STA461: Analysis of Covariance

Lynn Lin

November 9, 2018

Introduction

- ANCOVA is a very general term for the use of covariates in the analysis of experimental data
- ANCOVA is a technique that combines features of analysis of variance and regression
- We will focus on the motivation for ANCOVA and its interpretation
- Therefore, much of the discussion of ANCOVA will be devoted to understanding what the model means and what it assumes

Introduction

- Suppose that the response variable (Y) is linearly related to some other continuous variable (X) that the experimenter cannot control but can observe, along with Y
- Examples:
  - Native soil fertility in a yield trial
  - Initial weight of animals in a feeding trial
- ANCOVA uses X essentially as a continuous blocking variable to improve the precision of an experiment
- ANCOVA is a combination of ANOVA and Linear Regression

One-way model with one covariate

- The ANCOVA model simply includes the covariate as another predictor

$$Y_{it} = \mu + \tau_i + \beta X_{it} + \epsilon_{it}$$

- It is quite common to center the covariate at its sample mean $\bar{X}_{\cdot\cdot} = \frac{1}{n} \sum_i \sum_t X_{it}$

- The model becomes $Y_{it} = \mu + \tau_i + \beta(X_{it} - \bar{X}_{\cdot\cdot}) + \epsilon_{it}$

- We will be using this second model in this class

One-way model with one covariate

$$Y_{it} = \mu + \tau_i + \beta(X_{it} - \bar{X}_{..}) + \epsilon_{it}$$

- μ is an overall mean
- τ_i are the fixed treatment effects
- β is a regression coefficient for the relation between Y and X
- X_{it} are constants
- ϵ_{it} are independent $N(0, \sigma^2)$
- $i = 1, \dots, \nu; t = 1, \dots r_i$

One-way model with one covariate

- The intercepts have a different meaning between the two models
- First model, we interpret $(\mu + \tau_i)$ as the expected value of the response variable under treatment i when the covariate is held constant at zero
- Second model, we interpret $(\mu + \tau_i)$ as the expected value of the response variable under treatment i when the covariate is held constant at $\bar{X}_{..}$

One-way model with one covariate

Both of these models say that

- within each treatment, there is a linear relationship between the response and covariate
- these lines are parallel with a common slope β, and
- the vertical distance between the lines for treatment i and i' is $\tau_i - \tau_{i'}$

Understanding the meaning of ANCOVA

To help make a clear distinction between the parameters of ANOVA and ANCOVA, let's assume that

- the treatments were randomly assigned to units, and
- the covariate X_{it} was not causally affected by the treatment in any way. (This condition is usually satified if X_{it} is realized and measured before the treatments are applied.)

Now consider the ANOVA and centered ANCOVA models:

- ANOVA: $Y_{it} = \mu + \tau_i + \epsilon_{it}$
- ANCOVA: $Y_{it} = \mu + \tau_i + \beta(X_{it} - \bar{X}_{..}) + \epsilon_{it}$

Understanding the meaning of ANCOVA

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

Appropriateness of covariance model

1. Normality of error terms
2. Equality of error variances for different treatments
3. Equality of slopes of the different treatment regression lines
4. Linearity of regression relation with covariate
5. Uncorrelatedness of error terms

Inferences of interest

The key inferences of interest in ANCOVA are the same as with ANOVA:

- Whether the treatments have any effects

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_\nu = 0$$

$$H_1 : \text{ not all } \tau_i = 0$$

- If so what these effects are

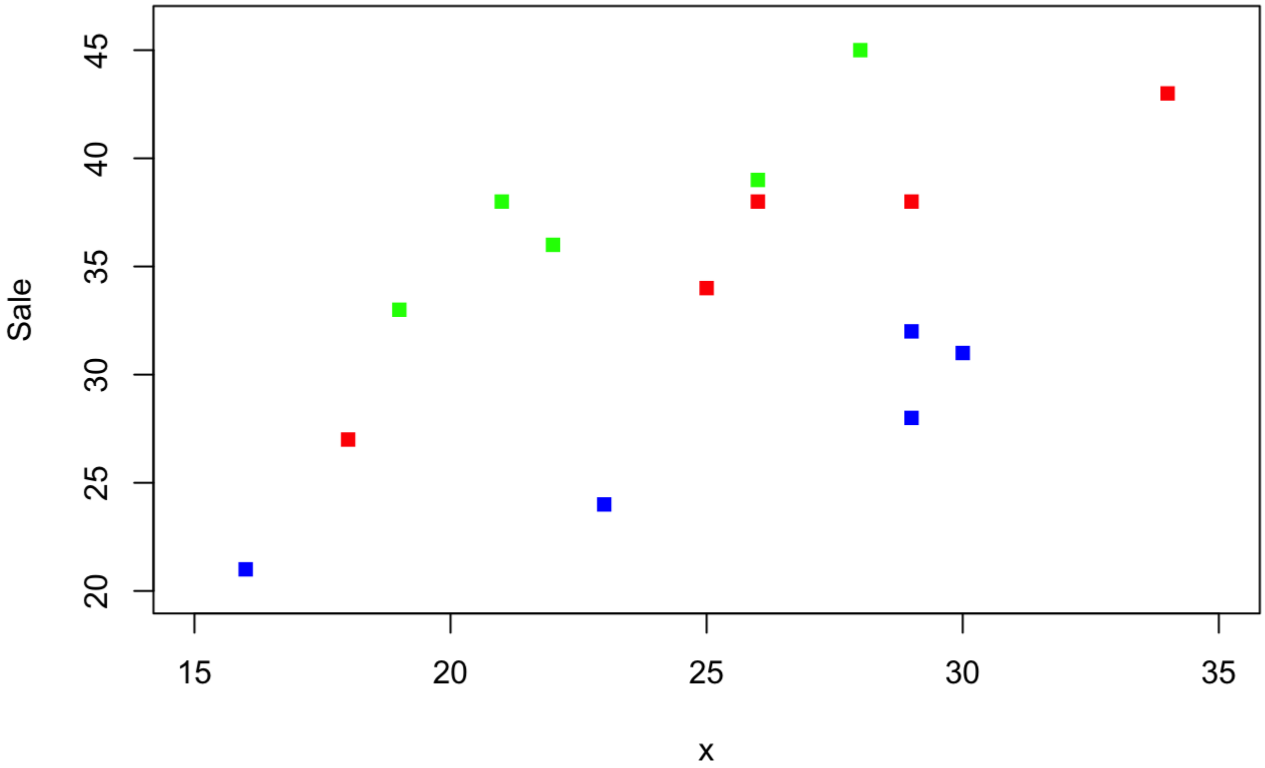
Example

```
str(dat)
```

```
## 'data.frame':    15 obs. of  4 variables:
##  $ sales    : int  38 39 36 45 33 43 38 38 27 34 ...
##  $ x        : int  21 26 22 28 19 34 26 29 18 25 ...
##  $ treatment: int  1 1 1 1 1 2 2 2 2 2 ...
##  $ store    : int  1 2 3 4 5 1 2 3 4 5 ...
```

```
dat$treatment = factor(dat$treatment)
```

Example



Example

It appears that both the linearity and equal slopes assumptions required for ANCOVA are valid

$$Y_{it} = \mu + \tau_i + \beta(X_{it} - \bar{X}_{..}) + \epsilon_{it}$$

Example

```
results = lm(sales ~ I(x-mean(x))  + treatment, dat)
anova(results)
```

```
## Analysis of Variance Table
##
## Response: sales
##                Df Sum Sq Mean Sq F value    Pr(>F)
## I(x - mean(x))  1 190.68 190.678  54.379 1.405e-05 ***
## treatment       2 417.15 208.575  59.483 1.264e-06 ***
## Residuals      11  38.57   3.506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

- The output tells us that the three cracker promotions differ in effectiveness

$$F = 59.48, p - value < 0.0001$$

- One can now continue by using multiple comparison techniques to determine how they differ

- Note that we also need to check the residuals to determine whether the other model assumptions hold

Example

```r
library(multcomp)
summary(glht(aov(sales ~ I(x-mean(x))  + treatment, dat), linfct=mcp(
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = sales ~ I(x - mean(x)) + treatment, data = dat)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0   -5.075      1.229  -4.130  0.00439 **
## 3 - 1 == 0  -12.977      1.206 -10.764  < 0.001 ***
## 3 - 2 == 0   -7.901      1.189  -6.647  < 0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Pairwise comparisons

- To obtain simultaneous coverage for all three comparisons, we may use Scheffe or Bonferroni CIs

$$\text{Sample estimate} \pm \text{ multiplier} \times \text{ standard error}$$

- The estimates and standard errors (square roots of the variances) have been found in the above output

Pairwise comparisons

- The multiplier for 95% Scheffe CIs are given by

$$\sqrt{(\nu-1)F(0.95;\nu-1,n-\nu-1)}$$

- The multiplier for 95% Scheffe CIs are given by

$$t(1-\alpha/(2g);n-\nu-1)$$

Test for parallel slopes

To test whether there are interactions between treatment and covariate

```
mod1 <- aov(sales ~ I(x-mean(x))  + treatment, dat)
mod2 <- aov(sales ~ I(x-mean(x))*treatment, dat)
anova(mod1, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ I(x - mean(x)) + treatment
## Model 2: sales ~ I(x - mean(x)) * treatment
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 38.571
## 2      9 31.521  2    7.0505 1.0065 0.4032
```

Two-factor covariance analysis

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

- Can extend this model to multiple factors or multiple covariates (or both)

Two-factor covariance analysis

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma(X_{ijt} - \bar{X}_{...}) + \epsilon_{ijt}$$

- $i = 1, \ldots, a; j = 1, \ldots, b; t = 1, \ldots, r$
- Basic idea remains the same. For each treatment combination we have a linear regression in which the slopes are the same, but the intercepts may differ
- We make comparisons using least-square means, with the covariates set to their mean values