Analysis of Covariance

when you compare treatment groups but know there's a continous "nuisance" factor (covariate) affecting your response, ANCOVA adjusts for it — so treatment comparisons are fairer + more powerful.

Hao Zhang

Introduction

In order to see the difference of treatment effects, we must control the nuisance factors if they are expected to be a major source of variation. If the values of the nuisance factors can be measured in advance of the experiment or controlled during the experiment, then they can be taken into account by using blocking factors. There are situations when some nuisance factors or variables may affect the response and they are also measured. For example, we would like to compare the effects of four diets on the weights of month-old piglets. The response variable is the weight three months after starting the diets, which is likely to be affected by the weight at the beginning of the experiment. We would use the starting weight as a covariate. Therefore the employment of covariates is another way to deal with nuance factors or variables. The analysis of covariance of the study to compare treatment effects in the presence of covariances. Note in terms of design in the example, it is a completely randomized design. What is new here we have additional measurements on the experimental units (piglets' starting weights).

Covariates are often continuous variables. Their effects on the response variable are assumed to follow some parametric form such as linear. → Ex) Depend on X, X^2... etc

Ex) Linear Quadratic...etc

Model with One Covariate:

For a completely randomized design with one covariate, the simplest model is

Linear Covariate

$$Y_{ij} = \mu + \underset{\text{Treatment factor}}{\tau_i} + \boxed{\beta x_{ij}} + \epsilon_{ij} \tag{1}$$

where Y_{ij} is the response variable of the jth unit from the ith treatment group; τ_i is the effect of the ith treatment, x_{ij} is the value of the covariate of the jth unit in the ith treatment group, ϵ_{ij} are i.i.d. $N\left(0, \sigma^2\right)$.

Here we assume that the effect of the covariate on the response is linear. It can be extended to more complex cases, e.g., quadratic, cubic, or polynomial effects.

Key assumptions

- **Linearity** between covariate and response within each treatment.
- **Equal slopes** (β) across treatments.
- **Covariate measured without error** and independent of treatments.
- **Normality** and **equal variance** of residuals.

Least Squares Estimates

Estimate the parameters by minimizing

$$\sum_{i=1}^{v}\sum_{t=1}^{r_i} e_{it}^2 = \sum_{i=1}^{v}\sum_{t=1}^{r_i}\left(y_{it} - \mu - \tau_i - \beta\left(x_{it} - \bar{x}_{..}\right)\right)^2.$$

centered around its grand mean $\bar{x}$.

Taking derivatives with respect to μ, τ_i and β, we get

$$y_{..} = n\hat{\mu} + \sum_{i=1}^{v} r_i\hat{\tau}_i$$

$$y_{i.} = r_i\left(\hat{\mu} + \hat{\tau}_i\right) + \hat{\beta}\sum_{t=1}^{r_i}\left(x_{it} - \bar{x}_{..}\right), \quad i = 1, \ldots, v$$

$$\sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it}\left(x_{it} - \bar{x}_{..}\right) = \sum_{i=1}^{v}\sum_{t=1}^{r_i}\left(\hat{\mu} + \hat{\tau}_i\right)\left(x_{it} - \bar{x}_{..}\right) + \sum_{i=1}^{v}\sum_{t=1}^{r_i}\hat{\beta}(x_{it} - \bar{x}_{..})^2$$

subject to $\sum_{i=1}^{v}\tau_i = 0$.

Solving for these equations to get the least squares estimators for the parameters. Note this estimators are linear functions of Y_{it} and therefore have normal distributions.

Analysis of Covariance

For a completely randomized design and analysis of covariance model (1), a one-way analysis of covariance is used to test the null hypothesis $H_0 : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$ against the alternative hypothesis H_A that at least two of the τ_i 's differ. As with other models, the test is based on the comparison of error sums of squares under the full and reduced models. If the null hypothesis is true with common treatment effect $\tau_i = 0$, then the reduced model is

$$Y_{it} = \mu + \beta\left(x_{it} - \bar{x}_{..}\right) + \epsilon_{it}.$$

Hence

$$SST = \underbrace{SSE_0}_{\text{Reduced}} - \underbrace{SSE}_{\text{Full}}$$

It can be shown that under H_0, SST/σ^2 has a χ^2 distribution with $(v-1)$ degrees of freedom, and $(v-1)$ is the number of parameters the reduced model reduced by.

If the null hypothesis is true, then

$$\boxed{\mathrm{MST}\ /\mathrm{MSE} \sim F_{v-1,n-v-1},}$$

Test treatment effect

so we can obtain a decision rule for testing $H_0: \{\tau_1 = \tau_2 = \cdots = \tau_v\}$ against $H_A: \{\tau_i$ not all equal$\}$ as reject H_0 if $\mathrm{msT}/\mathrm{msE} > F_{v-1,n-v-1,\alpha}$, where $n-v-1$ is the degrees of freedom for error.

Test Covariate effect

Similarly, the decision rule for testing $H_0: \{\beta = 0\}$ against $H_A: \{\beta \neq 0\}$, at significance level α, is reject H_0 if $ms\beta/msE > F_{1,n-v-1,\alpha}$.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio
T	$v-1$	ssT	$\frac{ssT}{v-1}$	$\frac{msT}{msE}$
β	1	$ss\beta$	$ss\beta$	
Error	$n-v-1$	ssE		
Total	$n-1$	ss_{yy}	msE	

F test: Compare to reduced models

Contrasts and Simultaneous Confidence Intervals

The inference for a single contrast and multiple contrasts is done similarly to ANOVA. Read Section 9.5 for more details.

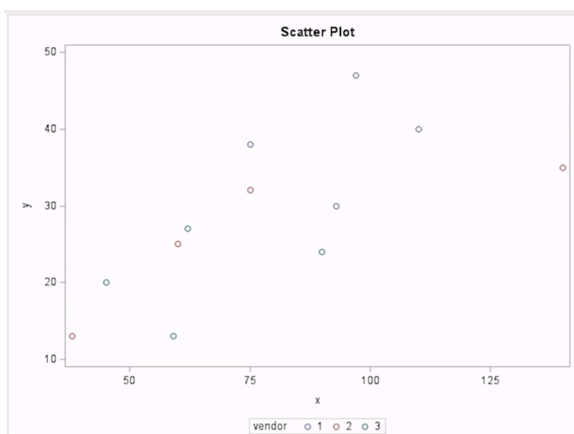
Example

The zink experiment was used by C. R. Hicks (1965), Industrial Quality Control, to illustrate the possible bias caused by ignoring an important covariate. The experimental units consisted of 12 steel brackets. Four steel brackets were sent to each of three vendors to be zinc plated. The response variable was the thickness of the zinc plating, in hundred-thousandths of an inch. The thickness of each bracket before plating was measured as a covariate. The data are reproduced in Table 9.8.

A. Plot the data to see if it is necessary to use the covariate.

```
data zinc;
  input vendor   x    y;
  lines;
          1    110  40
          1     75  38
          1     93  30
          1     97  47
          2     60  25
          2     75  32
          2     38  13
          2    140  35
          3     62  27
          3     90  24
          3     45  20
          3     59  13
;
run;

proc sgplot;
scatter Y=y X=x/group=vendor;
title "Scatter Plot";
run;
```



upward trend ↑
Use linear trend

(B). Fit the ANCOVA model and test the equality of vendor effects.

ANCOVA

```
proc glm data=zinc;
class vendor;
model y=vendor x;
run;
```

No x — class vendor; Do not include "x" in class.

Yes x — model y=vendor x;

tells "model" to adjust for "x" so vendor not confounded by "x" variation.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
vendor	2	292.4217278	146.2108639	3.57	0.0778
x	1	216.1015128	216.1015128	5.28	0.0506

For $\alpha = 0.05$, we fail to reject the null hypothesis. Therefore there is no significant evidence to indicate a difference among the vendors.

C. What happens if the covariate was not included?

Simple one-way ANOVA

```
proc glm data=zinc;
class vendor;
model y=vendor;
run;
```

Missing "x".

Source	DF	Type III SS	Mean Square	F Value	Pr > F
vendor	2	665.1666667	332.5833333	5.51	0.0274

Now it says there is a significant difference among the vendors! But actually, some of the differences are due to the thickness of the brackets, i.e., the covariate.

D. It is possible to extend the model so that the slope of the covariate is not assumed to be the same for all vendors. The following code is for illustration only because of the small sample size.

X

```
proc glm data=zinc;
class vendor;
model y=vendor x vendor*x;
run;
```

A B A*B Interaction

"X" here is not a class variable, it's a covariate.
↳ continuous variable.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
vendor	2	18.50178803	9.25089401	0.17	0.8450
x	1	62.68907040	62.68907040	1.17	0.3202
x* vendor	2	6.95209274	3.47604637	0.07	0.9376

The estimates are provided in the following table.

			Standard Error	t Value	Pr > $\|t\|$
Parameter	Estimate		0.91	0.3965	
Intercept	13.49530957	B	14.78397390	0.6464	
vendor 1	15.12056598	B	31.31879622	0.48	0.6464
vendor 2	−1.84416489	B	16.98609165	−0.11	0.9171
vendor 3	0.00000000	B			
x	0.11726079	B	0.22383267	0.52	0.6191
x * vendor 1	−0.00916346	B	0.36785165	−0.02	0.9809
x * vendor 2	0.06930605	B	0.24361107	0.28	0.7856
x * vendor 3	0.00000000	B			

*slope
Ex) Vendor 2:
0.11726079 +
0.0693605

It is interpreted as follows:

The slope estimate for vendor 3 is 0.1172. The slope estimate for vendor 2 is 0.1172+0.0693=0.1865. The slope estimate for vendor 1 is 0.1172-0.0092=0.108.

Download Print

```
proc glm data=zinc;
class vendor;
model y=vendor x/solution;
run;
```

Task: View this topic