Adjusting for Covariates, or Not

36-313

Lecture 9, 27 September 2022

Contents

1 Reminders and motivation **2**

2 All-else-equal in the linear model **5**
2.1 Counter-factual predictions with linear models 6
2.2 Why linear models? 6

3 An alternative to linear models: matching and nearest neighbors **8**
3.1 Contrasts and counterfactuals from the regression function 8
3.2 Estimating the regression function by matching and/or nearest neighbors 8
3.2.1 A very, very simple worked example of matching 10
3.2.2 Worked examples of matching and nearest neighbors 10
3.3 Other forms of nonparametric regression 10

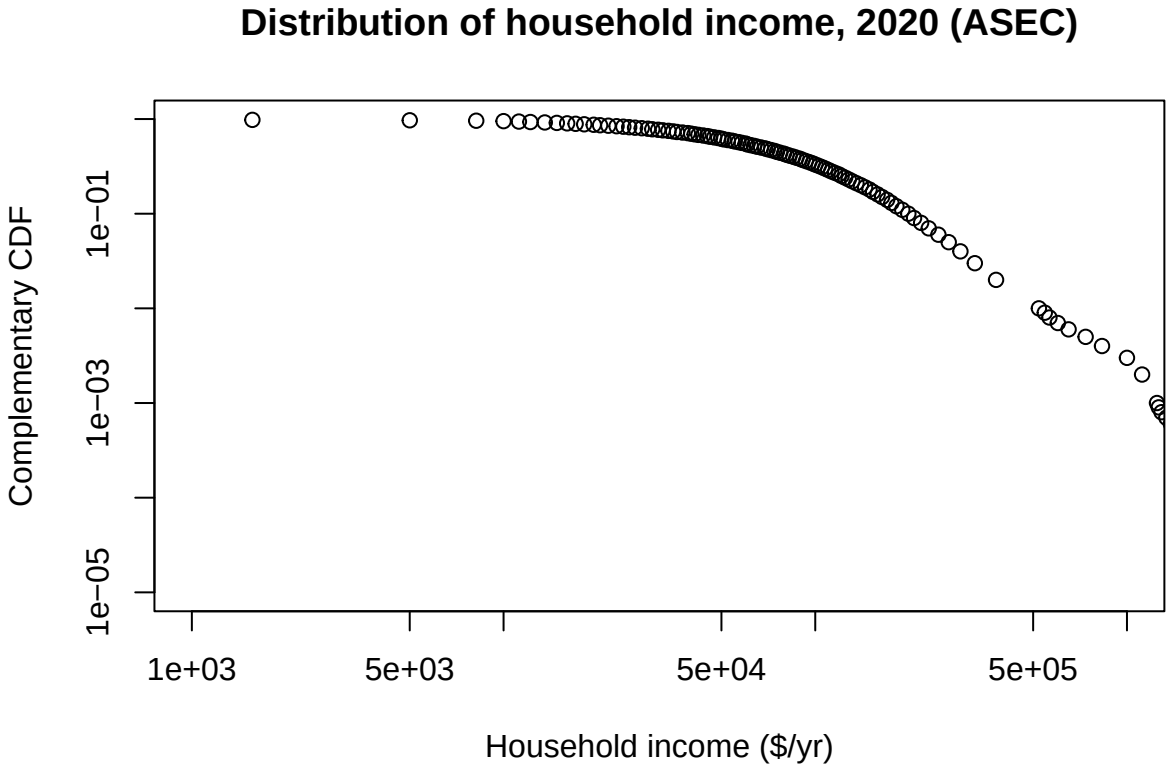
4 What do we adjust for? **11**

5 Further reading **12**

References **13**

1 Reminders and motivation

In previous lectures, we've looked at the distribution of variables like income and wealth across the population, and seen that they're very unequal. *How* unequal they are varies from country to country and year to year, with some interesting and important trends, but they're pretty much always right-skewed, heavy-tailed, and very unequal.



We've also seen that differences between social groups are very considerable compared to typical values.

Race	Mean	Median
Asian	133595.76	96847.0
Black	67063.79	44788.0
NBWA	78544.66	59364.0
White	98831.19	69814.0
Overall	96524.90	67002.5

College	Mean	Median
FALSE	70331.32	51145.0
TRUE	138412.67	102829.0
Overall	96524.90	67002.5

The largest raw difference in mean incomes between racial groups[1] is that between Asian Americans and African Americans, and amounts to 6.65×10^4 dollars/year, or 69% of the over-all mean. On the other hand, the difference between those who have and have not completed a bachelor's degree is 6.81×10^4 dollars/year,

[1] At this level of granularity; looking at a finer scale, people who self-identify as of mixed Native American and Hawaiian/Pacific Islander ancestry have the lowest mean household income in this survey.

which is 71% of the over-all mean.

On the *third* (or gripping) hand, whether or not someone has a college degree is strongly associated with race.

Race	Fraction
Asian	0.65
Black	0.28
NBWA	0.29
White	0.39
Overall	0.38

This is typical. In general, in *any* social-scientific problem, you will find that every variable is correlated with every other variable. Education is correlated with race, which is correlated with income, which is correlated with the area where you live[2], which is correlated with the type of job you have, which is correlated with age, which is correlated with...

The reason this is a *problem* for people interested in inequality is that it makes it harder to do fair comparisons. Take last three tables I just presented. If you think of education as (in some sense) primary[3], you have to wonder how much of the racial differences in the first table are *actually* reflecting the racial differences in educational attainment (in the third table) and the income differences by education level (in the second table). If you think of racial differences as primary[4], you could wonder how much of the differences in income by education in the second table are *actually* reflecting the racial differences in income (in the first table) and the racial differences in educational attainment (in the third table)[5]. If we are wishy-washy temporizers, we might hope that the *data* could tell us whether race or education or something else was primary.

Now why *does* anyone care about whether racial inequality drives educational inequality or vice versa?

1. *Scientific curiosity* Inequalities between groups are a conspicuous, persistent, nearly-universal phenomenon in human societies. Understanding how these inequalities arise is simple an important social-scientific issue, and worth investigating in its own right,
2. *Policy* Suppose (for the sake of argument) that it *did* turn out that racial inequalities in income were largely due to racial differences in educational attainment. If you wanted to reduce racial inequalities in economic outcomes, this is (potentially[6]) very useful information, because it suggests a *policy lever*, namely, doing something about educational inequalities. It could also inform you about the indirect consequences of other policies you might want to undertake[7].
3. *Measuring discrimination* In a lot of situations, people are *very* inclined to look at differences in outcomes (say, income) across groups and attribute those to discrimination (or prejudice, etc.). "Look, the average black family makes 67 thousand dollars a year less than the average Asian family!" A

[2]More specifically, average income levels increase with the number of people in the city. Economic geographers have known about this for a long time, as the "urban wage premium" (Thompson 1968). There are some claims that *average* income grows like a power of population (Bettencourt et al. 2007), but I am skeptical about that detail (Shalizi 2011). The flip side of this is that the cost of living is larger in larger cities, so their inhabitants do not necessarily have a higher standard of living.

[3]I think this is a fair paraphrase of authors like Adolph Reed, Jr., Walter Benn Michaels, and other serious scholars. Some of them, however, would insist on the primacy of "class" rather than "education". (But in contemporary societies, education and class are *very* strongly correlated.)

[4]I think this is a fair paraphrase of some authors who became extremely popular in 2020, and of some serious scholars as well.

[5]In principle, I suppose, you could ask how much of the racial differences in educational attainment in the third table are implied by the the combination of the racial differences in income (in the first table) and the educational differences in income (in the second table). But this seems so cart-before-the-horse that I can't think of anyone who'd seriously advocate this. (Doubtless one of you will inform me of counter-examples.)

[6]I say "potentially", because it could be that equalizing educational attainment across races is itself very hard to do. If it's *too* hard, our efforts might be better spent on some smaller cause of economic inequalities that's easier to manipulate. But we'd still want to know *how much* educational differences contributes to racial inequalities.

[7]Continuing with the (hypothetical) example, supposed you think there need to be more incentives for people to get educated. That means either increasing the rewards for the educated or reducing the rewards for the un-educated. Either way, you're talking about increasing economic inequality across educational levels. If you do that, and educational attainment is (still) unequal across races, you'll be increasing economic inequality across races. Even if you think that's a price worth paying, wouldn't you rather *know* that you're going to pay it, than be surprised by it?

natural rebuttal, which people are also very inclined to make, is that there are fair or appropriate reasons why outcomes should differ. E.g., maybe black people make less than Asians in part because they live in lower cost-of-living areas, and in part because they work in lower-paid occupations. "If you'd just compare black dentists in Chicago to Asian dentists in Chicago, rather than black janitors in Fayetteville, NC to Asian dentists in Chicago, you'd see that similar people are treated similarly!" The natural counter to the rebuttal is to want some way of setting "all else equal", and to say that *even so* there are differences by race (or gender, or whatever else).

2 All-else-equal in the linear model

What we *want*, then, is some way of saying what would happen if all else were equal, even though every variable is correlated with every other variable. This is where the linear model, a.k.a. multiple linear regression, has carved out a place for itself, and can *seem* to be the answer to the social scientist's prayers.

The way the model works is as follows. We're interested in some outcome variable Y (say, income), and how it relates to some **covariates** (or "features", "attributes", etc.) which for right now I'll call X and Z. Here you should think of X as the main variable of interest (e.g., race) and Z as the covariates or features which are also of interest (e.g., education)[8]. The model *assumption* is that

$$\mathbb{E}[Y|X=x, Z=z] = \beta_0 + \beta_1 x + \beta_2 z \tag{1}$$

Or (and this is strictly equivalent)

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 z + \epsilon & (2)\\ \mathbb{E}[\epsilon|X=x, Z=z] &= 0 & (3) \end{aligned}$$

I emphasize that this is an assumption because it's not always true; I'll come back to this.

Suppose for the moment that this *is* true. Then there is a *very* simple answer to the question of "what's the expected difference in Y if X changes, all else being equal?"

$$\begin{aligned} \mathbb{E}[Y|X=1, Z=z] - \mathbb{E}[Y|X=0, Z=z] &= (\beta_0 + \beta_1 + \beta_2 z) - (\beta_0 + \beta_1 0 + \beta_2 z) & (4)\\ &= \beta_1 & (5) \end{aligned}$$

Here for instance are the coefficients for the model where we use race *and* education to predict income:

	Estimate	Std. Error
(Intercept)	90000	3000
RACENAMEBlack	-42000	3000
RACENAMENBWA	-31000	4000
RACENAMEWhite	-17000	3000
COLLEGETRUE	66000	1000

This implies that if we compare a black person to a white person, then, *all else being equal*, then, on average, former's household income is 25 thousand dollars/year lower. Here "all else being equal" means "if they both have completed college, or both not completed college".

To repeat myself a little, we know that race and income are correlated, that race and education are correlated, and that education and income are correlated. If we know someone has a college degree, we can predict *something* about their race, and so *something* about their income. But if that was *all* that was going on, if we already knew someone's race, knowing their education shouldn't give us any more information about their income. The coefficient on `COLLEGE` in this model is telling us about how much we should adjust our prediction of someone's income from knowing about their education, *over and above* what we'd do from knowing their race, or from drawing inferences about their race from their education. Similarly, the coefficients for the racial contrasts are telling us about how to change our prediction of incomes even once we've accounted for education, or the inferences race lets us draw about education.

I am not, here, going to rehearse how we estimate a linear model like this, but will refer you to Lecture 8, and

[8] For categorical variables with k categories, we introduce $k-1$ "indicator" or "dummy" variables to keep track of the categories. I refer you to lecture 8 for details, and won't go over them here.

the references it gives. I will also refer you to that lecture for the notion of "interactions", which is how we can begin to include ideas like "the rewards of education vary by race" within the linear model framework.

2.1 Counter-factual predictions with linear models

A linear model is perfectly happy making predictions at any combination of x and z, it's just

$$\beta_0 + \beta_1 x + \beta_2 z \tag{6}$$

Whether there is any unit in the data with that particular value for the variables really doesn't matter. The assumption that $\mathbb{E}[Y|X=x, Z=z]$ is a linear function of x and z lets us make that prediction wherever we like.

I have said above that this lets us answer questions like "what would the expected difference in Y's be between $X=1$ and $X=0$, all else being equal?" But we can also answer questions like "What would the expected value of Y be, if $X=1$ had the same distribution of Z as $X=0$?", e.g., "what would the average income of black families be if they had the same educational attainment as Asians?" This'd just be

$$\int (\beta_0 + \beta_1 + \beta_2 z) p(z|X=0) dz \tag{7}$$

$$= \beta_0 + \beta_1 + \beta_2 \int z p(z|X=0) dz$$

$$= \beta_0 + \beta_1 + \beta_2 \mathbb{E}[Z|X=0] \tag{8}$$

$$\approx \beta_0 + \beta_1 + \beta_2 \frac{1}{n_0} \sum_{i\,:\,X_i=0} z_i \tag{9}$$

where in the last line I'm using the law of large numbers to approximate $\mathbb{E}[Z|X=0]$ with the data (and n_0 = the number of observed units where $X_i = 0$). If we're interested in what would happen if we imposed other distributions on Z, we can also answer that question from the linear model.

Similarly, the linear model tells us that the expected difference between two groups, say $X=1$ and $X=0$, is

$$\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] \tag{10}$$

$$= \mathbb{E}[\mathbb{E}[Y|X=1, Z]\,|X=1] - \mathbb{E}[\mathbb{E}[Y|X=0, Z]\,|X=0]$$

$$= \beta_0 + \beta_1 + \beta_2 \mathbb{E}[Z|X=1] - (\beta_0 + \beta_2 \mathbb{E}[Z|X=0]) \tag{11}$$

$$= \beta_1 + \beta_2 (\mathbb{E}[Z|X=1] - \mathbb{E}[Z|X=0]) \tag{12}$$

which lets us say how much of the difference between two groups can be accounted for by their having different covariates (namely, $\beta_2 (\mathbb{E}[Z|X=1] - \mathbb{E}[Z|X=0])$) and how much cannot (namely, β_1).

2.2 Why linear models?

There are three big reasons for using linear models in the study of social inequality.

1. *Rhetorical*[9]: The linear model gives us a fairly simple way of saying how much of the difference between groups is accounted for by to each of the covariates, and how much is *not* accounted for by the covariates. This is often extremely useful when it comes to communicating with people about inequality, and persuading of them of the merits of one or another position about that inequality (that it's not a big deal, that it's a very big deal, that we can solve it with education, that education won't fix it, etc.).

[9] I realize that people often use "rhetoric" as a dirty word (as in "mere rhetoric" or "cheap rhetoric"), or to suggest that someone is being deceptive or manipulative. But rhetoric really just means the art of persuasion, and often the best way to persuade your audience is to present them with good reasons, sound arguments and compelling evidence. In *this* sense, rhetoric is one of the key skills that an educated person needs to have. A big part of the discipline of statistics can in fact be seen as a branch of rhetoric, designed to persuade an audience who are skeptical and good with numbers. (This is a point I learned from Abelson (1995) a long time ago.)

2. *Computational*: Fitting a linear model by least squares boils down to some matrix algebra, and after two hundred plus years of work we have *very* good algorithms for doing that matrix algebra[10], and can do it at very large scales[11].
3. *Traditional*: Because the computational demands are so modest, for a very long time linear models were about the only *usable* statistical models. Statisticians developed *ideas* about nonlinear models from a very early date, but when computing had to be done by paper, pencil and people, those ideas couldn't really be *implemented*. While that's changed tremendously since programmable computers were invented in the 1940s, and especially since personal computers became common in the 1980s, there is still a *lot* of cultural and organizational inertia which pushes people to keep using linear models. (More positively, there is a lot of inherited wisdom about how to use linear models, what makes for a good linear model, etc., to draw on.)

You may have noticed that I do *not* list "accuracy" or "realism" or "matching the data well" as reasons people often use linear models. There are certainly situations where linear models are pretty accurate, but my experience is that they're rare. Generally speaking, there's no good mathematical reason why we should expect linear models to be accurate[12]. Our most trusted scientific theories also don't tell us to expect to use linear models, and in social questions about inequality we have precious few trusted scientific theories in the first place.

[10] Recall (from Lecture 8 if not elsewhere) that if we have p features in total, the vector of estimated parameters $\hat{\beta}$ is a $(p+1)$ vector (since we need to include the intercept), and $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$, where $\mathbf{x}$ is the $[n \times (p+1)]$ matrix of features (plus a column of 1s for the intercept) and $\mathbf{y}$ is an $[n \times 1]$ matrix of the y values. Multiplying an $[a \times b]$ matrix by a $[b \times c]$ matrix to get an $[a \times c]$ takes $O(abc)$ arithmetic operations (b multiplications and $b-1$ additions per entry in the product matrix). (There are cleverer algorithms which can do better for very large matrices.) Thus finding $\mathbf{x}^T\mathbf{x}$ takes $O(p^2 n)$ steps. Similarly, $\mathbf{x}^T\mathbf{y}$ can be computed in $O(pn)$ steps. Inverting a $[a \times a]$ matrix takes $O(a^3)$ operations if done straightforwardly from the definitions, though again there are more complicated algorithms which scale better for very large matrices. So inverting $\mathbf{x}^T\mathbf{x}$ once we have it takes $O(p^3)$ steps. The over-all time needed to find $\hat{\beta}$ is therefore $O(p^3 + p^2 n)$. Once we have $\hat{\beta}$, making a prediction for every data point takes an extra $O(pn)$ steps, so it hardly matters in comparison.

[11] Some of those algorithms for truly large data involve a certain amount of randomization and approximation (Mahoney 2011). But that's fine; "use only a small random subset of the data for any one step" is a basic design move for all kinds of big-data models.

[12] A partial exception is if we think we're dealing with a very smooth function over a very small range. Then basic calculus, in the form of Taylor's approximation theorem, says that a differentiable function is *approximately* a linear function, with the slopes being the partial derivatives. The range over which this approximation holds is determined by the curvature, i.e., the second derivatives.

3 An alternative to linear models: matching and nearest neighbors

The goal of the linear model is to approximate the conditional expectation function $\mathbb{E}[Y|X=x, Z=z]$. It's inconvenient to keep writing that out, so I'll abbreviate it:

$$\mu(x,z) \equiv \mathbb{E}[Y|X=x, Z=z] \tag{13}$$

The linear model *assumes* $\mu(x,z)$ is linear in x and z. But there are lots of other regression models we could try to use to estimate the **regression function** μ.

3.1 Contrasts and counterfactuals from the regression function

Suppose the Oracle, or Someone, were to simply *tell* us $\mu(x,z)$. We could then say what the expected difference was between any two individuals or groups defined by the values of their features, e.g.,

$$\mathbb{E}[Y|X=1, Z=z] - \mathbb{E}[Y|X=0, Z=z] = \mu(1,z) - \mu(0,z) \tag{14}$$

would be the expected difference in Y between those with $X=1$ and $X=0$ when Z is held equal to z. In the linear model, this was just β_1, regardless of z, but if there are nonlinearities, we can't just say there is *a* contrast between $X=1$ and $X=0$ regardless of Z, we have to specify the value of the covariates.

Of course, we *can* define an average or typical value for the contrast if we want to. The expected contrast would just be

$$\int \mu(1,z) - \mu(0,z) p(z) dz \approx \frac{1}{n}\sum_{i=1}^{n} (\mu(1,z_i) - \mu(0,z_i)) \tag{15}$$

where the approximation comes from using the data to approximate the true distribution of z. (I will let you work out the definition for the median contrast, etc.) Again, in the linear model, this is just β_1, but we don't have that simplification when we let the model be nonlinear.

If we want to know the average outcome for the $X=1$ group if they had the same distribution of Z as the $X=0$ group, that would be

$$\int \mu(1,z) p(z|X=0) dz \approx \frac{1}{n_0} \sum_{i:X_i=0} \mu(1,z_i) \tag{16}$$

If instead we want to know the average outcome for the $X=0$ group if they had the same distribution of Z as the $X=1$ group, that would be

$$\int \mu(0,z) p(z|X=1) dz \approx \frac{1}{n_1} \sum_{i:X_i=1} \mu(0,z_i) \tag{17}$$

In short, if we knew $\mu(x,z)$, we could come up with *an* answer to any of the questions the linear model lets us answer, though it might be a more context-dependent answer, and/or one that requires more calculation than just looking at the coefficient table.

3.2 Estimating the regression function by matching and/or nearest neighbors

Here is a *very* simple idea for how to estimate the regression function:

- We want to get an estimate of $\mu(x,z)$, say $\hat{\mu}(x,z)$.
- We go to our data and find all of the **matching** cases or units, i.e., all the i where $(x_i, z_i) = (x,z)$

- We average the y_i for those cases,

$$\hat{\mu}(x,z) = \frac{1}{n_{x,z}} \sum_{i:(x_i,z_i)=(x,z)} y_i \tag{18}$$

If X and Z are both categorical variables, this approach is in some sense guaranteed to work[13]. But if any covariate is continuous, or even if there are just a very large number of categories, there's the difficulty that we might not find an exact match. This suggests a slightly refined approach, called **nearest neighbors** or k-**nearest neighbors**.

- We want to get an estimate of $\mu(x,z)$, say $\hat{\mu}(x,z)$. The point (x,z) may or may not appear in our data set.
- For each $i \in 1:n$, we calculate the distance between (x,z) and the data point (x_i,z_i), say d_i. We then rank the points by distance.
- We average y_i for the k closest points,

$$\hat{\mu}(x,z) \equiv \frac{1}{k} \sum_{i:d_i \leq d_{(k)}} y_i \tag{19}$$

  where $d_{(k)}$ is the distance to the k^{th} nearest neighbor[14].

There is a trade-off here. If we make k very small, we can pick up a lot of detail in the regression function, but our estimates are also very noisy, because each predicted value derives from only a few, perhaps only one, observation. If we make k very large, we "smooth out" a lot of detail, but also gain more stability and reduce our vulnerability to noise. The right k isn't so much an aspect of the *world* as of our modeling procedure; we typically let $k \to \infty$ as $n \to \infty$, though with $k/n \to 0$.

There are a couple of points to make about this "nearest neighbor method" here.

1. *It assumes almost nothing*: The nearest neighbor method doesn't care at all about what the true regression function looks like. Any functional form, no matter how nonlinear, nonadditive, etc., is fine.
2. *It's generally consistent*: If $k \to \infty$ as $n \to \infty$ while $k/n \to 0$, then the estimated regression function converges on the true regression function, $\hat{\mu}(x,z) \to \mu(x,z)$, at least if $p(x,z) > 0$. The convergence slows down as the true function μ gets rougher and less smooth, but it still happens even for very rough, oscillating functions. More seriously, it slows down as the number of covariates increases. The nearest neighbor method isn't good at extrapolating beyond the data, or interpolating to an (x,z) point with 0 probability in the training data, but that sort of thing is intrinsically hard to do.
3. *"Coefficients? We don't need no stinking coefficients!"* There are no coefficients anywhere in the model. This is an example of a fully **nonparametric** regression, because there's nothing like the β slopes of the linear model at all. This is a strength (because it doesn't pre-commit us to any particular functional form), but also a weakness (because it makes the model harder to communicate).
4. *It's more computationally involved*: Once we have the linear model coefficients, making a prediction just involves the basic arithmetic operations of multiplication and addition. Finding the coefficients involves some matrix algebra, but is not much more complicated at its heart. Nearest neighbors involves calculating distances and then *searching* for the k smallest distances, and search or sorting is a different, and much trickier, kind of operation[15]. The upshot is that for all the *conceptual* simplicity of nearest neighbors, you don't want to write your own code for doing it, but rather rely on someone who's actually studied sorting algorithms, and finding matches in large data bases.

[13] As $n \to \infty$, $n_{x,z}/n \to p(x,z)$ (by the law of large numbers), and then the average of the y_is will $\to \mathbb{E}[Y|X=x,Z=z]$ (again by the law of large numbers).

[14] Break ties however you like.

[15] To make a prediction at a new point using k-nearest neighbors, I need to calculate the distance between my point of interest (x,z) and each (x_i,z_i). With p covariates, each distance takes $O(p)$ arithmetic operations, so that's $O(np)$ steps. I then need to sort that list of distances, which takes $O(n)$ time if I use an efficient algorithm, and longer if I do not. Then I need to average the k different y values for the nearest neighbors, which takes $O(k)$ steps. Overall, one prediction takes $O(np+n+k) = O(np+k)$ steps. Making n predictions thus takes $O(n^2p+nk)$ steps. The comparable time for the linear model is instead $O(p^3+p^2n)$, as discussed above. Which is faster depends on the balance between p and n.

3.2.1 A very, very simple worked example of matching

With only four racial categories and two educational levels, we can always find a pretty large number of exact matches for any of the eight possible values of (x, z). Exact matching would thus give us the following averages.

	Non-college	College
Asian	75700	164400
Black	50600	109200
NBWA	64000	113800
White	73200	139400

As we saw earlier, college completion rates are very different by race. 65% of Asians have at least a bachelor's degree, but only 28% of blacks do. The actual mean income of blacks is thus 67 thousand dollars/year, but if blacks completed college at the same rate as Asians, it would be 89 thousand dollars/year. This suggests that while the difference in educational attainment can account for *some* of the racial differences in income, it certainly can't make it all go away.

3.2.2 Worked examples of matching and nearest neighbors

Are deferred to homework 5.

3.3 Other forms of nonparametric regression

There's nothing magic about nearest neighbors and matching. It's one way of doing nonparametric regression, among many others[16] — kernel smoothing, splines, regression trees, neural networks, etc., etc. My experience in applied problems is that nearest neighbors is very easy to explain to non-statisticians[17], which is not to be sneezed at. (Rhetorical advantages are real advantages.)

[16] Take 36-402 and/or 36-462.

[17] Its only rival, in this respect, is regression trees, because lots of people get flow-charts. See the chapter on trees in Shalizi (n.d.).

4 What do we adjust for?

I have been using a very, very simple set of covariates in the running examples here, because I wanted to keep the technicalities to a minimum. In practice, deciding what to adjust for, and what to leave out, is a crucial and contentious issue. In this week's homework, for instance, you are looking at a linear model of racial differences in household income, which adjusts for education, occupation, and geography. Without spoiling the discussion from the next lecture, one could make a case for or against including any of these as covariates (Zs, basically) when the variable of primary interest is race (X). Take geography: black people still disproportionately live in the rural South, because that is where their ancestors were brought as slaves[18], whereas Asian Americans disproportionately live in big cities, because that's where they arrived as immigrants[19]. If we "control for" where people live, are we thereby arriving at a fairer comparison of more similar people, or are we covering up one of the *mechanisms* that produce racial inequality? But then again, even if the current patterns of where people live were produced by racism in the past, it is indeed now the case that anyone can live wherever they want (if they can afford it), so maybe it *is* appropriate to control for location.

What covariates should be controlled for in an analysis is one of the key questions when using data to study inequality. It involves substantive issues of causal structure (how variables influence each other), purely statistical concerns about measurement and modeling, and more philosophical issues about what our goals are in doing the analysis anyway, and what *exactly* we mean by "all else being equal". We'll devote the next lecture to it, but for now I just want to make you uneasy.

[18] A sizeable fraction of the African American population escaped from the rural South to northern cities in the early 20th century, in the "Great Migration" (Lemann 1991; Wilkerson 2010).

[19] Various legal, and extra-legal, means were used in the 19th and early 20th century to *keep* Asian immigrants, specifically, from moving into the countryside if they wanted to.

5 Further reading

I refer you (once more, for luck) to Lecture 8, and its references, on linear models, analysis of variance, and least squares.

Nearest neighbors are covered extensively in 36-462, e.g., here from Spring '22 (.Rmd). Those notes also include a fair amount of discussion on the history of nearest neighbors, and on computational techniques for working with truly large data sets, and references to the more advanced theory.

Matching is a fundamental technique in causal inference (Rubin 2006), though it can be used, as here, without necessarily making any causal assumptions. The connection to nearest neighbors seems to have originated as "folklore", but was formalized by Abadie and Imbens (2006). See also the causal inference chapters of Shalizi (n.d.).

References

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74:235–67. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Abelson, Robert P. 1995. *Statistics as Principled Argument*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bettencourt, Luís M. A., José Lobo, Dirk Helbing, Christian Künhert, and Geoffrey B. West. 2007. "Growth, Innovation, Scaling, and the Pace of Life in Cities." *Proceedings of the National Academy of Sciences (USA)* 104:7301–6. https://doi.org/10.1073/pnas.0610172104.

Lemann, Nicholas. 1991. *The Promised Land: The Great Black Migration and How It Changed America*. New York: Knopf.

Mahoney, Michael W. 2011. "Randomized Algorithms for Matrices and Data." *Foundations and Trends in Machine Learning* 2:123–224. https://doi.org/10.1561/2200000035.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge, England: Cambridge University Press.

Shalizi, Cosma Rohilla. 2011. "Scaling and Hierarchy in Urban Economies." arxiv:1102.4101. http://arxiv.org/abs/1102.4101.

———. n.d. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Thompson, Wilbur R. 1968. *Preface to Urban Economics*. Baltimore: Johns Hopkins University Press.

Wilkerson, Isabel. 2010. *The Warmth of Other Suns: The Epic Story of America's Great Migration*. New York: Random House.