Paper 261-25

Using Propensity Scores to Adjust for Group Differences: Examples Comparing Alternative Surgical Methods

David J. Pasta, The Lewin Group, San Francisco, CA

ABSTRACT

Comparing the effectiveness of two treatments on nonrandomized groups is difficult because there are almost always baseline differences between the groups. Propensity scores (Rosenbaum and Rubin, Biometrika 1983) are a valuable but underutilized approach for adjusting for group differences. When two groups are being compared, the propensity score can be calculated as the predicted probability of group membership from a logistic regression. It represents the 'propensity' for an observation to be in one group or the other. By adjusting for the value of the propensity score in a linear model, one effectively adjusts for any group differences attributable to the variables used to create the propensity score. In addition, the values of the propensity scores can serve as a diagnostic tool to evaluate the comparability of the groups in a quantitative way. In this paper, three practical examples are presented. In each example, propensity scores are used to adjust for differences between nonrandomized groups. Propensity scores were created using SAS System® PROC LOGISTIC or PROC CATMOD. For two examples, a linear model estimated using PROC GLM was used to compare groups; in the third example, Poisson regression was performed using PROC GENMOD. This paper will be of interest to individuals familiar with basic regression and ANOVA regardless of their level of SAS experience.

INTRODUCTION

In many fields, the question the statistician is asked to help answer can be summarized as "Which is better, Treatment A or Treatment B?" The "treatments" might be two alternative pedagogical methods in educational research, two ways of harvesting timber, two different medicines, or two different surgical methods. In some cases, a randomized experiment can be designed to answer the question of relative efficacy in a fairly precise and well-defined way. Sometimes the experiment is even double-blind. In other cases, a randomized experiment is not done, often because of ethical or practical considerations. In those cases, it is common to find differences in the characteristics of the two treatment groups.

What should a statistician do when faced with group differences? One kind of critic will say, "You can't prove anything – the groups are not comparable." Another kind of critic will say, "You can prove anything with statistics" which, of course, strongly implies that you can't prove anything with statistics. In fact, there are a variety of methods designed to "statistically adjust" for group differences. Propensity score methods have much in common with the other methods used to evaluate treatment differences in the face of other differences between groups.

One advantage of propensity score methods is that they can be used both to answer the original question – Which treatment is better? – and also to evaluate how well that question can be answered. Propensity score methods allow you to directly address the question of what ***can*** be learned from the data and what ***can not***.

Propensity score methods are not difficult to implement in practice. The necessary statistical procedures are readily available in SAS/STAT®. Creation of propensity scores will generally be done with the LOGISTIC or the CATMOD procedure. The propensity scores will then generally be used in a linear models procedure such as REG, GLM, or MIXED, or possibly another modeling procedure such as GENMOD.

This paper illustrates the use of propensity scores in three different examples, each of which compares alternative surgical methods. It begins with a brief discussion of the statistical theory underlying propensity scores, but the emphasis is on their use in practice.

WHAT ARE PROPENSITY SCORES?

In statistical terms, propensity scores are the estimated conditional probability that a subject will be assigned to a particular treatment, given a vector of observed covariates. In practice, propensity scores are typically estimated as the predicted probability of group membership from a logistic regression, using characteristics of the patient as predictors. It is also possible for other factors to be included in the propensity score, such as the characteristics of the physician or medical facility. (In other contexts, the characteristics of the physician might correspond to the characteristics of the teacher or technician, and the characteristics of the medical facility might correspond to the characteristics of the school district or geographical location.)

Controlling for the propensity score has the effect, in a very precise way, of controlling for any group differences due to any of the variables included in the creation of the propensity score. Why not, then, just include all those variables as covariates in a statistical model? One reason is for parsimony – keeping the model as simple as possible. Another reason is to better understand the extent to which the treatment groups include similar subjects. A final reason relates to the purpose for including covariates in statistical models. One reason covariates are included is to adjust for any group differences and another is to improve the prediction of the outcome. Those purposes can be in conflict with each other. Including a covariate in the calculation of propensity scores effectively adjusts for any group differences. It may still be useful to include the variable as a predictor in the final model if it significantly predicts the outcome.

Once you have calculated propensity scores, what do you do with them? One of the first things to do is to look at their distribution separately by treatment group. To the extent the predictions of treatment group are very good, there is substantial separation between the treatment groups (that is, little overlap) and comparisons between groups is suspect. If the predictions are not very good, adjusting for the propensity score will have a milder effect.

After examining the distribution of the propensity scores, there are two common approaches to using propensity scores to adjust for group differences. One approach is to use the propensity score to perform matching between groups. This might be a one-to-one or a many-to-one matching and it would be based on ranges of propensity score. This approach is similar to discriminant score matching, which was one of the predecessors to propensity score matching. It suffers from the usual objections to matching, most notably its failure to make use of all the available data.

Another approach, which I prefer, is to use the propensity score as a variable in the prediction model. The simplest model would include treatment group and the (continuous) propensity score. As there is no particular reason to believe that the propensity score would predict the outcome linearly, it is probably more appropriate to use quantiles to divide the propensity score into groups and include it as a categorical variable (include it in the CLASS statement). It is generally sufficient to divide the propensity score into five categories, although with large samples it may be valuable to have as many as ten categories. As mentioned above, it is reasonable to include other variables in this prediction equation to reduce the residual variability of the outcome, even if those variables are already included in the propensity score prediction equation.

The effect of adjusting for the propensity score is to weaken the effect of the treatment variable. If the propensity score is not a good predictor of treatment group, including it will have very little impact on the estimated treatment effect. If the propensity score is a nearly perfect predictor of treatment group, including it will greatly weaken the statistical significance of the estimated treatment effect (although it may not greatly affect the magnitude of the estimated treatment effect).

For additional information on the theoretical underpinnings of propensity scores, see the two papers by Rosenbaum and Rubin (Biometrika 1983 and JASA 1984) and the references therein.

EXAMPLE 1: LAPAROSCOPY VERSUS LAPAROTOMY FOR TREATMENT OF ENDOMETRIOSIS

Women suffering from endometriosis have reduced fertility. Infertility specialists commonly treat moderate and severe endometriosis by surgical eradication, either during laparoscopy (minimally invasive surgery using endoscopes and lasers) or laparotomy (open abdominal surgery). The research question in this example study was whether pregnancy rates were different for the two surgical approaches.

The study population consisted of 220 infertile women with moderate or severe endometriosis. There were 119 laparoscopy patients and 101 laparotomy patients. A variety of information about the women was available for the creation of propensity scores, including medical history and clinical assessments both pre- and post-treatment. Propensity scores were computed as the predicted probability that a patient is treated by laparotomy as opposed to laparoscopy using logistic regression. A total of 40 coded surgical variables were included in the logistic regression, including endometriosis scores and subscores, information about the percentage and density of adhesions, and function scores for the tubes, fimbria, and ovaries.

Some variables were found to improve the prediction of pregnancy rates even though they were not material contributors to group differences, and were included in the prediction equation in addition to propensity scores and treatment group. The variables were prior therapeutic abortion, medical treatment for endometriosis, and ovulation induction.

Life table analysis of pregnancy rates without taking account of group differences revealed that laparoscopy patients had somewhat better pregnancy rates than laparotomy patients.

However, there were substantial differences between the two groups. As expected, laparotomy patients had more severe disease and less functional tubes, fimbria, and ovaries. The differences were statistically significant, sometimes dramatically so.

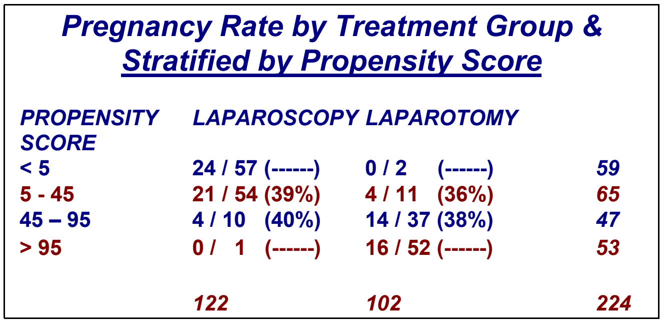
Life Table Analysis of Pregnancy Rates

	SCP (119)	LAP (101)	Total (220)
	mean (SEM)	*mean (SEM)*	*mean (SEM)*
1 year	28.1 (4.4)	25.0 (5.0)	26.8 (3.3)
2 years	47.9 (5.4)	41.8 (6.7)	45.7 (4.3)
3 years	54.4 (6.4)	51.0 (8.2)	53.6 (5.2)

Results Population Characteristics

VARIABLE	LAPAROSCOPY	LAPAROTOMY	P-VALUE
N	122	102	
AGE YEARS	34.5 (4.3)	33.4 (3.9)	0.0505
INFERTILE	4.3 (3.4)	4.1 (3.0)	0.58
GRAVIDITY	0.68 (0.95)	0.64 (1.0)	0.74
FOLLOW-UP	448 (379)	561 (540)	0.067
SIRE TOTAL	1.2 (2.1)	1.0 (1.5)	0.47
SIRE PARTNER	0.94 (2.0)	0.45 (1.1)	0.031
R-AFS TOTAL	35.2 (20.1)	50.7 (30.7)	0.0001
R-AFS ADHESIONS	12.8 (13.0)	25.7 (18.7)	0.0001
R-AFS ENDOMETRIOSIS	22.4 (17.0)	25.0 (21.8)	0.33
LF SCORE BEFORE SURGERY	3.7(1.4)	1.9 (1.6)	0.0001
LF SCORE AFTER SURGERY	5.4 (1.4)	4.4 (1.5)	0.0001

The distribution of the propensity scores indicated that for the vast majority of patients the surgical method could be accurately predicted. In other words, the measured characteristics of the patient predicted which surgical method would be used with nearly 100% accuracy. If the prediction were 100% accurate, that would imply that the two groups could not be fairly compared. Any level below 100%, though, allows at least some comparisons to be undertaken. It seemed reasonable to set aside the patients with extreme predicted probabilities (fully half of the patients had a propensity score less than 5% or over 95%), over 97% of whom were correctly classified. The remaining patients were divided into two strata based on propensity score (5 to 45 versus 45 to 95). The pregnancy rates were very similar for the two groups but still slightly favored laparoscopy. We can make only the weak statement that we are 95% confident that laparoscopy pregnancy rates are at least 80% as high as laparotomy pregnancy rates.



Pregnancy Rate by Treatment Group & Stratified by Propensity Score

PROPENSITY SCORE	LAPAROSCOPY	LAPAROTOMY	
< 5	24 / 57 (------)	0 / 2 (------)	59
5 - 45	21 / 54 (39%)	4 / 11 (36%)	65
45 – 95	4 / 10 (40%)	14 / 37 (38%)	47
> 95	0 / 1 (------)	16 / 52 (------)	53
	122	102	224

One of the questions that arises is how large a sample would be required to demonstrate any statistically significant difference between the groups after adjusting for propensity scores. We calculated that even if laparoscopy pregnancy rates are 50% higher than laparotomy, the sample size needed for statistical significance (p<0.05) in the face of such an extreme distribution of propensity scores is 650 to 1000 patients. In addition, there is some evidence that the propensity scores are becoming more extreme over time.

SUMMARY AND DISCUSSION: EXAMPLE 1

The propensity score can be used to adjust for differences between treatment groups. We compared treatment groups while controlling for the propensity score and for additional covariates which predicted pregnancy rates. Including the propensity score as a covariate tends to stabilize the estimated treatment effect, but increase its uncertainty (standard error). This is equivalent to reducing the effective sample size. In this example, we were very successful in predicting treatment group for about half the patients. Those patients are effectively eliminated from comparisons between the two groups.

A much larger sample would be needed to detect even a much larger difference between groups. To at least some extent, medical practice changes over time to reflect the conventional wisdom about the better surgical technique for patients with a given medical history and characteristics. This can make it essentially impossible to compare surgical alternatives except early in their history when they are still controversial.

EXAMPLE 2: THREE METHODS OF SURGICAL HYSTERECTOMY

There are three surgical methods commonly used to perform elective hysterectomy (removal of the uterus). The most common method is abdominal surgery, which requires an abdominal incision and substantial recovery time. More recently, surgeons are performing vaginal hysterectomy, which is a much less invasive procedure. There is also an even newer procedure that is intermediate between abdominal and vaginal hysterectomy, called laparoscopically assisted vaginal hysterectomy or LAVH. The research question was to compare the treatment groups on health care utilization and costs and the health-related quality of life of the patients. Of perhaps special interest in the analysis was the use of survival analysis to compare the speed of return to normal activity.

The study was performed at a large staff-model health maintenance organization in northern California (Kaiser). Women planning to undergo surgical hysterectomy were invited to participate in the study. Because the number of abdominal hysterectomies was so much larger than for the other two groups, only 30% of the abdominal hysterectomy patients were approached in order to keep the sample size more nearly balanced. The study lasted until 28 days post-surgery and included the collection of interviewer-administered patient information and medical records from the electronic records at the HMO. The primary endpoints were measured at days 7, 14, and 28 post-surgery.

There were, as expected, substantial differences between groups. Propensity scores were calculated using logistic regression. The patient characteristics found to be statistically significant predictors of treatment group were diagnosis of fibroids, number of fibroids, diagnosis of a uterine mass, uterine prolapse, uterine weight, parity, and age.

A difficult problem was whether to include in the propensity score variables representing the "style" of the surgeon. Some surgeons never perform LAVH. As potential patients, we would like to believe that patients “ideally suited for LAVH” who present to a surgeon who does not perform LAVH would be referred to a surgeon who does perform LAVH. In that case, it is not relevant that the surgeon does not perform LAVH. On the other hand, patients who are borderline choices between two surgical approaches are likely to have the type of surgery that the physician “usually” uses. After much deliberation, we decided to compute propensity scores both with and without surgeon "style” variables. We had the percentages of each type of surgery that surgeon had performed over the past year. We included two of those percentages (because the three percentages summed to 100%, the third one was redundant) along with a binary variable indicating whether the surgeon had done any LAVH in the past year.

We found statistically significant differences among treatment groups on health-related quality of life measures as well as on costs, even after adjusting for propensity scores and other key demographic variables. The results were similar for the two versions of the propensity score, whether or not the surgeon’s “style” variables were included.

An interesting feature of the analysis was the use of survival analysis methods on an interview question about return to normal activity. Patients were asked at each time point to compare their current activity level to their pre-surgery level on a five point scale, with the midpoint being the same as before surgery. We defined the date of return to normal activity as the interview at which the patient first provided a response indicating activity was as high or higher than before surgery. Some patients were lost to follow-up or had not returned to normal activity by day 28 and were therefore censored. Accordingly, we used survival methods to estimate the cumulative percentage of each group who had returned to normal activity and used the log-rank test to compare the speed of return to normal activity.

For additional details, see Van Den Eeden et al. (1998).

SUMMARY AND DISCUSSION: EXAMPLE 2

Propensity scores can be used even with more than two treatment groups. Including characteristics of the treating individual or organization is likely to overadjust for group differences. It may be worthwhile to compare treatment effects with no adjustment, adjusted for “conservative” propensity scores (based on the characteristics of the subject), and adjusted for “liberal” propensity scores (based on characteristics of the treating individual or organization as well as the subject). Survival techniques should be used whenever censoring is a concern, even if the outcome is not one generally associated with survival methods.

EXAMPLE 3: MODELING RARE EVENTS USING POISSON REGRESSION

A retrospective analysis of medical and pharmacy claims was designed to compare the frequency of rare cardiac events across five treatment arms, including one untreated arm. A total of 23,000 patients were included. The study included three years of data and considered both primary prevention (avoiding a cardiac event in a patient with no history of a previous cardiac event) and secondary prevention (avoiding an event in patients with a previous cardiac event either before or during the study). The outcomes analyzed were revascularization (CABG or PTCA), stroke, myocardial infarction, unstable angina requiring hospitalization, and any of the above.

Three patient cohorts were considered: (1) all patients with continuous coverage and at least six months of data and some evidence of a disorder; (2) the primary prevention cohort, the subset with no history of a cardiac event; and (3) the secondary prevention cohort, the subset with a prior event or an event during the study period. Note it is possible for patients to be in both the primary prevention cohort and the secondary prevention cohort. Events within 30 days of each other were counted as a single event based on an order of precedence (revascularization, stroke, MI, angina).

Propensity scores were formed from age (in five categories), gender, comorbidities, concomitant medications (anti-hypertensive and other cardiac), and type of insurance (fee-forservice versus other types). There were too many separate comorbidities and drugs included to allow all of them to be included in the propensity score; it would be easy to capitalize on chance differences between groups. Instead, the comorbidities and drugs were selected based on regression models that used backward selection to determine which conditions or medications impacted resource use. This was considered a test of pertinence.

The propensity scores were calculated using PROC CATMOD. The response was specified as logit and the treatment group was modeled as a function of gender, age, insurance, and comorbidity and drug dummy variables. For each unique combination of independent variables (profile), CATMOD yields a set of probabilities for each treatment arm. These probabilities reflect the distribution of patients by treatment and can be merged back into the full data set by the profile variables.

The analysis of the events was performed using Poisson regression as implemented in PROC GENMOD. For each of the outcomes, the probability of an event occurring by cohort was estimated separately for six-month time intervals. The analyses were run with and without propensity scores to see the effect of the adjustment.

Other factors included were the severity of illness according to the Ambulatory Care Group (ACG), the observation period (six months to two years), and the months of drug compliance (0 to 36 months). The GENMOD code for the Poisson regression is as follows:

```
proc genmod data=temp1;
model depv = tx_a tx_b tx_c tx_d
             pr_a pr_b pr_c pr_d
             acg pritime comply
  / dist=poisson
    link=log
    obstats;
contrast ‘a-b’ tx_a 1 tx_b -1;
contrast ‘a-c’ tx_a 1 tx_c -1;
contrast ‘a-d’ tx_a 1 tx_d -1;
contrast ‘b-c’ tx_b 1 tx_c -1;
contrast ‘b-d’ tx_b 1 tx_d -1;
contrast ‘c-d’ tx_c 1 tx_d -1;
make ‘obstats’ out=temp2 noprint;
run;
```

The calculation of appropriate group comparisons is somewhat problematic, as GENMOD does not include the equivalent of adjusted means (like LSMEANS). See Pasta, Cisternas, and Williamson (WUSS 1998). The approach used here is the one referred to as “predicted value of the mean” which is ***not*** the recommended method, which is referred to as the “mean of the predicted values.” It is, however, much easier to calculate for large data files. For the “predicted value of the mean” approach, one simply creates one “dummy” observation for each treatment group and sets the dependent variables to missing. The predictors (covariates) are all set at their respective means, and each observation is assigned to one of the treatment groups.

Comparisons among the four treated groups were obtained from likelihood ratio chi-squares. Comparisons of each treated group to the untreated group were obtained from parameter estimate chi-squares. Because the untreated group included about 50% of the sample, there is much greater statistical power for differences involving that group. The results showed no statistically significant differences among treated groups but significant differences between the untreated group and the treated groups. The propensity score adjustment was modest but in the expected direction.

Results

Event count		**p-values**			
Revascularization	**Adj Mean**	**B**	**C**	**D**	**No Tx**
➢ **A**	**.007**	**.66**	**.89**	**.82**	**<.001**
➢ **B**	**.007**		**.92**	**.58**	**<.001**
➢ **C**	**.005**			**.78**	**.041**
➢ **D**	**.006**				**.003**
➢ **No Tx**	**.004**				

♦ ***adjusting for propensity scores, ACG, observation time and compliance***

SUMMARY AND DISCUSSION: EXAMPLE 3

Propensity scores can be used with large data files and in the presence of many potentially spurious variables as long as some care is exercised. Here, we reduced candidate comorbidity and drug variables to those that showed significant association with costs. For events occurring at the same or nearly the same time, a hierarchy of precedence was used to break ties. Poisson regression is easy to perform with PROC GENMOD. The absence of adjusted means from GENMOD may require the use of another method to get a similar measure of group differences. Although the "predicted value of the mean" method is illustrated here, the "mean of the predicted values" method is recommended.

CONCLUSION

Propensity scores are an underutilized method for controlling for group differences. They are calculated easily using logistic regression procedures such as LOGISTIC or CATMOD. They provide not only a way of adjusting for group differences but also a diagnostic tool that shows the extent to which treatment groups are comparable.

REFERENCES

Pasta, David J., Cisternas, Miriam G., Williamson, Cynthia L. (1998), "Estimating Standard Errors of Treatment Effects for Probit Models and for Linear Models of Log-Transformed Variables using PROC IML," *Proceedings of the 6th Annual Western Users of SAS® Software Regional Users Group Conference,* 211-216.

Rosenbaum, Paul R. and Rubin, Donald B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41-55.

Rosenbaum, Paul R. and Rubin, Donald B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79(35), 516-524.

Van Den Eeden SK, Glasser M, Mathias SD, Colwell HH, Pasta DJ, Kunz K. (1998), "Quality of life, health care utilization and costs among women undergoing hysterectomy in a managed care setting. *American Journal of Obstetrics and Gynecology.* 178:91-100.

ACKNOWLEDGMENTS

This research described here was performed in collaboration with others too numerous to list. The presentation owes much to material prepared with Lori Potter of The Lewin Group and previously presented at a workshop on data analysis and at a separate workshop on propensity scores.

CONTACT INFORMATION

David J. Pasta
Senior Director, Data Management and Analysis
The Lewin Group
490 Second Street, Suite 201
San Francisco, CA 94107
(415) 495-8966 (phone)
(415) 495-8969 (fax)
dpasta@lewin.com