ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection on Observables Designs: Part I, Regression Adjustment

This set of lecture notes begins our discussion of what we refer to as "selection on observables" designs. The key assumption underlying these designs is that the treatment assignment is "ignorable" — which you can interpret as "as good as randomly assigned" — after you condition on a set of observable factors. There are a variety of estimation techniques available in this scenario: standard linear regression, flexible nonparametric regression, matching estimators, and propensity score estimators. The underlying (untestable) assumption of all of these estimators, however, is that you observe all of the factors that affect treatment assignment *and* are correlated with the potential outcomes. In other words, to the extent that there is systematic selection into treatment, this selection is only a function of the observable variables. Hence, if you can "control" for the effects of these variables on the probability of selection, then you can produce consistent estimates of causal effects. The flip side is that, if you don't observe all the determinants of selection, then these methods do not, in general, produce estimates with a causal interpretation. This important fact is often overlooked by applied practitioners who focus on the sophistication of the estimation technique (matching is somewhat en vogue these days). In my opinion, the underlying selection on observables assumption is too strong to hold in most cases, so these methods are probably applied more often than they should be. Nevertheless, in some cases the assumption is palatable (or at least defensible), and in those cases these techniques can be quite helpful.[1]

[1]For example, consider a case in which individuals apply for some program or job, and then are assigned to different areas/departments/treatments/whatever based upon the data in their applications. In this scenario, the researcher can observe all of the non-random factors that affected selection (i.e. the data in the applications), and the selection on observables assumption clearly makes sense.

1 Regression Adjustment

The key underlying assumption motivating regression adjustment (and the other selection on observables designs) is that the treatment is independent of the potential outcomes (particularly the untreated potential outcomes) after conditioning on the observable covariates, X_i. We write this assumption as:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

This assumption is referred to as the "unconfoundedness assumption," the "selection on observables assumption," or the "conditional independence assumption." When combined with an assumption about overlap, $0 < P(D_i = 1|X_i) < 1$, it is referred to as "strongly ignorable treatment assignment."

How does this fit in with what we previously learned about regression? We can translate the potential outcomes framework into a classical linear regression model by defining $Y_i(0)$ and $Y_i(1)$ as follows and assuming constant treatment effects.

$$Y_i(0) = \alpha + \varepsilon_i$$

$$Y_i(1) = Y_i(0) + \beta + \beta_i$$

Note that the ε_i here will now play the role of a structural residual, not (necessarily) a CEF residual. The constant treatment effects assumption amounts to $\beta_i = 0$. Under these definitions, we have

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

If D_i is randomly assigned, then $D_i \perp Y_i(0)$, so $D_i \perp \varepsilon_i$ (our typical regression orthogonality assumption). What happens when we relax the assumption to $Y_i(0) \perp D_i|X_i$? Consider rewriting our expression for Y_i as

$$Y_i = \alpha + \beta D_i + \delta h(X_i) + \eta_i$$

where $h(X_i) \equiv E[D_i|X_i]$ and $\eta_i = \varepsilon_i - \delta E[D_i|X_i]$. Note that $h(X_i)$ will not be estimated by running a regression of Y_i on D_i and X_i even if $E[D_i|X_i]$ is linear because the regression of Y_i on D_i and X_i estimates (or approximates) $E[Y_i|D_i, X_i]$. Using partitioned regression, however, we know that partialing out $h(X_i)$ from D_i and then running a bivariate regression of Y_i on the partialed-out D_i generates the same estimate of β as the multiple regression of Y_i on D_i and $h(X_i)$. So rewrite our expression for Y_i once again as:

$$Y_i = \alpha + \beta \tilde{D}_i + u_i$$

where[2]

$$\tilde{D}_i = D_i - E[D_i|X_i]$$

$$u_i = \varepsilon_i + \beta E[D_i|X_i]$$

Note that this identity comes straight from the fact that $Y_i = \alpha + \beta D_i + \varepsilon_i$. The standard orthogonality condition that we need for a regression of Y_i on $\tilde{D}_i$ to generate unbiased estimates of β can be written as $E[u_i \tilde{D}_i] = 0$. Under what circumstances does this condition hold?

$$E[u_i \tilde{D}_i] = E[\varepsilon_i \tilde{D}_i] + \beta E[E[D_i|X_i]\tilde{D}_i]$$

$$= E[\varepsilon_i \tilde{D}_i] + 0 = E[\varepsilon_i (D_i - E[D_i|X_i])]$$

$$= E[\varepsilon_i D_i] - E[\varepsilon_i E[D_i|X_i]]$$

$$= E[\varepsilon_i D_i] - E[E[\varepsilon_i|X_i]E[D_i|X_i]]$$

$$= E[E[\varepsilon_i D_i|X_i]] - E[E[\varepsilon_i|X_i]E[D_i|X_i]]$$

[2]Formally, $\tilde{D} = (I - P_{H(X)})D$, where $P_{H(X)} = H(H'H)^{-1}H'$ and $H_i = E[D_i|X_i]$. The fully transformed regression model would be $\tilde{Y}_i = \alpha + \beta \tilde{D}_i + \tilde{\varepsilon}_i$, where the $\tilde{}$ operator is defined such that $\tilde{A}$ represents variable A after partialing out H. However, since the projection matrix P_H is idempotent, the regression coefficient $(\tilde{D}'\tilde{D})^{-1}(\tilde{D}'\tilde{Y})$ is unaffected by whether or not we partial out H from Y.

$$= E[E[\varepsilon_i|X_i]E[D_i|X_i]] - E[E[\varepsilon_i|X_i]E[D_i|X_i]]$$

$$= 0$$

So *if* we have the unconfoundness assumption and *if* we know the CEF $h(X_i) = E[D_i|X_i]$ and *if* we have a homogeneous treatment effect, then we can get consistent estimates of β from a regression of Y_i on D_i and $h(X_i)$. Of course, we rarely know $h(X_i)$, but if the CEF is linear then we know that simply including X_i as a set of additional control variables will be sufficient to estimate the CEF (see Regression-CEF Theorem from previous notes).[3] Even if the CEF is *not* linear, including X_i as a set of regressors provides the MMSE linear approximation to the CEF (see Regression Approximation Theorem from previous notes). The accuracy of this approximation will often depend on how much extrapolation we are asking of the linear approximation.

For example, suppose that we assume a linear model of the form:

$$Y_i = \alpha + \beta D_i + \delta X_i + u_i$$

We can transform this model to look like our standard model that does not contain covariates:

$$Y_i - \delta X_i = Y_i^* = \alpha + \beta D_i + u_i$$

In this model we estimate the treatment effect as the difference in means between the treated and control groups. Replacing actual values of the coefficients with their estimates we have:

$$\hat{\tau} = \overline{Y}_T^* - \overline{Y}_C^* = (\overline{Y}_T - \hat{\delta}\overline{X}_T) - (\overline{Y}_C - \hat{\delta}\overline{X}_C)$$

[3]Note that the joint-normality case that gives you a linear CEF is guaranteed not to hold in this case because D_i is clearly not normal. Basically, the CEF $E[D|X]$ is not going to be linear unless it is saturated in X.

$$= (\overline{Y}_T - \overline{Y}_C) - \hat{\delta}(\overline{X}_T - \overline{X}_C)$$

From the last line, you can see that if $\overline{X}_T \approx \overline{X}_C$, then the precise specification of our linear approximation will generally not be that important. So loosely speaking, what matters is the overlap of the distributions of X_i in the treatment versus control groups. If $\overline{X}_T$ is far from $\overline{X}_C$, i.e. there is not much overlap in the distributions of X for the treated and control samples, then we are performing an extrapolation that will depend heavily on whether we get the functional form right. Of course, as X becomes multidimensional, it becomes harder to define what constitutes $\overline{X}_T \approx \overline{X}_C$. The overlap issue is one that we will revisit in nonparametric regression and propensity score matching. For the time being, however, note that the most important assumption is the unconfoundedness assumption $(Y_i(0) \perp D_i|X_i)$ — without this assumption we have nothing. In contrast, without the linear CEF assumption, we can still fall back upon the Regression Approximation Theorem.

2 How to Choose Controls

Researchers often work with data sets containing many potential covariates. In these scenarios, a frequent question that arises is, "How should I determine which covariates to include in my regression?" Include too many covariates in X and you consume degrees of freedom and run the risk of overfitting; include too few covariates in X and you may omit an important control variable. Simple rules such as "only include covariates with a t-statistic greater than 2 in absolute value" are ad hoc and have no sound theoretical basis. Recent "big data" algorithms, however, provide a data-driven way to select controls. Belloni, Chernozhukov, and Hansen (2014) summarize one method based on the Least Absolute Shrinkage and Selection Operator (LASSO). To understand how this method works we must first understand the LASSO.

The LASSO dates back to work by Frank and Friedman (1993) and Tibshirani (1996). Like linear regression, the LASSO postulates a linear model relating X and Y and attempts to find the coefficients that minimize the sum of squared residuals. However, unlike linear

regression, the LASSO applies a "penalty" for every nonzero coefficient that it fits. This strategy of penalizing model complexity, or overfitting, is generically known as "regularization." The basic intuition relates to one of the first properties you likely learned about regression: adding an additional regressor to a regression can only make the R^2 go up, not down. This fact represents an overfitting problem — when given additional parameters (i.e., coefficients), the model can always produce a better fit, regardless of whether those parameters are truly nonzero. The LASSO addresses this problem by penalizing the regression for every nonzero coefficient that it fits — with extra penalty for larger coefficients — thus incentivizing the regression to only include additional variables as regressors if the additional variables contribute meaningful predictive power. Because the penalty increases with coefficient size, the LASSO also tends to "shrink" coefficients towards zero (relative to OLS). Formally, in a case with p potential regressors, the LASSO solves the problem:

$$\hat{\beta} = argmin_b \ \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}b_j)^2 + \lambda \sum_{j=1}^{p} |b_j|\gamma_j$$

The first summation term in this objective function is the standard sum of squared residuals that OLS minimizes. The second summation term in the objective function is the "penalty" term that increases in the size of the fitted coefficients. It consists of three components: the absolute values of the coefficients, $|b_j|$; the "penalty level," λ; and the individual coefficient "penalty loadings," γ_j. The coefficient absolute values are the objects being penalized, and the penalty loadings exist to ensure that scale differences between the x's don't result in some variables having excessive influence in the penalty function (e.g., the simplest penalty loadings would be to rescale each x_j by its standard deviation). The penalty level, λ, determines the degree to which the LASSO penalizes nonzero coefficients. In contexts involving prediction λ is often set via cross-validation; i.e., the researcher estimates the LASSO for different values of λ using a subset of the data, and then chooses the value of λ that generates the best predictions of y_i in the data not used for estimation.[4] For our discussion, however, the exact choice of λ is not important.

[4]More precisely, the researcher chooses the value of λ that corresponds to the LASSO coefficients that generate the best predictions of y_i in the data not used in estimation.

One characteristic of the LASSO penalty function is that it has a "kink" at zero. This occurs because of the absolute value operator; the derivative of the penalty function with respect to b_j equals $-\lambda\gamma_j$ when b_j is just below zero (which encourages b_j to increase towards zero, since we have a minimization problem), and then reverses to $\lambda\gamma_j$ when b_j is just above zero (which encourages b_j to decrease towards zero). What this means in practice is that the LASSO tends to find "corner solutions" and set many coefficients equal to zero. This suggests that it could be useful in the variable selection problem — we might run the LASSO first to determine which covariates to include, and then estimate OLS using the subset of covariates that the LASSO assigns nonzero coefficients.[5]

A naïve application of the LASSO to the variable selection problem would be to apply the LASSO to the equation:

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{j=2}^{p} \beta_j X_{ij} + u_i$$

One issue here is that the LASSO might exclude D_i from the equation if it does not have sufficient predictive power, but this could be addressed by excluding D_i from the penalty term in the LASSO objective function. A bigger issue is that the LASSO might exclude some X_{ij} that are only moderately correlated with Y_i (and thus do not contribute much towards predicting Y_i) but are strongly correlated with D_i. The problem with excluding these covariates is that the degree of omitted variables bias from excluding a covariate X_{ij} depends on both X_{ij}'s correlation with D_i and X_{ij}'s correlation with u_i. (We show this in Section 4, but it should be intuitively clear that omiting a variable will cause bias only if the variable is correlated with both the treatment and the unobserved determinants of Y_i.) Excluding a covariate that is highly correlated with D_i can thus cause considerable bias.

As an alternative, Belloni, Chernozhukov, and Hansen suggest the following procedure:

[5]Naturally you might wonder why we don't simply run the LASSO and be done with it, rather than running OLS using the variables selected by the LASSO. One answer is that the LASSO is a "shrinkage" estimator (it's right there in the name!), which means that it "shrinks" the coefficient estimates towards zero. This can be desirable for the purposes of predicting y_i, but it does technically introduce bias into the coefficient estimates. That said, we're not that concerned about the actual coefficients on the x's if we're just going to use them as control variables, but technically we're still a little concerned because our ultimate goal is to eliminate bias in our estimate of the treatment effect, rather than to predict y_i.

1. Apply the LASSO to the equation $D_i = \pi_0 + \sum_{i=1}^{p} \pi_j X_{ij} + v_i$
2. Apply the LASSO to the equation $Y_i = \gamma_0 + \sum_{i=1}^{p} \gamma_j X_{ij} + \eta_i$
3. Regress Y_i on D_i and the set of covariates that the LASSO selects (i.e., variables with nonzero LASSO coefficients) in either 1 or 2. In other words, include X_{ij} as a covariate if the LASSO assigns it a nonzero coefficient in either Step 1 or 2. Note that this last regression is an OLS regression.

The intuition is simple here: variables that are useful to control for are those that are either correlated with D_i or correlated with Y_i (or, most importantly, correlated with both). The first step selects covariates correlated with D_i, and the second step selects covariates correlated with Y_i. The dissection of the two steps also illustrates a deeper point: the first step is entirely focused on eliminating bias in your treatment effect estimate; and the second step is entirely focused on increasing the precision of your treatment effect estimate. To see this, note that in a randomized controlled trial, the first step would generally select zero covariates, because none of the covariates should be correlated with D_i. In other words, the bias-reduction step is not necessary, because there is no bias to reduce. The second step, however, might select some covariates if these covariates correlate with the outcome. For example, if the outcome were a post-intervention test score, then an individual's baseline test score would likely get selected in the second step because it has a lot of predictive power for Y_i, even though it's uncorrelated with the randomly assigned treatment. Including these covariates in the regression isn't necessary for addressing bias (if D_i is randomly assigned), but it can generate more precise estimates of the treatment effect by reducing the final regression's mean squared error (i.e., the average value of $\hat{\varepsilon}_i^2$).

3 Regression Adjustment Application: Kreuger (1993) and DiNardo and Pischke (1997)

Krueger (1993) is an example of a carefully executed regression adjustment application that nevertheless fails to identify a causal effect (ideally I would have an example of a carefully executed regression adjustment application that does identify a causal effect, but those are regrettably rare and, regardless, it's generally impossible to know that a paper got the "right" answer).[6] Krueger uses CPS data to examine the wage premium for using computers at work. His primary specification is:

$$ln(W_i) = X_i\beta + \alpha C_i + \varepsilon_i$$

C_i is the treatment of interest, a dummy variable that is unity if an employee uses a computer at work and zero otherwise. W_i corresponds to the employee's hourly wage, and X_i contains other variables that might affect both wages and computer usage. When X_i contains no covariates, Krueger estimates $\hat{\alpha} = 0.33$ (using 1989 CPS data). When X_i contains a rich set of covariates — education, experience, race, gender, marital status, etc. — Krueger estimates $\hat{\alpha} = 0.19$. When X_i contains a rich set of covariates plus eight occupation dummies, Krueger estimates $\hat{\alpha} = 0.16$. Including 48 two-digit industry dummies reduces the coefficient by another 20 percent.

As a descriptive exercise, the results are surely valid. Do they have a causal interpretation, i.e. does $\hat{\alpha}$ represent the return to teaching a computer illiterate worker how to use a computer? Krueger cautions that "a critical concern in interpreting the OLS regressions reported above is that workers who use computers on the job may be abler workers, and therefore may have earned higher wages even in the absence of computer technology." (pp. 42-43) He presents four empirical facts to argue in favor a causal interpretation. First, he controls for computer use at home, arguing that this should reduce selection bias. The coefficient on computer use at work is unaffected. Second, he conducts a survey demonstrating

[6] More accurately, Krueger's regressions give us no reason to believe that he identifies a causal effect, although we may still believe that he has done so based on prior knowledge.

that temporary agencies report paying higher wages for computer-literate secretaries and find it profitable to offer computer training. Third, he uses a different data set to confirm the results. Fourth, he demonstrates that occupations which adopted computers most quickly showed higher wage growth. Based on this evidence, he argues, reasonably, that the computer use coefficient may have a causal interpretation (and that increased computer usage can account for one-third to one-half of the increase in the return to education during the late 1980s.).

In 1997, John DiNardo and Steve Pischke revisited the question in a paper entitled, "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?" DiNardo and Pischke replicate Krueger's methodology using a German data set. They find a similar association between computer use and wages, but unlike Krueger they have additional data on usage of office tools such as calculators, telephones, and pencils. Many of these tools demonstrate "returns" that are almost as high as the return associated with computer usage; in the first specification they present, for example, the coefficient on computer usage is 0.11 while the coefficient on pencil usage is 0.12. To clarify their argument, DiNardo and Pischke define the causal effect of computer usage on wages to be the effect of randomly assigning computer skills to some employees but not to others.[7] Within this framework, it becomes clear that the "return" to pencils (and hence possibly to computers) must be illusory — virtually every German worker knows how to use a pencil, whereas only 60 percent of jobs involve using a pencil, so the ability to use a pencil cannot be a scarce skill that demands a high premium. It is unlikely that the pencil-using jobs are paying higher wages simply because it is expensive to hire employees that know how to use pencils.

Although DiNardo and Pischke do not prove that the computer premium is illusory, they do demonstrate that Krueger's research design is unable to distinguish between causal relationships and relationships that are due to selection. Is this because the regression approximation of $E[C|X]$ is inaccurate or is it because the selection on observables assumptions

[7]Technically, even this "treatment" is not truly a treatment since you can't assign skills, you can only assign training that you hope will create those skills. Nevertheless, DiNardo and Pischke elucidate their argument greatly by clearly specifying what the counterfactual they have in mind is.

does not hold, i.e. it is not true that $W_i(0) \perp C_i|X_i$? Almost surely it is the latter. In the following lectures we will learn methods that can deal with the former, but it is important to keep in mind that these methods are no better than regression if the selection on observables assumption does not hold.

4 Detecting Bias: Altonji, Elder, and Taber (2005)

Altonji, Elder, and Taber (2005) examine the effects of attending a Catholic school on student outcomes. They lack random variation in Catholic school attendance, so they instead regression adjust their estimates based upon the observed characteristics of the students and their families. This approach is not novel, but they make a nice contribution by applying a variant of the omitted variables bias formula to estimate the potential bias from unobserved characteristics. The critical assumption underlying this estimate of potential bias is that, loosely speaking, the relationship between the treatment and the unobserved characteristics is no stronger than the relationship between the treatment and the observed characteristics.

Consider a model of the form:

$$Y_i = \alpha + \beta D_i + X_i\delta + u_i$$

D is the treatment of interest, and X are observed determinants of Y that may be correlated with D. Define δ and u_i such that u_i is orthogonal to X_i (i.e., δ reflects both the causal effect of X on Y and the projection of the unobserved determinants of Y onto X).

Using partitioned regression we can rewrite the regression of Y_i on D_i and X_i as:

$$\tilde{Y}_i = \alpha + \beta \tilde{D}_i + u_i$$

where $\tilde{Y}_i$ and $\tilde{D}_i$ are residuals from regressions of Y_i and D_i on X_i respectively. Note that $\tilde{u}_i = u_i$ because u_i is already orthogonal to X_i by construction. The bias due to the potential

correlation between the partialed-out treatment, $\tilde{D}_i$, and the unobserved determinants, u_i, is:

$$E[\hat{\beta} - \beta] = \frac{\text{Cov}(\tilde{D}, \tilde{Y})}{\text{Var}(\tilde{D})} - \beta = \frac{\text{Cov}(\tilde{D}, \beta\tilde{D} + u)}{\text{Var}(\tilde{D})} - \beta = \beta\frac{\text{Cov}(\tilde{D}, \tilde{D})}{\text{Var}(\tilde{D})} - \beta + \frac{\text{Cov}(\tilde{D}, u)}{\text{Var}(\tilde{D})}$$

$$= \beta\frac{\text{Var}(\tilde{D})}{\text{Var}(\tilde{D})} - \beta + \frac{\text{Cov}(\tilde{D}, u)}{\text{Var}(\tilde{D})} = \frac{\text{Cov}(\tilde{D}, u)}{\text{Var}(\tilde{D})}$$

What is a reasonable estimate for $\text{Cov}(\tilde{D}, u)/\text{Var}(\tilde{D})$? Altonji, Elder, and Taber suggest assuming that:

$$\frac{\text{Cov}(D, u)}{\text{Var}(u)} = \frac{\text{Cov}(D, X\delta)}{\text{Var}(X\delta)} \tag{1}$$

In other words, assume that the relationship between D_i and u_i (the unobserved determinants of Y_i) is no stronger than the relationship between D and $X_i\delta$ (the observed determinants of Y_i). This assumption is useful because we can estimate the latter relationship but not the former. Is the assumption reasonable? It implies that a one unit change in $X_i\delta$ is associated with the same change in D as a one unit change in u_i. This would be true if, for example, the observed determinants of Y_i were randomly chosen from the set of all determinants of Y_i. If the control variables X_i were chosen specifically because they were likely to be correlated with D_i, then the assumption is particularly likely to hold (in a bounding sense). Nevertheless, it is still an assumption.

Given equation (1), it is straightforward to estimate the potential bias. Note that $\text{Cov}(D, X\delta)/\text{Var}(X\delta) = \gamma$, where γ is the regression coefficient from regressing D_i on $X_i\delta$. Thus

$$\frac{\text{Cov}(\tilde{D}, u)}{\text{Var}(\tilde{D})} = \frac{\text{Var}(u)}{\text{Var}(\tilde{D})} \cdot \frac{\text{Cov}(D, u)}{\text{Var}(u)} = \frac{\text{Var}(u)}{\text{Var}(\tilde{D})} \cdot \frac{\text{Cov}(D, X\delta)}{\text{Var}(X\delta)} = \frac{\text{Var}(u)}{\text{Var}(\tilde{D})}\gamma$$

The equality of $\text{Cov}(\tilde{D}, u)$ and $\text{Cov}(D, u)$ arises from the fact that u_i is uncorrelated with X_i by construction, so its covariance with $\tilde{D}_i$ (the residuals of regressing D_i on X_i) is

identical to its covariance with D_i. This equality suggests a simple procedure for estimating the potential bias of $\hat{\beta}$:

1. Regress Y_i on X_i. Collect the fitted values from this regression, $X_i\hat{\delta}$. Also save the mean squared error from this regression, $\hat{\sigma}_u^2$.
2. Regress D_i on $X_i\hat{\delta}$. Save the coefficient from this regression, $\hat{\gamma}$.
3. Regress D_i on X_i. Save the mean squared error from this regression, $\hat{\sigma}_d^2$.
4. Calculate the potential bias as $\text{Cov}(\tilde{D}, u)/\text{Var}(\tilde{D}) = \hat{\sigma}_u^2\hat{\gamma}/\hat{\sigma}_d^2$.
5. Compare the potential bias to the actual coefficient estimate $\hat{\beta}$. If $\hat{\beta}$ is much larger than the potential bias (i.e., the ratio is much greater than 1), then it is unlikely that the observed relationship between D_i and Y_i is due solely to selection bias. If $\hat{\beta}$ is of similar magnitude to (or smaller than) the potential selection bias, then it is more plausible that the observed relationship between D_i and Y_i is due solely to selection bias.

Note that we impose the null hypothesis that $\beta = 0$ in the calculations above. But even if this hypothesis is violated, it shouldn't affect the potential bias too much unless β is large enough to generate a high partial R^2 (uncommon in applied work).

Unfortunately, this procedure is of limited use when the regression of Y_i on X_i has a low R^2 (a common scenario in applied work). A low R^2 implies that the ratio of the unexplained variation to explained variation is very high; in other words, the ratio of $\text{Var}(u)$ to $\text{Var}(X\delta)$ is very large. In this scenario, even a weak relationship between D_i and $X_i\delta$ can still translate into a large potential bias from u_i. Ultimately, you may need to argue that the unobserved factors determining Y_i are less correlated with D_i than the observed factors determining Y_i — but this is not much different from arguing that the selection on observables assumption holds.

5 Additional References

Frank, Ildiko, and Jerome Friedman. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, 1993, 35, 109-135.

Tibshirani, Rob. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B*, 1996, 58, 267-88.