

Comments on McCandless et. al.
Bayesian Propensity Score Analysis for
Observational Data

Statistics in Medicine, **28**, 94-112, 2009

- ▶ Bayesian formulation of propensity score models
- ▶ Example
- ▶ Comparison of classical and Bayesian procedures
- ▶ Computation

Setting is a binary treatment, binary response problem; real example compares cardiac outcomes amongst patients prescribed/not prescribed statins, in a large observational hospital-originated study in Ontario.

Why did this paper get published ?

- ▶ First (published) paper using Bayesian inferential methods for propensity scores
- ▶ Some insights into performance of Bayesian estimators versus classical estimators
- ▶ Interesting real example

Central thesis: Bayesian estimators allow for propagation of uncertainty through all components of the model, rather than using point estimates of propensity scores favoured in the classical analysis.

Model details

From p.98: X records treatment, Y records outcome, C records potential confounders.

- **Observation model:** Logistic regression

$$\text{logit } \mathbb{P}[Y = 1|X, C] = \beta X + \xi^\top g(z(C, \gamma))$$

where $z(C, \gamma)$ is the propensity score

- **Propensity score model:** Logistic regression

$$\text{logit } \mathbb{P}[X = 1|C] = \gamma^\top C$$

The propensity score is denoted Z , a random variable defined by $Z = z(C, \gamma)$. The *observed* propensity score is $z = z(c, \gamma)$.

The effect of interest is β , the log-odds ratio

$$\beta = \log \left\{ \frac{\mathbb{P}[Y = 1|X = 1, C]/\mathbb{P}[Y = 0|X = 1, C]}{\mathbb{P}[Y = 1|X = 0, C]/\mathbb{P}[Y = 0|X = 0, C]} \right\}$$

Function g is a quintile indicator vector (p. 98)

$$g(z)^{\top} = \begin{cases} [1, 0, 0, 0, 0] & 0 < z < q_1 \\ [1, 1, 0, 0, 0] & q_1 \leq z < q_2 \\ [1, 0, 1, 0, 0] & q_2 \leq z < q_3 \\ [1, 0, 0, 1, 0] & q_3 \leq z < q_4 \\ [1, 0, 0, 0, 1] & q_4 < z < 1 \end{cases}$$

Model details

Thus in the linear predictor

$$\xi^T g(z) = \begin{cases} \xi_1 & 0 < z < q_1 \\ \xi_1 + \xi_2 & q_1 \leq z < q_2 \\ \xi_1 + \xi_3 & q_2 \leq z < q_3 \\ \xi_1 + \xi_4 & q_3 \leq z < q_4 \\ \xi_1 + \xi_5 & q_4 < z < 1 \end{cases}$$

Note that $q_1, \dots, q_4$ are chosen and **fixed** throughout the analysis.

The authors point out (p. 99) that a more flexible model could be used at this stage, but do not really elaborate on precisely how they would use the more flexible model.

Model details

Under this model specification, within *each* strata of z ,

$$\beta = \log \left\{ \frac{\mathbb{P}[Y = 1|X = 1, z]/\mathbb{P}[Y = 0|X = 1, z]}{\mathbb{P}[Y = 1|X = 0, z]/\mathbb{P}[Y = 0|X = 0, z]} \right\}$$

therefore in fact

$$\beta = \log \left\{ \frac{\mathbb{P}[Y = 1|X = 1]/\mathbb{P}[Y = 0|X = 1]}{\mathbb{P}[Y = 1|X = 0]/\mathbb{P}[Y = 0|X = 0]} \right\}$$

is a more “legitimate” causal parameter, the population log-odds ratio.

Model details

Note: The risk difference (average treatment effect)

$$\mu = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_Z [\mathbb{P}[Y = 1|X = 1, Z] - \mathbb{P}[Y = 1|X = 0, Z]]$$

might be another causal effect of interest. We have

$$\begin{aligned}\beta(Z) &= \mathbb{P}[Y = 1|X = 1, Z] - \mathbb{P}[Y = 1|X = 0, Z] \\ &= \frac{\exp\{\beta + \xi^\top g(Z)\}}{1 + \exp\{\beta + \xi^\top g(Z)\}} - \frac{\exp\{\xi^\top g(Z)\}}{1 + \exp\{\xi^\top g(Z)\}}\end{aligned}$$

The expectation over Z becomes a weighted average:

$$\mu = \sum_{j=1}^5 w_j \beta_j$$

with w_j is the (estimated) probability of being in stratum j , and β_j is the effect size in stratum j .

Estimating the causal effect

Note: Regressing (rather than stratifying) on the propensity score can complicate the subsequent computation of the estimated causal effect

$$\begin{aligned}\beta &= \mathbb{E}_Z [\text{logit } \mathbb{P}[Y = 1|X = 1, Z] - \text{logit } \mathbb{P}[Y = 1|X = 0, Z]] \\ &= \mathbb{E}_Z [\log \psi(Z)]\end{aligned}$$

where

$$\psi(Z) = \frac{\mathbb{P}[Y = 1|X = 1, Z]/\mathbb{P}[Y = 0|X = 1, Z]}{\mathbb{P}[Y = 1|X = 0, Z]/\mathbb{P}[Y = 0|X = 0, Z]}$$

varies continuously with Z .

Estimating the causal effect

In a regression version, with regression function $h(Z, X; \gamma)$,

$$\psi(Z) = \exp\{\beta + h(Z, 1; \gamma) - h(Z, 0; \gamma)\}$$

which does not simplify to $\exp\{\beta\}$ unless $h(Z, X; \gamma)$ does not depend on X (ie there is no interaction).

The usual estimator would be

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n \log \hat{\psi}(\hat{z}_i)$$

This is not really addressed in this paper (but, see Moodie and Stephens (2010), SMMR, *Generalized Propensity Score* paper)

Assumptions

The usual assumptions

- ▶ No unmeasured confounders
- ▶ Experimental treatment assignment assumption
- ▶ Propensity score model correctly specified

are made.

Point 1

In a classical PS analysis, checking for balance is imperative; without balance, unbiased estimation of the causal effect is not guaranteed.

No real attempt is made to check balance here, apart from via Figure 1 (p. 102)

The Propensity Score is a Balancing Score

The original motivation for the propensity score is that it is the coarsest function of the confounders that has the *balancing property*. Under the usual no unmeasured confounders assumption

$$Y(x) \perp X|C \quad \text{for all treatments } x$$

ie counterfactual outcomes are independent of the treatment allocation given the confounders C . It can be shown that

$$Y(x) \perp X|Z \quad \text{for all treatments } x$$

ie counterfactual outcomes are independent of the treatment allocation given the propensity score Z .

The Propensity Score is a Balancing Score

This follows from the balancing property, namely that

$$C \perp \{X = j\} | Z \quad \text{for } j = 0, 1.$$

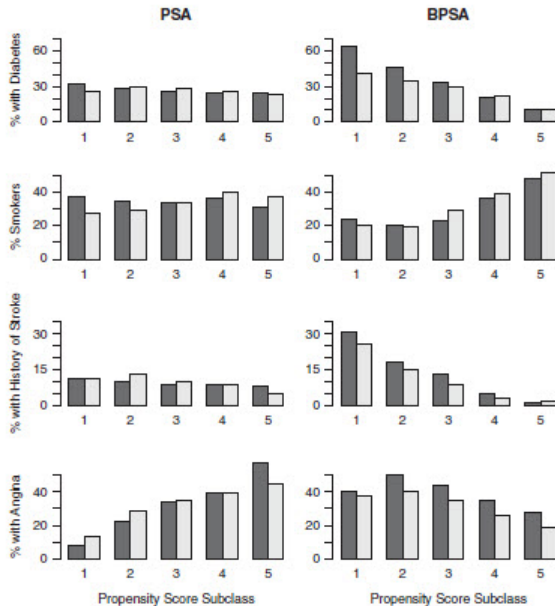
ie for fixed Z (or within strata of Z), the joint covariate distribution is the same for $X = 0$ as for $X = 1$.

Other balancing scores may exist, but the propensity score is the coarsest (ie the partition of the covariate space induced by the propensity score contains all other partitions induced by balancing scores).

Point 2

If we assume the PS is correctly specified, it is automatically a balancing score.

Balanced ?



Balanced ?

It is not clear here whether balance has been achieved on these key covariates. If balance is not achieved, the causal parameters are likely to be estimated with bias.

Note: here the posterior expected propensity score is used

$$\hat{Z}_i^{\text{BPSA}} = \mathbb{E}_{\gamma|\text{data}}[z(C_i, \gamma)] = \int \text{expit}\{\gamma^\top C_i\} p(\gamma|\text{data}) d\gamma$$

computed direct from the posterior samples.

Point 3

Is this really a balancing score ? No evidence is presented to support this implicit assertion. It might be, but the averaging over the posterior makes checking difficult.

Using a point estimate may be preferable.

Analysis

The analysis is based on the likelihood

$$L(\beta, \gamma, \xi; \mathbf{y}, \mathbf{x}, \mathbf{c}) = \prod_{i=1}^n \{f(y_i|x_i, c_i, \beta, \eta)f(x_i|c_i, \gamma)\}$$

where

$$f(y_i|x_i, c_i, \beta, \eta)$$

depends on confounders c_i only through the PS “link” function $g(z(c, \gamma))$.

Point 4

This is just a standard likelihood. No use of the balancing property of the PS is alluded to. This means that the “Bayesian” part of the analysis is not essential (you could use ML)

Analysis: Implications

There is “feedback” between the observation model and the treatment model, that is, the two models are estimated jointly. In a typical classical PS analysis, feedback is not really encouraged.

- ▶ many authors recommend treating the observation and PS model in isolation (ie build the PS model without reference to the outcome data - eg Rubin, AOAS, 2007)
- ▶ precisely what is the right thing to do is the subject of debate; for unbiased estimation, getting the PS model correct is sufficient, irrespective of the outcome model (provided there are no unmeasured confounders).

Point 5

In this likelihood-based setting, the answer is clear; there should be feedback.

Proper priors on the regression/PS parameters are used, quantified by considering impact on the log-odds ratio.

Point 6

It is not legitimate to use flat priors (proportional to 1) for these parameters; if you do, the posterior is improper.

The authors adopt a sensible and pragmatic approach that is not always adopted by authors using Bayesian methods for GLMs.

Posterior distributions are computed using MCMC (some details are given in the appendix). This is pretty straightforward, and convergence is likely to be quite rapid.

Given samples from the posterior distribution, all the required quantities can be computed easily.

As pointed out in the appendix, you have to remember to use the full likelihood when updating the γ parameters (feedback).

EFECT study

The EFECT study recruited patients after hospitalization due to acute myocardial infarction

- ▶ Outcome is death (all causes) within three years of hospital discharge.
- ▶ Treatment is statin **prescription** ($X = 1$) at discharge.
- ▶ Covariates/confounders (demographics, cardiac risk factors, vital signs at admission etc.).
- ▶ $n = 4572$ (p. 96) or 5472 (p. 98). Table 1 indicates 1354 prescribed statin ($X = 1$), 3218 not prescribed ($X = 0$).

Point 7

Is all-cause mortality the right outcome to look at ? Surely cardiac-related mortality would be more sensible ?

Point 8

The analysis addresses the causal effect of statin **prescription**, not the causal therapeutic effect of statins. These two things are different.

Non-compliance is a potential issue; if compliance is confounded by the same confounders, then there is a potential bias in estimating the therapeutic effect of the treatment itself.

Results: Table I (p. 97)

Table I. Baseline characteristics of 4572 patients discharged from hospital following acute myocardial infarction.

	Statin prescribed (<i>n</i> = 1354)	Statin not prescribed (<i>n</i> = 3218)
Characteristic	Number (per cent) or mean ± SD	
<i>Demographic characteristics</i>		
Age (years)	63 ± 12	68 ± 14
Female sex	398 (29)	1201 (37)
<i>Presenting characteristics</i>		
Shock	≤5 (≤1)	24 (1)
<i>AMI risk factors</i>		
Family history of CAD	525 (39)	973 (30)
Diabetes	459 (26)	1060 (26)
CVA/TIA	122 (9)	312 (10)
High BP	548 (48)	1386 (43)
Current smoker	459 (34)	1060 (33)
Hyperlipidaemia	794 (59)	604 (19)
<i>Comorbidities</i>		
Angina	504 (37)	999 (31)
Renal disease	10 (1)	13 (≤1)
<i>Vital signs on admission</i>		
Systolic BP (mmHg)	149 ± 31	148 ± 32
Diastolic BP (mmHg)	85 ± 18	84 ± 18
Heart rate (beats/min)	81 ± 23	84 ± 23
Respiratory rate (beats/min)	20 ± 5	21 ± 6
<i>Laboratory values</i>		
White blood count (cell 10 ⁹ /L)	10 ± 5	10 ± 5
Haemoglobin (g/L)	141 ± 17	137 ± 19
Sodium (mmol/L)	139 ± 3	139 ± 4
Glucose (mmol/L)	9 ± 6	9 ± 5
Creatinine (μmol/L)	101 ± 54	104 ± 60

Results: Table II (p. 105)

Table II. Analysis of the EFFECT data. Point and 95 per cent interval estimates for the regression coefficients β , γ , and ξ , calculated from BPSA or PSA.

Description	Parameter		BPSA		PSA
Statin effect	β	-0.30	(-0.50, -0.10)	-0.36	(-0.54, -0.18)
<i>Regression coefficient for linear predictor g(.)</i>					
y-intercept	ξ_1	0.6	(0.4, 0.9)	-0.1	(-0.3, 0.0)
Subclass number 2	ξ_2	-1.0	(-1.4, -0.7)	-0.8	(-1.0, -0.6)
Subclass number 3	ξ_3	-1.9	(-2.3, -1.6)	-1.4	(-1.7, -1.2)
Subclass number 4	ξ_4	-3.0	(-3.4, -2.6)	-1.7	(-1.9, -1.5)
Subclass number 5	ξ_5	-4.0	(-4.4, -3.6)	-2.2	(-2.5, -1.9)
<i>Demographic characteristics</i>					
Age (years)	γ_1	-0.02	(-0.03, -0.02)	-0.03	(-0.04, -0.02)
Female sex	γ_2	0.02	(-0.05, 0.09)	-0.17	(-0.33, -0.01)
<i>Presenting characteristics</i>					
Shock	γ_3	-0.37	(-1.01, 0.26)	-0.55	(-1.55, 0.45)
<i>AMI risk factors</i>					
Family history of CAD	γ_4	0.05	(-0.02, 0.12)	0.16	(0.02, 0.31)
Diabetes	γ_5	-0.10	(-0.18, -0.02)	-0.05	(-0.22, 0.12)
CVA/TIA	γ_6	-0.09	(-0.19, 0.00)	0.17	(-0.07, 0.40)
High BP	γ_7	0.09	(0.02, 0.16)	0.31	(0.17, 0.44)
Current smoker	γ_8	-0.10	(-0.19, -0.02)	-0.27	(-0.42, -0.12)
<i>Comorbidities</i>					
Angina	γ_{10}	-0.03	(-0.09, 0.04)	0.37	(0.23, 0.51)
Renal disease	γ_{11}	0.08	(-0.59, 0.76)	1.06	(0.01, 2.11)
<i>Vital signs on admission*</i>					
Systolic BP	γ_{12}	0.06	(0.02, 0.11)	0.02	(-0.08, 0.12)
Diastolic BP	γ_{13}	-0.00	(-0.05, 0.04)	-0.02	(-0.12, 0.08)
Heart rate	γ_{14}	-0.09	(-0.12, 0.06)	-0.05	(-0.12, 0.03)
Respiratory rate	γ_{15}	-0.05	(-0.08, -0.02)	-0.06	(-0.13, 0.02)
<i>Laboratory values*</i>					
White blood count	γ_{16}	-0.04	(-0.08, -0.01)	0.00	(-0.07, 0.07)
Haemoglobin	γ_{17}	0.075	(0.04, 0.11)	0.04	(-0.04, 0.12)
Sodium	γ_{18}	0.04	(0.01, 0.08)	0.05	(-0.02, 0.12)
Glucose	γ_{19}	-0.03	(-0.07, 0.01)	0.05	(-0.03, 0.12)
Creatinine	γ_{20}	-0.14	(-0.20, -0.09)	-0.07	(-0.15, 0.02)

*Continuous variables, re-scaled to have mean zero and unit variance.

Results

Description	Parameter		BPSA		PSA
Statin effect	β	-0.30	(-0.50, -0.10)	-0.36	(-0.54, -0.18)
<i>Regression coefficient for linear predictor g{.}</i>					
y-intercept	ζ_1	0.6	(0.4, 0.9)	-0.1	(-0.3, 0.0)
Subclass number 2	ζ_2	-1.0	(-1.4, -0.7)	-0.8	(-1.0, -0.6)
Subclass number 3	ζ_3	-1.9	(-2.3, -1.6)	-1.4	(-1.7, -1.2)
Subclass number 4	ζ_4	-3.0	(-3.4, -2.6)	-1.7	(-1.9, -1.5)
Subclass number 5	ζ_5	-4.0	(-4.4, -3.6)	-2.2	(-2.5, -1.9)

- ▶ Significant statin prescription effect
- ▶ Shrinkage towards zero of estimate of β , with longer credible interval, for BPSA.
- ▶ Monotonic dependence on propensity score strata coefficients (ie upper strata have more negative intercepts) - this may be a “prescribing effect” (p. 100).

The unadjusted (observational) effect size is (p. 96) is $\log 0.5 = -0.693$ with 95% interval $(-0.867, -0.511)$. So the adjusted “causal” effect size is smaller (however, it is not clear that the multiple regression-adjusted estimate is any different).

Point 9

Is there a danger that the confounding has not been accounted for fully ? For example, **hyperlipidemia** at baseline is clearly a strong predictor of treatment, and is arguably a predictor of all-cause mortality in the three years after admission, yet it is not in the propensity score.

This is not an issue **if** statins cure hyperlipidemia with 100 % efficacy.

Point 10

Non-compliance: It is already suspected that treatment allocation favours healthier subjects. Perhaps healthier subjects are also better (or worse) compliers ?

Quintiles

The analysis is carried out with fixed quintiles chosen by performing an ML estimation and using fitted values. This does not produce equal sample sizes in the propensity score strata for the BPSA.

Point 11

Some individuals who were in the same strata under the ML analysis maybe in different strata for the Bayesian analysis. Is this a problem ? It may lead to contradictory results (from a matching perspective).

Propensity score regression seems a safer strategy.
Alternatively, a matching model that clusters propensity scores might be feasible.

Section 4 contains simulation studies, summarized in tables IV-VI.

Perhaps I am misunderstanding something, but

Point 12

Whereas for the causal parameter, performance of the BPSA is as expected (larger variance, coverage about right), the frequentist performance for the other parameters is not good (see pp 105, 107).

There seems to be no reason why this should be the case, unless it is the quintile stratification.

Simulation study

Some questionable statements

- ▶ p. 105: *“Remarkably, ignoring uncertainty in the estimated propensity scores appears to have no adverse effect on coverage probability. PSA interval estimates are shorter on average compared with BPSA, and yet they retain nominal 80 per cent coverage levels. PSA point estimates of β are more efficient and have smaller MSE. Contrary to intuition, substituting estimates in place of the true propensity scores does not yield inferences(,) which are falsely precise.”*

The PSA intervals for β are shorter than, and have the same coverage as, BPSA intervals. We would expect this, and would expect the difference to diminish in large sample sizes.

The BPSA intervals for the nuisance parameters do not exhibit 80% coverage, and are much shorter. It is hard to see how this happens. For the parameter of interest, the difference is negligible - there are only 400 replicates.

- ▶ p. 106: *“Table IV shows that Bayesian point estimates of γ are more efficient compared with PSA”*

It does not, it shows that the variance of the estimator in finite samples is smaller. However, ignoring that, given the frequentist coverage probabilities, it could be argued that the Bayesian posterior variances are too small.

Are they, too, “falsely precise” ?

Predictive Performance

Section 5 examines out-of-sample performance of the estimation procedure on by pooling the data with 4599 patients discharged in the following year.

Conclusion is that BPSA gives a more accurate prediction mechanism (by 7%) than PSA.

Point 13

This is a compelling reason for using the Bayesian approach (I would argue the most compelling in this setting).

This should have been studied in simulation as well, however.

Questions not addressed

- ▶ **Model selection:** how can we ensure the model selected for the propensity score is “correct” (ie encapsulates the correct functional relationships).
- ▶ **Theoretical properties of the Bayesian estimator:**
Asymptotically, do the Bayesian and frequentist estimators have identical properties ? It seems like they should do.

Conclusions

- ▶ Despite being grounded in the Bayesian framework, it is relatively uncontroversial, and would be largely acceptable to frequentists. The only point of contention might be the priors selected, but the impact of these could be studied in a sensitivity analysis.
- ▶ Some important questions are left unanswered (the importance of balancing, the means of selecting the variables entering into the PS etc.).
- ▶ In terms of the analysis, the predictive analysis is a good innovation.

Conclusions

- ▶ There seem to be inconsistencies in the way the simulation results are interpreted.
- ▶ Fixing the quintiles based on the ML-fitted propensity score seems like a weak point.
- ▶ The Bayesian analysis itself is not absolutely necessary, inference could be addressed in a pure likelihood setting.
- ▶ The MCMC implementation is relatively straightforward.
- ▶ In terms of the study results, the biggest flaw is possibly the issue of informative non-compliance, although important confounders may have been missed.