Stat 305A: Linear Models **Autumn 2019/20**

Lecture 14: Experiments, A/B tests and ANCOVA

Lecturer: Art B. Owen *November 12*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students who took stat 305A at Stanford University. They appear in* `http://statweb.stanford.edu/~owen/courses/305a/` *where you may find the most recent version. Stanford University holds the copyright.*

14.1 Brief remarks on outliers

The lecture began with some very brief comments on outliers from the prior lecture:

- when presenting a data analysis, document which outliers were removed and why they were removed. Sometimes you have a perfect explanation. Other times, one might not be available.
- Outlier deletion is potentially a 'researcher degree of freedom'. Some people might select the outliers to delete based on how the model ends up. This creates a catch-22 for the analyst. Leaving outliers in spoils the model, taking them out raises doubts. Choosing a robust method ahead of the analysis would reduce the doubt. Choosing it partway through is different.
- sometimes the outlier is the one good and informative point and the other $n-1$ have less value. For instance, if only one person recovers from a severe infection, there could be more scientific value in that case than in all the others.

14.2 Causal inference

We know that correlation is not causation. There is more to the story. The field of causal inference is growing recently and there are many good things coming out now from right here at Stanford. This lecture and the next describe some of them that involve linear regression.

Sometimes we can get a causal conclusion out of observational data. We do that by merging the observational data with some causal assumption. Examples to follow. The causal assumption can often be questioned.

14.3 A/B testing

Suppose that we have two treatment options, call them A and B. If we select people at random to give A and the rest get B, then any difference we see between those two groups can be attributed to either the treatment then got, or plain luck from the randomization. As the data set grows larger the probability that randomization would explain the difference can be made to diminish, leaving only A vs B.

If some future response Y correlates with A vs B, it cannot be that the response was causing A/B. If some other variable Z was going to cause both Y and A/B we would have to ask how it could have affected the coin toss or other randomization we used. That seems like an unreasonable doubt to hold.

© Stanford University 2019

There is a long history of randomized controlled trials like this. There was an early randomization of which ships in the British navy should get citrus. That showed conclusively that citrus reduced scurvy. We know that vitamin C is the cause; that explanation might not have been known, but the result was.

We considered the Lanarkshire milk experiment where in some schools the students who were thought to need milk the most got it. That muddles any causal conclusion that might have otherwise been attainable.

In medicine, using historical controls for a new treatment is usually considered questionable. The old results might have had a different standard of care or a different patient population or both. Then the differences could as well be due to those factors instead of the changed treatment.

In medicine it is best if the treatment is double blind. If a patient knows that they're getting placebo vs active drug, they might stop taking it, so the allocation should be blinded. If the medical staff know which treatment somebody is getting they might treat them differently. If they also don't know, the study is double blind.

If the study is not blinded then we could learn that there is a causal difference between being assigned A vs B but maybe it is due to how the patients reacted or the staff and that is not what we were hoping to measure.

A common issue there is that compliance (ie taking the medicine) could be different in both groups. The experiment measures the causal impact of prescribing which is not the same as taking. It is called an 'intent to treat' analysis. The FDA often insists on intent to treat analyses because going forward, what will happen is presecription followed by some less than predictable compliance. Ethical considerations are important in this context. Experiments are often stopped early if it becomes clear which of A or B is best.

In electronic commerce, A/B tests are very commonly used to tune websites. Those can involve very large sample sizes aimed at very small effects that are nonetheless valuable on super large scale problems.

14.4 Randomization analysis (or not)

We looked earlier at permutation tests that make confidence intervals and p-values using the randomizations.

If you have randomized the allocation of A vs B, then you have made data for which the plain old linear model analysis safer. This is the approach taken in the book "Statistics for experimenters" by Box, Hunter and Hunter. The random allocation removes the worry about your treatment being highly correlated with some important missing value. The extreme version of that correlation is ***confounding***. Suppose that all of the patients getting A were under 40 and all those getting B were over 40. Then any difference could as easily be due to age as A/B. They are confounded.

You could also balance the experiment giving each age group half A and half B treatment. That becomes cumbersome when you have lots of variables to split on. Splitting on the important ones and randomizing on the rest is a reasonable hybrid.

Randomization also protects you somewhat against correlated errors as might happen if your data arrive steadily over time. Let $X_i \in \{-1, 1\}$ for $i = 1, \dots, n$ represent A vs B. If

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and

$$\operatorname{cor}(\varepsilon_i, \varepsilon_j) = \begin{cases} 1, & i = j \\ \rho, & |i - j| = 1 \\ 0, & \text{else} \end{cases}$$

then if half the X_i are 1s and half -1s we can find that

$$\hat{\beta}_1 = \frac{1}{n}\sum_i X_i Y_i = \frac{1}{n} X^\mathsf{T} Y$$

so

$$\mathrm{var}(\hat{\beta}_1) = \frac{1}{n^2} X^\mathsf{T} \mathrm{cov}(Y) X = \frac{1}{n^2}\Big(n\sigma^2 + 2(n-1)\rho \sum_{i=1}^{n-1} X_i X_{i+1}\Big).$$

The second term in brackets gets lots of cancellations (on average) if you randomized the X_i.

The randomization analysis gets complicated when there are many things to adjust for or not just binary Xs and so on. There is a forthcoming book by D. Rubin and T. Dasgupta about how to do randomization for complicated problems.

14.5 Neyman-Rubin causal model

In this model, which requires us to think about new notation, there are subjects $u \in \mathcal{U}$. Let's suppose that $\mathcal{U}$ is a finite set. For each $u \in \mathcal{U}$ there is a value $Y_t(u)$ that would be observed if u got 'treatment' and a value $Y_c(u)$ that would be observed if u got 'control'.

The treatment effect for u is $Y_t(u) - Y_c(u)$. We cannot measure it because we will not give subject u both treatment and control. Instead there is a variable $W(u) \in \{c, t\}$. We observe

$$\begin{aligned} Y(u) &= \begin{cases} Y_t(u), & W(u) = t, \\ Y_c(u), & W(u) = c \end{cases} \\ &= Y_{W(u)}(u). \end{aligned}$$

The observed vs unobserved pattern is depicted below.

u	$Y_t(u)$	$Y_c(u)$
1	✓	×
2	✓	×
3	×	✓
⋮	⋮	⋮
N	×	✓
Avg	$\bar{Y}_t$	$\bar{Y}_c$

This is sometimes called the ***science table***. If we had that, we'd know all about the causal impact. In a more complicated problem there are more different kinds of treatments and treatment combinations and a bigger science table.

We might want the average treatment difference over subjects. The average causal effect is

$$\frac{1}{N}\sum_u Y_t(u) - Y_c(u)$$

but we don't get to see any of the values in that average. This is the fundamental problem of causal inference. We didn't do it both ways.

Let $\mathcal{W} = (W(1), W(2), \ldots, W(N))$ denote all of the random treatments we made. We want

$$\mathbb{E}(Y_t - Y_c) = \mathbb{E}(Y_t) - \mathbb{E}(Y_c)$$

with expectation averaging over $u \in \mathcal{U}$. We can get

$$\mathbb{E}(Y_W \,|\, W = t) - \mathbb{E}(Y_W \,|\, W = c) = \mathbb{E}(Y_t \,|\, W = t) - \mathbb{E}(Y_c \,|\, W = c).$$

If we make $\mathcal{W}$ independent of $\{(Y_t(u), Y_c(u), \mid u \in \mathcal{U}\}$ then the expectations we estimate are the ones that we want. Notice that $W(u)$ for subject u has to be independent of everybody else's (Y_t, Y_c) values, not just their own.

We could toss a coin for each subject u. Usually a fair coin but it does not have to be. Or we could decide that T of the N will get treatment and choose so that every subset of T subjects has probability $1/\binom{N}{T}$ of being selected.

14.6 Internal vs external validity

A randomized analysis is very convincing about the causal effect of a treatment for the set of subjects included in that experiment. That is, it has high ***internal validity***. If we want to extend it to subjects ouside that experiment then we are concerned about ***external validity***. The subjects included in the study might differ in place or time or health conditions from the ones we might want to apply the findings too. If the included subjects are a random sample from the ones we want to apply conclusions too, then we are in a stronger position to generalize.

14.7 Analysis of covariance

The analysis of covariance (ANCOVA) model is one where Y is predicted by an ANOVA plus some other elements of a linear model. For causal inference issues, the ANOVA could be at two levels. (Or more, but two is the minimum where we see how the ideas work out.) The sum of two linear models is a linear model.

Suppose that we use $W \in \{-1, 1\}$ as our treatment variable and assign it at random. We could write

$$Y_i = \alpha + \tau W + X_i^{\mathsf{T}}\beta + \varepsilon_i$$

where τ is commonly used for the treatment effect we care about.

If the variables in X_i are ones that we measure before allocating the random treatment, then we are looking at the effect of W on $Y - X^{\mathsf{T}}\beta$ instead of on Y. If $X^{\mathsf{T}}\beta$ describes something with high explanatory power for Y, then by fitting that model we get a better measure of the effect of W. The $X_i^{\mathsf{T}}\beta$ values will be random between the two treatment groups. One of them may have been 'lucky'. A good regression reduces that effect and lets us discern a possibly small incremental value of W after adjusting for potentially large effects of other variables.

In a very common form of ANCOVA, we have just one component in X_i and it is the 'before' version of Y_i. In class we looked at measures of gingivitis made both before and after two treatments were applied. Then

$$Y_{i,post} = \alpha + \tau W_i + Y_{i,pre}\beta + \varepsilon_i.$$

As above suppose that half the W_i are 1 and half are -1. We can do a regression to see if W_i is significant when added to the model after $Y_{i,pre}$. For $\beta = 0$, the model looks at whether $Y_{i,post}$ is related to W_i. Do

people end up better in treatment vs control? For $\beta = 1$, the model looks at whether the change $Y_{i,post} - Y_{i,pre}$ is related to W_i. Do people improve more under treatment vs control?

The least squares β could well be between these two cases and it then gives a more efficient estimate of the treatment effect.