Chapter 13 Estimating the Propensity Score

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

August 20, 2021

Introduction

It is rare that we know the PS *a priori* in settings other than those involving randomized experiments.

A common situation is where a researcher views unconfoundedness as a reasonable approximation to the actual assignment mechanism, with only vague *a priori* information about the form of the dependence of the PS on the observed pre-treatment variables.

For example, in many medical settings, decisions are based on a set of clinically relevant patient characteristics observed by doctors and entered in patients' medical records.

However, there is typically no explicit rule that requires physicians to choose a specific treatment based on particular values of the pre-treatment variables, why there is no explicitly known form for the PS, and, thus, that it need to be estimated.

Introduction

It is important to note that some of the methods for causal analysis rely more heavily than others on an accurate approximation of the true PS by the estimated PS.

For example, the exact specification will likely matter less than when using the stratification methods then when using the inverse of the PS using weighting methods.

As a consequence, estimators for the treatment effects may be more or less sensitive to the decisions made in the specification of the PS.

In the following, we assume have a random sample of N units from an infinite super-population, that is either exposed to, or not exposed to, the treatment.

In the sample N_c units are exposed to the control treatment $W_i = 0$ and N_t units exposed to the active treatment $W_i = 1$, with $N = N_c + N_t$.

Introduction

The sole focus is on the statistical problem of estimating the conditional probability of receiving the treatment given the observed covariates, X_i, $i = 1, ...N$:

$$\Pr(W_i = 1|X_i = x) = \mathbb{E}\left[W_i|X_i = x\right], \tag{1}$$

which is equal to the super-population PS, $e(x)$, and we will use that notation here.

If the covariate X_i is a binary scalar, or more generally takes on only a few values, the statistical problem of estimating the PS is straightforward: we can simply partition the sample into subsamples that are homogeneous in the covariates, and estimate the PS for each subsample as the proportion of treated units in that subsample.

Using such a fully saturated model is not feasible in many realistic settings.

Introduction

Here we explicitly focus on settings where the covariates take on too many values to allow for a fully saturated model, so that some form of smoothing is essential.

The goal is to obtain estimates of the PS that balance the covariates between treated and control subsamples.

It is important to note, that the goal is *not* simply to get the best estimate of the PS in terms of mean-integrated-squared-error, or a similar criterion based on minimizing the difference between the estimated and true PS.

Such a criterion would always suggest that using the true PS is preferable to using an estimated PS.

Introduction

In contrast, for our purposes, it is often preferable to use the estimated PS.

The reason is that using the estimated PS may lead to superior covariate balance in the sample compared to that achieved when using the true super-population PS.

For example, in a CRE with a single binary covariate using the estimated PS to stratify units would lead to perfect within-stratum balance on the covariates in the sample, whereas using the true PS generally would not.

The difficulty is that our criterion, in-sample balance in the covariates given the (estimated) PS, is not as easy to formalize and operationalize as some of the conventional goodness-of-fit measures,

Introduction

There are two parts to the proposed algorithm of specifying the PS.

First we specify an initial model, motivated by substantive knowledge. Second, we assess the statistical adequacy of an estimate of that initial model, by checking whether the covariates are balanced within strata defined by the estimated PS.

In principle, one can iterate back and forth between these two stages, (i) specification of the model and (ii) assessment of that model, each time refining the specification of the model.

This lecture will describe a procedure (*i.e.,* an algorithm) for selecting a specification that can, at the very least, provide a useful starting point for such an iterative procedure.

Introduction

Three general comments are in order.

1. We do not use the outcome data. Consequently, there is no concern regarding the statistical properties of the ultimate estimates of the average treatment effects obtained from iterating back and forth between (i) and (ii).
2. It is difficult to give a fully automatic procedure for specifying the PS in a way that leads to a specification that passes all the tests and diagnostics that we may subject that specification to in the second stage.
3. The primary goal is to find an adequate specification of the PS, in the sense of a specification that achieves statistical balance in the covariates.

The Reinisch Barbituate Exposure Data

The data we use to illustrate the methods come from a study of the effect of prenatal exposure to barbituates (Reinisch, Sanders, Mortenson, and Rubin, 1995).

The data set contains information on $N = 7,943$ men and women born between 1959 and 1961 in Copenhagen, Denmark.

Of these, $N_t = 745$ men and women had been exposed *in utero* to substantial amounts of barbituates due to maternal medical conditions, and thus $N_c = 7,198$.

The substantive interest is in the effect of the barbituate exposure on cognitive development measured many years later.

The data set contains information on seventeen covariates that are potentially related to both the outcomes of interest, reflecting cognitive development, and the likelihood of having been prescribed and taking, barbituates.

Table 13.1: Summary Statistics Reinisch Data Set

Label	Variable Description	Controls (N_c =7198) mean	(s.d.)	Treated (N_t =745) mean	(s.d)	t-stat difference
gender	gender of child (female is 0)	0.51	0.50	0.50	0.50	-0.3
antih	exposure to antihistamine	0.10	0.30	0.17	0.37	4.5
hormone	exposure to hormone treatment	0.01	0.10	0.03	0.16	2.5
chemo	exposure to chemotherapy agents	0.08	0.27	0.11	0.32	2.5
cage	calendar time of birth	-0.00	1.01	0.03	0.97	0.7
cigar	mother smoked cigarettes	0.54	0.50	0.48	0.50	-3.0
lgest	length of gestation (10 ordered categories)	5.24	1.16	5.23	0.98	-0.3
lmotage	log of mother's age	-0.04	0.99	0.48	0.99	13.8
lpbc415	first pregnancy complication index	0.00	0.99	0.05	1.04	1.2
lpbc420	second pregnancy complication index	-0.12	0.96	1.17	0.56	55.2
motht	mother's height	3.77	0.78	3.79	0.80	0.7
motwt	mother's weight	3.91	1.20	4.01	1.22	2.0
mbirth	multiple births	0.03	0.17	0.02	0.14	-1.9
psydrug	exposure to psychotherapy drugs	0.07	0.25	0.21	0.41	9.1
respir	respiratory illness	0.03	0.18	0.04	0.19	0.7
ses	socioeconomic status (10 ordered categories)	-0.03	0.99	0.25	1.05	7.0
sib	if sibling equal to 1, otherwise 0	0.55	0.50	0.52	0.50	-1.6

Selecting the Covariates and Interactions

With many covariates it is is not always feasible simply to include all covariates in a model for the PS.

Moreoever, for some of the most important covariates, it may not be sufficient to include them only linearly, and we may wish to include functions, such as logarithms, and higher-order terms.

Here we describe a step-wise procedure for selecting the covariates and higher-order terms for inclusion in the PS.

We focus on logistic regression models and estimate the coefficients by maximum likelihood

The main question now concerns the selection of the functions of the basic covariates to include in the specification.

Selecting the Covariates and Interactions

We start by selecting a subset of the K covariates to be included linearly when estimating the log odds ratio of the PS, as well as a subset of all $K \cdot (K+1)/2$ second order terms (both quadratic and interactions terms). Thus a total of $K + K \cdot (K+1)/2 = K \cdot (K+3)/2$ included predictors.

As the the number of such subsets is $2^{K \cdot (K+3)/2}$ we cannot compare all possible subsets of this set. Instead we follow a stepwise procedure with three stages.

Selecting the Covariates and Interactions

1. select a set of K_{B} basic covariates that will be included in the PS, regardless of their statistical association with the treatment indicator, because they are viewed as important on substantive grounds.
2. decide which of the K_{L} of the remaining $K - K_{\mathrm{B}}$ covariates that will also be included linearly to estimate the log odds ratio.
3. decide which of the $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ interactions and quadratic terms involving only the K_{L} selected covariates to include.

Stage 3 will lead to the selection of K_{Q} second-order terms, leaving us with a vector of covariates with $K_{\mathrm{L}} + K_{\mathrm{Q}}$ components to be included linearly in the specification of the log odds ratio.

Now let us consider each of these three stages in more detail.

Step 1: Basic Covariates

The K_{B} basic covariates include those that are *a priori* viewed as important for explaining the assignment and plausibly related to some outcome measures. It may also be that $K_{\mathrm{B}} = 0$.

In evaluations of JTP, this step might lead to including covariates, viewed as important for the decision of the individual to participate, such as recent labor market experiences. With regard to the association with the outcomes, prior earnings or education levels should be highly relevant.

In the barbituate exposure example, this set includes three pretreatment variables, mother's age (`lmotage`), which is plausibly related to cognitive outcomes for the child, as well as socio-economic status (`ses`), which is strongly related to the number of physician visits during pregnancies and, thus exposes the mother to greater risk of barbituate prescriptions, and finally, gender of the child (`gender`), which may be associated with measures of cognitive outcomes.

Step 2: Additional Linear Terms

There are $K - K_{\mathrm{B}}$ covariates not included yet. We only consider at most $(K - K_{\mathrm{B}})!$ of the $2^{K-K_{\mathrm{B}}}$ different subsets involving these covariates.

Exactly how many and which of the subsets we consider depends on the configuration of the data.

We add one of the remaining covariates at a time, each time checking whether we wish to add it.

More specifically, suppose that at some point in the covariate selection process, we have selected $\tilde{K_{\mathrm{L}}}$ linear terms, including the K_B terms selected in the first step.

At that point we are faced with the decision whether to include an additional covariate from the set of $K - \tilde{K_{\mathrm{L}}}$ covariates, and if so, which one.

Step 2: Additional Linear Terms

This decision is based on the results of $K - \tilde{K}_{\mathrm{L}}$ additional logistic regression models.

In each of these $K - \tilde{K}_{\mathrm{L}}$ additional logistic regression models, we add to the basic specification with $\tilde{K}_{\mathrm{L}}$ covariates and an intercept, a single one of the remaining $K - \tilde{K}_{\mathrm{L}}$ covariates at a time.

For each of these $K - \tilde{K}_{\mathrm{L}}$ specifications, we calculate the likelihood ratio statistic assessing the null hypothesis that the newly included covariate has a zero coefficient.

If all the likelihood ratio statistics are less than some pre-set constant C_{L}, we stop, and we include only the $\tilde{K}_{\mathrm{L}}$ covariates linearly.

Step 2: Additional Linear Terms

If at least one of the likelihood ratio test statistics is greater than C_{L}, we add the covariate with the largest likelihood ratio statistic.

We now have $\tilde{K_{\mathrm{L}}} + 1$ covariates, and check whether any of the remaining $K - \tilde{K_{\mathrm{L}}} - 1$ covariates should be included by calculating likelihood ratio statistics for each of them.

We continue this process until none of the remaining likelihood ratio statistics exceeds C_{L}.

This second stage leads to the addition of $K_{\mathrm{L}} - K_{\mathrm{B}}$ covariates to the K_{B} covariates already selected for inclusion in the linear set in the first stage, for a total of K_{L} covariates.

Step 3: Quadratic and Interaction Terms

Given that $K_{\mathrm{L}} \leq K$ covariates in the linear stage, we now decide which of the $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ quadratic and interactions terms involving these K_L covariates to include.

Note that if some of the covariates are binary the effective set of possible second-order terms may be smaller than $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$.

We follow essentially the same procedure as for the linear stage. Suppose at some point we have added $\tilde{K_{\mathrm{Q}}}$ of the $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ possible interactions.

Step 3: Quadratic and Interaction Terms

We estimate $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ logistic regressions, each of which includes the intercept, the K_{L} linear terms (including the K_{B} basic ones), the $\tilde{K_{\mathrm{Q}}}$ second-order terms already selected, and one of the remaining $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ terms.

For each of these $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ logistic regressions, we calculate the likelihood ratio statistic for the null hypothesis that the most recently added second order term has a coefficient of zero.

If the largest likelihood ratio statistic is greater than some predetermined constant C_{Q}, we include that interaction term in the model.

Step 3: Quadratic and Interaction Terms

Then we re-calculate the likelihood ratio statistics for the remaining $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}} - 1$ interaction terms, and we keep including the term with the largest likelihood ratio statistic until all of the remaining likelihood ratio statistics are below C_{Q}.

This algorithm leaves us with a selection of K_{L} linear covariates and a selection of K_{Q} second order terms (plus an intercept).

We estimate the PS using this vector of $1 + K_{\mathrm{L}} + K_{\mathrm{Q}}$ terms.

To illustrate the implementation of this strategy, we use the threshold value for the likelihood ratio statistic of $C_{\mathrm{L}} = 1$ and $C_{\mathrm{Q}} = 2.71$, corresponding implicitly to z-statistics of 1 and 1.645.

Choosing the Specification of the PS for the Barbituate Data

The ultimate interest is in the effect of *in utero* barbituate exposure on cognitive outcomes when young adults.

Based on the substantive argument in the original papers using these data, it was argued that the child's sex, the mother's age, and mother's socio-economic status (`sex`, `lmotage`, and `ses` respectively) are particularly important covariates, the first two because they are likely to be associated with the outcomes of interest, and the last two because they are likely to be related to barbituate exposure.

With these three basic covariates in the specification of the PS, $K_{\mathrm{B}} = 3$.

As the first step towards deciding which other covariates to include linearly, we estimate the baseline model with an intercept and these three covariates.

Table 13.2: Estimated Parameters of PS: Baseline Case

Variable	est	s.e.	t-stat
intercept	-2.38	0.06	-41.0
`gender`	-0.01	0.08	-0.2
`lmotage`	0.48	0.04	11.7
`ses`	0.10	0.04	2.6

Both `lmotage` and `ses` are statistically significantly (at the 0.05 level) associated with *in utero* exposure to barbituates.

Choosing the Specification of the PS for the Barbituate Data

Next we estimate fourteen logistic regression models where we always include an intercept, `sex`, `lmotage`, `ses`, and additionally include, one at a time, the remaining 14 covariates.

For each specification, we calculate the likelihood ratio statistic for the test of the null hypothesis that the coefficient on the additional covariate is equal to zero.

For example, for the covariate `lpbc420`, the second pregnancy complication index

Variable	est	s.e.	t-stat
intercept	-3.71	0.10	-36.3
`gender`	0.07	0.09	0.8
`lmotage`	0.22	0.05	4.7
`ses`	0.15	0.05	3.3
`lpbc420`	2.11	0.08	27.2
LR Statistic	1308.0		

Choosing the Specification of the PS for the Barbituate Data

We do this for each of the fourteen remaining covariates (seventeen covariates minus the three pre-selected).

We report the fourteen likelihood ratio statistics in the first column of Table 13.4.

We find that the covariate that leads to the biggest improvement in the likelihood function is `lpbc420`.

The likelihood ratio statistic for that covariate is 1308.0. Because this value exceeds our threshold of $C_{\mathrm{L}} = 1$, we include the second pregnancy complication index `lpbc420` in the specification of the PS.

Table 13.4: Likelihood Ratio Statistics for Sequential Selection of Covariates to Enter Linearly

Covariate	Step →										
gender	–	–	–	–	–	–	–	–	–	–	–
antih	17.5	0.5	1.6	1.3	2.1	1.8	1.6	1.6	1.7	1.3	–
hormone	3.9	0.3	0.7	0.7	0.4	0.8	0.7	0.7	0.7	0.8	0.9
chemo	10.0	36.6	41.9	–	–	–	–	–	–	–	–
cage	0.8	5.8	6.4	7.2	7.6	7.9	–	–	–	–	–
cigar	4.3	2.3	3.5	3.7	3.0	2.1	2.1	1.7	2.1	–	–
lgest	0.4	11.1	5.0	6.4	7.3	5.5	5.6	–	–	–	–
lmotage	–	–	–	–	–	–	–	–	–	–	–
lpbc415	0.6	0.0	0.2	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
lpbc420	1308.0	–	–	–	–	–	–	–	–	–	–
motht	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
motwt	6.1	1.5	0.6	1.2	2.5	2.7	2.4	3.4	–	–	–
mbirth	4.6	66.1	–	–	–	–	–	–	–	–	–
psydrug	93.1	29.8	38.9	46.8	–	–	–	–	–	–	–
respir	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ses	–	–	–	–	–	–	–	–	–	–	–
sib	21.0	13.8	12.5	15.0	15.7	–	–	–	–	–	–

Choosing the Specification of the PS for the Barbituate Data

Next we estimate thirteen logistic regression models where we always include an intercept, `sex`, `lmotage`, `ses`, and `lpbc420`, and additionally include, one at a time, the remaining thirteen covariates.

The likelihood ratio statistics for the inclusion of these thirteen covariates are reported in the second column of the previous Table.

Now `mbirth`, the indicator for multiple births, is the most important covariate. Because the likelihood ratio statistic for the inclusion of `mbirth`, 66.1, exceeds the threshold of $C_{\mathrm{L}} = 1$, `mbirth` is added to the specification.

We keep checking whether there is any covariate that, when added to the baseline model, improves the likelihood function sufficiently, and if so, we include it in the specification of the PS.

Proceeding this way leads to the inclusion to ten additional covariates.

In the order they were added to the specification, these are, `lpbc420`, `mbirth`, `chemo`, `psydrug`, `sib`, `cage`, `lgest`, `motwt`, `cigar`, and `antih`.

With thirteen covariates there are potentially $13 \times (13 + 1)/2 = 91$ second order terms.

Not all 91 potential second-order terms are feasible, because some of the thirteen covariates selected in the first two steps are binary indicator variables, so that the corresponding quadratic terms are identical to the linear terms.

We select a subset of the non-trivial second-order terms in the same way we selected the linear terms, with the only difference being that the threshold for the likelihood ratio statistic is now 2.71, which corresponds to nominal statistical significance at the 10% level.

Choosing the Specification of the PS for the Barbituate Data

Following this procedure, adding one second-order term at a time, leads to the inclusion of seventeen second order terms.

Table 3.6 reports the parameter estimates for the PS with all the linear and second order terms selected, with the variables in the order in which they were selected for inclusion in the specification of the PS.

Table 13.6: Estimated Parameters of PS – Final Specification

Variable	est	s.e.	t-stat
intercept	-5.67	0.23	-24.4
linear terms			
sex	0.12	0.09	1.3
lmotage	0.52	0.11	4.7
ses	0.06	0.09	0.6
lpbc420	2.37	0.36	6.6
mbirth	-2.11	0.36	-5.9
chemo	-3.51	0.67	-5.2
psydrug	-3.37	0.55	-6.1
sib	-0.24	0.22	-1.1
cage	-0.56	0.26	-2.2
lgest	0.57	0.23	2.5
motwt	0.49	0.17	2.9
cigar	-0.15	0.10	-1.5
antih	0.17	0.13	1.3
second order terms			
lpbc420 × sib	0.60	0.19	3.1
motwt× motwt	-0.10	0.02	-4.5
lpbc420 × psydrug	1.88	0.39	4.8
ses× sib	-0.22	0.10	-2.2
cage× antih	-0.39	0.14	-2.8
lpbc420× chemo	1.97	0.49	4.0
lpbc420 × lpbc420	-0.46	0.14	-3.3
cage × lgest	0.15	0.05	3.0
lmotage × lpbc420	-0.24	0.10	-2.5
mbirth× cage	-0.88	0.39	-2.3
lgest × lgest	-0.04	0.02	-2.0
ses× cigar	0.20	0.09	2.2
lpbc420× motwt	0.15	0.07	2.0
chemo × psydrug	-0.93	0.46	-2.0
lmotage× ses	0.10	0.05	1.9
cage × cage	-0.10	0.05	-1.8
mbirth × chemo	$-\infty$	0.00	$-\infty$

Constructing Propensity-Score Strata

Next we wish to assess the adequacy of estimated PS at each value x, $\hat{e}(x)$, by exploiting a property of the true PS, namely

$$W_i \perp\!\!\!\perp X_i \,\Big|\, e(X_i). \tag{2}$$

We substitute $\hat{e}(X_i)$ for $e(X_i)$ and investigate whether, at least approximately,

$$W_i \perp\!\!\!\perp X_i \,\Big|\, \hat{e}(X_i), \tag{3}$$

Ideally we would do this by stratifying the sample into subsamples or blocks within each of which all units would have the exact same value of $\hat{e}(x)$, and then assessing whether W_i and X_i within each resulting block are independent.

Constructing Propensity-Score Strata

This plan is only feasible if $\hat{e}(x)$ takes on a relatively small number of values, and thus if the covariates jointly only take on a relatively small number of values in the sample.

Typically, in practice, that is not the case, and so we coarsen $\hat{e}(x)$ by constructing blocks (*i.e.*, strata or subclasses) within which $\hat{e}(x)$ vary only little.

For a set of boundary points, $0 = b_0 < b_1 < \ldots < b_{J-1} < b_J = 1$, define the block indicator B_{ij}, for the i-th unit, as

$$B_{ij} = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j, \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, J$. (Here we ignore the possibility that there are units with $\hat{e}(X_i)$ exactly equal to $B_{iJ} = 1$.)

Constructing Propensity-Score Strata

Then we will assess adequacy of $\hat{e}(x)$ by assessing whether

$$W_i \perp\!\!\!\perp X_i \,\Big|\, B_{i1}, \ldots, B_{iJ}. \tag{4}$$

We operationalize the assessment of independence by examining whether the treatment indicator and the covariates are uncorrelated within each of these blocks:

$$\mathbb{E}\left[X_i | W_i = 1, B_{ij} = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_{ij} = 1\right], \tag{5}$$

for all blocks $j = 1, \ldots, J$.

The first step in implementing this procedure is the choice of boundary values b_j, for $j = 0, \ldots, J$.

Constructing Propensity-Score Strata

We want to choose the boundary values in such a way that within each stratum, the variation in $\hat{e}(x)$ is modest.

The reason is that, if the $e(x)$ itself varies substantially within a stratum, then any evidence that the covariates are correlated with the treatment indicator within that same stratum is not compelling evidence of misspecification of $\hat{e}(x)$.

Thus, we choose the boundary values in such a way that, within any stratum, the indicator of receiving the treatment appears statistically unrelated to $\hat{e}(x)$.

We implement the selection of boundary points by an iterative procedure as follows.

Constructing Propensity-Score Strata

First we drop from this analysis all 'controls' with $\hat{e}(X_i) < \underline{e}_t$ and all 'treated' with $\hat{e}(X_i) > \overline{e}_c$, where

$$\underline{e}_t = \min_{i:W_i=1} \hat{e}(X_i),$$

and

$$\overline{e}_c = \max_{i:W_i=0} \hat{e}(X_i).$$

This trimming ensures some overlap between the two groups.

We start with a single block: $J = 1$, with boundaries $b_0 = \underline{e}_t$ and $b_1 = b_J = \overline{e}_c$. We then iterate through two steps.

Constructing Propensity-Score Strata

Step 1. Assessment of Adequacy of Blocks.

We use the linearized PS and its estimated counterpart, defined respectively as:

$$\ell(x) = \ln\left(\frac{e(x)}{1-e(x)}\right) \qquad \text{and} \quad \hat{\ell}(x) = \ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right).$$

The main reason to focus on the linearized PS is that it is more likely than the PS to have a distribution that is well approximated by a normal distribution.

Using the linearized PS we check the following two conditions for each block $j = 1, \ldots, J$.

1.A Independence

Let $N_c(j)$ and $N_t(j)$ denote the subsample sizes for controls and treated in block j,

$$N_c(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_{ij}, \qquad \text{and} \qquad N_t(j) = \sum_{i=1}^{N} W_i \cdot B_{ij},$$

and let $\overline{\ell}_c(j)$ and $\overline{\ell}_t(j)$ denote the average values for the estimated linearized PS, by treatment status and block,

$$\overline{\ell}_c(j) = \frac{1}{N_c(j)} \sum_{i=1}^{N} (1 - W_i) \cdot B_{ij} \cdot \hat{\ell}(X_i), \qquad \overline{\ell}_t(j) = \frac{1}{N_t(j)} \sum_{i=1}^{N} W_i \cdot B_{ij} \cdot \hat{\ell}(X_i),$$

1.A Independence

We the calculate the sample variance of the estimated linearized PS within block j,

$$s^2_\ell(j) = \frac{1}{N_t(j) + N_c(j) - 2} \cdot \left(\sum_{i:B_{ij}=1} (1 - W_i) \left(\hat{\ell}(X_i) - \overline{\ell}_c(j) \right)^2 + \sum_{i:B_{ij}=1} W_i \cdot \left(\hat{\ell}(X_i) - \overline{\ell}_t(j) \right)^2 \right).$$

The t-statistic for block j is then defined as

$$t_j = \frac{\overline{\ell}_t(j) - \overline{\ell}_c(j)}{\sqrt{S^2_\ell(j) \cdot (1/N_c(j) + 1/N_t(j))}}. \tag{6}$$

We compare this t-statistic for each stratum to a threshold value, which we fix at $t_{\max} = 1$. If the t-statistic is less than or equal to $t_{\max}$, we assess the estimated PS as varying little within the block, and if the t-statistic exceeds $t_{\max}$, we assess the block as exhibiting substantial variation in the PS.

1.B New Strata Size

If we were to split the current j-th stratum into two substrata, what would the new boundary value be, and how many observations would fall in each of the new substrata?

We compute the median value of the PS among the $N_c(j) + N_t(j)$ units with an estimated PS in the interval $[b_{j-1}, b_j)$. Denote this median by b'_j.

Let l and h denote the low and high substratum respectively, then

$$N_c^l(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_{ij} \cdot \mathbf{1}_{\hat{e}(X_i) < b'_j}, \qquad N_c^u(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_{ij} \cdot \mathbf{1}_{\hat{e}(X_i) \geq b'_j}$$

$$N_t^l(j) = \sum_{i=1}^{N} W_i \cdot B_{ij} \cdot \mathbf{1}_{\hat{e}(X_i) < b'_j}, \qquad \text{and} \quad N_t^u(j) = \sum_{i=1}^{N} W_i \cdot B_{ij} \cdot \mathbf{1}_{\hat{e}(X_i) \geq b'_j},$$

is the number of control and treated units with estimated PS in the lower subinterval $[b_{j-1}, b'_j)$ and in the upper subinterval $[b'_j, b_j)$ respectively.

1.B New Strata Size

The current block j is assessed to be inadequately balanced if the t-statistic is too high, $|t_j| > t_{\max}$.

It also needs to be and amenable to splitting at the median. By choosing

$$\min(N_c^l(j), N_t^l(j), N_c^u(j), N_t^u(j)) \geq 3, \text{ and } \min(N_c^l(j) + N_t^l(j), N_c^u(j) + N_t^u(j)) \geq K + 2,$$

this enables us to compare mean covariate values within blocks, and so that later we can do at least some adjustment for remaining covariate differences within blocks.

Constructing Propensity-Score Strata

Step 2: Split blocks that are both inadequately balanced and amenable to splitting.

If block j is assessed to be inadequately balanced and amenable to splitting, then this block is split into two new blocks, corresponding to PS values in $[b_{j-1}, b'_j)$ and in $[b'_j, b_j)$, and the number of strata is increased by 1.

We iterate between assessment (1.A) and the splitting (1.B) until all blocks are assessed to be either adequately balanced, or too small to split.

Choosing Strata for the Barbituate Data

We use the PS score estimated using the specification in Table 13.6.

With a single block, $J = 1$ the the lower and upper boundaries are equal to $b_0 = \underline{e}_t = 0.0080$, and $b_1 = \overline{e}_c = 0.9252$ respectively.

Out of the 7,198 individuals who were not exposed to barbituates, 2,737 have estimated PS less than b_0, and out of the 745 individuals who were exposed to barbituates before birth, 3 have estimated PS exceeding b_1. These individuals are discarded.

Hence, in this first stratum we have $N_c(1) = 4,461$ controls and $N_t(1) = 742$ treated individuals left with estimated PS between $b_0 = 0.0080$ and $b_1 = 0.9252$.

For this first block we calculate the t-statistic, t_1. This leads to $t_1 = 36.3$, which exceeds $t_{\max} = 1$.

Choosing Strata for the Barbituate Data

We split the block at the median of the estimated PS within this stratum (equal to 0.06) and find: $N_c^l(1) = 2,540$, $N_t^l(1) = 61$, $N_c^u(1) = 1,921$, and $N_t^u(1) = 681$.

Therefore the current single-block subclassification is deemed inadequate, and the single block is split into two new blocks, with the new boundary equal to the median in the original subclass, equal to 0.06. These results are in the first panel of Table 13.7

Table 13.7: Determination of the Number of Blocks and Their Boundaries. Boldface block numbers indicate blocks that were split at this step.

step	block	lower bound	upper bound	width	# controls	# treated	t-stat
1	1	0.00	0.94	0.94	4462	742	36.3
2	**1**	0.00	0.06	0.06	2540	61	3.2
	2	0.06	0.94	0.88	1922	681	23.7
3	**1**	0.00	0.02	0.01	1280	20	2.2
	2	0.02	0.06	0.05	1260	41	0.5
	3	0.06	0.20	0.14	1163	138	3.9
	4	0.20	0.94	0.74	759	543	10.9
4	**1**	0.00	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	-0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.94	0.57	301	351	5.6
5	1	0.00	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	-0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.50	0.13	181	144	2.5
	8	0.50	0.94	0.44	120	207	2.3
6	1	0.00	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	-0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.42	0.05	101	61	0.3
	8	0.42	0.50	0.08	80	83	0.7
	9	0.50	0.61	0.11	73	90	0.8
	10	0.61	0.94	0.34	47	117	-0.3

Choosing Strata for the Barbituate Data

The first block with boundaries 0.01 and 0.06 has $N_c(1) = 2,540$ individuals in the control group and $N_t(1) = 61$ individuals in the treatment group.

The t-statistic is 3.2. Splitting the block into two parts at the median value of the estimated PS we find 1,280 control and 20 treated units in the first sub block, and 1,260 control and 41 treated units in the second sub block.

The number of units in each sub class is sufficiently large, and therefore the original block will be split into two parts, at the median value of 0.02.

Choosing Strata for the Barbituate Data

For the second block with boundary values 0.06 and 0.9252, we again find that the stratification is inadequate, with a t-statistic of 23.7.

These results are in the second panel. As a result, we split both blocks, leading to four new blocks.

When we continue this procedure with the four new blocks, we find that the second of the four blocks was sufficiently balanced. The remaining three blocks were not well balanced, leading a total of seven blocks in the next round. See the third panel.

Choosing Strata for the Barbituate Data

We continue checking the adequacy of the blocks until either all the t-statistics are below $t_{\max} = 1$, or splitting a block would lead to a new block that would contain an insufficient number of units of one treatment type or another.

This algorithm leads to ten blocks, with the block boundaries, block widths, and the number of units of each type in the block presented in the last panel of the Table.

In the last column we also present the t-statistics. One can see that most of the blocks are well-balanced in the linearized PS, with only two blocks somewhat unbalanced with t-statistics exceeding $t_{\max} = 1$.

For example, the second block is not particularly well balanced in the linearized PS, with a t-statistic of 1.7, but splitting it would lead to a new block with no treated units, and therefore this block is not split further.

Assessing Balance Conditional on the Estimated PS

One problem assessing the within-block equality of means of the covariates across the treatment groups is the large amount of relevant information.

We may have a large number of covariates (in the barbituate study, there are seventeen covariates), and a substantial number of blocks (ten in our application).

Even if we were to have data from a randomized experiment, in a finite sample one would expect some covariates, in at least some strata, to be sufficiently correlated with W_i that some statistical tests ignoring the multiplicity of comparisons would suggest statistical significance of some comparisons at conventional single-test levels.

Assessing Balance Conditional on the Estimated PS

Here we propose a method for assessing the overall balance that allows for comparisons of balance across specifications of the PS and across strata definitions.

We are interested in assessing

$$W_i \perp\!\!\!\perp X_i \,\Big|\, B_{i1}, \ldots, B_{iJ},$$

implemented through an assessment of the equality,

$$\mathbb{E}\left[X_i | W_i = 1, B_{ij} = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_{ij} = 1\right], \quad \text{for } j = 1, \ldots, J.$$

We discuss three sets of tests for each covariate.

The first two are based on single statistics while the third is a set of tests based on separate within-stratum comparisons.

Assessing Balance Conditional on the Estimated PS

For the first two tests, we analyze the data as if they arose from a SRE.

Define

$$\overline{X}_{c,k}(j) = \frac{1}{N_c(j)} \sum_{i:W_i=0} B_{ij} \cdot X_{ik}, \qquad \text{and} \quad \overline{X}_{t,k}(j) = \frac{1}{N_t(j)} \sum_{i:W_i=1} B_{ij} \cdot X_{ik},$$

respectively, for $k = 1, \ldots, K$, and $j = 1, \ldots, J$.

Assessing Global Balance for Each Covariate Across Strata

Take the k-th component of the vector covariate X_i, X_{ik}. In stratum j the pseudo average causal effect of the treatment on this covariate can be estimated by

$$\hat{\tau}_k^X(j) = \overline{X}_{t,k}(j) - \overline{X}_{c,k}(j),$$

The sampling variance of this estimator $\hat{\tau}_k^X(j)$ is estimated as

$$\hat{\mathbb{V}}_k^X(j) = s_k^2(j) \cdot \left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)} \right),$$

Assessing Global Balance for Each Covariate Across Strata

where

$$s_k^2(j) = \frac{1}{N_c(j) + N_t(j) - 2}$$
$$\times \left(\sum_{i=1}^{N} (1 - W_i) \cdot B_{ij} \cdot \left(X_{ik} - \overline{X}_{c,k}(j) \right)^2 + \sum_{i=1}^{N} W_i \cdot B_{ij} \cdot \left(X_{ik} - \overline{X}_{t,k}(j) \right)^2 \right).$$

The estimate of the pseudo average causal effect is then the weighted average of these within-block estimates,

$$\hat{\tau}_k^X = \sum_{j=1}^{J} \frac{N_c(j) + N_t(j)}{N} \cdot \hat{\tau}_k^X(j),$$

with estimated sampling variance

$$\hat{\mathbb{V}}_k^X = \sum_{j=1}^{J} \left(\frac{N_c(j) + N_t(j)}{N} \right)^2 \cdot \hat{\mathbb{V}}_k^X(j).$$

Assessing Global Balance for Each Covariate Across Strata

Finally we convert these into

$$z_k = \frac{\hat{\tau}_k^X}{\sqrt{\hat{\mathbb{V}}_k^X}}.$$

If we find that the z-values are substantially larger in absolute values than one would expect if they were drawn independently from a normal distribution, we would conclude that the stratification does not lead to satisfactory balance in the covariates.

The average pseudo causal effects τ_k^X may be zero, even if some of the $\tau_k^X(j)$ are not.

Assessing Balance for Each Covariate Within All Blocks

Next we therefore assess overall balance by calculating F-statistics across all strata, one covariate at a time.

To this end we use a two-way Analysis Of Variance (ANOVA) procedure to test the null hypothesis that its mean for the treated subpopulation is identical to that of the mean of the control subpopulation in each of the J strata.

One way to calculate the F-statistic is through a linear regression of the form

$$\mathbb{E}\left[X_{ik} | \, W_i, B_{i1}, \ldots, B_{iJ}\right] = \sum_{j=1}^{J} \alpha_{kj} \cdot B_{ij} + \sum_{j=1}^{J} \tau_k^X(j) \cdot B_{ij} \cdot W_i.$$

Assessing Balance for Each Covariate Within All Blocks

First we estimate the unrestricted estimates $(\hat{\alpha}^{\text{ur}}, \hat{\tau}^X)$ by minimizing

$$(\hat{\alpha}^{\text{ur}}, \hat{\tau}^X) = \arg\min_{\alpha,\tau} \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \alpha_{kj} \cdot B_{ij} - \sum_{j=1}^{J} \tau_k^X(j) \cdot B_{ij} \cdot W_i \right)^2,$$

which leads to

$$\hat{\alpha}_{kj}^{\text{ur}} = \overline{X}_{c,k}(j), \qquad \text{and} \quad \hat{\tau}_k^X(j) = \overline{X}_{t,k}(j) - \overline{X}_{c,k}(j).$$

Next we estimate the restricted estimates $\hat{\alpha}^{\text{r}}$ (under the restriction that all the $\tau_k^X(j) = 0$) by minimizing

$$\hat{\alpha}^{\text{r}} = \arg\min_{\alpha} \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \alpha_{kj} \cdot B_{ij} \right)^2,$$

leading to

$$\hat{\alpha}_{kj}^{\text{r}} = \frac{N_c(j)}{N_c(j) + N_t(j)} \cdot \overline{X}_{c,k}(j) + \frac{N_t(j)}{N_c(j) + N_t(j)} \cdot \overline{X}_{t,k}(j).$$

Assessing Balance for Each Covariate Within All Blocks

The F-test of interest is then the statistic for testing the null hypothesis that all $\tau_k^X(j) = 0$, for $j = 1, \ldots, J$.

The form of the F-statistic for covariate X_{ik} is

$$F_k = \frac{(\mathrm{SSR}_k^{\mathrm{r}} - \mathrm{SSR}_k^{\mathrm{ur}})/J}{\mathrm{SSR}_k^{\mathrm{ur}}/(N-2J)},$$

where the restricted sum of squared residuals is

$$\mathrm{SSR}_k^{\mathrm{r}} = \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \hat{\alpha}_{kj}^{\mathrm{r}} \cdot B_{ij} \right)^2,$$

and the unrestricted sum of squares is

$$\mathrm{SSR}_k^{\mathrm{ur}} = \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \hat{\alpha}_{kj}^{\mathrm{ur}} \cdot B_{ij} - \sum_{j=1}^{J} \hat{\tau}_k^X(j) \cdot B_{ij} \cdot W_i \right)^2.$$

Assessing Balance for Each Covariate Within All Blocks

We then convert the p-value associated with this F-stastistic, under normality of the covariates nominally from an F-distribution with J and $N - 2 \cdot J$ degrees of freedom, to a z-value.

For each of the K covariates X_{ik}, we obtain a set of K z-values, z_k, $k = 1, \ldots, K$.

If the covariates are well balanced between treatment and control groups conditional on the PS, we would expect to find the z-values to be concentrated towards smaller (more negative) values relative to a normal distribution.

Finding large positive values would suggest that the covariates are not balanced within the strata.

Assessing Balance within Strata for Each Covariate

The third approach focuses on a single covariate in a single stratum at a time.

For each covariate X_{ik}, for $k = 1, \ldots, K$, and for each stratum $j = 1, \ldots, J$, we test the null hypothesis

$$\mathbb{E}\left[X_{ik} | W_i = 1, B_{ij} = 1\right] = \mathbb{E}\left[X_{ik} | W_i = 0, B_{ij} = 1\right] \quad \text{for } j = 1, \ldots, J.$$

against the alternative hypothesis that the two averages differ.

For the k-th covariate, and for this stratum j, we calculate

$$z_{jk} = \frac{\overline{X}_{t,k}(j) - \overline{X}_{c,k}(j)}{\sqrt{s_k^2(j) \cdot (1/N_c(j) + 1/N_t(j))}}, \tag{7}$$

where

Assessing Balance within Strata for Each Covariate

$$s_k^2(j) = \frac{1}{N_c(j) + N_t(j) - 2}$$
$$\times \left(\sum_{i=1}^{N} (1 - W_i) \cdot B_{ij} \cdot \left(X_{ik} - \overline{X}_{c,k}(j) \right)^2 + \sum_{i=1}^{N} W_i \cdot B_{ij} \cdot \left(X_{ik} - \overline{X}_{t,k}(j) \right)^2 \right).$$

If the covariates are well balanced, we would expect to find the absolute values of the z-values to be concentrated towards smaller (less significant) values relative to a normal distribution.

To summarize the $K \times J$ z-values it is useful to present Q-Q plots, comparing the z-values against their expected values under independent draws from the normal distribution. A Q-Q plots flatter than a 45^0 line is a sign of balance.

Assessing Covariate Balance for the Barbituate Data

Given the previous stratification for the barbituate data with ten blocks we calculate a number of statistics to assess the adequacy of the PS specification.

The results are presented in Table 13.8.

The **Within-blocks** panel presents the z-values given in equation (7)

In addition, there are two columns for the two overall tests.

For comparison, the last column present the t-statistic for the null hypothesis that the overall average covariate values are equal in the two treatment groups, not adjusted for the blocks.

Table 13.8: z-values For Balancing Tests – Final PS Specification

	Within Blocks										Overall		1 block
Block →	1	2	3	4	5	6	7	8	9	10	t-test	F-test (z-value)	t-test
Cov ↓													
sex	-0.05	-2.27	1.97	0.81	0.89	-1.28	0.04	-0.39	-1.42	1.14	0.13	1.22	-0.73
antih	-0.67	-0.47	0.67	0.03	0.37	-0.25	0.38	-0.53	-0.11	0.27	-0.17	-2.88	3.21
hormone	-0.14	-0.42	-0.65	-1.00	0.25	0.71	-0.22	-1.05	-1.10	0.21	-0.99	-0.66	1.66
chemo	0.55	-0.39	-0.78	-0.75	-1.17	1.47	-0.94	0.61	0.66	0.29	-0.27	-0.61	1.76
cage	-1.41	-0.29	-1.04	-0.46	2.11	0.28	0.20	0.46	-1.48	-0.74	-1.38	0.34	1.15
cigar	-0.37	0.55	0.58	1.50	0.31	-0.93	0.21	-0.99	0.25	-0.39	0.52	-1.17	-3.13
lgest	0.90	0.58	-0.07	-0.82	0.79	-0.36	0.05	-0.33	-1.14	1.21	0.71	-1.48	0.12
lmotage	-2.20	-1.37	0.56	1.64	0.95	0.60	-0.96	-1.73	-1.47	0.36	-1.26	1.45	8.56
lpbc415	-0.48	-1.84	-1.00	-0.34	0.59	0.44	-0.20	-0.16	1.07	-0.10	-1.49	-0.82	0.75
lpbc420	1.04	0.84	-0.67	-0.86	-1.61	1.80	-0.39	1.62	1.14	-1.80	0.51	0.59	32.04
motht	-0.84	0.45	-0.67	0.75	0.64	0.09	0.30	-1.37	-0.60	-0.13	-0.50	-1.37	0.90
motwt	1.23	1.14	0.12	-1.23	-0.05	-0.45	-0.32	1.94	-0.01	-0.47	1.08	-0.18	1.44
mbirth	-0.44	-0.80	-1.54	-0.37	1.80	0.20	0.00	2.25	-1.58	-1.60	-1.28	1.00	-2.93
psydrug	-0.66	-1.01	1.05	-0.15	-0.78	0.06	-0.18	0.08	0.09	0.89	-0.29	-1.40	6.32
respir	-0.49	0.53	-0.21	0.98	1.38	0.24	-0.78	-1.51	0.22	-0.28	0.24	-0.49	0.19
ses	-0.60	-0.31	-0.74	1.16	0.82	-0.08	-0.03	-0.82	-0.91	0.36	-0.56	-1.37	5.19
sib	1.42	2.37	-1.09	-1.58	-1.53	0.11	0.63	1.63	1.19	0.23	0.98	1.64	1.48

Assessing Covariate Balance for the Barbituate Data

The first ten columns of the table give the z-values separately for each block and each of the seventeen covariates.

The largest of these 170 z-values is 2.37. To facilitate the overall assessment of these z-values we construct a Q-Q plot, where we plot the ordered z-values, against the corresponding quantiles of the normal distribution. The Q-Q plot is presented in the top panel in the Figure in the next slide.

Q-Q plots

Fig 1: balance in covariates (C_L=1,C_Q=2.78)

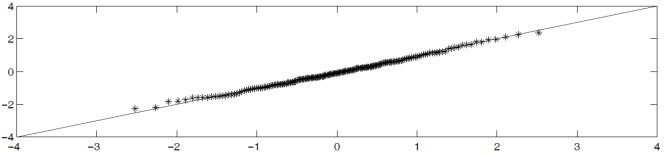


Fig 2: balance in covariates (C_L=0.C_Q=infty)

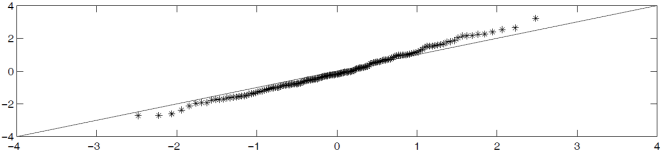
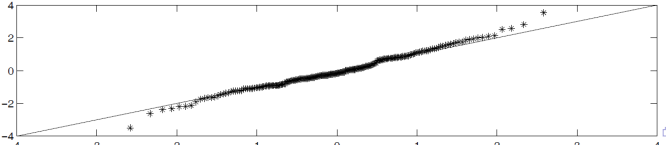


Fig 3: balance in covariates (lasso)



Assessing Covariate Balance for the Barbituate Data

The column with the heading "t-test" presents z-values for the test of zero average pseudo causal effects for each of the seventeen covariates, after stratification on the estimated propensity score.

The largest of the absolute values of the seventeen t-statistics is 1.49, suggesting excellent balance.

For the alternative F-test we find that the largest value is 2.88, with all the others below 2.00 again suggesting excellent balance conditional on the propensity score.

Based on these balance assessments, we conclude that the specification of the PS is adequate in the sense that it leads to somewhat better balance than one would expect to see if assignment were randomized within blocks.

Assessing Covariate Balance for the Barbituate Data

If we had found that the balance was poor, we might have attempted to improve balance by changing the specification for the PS.

We propose no general algorithm to improve balance beyond providing some general guidelines.

For example, if one finds that many of the t-statisics for a particular covariate are large in absolute value, one may wish to include more flexible functional forms for that covariate, possibly piecewise linear components, or indicator variables for particularly important regions of its values.

Assessing Covariate Balance for the Barbituate Data

To put the extent of the covariate balance given our preferred specification in perspective, we consider two alternative specifications of the PS.

In the first alternative specification we include all seventeen linear terms but no second order terms. Within our algorithm this corresponds to $C_{\mathrm{L}} = 0$, $C_{\mathrm{Q}} = \infty$.

This specification appears to be common in empirical work, where researchers often simply include all covariates in the PS without investigating whether that specification of the PS leads to adequate balance in the covariates.

Constructing the blocks with this specification of the PS leads to nine blocks.

Assessing Covariate Balance for the Barbituate Data

We find that 15 out of 153 z-values exceed 2.0, compared to only 2 out of 170 with our preferred specification of the PS (for details see Table 13.9).

The Q-Q plot for the 153 z-values is displayed in the second panel of the last figure in the previous slide.

It is clear that including some second order terms leads to substantially better balance in the covariates.

In the second alternative specification we use lasso methods to select among all seventeen linear terms and 153 second order terms. We use ten-fold cross-validation to select the penalty term.

Assessing Covariate Balance for the Barbituate Data

The lasso procedure selects fourteen covariates, three linear ones (`chemo`, `lpbc420`, and `mbirth`), and eleven second order terms.

We find that there are now 14 out of 204 z-values exceeding 2.0, again, compared to 2 out of 170 with our preferred specification of the PS (for details see Table 13.10).

The Q-Q plot for the 153 z-values is displayed in the last panel of the last figure.

It appears that the lasso does not lead to as good an in-sample fit as our proposed specification, possibly due to its focus on out-of-sample prediction.

Assessing Covariate Balance for the Barbituate Data

The correlation between the linearized PS based on our proposed specification and the linear specification is 0.95, between the proposed specification and the lasso specification the correlation is 0.96, and the correlation between the linear and the lasso specification is 0.98.

The log likelihood values for the three specifications are -1,556.3 for the proposed specification, -1.627.7 for the linear specification, and -1,614.7 for the lasso specification.