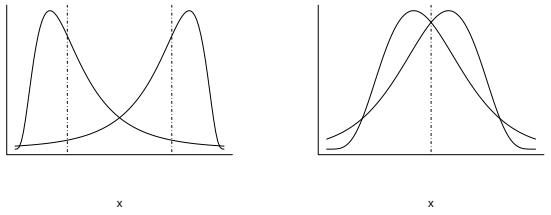
Ideally we have two groups similar in everything but the treatment experienced (and treatment assignment is ignorable).

Imbalance: Distributions of relevant pre-treatment variable differ. Can't just compare means: $\overline{y}_1 - \overline{y}_0$; instead must adjust for pre-treat differences.

Incomplete Overlap: For some regions in the pre-treatment space there are no controls but some treatment subjects, or vice versa.
Limits information available in regions of non-overlap. Must either scale back to get overlap, or model the missing pretreatment values via extrapolation.

Must rely more heavily on models, less on direct info in the data.

Imbalance and Model Sensitivity



Both plots show imbalance – any difference in distribution, not just a means shift.
Try to make inference on y controlling for pretreatment x.

$$\begin{aligned} \text{trt: } y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \theta + \epsilon_i \\ \text{cntrl: } y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \end{aligned}$$

Average and subtract (average error is close to 0?):

$$\overline{y}_1 - \overline{y}_0 = \beta_1(\overline{x}_1 - \overline{x}_0) + \beta_2(\overline{x_1^2} - \overline{x_0^2}) + \theta$$

Estimating θ with Imbalance

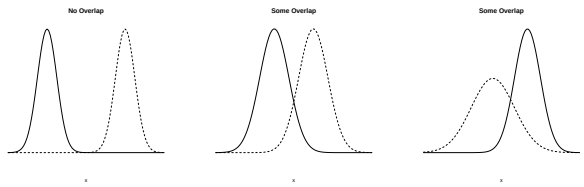
Solve for $\theta = \overline{y}_1 - \overline{y}_0 - \beta_1(\overline{x}_1 - \overline{x}_0) - \beta_2(\overline{x_1^2} - \overline{x_0^2})$

What's the bias if we ignore x?

What makes it big/small?

What should happen under randomized assignment?

Non-complete Overlap



Complete overlap means the range of pretreatment x is the same for both groups.
Relates how to counterfactuals?
Incomplete overlap means we must:

- Limit range of any causal inference or
- Trust some model extrapolations

Child Care Example

Low birth weight babies may have lower IQ scores, so an intervention was designed to stimulate them – high quality child care 5 days/week.
Response: intelligence at age 3
290 kids in treatment group weighted < 2500g at birth,
4901 controls.
Missing data imputed once meaning you build a model for each predictor which has missing values based on other predictors and fill in the hole with a random draw having the right mean and variance. (Better to repeat analysis many times imputing new random values each time.)

Child Care Plot 10.4

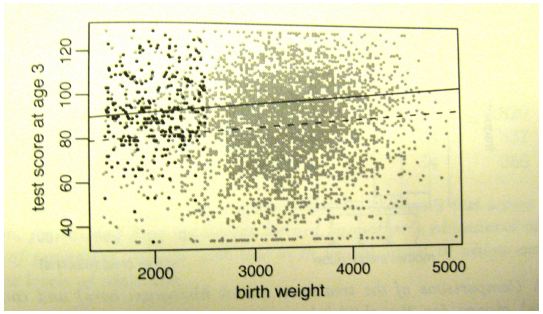


Figure 10.4

Plot exhibits what problem for covariate "birth weight"?

Desired inference is for kids who might have been given the treatment, so we can limit scope of inference to those we "intended to treat"

Imbalance vs Non-overlap

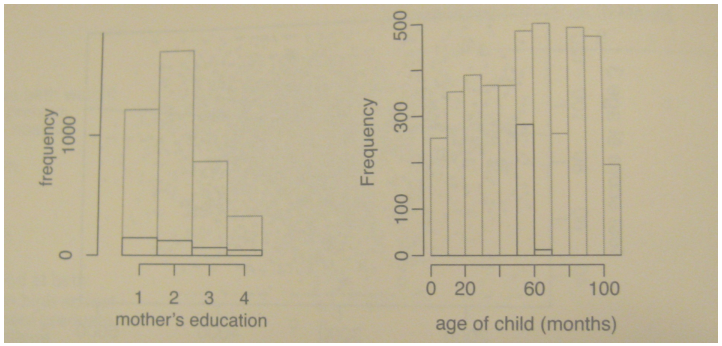


Figure 10.5 (a) complete overlap for mom's education, but largest category differs. (b) Balance in age, but treated kids are all very close. If age is important, we should drop most of the controls.

One problem does not mean the other also exists for these data. Lack of overlap is more serious – limits what type of inference?

§10.2 Subclassification

If we assume ignorability, how to check if we are relying to much on modeling assumptions? Or what if we question the assumptions? Suppose we have one predictor (mom's education) associated with both participation and response. How to get causal effects? Compute treatment effect within each education category.

Mother's education	Treatment effect estimate $\pm$ s.e.	Sample size treated	controls
Not a high school grad	9.3 ± 1.3	126	1358
High school graduate	4.0 ± 1.8	82	1820
Some college	7.9 ± 2.3	48	837
College graduate	4.6 ± 2.1	34	366

Figure 10-6. Treatment effects within strata of education.

Matching
Classic Example: Case – Control Studies

Question: Does exposure to a risk, E, cause disease D?
(Not ethical to run an experiment)
Prospective study: follow people thru time recording exposure E and disease D.
But if D is rare, the number of **cases** with disease will be small.
Retrospective study: Sample from people with disease (**cases**) and without (**controls**). Look back at histories to see who was exposed or not.
Typically, controls are easy to find so we might match one case to several controls.
Cases are sampled from people entering hospital with disease D.
Controls are sampled from same hospital with unrelated reason for admission. Match each case to a control of similar gender, age, SES, region of country, ethnicity, ...

§10.3 Matching

Gelman & Hill start with a big sample and match within to create overlap. Note: match on pretreatment variables only.

One-to-one: divide sample into pairs of similar units - one treated, the other control.

May not be able to match everyone due to lack of overlap, so then some subjects are not used (control, or treatment or both).

Use differences in means or regression to estimate treatment effect.

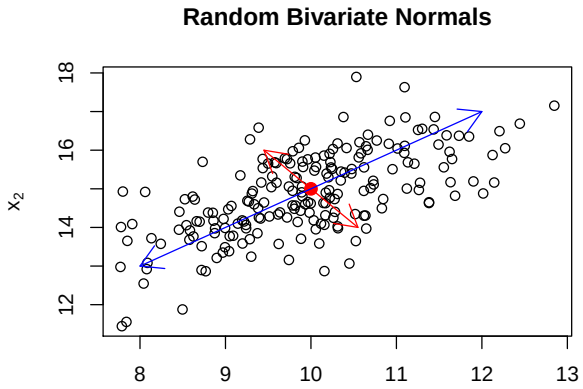
Similar to stratifying, but more precise when x is continuous.

Distance Based Matching

Build a vector of covariates – a row of an $\mathbf{X}$ matrix.

How similar are these vectors for two different people? We'd like to pair people who are close together.

What do we do when different covariates have different amounts of spread? When covariates are correlated?



Propensity Score Matching

We want to assume ignorability, so look for variables which differ between the control and treatment groups. Combine them all together in some way to get a single score telling us how different the 2 groups are.
Logistic regression of pretreatment covariates on $T = 1$ if in treatment group, 0 otherwise. The combination of covariates which best models the logistic regression is the propensity score.

$$\widehat{Pr(T = 1|\mathbf{x})} = \beta_0 + \widehat{\beta_1 x_1} + \widehat{\beta_2 \mathbf{x}_2} + \widehat{\beta_3 \mathbf{x}_3}$$

Build a model, compute propensities for each treated unit, pair it to one control which is closest in propensity.
Goal: to get matched groups which are similar on average across all covariate values. (Not: each pair of matched obs's is very similar in all covariates.)

Example with Lalonde Data

```
data(lalonde);  attach(lalonde)
fit <- glm(treat ~ re74 + re75 + black + hisp + married + u74 + u75,
           family=binomial)
pscores <- predict(fit, type="response")
matches <- matching(z=lalonde$treat, score=pscores)
matched <- lalonde[matches$matched,]
b.stats <- balance(lalonde, matched, fit) # print(b.stats)
```

Predict on the "response" scale gives `invlogit(X.beta.hat)`, a value between 0 and 1.
Alternatively the "linear" scale is $\mathbf{X}\widehat{\beta}$, in the linear (logit) scale.
The `matching` function from `arm` pairs up observed treatment (0 or 1) with propensity scores.

Lalonde Data Output

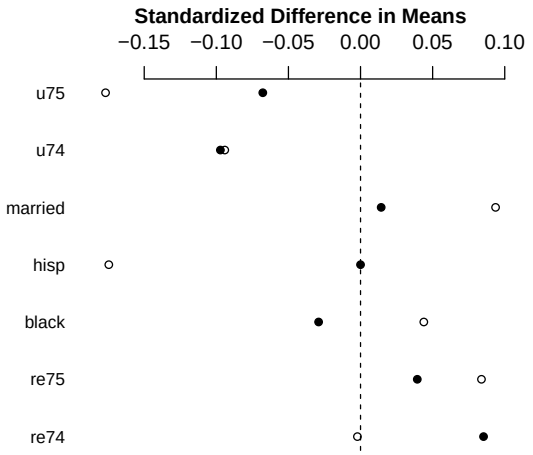
Unmatched Mean differences

	Treat	control	diff	diff.std	se	sd
re74	2095.57	2107.03	-11.45	-0.00	503.50	5302.42
re75	1532.06	1266.91	265.15	0.08	305.04	3161.65
black	0.84	0.83	0.02	0.04	0.04	0.37
hisp	0.06	0.11	-0.05	-0.17	0.03	0.28
married	0.19	0.15	0.04	0.09	0.04	0.38
u74	0.71	0.75	-0.04	-0.09	0.04	0.44
u75	0.60	0.68	-0.08	-0.18	0.05	0.48

Matched Mean differences

	Treat	control	diff	diff.std	se	sd
re74	2095.57	1643.29	452.28	0.09	503.50	5302.42
re75	1532.06	1407.61	124.44	0.04	305.04	3161.65
black	0.84	0.85	-0.01	-0.03	0.04	0.37
hisp	0.06	0.06	0.00	0.00	0.03	0.28
married	0.19	0.18	0.01	0.01	0.04	0.38
u74	0.71	0.75	-0.04	-0.10	0.04	0.44
u75	0.60	0.63	-0.03	-0.07	0.05	0.48

Lalonde Plot



More Propensity

Propensity score does not involve response y at all.
We do build a model, and can check it by seeing how well it balances treatment to control.

Note: Figure 10.7 p 209 Distribution of Propensities after matching are very similar in treated and control groups.

Gives another option to an extrapolated regression.

What if some treated subjects are not well matched in controls? Could drop them, but that changes the scope of inference to a subpopulation. How are the ones we're dropping different?

Matching this way does affect variance calculations.

Other Matching Methods

We have used *matching without replacement* using each control at most one time. Can instead allow multiple treated units to match the same control.

Other Uses of Propensity

Use inverse propensity as weights in weighted regression and use the whole sample.

Standard errors are not quite right when we dropped some observations.

- Pairs matched are correlated.
  Fix by using the same predictors in the linear model estimating treatment effects.
- Propensity score is an estimate, but is treated like observed data. How do we incorporate the uncertainty?

§10.4 Regression Discontinuity

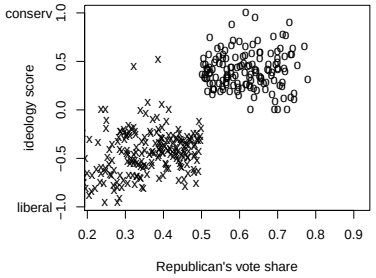
Regression of pretreatment covariates provides a clean and more precise estimate of treatment effect if we have complete overlap.

Special case of "no overlap": when another variable explains the lack of overlap.
Example: only apply a medical treatment to patients whose blood pressure is below threshold C.
Or only try experimental new treatment if their cancer has advanced to stage C and they have no other recourse.
Elections: candidate wins only if % vote > 50. "Treatment" is candidate's party, x is Republican vote share.

Can we assume regression function is continuous at C?
We don't have to assume ignorability since the mechanism for assigning treatment is known.

May only apply for x values near C.

Political Ideology



Why is there a bump up at 0.5? Consider $x \in (0.45, 0.55)$ what is the "party" effect?

	Estimate	Std. Error	t value
(Intercept)	-1.21	0.62	-1.94
party	0.73	0.07	9.78
x	1.65	1.31	1.26

Table: n = 68 rank = 3 resid sd = 0.151 R-Squared = 0.883

S 10.5 Instrumental Variables

Assumption of ignorability rests on having adjusted out all inputs which predict both treatment and response outcome. What if we don't have access to the "lurking" variables?

An **Instrument Variable** (IV) is thought to randomly induce variation in the treatment variable.

Example: With the Sesame Street data we really cared about the effect of *regular* watching of the program. Could use the randomly assigned *encouragement* as an instrumental variable on *regular* watching.

Assumptions for Instrumental Variables

(Using a binary treatment and binary instrument.)

- Instrument is ignorable
- Nonzero association between IV and treatment
- Monotonicity
- Exclusion restriction

Assumption 1: Ignorability

Earlier we assumed treatment assignment was ignorable, that is, not dependent on pre-treatment variables – known or unknown.

Now make the same assumption for the IV – that it is not explained by some lurking variable.

Assumption 2: Association

For the IV to be useful it must be associated with the treatment.

With the Sesame Street study, encourage was the treatment, but if we want to consider causal effects of regular watching, encourage is useful because it is associated with regular. Of those encouraged, 90% were regular watchers, and of those not encouraged, only 55% were regular watchers.

We observed an increase of 90 - 55 = 35% in the percent of regular viewers which may be attributed to encouragement. These people seemed to have become regular viewers due to the intervention.

Assumption 3: Monotonicity

Children in the Sesame study may be classified as "always-watchers", "never-watchers" or "induced-watchers". We assume no backtracking – that encouragement does not influence some watchers to stop watching.

Often defensible, but need not hold in practice.

Assumption 4: Exclusion Restriction

- always-watchers were not influenced by encouragement
- never-watchers were not influenced by encouragement

Assume all the effect of the IV happens in the induced-viewers, not the other two groups.
Violation of Exclusion Restriction:
Parents believe TV is harmful to kids and won't let them watch Sesame Street or anything else. If they are encouraged with info about the beneficial effects of the show, they might compensate by obtaining other educational materials, or by more reading aloud with the kids.

Deriving an Instrumental Variable
Assume the exclusion restriction and complete data

Figure 10.9,p 218: for each kid in the study we know

- regular (T_i)
- outcome test score
- c_i = always-watcher or never-watcher or induced-watchers
- encouragement (randomly applied) and
- counterfactuals y^0, y^1 (equal for groups 1 and 2)

True Intent-to-Treat effect is based on mean differences, $\overline{\Delta}$, in the induced watchers group.

$$ITT = \overline{\Delta} \cdot \text{proportion of induced watchers} + 0$$

Here that's $8.5 \cdot 0.4 = 3.4$

Instrumental Variable via Regression

Estimate proportion induced to watch the show.

	Estimate	Std. Error	t value
(Intercept)	0.55	0.04	13.43
encour	0.36	0.05	7.10

Table: n = 240 rank = 2 resid sd = 0.381 R-Squared = 0.175

	Estimate	Std. Error	t value
(Intercept)	24.92	1.42	17.54
encour	2.88	1.79	1.61

Table: n = 240 rank = 2 resid sd = 13.331 R-Squared = 0.011

Intent-to-treat estimate is 2.88. Divide by 0.36 to get Wald estimate of the effect of regular viewing: 7.93

Local Average Treatment Effects

Causal inference on effect of Sesame Street watching does not depend on always-watchers or never-watchers.
Causal inference is valid only for induced watchers.

Local average treatment effect (LATE) is the average treatment effect for those induced to change behavior. Estimate above is a special case of a LATE estimator.

ITT may be better for setting policy because they more accurately reflect the fact that not all targeted subject will participate. But that assumes compliance rates don't change.
G & H recommend estimating ITT and LATE.

IV in regression models

Let IV $= z$. We have two equations:

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 z + \epsilon \\ T &= \gamma_0 + \gamma_1 z + \nu \end{aligned} \tag{4}$$

Assuming $\epsilon \perp\!\!\!\perp \nu$ is equivalent to ignorability and exclusion restriction assumptions.
"The IV affects response only through its effect on treatment variable."
Nonzero correlation between T and z.

Identifiability

Do the data contain enough info to provide a unique estimate of all parameters? Without exclusion restriction, we can't separate T and z effects.

Substitution in equation (4) gives

$$\begin{aligned} y &= (\beta_0 + \beta_1\gamma_0) + (\beta_1\gamma_1 + \beta_2)z + \text{error} \\ &= \delta_0 + \delta_1 z + \text{error} \end{aligned}$$

Estimate δ_1 by regressing y on z, then solve for β_1

$$\beta_1 = (\delta_1 - \beta_2)/\gamma_1 \text{ typo in book: not } \gamma_2$$

γ_1 is estimated from regression of T on z.
What about β_2? (Effect of IV given T is in the model) By the exclusion restriction, it is 0.

Two Stage Least Squares

1. Regress T on z, here regular on encour and get fits, $\widehat{T}$.
2. Regress y on $\widehat{T}$, test score on predicted watching.

```
fit.3a <- lm (regular ~ encour + prelet + factor(site) + se
watched.hat <- fit.3a$fitted
fit.3b <- lm (postlet ~ watched.hat + prelet + factor(site)
display.xtable(fit.3b)
```

	Estimate	Std. Error	t value
(Intercept)	1.22	4.77	0.26
watched.hat	14.03	4.04	3.47
prelet	0.70	0.08	8.91
factor(site)2	8.40	1.83	4.60
factor(site)3	-3.94	1.81	-2.18
factor(site)4	0.94	2.45	0.38
factor(site)5	2.76	2.89	0.96
setting	1.60	1.48	1.08

Continuous IV and Treatments

Theoretically no problem, we just regress the treatment on the IV, then the response on predicted treatment. Interpretation?
Cannot estimate dosage effects like "How much Sesame Street does a kid need to watch to raise IQ 10 points?"
When do we get **ignorability** of the instrument?

- Draft Lottery in late 1960's as instrument for treatment "military service" on response "earnings" or "health".
- Weather as instrument for supply of fish on price.
- Sex of 2nd child on effect of # of kids on labor supply

Plausibility of Assumptions

First stage model regresses treatment on IV. This should show a strong relationship – in the right direction. We can check this.
Cannot check ignorability of instrument or exclusion restriction.

Structural Equations Models (SEM)

Sometimes used to estimate "causal" relationships in complex observational data.
We've seen enough issues with untestable assumptions to feel uneasy about these approaches.

§10.7 Within or Between Groups

Can we use repeated measurement on individuals over time to control for lurking variables?
Hard to focus on the health effects of a baby being born at low weight – tied up with SES and other factors. But twins can act as counterfactuals to each other, so if we look at twins of different birth weights, we could perhaps estimate the effects of low weight.
Fixed Effects: are repeatable. are the only coefficients of interest. whereas
Random effects represent some population of effects. are not repeatable. coefficient values are not of interest, just variances are.
Multilevel models: very useful for repeated measurements on the same units. Some variables stay constant across time, others vary with time.
Finally: Control for pretreatment variables only.

Continuous IV and Treatments

Theoretically no problem, we just regress the treatment on the IV, then the response on predicted treatment. Interpretation?
Cannot estimate dosage effects like "How much Sesame Street does a kid need to watch to raise IQ 10 points?"
When do we get **ignorability** of the instrument?

- Draft Lottery in late 1960's as instrument for treatment "military service" on response "earnings" or "health".
- Weather as instrument for supply of fish on price.
- Sex of 2nd child on effect of # of kids on labor supply

Plausibility of Assumptions

First stage model regresses treatment on IV. This should show a strong relationship – in the right direction. We can check this.
Cannot check ignorability of instrument or exclusion restriction.

Structural Equations Models (SEM)

Sometimes used to estimate "causal" relationships in complex observational data.
We've seen enough issues with untestable assumptions to feel uneasy about these approaches.

§10.7 Within or Between Groups

Can we use repeated measurement on individuals over time to control for lurking variables?
Hard to focus on the health effects of a baby being born at low weight – tied up with SES and other factors. But twins can act as counterfactuals to each other, so if we look at twins of different birth weights, we could perhaps estimate the effects of low weight.
Fixed Effects: are repeatable. are the only coefficients of interest. whereas
Random effects represent some population of effects. are not repeatable. coefficient values are not of interest, just variances are.
Multilevel models: very useful for repeated measurements on the same units. Some variables stay constant across time, others vary with time.
Finally: Control for pretreatment variables only.