

Causal Inference Methods and Case Studies

STAT24630

Jingshu Wang

Lecture 10

Topic: confounding in observational data, propensity score estimation

- Conditional randomized experiment
 - Simpson's paradox
 - Balancing score
 - Estimators: outcome regression, IPW, matching
- Observational data
 - Causal inference rationale
 - Propensity score estimation

Conditional randomized experiment

- Treatment assignment mechanism depends on pre-treatment covariates \mathbf{X}_i
 - Example: stratified randomized experiment, proportion of treated units can be different in different strata
- **Unconfoundedness property:** $W_i \perp (Y_i(0), Y_i(1)) \mid \mathbf{X}_i$
 - Assignment mechanism does not depend any unobserved \mathbf{U} pretreatment confounders
 - \mathbf{X}_i can either be continuous or discrete
 - If \mathbf{X}_i is discrete or discretized \rightarrow stratified randomized experiment
- Propensity score: $e(\mathbf{X}_i) = P(W_i = 1 \mid \mathbf{X}_i) \in (0,1)$
 - **Overlap assumption:** $e(\mathbf{x}) \neq 0$ or 1 for any \mathbf{x} (otherwise we won't have data to identify $\tau(\mathbf{x})$)
 - In stratified randomized experiment: $e(\mathbf{X}_i = j) = P(W_i = 1 \mid \mathbf{X}_i = j) = N_t(j)/N(j)$
- Identify conditional average treatment effect under unconfoundedness

$$\begin{aligned}\tau(\mathbf{x}) &= \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}) \\ &= \mathbb{E}(Y_i(1) \mid \mathbf{X}_i = \mathbf{x}, W_i = 1) - \mathbb{E}(Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, W_i = 0) \\ &= \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 0)\end{aligned}$$

Conditioning on confounded covariates

- (Population) average treatment effect

$$\begin{aligned}\tau &= \mathbb{E}(\tau(\mathbf{X}_i)) = \mathbb{E}\left(\mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i, W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i, W_i = 0)\right) \\ &= \sum_{\mathbf{x}} \left(\mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 0) \right) P(\mathbf{X}_i = \mathbf{x})\end{aligned}$$

Shared weights

- Conditioning on the confounding covariates \mathbf{X}_i is important

$$\begin{aligned}\mathbb{E}(Y_i^{\text{obs}} \mid W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid W_i = 0) \\ = \sum_{\mathbf{x}} \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 1) P(\mathbf{X}_i = \mathbf{x} \mid W_i = 1) - \sum_{\mathbf{x}} \mathbb{E}(Y_i^{\text{obs}} \mid \mathbf{X}_i = \mathbf{x}, W_i = 1) P(\mathbf{X}_i = \mathbf{x} \mid W_i = 0)\end{aligned}$$

Different weights

- If $e(\mathbf{X}_i) = P(W_i = 1 \mid \mathbf{X}_i) \equiv c$, then $W_i \perp \mathbf{X}_i \Rightarrow P(\mathbf{X}_i = \mathbf{x} \mid W_i = 1) = P(\mathbf{X}_i = \mathbf{x} \mid W_i = 0)$

Simpson's paradox: kidney stone treatment

- Compare the success rates of two treatment of kidney stones
- Treatment A: open surgery; treatment B: small puctures

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

$$P(X_i = x)$$

$$(87 + 270)/700 = 0.51$$

$$(263 + 80)/700 = 0.49$$

- What is the confounder here? Size of the stone
 - Small stone: propensity score is $\frac{87}{87+270} = 0.24$
 - Large stone: propensity score is $\frac{263}{263+80} = 0.77$
- True average causal effect: $83.2\% - 78.2\% : (93\% \times 0.51 + 73\% \times 0.49) - (87\% \times 0.51 + 69\% \times 0.49)$
- We also mentioned Simpson's paradox in Lecture 6 when choosing test statistics for Fisher's exact p-value in stratified randomized experiment

Simpson's paradox: UC Berkeley gender bias

- In the early 1970s, the University of California, Berkeley was sued for gender discrimination over admission to graduate school.
- “Causal” effect of sex on application admission (data of Year 1973 admission)

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

- Confounding covariate: department

Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

$e(X_i)$	$P(X_i)$
0.12	0.21
0.04	0.13
0.65	0.21
0.47	0.18
0.67	0.13
0.56	0.14

For data from departments A-F:

- Raw average admission rate between men and women:
46% V.S. 30%
- After adjusting for department:
40% V.S. 44%

Balancing score

- Under unconfoundedness, we can remove all biases in comparing treated and control units by conditioning on each level of \mathbf{X}_i
- Too few samples to compare at each level if too many variables in \mathbf{X}_i
- Balancing score $b(\mathbf{X}_i)$: lower-dimensional functions of \mathbf{X}_i that remove differences between treatment and control groups

$$W_i \perp \mathbf{X}_i \mid b(\mathbf{X}_i)$$

- Balancing scores are not unique: any one-to-one mapping of a balancing score is a balancing score
- **Propensity score $e(\mathbf{X}_i)$ is a balancing score**
 - We want to show that $P(W_i = 1|\mathbf{X}_i, e(\mathbf{X}_i)) = P(W_i = 1|e(\mathbf{X}_i))$
 - $P(W_i = 1|\mathbf{X}_i, e(\mathbf{X}_i)) = P(W_i = 1|\mathbf{X}_i) = e(\mathbf{X}_i)$
 - By the law of total expectation
$$\begin{aligned} P(W_i = 1|e(\mathbf{X}_i)) &= \mathbb{E}[W_i|e(\mathbf{X}_i)] = \mathbb{E}[\mathbb{E}[W_i|\mathbf{X}_i, e(\mathbf{X}_i)]|e(\mathbf{X}_i)] \\ &= \mathbb{E}[\mathbb{E}[W_i|\mathbf{X}_i]|e(\mathbf{X}_i)] = \mathbb{E}[e(\mathbf{X}_i)|e(\mathbf{X}_i)] = e(\mathbf{X}_i) \end{aligned}$$
 - Propensity score the coarsest balancing score (Lemma 12.3): $e(\mathbf{X}_i)$ is a function of any $b(\mathbf{X}_i)$

Unconfoundedness with balancing score

- Why do we care about balancing score?

$$W_i \perp (Y_i(0), Y_i(1)) \mid \mathbf{X}_i \Rightarrow W_i \perp (Y_i(0), Y_i(1)) \mid b(\mathbf{X}_i)$$

- Given a vector of covariates that ensure unconfoundedness, adjustment for differences in propensity scores removes all biases associated with differences in the covariates
- For the propensity score $W_i \perp (Y_i(0), Y_i(1)) \mid e(\mathbf{X}_i)$
- $e(\mathbf{X}_i)$ can be reviewed as a summary score of the pre-treatment covariates

$$\tau = \mathbb{E} \left(\mathbb{E}(Y_i^{\text{obs}} \mid e(\mathbf{X}_i), W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid e(\mathbf{X}_i), W_i = 0) \right)$$

- The proof can be found on Page 267, Chapter 12.3

Estimate ATE under unconfoundedness

- Adjust for confounding variables when estimating the average treatment effect τ
- Three strategies
 - Outcome regression
 - Inverse probability weighting
 - Matching
- We are not introducing new methods to estimate ATE for randomized experiments, we review the estimators we discuss in previous lectures from a different angle, to prepare us to perform causal inference in observation studies

Outcome regression estimator

- $\tau = \mathbb{E}(\mathbb{E}(Y_i^{\text{obs}} | \mathbf{X}_i, W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} | \mathbf{X}_i, W_i = 0))$
- Define the conditional expectations $\mu_w(\mathbf{x}) = \mathbb{E}(Y_i^{\text{obs}} | \mathbf{X}_i = \mathbf{x}, W_i = w)$
- We can estimate the conditional expectations via a regression model and obtain $\hat{\mu}_w(\mathbf{x})$
- Estimator for the ATE: $\hat{\tau}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^N (\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i))$
- For example, if we assume a linear regression model $\mathbb{E}(Y_i^{\text{obs}} | \mathbf{X}_i, W_i) = \alpha + \tau W_i + \boldsymbol{\beta}^T \mathbf{X}_i$
 - $\hat{\mu}_w(\mathbf{x}) = \hat{\alpha} + \hat{\tau}w + \hat{\boldsymbol{\beta}}^T \mathbf{x}$, $\hat{\tau}_{\text{reg}} = \hat{\tau}$
- In practice, we can use any kinds of machine learning approaches (linear regressions, logistic regression, random forest, SVM, deep learning, ...) to obtain $\hat{\mu}_w(\mathbf{x})$
- Drawback of outcome regression approach: interpretability of the assumption
 - a regression model on $\mathbb{E}(Y_i^{\text{obs}} | \mathbf{X}_i = \mathbf{x}, W_i = w)$ is modeling the observed data
 - Need to explain the underlying model assumptions on the potential outcomes (like what we did in Lecture 6) : a model for $\mathbb{E}((Y_i(0), Y_i(1)) | \mathbf{X}_i, W_i)$

Inverse probability weighting (IPW)

- What if we don't want to put a model assumption on the observed (potential) outcome?
 - If X_i is unconfounded ($W_i \perp X_i$) and the model assumption is wrong, we may lose efficiency, but $\hat{\tau}_{\text{reg}}$ is likely still unbiased for τ
 - If X_i are confounding covariates and the model assumption is wrong, $\hat{\tau}_{\text{reg}}$ is often be a biased estimator of τ
- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)} \right] = \mathbb{E}_{\text{sp}} [Y_i(1)], \quad \text{and} \quad \mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)} \right] = \mathbb{E}_{\text{sp}} [Y_i(0)].$$

Proof:

$$\begin{aligned} \mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)} \right] &= \mathbb{E}_{\text{sp}} \left[\mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)} \middle| X_i \right] \right] = \mathbb{E}_{\text{sp}} \left[\mathbb{E} \left[\frac{Y_i(1) \cdot W_i}{e(X_i)} \middle| X_i \right] \right] = \mathbb{E}_{\text{sp}} \left[\frac{\mathbb{E}_{\text{sp}}[Y_i(1)|X_i] \cdot \mathbb{E}_W[W_i|X_i]}{e(X_i)} \right] \\ &= \mathbb{E}_{\text{sp}} [\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]] = \mathbb{E}_{\text{sp}} [Y_i(1)] \end{aligned}$$

Same derivation for the second equation.

Inverse probability weighting (IPW)

- What if we don't want to put a model assumption on the observed (potential) outcome?
 - If X_i is unconfounded ($W_i \perp X_i$) and the model assumption is wrong, we may lose efficiency, but $\hat{\tau}_{\text{reg}}$ is likely still unbiased for τ
 - If X_i are confounding covariates and the model assumption is wrong, $\hat{\tau}_{\text{reg}}$ is often be a biased estimator of τ
- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)} \right] = \mathbb{E}_{\text{sp}} [Y_i(1)], \quad \text{and} \quad \mathbb{E} \left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)} \right] = \mathbb{E}_{\text{sp}} [Y_i(0)].$$

- We give a weight $\lambda_i = 1/P(W_i = w | X_i)$ to each unit i , inversely proportional to the probability of being assigned to the group w
- Intuitively, unit that has a smaller $e(X_i)$ has less chance to appear in the treatment group, so we should give it a higher weight

Inverse probability weighting estimator

$$\begin{aligned}\hat{\tau}_{\text{IPW}} &= \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N} \sum_{i=1}^N \frac{(1 - W_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)} \\ &= \frac{1}{N} \sum_{i: W_i=1} \lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N} \sum_{i: W_i=0} \lambda_i \cdot Y_i^{\text{obs}},\end{aligned}$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i}} = \begin{cases} 1/(1 - e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

IVW estimator in stratified randomized experiment

- Propensity score in each strata is $e(X_i = j) = P(W_i = 1 | X_i = j) = \frac{N_t(j)}{N(j)}$
- $\hat{\tau}_{\text{IPW}} = \frac{1}{N} \sum_{j=1}^K \left(\sum_{i: B_i=j} \frac{N(j)}{N_t(j)} W_i Y_i^{\text{obs}} - \sum_{i: B_i=j} \frac{N(j)}{N_c(j)} (1 - W_i) Y_i^{\text{obs}} \right) = \frac{1}{N} \sum_{j=1}^K N(j) (\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}})$
- Same as the estimator from Neyman's repeated sampling approach

Matching estimator

- In conditional randomized experiments, the IVW estimator do not have any further assumptions as the propensity scores $e(\mathbf{X}_i)$ are known.
- Instead of weighting based on $e(\mathbf{X}_i)$, we can also perform matching based on $e(\mathbf{X}_i)$
- We can match treatment and control unit to form a pair if their propensity scores are very close to each other
 - To assess the effect of job-training program on a thirty-year-old women with two children under the age of six, with a high school education and four months of work experience in the past 12 months, we want to compare her with a thirty-year-old women with two children under the age of six, with a high school education and four months of work experience in the past 12 months, **who did not attend the program**
- As $W_i \perp (Y_i(0), Y_i(1)) \mid e(\mathbf{X}_i)$, we can treat the matched data as from a paired randomized experiment

Causal inference with observational data

- The core rationale is to conceptualize observational studies as conditional randomized experiments
 - Analyze observational data as if treatment has been randomly assigned conditional on measured pre-treatment covariates X_i (unconfoundedness: $W_i \perp (Y_i(0), Y_i(1)) \mid X_i$)

Therefore “what randomized experiment are you trying to emulate?” is a key question for causal inference from observational data. For each causal effect that we wish to estimate using observational data, we can describe (i) the target trial that we would like to, but cannot, conduct, and (ii) how the observational data can be used to emulate that target trial.

-- *Causal Inference: What If* (Herman and Robins, 2020)

- Not all observational data can be conceptualized as a conditional randomized experiment!

Observational study V.S. conditional randomized experiments

1. Conditional randomized experiment: $W_i \perp (Y_i(0), Y_i(1)) \mid X_i$ is a fact as we control treatment assignment mechanism
Observational study: $W_i \perp (Y_i(0), Y_i(1)) \mid X_i$ is **an assumption**. It is always possible that this assumption is violated.
2. Conditional randomized experiment: $e(X_i) = P(W_i = 1 \mid X_i)$ is known
Observational study: $e(X_i) = P(W_i = 1 \mid X_i)$ needs to be estimated. Can introduce bias and suffer from estimation uncertainty

Need to evaluate identifiability assumptions carefully

- SUTVA
 - Can any variable have a causal effect? Are there multiple versions of assignment?
We need “sufficiently well-defined interventions”
Example: effect of sex, heart transplant by different techniques
 - Interventions may not be well defined as the experiment is not really conducted
- Overlap
$$e(\mathbf{X}_i) = P(W_i = 1 | \mathbf{X}_i) \in (0,1) \text{ or } P(W_i = w | \mathbf{X}_i = \mathbf{x}) > 0 \text{ for all } \mathbf{x} \text{ and } w$$
 - Guaranteed by the nature of experiments
 - Not guaranteed in observational studies
- L only contains pre-treatment covariates
- **Unconfoundedness:** $W_i \perp (Y_i(0), Y_i(1)) | \mathbf{X}_i$ is an **untestable assumption!!**

Estimate ATE with observation data

- We can still use outcome regression, IPW and matching estimators
- For IPW and matching estimators, as the propensity scores are unknown, we need to estimate the propensity scores from data first
- Once we estimate the propensity scores, we can replace the true propensity scores by their estimates in IPW or matching
- We need good estimates of the true propensity scores → not an easy task!
- We will also discuss other estimators that are more robust to a poor estimate of the propensity scores: blocking, trimming, doubly robust estimator

Propensity score estimation procedure

What is the criteria of a good estimated propensity score?

- Estimate $e(\mathbf{X}_i) = P(W_i = 1 | \mathbf{X}_i)$: a classification problem but not exactly a classification problem
 - The goal is not simply minimizing the mean square error or classification error
 - A good propensity score needs to achieve covariates balancing $W_i \perp \mathbf{X}_i | \hat{e}(\mathbf{X}_i)$
 - Even if $\hat{e}(\mathbf{X}_i)$ is NOT an accurate estimate of the true $e(\mathbf{X}_i)$, as long as it achieves covariates balancing, $\hat{e}(\mathbf{X}_i)$ is at least a balancing score which leads to unconfoundedness given $\hat{e}(\mathbf{X}_i)$
- Two stages to estimate the propensity score:
 - 1) Use an initial specified model, such as logistic regression, to obtain $\hat{e}(\mathbf{X}_i)$
 - 2) Check covariate balancing based on weights or matched sets defined by $\hat{e}(\mathbf{X}_i)$
 - 3) We can iterate back and forth between the above two stages, each time refining the specified model
- During the whole process, we do not use the outcome data Y_i^{obs}

The Barbiturate exposure data

- We aim to evaluate the effect of prenatal exposure to barbiturates
- The data set contains information on $N = 7,943$ men and women born between 1959 and 1961 in Copenhagen, Denmark.
- $N_t = 745$ men and women had been exposed in utero to substantial amounts of barbiturates due to maternal medical conditions. The comparison group consists of $N_c = 745$ individuals from the same birth cohort who were not exposed in utero to barbiturates.
- Outcome: barbiturate exposure on cognitive development in later years
- Treatment and control group can be systematically different: dataset contains 17 **pre-treatment** covariates that can potentially relate to both cognitive development and likelihood of being exposed to barbiturates

The Barbiturate exposure data

Table 13.1. Summary Statistics Reinisch Data Set

Label	Variable Description	Controls		Treated		t-Stat
		($N_c = 7198$)	Mean (S.D.)	($N_t = 745$)	Mean (S.D.)	
sex	Sex 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	

Logistic regression: specify a model to obtain $\hat{e}(\mathbf{X}_i)$

- Logistic regression is an extension of linear regression to regression binary response variable W_i on the predictors $\tilde{\mathbf{X}}_i$
 - Here, the predictors $\tilde{\mathbf{X}}_i$ is not necessary the original set of pre-treatment covariates \mathbf{X}_i , we may drop some irrelevant covariates and add interaction terms
 - Logistic regression assumes the model

$$\pi_i = P(W_i = 1 | \tilde{\mathbf{X}}_i) = \frac{e^{\alpha + \boldsymbol{\beta}^T \tilde{\mathbf{X}}_i}}{1 + e^{\alpha + \boldsymbol{\beta}^T \tilde{\mathbf{X}}_i}}$$

or equivalently, $\text{logit}\left(P(W_i = 1 | \tilde{\mathbf{X}}_i)\right) = \alpha + \boldsymbol{\beta}^T \tilde{\mathbf{X}}_i$

- It also assumes that $W_i \sim \text{Bernoulli}(\pi_i)$
- The log-likelihood function of the above model is

$$\sum_{i=1}^N W_i(\alpha + \boldsymbol{\beta}^T \tilde{\mathbf{X}}_i) - \ln(1 + \exp(\alpha + \boldsymbol{\beta}^T \tilde{\mathbf{X}}_i))$$

- We maximize the likelihood to obtain estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$, and $\hat{e}(\mathbf{X}_i) = \frac{e^{\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{X}}_i}}{1 + e^{\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{X}}_i}}$

Selecting the covariates and interactions

- We can not include all 17 covariates and their $17*18/2 = 162$ quadratic and interactions terms in the logistic regression, and want to select a subset of these terms

Step 1: select a subset of basic covariates based on scientific understanding

- Basic covariates: covariates that are a priori viewed as important for explaining the assignment and plausibly related to some outcome measures
- In the Barbiturate exposure data
 - lmotage: mother's age, which is plausibly related to cognitive outcomes for the child
 - ses: mother's socio-economic status, which is strongly related to the number of physician visits during pregnancies and thus exposes the mother to greater risk of barbiturate prescriptions
 - sex: sex of the child, which may be associated with measures of cognitive outcomes

Variable	EST	(s.e.)	t-Stat
Intercept	-2.38	(0.06)	-41.0
sex	-0.01	(0.08)	-0.2
lmotage	0.48	(0.04)	11.7
ses	0.10	(0.04)	2.6

Selecting the covariates and interactions

Step 2: add additional linear terms

- For each of the covariate not yet added, calculate the likelihood ratio statistics assessing the null hypothesis that the newly included covariate has a zero coefficient
- Add the covariate with the largest likelihood ratio statistics
- Stop if all likelihood ratio statistics of the remaining covariates are smaller to a cutoff (Say $C_L = 1$)
- Similar to forward stepwise regression

Selecting the covariates and interactions

Table 13.4. Likelihood Ratio Statistics for Sequential Selection of Covariates to Enter Linearly; Barbiturate Data

Covariate	Step →											
	1	2	3	4	5	6	7	8	9	10	11	12
sex	–	–	–	–	–	–	–	–	–	–	–	–
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	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	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	0.0
ses	–	–	–	–	–	–	–	–	–	–	–	–
sib	21.0	13.8	12.5	15.0	15.7	–	–	–	–	–	–	–

Selecting the covariates and interactions

Step 3: add additional quadratic and interaction terms

- Say we now have K_L linear covariates selected
- Quadratic and interaction terms are $K_L(K_L + 1)/2$
- Actual quadratic and interaction terms can be less as quadratic of binary covariate is itself
- Follow the same procedure in step 2 to add these terms sequentially
- There can be a different cutoff for the likelihood ratio statistics C_Q (say $C_Q = 2.71$, corresponding to a 10% significance level)

Final model for the estimated propensity score

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