Simulation and Regression

Drew Dimmery drewd@nyu.edu

February 7, 2014

Today's Plan

- Homework
- Regression Adjustment
- Simulation
- Regression in R

Homework

- [C]onsider a stratified estimator that controls for Z_i by
  - (i) partitioning the sample by values of Z_i, then
  - (ii) taking the difference in treated and control means within each of these strata, and then
  - (iii) combining these stratum-specific estimates with a weighted average, where we weight each stratum contribution by the share of the P in each stratum

Notation and Setup

- So we consider the following two expectations:
  - $E[Y_i(1) - Y_i(0)|Z_i = 1]$ weighted by $p_Z = P(Z_i = 1)$
  - $E[Y_i(1) - Y_i(0)|Z_i = 0]$ weighted by $1 - p_Z$
- Then we want the weighted sum to be $E[Y_i(1) - Y_i(0)]$
- Homework: Is this possible?

The kink

- We **do not** observe principal Strata (counterfactual treatments)
- But we still need to think about them.
- If you talked about them on the homework, you were probably on the right track.

Decompose to Principal Strata

- Within the stratum $Z = 1$, we have the following:
  - $p_{comp} = P(D_i(1) - D_i(0) = 1)$
  - $p_{NT} = P(D_i(1) = D_i(0) = 0)$
  - $p_{AT} = P(D_i(1) = D_i(0) = 1)$
  - $p_{def} = P(D_i(1) - D_i(0) = -1) = 0$
- And these probabilities are equal (in expectation) across strata defined by Z due to random assignment

Principal Strata TEs

- Each principal strata may have its own conditional average treatment effect
  - $\rho_{comp} = E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1]$
  - $\rho_{NT} = E[Y_i(1) - Y_i(0)|D_i(1) = D_i(0) = 0]$
  - $\rho_{AT} = E[Y_i(1) - Y_i(0)|D_i(1) = D_i(0) = 1]$
  - $\rho_{def} = E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = -1]$
- We don't assume anything about these effects.
- Also note that these are equal across strata in Z due to random assignment of Z.

Counterfactuals and Principal Strata

- But those effects assume counterfactual conditions in treatment that we don't observe.
- For instance, for never takers:
  - $E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) = D_i(0) = 0]$
  - This observed quantity may be simplified: $E[Y_i(0) - Y_i(0)|D_i(1) = D_i(0) = 0]$
  - Which is equal to zero.

  - The same is true for always takers.

- This isn't to say that Always-Takers wouldn't be affected by treatment: just that we never see them affected by treatment.

Complier TEs

- This is not the case for compliers, though.
  - $E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) - D_i(0) = 1]$
  - This can be similar simplified to:
  - $E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1]$
- And we've assumed that there are no defiers.

Intention to Treat Effect

- This part isn't necessary to fully grok, yet.
- This shows what we can get with a simple difference in means:
  - $ITT = E[Y_i(D_i(1))] - E[Y_i(D_i(0))]$
  - $ITT = E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) = D_i(0) = 0] \times p_{NT} +$ $E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) = D_i(0) = 1] \times p_{AT} +$ $E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) - D_i(0) = 1] \times p_{comp} +$ $E[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1) - D_i(0) = -1] \times p_{def}$
- Taking into account some things we know (observed effects of AT & NT is zero, $p_{def} = 0$):
  - $ITT = E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1] \times p_{comp}$
  - We're close, now!
  - We just need to think about p_{comp}.

Intention to Treat Effect (on D)

- Given what we know, we can look at the following:
  - $ITT_D = E[D_i(1) - D_i(0)] =$ $E[D_i(1) - D_i(0)|D_i(1) = D_i(0) = 0]p_{NT} +$ $E[D_i(1) - D_i(0)|D_i(1) = D_i(0) = 1]p_{AT} +$ $E[D_i(1) - D_i(0)|D_i(1) - D_i(0) = 1]p_{comp} +$ $E[D_i(1) - D_i(0)|D_i(1) - D_i(0) = -1]p_{def}$
  - Or simply:
  - $ITT_D = 1 \times p_{comp}$
  - Which is what we want.

And finally

- The observed difference in treatment across Z gives us p_{comp}.
- So we can simply take $\frac{ITT}{ITT_D} = E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1]$
- This is a LATE (Local Average Treatment Effect) or CACE (Complier Average Causal Effect) depending on who is talking about it.
- It's the best we can do in the case of non-compliance like this. (More on this stuff later in the semester)

Back to the homework!

- But the homework had even more significant issues, as we were looking WITHIN strata.
- This essentially gets rid of the benefits of randomization.
- To get a good estimate for the population using this method, we have to get a good estimate WITHIN each strata, too.
- In other words, we must be able to recover $E[Y_i(1) - Y_i(0)|Z = 1]$ and vice versa within each strata. This would allow:
  - $E[Y_i(1) - Y_i(0)|Z = 0](1 - p_Z) + E[Y_i(1) - Y_i(0)|Z = 1]p_Z$
- But can we get that?
- No. Not even a little bit.

What's in a strata?

- For $Z = 0$, and our three principal strata, we have:
  - Always Takers will be $D_i = 1$
  - Never takers will be $D_i = 0$
  - Compliers will be $D_i = 0$
- So we can decompose the difference in means is as follows:
  - $E[Y_i(1)|Z = 0] - E[Y_i(0)|Z = 0] = E[Y_i(1)|D_i(1) = D_i(0) = 1] - E[Y_i(0)|D_i(1) - D_i(0) = 1]\frac{p_{comp}}{p_{NT}+p_{comp}} - E[Y_i(0)|D_i(1) = D_i(0) = 0]\frac{p_{NT}}{p_{NT}+p_{comp}}$
- The key point is that these counterfactuals are not the ones we want.
- Even if they were, we still wouldn't know what we were estimating without knowing proportions in each strata (which we wouldn't).
- For this to equal $E[Y_i(1) - Y_i(0)|Z = 0]$, we would need to make some strong assumptions directly on the potential outcomes.

What sort of assumptions work?

- If complier and never taker proportions are equal, then we get:
    - $E[Y_i(1)|D_i(1) = D_i(0) = 1] -$ $\frac{1}{2}E[Y_i(0)|D_i(1) - D_i(0) = 1] + \frac{1}{2}E[Y_i(0)|D_i(1) = D_i(0) = 0]$
    - This isn't enough.
- The assumption we'd need would be on the equality of potential outcomes across all principal strata (ludicrously strong):
    - $E[Y_i(1)|D_i(1) = D_i(0) = 1] = E[Y_i(1)|D_i(1) = D_i(0) = 0] = E[Y_i(1)|D_i(1) - D_i(0) = 1] =$ $E[Y_i(1)]$
    - And $E[Y_i(0)|D_i(1) = D_i(0) = 1] = E[Y_i(0)|D_i(1) = D_i(0) = 0] = E[Y_i(0)|D_i(1) - D_i(0) = 1] =$ $E[Y_i(0)]$
    - (Since we randomized over Z_i, it doesn't help us to only assume this only in strata of Z_i)
    - This WOULD allow us to get at the common causal effect. (But at what cost?)
    - For all practical purposes, this estimation strategy is DOUBLY not identified.

Graphically

. . .

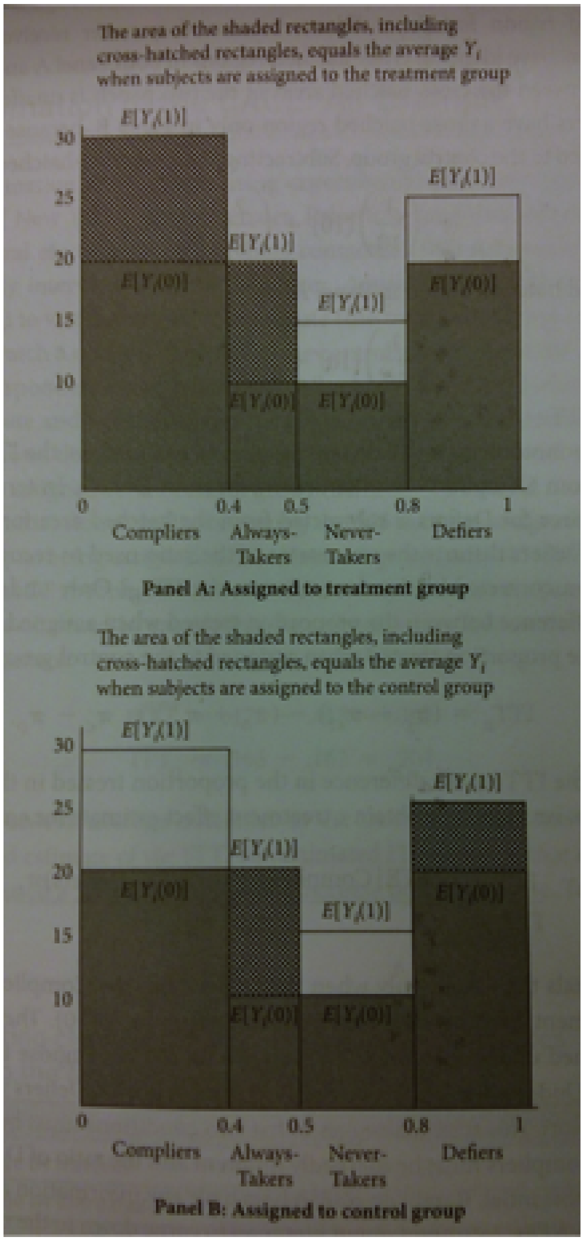
This is a ridiculous amount of regularity to assume, though.

Example

- What if we had **all** the data? (- Don Rubin)
- Assume a balanced design $p_Z = \frac{1}{2}$ and constant TE $\rho = 20$, with expected potential outcomes as follows:

. . .

	Y_1	Y_2	size
Always-taker	10	30	20%
Never-taker	0	20	30%
Complier	65	85	50%



The area of the shaded rectangles, including cross-hatched rectangles, equals the average Y_i when subjects are assigned to the treatment group

Panel A: Assigned to treatment group

The area of the shaded rectangles, including cross-hatched rectangles, equals the average Y_i when subjects are assigned to the control group

Panel B: Assigned to control group

- The given estimator will target the following parameter (using the decomposition from before):

. . .

	$Z = 1$	$Z = 0$
Y_1	$85 \cdot \frac{.2}{.7} + 30 \cdot \frac{.2}{.7}$	
	`69.286`	`30`
Y_0		$65 \cdot \frac{.5}{.8}$
	`0`	`40.625`
ρ	**69.286**	**-10.625**

- This setup gives an estimand of 29.33. This is *not* the ATE (20).
- That is, we've derived the population estimand under this stratified estimator.
- And we already assumed a lot of common issues away (balanced design, constant effects)
- If we knew population sizes within principal strata, would this help?

Now for something completely different

- Now we're going to switch gears to regression and covariate adjustment.
- We'll also be getting some examples of how to work with linear models (in R)
- This will include some consideration of partial regression.
- All of which should make your homework easier to work with.

Covariate Adjustment in sampling

- Lin, Winston. (2013) "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique" *Annals of Applied Statistics.* 7(1):295-318.
- Imagine that we are biologists. We are interested in leaf size.
- Finding the size of leaves is hard, but weighing leaves is easy.
- Key insight is that we can use auxilliary information to be smarter:
  - Sample from leaves on a tree.
  - Measure their size and weight.

- Let $\hat{y}_s$ be the average size in the sample.
- Let $\hat{x}_s$ be the average weight in the sample.
- Population averages drop the subscript.
- We know that $\hat{y}_s$ unbiased and consistent for $\hat{y}$
- But we have extra information!
- We also have $\hat{x}$ (all the weights)
- This motivates the regression estimator: $\hat{\bar{y}}_{reg} = \hat{y}_s + \beta(\hat{x} - \hat{x}_s)$
- We get β by a regression of leaf area on weight in the sample.

Connection to Multiple Regression

- In the case of OLS for the analysis of experiments, we have nearly the same setup.
- Only difference is that we are sampling for both treatment and control.
- This means that we must adjust both groups separately.
- This motivates the use of covariate-treatment interactions.
- Remember! Freedman (2008) showed that regression is **biased** and can be **inconsistent** for an experimental parameter in the case when interactions aren't included.
- This is something to consider in many different situations. There's no reason to expect treatment and control groups to exhibit identical effects (even ones that are orthogonal to the causal parameter of interest)
- This is just a particular sort of omitted variable bias, which you should already be familiar with.

Covariate Adjustment in Experiments

- Now imagine we are social scientists (hopefully this isn't hard)
- We are interested in the effects of a binary treatment on education, measured by a test.
- Let's set up a simulation.
- 250 students. Ten classes of 25 students each. Observed over two years.
- First year has half good teachers and half bad.
- We want to estimate the effect of the intervention in year 2.
- Treatment is assigned randomly by **individual**
- Note: This setup usually demands an accounting of clustering, which I'm ignoring. Maybe I'll bring it back later in the semester when we discuss SUTVA.

Simulation

```r
# Variables which govern the size of the simulation (and our causal effects)
nclass <- 5
nstudent <- 25
Eff <- 5
EffSD <- 3
# Simulate data
set.seed(1977)
Yr1ClassType <- rep(c(1, 0), nclass * nstudent)
Yr2ClassType <- sample(Yr1Class, replace = FALSE)
Yr1Score <- rnorm(2 * nclass * nstudent, 76 + Yr1ClassType * 5, 9)
# Fixed margins randomization
Trt <- sample(Yr1ClassType, replace = FALSE)
# There is an independent effect of class type in each year Variance is
# different across class types in year 2
CtlOutcome <- rnorm(2 * nclass * nstudent, Yr1Score + Yr2ClassType * 3, 9 -
    Yr2ClassType * 4)
# Treatment effect is random, but with expectation Eff
Yr2Obs <- CtlOutcome + Trt * rnorm(2 * nclass * nstudent, Eff, EffSD)

summary(lm(Yr2Obs ~ Trt))$coefficients[2, ]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##   4.307282   1.558184   2.764296   0.006133
```

```r
summary(lm(Yr2Obs ~ Trt + Yr1Score))$coefficients[2, ]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  3.5206647  1.0064194  3.4982082  0.0005554
```

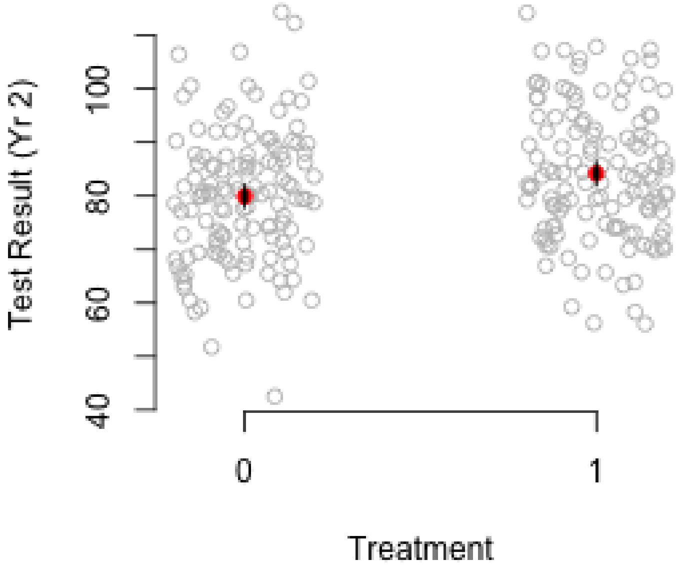
```r
# *IF* we trust the model-based SEs, then we could pull SEs for a TE
# estimate with:
summary(lm(Yr2Obs ~ Trt))$coefficients["Trt", 2]
```

```
## [1] 1.558
```

Plot Data

```r
plot(jitter(Trt), Yr2Obs, axes = F, xlab = "Treatment", ylab = "Test Result (Yr 2)",
    col = "grey")
```

```
axis(2)
axis(1, at = c(0, 1))
# Calculate quantities for plotting CIs
mns <- tapply(Yr2Obs, Trt, mean)
# SEs could also be pulled from the linear models we fit above with:
ses <- tapply(Yr2Obs, Trt, function(x) sd(x)/sqrt(length(x)))
points(c(0, 1), mns, col = "red", pch = 19)
# Note the loop so that I only write this code once
for (tr in unique(Trt)) {
    for (q in c(0.25, 0.025)) {
        upr <- mns[as.character(tr)] + qnorm(1 - q) * ses[as.character(tr)]
        lwr <- mns[as.character(tr)] - qnorm(1 - q) * ses[as.character(tr)]
        segments(tr, upr, tr, lwr, lwd = (-4/log(q)))
    }
}
```



Partial Regression

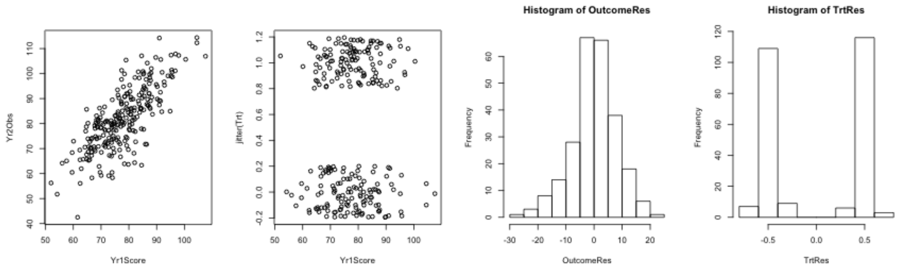
- Can we make that plot a little more friendly?
- Let's residualize our outcome based on scores in the first period. This should remove a substantial amount of the variance in the outcome.
- A few things to check, though.

. . .

```
OutcomeRes <- residuals(lm(Yr2Obs ~ Yr1Score + 0))
TrtRes <- residuals(lm(Trt ~ Yr1Score + 0))
c(sd(OutcomeRes), sd(Yr2Obs))
```

```
## [1]  8.131 12.482
```

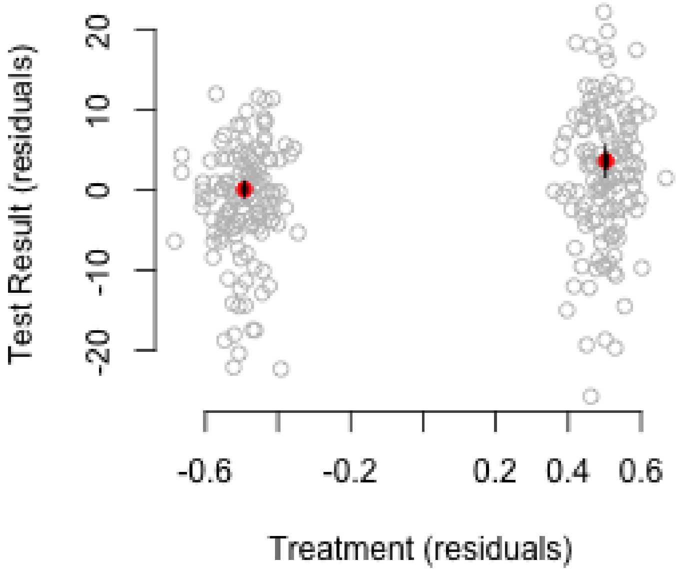
```
# Diagnostics
par(mfrow = c(1, 4))
plot(Yr1Score, Yr2Obs)
plot(Yr1Score, jitter(Trt))
hist(OutcomeRes)
hist(TrtRes)
```



Residualized Plot

```
par(mfrow = c(1, 1))
plot(jitter(TrtRes), OutcomeRes, axes = F, xlab = "Treatment (residuals)", ylab = "Test Resu
    col = "grey")
axis(2)
axis(1)
# Pull information from the new bivariate model
mns <- coef(lm(OutcomeRes ~ TrtRes))
```

```
ses <- summary(lm(OutcomeRes ~ TrtRes))$coefficients[, 2]
TrtResMns <- tapply(TrtRes, Trt, mean)
names(ses) <- names(mns) <- names(TrtResMns)
points(TrtResMns, mns, col = "red", pch = 19)
for (tr in names(TrtResMns)) {
    for (q in c(0.25, 0.025)) {
        upr <- mns[tr] + qnorm(1 - q) * ses[tr]
        lwr <- mns[tr] - qnorm(1 - q) * ses[tr]
        segments(TrtResMns[tr], upr, TrtResMns[tr], lwr, lwd = (-4/log(q)))
    }
}
```



Partial Regression for FEs

- We'll get to this later in the semester.
- The point is, partial regression is a fundamentally important tool that let's us do things that would otherwise be very hard.

. . .

```
sweeplm <- function(formula, dat, ind) {
    newd <- model.matrix(~., model.frame(formula, dat, na.action = na.pass))
    newd <- newd[, -1]
    ok <- complete.cases(newd)
    newd <- newd[ok, ]
    ind <- ind[ok]
    newd <- apply(newd, 2, function(x) unlist(tapply(x, ind, function(z) z -
        mean(z, na.rm = TRUE))))
    list(lm(newd[, 1] ~ newd[, -1] - 1, as.data.frame(newd)), newd, as.character(ind))
}
```

Coefficient Plots

- We only really care about one causal parameter at a time, usually.
- When we're in the causal inference mindset, we very rarely want to see long lists of covariates that are causally meaningless.
- In general, it is often worthwhile to make simple plots of your coefficients.
- I'll put some example code on the next page.
- It assumes a couple vectors exist:

. . .

```
ests <- c(coef(lm(Yr2Obs ~ Trt))[2], coef(lm(Yr2Obs ~ Trt + Yr1Score))[2])
ses <- c(summary(lm(Yr2Obs ~ Trt))$coefficients[2, 2], summary(lm(Yr2Obs ~ Trt +
    Yr1Score))$coefficients[2, 2])
var.names <- c("Unadjusted", "Adjusted")
```

Coefficient Plot Code

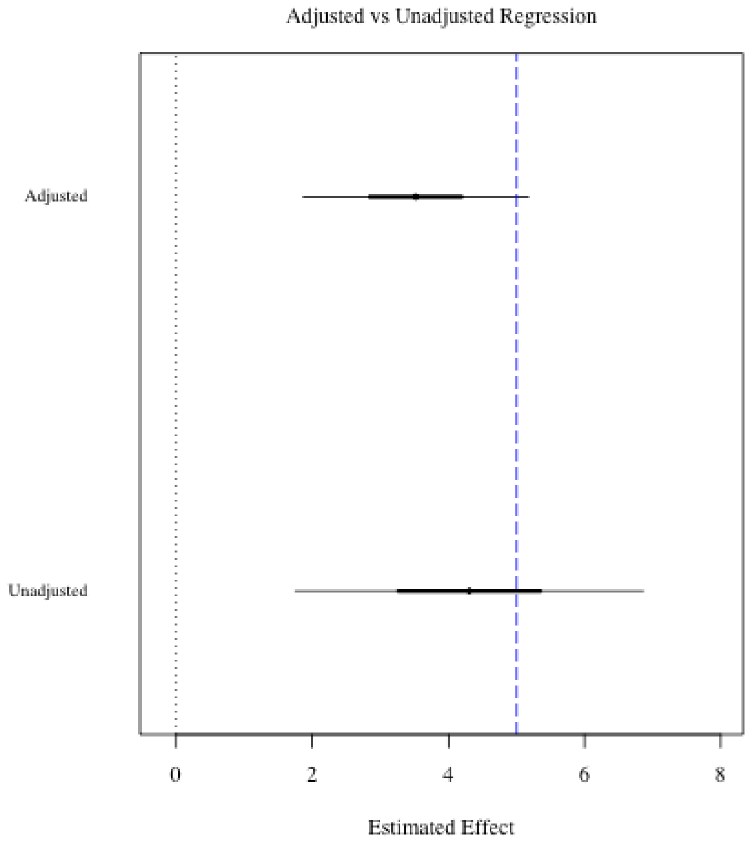
```
par(family = "serif", oma = c(0, 0, 0, 0), mar = c(5, 10, 4, 2))

plot(NULL, xlim = c(-0.2, 8), ylim = c(0.7, length(ests) + 0.3), axes = F, xlab = NA,
    ylab = NA)

for (i in 1:length(ests)) {
    points(ests[i], i, pch = 19, cex = 0.5)
    lines(c(ests[i] + 1.64 * ses[i], ests[i] - 1.64 * ses[i]), c(i, i))
    lines(c(ests[i] + 0.67 * ses[i], ests[i] - 0.67 * ses[i]), c(i, i), lwd = 3)
    text(-1.1, i, var.names[i], xpd = T, cex = 0.8, pos = 2)
}

axis(side = 1)
abline(v = 0, lty = 3, col = "black")
mtext(side = 1, "Estimated Effect", line = 3)
mtext(side = 3, "Adjusted vs Unadjusted Regression", line = 1)
box()
```

Plotted



- Think about how these two might differ for different starting parameters (ex. sample size)