Contents

8

Analysis of covariance and crossover designs

What if we have a categorical treatment variable and some continuous ones? It is easy to fold them all into one regresson. The combination is known as the analysis of covariance ***ANCOVA***. The name derives from some specialized shortcut algorithms to fit this model. Those were very useful before computers became routinely used. Now tt would take a staggeringly large data set for those algorithms to bring a meaningful savings now. Cutting 10^{-6} seconds in half is not interesting. What is more interesting and useful is the statistical thinking involved in using this method, especially in regard to measurements taken both before and after at treatment is applied.

In this chapter we also consider a related way compare before and after measurements. It is known as the cross-over design and it applies two or more treatments to the same subject, one treatment at a time.

8.1 Combining an ANOVA with continuous predictors

Suppose that we have Y_{ij} for treatments $i = 1, \dots, k$ and replicates $j = 1, \dots, n$. Now suppose that we additionally have continuously distributed covariates $\boldsymbol{x}_{ij} \in \mathbb{R}^p$. We can combine the ANOVA model with the continuous predictors by writing

$$Y_{ij} = \mu + \alpha_i + \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + \varepsilon_{ij}.$$

It seems better to center the $\boldsymbol{x}_{ij}$ and fit

$$Y_{ij} = \mu + \alpha_i + (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij}$$

where

$$\bar{\boldsymbol{x}}_{\bullet\bullet} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}\boldsymbol{x}_{ij}.$$

This way $\mu = \mathbb{E}(\bar{y}_{\bullet\bullet})$. We also still impose $\sum_i \alpha_i = 0$.

We can similarly add the $(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta$ term to any of our other models: blocked designs, Latin squares, factorials and fractional factorials. For instance, in a randomized block experiment we would fit

$$Y_{ij} = \mu + \alpha_i + b_j + (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij}$$

where b_j are block effects.

8.2 Before and after

An extremely important special case arises when for $x_{ij} \in \mathbb{R}$ is the pre-treatment version of the same quantity we measure as Y_{ij} after the treatment. Even with a random assignment of treatments to experimental units there can still be a meaningfully large amount of variance in x_{ij} between treatment groups $i = 1, \dots, k$. The randomization still means that we get valid/reliable variance estimates, confidence intervals for the treatment effect and p-values for H_0 : $\mathbb{E}(Y \mid A) = \mathbb{E}(Y \mid B)$ for treatments A and B. We can often still improve our analysis by getting reliable inferences with smaller variance by taking account of the x_{ij}.

Consider treatments to help people lose weight. The subjects might vary considerably in weight prior to the treatment. A useful weight reduction could be much smaller than the person to person differences. Those would then make it hard to properly compare the treatments. For two treatments, if we did not take account of x_{ij} and used a model like

$$Y_{ij} = \alpha + \mu_i + \varepsilon_{ij}$$

then $\mathrm{var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})$ under random assignment would be very large.

In a setting like this we might opt to analyze differences $D_{ij} = Y_{ij} - x_{ij}$ and then test $H_0 : \mathbb{E}(D \mid A) = \mathbb{E}(D \mid B)$. This can be a great improvement over comparing unadjusted post-treatment means.

In class we saw a dental example from Fleiss (1986). There x_{ij} and Y_{ij} were measurements of gingivitis (gum inflammation) prior to and following upon one of two treatments. Fleiss analyzes $D_{ij} = x_{ij} - Y_{ij}$, the opposite of the difference described above. With this choice positive values indicate better outcomes, since gingivitis is undesirable.

The post treatment data are as follows:

Trt	n	Mean	s.dev.
1	74	0.5514	0.3054
2	64	0.3927	0.1988

Fleiss gives four significant figures for most numerical values in this example. These data lead to an estimated benefit for treatment 2 of 0.1587. A t-test (pooling the variance estimates) gives $t = 3.56$ which is significant enough. [Exercise: figure out the degrees of freedom and the p-value.]

Prior to the treatment, however, the subjects started off as follows:

Trt	n	Mean	s.dev.
1	74	0.6065	0.2541
2	64	0.5578	0.2293

So group 2 started out with an advantage of 0.0487. That is roughly one third of the apparent post-treatment gain.

We could reasonably want to get a sharper estimate of the treatment benefit. In some problems it might be enough to simply be confident about which of two treatments is best. For other purposes it is important to estimate how much better that treatment is.

The table of treatment differences $x_{ij} - Y_{ij}$ is

Trt	n	Mean	s.dev.
1	74	0.0551	0.2192
2	64	0.1651	0.2235

The estimated benefit from treatment 2 is now 0.1100 with $t = 2.91$ (still significant).

Fleiss articulates two goals. One is to properly account for pre-treatment differences. Another is to reduce the variance of the treatment effect estimate.

Differences do not alway achieve the second goal. To see why, let $\rho = \mathrm{corr}(x_{ij}, Y_{ij})$ and $\sigma_y^2 = \mathrm{var}(Y_{ij})$ and $\sigma_x^2 = \mathrm{var}(x_{ij})$. Then $\mathrm{var}(x_{ij} - Y_{ij}) = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$. Now $\mathrm{var}(x_{ij} - Y_{ij}) < \mathrm{var}(Y_{ij})$ if and only if

$$\rho > \frac{1}{2}\frac{\sigma_x}{\sigma_y}.$$

If $\sigma_x \approx \sigma_y$, then we need $\rho \gtrapprox 1/2$ for differencing to improve accuracy. In the hypothetical weight loss example it seems quite plausible that ρ would be large enough to make differences more accurate than using post-treatment weight.

Now let's look at a regression model

$$Y_{ij} = \mu_i + \beta(x_{ij} - \bar{x}_{\bullet\bullet}) + \varepsilon_{ij}, \quad i = 1, 2 \quad j = 1, \ldots, n_i.$$

If we take $\beta = 0$, we get the post-treatment analysis. If we take $\beta = 1$, we get an analysis of differences (in this case future minus past). There is a value of β that would minimize the variance of $Y_{ij} - \beta(x_{ij} - \bar{x}_{\bullet,\bullet})$. If we knew that β we could test $H_0 : \mathbb{E}(Y - \beta x \mid A) = \mathbb{E}(Y - \beta x \mid B)$ which is the same as $H_0 : \mathbb{E}(Y - \beta(x - \bar{x}) \,|\, A) = \mathbb{E}(Y - \beta(x - \bar{x}) \,|\, B)$ so centering does not affect the null hypothesis we are testing. It is also the same as

$$H_0 : \mathbb{E}(Y - \mu - \beta(x - \bar{x}) \,|\, A) = \mathbb{E}(Y - \mu - \beta(x - \bar{x}) \,|\, B).$$

In this version we can think of $\mu + \beta(x - \bar{x})$ as an estimate of where the subject would be post-treatment and then the experiment is comparing the average amount that Y exceeds this baseline, between levels A and B of the treatment. We don't know β, but we can estimate it by least squares, and that's what ANCOVA does.

For the gingivitis data, the estimated treatment difference from ANCOVA was 0.1263 and the$t$ statistic was 3.57 (vs 3.56 for the original data and 2.91 for differences). The standard error of the estimate was 0.0354 vs 0.0446 (for Y) or 0.0374 (for $x - Y$). In this instance the precision was improved by regression on the prior measurement.

8.3 More general regression

The regression does not have to be on a prior version of Y. Montgomery (1997) considers a setting where Y_{ij} is the strength of a cable and x_{ij} is the diameter of the cables prior to treatment. Given multiple variables $\boldsymbol{x}_{ij} \in \mathbb{R}^p$, measured prior to treatment, we could put them all into the ANCOVA. If p becomes large then it is possible that ANCOVA would increase the variance of the adjusted responses $Y_{ij} - (\boldsymbol{x}_{ij} - \boldsymbol{x}_{\bullet\bullet})^{\mathsf{T}}\hat{\beta}$ beyond the variance of Y_{ij} and give a less precise comparison. In some cases, even the reduced degrees of freedom could widen the confidence interval on the treatment difference. Deciding the number and identity of predictors to include is a common tradeoff in statistics.

We can think of ANCOVA as adjusting for pre-treatment differences in much the same way that blocking does. The difference is that blocking is used for categorical prior variables while ANCOVA can handle continuous ones.

Suppose that

$$Y_{ij} = \mu_i + g(\boldsymbol{x}_{ij}) + \varepsilon_{ij}$$

where $g(\cdot)$ is not a linear function. In his causal inference class notes, Wager shows that regressing on prior variables will be asymptotically better than not doing it even if the linear model used is not right. We would like to compare $\mathbb{E}(Y_{ij} - g(\boldsymbol{x}_{ij}) \,|\, A)$ and $\mathbb{E}(Y_{ij} - g(\boldsymbol{x}_{ij}) \,|\, B)$. We get to compare $\mathbb{E}(Y_{ij} - \tilde{g}(\boldsymbol{x}_{ij}) \,|\, A)$ and $\mathbb{E}(Y_{ij} - \tilde{g}(\boldsymbol{x}_{ij}) \,|\, B)$ for some imperfect function $\tilde{g}(\cdot)$, such as a linear model approximation to $g(\cdot)$.

One way to see this is to consider just doing a plain ANOVA without the regression on x_{ij}. That will have validity from the randomization while it also corresponds to taking $\beta = 0$, or $\tilde{g}(\cdot) = 0$, which will ordinarily not be the best regression model to have used.

8.4 Post treatment variables

Suppose that we measure x_{ij} prior to treatment, then apply the treatment and then measure a response Y_{ij} and additional post-treatment variables z_{ij}. We might contemplate the model

$$Y_{ij} = \mu_i + (x_{ij} - \bar{x}_{\bullet\bullet})^{\mathsf{T}}\beta + (z_{ij} - \bar{z}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij} \qquad \text{(bad idea!)}.$$

This model is a very bad choice if our goal is to test whether the treatment has a causal impact on Y. Suppose that the treatment has a causal impact on Z which then has a causal impact on Y. We might then find that Z explains Y so well that the treatment variable appears to be insignificant. For instance, the causal implications in an agriculture setting might be that a pesticide treatment has a causal impact on the quantity X of insects in an orchard, and that in turn has a causal impact on the size Y of the apple harvest. Including Z in the regression could change the magnitude of the treatment differences and lead us to conclude that the treatment does not causally affect the size of the apple harvest. This could be exactly wrong if the treatment really does increase the apple harvest ***because*** of its impact on the pests.

With data like this we could do two ANCOVAs. One relates Y to the treatment and any pre-treatment predictors. Another relates z to the treatment and any pre-treatment variables. We could then learn whether the treatment affects z and that might be a plausible mechanism for how it affects Y.

If there is a strong predictive value in $\boldsymbol{x}_{ij}^{\mathsf{T}}\beta$ then we get a more precise comparison of treatment groups from the ANCOVA than we would from an ANOVA that ignored $\boldsymbol{x}_{ij}$.

8.5 Crossover trials

Consider two painkillers A and B. In a regular trial, n people get A and n people get B by a random assignment. Then later we compare the two groups of people.

In a cross-over trial, each of the people would get both A and B and we would then compare their experiences with the two medications. The potential advantage is that now the different people are analogous to blocks. If there are strong person to person variations in the response we can balance them out. It is like the running shoe example from before, except that while kids can wear two kinds of running shoes at once, we cannot have people take two medicines at once because we would not know how to separate the effects that they had on the people. For the shoes, the response was the effect that the person had on two separate shoes and that's different.

In the cross-over trial n people get A for a period (say one week) then they wait some times (perhaps also a week) for the effects to wash out and then they get B for the same length of time that they got A. Another n people go through the trial in the opposite order taking B first then waiting for it to wash out and then taking A. We can sketch the layout like this:

	Period 1	Period 2
Subject Group 1	A	B
Subject Group 2	B	A

The complication in cross-over designs is that the treatment in period one might affect the measurement in period two. If we are confident that the washout

period is long enough then we might bet on a cross-over design. Clearly if the first period treatment can bring a permanent cure (or irreversible damage) then the second period is affected and a cross-over is problematic. Cross-over designs are well suited for chronic issues that require continual medication.

Here is a sketch from Brown (1980) of how cross-over data looks.

		Group 1		Group 2
Period	Trt	Subjects $S_{11}, \cdots, S_{1n_1}$	Trt	Subjects $S_{21}, \ldots, S_{2n_2}$
1	A	$Y_{111}, \cdots, Y_{1n_11}$	B	$Y_{211}, \cdots, Y_{2n_21}$
2	B	$Y_{112}, \cdots, Y_{1n_12}$	A	$Y_{221}, \cdots, Y_{2n_22}$

There are two groups of subjects, two treatments and two periods. For instance, in period 1, group 1 gets treatment A and yields observations $Y_{111}, \cdots, Y_{1n_11}$.

Here is his regression model after the carry-over term has been removed.

$$\begin{aligned} Y_{ijk} &= \mu + \pi_k + \tau_{u(i,k)} + \eta_{ij} + \varepsilon_{ijk} \\ i &= \text{group 1 or 2} \qquad j = \text{subject 1 to } n_i \qquad k = \text{period 1 or 2} \\ u &= u(i,k) = \text{treatment A or B} \qquad \text{A iff } i = k \\ \tau_u &= \text{treatment effect} \\ \eta_{ij} &= \text{subject effect.} \end{aligned}$$

Notice the subscript on the treatment effect τ. The treatment level is $u(i,k)$ which is A if it is group $i = 1$ in period $k = 1$ or group $i = 2$ in period $k = 2$. If $i \neq k$ then the treatment is B.

As in ANCOVA, we will consider some differences. Let D_j be the period 2 value minus the period 1 value for subject j. Then

$$\mathbb{E}(D_j) = \begin{cases} \underbrace{\pi_1 - \pi_2}_{\text{period effect}} + \underbrace{\tau_1 - \tau_2}_{\text{treatment effect}} & \text{Group } i = 1 \\ \underbrace{\pi_1 - \pi_2}_{\text{period effect}} + \underbrace{\boldsymbol{\tau_2 - \tau_1}}_{\text{treatment effect}} & \text{Group } i = 2 \end{cases}$$

and so

$$\mathbb{E}(D_j \,|\, i = 1) - \mathbb{E}(D_j \,|\, i = 2) = 2(\tau_1 - \tau_2).$$

That is, the group differences of D_j inform us about treatment difference $\tau_1 - \tau_2$. We can use this to test for treatment differences via a t-statistic

$$t = \frac{\bar{D}_{gp1} - \bar{D}_{gp2}}{s_D} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad s_D^2 = \frac{(n_1 - 1)s_{Dgp1}^2 + (n_2 - 1)s_{Dgp2}^2}{n_1 + n_2 - 2}.$$

Or we could do an unpooled t-test.

Similarly

$$\mathbb{E}(D_j \,|\, i = 1) + \mathbb{E}(D_j \,|\, i = 2) = 2(\pi_1 - \pi_2).$$

Therefore we can test for period differences by summing group differences of D_j.

From the Table that Brown gives we can see that the A/V treatment effect is exactly the same as an interaction between periods and groups. A crossover would be bad choice if we thought such an interaction likely.

Now suppose that we want to test for a carryover effect. We could add a predictor variable to the regression model. We would construct a predictor variable that takes the value 0 for all period one data, takes the value +1 in period 2 if the period 1 treatment was A and itakes the value −1 in period 2 if the period 1 treatment was B. This predictor times its coefficient does nothing in period 1, as appropriate, because there was not prior effect that could carry over to period 1. In period 2 it models twice the difference at period 2 from having period A versus B at period 1.

Some authors advise against crossovers if carryover is at all possible. Some revert to period 1 after testing for carryover and finding that it is present. There is some controversy over this method. The test for carryover may not be reliable enough to use. See Brown (1980). There is a numerical example, also from dentistry, in Brown's paper at `https://www.jstor.org/stable/2530496`.

8.6 More general crossover

Suppose that we have 3 treatments. We could arrange them in a 3 period cross-over as follows.

$$\begin{matrix} Per1 & Per2 & Per3 \end{matrix}$$
$$\begin{bmatrix} A & B & C \\ A & C & B \\ B & A & C \\ B & C & A \\ C & B & A \\ C & A & B \end{bmatrix}$$

This uses 6 groups. Cox (1958) give an example using 3 groups

$$\begin{matrix} Per1 & Per2 & Per3 \end{matrix}$$
$$\begin{bmatrix} B & A & C \\ C & B & A \\ A & C & B \end{bmatrix}$$

This is a Latin square. Each treatment appears once in each of the three periods. It misses one kind of balance that we might want. If we thought that carryover were possible then we might want every treatment to have a balanced set of immediate predecessors. For instance each of AB, AC, BC, BA, CA, CB should appear the same number of times in consecutive time periods. The Latin square above has BA twice but never has BC. The top design has all 6 possible permutations once and so by symmetry it is balanced. For instance A is first twice, follow B twice and follows C twice. Letters B and C are similarly balanced.

If we were interested in carryover effects and wanted to measure them well, we could add more groups. With two treatments, we could have four groups: AB, BA, AA and BB. Or we could have groups like: ABB, ABA, BAB, and BAA.

These and similar ideas are taken up at length in the book Jones and Kenward (2014).

Bibliography

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, pages 69–79.

Cox, D. R. (1958). *Planning of experiments.* Wiley.

Fleiss, J. L. (1986). *Design and analysis of clinical experiments*, volume 73. John Wiley & Sons.

Jones, B. and Kenward, M. G. (2014). *Design and analysis of cross-over trials.* CRC press.

Montgomery, D. C. (1997). *Design and analysis of experiments.* John wiley & sons, 4 edition.