or exposure to outcome without facing the influence of other factors, we create or identify *a slightly different 'treatment' variable* that has *not* been influenced by other factors and is related to the original treatment of interest. For example, the original exposure to smoking can not be controlled by the researcher, but the related variable 'encourage to stop smoking' can be controlled by randomization before collecting the data on smoking.[19] In this situation the method of *instrumental variables* can be used, where the 'alternative treatment variable' is called the instrument or instrumental variable. For treatments with only two classes, this is similar to saying that it is not necessary for the treated group to contain *only treated* individuals and for the untreated group to contain *only untreated* individuals in order to compare these groups. The amount of 'noise' or better, the association between the original and alternative treatment groups, should be known in order to use this method for validly assessing treatment effects. Application of this method is not limited to situations in which a suitable instrumental variable has to be *created* by the researcher (as in the previous example of encouragement to stop smoking), but also in situations where a variable is available in the data and can be used as instrumental variable. Examples of such variables are the distance to a hospital that performed cardiac catheterizations[21] and the physician-specific prescribing preference to certain drugs.[22]

When the assumptions of this method are fulfilled it has the potential to adjust for *all confounders*, whether observed or not. This clearly distinguishes this approach from the other two in which it is only possible to adjust for observed confounders.

STRENGTHS AND LIMITATIONS OF ADJUSTMENT METHODS

The first approach in which the factor of interest and all possible confounders are included in a regression model, is still the standard in observational studies, while the methods of propensity scores and instrumental variables can be considered as *alternative methods* to adjust for confounding. Whenever a conventional regression-based technique can be adopted, the method of propensity scores is also applicable. This is not the case for instrumental variable methods, because at least one suitable instrumental variable is needed. An important strength of both alternative methods is that these are developed with a randomized controlled experiment in mind. That means that before relating treatment to outcome, these methods direct their attention towards the relationship between treatment and potential confounders. An advantage of regression-based methods is that also the effects of other factors on the outcome can be estimated instead of just the effect of treatment. However, when the focus is to estimate a valid treatment effect, the effect estimates of other factors are usually not of interest.

A specific advantage of propensity score methods is its *transparency* on the similarity of treatment groups, which informs the user whether the method was successful in creating this similarity. Another advantage is the *larger number of confounders* that can be adjusted for in the analysis compared to regression-based methods, because only the factor of interest and the propensity score are used in the final outcome model.

In Section 3.1 and 3.2 we pointed at another, frequently overlooked advantage of propensity scores when dealing with a dichotomous outcome. With such data it is common to use logistic regression (or Cox proportional hazards regression with survival data) and express treatment effects as odds ratios (or hazard ratios). The important advantage of propensity scores is that the treatment effect estimator is closer to the true average treatment effect than in logistic or Cox proportional hazards regression. The reason is that in a multivariable logistic regression or Cox proportional hazards regression analysis the effect of treatment, averaged over for instance men and women, is *not equal* to the treatment effect for the whole population, even when the proportion of men is similar in both treatment groups. The difference is systematic and can lead to a serious overestimation of the average treatment effect, especially when the number of prognostic factors is more than 5, the treatment effect is larger than an odds ratio of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95.

The most important advantage of instrumental variable methods is that it adjusts for *all possible confounders*, whether observed or not. In fact, this is a similar objective as in randomized experiments, but is in observational studies rather ambitious. In situations where little information on confounding factors is available, instrumental variables might be considered. The price to pay are the strong assumptions for an instrumental variable to be valid, which means that 1. the instrumental variable should have no direct causal effect on the outcome, 2. its categories are similar with respect to other characteristics of individuals (for instance by randomization), and 3. there exists a relationship between the instrumental variable and the original treatment variable that should not be weak. While it is difficult to test whether the two first assumptions are fulfilled, the third assumption implies also that a strong instrument is wanted. In Section 3.3 we showed that in cases of strong confounding (for instance confounding by indication), a strong valid instrumental variable *can not be found* because of its inherently strong relationship with the confounders. One should either rely on a weak instrument or one is likely to violate the first assumption. This limitation is inherent to this method.

IMPROVEMENT OF PROPENSITY SCORE METHODS

In applying the method of propensity scores (Section 4.1) we recognized that in the methodological literature and in applications the step of creating and checking the balance between treatment groups has not been given much attention. Often no effort has been made to improve a certain propensity score model in order to reach a better balance on prognostic factors and consequently a better estimate of the treatment effect. Therefore, we focused in Chapter 5 on the similarity or balance on confounders between treatment groups. In Section 5.1 and 5.2 we introduced different ways to *measure the amount of balance* in propensity score methods to inform the reader about the quality of the model and to help the researcher to choose among a number of possible propensity score models. One of these measures is the *overlapping coefficient*, which suits the objective of propensity scores: it directly measures the overlap in two distributions and that is exactly what this method tries to achieve. The better the overlap of covariate distributions within strata of the propensity score, the better the balance on those distributions. When compared to a reference distribution of expected values of the overlapping coefficient, we showed that this measure can be used to assess whether sufficient balance on covariates has been reached by propensity score modelling. In simulation studies we showed that there exists an inverse relationship between the weighted average overlapping coefficient (the balance) and the bias that remains after adjustment. For larger data sets this relationship is stronger than for other commonly used methods to check the balance. Compared to simple propensity score models a better reduction of the mean squared error can be reached when the overlapping coefficient is used to improve the propensity score model. We also studied some alternative ways of measuring balance (Section 5.2), i.e. the Kolmogorov-Smirnov distance and the Lévy metric, of which the first has, in the chosen setting, similar characteristics as the overlapping coefficient.

IMPLICATIONS AND FUTURE RESEARCH

This work is relevant to researchers in medical sciences and many other disciplines who face the problem of confounding. We showed the advantages, limitations and applications of two alternative methods to support researchers who want to apply these methods. For the method of instrumental variables we focused on the strength of the correlation between the instrumental variable and treatment: this correlation ought to be strong, but has a maximum which can be a problem for using this method in case of strong confounding. For propensity scores we pointed at a general overlooked advantage of the method compared to logistic regression analysis and Cox proportional hazards regression: it gives in general treatment effect estimates that are closer to the true average treatment effect.

In this thesis there are several leads for future research. One intriguing question is the effect of the *simulation procedure* on the results. In Section 3.2 and 5.1 a method was used in which the average marginal effect was known in advance and where weights were used to create confounding. In Section 5.2 the average marginal effect was calculated using the a model-based procedure described by Austin.[23] More research is needed to compare both simulation procedures in order to assess its influence on the results.

A limitation of our work is that we mainly used *stratification* on the propensity score although also matching, covariate adjustment or inverse probability weighting are also possible ways to use the propensity score. This choice was partly based on the literature and on the study in Section 4.1, but there are still no convincing and conclusive research results that indicate which method is to be preferred. Conceptually stratification or matching is preferred, but more research is needed to investigate the situations in which these methods are favored.

Another limitation is the *fixed number of five strata* that were mostly used in this thesis for stratification on the propensity score. In Section 4.1 also 7 and 10 strata were explored, but the information is still insufficient for general remarks on the use of the number of strata. Obviously, the number of strata should be dependent on the number of observations, but guidelines are not available for determining the optimal number of strata. In Chapter 3 and 5 we conformed ourselves to the convention of using five strata.

In our work we recognized another 'problem' when propensity score methods are evaluated with simulated data. This is the situation in which there are *no or only a few* treated or untreated subjects in one of the strata. Such situations are difficult to handle in simulation studies because then results apply only to a subgroup of the original population. Comparison with another method that estimates a treatment effect for the whole population, will be difficult. In real data sets those situations are common when strong confounding is present and more research is needed to provide guidelines in case strata are 'empty' for one the groups.

In Sections 5.1 and 5.2 we mainly concentrated on *continuous covariates*, although the overlapping coefficient also can be calculated for dichotomous covariates. How the overlapping coefficient will behave in case of a mixture of continuous and dichotomous covariates, could be subject for further research.

For the method of instrumental variables we recognized that application in survival analysis is scarce (Section 4.2). We concentrated on *differences in survival probabilities* and did not estimate the more common hazard ratio using instrumental variables. More research and more medical research examples are needed concerning instrumental variables in general and the use with survival analysis in particular.

Except for Section 2.1 in which many methods are reviewed, we only focused on *propensity scores* and *instrumental variables* as methods to adjust for confounding. The most important reasons for this choice are that these methods are increasingly used in the medical literature in the last decade and that there are several questions to be answered when applying these methods. This choice does not mean that other possible methods are not interesting or not applicable. We mention for example *sensitivity analyses*, *propensity score calibration*, *G-estimation*, propensity score methods when treatment is *time dependent* and the application of propensity score methods in *case-control studies*.

Another subject that is not covered by this thesis is a *simulation study* in which methods of propensity scores and instrumental variables *are compared*. This is an interesting challenge for future research, although practical situations in which both methods are applicable are limited.

CONCLUSION

In conclusion, propensity scores to adjust for confounding have several advantages compared to conventional regression-based techniques. One of these is that the estimated treatment effect is closer to the true average treatment effect, mainly when there are numerous confounders, the treatment effect is substantial and incidence proportions are not too low. Therefore, this method should be considered more often when adjustment for confounding is needed. When propensity scores are considered, more attention should be given to the building of the propensity score model, based on a measure of balance such as the overlapping coefficient. For instrumental variables, researchers should be more aware of the possibility of using this method to adjust for confounding, while awareness of its limitations is equally important. Future research should be directed to further assess in different situations how propensity scores perform and to formulate concrete recommendations on when and how to apply this method. With respect to instrumental variables, it would be helpful when good examples of applications come available as researchers realize that randomization could be used in non-randomized settings. Also simulation studies that show the valid treatment effect estimates that can be reached, are important for future development and application of this method.

REFERENCES

[1] Pearson K. Regression, heredity and panmixia. *Phil Transactions of the Royal Society of London, Series A*, 187:253–318, 1896.

[2] Galton F. Types and their Inheritance [Presidential address, Section H, Anthropology]. *Nature*, 32:507–510, 1885.

[3] Yule GU. On the theory of correlation. *J Royal Stat Society*, 60:812–854, 1897.

[4] Cox DR. The regression analysis of binary sequences. *J Royal Stat Society Series B*, 20:215–242, 1958.

[5] Cox DR. Regression models and life tables. *J Royal Stat Society Series B*, 34:187–220, 1972.

[6] Fisher RA. *Statistical Method for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[7] Fisher RA. The arrangement of field experiments. *J Ministry Agric*, 33:503–513, 1926.

[8] Neyman J. On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Roczniki Nauk Roiniczych, Tom X*, 19:1–51, 1923. Reprinted in *Statistical Science* 1990;5:463-480, with discussion by T. Speed and D. Rubin.

[9] Neyman J. Statistical problems in agricultural experimentation. *Sup J Royal Stat Society*, 2:107–180, 1935.

[10] Peirce CS, Jastrow J. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885. Reprinted in Burks AW (ed.). Collected Papers of Charles Sanders Peirce. Cambridge: Harvard University Press, 1958; 7:1334.

[11] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[12] Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:184203, 1973.

[13] Rubin DB. *The use of matched sampling and regression adjustment in observational studies (Ph.D. Thesis)*. Department of Statistics, Harvard University: Cambridge, 1970.

[14] Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya-A*, 35:417–446, 1973.

[15] Wright PhG. *The tariff on Animal and vegetable oils*. Macmillan, New York, 1928.

[16] Stock JH, Trebbi F. Who invented IV regression? *J of Economic Perspectives*, 17:177–194, 2003.

[17] Reiersøl O. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik, Astronomi och Fysik*, 32:1–119, 1945.

[18] Aldrich J. Reiersøl, Geary and the idea of instrumental variables. *Economic and Social Review*, 24:247–274, 1993.

[19] Permutt Th, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[20] Fisher RA. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.

[21] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*, 272:859–866, 1994.

[22] Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiol*, 17:268275, 2006.

[23] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 2007. On line: DOI: 10.1002/sim.2781.