Chapter 13: Analysis of Covariance

Gunwoong Park

Lecture Note

University of Seoul

Analysis of Covariance

- Categorical variables in regression
- Two-level factor example
- Multi-level factor example

Categorical Variables

- Qualitative variables a.k.a. categorical variables a.k.a. factors: e.g., eye color
- Analysis of covariance: regression problems with **both** quantitative and qualitative predictors.

A simple example

- Two predictors: x_1 age, x_2 indicates whether taking medication or not:
$$x_2 = \begin{cases} 0 & \text{No medication} \\ 1 & \text{Taking medication} \end{cases}$$
- Response y: cholesterol level
- A column of 0s and 1s in the design matrix X for x_2

Possible models

- Same model for both groups:

$$y = \beta_0 + \beta_1 x_1$$

- Same slope but different intercepts:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Different slopes and intercepts (interactions)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Two-level Example

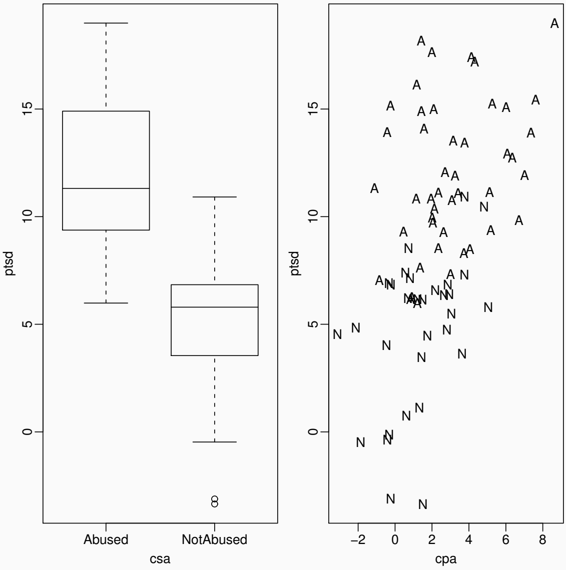
- Study of the effects of childhood sexual abuse on adult females
- Response: post traumatic stress disorder (`ptsd`) – continuous standardized scale
- Predictors:
  - Existence of childhood sexual abuse (`csa` – 0 or 1)
  - Childhood physical abuse (`cpa` – continuous standardized scale)
- 45 cases `csa = 1`, 31 cases `csa = 0`.

Data summaries

```
> library(faraway)
> data(sexab)
> sexab
cpa     ptsd       csa
1   2.04786  9.71365    Abused
2   0.83895  6.16933    Abused
... ...
75  2.85253  6.84304 NotAbused
76  0.81138  7.12918 NotAbused

> ## Summary of the data
> by(sexab, sexab$csa, summary)

sexab$csa: Abused
cpa              ptsd                csa
Min.   :-1.115   Min.   : 5.985   Abused   :45
1st Qu.: 1.415   1st Qu.: 9.374   NotAbused: 0
Median : 2.627   Median :11.313
Mean   : 3.075   Mean   :11.941
3rd Qu.: 4.317   3rd Qu.:14.901
Max.   : 8.647   Max.   :18.993
---------------------------------------------------
```

```
sexab$csa: NotAbused
cpa                 ptsd                 csa
Min.   :-3.1204   Min.   :-3.349   Abused   : 0
1st Qu.:-0.2299   1st Qu.: 3.544   NotAbused:31
Median : 1.3217   Median : 5.794
Mean   : 1.3088   Mean   : 4.696
3rd Qu.: 2.8309   3rd Qu.: 6.838
Max.   : 5.0497   Max.   :10.914

> ## t-test for the difference between groups
> attach(sexab)
> t.test(ptsd[csa=="Abused"], ptsd[csa=="NotAbused"])
data: ptsd[csa == "Abused"] and ptsd[csa == "NotAbused"]
t = 8.9006, df = 63.675, p-value = 8.803e-13
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
5.618873 8.871565
> ## summary plots
> plot(ptsd ~ csa, sexab)
> plot(ptsd ~ cpa, pch=as.character(csa), sexab)
```

Summary plots

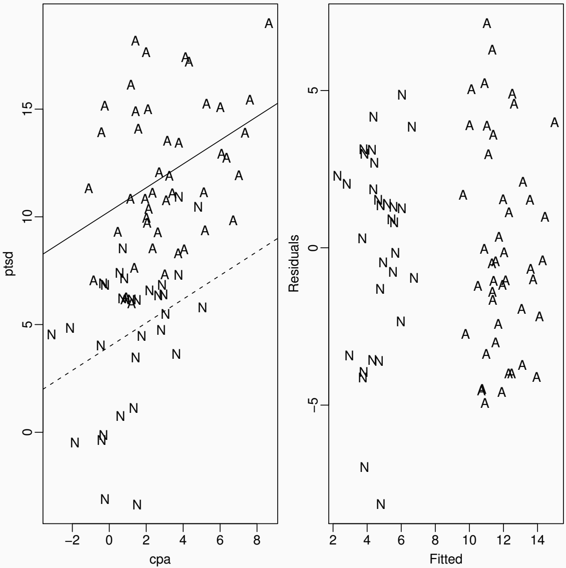


```
> ## Want to estimate interaction between CSA and CPA
> g = lm(ptsd ~ cpa + csa + cpa:csa, sexab)
> summary(g)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.5571     0.8063  13.094  < 2e-16
cpa                0.4500     0.2085   2.159   0.0342
csaNotAbused      -6.8612     1.0747  -6.384 1.48e-08
cpa:csaNotAbused   0.3140     0.3685   0.852   0.3970
---
Residual standard error: 3.279 on 72 degrees of freedom
Multiple R-Squared: 0.5828, Adjusted R-squared: 0.5654
F-statistic: 33.53 on 3 and 72 DF,  p-value: 1.133e-13
```

```
> ## What is the factor coding?
> model.matrix(g)
Intercept      cpa csaNoAbuse cpa:csaNoAbuse
1          1  2.04786          0        0.00000
2          1  0.83895          0        0.00000
... ...
75         1  2.85253          1        2.85253
76         1  0.81138          1        0.81138
> ## "Abused" is the reference level

> ## The interaction term is not significant
> g = lm(ptsd ~ cpa + csa, sexab)
> summary(g)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.2480     0.7187  14.260  < 2e-16
cpa            0.5506     0.1716   3.209  0.00198
csaNotAbused  -6.2728     0.8219  -7.632  6.9e-11
---
Residual standard error: 3.273 on 73 degrees of freedom
Multiple R-Squared: 0.5786, Adjusted R-squared: 0.5671
F-statistic: 50.12 on 2 and 73 DF,  p-value: 2.002e-14
```

```
## add regression lines
> plot(ptsd ~ cpa, pch=as.character(csa))
> abline(10.248, 0.551)
> abline(10.248-6.273, 0.551, lty=2)
## The "Abused" line is 6.273 higher than "NonAbused"
## Diagnostics
> plot(fitted(g), residuals(g), pch=as.character(csa),
xlab="Fitted", ylab="Residuals")
```

Regression and diagnostic plots



Multi-level Coding

For a K-level predictor, $K-1$ dummy variables are needed. Treatment coding is commonly used:

```
## 4-level example
> contr.treatment(4)
2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

Treat level one as the reference level to which all other levels are compared.

Multi-level Example

- NELS 88: a large longitudinal study of schoolchildren
- Response: `math` test score in 8th grade
- Predictors:
  - Parents education (`paredu`)
  - Socioeconomic status (`ses`)
- `paredu` is a 6-level factor: high school dropout, high school, some college, BA, MA, PhD

Multi-level Example

```
> data(nels88)
> nels88
sex     race   ses  paredu math
1   Female    White -0.13      hs   48
2     Male    White -0.39      hs   48
... ...
> dim(nels88)
[1] 260   5

> ## Fit the full model with all terms
> g = lm(math ~ ses*paredu, nels88)
> summary(g)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        55.6757     2.7538  20.218  < 2e-16
ses                 7.6058     3.7301   2.039  0.04251
pareducollege      -4.3727     2.9458  -1.484  0.13898
pareduhs          -10.7327     3.6660  -2.928  0.00373
paredulesshs      -15.9411     4.9613  -3.213  0.00149
pareduma            2.9368     8.1867   0.359  0.72010
pareduphd          -6.0150     8.3488  -0.720  0.47192
```

Multi-level Example

```
ses:pareducollege  -3.6922     4.3041  -0.858  0.39181
ses:pareduhs       -6.8248     4.7367  -1.441  0.15089
ses:paredulesshs   -8.7842     4.6811  -1.877  0.06176
ses:pareduma       -5.6616     7.9604  -0.711  0.47762
ses:pareduphd       0.1436     6.7020   0.021  0.98292

Residual standard error: 8.451 on 248 degrees of freedom
Multiple R-Squared: 0.4485,    Adjusted R-squared: 0.424
F-statistic: 18.34 on 11 and 248 DF,  p-value: < 2.2e-16

## Example: BA: math = 55.7 + 7.61 * ses
## PhD: math = 55.7 - 6.0 + (7.61 + 0.14) * ses
##           = 49.7 + 7.75 * ses
```

```
## Go back to conventional parametrization (model g)
## Sequential analysis of variance table
> anova(g)
Analysis of Variance Table
Response: math
Df  Sum Sq Mean Sq  F value    Pr(>F)
ses        1 12391.4 12391.4 173.5021 < 2.2e-16
paredu     5  1642.4   328.5   4.5994 0.0004959
ses:paredu  5   370.7    74.1   1.0381 0.3957126
Residuals 248 17712.0    71.4
## Interaction term is not significant
```

```
> ## Refit model without interaction
> gb = lm(math ~ ses + paredu, nels88)
> summary(gb)
Estimate Std. Error t value Pr(>|t|)
(Intercept)    58.5712     1.7813  32.882  < 2e-16
ses             2.7913     1.3096   2.132 0.034012
pareducollege  -7.5210     2.1414  -3.512 0.000526
pareduhs      -12.1792     2.6420  -4.610  6.4e-06
paredulesshs  -13.3645     3.3328  -4.010  8.0e-05
pareduma       -0.8709     2.2455  -0.388 0.698449
pareduphd      -2.0494     2.5202  -0.813 0.416885
Residual standard error: 8.454 on 253 degrees of freedom
Multiple R-Squared: 0.437,   Adjusted R-squared: 0.4236
F-statistic: 32.73 on 6 and 253 DF,  p-value: < 2.2e-16
```

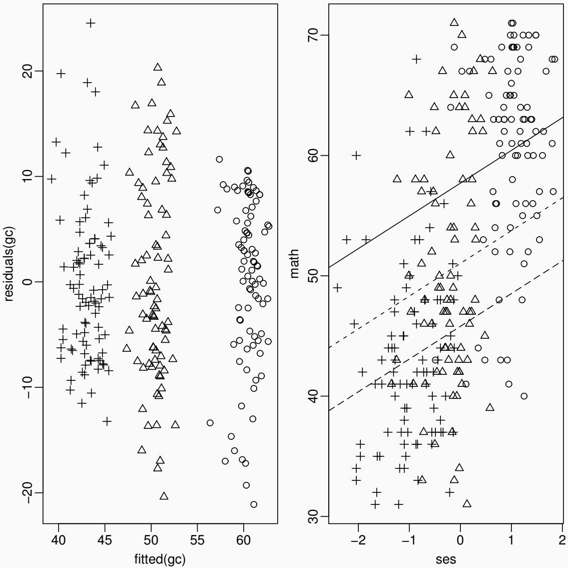
```
## BA, MA and PhD are not significantly different
## High school and less are similar
## Re-group into three levels
> nels88$edupar = nels88$paredu
> levels(nels88$edupar)
[1] "ba"      "college" "hs"      "lesshs"
[5] "ma"      "phd"
> levels(nels88$edupar) = c("degree",
"college", "highsch", "highsch", "degree",
"degree")
```

```
# fit new model with 3-level factor
> gc = lm(math ~ ses + edupar, nels88)
> summary(gc)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept     57.703     1.427  40.437  < 2e-16
ses            2.719     1.091   2.492 0.013338
eduparcollege -6.669     1.870  -3.566 0.000432
eduparhighsch-11.903     2.547  -4.673  4.8e-06
Residual standard error: 8.425 on 256 degrees of freedom
Multiple R-Squared: 0.4342,  Adjusted R-squared: 0.4276
F-statistic:  65.5 on 3 and 256 DF,  p-value: < 2.2e-16
```

```
> ## Compare the 3-level model to 6-level model
> anova(gc, gb)
Analysis of Variance Table

Model 1: math ~ ses + edupar
Model 2: math ~ ses + paredu
Res.Df     RSS  Df Sum of Sq      F Pr(>F)
1    256 18170.1
2    253 18082.7   3      87.4 0.4075 0.7478
# Merging levels is justifiable
```

```
> ## Check whether each term is significant
> ## after the other has been taken into account
> drop1(gb, test="F")
Single term deletions

Model:
math ~ ses + paredu
Df      SS     RSS     AIC F value Pr(F)
<none>             18082.7  1116.9
ses     1   324.7 18407.5  1119.6  4.5433 0.034
paredu  5  1642.4 19725.2  1129.5  4.5960 0.001

> ## Diagnostics
> par(mfrow=c(1, 2))
> plot(fitted(gc), residuals(gc),
pch=as.numeric(nels88$edupar))
> ## Regression lines
> plot(math ~ ses, nels88,
pch=as.numeric(nels88$edupar))
> abline(57.7, 2.72)
> abline(57.7 - 6.67, 2.72, lty=2)
> abline(57.7 - 11.9, 2.72, lty=5)
```

Multi-level Example Continued



Summary

- Factors are easy to incorporate into regression
- All usual diagnostic and other procedures should be followed
- With many levels and interaction terms, parameters add up very quickly – be careful not to include too many
- Confounding is still an issue except in randomized experiments