Analysis of Covariance

Nathaniel E. Helwig

Assistant Professor of Psychology and Statistics
University of Minnesota (Twin Cities)

Updated 04-Jan-2017

Copyright

Outline of Notes

1) ANCOVA Overview:

- Model form
- Model assumptions
- Estimating parameters
- Significance testing
- Multiple comparisons

2) Auditing Example:

- Overview of data
- Testing assumptions
- Fitting ANOVA
- Fitting ANCOVA
- Multiple comparisons

Overview of ANCOVA

General Idea of Analysis of Covariance

Suppose we have a one-way ANOVA situation, but we also have one (or more) continuous variables that are associated with the response.

In Analysis of Covariance (ANCOVA) we want to incorporate additional variable(s) into the model to reduce the error variance.

- Goal is to get a better estimate of the treatment effect
- We accomplish this by including additional predictors (covariates)
- If covariates are related to Y, error variance σ^2 is reduced, so we have more power to examine treatment differences

Analysis of Covariance Model (uncentered)

The uncentered Analysis of Covariance (ANCOVA) model has the form

$$y_{ij} = \tilde{\mu} + \alpha_j + \beta x_{ij} + e_{ij}$$

for $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, g\}$ where

- $y_{ij} \in \mathbb{R}$ is response for i-th subject in j-th treatment level
- $\tilde{\mu} \in \mathbb{R}$ is a constant common to all individuals ($\tilde{\mu} \neq \bar{y}$)
- $\alpha_j \in \mathbb{R}$ is the treatment effect of the j-th treatment level
- $x_{ij} \in \mathbb{R}$ is the covariate for the i-th subject in the j-th treatment level
- $\beta \in \mathbb{R}$ is the regression slope corresponding to the covariate x_{ij}
- $e_{ij} \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

Implies that $y_{ij} \stackrel{\text{ind}}{\sim} \mathrm{N}(\tilde{\mu} + \alpha_j + \beta x_{ij}, \sigma^2)$.

Analysis of Covariance Model (centered)

The centered Analysis of Covariance (ANCOVA) model has the form

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}) + e_{ij}$$

for $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, g\}$ where

- $y_{ij} \in \mathbb{R}$ is response for i-th subject in j-th treatment level
- $\mu \in \mathbb{R}$ is population mean common to all individuals ($\mu = \bar{y}$)
- $\alpha_j \in \mathbb{R}$ is the treatment effect of the j-th treatment level
- $x_{ij} \in \mathbb{R}$ is the covariate for the i-th subject in the j-th treatment level
- $\beta \in \mathbb{R}$ is the regression slope corresponding to the covariate x_{ij}
- $e_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

Implies that $y_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(\mu + \alpha_j + \beta(x_{ij} - \bar{x}), \sigma^2)$.

ANCOVA Assumptions: Overview

The fundamental assumptions of the ANCOVA model are:

1. Relationship between x and y is linear
2. Relationship between x and y is same for each treatment level; note: parallel slopes (homogeneity of slopes)
3. x_{ij}, y_{ij} and z_{ij} are observed random variables (known constants)
4. $e_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
5. μ, $\{\alpha_j\}_{j=1}^{g}$, and β are unknown constants
6. $(y_{ij}|x_{ij}) \stackrel{\text{ind}}{\sim} \mathrm{N}(\mu + \alpha_j + \beta(x_{ij} - \bar{x}), \sigma^2)$; note: homogeneity of variance

Criteria for the Covariate

For a proper ANCOVA, the covariate X...

1. Should be (linearly) associated with the response Y
2. Should NOT be associated with the treatment Z
3. Should (ideally) be collected/observed before the study

ANCOVA is designed for experiments where treatments are randomly assigned to experimental units.

- Assume that each treatment group has approximately the same mean on the covariate X.

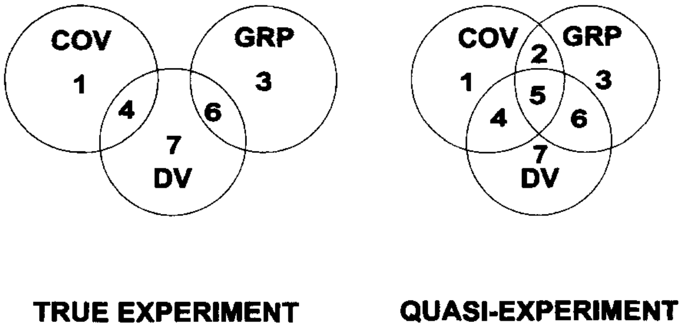
Collection of the Covariate

Collection of the covariate X should NOT influence or be influenced by the treatment Z!

Note that the “treatment levels” should NOT be preexisting groups

- If the preexisting groups truly differ on the response variable and if the covariate is truly related to the response variable, then there are likely preexisting group differences on the covariate as well
- The variance in the response variable that is explained by the covariate now overlaps with the variance in the response variable that is explained by the treatment
- Including the covariate in the model inherently alters what your group (or treatment) effect now represents

Helpful Figure of True versus Quasi ANCOVA



Miller, G.A., & Chapman, J.P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*, 40–48.

Analysis of Covariance Model (effect coding)

Effect coding uses $g-1$ variables to code a factor:

$$z_{ij} = \begin{cases} 1 & \text{if } i\text{-th observation is in } j\text{-th level} \\ -1 & \text{if } i\text{-th observation is in } g\text{-th level} \\ 0 & \text{otherwise} \end{cases}$$

for $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, g-1\}$.

Analysis of Covariance model becomes

$$y_{ij} = \mu + \sum_{j=1}^{g-1} \alpha_j z_{ij} + \beta(x_{ij} - \bar{x}) + e_{ij}$$

where $\alpha_g = -\sum_{j=1}^{g-1} \alpha_j$ because $\sum_{j=1}^{g} \alpha_j = 0$

Analysis of Covariance Model (matrix form)

In matrix form, the ANCOVA model for the *j*-th treatment level is

$$\mathbf{y}_j = \mathbf{X}_j \mathbf{b} + \mathbf{e}_j$$

$$\begin{pmatrix} y_{1j} \\ y_{2j} \\ y_{3j} \\ \vdots \\ y_{n_j j} \end{pmatrix} = \begin{pmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1(g-1)} & x_{1j} \\ 1 & z_{21} & z_{22} & \cdots & z_{2(g-1)} & x_{2j} \\ 1 & z_{31} & z_{32} & \cdots & z_{3(g-1)} & x_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & z_{n_j 1} & z_{n_j 2} & \cdots & z_{n_j(g-1)} & x_{n_j j} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{g-1} \\ \beta \end{pmatrix} + \begin{pmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \\ \vdots \\ e_{n_j j} \end{pmatrix}$$

Analysis of Covariance Model (matrix form continued)

We can write the model for all treatment levels simultaneously using

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_g \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \vdots \\ \mathbf{X}_g \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \vdots \\ \mathbf{e}_g \end{pmatrix}$$

where **b** is defined as it was on the previous slide.

Estimating Treatment Effects and Slope

Note that $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ has the General Linear Model (GLM) form.

- ANCOVA is a particular type of multiple regression

$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is our ordinary least squares estimate of $\mathbf{b}$

- $\hat{\mathbf{b}} \sim \mathrm{N}(\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ from our assumptions

$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{H}\mathbf{y}$ are the fitted values where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

- $\hat{\mathbf{y}} \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{H})$ from our assumptions

$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ are the residuals

- $\hat{\mathbf{e}} \sim \mathrm{N}(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ from our assumptions

Testing Significance of Effects

Significance tests can be formulated using the GLM F test statistic:

$$F = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$
$$\sim F_{(df_R - df_F, df_F)}$$

where

- SSE_R is sum-of-squares error for reduced model
- SSE_F is sum-of-squares error for full model
- df_R is error degrees of freedom for reduced model
- df_F is error degrees of freedom for full model

Testing Significance of Covariate

To test a significant linear relationship between X and Y

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

we could use two different sets of full and reduced models.

Type I SS (with covariate entered first)

- F: $y_{ij} = \mu + \beta(x_{ij} - \bar{x}) + e_{ij}$
- R: $y_{ij} = \mu + e_{ij}$

Type II/Type III SS

- F: $y_{ij} = \mu + \sum_{j=1}^{g-1} \alpha_j z_{ij} + \beta(x_{ij} - \bar{x}) + e_{ij}$
- R: $y_{ij} = \mu + \sum_{j=1}^{g-1} \alpha_j z_{ij} + e_{ij}$

Testing Significance of Treatment

To test the significance of the treatment effect Z

$$H_0 : \alpha_j = 0 \; \forall \; j \quad \text{versus} \quad H_1 : \alpha_j \neq 0 \text{ for some } j$$

there is only one reasonable test for the ANCOVA model.

Type II/Type III SS

- F: $y_{ij} = \mu + \sum_{j=1}^{g-1} \alpha_j z_{ij} + \beta(x_{ij} - \bar{x}) + e_{ij}$
- R: $y_{ij} = \mu + \beta(x_{ij} - \bar{x}) + e_{ij}$

Comparing Treatment Effects

If there is a significant treatment effect, we need to conduct a follow-up test to determine which treatment levels significantly differ.

- Paired comparisons and/or other contrasts

The adjusted means are the least squares means, which are the treatment level means adjusted for the average covariate values.

- $\bar{y}_j^* = \bar{y}_j - \hat{\beta}(\bar{x}_j - \bar{x})$

Typically want to compare differences in adjusted means in ANCOVA.

Comparing Treatment Effects in R

Can obtain adjusted means using `predict` function.

- Just need the least-squares mean for each treatment level
- Need to obtain predictions at average covariate value

Multiple comparisons can be performed using various procedures

- Bonferroni adjustment is a flexible option
- The `multcomp` package in R has many options

Auditing Example

Auditing Data (Kutner, Nachtsheim, Neter, & Li, 2005)

An accounting firm wants to study the effectiveness of three different methods for training statistical auditors.

- Method A: study at home with provided materials
- Method B: train in local offices with local staff
- Method C: train in Chicago office with national staff

A total of $n = 30$ auditors were blocked into ten groups of three, and the three training methods were randomly assigned within each block.

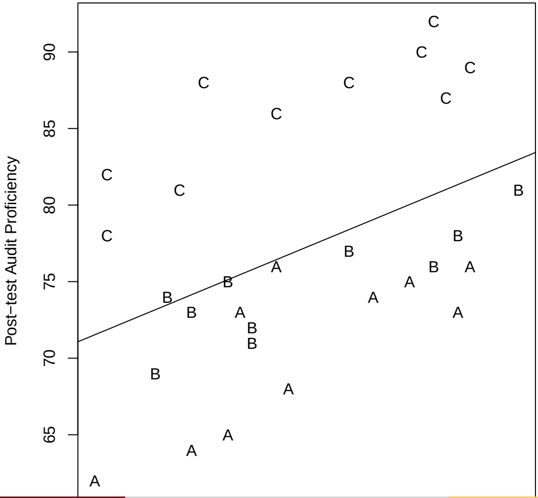
- Note that $n_j = 10$ for $j \in \{1, 2, 3\}$

Pre- and post-test measures of auditing proficiency were collected before and after the training (units of measurement not comparable).

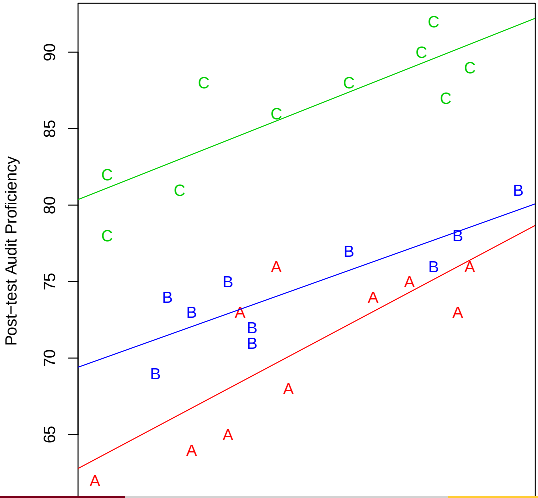
Look at Data in R

```
> head(audit)
  method pre post
1      A  93   73
2      B  98   81
3      C  91   92
4      A  94   76
5      B  93   78
6      C  94   89
> tapply(audit$pre,audit$method,mean)
   A    B    C
80.2 80.0 79.9
> tapply(audit$post,audit$method,mean)
   A    B    C
70.6 74.6 86.1
> audit=cbind(audit,cpre=(audit$pre-mean(audit$pre)))
```

Plot Data with Overall Regression Line



Plot Data with Treatment-Specific Regression Lines



Overview of ANCOVA Assumption Testing

Before fitting an ANCOVA, we will check...

- If there is a linear relationship between X and Y
- If the parallel slopes assumption is reasonable
- If the covariate X and treatment Z are related
- If data meet homogeneity of variance assumption (for ANOVA)

Testing for Linear Relationship between X and Y

```
> rmod=lm(post~pre,data=audit)
> rmod$coef
(Intercept)         pre
 50.9216933   0.3270925
> anova(rmod)
Analysis of Variance Table

Response: post
          Df Sum Sq Mean Sq F value  Pr(>F)
pre        1  344.4  344.40  6.4446 0.01697 *
Residuals 28 1496.3   53.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that there is a positive linear relationship between `pre` and `post`, so the pre-test audit proficiency scores could be a useful covariate.

Testing Parallel Slopes Assumption

```
> contrasts(audit$method)=contr.sum(3)
> imod=lm(post~pre*method,data=audit)
> anova(imod)
Analysis of Variance Table

Response: post
           Df  Sum Sq Mean Sq F value    Pr(>F)    
pre         1  344.40  344.40 46.8127 4.469e-07 ***
method      2 1309.59  654.80 89.0046 7.908e-12 ***
pre:method  2   10.14    5.07  0.6894    0.5115    
Residuals  24  176.57    7.36                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that there is no significant interaction between `pre` and `method`, so the parallel slopes assumption is reasonable.

Testing for Relationship between X and Z

```
> anova(lm(pre~method,data=audit))
Analysis of Variance Table

Response: pre
          Df Sum Sq Mean Sq F value Pr(>F)
method     2    0.5   0.233   0.002  0.998
Residuals 27 3218.5 119.204
```

Note that there is no significant relationship between `pre` and `method`, so the covariate meets the necessary criteria for a true ANCOVA.

Testing ANOVA Homogeneity of Variance Assumption

```
> library(car)
> leveneTest(post~method,data=audit)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.5241  0.598
      27
```

Note that we retain the null hypothesis that the `post` scores have homogeneous variance across the different levels of `method`.

- So data meet assumptions necessary for ANOVA

Fitting One-Way ANOVA Model

```
> amod=lm(post~method,data=audit)
> summary(amod)$sigma
[1] 4.495677
> anova(amod)
Analysis of Variance Table

Response: post
          Df Sum Sq Mean Sq F value    Pr(>F)
method     2 1295.0  647.50  32.037 7.441e-08 ***
Residuals 27  545.7   20.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that we have a significant effect of audit training `method` on the `post` audit proficiency scores, and note that $\hat{\sigma} = 4.5$.

Multiple Comparisons using Tukey's Method

```
> TukeyHSD(aov(post~method,data=audit))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = post ~ method, data = audit)

$method
    diff        lwr       upr     p adj
B-A  4.0 -0.9849383  8.984938 0.1341187
C-A 15.5 10.5150617 20.484938 0.0000001
C-B 11.5  6.5150617 16.484938 0.0000129
```

Training `method C` results in larger `post` audit proficiency scores, whereas there are no significant differences between `A` and `B`.

Fitting Uncentered ANCOVA Model

```
> umod=lm(post~pre+method,data=audit)
> umod$coef
(Intercept)         pre     method1     method2
 50.3708311   0.3339755  -6.5556626  -2.4888675
> summary(umod)$sigma
[1] 2.679766
> anova(umod)
Analysis of Variance Table

Response: post
          Df  Sum Sq Mean Sq F value    Pr(>F)
pre        1  344.40  344.40  47.958 2.366e-07 ***
method     2 1309.59  654.80  91.183 1.778e-12 ***
Residuals 26  186.71    7.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a significant effect of training `method` on the `post` test scores after controlling for `pre` test scores, and note that $\hat{\sigma} = 2.7$.

Fitting Centered ANCOVA Model

```
> cmod=lm(post~cpre+method,data=audit)
> cmod$coef
(Intercept)        cpre     method1     method2
 77.1000000   0.3339755  -6.5556626  -2.4888675
> summary(cmod)$sigma
[1] 2.679766
> anova(cmod)
Analysis of Variance Table

Response: post
          Df  Sum Sq Mean Sq F value    Pr(>F)
cpre       1  344.40  344.40  47.958 2.366e-07 ***
method     2 1309.59  654.80  91.183 1.778e-12 ***
Residuals 26  186.71    7.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The only difference between this and previous solution is the intercept (which is $\hat{\beta}\bar{x} = 26.72917$ larger than uncentered intercept).

Obtaining Adjusted Means with `predict` Function

```
> postmean=tapply(audit$post,audit$method,mean)
> cpremean=tapply(audit$cpre,audit$method,mean)
> postmean - cmod$coef[2]*cpremean
       A        B        C
70.54434 74.61113 86.14453
> newdata=data.frame(method=c("A","B","C"),cpre=rep(0,3))
> yhat=predict(cmod,newdata=newdata)
> yhat
       1        2        3
70.54434 74.61113 86.14453
> BmA=yhat[2]-yhat[1]
> CmA=yhat[3]-yhat[1]
> CmB=yhat[3]-yhat[2]
> difs=c(BmA,CmA,CmB)
> names(difs)=c("B-A","C-A","C-B")
> difs
      B-A       C-A       C-B
 4.066795 15.600193 11.533398
```

Obtaining Multiple Comparisons with `multcomp`

```
> library(multcomp)
> pwc=glht(cmod,linfct=mcp(method="Tukey"))
> summary(pwc)

  Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = post ~ cpre + method, data = audit)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0    4.067      1.198   3.393  0.00591 **
C - A == 0   15.600      1.199  13.016  < 0.001 ***
C - B == 0   11.533      1.198   9.624  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```