Data Analysis for Social Scientists (14.1310x) review notes.

David G. Khachatrian

September 28, 2019

1 Preamble

This was made a good deal after having taken the course. It will likely not be exhaustive. It may also include some editorializing: bits of what I believe are relevant observations and/or information I have come across.

Also, many of the earlier topics of probability and "theoretical" statistics were covered in detail in the Probability (6.431x) and Fundamentals of Statistics (18.6501x) notes, so those topics will either be skipped or mentioned very quickly in these notes.

2 Probability Terms

A *probability* on a sample space S is a collection of numbers $P(A)$ that satisfy sigma-algebra properties.

3 Exploratory data analysis.

One can often get a sense of the data by plotting *histograms* of the features in question.

You may want to smooth out your plots. One can use *kernel density estimation* to achieve this. We extrapolate from known datapoints x_i using a kernel function K:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K(\frac{x - x_i}{h})$$

K is a probability density with mean 0, and $h > 0$ is the bandwidth. Increased h increases the effect each point x_i on faraway points (it "smooths/flattens things out"). Common choices for K are Epanechnikov or Normal (here, σ^2 is essentially the bandwidth parameter).

(Note that kernel density estimation is a distortion added to the visualization of the data).

Sometimes you'll want to plot the CDF. This is especially helpful to determine the existence of *first-order stochastic dominance (FOSD)*. X dominates Y to the first-order if X's α-quantile is greater than or equal to Y's α-quantile for all $\alpha \in (0, 1)$, i.e., $F_X^{-1}(\alpha) \geq F_Y^{-1}(\alpha)$. This corresponds to $Pr[X \leq k] \leq Pr[Y \leq k] \forall\, k \in \mathbb{R}$.

Intuitively, "X is always more likely to yield larger numbers than Y". Visually, the CDF of X is always "to the right of" the CDF of Y.

4 Auctions

The k'th order statistic of n realizations of a random variable X, denoted $X_n^{(k)}$ (n usually suppressed) describes the probability law governing the distribution of the $\frac{k}{n}$-quantile of X. Perhaps better explained in mathematical terms (assuming iid draws),

$$Pr[X_n^{(k)} \leq k] = Pr[\text{k of n realizations of X are } \leq k] \times Pr[\text{(n-k) realizations of X are } > k]$$
$$F_Y(k) = \binom{n}{k}(F_X(k))^k(1 - F_X(k))^{n-k}$$

We can compare two scenarios, both times of which will have N potential buyers of a good with distribution of valuation/offers O (and we assume zero transaction costs):

1. The seller fixes a price $P_{threshold} = P_t$ (which can be chosen optimally if the seller knows both N and the distribution of offers O), and sells at the first offer at or above P_t (so $P \geq P_t$).
2. The seller sets up an auction; they sell to the second-highest bid (the $\frac{N-1}{N}$-quantile of realized values), P.

The goal is to maximize expected profit/revenue $E[P]$.

One can go through the calculations (assuming, for example, that $O \sim U[0,1]$) to find that the auction scenario benefits the seller for $N > 2$. In general, for large enough N, an auction leads to a better outcome for the seller, even though they don't need to know O in the auction scenario! This is essentially *price discovery* in action.

5 Some new distributions

5.1 Hypergeometric Distribution

The *hypergeometric distribution* arises when you sample without replacement (in comparison to the binomial distribution, where you sample *with* replacement). Say you sample n items without replacement from a pool with s "successes" and f "failures" (with a total of $s + f = N$ items). Then the number of successes in that sample X follows the hypergeometric distribution:

$$p_X(x; s, f, n) = \frac{\binom{s}{x}\binom{f}{n-x}}{\binom{s+f}{n}}$$

There is a similarity to the binomial in terms of expectation and variance as well. Call the fraction of successes before first draw $p = \frac{s}{s+f}$ and the fraction of failures before first draw $q = \frac{f}{s+f} = 1 - p$:

$$E[X] = n\frac{s}{s+f} = np$$

$$Var(X) = n\frac{s}{s+f}\frac{f}{s+f}\frac{s+f-n}{s+f-1} = np(1-p)\frac{s+f-n}{s+f-1} \leq np(1-p)$$

5.2 Negative binomial distribution

The negative binomial distribution of order r, $NB(p, r)$ is the sum r iid $Geom(p)$ r.v.'s. (This is the discrete analog to "the Erlang distribution of order k, $Erlang(\lambda, k)$, is the sum of k iid $Exp(\lambda)$ r.v.'s.)

5.3 F distribution

If $X \sim \chi^2_n$, $Y \sim \chi^2_m$, and X and Y are independent then

$$\frac{X/n}{Y/m} \sim F_{n,m}$$

This is useful when using an *F-test* to compare how good two models compare to one another. (Longer explanation about F-test to come.)

6 RCT design and power calculations.

The *power* of a test is the probability of a false negative.

Say you have N individuals, you set γ fraction of them to the treatment group and the rest as control (both samples large enough to be able to invoke CLT). You choose significance level α. You assume that the treatment has a constant effect (of unknown magnitude) τ on the population. That is to say, you assume that:

$$\bar{X}_c \sim N(\mu, \frac{\sigma^2}{\gamma N}), \quad \bar{X}_t \sim N(\mu + \tau, \frac{\sigma^2}{(1-\gamma)N}$$

for some shared σ^2 between the two populations (note: this is another assumption, which you can relax if you have reason to). Now in total there are the following parameters in our experimental design:

1. the effect size of treatment τ
2. the variance of the underlying population distribution σ^2
3. the probability of Type II error (for a given threshold/effect size) β
4. the probability of Type I erro α
5. the number of observations/participants in the study N
6. the fraction of participants that join the treatment group γ

When you fix all the other parameters, you can calculate the induced value for the final parameter. $\gamma = 0.5$ is always the most "efficient" partition but may not always be reasonable/ethical (e.g., medical study for new disease treatment that turns out to be really effective). α is usually fixed at a low value to limit false positives. The higher the N, the better (but usually increasing N is costly).

Usually, you choose an effect size τ' such that, if $\tau < \tau'$, you "may as well not even bother". You usually also estimate σ^2 from previous studies. (Both τ and σ^2 could also be estimated from a smaller pilot study of the intervention under consideration.) From there, you can calculate the final value.

The relevant test statistic is

$$T = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{V_{Neyman}}} \approx N(0,1)$$

where the *Neyman variance assumes independence between the populations*, i.e. $V_{Neyman} = \frac{\sigma^2}{N_t} + \frac{\sigma^2}{N_c}$.

6.1 Experimental design effects on variance

In a *matched design*, you have paired participants exactly by matching them on other features which may confound treatment effect (socioeconomic status, location, age, etc). You then randomly choose one of the pair to join the control arm and the other to join the treatment arm.

The matched design ends up lowering the variance of the estimated treatment effect (which is good!).

The slightly less beefy version of matched design is *stratified design*, where you end up having more than two people in each "group", and you randomly send γ of them to the treatment arm and the rest to the control arm.

Stratified design still lowers the variance of the estimated treatment effect, but not as much as matched design. (Stratified design where you end up only having one stratum basically leads to a "regular" split.)

Clustered design involves combining clusters of participants into just 1 datapoint.

Clustered design *increases* the variances of estimated treatment effect, so it's not recommended if you can avoid it. Clustered design tends to be by necessity rather than desire. For example: you stage an intervention in different classrooms in a school. Ideally, you'd get information about all the students in each classroom and compare their treatment effect; in reality, the school can only provide information about aggregate effects across classes with no per-student breakdown. What would otherwise be $\sim$ 30 datapoints/students become 1 datapoint/class. (So with lower n and no improvement anywhere else, variance will go up.)

7 Causality.

Often, we are interested in trying to tease out *causality* among different factors. A common method is *structural equation modeling (SEM)*, where we assume a structure between observable variables (e.g test scores) and latent variables (e.g. intelligence), and then calculate the parameters that link these variables together based on e.g. the covariance matrices.

This course looks more at the *Rubin causal model*. We will want to design our experiment so that the *stable unit treatment value assumption (SUTVA)* holds. SUTVA assumes that "my choice of treat/no-treat for Person A does not affect the outcome for Person B, and vice-versa". (An example where SUTVA would not hold is if you choose parts of a local population to get immunized and others don't – Person A getting immunized improves Person B's prospects regardless of whether B gets immunized.)

Another important aspect is to *randomly assign individuals* to the different arms of the study. This ensures no built-in *selection bias*.

Why? A key idea is that "we could have gotten different outcomes for the same person i if they had been given a different intervention $k \neq j$, so in general $Y_{i,j} \neq Y_{i,k}$."

For an individual i and a study with A treatment arms, let $Y_{i,j}$ be what you would observe (for whatever it is you're measuring with this study) if you put person i in arm j. We can then assign people i to arm j; denote the choice of assignment W_i (which equals $0, 1, \cdots, j$, depending on which arm Person i is assigned to). Since we don't have A clones of people, we'll only be able to observe one branch. For simplicity, let's have $A = 2$ ("either you get a treatment or you get a control"). So $Y_i^{(obs)} = (Y_{i,j} \mid W_i = a) = Y_{i,a}$.

We want to estimate the effect of treatment, so what would make sense is to use sample estimators for $E[Y_{i,1}] - E[Y_{i,0}]$. We expand and rearrange:

$$\begin{aligned} E[Y_{i,1}] - E[Y_{i,0}] &= (E[Y_{i,1} \mid W_i = 1] + E[Y_{i,1} \mid W_i = 0]) - (E[Y_{i,0} \mid W_i = 0] + E[Y_{i,0} \mid W_i = 1]) \text{ (law of total expectation)} \\ &= (E[Y_{i,1} \mid W_i = 1] - E[Y_{i,1} \mid W_i = 0]) + (E[Y_{i,0} \mid W_i = 1] - E[Y_{i,0} \mid W_i = 0]) \\ &= \text{effect of treatment on the treated} + \text{selection bias} \end{aligned}$$

What happened here? What does $Y_{i,0} \mid W_i = 1$ mean? This refers to what response the people in the treatment arm would have given if they had instead been in the control arm. If there's a systematic difference between the groups (their baseline responses would differ, i.e., $(E[Y_{i,0} \mid W_i = 1] - E[Y_{i,0} \mid W_i = 0]) \neq 0$), selection bias is at play. No good.

How can we prevent selection bias from occurring? The simplest way is to use *random assignment* (so any bias decreases with increasing n due to Law of Large Numbers). Another benefit is that if we assume that there is no selection bias, that means the two groups are essentially "from the same population". So our conditional estimates $Y_{i,j} \mid W_i = j$ can describe the whole population and not just "those people with the characteristics we selected for inclusion in our treatment arm", etc. A double-win!

Note that if the above is true, we can view all of this as a linear model (if we're comfortable making the linear model assumptions, e.g. Gaussian homoskedastic noise with no serial correlation). We are trying to estimate $E[Y_i \mid X_i]$ with the following:

$$Y_i = \beta_0 + \beta_1 \text{IsTreatment}_i$$

In this case, β_0 estimates the average response for the control group and β_1 estimates the average marginal benefit from undergoing treatment. (The point estimates for β will correspond to the sample means/difference of sample means of the two groups.)

7.1 Fisher and the Sharp Null

We could proceed as normal. Our natural estimator for the treatment takes sample means: $(\bar{Y}_i \mid W_i = 1) - (\bar{Y}_i \mid W_i = 0)$. Like before, we assume that the two populations are independent (knowing how one group doesn't tell us how the other group did – this is essentially the assumption behind SUTVA), so we construct the unbiased sample variance. We can run things as normally from there; if we assume $Y_{i,j} \mid W_i = j$ is Normally distributed, we can run a t-test, etc.

But we can decide to view things differently. Rather than having sampled randomly from a larger population, what if the data we collected *is* our population of interest? Then we'd have a good deal of information; if there are a treatment arms, we have $1/a$ of *all* the information about our population. Let's return to just two arms; then we have half of all the information.

So here's a question: Can we test whether the intervention has any effect on *anyone*? This is a hypothesis test:

$$\begin{cases} H_0 : Y_{i,0} - Y_{i,1} = 0 \forall\ i \\ H_1 : \exists\ i\ s.t.\ Y_{i,0} - Y_{i,1} \neq 0 \end{cases}$$

Note that this is a pretty intense assumption. We aren't saying "maybe the mean effect is zero"; we're saying "this doesn't do *anything* to *anyone*". Hence why we call it a *sharp null hypothesis*.

Under the null, we now have *all* the relevant information and can "fill in" our empty rows. We now need to choose a relevant test-statistic: one choice is the absolute difference in sample means: $T = \left|\bar{Y}_{i,1} - \bar{Y}_{i,0}\right|$.

How do we construct a p-value/measure of significance? We have a measure of T_{obs} for how our experiment split up individuals W_i. What are all the other T corresponding to different treatment-control splits? That gives us a range of possible T values. We can then see how extreme T_{obs} was and calculate the p-value from that.

To be clear, let's say you had N_c people in the control arm and N_t in the treatment arm. This means you calculate $\binom{N_c}{N_c+N_t}$ combinations of hypothetical assignments, calculate T for each one, and then compare how extreme the actual assignment's T_{obs} was. A combinatorial number of calculations for a test isn't great – accordingly, this sort of test is usually only really considered for small sample sizes. And often, one would use *simulations* rather than calculate the distribution of T exactly.

8 Nonparametric comparisons/regressions.

Say you assume a model $Y = E[Y \mid X] + \epsilon = g(X) + \epsilon$, and you want to estimate g without imposing a functional form. We can use a kernel to describe an estimator based on the observation:

$$E[Y \mid X] = \frac{E[Y \mid X]}{E[1 \mid X]} \to \hat{g} = \frac{\sum_i y_i K(\frac{x - x_i}{h})}{\sum_i K(\frac{x - x_i}{h})}$$

As $h \to 0$, the bias in your estimator goes to 0. As $nh \to \infty$, the variance in your estimator goes to zero. (So, "the less you smear each point everywhere, the less you're aiming at the wrong location. And the more points you have, the less spread you'll have.")

You have to choose both your kernel and your bandwidth. If you've chosen the kernel, how do you choose an optimal bandwidth? Cross-validation. You fit to a random subset of points and find the h that minimizes the sum of squared residuals on the holdout set of points.

Other nonparametric fitting methods include series estimation, spline interpolation, local linear regression (LOESS).

9 R^2 and the F-test

We can view R^2 as answering "How much of the variance in Y does our model $M = \hat{y} = g(x)$ explain?" From this explanation, the formula is hopefully clearer:

$$R^2 = 1 - \text{fraction of variance not explained by } M = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

As it turns out, these sum of squared residuals follow χ^2 distributions with differing degrees of freedom, so with a bit of finagling, you can construct an F-statistic of your model M and run an F-test to see whether your R^2 is statistically significant. (Whether that's of practical use depends on your model, the value of R^2, etc.)

In general, you can run F-tests based on comparing a "full" model $\hat{y}_f$ and a "restricted" model $\hat{y}_r$ with f and r degrees of freedom, respectively. Then you can write:

$$T = \frac{\text{scaled improvement in SSR using fuller model}}{\text{SSR not explained even by the fuller model}} = \frac{(SSR_r - SSR_f)/((n-r)-(n-f))}{(SSR_f)/(n-f)} = \frac{\frac{SSR_r - SSR_f}{f-r}}{(SSR_f)/(n-f)} \sim F_{f-r,n-f}$$

10 Considerations and methods for running regressions.

10.1 Interaction effects and their uses (e.g., "Difference-in-Differences").

Consider a regression of the form:

$$Y_i = \delta + \alpha A_i + \beta B_i + \gamma A_i * B_i + \epsilon_i$$

A_i and B_i are two features of an observation. What does (the parameter γ for) $A_i * B_i$ represent? The *interaction effects* of having both A_i and B_i at the same time. For example, A_i can represent the effect of a treatment for female participants, B_i can represent the baseline difference between males and females, and $A_i * B_i$ would represent the difference in effect of the treatment for males compared to females. It's a "difference between male and female" on top of a "difference between treatment and control" – a *difference-in-differences* model.

An assumption behind this model is that the difference between two groups (B_i/β for male vs. female in this example) would have remained stable over time if no intervention had been applied (A_i/α). Without this assumption, there is "mixing" between β and γ (which ultimately leads to higher variances due to having correlated features, plus inaccurate point estimates since γ "steals" some of the change that would have happened over time anyway).

10.2 Transformations of the dependent variable.

In the above regressions, we have assumed a linear dependence between target variables and features: $Y_i = X_i^T \beta + \epsilon_i$. What if we assume different functional forms? We can adjust our expressions to still perform a sort of linear regression! Examples include:

Logarithmic transformation Assume the correct form between X_i and Y_i is $Y_i = AX_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{\epsilon_i}$. Take the log of both sides to get a linear form: $\log(Y_i) = \beta_0 + \beta_1 \log(X_{1i}) + \beta_2 \log(X_{2i}) + \epsilon_i$.

Box Cox Transformation In this case we assume the correct form is $Y_i = (X_i^T \beta + \epsilon_i)^{-1}$. Invert to have $\frac{1}{Y_i} = X_i^T \beta + \epsilon_i$.

Discrete Choice Model In this case we assume a sigmoidal shape/softmax: $P_i = \frac{\exp(X_i^T \beta + \epsilon_i)}{1+\exp(X_i^T \beta + \epsilon_i)}$. Then we can get to a linear form via $Y_i = \log\left(\frac{P_i}{1-P_i}\right) = X_i^T \beta + \epsilon_i$.

Note that this is highly reminiscent of *generalized linear models* using link functions g so that $g(E[Y \mid X]) = X^T \beta$. Though we have not explicitly written our probability distribution for $Y \mid X$, we are implying said distribution and are essentially performing the same sort of optimization, which likely lacks a closed-form solution (and so would using, e.g., iteratively reweighted least squares).

10.3 Nonlinear transformations of the independent variables.

Let's say you visualize the data and you suspect that Y_i is not linearly related to X_i, but instead to $f(X_i)$? Well, you can calculate that feature and put that in the linear model! This is *feature engineering* and is an important part of any model creation process. This can also include interaction terms, e.g. $X_{1i} \times \log(X_{3i})$ or $X_{2i} \times \text{IsTreatment}_i$. (Don't go too crazy here though – ideally you can interpret the features you're creating.)

If you're not sure what form Y_i takes, you can try to perform a regression on a series expansion of $X_1 i$: $Y_i = \sum_{j=0}^{k} \beta_j X_{1i}^j + \epsilon_i$. This is a non-parametric (distribution-free) method called *series regression*.

10.4 Locally Linear Regression

A nonparametric estimation method that is generally better than kernel estimation is *locally linear regression*. It still involves kernels, but rather than trying to fit the best "flat-lines" around each point, it tries to fit the best line around each point (weighting each observation based on the value of the kernel at each point). This gives us an estimated slope at each point (alongside the estimated value), which can be of interest. (More rigorous mathematical discussion here.

10.5 Dummy variables and their uses (e.g., controlling for group fixed effects).

In general, categorical variables that take the form of indicator variables (1 if true, 0 otherwise) are called *dummy variables*. It can be worth including dummy variables that capture, for example, the location of an applicant or the number of applications an individual sent, grouping those from the same location/same number of applications together. The idea is that such information may contain variability that is not of interest to you but could affect the parameters of your features of interest.

For example, say you're interested in the effect of (only) SAT score on future earnings. Perhaps people with higher SAT scores send more applications, or people with higher SAT scores went to better schools. You could capture these into separate groups so that the SAT_i parameter only captures the effects of SAT score, and not any subsequent or surrounding effects. (I suspect this will greatly diminish the magnitude of the effect, and more of it would be captured by $College_i$ – but perhaps not!)

This notion of creating dummy variables to capture the effects of being in a certain group can be called *controlling for group fixed effects*.

10.6 Regression Discontinuity Design

One form of nonlinear transformation of X_i could be partitioning the range of X_i and considering in which partition the realized value is located: $S_{(k_0,k_1);i} = \text{IsFeatureWithinInterval}_i = S_{0,i}$, etc.

This can be used in circumstances where an intervention's effect has a discontinuity/"jump" based on the value of the *running variable*, a. (When would such a discontinuity exist? e.g., "Candidate only under consideration if they achieve at least $P = 70$ points on the exam." There is discontinuity of effect between $P = 69$ and $P = 70$.)

Under such circumstances, a *regression discontinuity (RD) design* to evaluate causal effects (according to one's model of the world) is of interest. As an example: Is there an increase to all-cause mortality that can be attributed to individuals reaching legal drinking age? Construct the equation:

$$Y_i = \beta_0 + \beta_1 D_{ai} + \beta_2 a_i + \epsilon_i$$

Y_i represents all-cause mortality, a_i is an individual's age, and D_{ai} represents whether they are of legal drinking age. Note we include both a_i and D_{ai}; in this way, the coefficient of D_{ai} captures the *excess change* of all-cause mortality for reaching legal drinking age, while a_i captures the (continuous) effect of increasing age on all-cause mortality.

Note that an RD design might look statistically significant, but you may in fact just be missing the correct feature (perhaps a nonlinear transformation of obtained data) in your model. Always inspect your data!

10.7 Omitted Variable Bias

An unfortunate fact of life is that very often (if not always), we will not have all the relevant features that causally affect Y_i. Say the true model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

But we don't have access to X_{2i} or any proxy for it, so we have to fit our model without it:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + w_i$$

If we *did* have both X_{1i} and X_{2i}, we could see the relationship between the two by running an *ancillary (or auxiliary) regression*:

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \xi_i$$

Intuitively, we can expect that if X_{2i} has an effect on Y_i (i.e. $\beta_2 \neq 0$, its "effect" will have to be "transferred" over to some other part of the partial model (because we're still forming an unbiased model of Y_i). If it's uncorrelated with all the other variables, it will transfer into the error term, making $w_i := \beta_2 X_{2i} + \epsilon_i$. But if there *is* a correlation between features, then that sneaks into the coefficients of the partial model. For example, say $\delta_1 > 0$, i.e., X_{1i} and X_{2i} are positively correlated. Then in our partial model without X_{2i}, the model can "make up" for the missing variable by increasing $|\hat{\alpha}_1|$. More specifically, we could break apart the partial model's parameter as:

$$\begin{aligned}\hat{\alpha}_1 &= \text{part actually due to } X_1 + \text{part that } X_{2i} \text{ snuck in because they're correlated} \\ &= \beta_1 + \delta_1 \beta_2\end{aligned}$$

So there's an *omitted variable bias (OVB)* due to the missing X_{2i}, specifically:

$$OVB = \hat{\alpha}_1 - \beta_1 = \delta_1 \beta_2$$

You can derive this intuition more rigorously by using the expression for $\hat{\alpha}_1$, a coefficient for a linear model:

$$\begin{aligned}\hat{\alpha}_1 &= \frac{Cov(Y_i, X_{1i})}{Var(X_{1i})} \\ &= \frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, X_{1i})}{Var(X_{1i})} \quad \text{(plug in full/true model of } Y_i\text{)} \\ &= \cdots \text{(break apart into separate covariance terms, evaluate, etc.)}\end{aligned}$$

If dealing with vectors, a similar analysis can be performed with $\hat{\alpha}_1 = (X_1^T X)^{-1} X_1^T Y$.

What is all this useful for? (Presumably we don't *have* the omitted variables to place in.) This guides our thinking and makes sure we frame our analysis properly. We should ask ourselves:

1. Are we missing features X_{2i} that are relevant to our target variable Y_i?
2. Would this feature have a strong effect on Y_i? (β_2).
3. Would this missing feature likely be strongly correlated with *a feature whose effect we are interested in*? (δ_1; if it's only correlated with features we have added in to avoid bias in our estimates, then it shouldn't be a huge problem)
4. If there is a correlation with both outcome variable and feature of interest, which way does it bias the parameter associated with our feature of interest?

11 Machine Learning and Econometrics

In econometrics, we focus on *estimation*. With machine learning, we tend to focus on *prediction*. What's the difference? In econometrics, we make fairly strict assumptions about how the data is generated ($Y_i = X_i^T \beta + \epsilon_i$, for example), which limits our search space to a specific family of functions, for which we can find the "best" coefficients $\hat{\beta}$. With machine learning, we provide very loose specifications or provide free reign on functional form, which diminishes interpretability a good deal but gives the flexibility to be able to make the "best" predictions $\hat{y}$.

Aside: This is reminiscent of M-estimation with the appropriate choice of ρ and Q (except M-estimation is still *estimation*, with a theoretically derived estimator and no "train-test split"). If you felt comfortable assuming a probability law for Y and assuming the necessary conditions, you could make asymptotic Normality guarantees for $\hat{y}$ based on the choice of ρ/Q.

With completely free reign, you can choose an f that just perfectly memorizes the input data. But intuitively, if you "overlearned" your data, you probably learned some random idiosyncrasies of that sample rather than "the world as a whole". How do you handle this? Impose regularization conditioning: $\min_{f \in \mathcal{F}} \sum_i Loss(y_i, f(x_i)) + \lambda R(f)$. (Ridge regression: $R(f(\theta)) = \|\theta\|_2^2$. LASSO regression: $R(f(\theta)) = \|\theta\|_1^1$.) Also, have a holdout set to assist in parameter turning, and a holdout set only used at the very end to get a sense for how the finished model will perform "in the real world".

All of this might seem ad-hoc. Is any of this legitimate? Remember, our current focus is on *prediction*, not *estimation*. We're not concerned with whether the model reflects how the world truly generated the data, which is in the end never directly observable (estimation; the benefit is attributing importance to different factors). Rather, we only care about the quality of the model's predictions outside our sample data, which we can directly observe – after all, we can leave aside data to be "out of sample" and see how it does. If we were to draw graphical models, causal models would point features to *latent variables* that then generate the *observed outcomes*, while predictive models would point all the features directly to the observed outcomes.

12 Data Visualization for others

You may want to visualize data for yourself ("exploratory data analysis"). However, there is a different set of considerations when you are visualizing data for others. Such visualizations must be clear, uncluttered, and convey important information. (By necessity, you can't put all of the information into a single visual – but you shouldn't be lying by omission!)

A starting point for considering one's visualizations is *Tufte's principles*, which emphasizes minimalism (in short, "as little ink as is needed to clearly convey your message"). (That may be overboard, but it's better to start small and build up vs. including everything than paring down.)

For tables, the same general considerations apply: include only what's necessary and be clear. In R, the package *stargazer* is a good starting point.

13 Endogeneity and Instrumental Variables

Let's say we have our linear model:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$$

Endogeneity refers to when the error term is correlated with a feature – in this bivariate case, $Cov(X_{1i}, \epsilon_i) \neq 0$. This is bad; it muddies the waters. We would want $\hat{\alpha_1} \approx \frac{\delta_{X_{1i}} y}{\delta X_{1i}} = \alpha_1$, but if there's a correlation, $\hat{\alpha_1} \approx \frac{\delta_{X_{1i}} y + \delta_{\epsilon_i} y}{\delta X_{1i}} \neq \alpha_1$ then you'll have Endogeneity can be caused from a number of methods:

1. Omitted Variable Bias. (Discussed earlier – but we never gave an answer as to how to deal with it. We'll get there now.)
2. Measurement error in the features. Since we assume the measurements are precise, any measurement error in X_{1i} would go into ϵ_i, making the two correlated.
3. Reverse causality. If, rather than X_{1i} causing Y_i, Y_i in fact causes X_{1i}, then our whole model is incorrectly specified. (In a regular model, $Cov(Y_i, \epsilon_i) = Var(\epsilon_i)$. If X_{1i} is what should actually be the target variable, it would make sense for $Cov(X_{1i}, \epsilon_i) \propto Var(\epsilon_i) \neq 0$.)

This is a bit of a mess for our estimation methods – our estimates $\hat{\beta}$ will be both biased and inconsistent. How can we deal with it? Using *instrumental variables (IVs)*.

Say X_{1i} is an endogenous variable in our linear model. An instrumental variable Z_i for X_{1i} satisfies the following properties:

1. $Cov(X_{1i}, Z_i) \neq 0$. (There is some relationship between the instrumental variable and the variable it's intended to proxy for.)
2. $Cov(\epsilon_i, Z_i) = 0$. (The instrumental variable is otherwise unconnected with the other unobserved features that had been causing the endogeneity with X_{1i}.)
3. Z_i is not a causal determiner of Y_i. (Z_i isn't actually just a feature that should have been in your model in the first place. Its effects on Y_i will only be through its ability to proxy X_{1i}.)

The estimators for the instrumental variable $\hat{beta}_{IV} = \frac{Cov(Y_i, Z_i)}{Cov(X_{1i}, Z_i)}$ is biased but consistent!

Before we go further: an IV kind of sounds magical. If Z_i is correlated with X_{1i}, won't it *have* to also be correlated with ϵ_i (since X_{1i} is correlated with ϵ_i)? That is to say, wouldn't Z_i have the exact same problem?

Let's consider an example: We want to see the effect of education on log-wages.

$$ln(wage)_i = \alpha + \beta educ_i + \epsilon_i$$

One could expect that many other factors affect earnings as well as years of education, for example some measure of innate ability:

$$ln(wage)_i = \alpha + \beta educ_i + \gamma abil_i + \epsilon_i$$

If we assume that innate ability correlates with education and with earnings, we would have omitted variable bias and therefore endogeneity in $educ_i$. We can't reasonably get measures of everyone's innate ability, so what do we do? How do we get an IV ("proxy") for education that isn't correlated with innate ability or anything else?

As it turns out, if someone is born in the fourth quarter of the year, they tend to join school at around 5 3/4 years old, whereas people born in the first quarter start around the age of 6 3/4 (due to school regulations). Similarly, one couldn't end their school (to e.g. join an apprenticeship) until they were 16 years old. So on average, people born in the fourth quarter were "forced" into more schooling than those in the first quarter. At the same time, one can reasonably assume that the quarter of the year in which you're born is not correlated with your innate ability, etc. In this way, $Is4thQuarter_i$ can serve as an instrumental variable for $educ_i$.

How can we deal with this in general? One can perform *two-stage least-squares*. Say you have a model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_I$, and X_{1i} is our suspected endogenous variable for which we have gotten instrumental variables Z_1 and Z_2. Then:

1. First Stage: $X_{1i} = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_{-2} X_{2i} + w_i$. Get estimators for π, $\hat{X}_{1i}$. We include the other exogenous variables to keep any correlations with the other features. Remember: we're trying to make a proxy for X_{1i} that is uncorrelated with ϵ_i – if X_{1i} is correlated with X_{2i}, we should try to keep that behavior in our estimator. (The more we deviate our proxy from its target, the less good a proxy it becomes.)
2. Second Stage: $Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \beta_2 X_{2i} + \epsilon_i$. This time, our estimator $\hat{\beta_1}$ is consistent (but still biased).

If our instrumental variable is a binary/indicator variable, we can estimate our $\hat{\beta}_{IV}$ quite simply with the *Wald estimate*:

$$\hat{\beta}_{Wald} = \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid z_i = 0}{E[A_i \mid Z_i = 1] - E[A_i \mid Z_i = 0}$$

In general, we can write:

$$\hat{\beta} = (Z^T X)^{-1} Z^T Y$$

(If there are more instruments than endogenous variables, you can write something similar, just with some projection matrices.)

If an instrument is meant to compel subjects to get a treatment, the estimator captures the effect of treatment on those who are in fact compelled to get treatment because of the instrument. This is called a *local average treatment effect (LATE).* (Note that this suggests that this does not describe the population as a whole, only people that would be swayed by the instrument.)

14 Experimental design.

There are a few important things to keep in mind:

1. Be clear about what question exactly you're trying to answer.
2. Try to introduce randomness in your experiments (even if it may not seem feasible at first).

14.1 Randomization methods

Ideally, you can *stratify* your experimental design. Maybe you can do simple random assignment. You may be forced to *cluster*, but try to avoid it if possible.

What about cases where "ideal" randomization is not ideal? You can perform a *randomized phase-in* of subgroups in your sample over time; those who haven't phased in can be your point of comparison. You can *randomize around a cutoff* to see how important that cutoff actually is. You can set up an *encouragement design* and estimate the local average treatment effect (LATE) of the intervention (rather than force people to join or be excluded from an intervention).

14.2 Clarify your design.

Try to drill down on not only whether an intervention works, but *what aspects of the intervention causes a change* (for example, is it the fact that some information is made openly public? That you have a physical reminder? That there is an appearance of accountability?). Be aware of potential unintended consequences of your interventions, and attempt to estimate them (e.g. "crowding-out"/displacement effects).