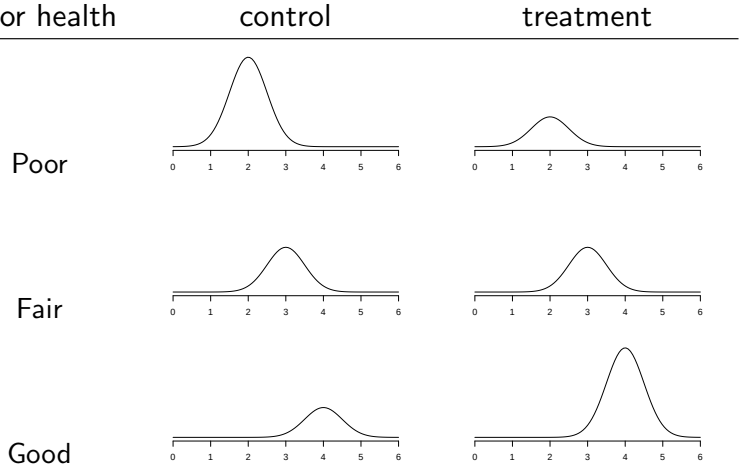
Chapter 9 Causal Inference in Regression

So far *predictive inferences*: building models to predict **y** given **X**.
Consider treatment variable $T_i = 1$ for those getting treatment, 0 for controls. (Or continuous level of treatment, T_i)
I want to know if the treatment is worth the risk of side effects. Will it improve **my** condition? Not: "On average does treated group do better than controls?" but I want to compare within one patient: "Will the treatment cause an improvement for me?" This is comparing what would happen to the same person under different treatment options, yet only one will be assigned.

Assignment Issues 1

Reasons to Randomize

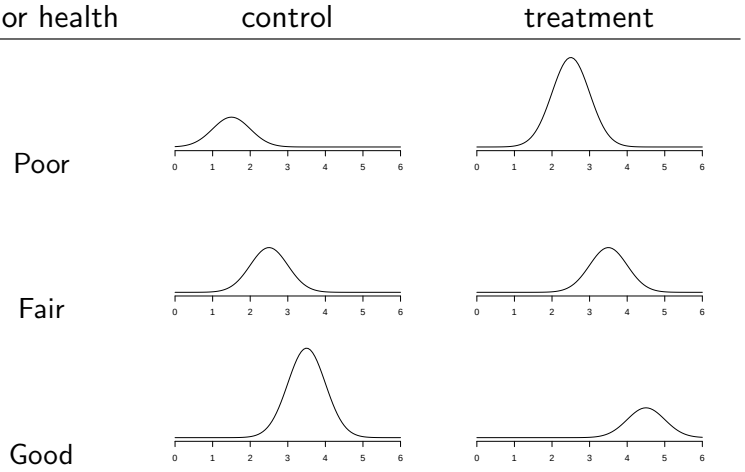
Scenario 1: No treatment effect, but sicker patients get control, healthier ones the treatment. It then looks as if the treatment is effective.



Assignment Issues 2

Reasons to Randomize

Scenario 2: There is a treatment effect, but sicker patients get treatment, healthier ones the control. It then looks as if the treatment is ineffective.



Lurking Variables

Above we could use pre-post measures to account for initial conditions as a covariate. Or compare only within the same health class.

But what if we fail to measure an important predictor? Then we have a bias in the results.
Suppose the correct model is: $y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$ where T_i is a treatment indicator.
Fail to measure covariate x_i and fit star model:
$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i$.
If x is also affected by T, then $x_i = \gamma_0 + \gamma_1 T_i + \nu_i$ and the full model is also: $y_i = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1) T_i + \epsilon_i + \beta_2\nu_i$
Conclusion: the treatment effect in the star model is $\beta_1^* = \beta_1 + \beta_2\gamma_1$. When is it β_1? If $x \perp\!\!\!\perp T$?

Bias from underfitting

Traditional explanation of underfitting where the "true" model is

$$\mathbf{y} = [\mathbf{X}_1\ \mathbf{X}_2] \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$$

but we omit $\mathbf{X}_2$ and $\boldsymbol{\beta}_2$. Then the estimate for $\boldsymbol{\beta}_1$ is $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1\mathbf{X}_1{}^\mathsf{T})^{-1}\mathbf{X}_1{}^\mathsf{T}\mathbf{y}$ which has expected value

$$E(\hat{\boldsymbol{\beta}}_1) = (\mathbf{X}_1\mathbf{X}_1{}^\mathsf{T})^{-1}\mathbf{X}_1{}^\mathsf{T}[\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2] = \boldsymbol{\beta}_1+(\mathbf{X}_1\mathbf{X}_1{}^\mathsf{T})^{-1}\mathbf{X}_1{}^\mathsf{T}\mathbf{X}_2\boldsymbol{\beta}_2.$$

If $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ then $\mathbf{X}_1{}^\mathsf{T}\mathbf{X}_2 = \mathbf{0}$ and there is no bias, but in observational studies, that is never the case, so there is some bias in the estimation of $\boldsymbol{\beta}_1$ dependent on $\mathbf{X}_1$, $\mathbf{X}_2$, and the true $\boldsymbol{\beta}_2$.

Gelman suggests this is not a problem unless we are wanting to make causal inference. I see it as a problem when the "true" $\boldsymbol{\beta}_1$ needs to be adjusted for $\boldsymbol{\beta}_2$ and $\mathbf{X}_2$ belongs in the model. Does Nature build complex models with lots of predictors?

§9.2 The Fundamental Problem of Causal Inference

In an experiment,

- assign one treatment to one subject
- measure the response
- we'll never know what would have occurred under another treatment.

But we could imagine that other response. (a counterfactual)
Fundamental Problem:
We'd like to observe two responses: y_i^1 under treatment and y_i^0 under control, then take the difference $y_i^1 - y_i^0$ = treatment effect within patient i.
Can't do that, as we observe one or the other, not both.

Conceptual Model

unit	Covariates			Treatment	y_i^1	y_i^0	$y_i^1 - y_i^o$
1	2	1	59	0	69	75	6
2	3	1	98	0	111	108	-3
3	2	2	80	1	92	102	10
4	3	1	98	1	112	111	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	111	114	3

Circle the responses we will actually observe.

Observed data

unit	Covariates			Treatment	y_i^1	y_i^0	$y_i^1 - y_i^o$
1	2	1	59	0	69	?	?
2	3	1	98	0	111	?	?
3	2	2	80	1	?	102	?
4	3	1	98	1	?	111	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	?	114	?

What to Do?

We need:

- **Close substitutes for the unobservable values**
  A cross-over experimental design does provide both y_i^0 and y_i^1, but we must assume that there is no carry over effect and that the order of application does not matter (no learning effect). Strong assumptions are needed.
- **Randomization**
  Use a random sample of subjects from the population of interest.
  Randomly allocate treatments. Difference in mean responses is an unbiased estimator of the true treaatment effect.
- **Statistical adjustment**
  Sometimes randomization is impossible or unethical.
  Build a model based on covariates to predict how subjects would have reponded to the other treatment. Again strong assumptions (and expert opinion) are needed. See Chapter 10.

§9.3 Randomized Experiments

- Obtain a random sample of units from the population of interest.
- Randomly allocate treatments

Then $\overline{y}_{\cdot}^1 - \overline{y}_{\cdot}^0$ is an unbiased estimator of average treatment effect $\text{avg}(y_i^1 - y_i^0)$.
β_1 is estimable in $y_i = \beta_0 + \beta_1 T_i + \epsilon_i$.
Assuming:

- Every unit had a non-zero probability of being assigned each treatment.
- Control group is just like the treated group except that they did not receive treatment.

Sampling Scenarios

Scenario 1:

- Take a sample of size n_0 from the population of all control observations (and these people get the control).
- Take a sample of size n_i from all treated observations (and give these guys treatment)

You end up with a clean comparison of two subpopulations.
$\overline{y}_{\cdot}^1 - \overline{y}_{\cdot}^0$ has SE $\sqrt{s_0^2/n_0 + s_1^2/n_1}$

Scenario 2: Sample $n_0 + n_1$ from the population and randomly assign treatments. We end up with the same estimator. Each group acts as counterfactual for the other.

Real Life

More typically, the subjects available are a convenience (not random) sample. E.g. clinical studies, experiments on lab rats or Psych 101 students.

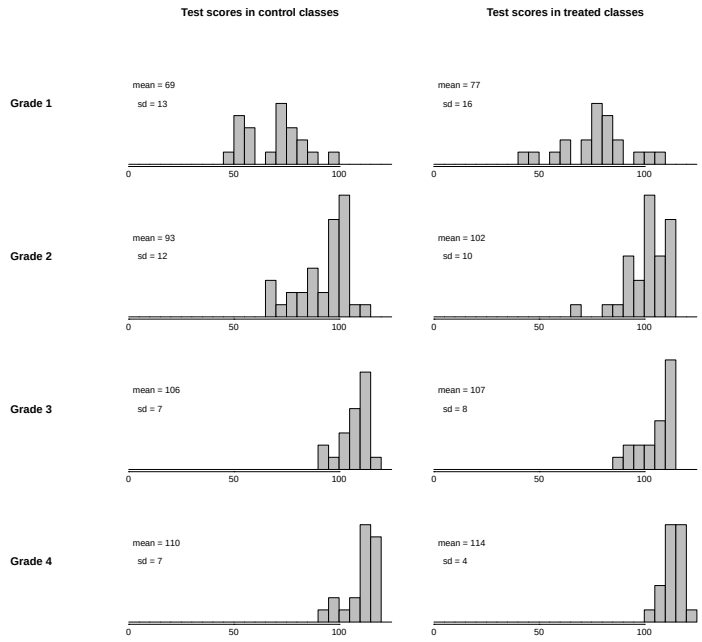
Conclude: experiment can show there is a causal relationship within this hypothetical subpopulation (internal validity). Extension to broader population is possible through modeling (external validity).

Electric Company 1

Does *The Electric Company* TV show improve test scores of primary grade children?
Design:
Randomized Block Design with two cities = blocks.
Within each city and for each of grades 1, 2, 3, 4, pick 20 schools.
Within each school/grade pick 2 classes.
Randomly assign treatment or control within pairs.
Note: each observation is an average test score for a classroom.
For now ignore the pairing within school and town, pretend they are independent, completely randomized design.
Fit model:

$$y_i = \alpha + \theta T_i + \epsilon_i$$

Electric Company Response Data



Fit 4 models

```
results <- NULL
 for (k in 1:4){
    results <- rbind(results, summary(lm(post.test ~ treatment, subset=
 results[c(1,3,5,7),4] <- NA
 rownames(results) = paste( rep("Grade",8), rep(1:4,each=2),  rownames(
 xtable(results)
```

	Estimate	Std. Error	t value	Pr(>|t|)
Grade 1 (Intercept)	68.79	3.27	21.05	
Grade 1 treatment	8.30	4.62	1.80	0.08
Grade 2 (Intercept)	93.21	1.91	48.88	
Grade 2 treatment	8.36	2.70	3.10	0.00
Grade 3 (Intercept)	106.17	1.66	63.86	
Grade 3 treatment	0.33	2.35	0.14	0.89
Grade 4 (Intercept)	110.36	1.30	84.98	
Grade 4 treatment	3.71	1.84	2.02	0.05

Include Pre-treatment Covariates

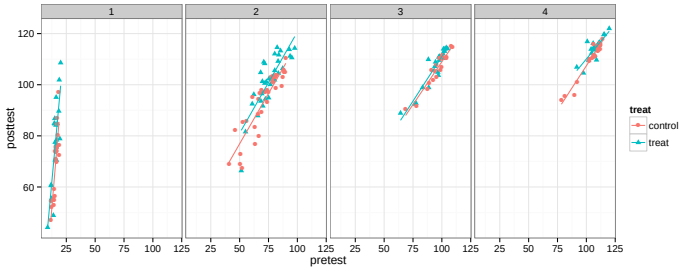
Large SE's above. Reduce noise by fitting covariates. New model

$$y_i = \alpha + \theta T_i + \beta x_i + \epsilon_i$$

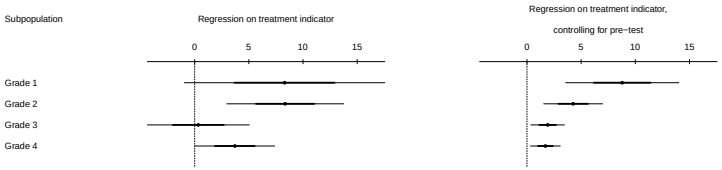
```
tallElectric <- with(electric,data.frame( city=rep(City,2), grade=fac
```

```
qplot(x=pretest, y=posttest, data = tallElectric, shape=treat, col=trea
      geom = c("point", "smooth"), method="lm", se = FALSE ) + theme_bw
   facet_grid(. ~ grade)
```

Pre-treatment Plot



Compare Effect Estimates



Estimates have much less error after adjusting for **pre-treatment** predictors. Include covariates measured before treatment (or unaffected by treatment).
The treatment is **causing** improvements in test scores, though the improvement is less in the higher grades. (Same test throughout, so less room for improvement in grades 3 and 4.)

Gain Scores

Gain = post-test - pretest. $g_i = y_i = x_i$. Assumes $\beta = 1$.
Model: $g_i = \alpha + \theta T_i + \epsilon_i$
Regression estimate: $\widehat{\theta} = \overline{g}^T - \overline{g}^C$

Or use $g_i = \alpha + \theta T_i + \gamma x_i + \epsilon_i$ to adjust for non-one β.

More Levels of Treatment

With more treatment levels, label one baseline, and compare each other level to baseline just like treatment to control above.

Numerical treatments (level of fertilizer chosen at random from an interval)

No Interference

Often overlooked assumption:
Treatment assignment for subject i does not affect the outcome for any other subject. SUTVA = Stable Unit Treatment Value Assumption. Otherwise life is much more complicated – we have to consider the entire list of who got which treatments in our model.

Ag experiments, plants do interact within root extent. Need to model spatial interactions.

Interactions

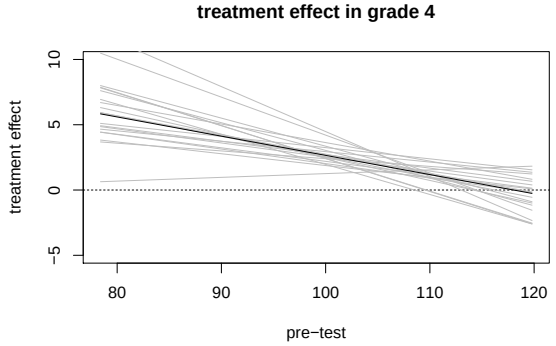
Why assume the treatment effect is the same for all units? It might depend on the pre-treatment predictor.
Our model for Electric Company experiment used two parallell lines within each grade. They might have different slopes. How do we interpret treatment effects in this case? Grade 4, 3 models:

	Estimate	Std. Error	t value	Pr(>|t|)
1 Intercept	110.36	1.30	84.98	
1 Treatment	3.71	1.84	2.02	0.05
2 Intercept	41.99	4.28	9.81	
2 Treatment	1.70	0.69	2.48	0.02
2 PreTest	0.66	0.04	16.07	0.00
3 Intercept	37.84	4.90	7.72	
3 Treatment	17.37	9.60	1.81	0.08
3 PreTest	0.70	0.05	14.86	0.00
3 Interaction	-0.15	0.09	-1.64	0.11

Interpretation

Treatment effect is $17.37 - 0.15x$. Note x varies from 80 to 120, treatment effect is 5 on the left to -1 on the right (assuming the coefficients are exactly right).

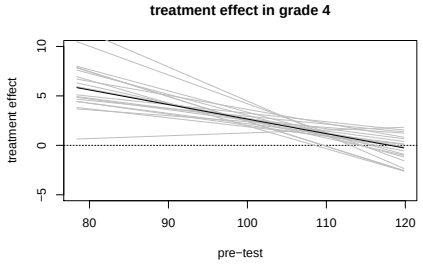
Include variation from sampling distribution of $\widehat{\beta}$.



Treatment effect is greater for those who start lower in pretest.

Post-stratification

With interactions, treatment effect depends on the value of the covariate, as with:



Pretest	Treatment Effect
80	5
120	-1

Could use the "average" treatment effect for all x values in the sample. That's post-stratification. Example:

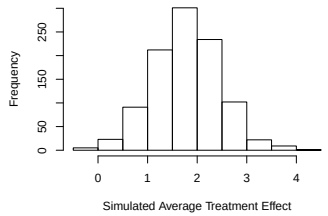
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 T + \beta_4 x_1 T + \beta_5 x_2 T + \epsilon$$

Treatment effect is: $\beta_3 + \beta_4 x_1 + \beta_5 x_2$. Plug in sample means $\overline{x}_1$ and $\overline{x}_2$ Or, if known, μ_{x_1}, μ_{x_2}. Get SE's by simulation.

Post-Stratify Electric Company

Compute the average treatment effect & summarize

```
lm.4.sim <- sim(lm(post.test ~ treatment*pre.test,subset= grade == 4 )
effect <- array (NA, c(1000, sum(grade==4)))
for (i in 1:1000){
 effect[i,] <- lm.4.sim@coef[i,2] + lm.4.sim@coef[i,4]*pre.test[grade=
}
avg.effect <- rowMeans (effect)
## c(mean(avg.effect), sd(avg.effect))            1.791 0.659
hist(avg.effect, main="", xlab="Simulated Average Treatment Effect")
```



§9.5 Observational Studies

Experiments on random samples with random treatment allocation are "gold standard" in terms of causal inference. We can then estimate average differences of counterfactuals.
BUT, observational studies are more common in many areas due to ethical, logistic, and other problems. (More realistic)
Problem:
Lurking variables may differ between the control and treatment groups.
Not random samples from a common population.
Can we control for confounding variables?

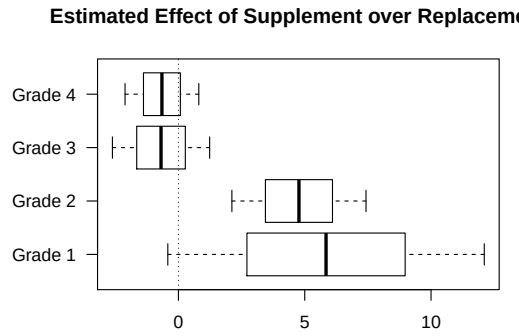
More E.Co.

After assigning treatment at random, teachers chose to use E.Co. as supplement to or replacement of reading program.
Initially, pretend this was randomly assigned as well. 2-factor (2^2) design.

```
est1 <- se1 <- rep(NA,4)
for (k in 1:4){
 lm.supp <- lm (post.test ~ supp + pre.test, subset= grade==k & !is.na
 est1[k] <- lm.supp$coef[2]; se1[k] <- se.coef(lm.supp)[2]
}
df <- data.frame(grade=factor(paste("Grade",rep(1:4,each=5))), effect
```

```
boxplot(effect~grade, df, horizontal=TRUE,las=TRUE,main="Estimated Eff
abline(v=0.0,lty=3)
```


E.Co Boxplot



Ignorable Treatment Assignment

Analysis above works for observational studies if:
Dist'n of units across treatment conditions is random in x w.r.t. y.

$$y^0,\ y^1 \perp\!\!\!\perp T|X$$

Dist'n of $(y^0,\ y^1)$ is the same across levels of T after conditioning on X.
Violations:

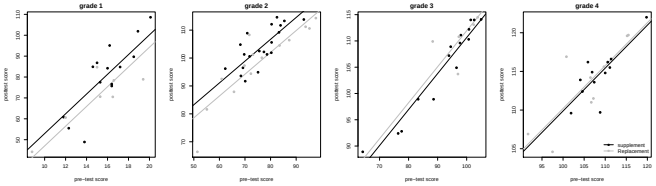
- Teachers assign Supp vs Repl based on class attributes which affect y scores.
- Motivated teachers favor Supp and also increase y scores.

A fix for the latter: measure teacher's motivation and include it as another predictor.

Reasonableness of Regression

Can never "prove" ignorability holds, as it depends on having measured all the right lurking variables.
If it does hold, SLR may not be adequate to describe it due to:

- Lack of complete overlap
  Those on one end of the covariate range fall in the same treatment group.



- Lack of balance

§9.6 Causal Inference in Observational Studies

Setting 1: Intervention occurred, but was not randomly allocated.
(above) Need to assume ignorability
Setting 2: No intervention. Observe "treatment", response and any covariates.
Define treatment.
Height affects earnings. Can we allocate heights?
Study children of "single-moms" versus "mom-and-dad" families. Could we change mom's marital status by changing tax laws, divorce laws, child support rules, marriage encouragement counseling? Pre- or post-birth? Many hypothetical interventions.
How would we design such an experiment?
Adoptions of Korean kids into Am. families groups together similar kids, not similar parents.

Multiple Treatment Factors

If A and B are both treatments, we need to think of changing A while B stays the same and vice versa. Not possible if one occurs prior in time.

Example of a poor design:
Ask why people live in poverty. Collect data on many measures and look for those with strongest associations. "Searching for causes of an effect" is ill defined, will be avoided.

Thought experiment:
Ask, "What kind of experiment ... could be performed?" or " ... could give rise to these data?"
Does breastfeeding increase intelligence of kids?
Diet effects on Weight Loss. What is control?
Does size of police force affect crime rate or vice versa?

§9.7 Control for post-treatment variables? NO!

unit I	trt T_i	observed intermediate outcome, z_i	potential Intermediate Outcome z_i^0	z_i^1	final outcome y_i
1	0	0.5	**0.5**	*0.7*	y_1
2	1	0.5	*0.3*	**0.5**	y_2

Units 1 and 2 have the same intermediate observed $z = .5$, but for one it's low due to being control, the other it's high from treatment.
Not comparable, even though z_i's are equal.

$$z = 0.3 + 0.2T + \gamma x + \text{error}$$

$$y = \theta T + \beta x + \epsilon = \theta^* T + \beta^* x + \delta^* z + \epsilon^*$$

Stars indicate changed parameters. θ^* is not the true effect of T.

§9.8 Causal Paths

Mediating (post-treatment) variables may have a causal path to outcomes.

Child Care Example
Suppose a 3-year intervention with parents shows a 10 pt gain in kids IQs over control. Is it because of improved parenting?

Parenting Potential	Parenting Quality after assigned to control	treated	Child's IQ after assigned to control	treated	Proportion of sample
Poor either way	Poor	Poor	60	70	0.1
Good if treated	Poor	Good	65	80	0.7
Good either way	Good	Good	90	100	0.2

Note top and bottom lines show an improvement of 10. Middle line gets 15, so intervention is $5/15 = 1/3$ of the effect.

Child Care 2

If we include both parenting and treatment in the model we get a negative treatment effect. Why? Treatment effect contrasts trt vs control when parenting is fixed.
Generate data by taking 20 kids (10 trt, 10 control, 2 line 1, 14 line 2, 4 line 3)

	poor	good	poor	good
control	64.38	90.00	8	2
treated	70.00	84.44	1	9

Child Care 3

```
iqs <- rep(c(60,65,90,70,80,100),c(1,7,2,1,7,2))
trt <- factor(rep(0:1,each=10),labels=c("control","treated"))
parenting <- factor(rep(c(0,0,1,0,1,1),c(1,7,2,1,7,2)),labels=c("poor
coef( lm(iqs ~ trt + parenting))

##   (Intercept)    trttreated parentinggood
##         65.18         -1.62         21.60
```

Treatment does NOT have a negative effect on IQ.
Using the mediating variable "parenting" creates a problem.
Randomization allows us to find causal effects of the treatment, not of mediators (unless you want to add a boat load of assumptions)
Cannot answer: "What proportion of the treatment effect works through variable Z?"

Stratification

Ideally we would divide the study into the 3 levels in the table above and study each in turn. (Line 2 is of the most interest and gives strongest results). BUT these strata are unknown.

Can we ever find substitutes – principal strata?
Need asumptions akin to those for ignorablity.

More in Chapter 10.

Intermediate Outcomes in Observational Studies

Still a problem, now doubles the ignorability problem.

Wells again
Pr(switch) =
$\text{logit}^{-1}(-0.21 - .90 \cdot \text{dist100} + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4})$
Can we interpret it causally for distance, arsenic level and education?
Only if we are changing one while holding the other 2 fixed. Can we do that?

Think about digging a new well at distance x from a house with an arsenic problem. What would make them switch to it? Can we model cost and benefit?