LECTURE 25b
Analysis of covariance (ANCOVA) (Chapter 22)

ANCOVA is a very general term for the use of covariates in the analysis of experimental data. The computations required for ANCOVA are simple. In most cases, we just include the covariate(s) on the right-hand side of the linear model and apply ordinary least squares and partial F-tests. For students who understand regression, the mechanical aspects of ANCOVA are trivial. But the motivation for ANCOVA and its interpretation are often poorly understood. Therefore, much of our discussion of ANCOVA will be devoted to understanding what the model means and what it assumes.

ANCOVA model for one-way design. In the usual one-way ANOVA for comparing k groups in a randomized experiment, we apply the linear model

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu_\cdot + \alpha_i + \epsilon_{ij} \end{aligned} \tag{1}$$

where

$$\begin{aligned} \mu_i &= \text{mean for group } i \ (i = 1, \ldots, k) \\ \mu_\cdot &= \frac{1}{k}\sum_{i=1}^{k} \mu_k \\ \alpha_i &= \mu_i - \mu_\cdot \end{aligned}$$

and $\epsilon_{ij} \sim N(0, \sigma^2)$. The within-group sample sizes $n_1, \ldots, n_k$ do not have to be equal, and the total sample size is $N = \sum_i n_i$.

Now suppose that, in addition to the response variable Y_{ij}, we also observe another variable X_{ij} for each unit which

- was not used in the randomization (i.e., was not taken into consideration when assigning units to treatment groups),
- is possibly correlated with the response, and
- is causally unaffected by the treatment.

This variable X_{ij} is often called a *covariate* or a *concomitant variable*.

Most presentations of ANCOVA suppose that X_{ij} is a single continuous variable, but that is not necessary. One can do ANCOVA with multiple covariates, and the covariates may be binary or categorical. X_{ij} may be a vector of covariates including dummy codes, effect codes, etc. This discussion will proceed as if X_{ij} is one-dimensional, but the extension to vector-valued covariates is straughtforward; we simply change every occurrence of the coefficient β in the expressions below by β^T, where β is then a vector.

The classical ANCOVA model simply includes the covariate as another predictor, so that the model becomes

$$Y_{ij} = \mu_\cdot + \alpha_i + \beta\, X_{ij} + \epsilon_{ij} \tag{2}$$

It is quite common to center the covariate at its sample mean, $X_{\cdot\cdot} = N^{-1}\sum_i \sum_j Y_{ij}$. In that case, the model becomes

$$Y_{ij} = \mu_\cdot + \alpha_i + \beta\,(X_{ij} - \bar{X}_{\cdot\cdot}) + \epsilon_{ij} \tag{3}$$

These two models give the same fit, but the intercepts have a different meaning. In (2), we interpret $(\mu_\cdot + \alpha_i)$ as the expected value of the response variable under treatment i when the covariate is held constant at zero. In (3), we interpret $(\mu_\cdot + \alpha_i)$ as the expected response under treatment i when the covariate is held constant at $\bar{X}_{\cdot\cdot}$. Either way, it is obvious that these models say that (a) within each treatment, there is a linear relationship between the response and covariate, (b) these lines are parallel with a common slope β, and (c) the vertical distance between the lines for treatments i and i' is $\alpha_i - \alpha_{i'}$.

Example. For example, consider a study to compare k different medications for controlling blood pressure. The response Y_{ij} could be the blood pressure of subject i in group j at the end of the study, and X_{ij} could be the subject's blood pressure at the beginning of the study. A pre-treatment (baseline) measure of the outcome variable is an excellent covariate, because it is likely to be highly correlated with the response. But it doesn't have to be the same variable as the response. It could be age, body mass index (hopefully measured prior to the treatment), or any other variable that is not causally affected by the treatment.

Note on difference scores. If the covariate is a baseline measurement of the outcome variable, then researchers will sometimes redefine the outcome to be the difference between the post-treatment measurement and the pre-treatment measurement. By doing this, they are effectively setting the slope to $\beta = 1$ and moving the term βX_{ij} to the left-hand side of the equation, so that the model becomes a one-way ANOVA with $Y_{ij} - X_{ij}$ as the response:

$$(Y_{ij} - X_{ij}) = \mu_{\cdot} + \alpha_i + \epsilon_{ij}$$

If time permits, we may discuss the relative merits of using difference scores versus traditional ANCOVA when we cover repeated measures designs.

Difference scores have a nice interpretation as a gain or loss, but the assumption that $\beta = 1$ might be worrisome. If we are unsure of that assumption, we can always fit the more general ANCOVA model (2) or (3) and test the null hypothesis $H_0 : \beta = 1$. If we like the interpretation of using using the difference score $Y_{ij} - X_{ij}$ as the outcome but we don't want to assume that $\beta = 1$, then we can expand the difference-score model by including the covariate, like this

$$Y_{ij} - X_{ij} = \mu_{\cdot} + \alpha_i + \beta X_{ij} + \epsilon_{ij} \tag{4}$$

or like this:

$$Y_{ij} - X_{ij} = \mu_{\cdot} + \alpha_i + \beta (X_{ij} - \bar{X}_{\cdot\cdot}) + \epsilon_{ij} \tag{5}$$

Note, however, that models (4) and (5) give exactly the same fit as models (2) and (3), and putting $-X_{ij}$ on the left-hand side of the equation merely changes the value of β by the constant amount -1.

Understanding the meaning of ANCOVA. Presentations of ANCOVA in some textbooks are difficult to understand because they fail to make a distinction between the parameters of ANOVA and ANCOVA. For purposes of this discussion, let's assume that

- the treatments were randomly assigned to units, and
- the covariate X_{ij} was not causally affected by the treatment in any way.

The latter condition is usually satisfied if X_{ij} is realized and measured before the treatments are applied. Now consider the ANOVA and centered ANCOVA models:

$$\begin{aligned} \text{ANOVA:} \quad Y_{ij} &= \mu_{\cdot} + \alpha_i + \epsilon_{ij} \\ \text{ANCOVA:} \quad Y_{ij} &= \mu_{\cdot} + \alpha_i + \beta (X_{ij} - \bar{X}_{\cdot\cdot}) + \epsilon_{ij} \end{aligned}$$

The appearance of the same symbols $\mu_{\cdot}$ and α_i in both models is potentially confusing because, unless $\beta = 0$, they do not mean the same thing. The ANOVA model describes the mean of the response, but the ANCOVA model describes the conditional mean of the response at fixed values of the covariate.

- In the ANOVA model, $\mu_{\cdot} + \alpha_i$ is the average response under treatment i.
- In the ANCOVA model, $\mu_{\cdot} + \alpha_i$ is the average response under treatment i when the covariate is held fixed at $\bar{X}_{\cdot\cdot}$.

If the goal of the experiment is to compare the effectiveness of the treatments (and it usually is), then *we do not need the covariate to achieve that goal.* The ANOVA model will always yield unbiased estimates of the treatment effects, where "uniased" means over repeated runs of the randomized experiment (i.e., the frequentist sense). The fact that the covariate X_{ij} may be related to the response doesn't change that fact. The randomization ensures

that we get unbiased estimates of treatment effects whether the covariate is measured or not. The mean response under treatment i is $\mu_. + \alpha_i$ from the ANOVA model, not the ANCOVA model. And the average effect of treatment i versus treatment I' is $\alpha_i - \alpha_{i'}$ from the ANOVA model.

So why would we ever use ANCOVA? The parameter $\mu_. + \alpha_i$ in the ANCOVA model is the mean response under treatment i among units having $X_{ij} = \bar{X}_{..}$. This is not the full population of interest, but a very special and select group of units. Now consider the ANCOVA parameter $\alpha_i - \alpha_{i'}$, which us the difference between the mean response under treatment i and the mean response under treatment i' among units having $X_{ij} = \bar{X}_{..}$. Now here comes the interesting part.

1. If the ANCOVA model is actually true — that is, if the relationship between the response and the covariate within each treatment group is linear with the same slope — then the lines are parallel. Parallelism means that we can also interpret $\alpha_i - \alpha_{i'}$ as the difference between the mean response under treatment i and the mean response under treatment i' among units having $X_{ij} = x$, where x is any fixed value of the covariate. In other words, $\alpha_i - \alpha_{i'}$ is the difference in mean response under treatments i and i' *among units with the same values for the covariate.*

2. Now if the experiment was truly randomized, and if the covariate is not causally affected by the treatment, then, on average (again, over repeated runs of the experiment), the values of the covariate under treatments i and i' will be exactly the same. Because of randomization, the populations of units receiving the two treatments are the same population, so the distributions of the covariate within the treatment groups are, on average (again, over repeated runs of the experiment), exactly the same.

3. As a consequence of 1 and 2, we can also interpret $\alpha_i - \alpha_{i'}$ from the ANCOVA model as *the difference in mean response under treatments i and i' in the entire population, regardless of the covariate.* This follows from application of the well known principle $E(Y) = E(E(Y \mid X))$. *Therefore, the treatment effect $\alpha_i - \alpha_{i'}$, or any other contrast among the α_i's, has the same meaning in the ANCOVA model as it does in the ANOVA model, and is a genuine treatment effect in the population.*

Importance of the modeling assumptions. Note that the truthfulness of the italicized statements in point 3 depends on the ANCOVA model being correct. That is, it assumes that the regression of the response on the covariate in each group is linear with the same slope. And it assumes that the population distribution of the covariate under each treatment is the same (which is satisfied if the treatment was assigned at random after the covariate was realized.)

What if the ANCOVA model does not hold? For example, what if the regression lines are not parallel? If they are not, then we can generalize the ANCOVA model by fitting different slopes $\beta_1, \ldots, \beta_k$ to the treatment groups:

$$Y_{ij} = \mu_. + \alpha_i + \beta_i (X_{ij} - \bar{X}_{..}) + \epsilon_{ij} \tag{6}$$

And we will be able to test the hypothesis of parallelism by comparing the fit of this model to that of the common-slopes model.

Now if we use different slopes, then we are acknowledging that the treatment effects in subpopulations with different fixed values of the covariate will be different. That is, we acknowledge that the conditional treatment effects given X depend on X. But what about the treatment effects in the overall population regardless of X? In that case, if we have centered the covariate X_{ij} at $\bar{X}_{..}$, then we can interpret $\alpha_i - \alpha_{i'}$ as the average effect of treatment i versus i' in a population for which the mean value of the covariate is $\bar{X}_{..}$. (Once again, this is the principle $E(Y) = E(E(Y \mid X))$ at work). And since $\bar{X}_{..}$ is an unbiased estimate of the population mean of X, we can once again interpret the estimate of $\alpha_i - \alpha_{i'}$ as an estimate of the average effect of treatment i versus i' in the full population. For this interpretation to hold, it is now essential that we center the covariate at $\bar{X}_{..}$. *With parallel slopes, we do not have to center the covariate. But with non-parallel slopes, we must center the covariate, because if we do not then the treatment effects will not be defined correctly.*

With real data, the assumptions of linearity and parallel slopes will never exactly hold. At best, they are only an approximation. If the slopes truly vary across the groups but we only estimate a common slope, then our estimates of the treatment effects will be biased. In practice, however, these biases tend to be small if the experiment is

balanced or nearly balanced. That is, if the sample sizes $n_1, \ldots, n_k$ within the groups are not drastically different, violations of the common-slopes assumption do not have a dramatic impact on the performance of the method, as long as the data come from a randomized experiment.

So why do we need ANCOVA anyway? In these discussions, we have argued that, as long as the assumptions of ANCOVA hold, then ANOVA and ANCOVA are estimating the same treatment effects. Now if we can estimate the same thing either way, and if ANCOVA is only introducing extra assumptions that may or may not be valid, why would anyone choose to use the more complicated ANCOVA instead of the simpler ANOVA?

The main reason for using ANCOVA is efficiency. The estimated treatment effects from ANCOVA can be substantially more precise. The variance of the estimated treatment effects in the two methods are proportional to their respective σ^2's. If the correlation between X_{ij} and Y_{ij} is substantial, then including X_{ij} in the regression can dramatically decrease the MSE, which greatly increases the power of treatment comparisons. If the covariate X_{ij} is completely unrelated to Y_{ij} (i.e., if the true β is zero), then including X_{ij} is unnecessary. But if we include it anyway, we do not change the expected MSE, and we have only lost one degree of freedom in the estimation of σ^2 which usually has very little impact on the power. The *relative efficiency* of the procedures can be estimated by examining the ratio of the MSE's with and without the covariate:

$$\text{relative efficiency} = \frac{\text{MSE from ANOVA}}{\text{MSE from ANCOVA}}$$

We can interpret the relative efficiency as a ratio of sample sizes needed to achieve similar power by the two methods. For example, if the relative efficiency is 1.8, performing ANCOVA instead of ANOVA produces an increase in power that is roughly equivalent to increasing the sample size by 80%.

Another benefit of using ANCOVA is that *it helps to protect us from "bad" outcomes in the randomization.* Suppose we are doing an experiment to compare the effectiveness of two drugs for reducing blood pressure. Suppose we randomly assign half of the patients to receive Drug 1 and the other half to receive Drug 2. Before starting the treatment, we measure the body-mass index of each patient, and we find that by chance we have assigned most of the patients who are obese (having high BMI) to receive Drug 1. And after the experiment is completed, we find that the average final blood pressure in group 1 is higher than group 2. Does that mean that Drug 2 is more effective than Drug 1? Is the difference seen in the outcomes really due to the treatment, or is it simply a reflection of the fact that by bad luck most of the obese patients happened to fall in group 1? If we ignore the covariate and proceed with ANOVA, then the estimated treatment effect will be unbiased *over repeated experiments.* But that is little consolation, because in fact we are not running repeated experiments; we are only doing it this one time. In this case, it is highly recommended that we use ANCOVA, because it adjusts the estimated treatment effects to account for the differences ar baseline. In fact, if we peek at the data and see a big difference at baseline in an important covariate but do not adjust for it using ANCOVA, then we are (arguably) being naive or dishonest, because we know that some of the difference in the outcomes that we see between the groups should be attributed to the difference at baseline.

Covariance adjustment versus "blocking." In the previous example, we could have avoided the bad randomization outcome by changing the design and blocking on obesity. That is, we could have measured BMI ahead of time and randomly assigned treatments in such a way that the proportions of obese patients were identical in the two drug groups. In that case, obesity would be considered a "blocking factor," and we would proceed with a 2×2 factorial analysis (obesity crossed with drug). Blocking guarantees that the treatment groups are balanced with respect to the blocking variable, and if the blocking variable is related to the outcome, it will reduce the variance in the estimated treatment effect in much the same way that ANCOVA does.

If you are designing a study and you know of a baseline variable that is closely related to the outcome, you have a choice: You can either use it as a blocking variable in the randomization, or you can adjust for the variable later using ANCOVA. Which is better? As a general rule, many statisticians would say that blocking is preferable to covariate adjustment, because it makes fewer assumptions. They would say, "Block on the variables that you believe are most important, and use covariance adjustment for the rest." Another consideration is the nature of the baseline variable in question. If the baseline variable is continuous, then it often seems more natural to use it as a covariate in ANCOVA rather than a blocking factor. A continuous variable can be coarsened into a few categories, but that may produce loss of information. If we are talking about a pre-test score, i.e., a baseline measure of the continuous response variable, then it is probably more common to use that variable in a covariance adjustment rather than

categorize it and treat it as a blocking variable.

Another interpretation of the ANCOVA model. We have argued that ANCOVA gives estimated treatment effects in the full population without regard for X. In a randomized experiment, the primary research questions usually pertain to average treatment effects in the population.

Suppose that we perform ANCOVA and discover that the assumption of parallel slopes does not hold. This provides evidence that the treatment effects are not constant in the population but vary in relation to the covariate. In that case, we may fit the model with k different slopes and interpret it *as a model for how the treatment effect varies by* X. This interpretation is in addition to — not in place of — the interpretation given earlier. The estimated value of $\alpha_i - \alpha_{i'}$ is still a valid estimate of the average effect of treatment i versus I' in the population. But the model can also be used to estimate the average treatment effects in subpopulations defined by X. Under the k-slopes model, the mean response under treatment i when $X = x$ is

$$\mu_{\cdot} + \alpha_i + \beta_i(x - \bar{X}_{\cdot\cdot})$$

and the mean response under treatment i' when $X = x$ is

$$\mu_{\cdot} + \alpha_{i'} + \beta_{i'}(x - \bar{X}_{\cdot\cdot})$$

so the effect of treatment i versus i' when $X = x$ is

$$(\alpha_i - \alpha_{i'}) + (\beta_i - \beta_{i'})(x - \bar{X}_{\cdot\cdot})$$

By computing and plotting the estimates of this quantity for different values of x, we can see how the treatment effect varies across these subpopulations.