Statistica Sinica

ESTIMATION UNDER MODEL UNCERTAINTY

Nicholas T. Longford

SNTL and Imperial College, London, UK

Supplementary Materials

A. Balance in ordinary regression and post-observation design

This section expands on the text presented in Section 2 of the paper.

(Approximate) balance of a background variable X, irrespective of whether observed or not, can be arranged by randomization. It enables us to ignore all covariates (background) with no loss of efficiency. Thus, for estimating the treatment effect $\Delta\mu$, we can dispense with modelling (considering several alternative models) thanks to the design. This can be interpreted as using a *bad* model that ignores some important covariates. However, modelling is essential for estimating some other quantities, such as the expected outcome for a particular value of X and a treatment.

Methods that use matching to estimate a treatment effect in observational studies (Rosenbaum, 2010) seek arrangements in the design (selection of units based on their values of X) that enable the analyst to ignore the values of X. Balance of the within-treatment distributions of the covariates, arranged by subsetting (e.g., forming matched pairs) or weighting, is the key intermediate goal in such an analysis. So, model uncertainty is addressed, and eliminated, by post-

observation design, and dispensed are all the model-related assumptions: distribution of the outcomes, linearity (appropriate scale for X) and even homoscedasticity. Our approach to estimating a treatment effect in an observational study would be more effective if the posited regression model were valid, an assumption that, admittedly, can rarely pass a thorough examination. Post-observation design has a weaker assumption, namely that all the relevant background variables are available, but it usually cannot be confirmed either. Some other differences, related to how the (average) treatment effect is defined, are discussed in an application presented in Section D below.

In our proposal for estimating $\Delta\mu$, the relative importance of the candidate estimators is moderated by the departure of X from group-level balance. We should not ignore the configuration of the values of X in estimation in general. If an observational dataset happens to have a near-balance of covariate A with respect to covariate B, then A is a distraction for estimating some targets related to B. In summary, some invalid models are sometimes useful, and the combination of validity and parsimony, the goal of the information criteria, such as Akaike (1976), Schwartz (1978) and Spiegelhalter et al. (2002), and other derivatives of the likelihood ratio statistic, may sidetrack us from the pursuit of efficiency.

B. Variance estimation

This section expands on the text presented in Section 4 of the paper.

Composite estimation of the residual variance σ^2 in ordinary regression does not have the potential of its counterpart for linear predictors ($\mathbf{x}_0\boldsymbol{\beta}$) when the candidate submodels of M^* are associated with only a few degrees of freedom

more than the unbiased estimator $\hat{\sigma}^2_*$. This is borne out by analytical derivations.

Let $\hat{\sigma}^2_k$ be the established estimator of σ^2 for model k in a basis of models nested from M_0 to $M_K = M^*$. Then $\hat{\sigma}^2_k$ has bias $b_k = \lambda_k/(n-p_k)$ and variance

$$v_k \;=\; \frac{2\sigma^4}{n-p_k} + \frac{4\lambda_k\,\sigma^2}{(n-p_k)^2}\,,$$

where $\lambda_k = \boldsymbol{\beta}^\top_{k,\text{out}}\,\mathbf{X}^\top_{k,\text{out}}\,\mathbf{X}_{k,\text{out}}\,\boldsymbol{\beta}_{k,\text{out}}$ and p_k is the number of parameters in $\boldsymbol{\beta}_{k,\text{in}}$; see Seber (1977), Theorem 1.8. The (additional) indices 'in' and 'out' refer to the elements of the vector or columns of the matrix linked to the covariates included in and excluded from a model. Further, the covariance of basis estimators $\hat{\sigma}^2_h$ and $\hat{\sigma}^2_k$, $h < k$, is

$$v_{hk} \;=\; \frac{2\sigma^4}{n-p_h} + \frac{4\lambda_k\,\sigma^2}{(n-p_h)(n-p_k)}\,.$$

The variance matrix of the basis estimators $\hat{\boldsymbol{\sigma}}^2$ has the same pattern, $v_{hk} = v_{\max(h,k)}$, as its counterpart for $\hat{\boldsymbol{\theta}}$ in Section 4.1 of the paper only approximately when $n \gg p_k$. Unlike in estimating $\mathbf{x}_0\boldsymbol{\beta}$, the bias terms λ_k accumulate with model simplification, and they inflate both the bias and the sampling variance of $\hat{\sigma}^2_k$. The results from Section 4.1 carry over only subject to some approximation, leading to the estimator

$$\tilde{\sigma}^2 \;=\; \hat{\sigma}^2_K + \frac{\hat{b}_0}{1+R}\,,$$

where $R = \sum_{k=1}^K \Delta b_k^2/\Delta v_k$, with $b_0 = \lambda_0/(n-p_0)$, $\Delta b_k = \lambda_k/(n-p_k) - \lambda_{k+1}/(n-p_{k+1})$ and $\Delta v_k = v_{k+1} - v_k$. In the terms contributing to R, the squared bias in the numerator is a quartic function of the regression parameters and is often far greater than the denominator. Thus, R is large and $\tilde{\sigma}^2 \doteq \hat{\sigma}^2_K$. The same conclusion is arrived at even with the simple composition of $\hat{\sigma}^2_0$ and $\hat{\sigma}^2_*$ when

$K \ll n$. In brief, increasing the number of degrees of freedom in $\hat{\sigma}^2$ is too risky, given the uncertainty about $\boldsymbol{\beta}$ and its propagation in both the bias and the variance of a basis estimator.

In simple composition of $\hat{\sigma}_0^2$ and $\hat{\sigma}_*^2$, the optimal coefficient for $\hat{\sigma}_0^2$ is

$$c^* \;=\; \frac{K(n-1)\sigma^4}{K(n-1)\sigma^4 + (n-p)(\lambda_0^2 + 4\sigma^2\lambda_0)}\,.$$

C. Optimal composition of simple regressions (basis B)

This section complements Section 4.1 of the paper.

Suppose the basis comprises the estimators based on simple regressions, estimator $\hat{\theta}_k$ based on covariate X_k, and the intercept-only model 0 with $\hat{\theta}_0 = \bar{y}$. The covariates are pairwise orthogonal. Denote $\nabla b_k = b_0 - b_k$ and $\nabla v_k = v_k - v_0$. The diagonal entries of $\mathbf{V} = \mathrm{var}(\hat{\boldsymbol{\theta}})$ are $v_0, v_1, \ldots, v_K$ and every off-diagonal element is equal to $v_0 = \sigma^2/n$. The inverse of $\mathbf{V}$ is an arrow-shaped matrix; its diagonal elements are $u_{00} = 1/v_0 + 1/\nabla v_1 + \cdots + 1/\nabla v_K$ and $u_{kk} = 1/\nabla v_k$, $k = 1, \ldots, K$, and off-diagonal elements are all equal to zero except for the elements in row and column of $\hat{\theta}_0$, which are equal to $u_{0k} = u_{k0} = -1/\nabla v_k$, $k = 1, \ldots, K$. For example,

$$\mathbf{V}^{-1} \;=\; \begin{pmatrix} \frac{1}{v_0} + \frac{1}{\nabla v_1} + \frac{1}{\nabla v_2} + \frac{1}{\nabla v_3} & -\frac{1}{\nabla v_1} & -\frac{1}{\nabla v_2} & -\frac{1}{\nabla v_3} \\ -\frac{1}{\nabla v_1} & \frac{1}{\nabla v_1} & 0 & 0 \\ -\frac{1}{\nabla v_2} & 0 & \frac{1}{\nabla v_2} & 0 \\ -\frac{1}{\nabla v_3} & 0 & 0 & \frac{1}{\nabla v_3} \end{pmatrix}$$

for $K = 3$. Hence $B_0 = 1/v_0$, $B_1 = b_0/v_0$ and $B_2 = R' + b_0^2/v_0$, where $R' = r'_1 + \cdots + r'_K$ and $r'_k = \nabla b_k^2/\nabla v_k$; compare with R in Section 4.1. In fact, for complete bases A and B, which generate identical spaces, $R = R'$. The optimal

composition is $\mathbf{c}^{*\top}\hat{\boldsymbol{\theta}}$, with

$$\mathbf{c}^* \;=\; \frac{1}{1+R'}\left(1+R'+b_0\sum_{k=1}^{K}\rho'_k\,,\;\; -b_0\,\rho'_1\,,\;\; \ldots\,,\;\; -b_0\,\rho'_K\right)^{\top}\,,$$

where $\rho'_k = \nabla b_k/\nabla v_k$. Substitution of $\hat{\theta}_0 - \hat{\theta}_k = \nabla\hat{b}_k$ in $\hat{\boldsymbol{\theta}}$ and replacing B_0, B_1 and B_2 by their estimates yield the estimator $\tilde{\theta} \;=\; \hat{\theta}_0 - \hat{R}'\hat{b}_0/(1+\hat{R}')$, leading to a discussion similar to that following equation (4.1) of the paper. In particular, $\mathrm{MSE}^{\circ}(\tilde{\theta};\theta)$ would attain its minimum for $R'_* = b_0^2/(v^* - v_0)$, where $v^* = \mathrm{var}(\hat{\theta}^*)$, and it corresponds to the optimal composition of $\hat{\theta}_0$ and $\hat{\theta}^*$. The presence of model M_0 in the basis is essential for the analytical convenience of the arrow shape of $\mathbf{V}^{-1}$. However, analytical tractability is retained if a nontrivial model M_0 is declared, so long as it is a submodel of all the other basis models and they differ from M_0 by disjoint sets of covariates, so that $\nabla\hat{b}_k$, $k = 1, \ldots, K$, are mutually independent. When a covariate h is included in M_0, every remaining model in the basis has to be supplemented with covariate h. The value of v_0 is increased but R' is reduced by the eliminated term r'_h.

D. Composition and propensity matching with prostate cancer data

This section is related to Section 5.1 of the paper.

We contrast composite estimation with propensity matching analysis on the example of estimating the (causal) effect of the variable svi. In the counterfactual formulation, our target is the average difference of the outcomes of the 21 subjects with svi equal to 1, if their values of svi were all switched to zero, without affecting the values of any other background variables, or the outcomes for any other subjects. First, every subject with svi = 1 is matched with a similar subject

with svi = 0. No subject can be reused in the matched pairs, but some may end up without a match. Propensity matching, Rosenbaum and Rubin (1983) and Rosenbaum and Rubin (1985), yields the scores plotted in Figure 1. They are the linear predictors fitted by the logistic regression of svi on the other background variables. The vertical dashes delimit the overlap of the ranges of the scores, with a bit of leeway left for inexact matching. There are only $32 + 14$ subjects (47%) within the range from the respective groups 0 and 1. Further, the distributions of the two groups in this range are uneven (excess of group 0 around -3 and shortage around 1), and not all the 14 subjects with svi = 1 could be paired. Dots connect the 11 pairs formed by caliper matching (Cochran and Rubin, 1973) with half-width of 0.2. Ten units in group 1 (marked by large gray discs) are not matched, three of them inside the overlap. In summary, very few subjects in the data are useful for comparing the two groups defined by the values of svi. With 11 matched pairs, the expected value of the standard error is $\sigma/\sqrt{5.5} \doteq 0.30$.

In the regression formulation, the parameter of interest is the slope on svi, without a reference to a particular set of subjects. The subjects are represented in both analyses only through their vectors of covariates $\mathbf{x}$, which in the regression approach cancel out owing to linearity. The composite estimate with the intercept-only model 0 is 0.766, with standard error 0.232, similar to the estimate based on the the unconstrained model, 0.695 (0.233). Composite estimation, or regression in general, appears to be more efficient than propensity matching analysis. However, the assumption that model M^* is valid is rather onerous because estimation of the relevant parameter entails substantial extrapolation, and this

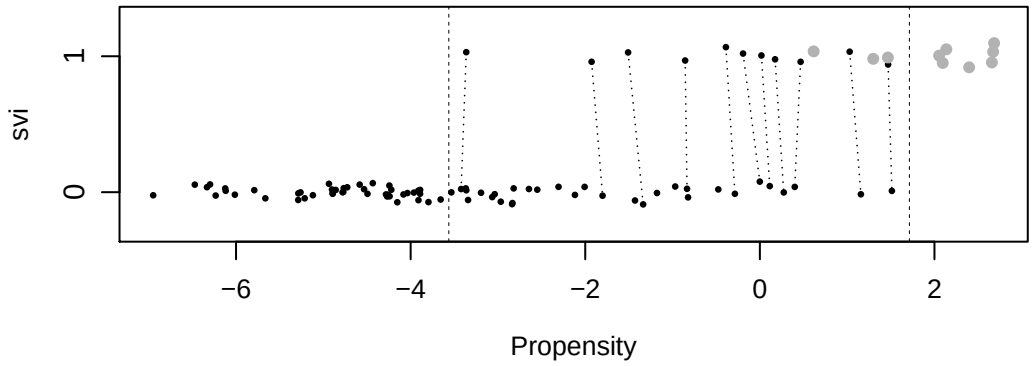

Figure 1: Propensity scores (values of the fitted linear predictor $\mathbf{X}\hat{\boldsymbol{\beta}}$ in logistic regression) for estimating the causal effect of svi on lpsa for subjects with svi = 1. Vertical noise is added to the points to distinguish subjects with very similar values of $\mathbf{X}\hat{\boldsymbol{\beta}}$. The matched pairs are connected by dots.

may not be apparent in a cursory inspection of the data or of the results. In many settings, the cautious attitude implied by the causal analysis, and placing no faith in a model (M^*) in particular, is well justified, although it has to be supplemented by an extrapolation from the realized sample to the relevant population, a problematic task with a small (selected) subset of a dataset. Note that the outcomes have no role in the selection of the propensity model, nor in forming matched pairs. Thus, no ps issues arise.

E. Guide dogs for the blind

This section presents an application related to Section 6 of the paper.

A guide dog is invaluable for a blind person when the dog is well trained, the person well instructed and well disposed, and the two are well matched. A survey was conducted to assess factors that contribute to the successful match of a blind person with a guide dog. The dataset is from a donor who wishes to

Table 1: Summary of the data about guide dogs.

	Success		Living alone		Past dog owner		
Age (years)	No	Yes	No	Yes	No	Yes	All
20 – 29	285	1089	698	676	425	949	1374
30 – 39	125	545	352	318	261	409	670
40 – 49	75	441	333	183	223	293	516
50 – 59	69	368	340	97	227	210	437
60 +	59	344	357	46	229	174	403
All	613	2787	2080	1320	1365	2035	3400

remain anonymous. The variables recorded are success of the match, person's age (in years, between 20 and 70), whether the person lives alone and whether he or she has previously owned a dog, not necessarily a guide dog. The three variables other than age are binary. Success is defined as retaining the dog for more than 90 days. Most of the failed matches lasted for less than three weeks. A person with a failure appears in the dataset only once, with the first match.

The marginal success rate in the 3400 matches (records) is 82.0%; 38.8% of the persons live alone and 59.9% have previously owned a dog. The mean age in the data is 37.7 years and the median is 34 years. Table 1 gives some of the data summaries (counts). It shows that the rate of living alone declines with age and past dog ownership among older persons is less prevalent. The failure rate declines with age (from 21% in the twenties to 14.5% in the sixties). We estimate by logistic regression the advantage that can be attributed to past dog ownership.

In the logistic regression fit with 'linear' age and no interactions, the covariate

'living alone' (LA) is nominally not significant (t ratio 0.45) and the other two covariates are highly significant. Model selection would conclude that LA should be dropped and the two other covariates retained. The model with all three covariates which, for illustration, we regard as valid, yields the estimated log-odds of success with respect to past dog ownership 0.959, with estimated standard error 0.095. If the model with LA dropped were adopted unconditionally, the estimate would be 0.950 (0.086). The composition of these two estimators yields the estimate 0.952, with the shrinkage coefficient (the weight on the reduced model) 0.837. The estimated rMSE is 0.0935, but it should be adjusted for the deception due to ignoring the uncertainty about the extent of shrinkage. Composition of the unconstrained model with other submodels is not useful; next to no shrinkage takes place ($\hat{c} < 0.05$) and the estimated standard error is 0.095, not adjusted for deception. In conclusion, we should estimate the log-odds of success with respect to past dog ownership by the unconstrained model, because the variance reduction is not worth the bias likely to be incurred by composition.

Caliper matching analysis with guide dogs data

In the Rubin's causal model (Holland, 1986), previous dog ownership is a cause (treatment) that could be, at least in principle, altered (manipulated) in advance of the onset of blindness. The effect of this cause is estimated by forming a set of matched pairs; each pair comprises a subject with previous ownership and one without. The selection of such a matched subset for analysis is an alternative to adjusting for confounders (age and LA) by logistic regression. The paucity of background information is a problem common to the two approaches.

The result of caliper matching is a set of 1323 pairs of subjects (a match is not found for 42 subjects in the focal group), so $3400 - 2646 = 754$ records (22.2%) are discarded. The table of successes for them is

	Failure	Success
Not past owner	343	980
Past dog owner	156	1167

so the estimated difference of the rates is 14.1% and the log-odds ratio is estimated by 0.963. In 100 replications of the matching process, the number of matched pairs was in the range 1288 – 1349 (mean 1316.4 and standard deviation 13.4).

The replicate estimates of the difference of the rates have mean 13.7% and standard deviation 0.6%. The replicate estimates of the log-odds ratio have mean 0.929 and standard deviation 0.049. These results differ somewhat from the regression analysis, but the targets of the two analyses are also different. The regression refers to a superpopulation, whereas the matched-pairs analysis is for the subset of persons in the survey who have not owned a dog in the past. Arguably, the target of the regression analysis is more relevant. On the other hand, the matched-pairs analysis does not rely on linear dependence on age (on the logit scale), and the estimated probability is easier to interpret than the log-odds ratio.

References

Cochran, W. G., and Rubin, D. B. (1973). Controlling bias in observational studies. A review. *Sankhya 35*, 417–466.

REFERENCES

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–970.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* AU–*19*, 716–722.

Rosenbaum, P. R. (2010). *Design of Observational Studies.* Springer-Verlag, New York.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*, 33–38.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Seber, G. A. F. (1977) *Linear Regression Analysis.* Wiley, New York, NY.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society* Series B *64*, 583–639.