Analysis of Covariance with R

Anne Boomsma

Department of Statistics & Measurement Theory
University of Groningen

April 26, 2012

`ANCOVA_R.tex`

Analysis of Covariance with R

Anne Boomsma

Department of Statistics & Measurement Theory, University of Groningen

1. Introduction

Analysis of covariance analysis (ANCOVA) can be applied as a statistical tool for the adjustment of treatment effects in causal inference. Such analyses can be performed with the help of statistical packages, for example the S-PLUS package or the SPSS program, often used in behavioral and social sciences. As for the latter, for the simplest case we could, for example, start by choosing the options `Analyze -> General Linear Model -> Univariate`. An illustration of how to proceed on a covariance analysis with the SPSS program can be found in Field (2005, Chapter 9), providing a step-by-step program guidance — not very insightful though.

For the example of covariance analysis presented below we are using `R` software instead. The advantage of `R` usage is that we can now far more clearly see what we are doing, and we have numerous graphical options to examine the data and the estimated statistical models along with it.

2. The R environment

`R` is an open source environment for statistical computing. For an introduction we refer to The R Project for Statistical Computing at Internet site `http://www.r-project.org`. Verzani's (2005) book, also available at `http://wiener.math.csi.cuny.edu/UsingR`, provides a nice introduction; the books written by Cohen and Cohen (2008) and Faraway (2005) are more elaborate on statistical modeling. An overview of `R` commands for introductory statistics — linking Moore and McCabe's (2004) statistics book to that of Verzani — is listed at website `http://www.gmw.rug.nl/~boomsma/apstat.htm` in the portable document format file `MoorR.pdf`.

3. Data on sex abuse

The ANCOVA example was taken from Faraway (2005, Chapter 13), and the data set we consider, `sexab`, was obtained from the `R` library package `faraway`. The related subject of research deals with the post-traumatic stress disorder in abused adult females (see Rodriguez, Ryan, Vande Kemp & Foy, 1997).

```
> library(faraway)              # loading library 'faraway'
> ? sexab                       # documentation on data set 'sexab'
```

3.1. Description of the data

The sample data in the present example come from a study of the effects of childhood sexual abuse on adult females. 45 women being treated at a clinic, who reported childhood sexual abuse (`csa`), were measured for post-traumatic stress disorder (`ptsd`) and childhood physical abuse (`cpa`), both on standardized scales. Also measured were 31 women treated at the same clinic, who did not report childhood sexual abuse, were also measured. All women were treated for problems in their committed relationships with male living partners. The full study was more complex than reported here. Interested readers are referred to the original article (Rodriguez et al., 1997).

We show the three variables in the data frame `sexab` in summary:

`cpa`	Childhood physical abuse on a standard scale
`ptsd`	Post-traumatic stress disorder on a standard scale
`csa`	Childhood sexual abuse − `Abused` or `NotAbused`

3.2. Specific research question

The research question is about a comparison of treatment means. Is there a difference in the population means of `ptsd` for the `Abused` and the `NotAbused`? Phrasing the research question in terms of causal effects—with cautious reservations: Is there a treatment effect of `csa` on `ptsd`?

For the comparison of population means of `ptsd`, or for estimating the treatment effect of `csa` on `ptsd`, notice that childhood sexual abuse (`csa`) is seen as an explanatory variable for scores on the response variable of post-traumatic stress disorder (`ptsd`). The explanatory variable `csa` is of categorical type (dichotomous)—it is a qualitative predictor. The response variable `ptsd` is a continuous numerical variable.

Regarding the research question from a statistical analysis perspective, we immediately think in terms of an analysis of variance (ANOVA) model, or a linear regression model for that matter. The question is whether the two groups (`Abused` and `NotAbused`) differ in the means of post-traumatic stress disorder (`ptsd`). Which is similar to the question: How well can we predict the score on `ptsd` if we know which group, the `Abused` or `NotAbused`, a subject belongs to?

3.3. Covariate or concomitant variable `cpa`

The continuous numerical variable of childhood physical abuse (`cpa`) is considered as a concomitant variable, a confounder or a covariate, for which we want to control in making comparisons between the two groups. That is where analysis of covariance ANCOVA comes into play. We might find an effect of `csa` on `ptsd`, but it is possible that the estimated effect is — to some degree, yet unknown — due to effects of reported childhood physical abuse (`cpa`). An analysis of covariance allows us to disentangle two competing explanations for the treatment effect to be estimated, as we will see below. By an ANCOVA we can, in principle, make proper adjustments of estimates of treatment effects as well, i.e., adjust for the confounding influence of the covariate `cpa` in mean-difference or causal effect estimation.

4. Looking at the data: The first thing to do

Data inspection is the foremost job in any statistical analysis, no matter the research question. So we better have a look at the data first.

```
> data(sexab)                    # getting the data frame 'sexab'
> attach(sexab)                  # attaching object names of 'sexab'
> names(sexab)                   # names of objects in 'sexab'
> sexab                          # list complete data frame

       cpa   ptsd       csa
  1   2.048  9.714    Abused
  2   0.839  6.169    Abused

  44  1.353  7.622    Abused
  45  5.119 11.128    Abused
  46  1.492  6.142 NotAbused
  47  0.610  0.745 NotAbused

  75  2.853  6.843 NotAbused
  76  0.811  7.129 NotAbused

> length(sexab)                  # number of variables
  [1] 3

> length(cpa)                    # number of respondents
  [1] 76
```

The output shows that we have a total number of $n = 76$ observations on the three variables, `cpa`, `ptsd` and `csa`. We can also observe that there are 45 subjects `[1:45]` in the `Abused` group ($n_A = 45$) and $76 - 45 = 31$ subjects `[46:76]` in the `NotAbused` group ($n_N = 31$).

4.1. Summary statistics

We can get summary statistics of the data for each of the two categories of the categorical variable `csa`, the two groups under comparison.

```
> by(sexab, sexab$csa, summary)           # data summary statistics

  sexab$csa: Abused
       cpa              ptsd              csa
  Min.   :-1.11   Min.   : 5.98   Abused   :45
  1st Qu.: 1.41   1st Qu.: 9.37   NotAbused: 0
  Median : 2.63   Median :11.31
  Mean   : 3.08   Mean   :11.94
  3rd Qu.: 4.32   3rd Qu.:14.90
  Max.   : 8.65   Max.   :18.99
  ------------------------------------------------------------
  sexab$csa: NotAbused
       cpa              ptsd              csa
  Min.   :-3.12   Min.   :-3.35   Abused   : 0
  1st Qu.:-0.23   1st Qu.: 3.54   NotAbused:31
  Median : 1.32   Median : 5.79
  Mean   : 1.31   Mean   : 4.70
  3rd Qu.: 2.83   3rd Qu.: 6.84
  Max.   : 5.05   Max.   :10.91
```

Notice the difference in means on `ptsd`: 11.94 versus 4.70, or the difference in medians on `ptsd`: 11.31 versus 5.79. Sample medians are more robust against outliers than sample means. There are differences in the sample frequency distributions of `cpa` too, which is rather important from a balancing point of view (Rosenbaum, 2004). If there were hardly any distributional differences between the two groups regarding these two variables, it would not make sense to continue the analysis.

4.2. Box plots

A box plot helps to give us a quick impression of the sample frequency distribution of the response variable `ptsd` and the covariate `cpa` in each group.

```
> plot(ptsd~csa, sexab)      # box plot 'ptsd' for each 'csa' group
> identify(ptsd~csa, n=2)    # identifying two outliers

> plot(cpa~csa, sexab)       # box plot 'cpa' for each 'csa' group
```

5. Relationship between covariate and response variable

It is also necessary to inspect the relationship between the covariate `cpa` and the response variable `ptsd`. If there is no relationship between a covariate and a response variable, there is no point in controlling for the covariate, or to incorporate it in a response model. We need to inspect the direction, size and form of the relationship between `cpa` and `ptsd` for each group separately.

5.1. Graphical inspection

We start by a graphical inspection of that relationship. To that purpose, each plotting point is indexed as either belonging to group `A` (`Abused`) or to group `N` (`NotAbused`). We can then inspect the scatterplot of `ptsd` and `cpa`, using the command

```
> plot(ptsd~cpa, pch=as.character(csa), sexab)
```

5.2. Correlation coefficients

Karl Pearson's product-moment correlation coefficient between the response variable `ptsd` and the control variable `cpa` can be calculated for all subjects, and within each group separately, for example as follows:

```
> cor(ptsd, cpa)                        # correlation coefficient r
  [1] 0.492

> cor(ptsd[1:45], cpa[1:45])            # r for the 'Abused'
  [1] 0.310

> cor(ptsd[46:76], cpa[46:76])          # r for the 'NotAbused'
  [1] 0.428
```

Recall that these correlations should be substantive for any covariate adjustment to be meaningful and effective (Cochran, 1983). Clearly, there is a substantive positive overall correlation here, and the estimated correlations in the two groups are not extremely different.

To simplify subsequent R syntax, we define a new treatment factor, T, which is equivalent to the factor variable csa, and we assign new, short level labels to that factor: A for Abused, and N for NotAbused.

```
> T <- sexab$csa                    # new treatment factor name 'T'
> levels(T)=c("A","N")              # new names of factor levels

> cor(ptsd, cpa)                    # correlation coefficient r
> cor(ptsd[T=="A"], cpa[T=="A"])    # r for the 'Abused'
> cor(ptsd[T=="N"], cpa[T=="N"])    # r for the 'NotAbused'
```

5.3. Regression coefficients

For the total group of respondents ($n = 76$) the relationship between ptsd and cpa can be illustrated by constructing a linear model for the response variable: $\texttt{ptsd}_i = \beta_0 + \beta_1 \texttt{cpa}_i + \epsilon_i$, for $i = 1, 2, \ldots, n$. In R this linear model equation is specified as lm(ptsd ~ cpa). We can write the results of the estimated model, for example to object rslt_t (short for: results for the total group), as follows:

```
> rslt_t <- lm(ptsd ~ cpa)     # linear model 'ptsd' for total group
> plot(cpa, ptsd)              # scatterplot of 'cpa' vs. 'ptsd'

> rslt_t                       # estimated regression coefficients

  Call:
  lm(formula = ptsd ~ cpa)

  Coefficients:                # unstandardized regr. coefficients
  (Intercept)          cpa
        6.55          1.03

> abline(rslt_t, col="green")  # plotting linear regression line
```

We can also do this for each group separately, after making proper arrangements for the plotting device.

```
> par(mfrow=c(2,1))                           # 2 plots in 1 device
> rslt_A <- lm(ptsd[T=="A"] ~ cpa[T=="A"])  # linear model 'Abused'
> rslt_A                                      # results linear model
```

```
  Call:
  lm(formula = ptsd[T=="A"] ~ cpa[T=="A"])

  Coefficients:
  (Intercept)    cpa[T=="A"]
        10.56           0.45

> plot(cpa[T=="A"], ptsd[T=="A"])     # scatterplot 'cpa' vs. 'ptsd'
> abline(rslt_A, col="red")           # regression line 'Abused'


> rslt_N <- lm(ptsd[T=="N"] ~ cpa[T=="N"])  # lin. model 'NotAbused'
> rslt_N

  Call:
  lm(formula = ptsd[T=="N"] ~ cpa[T=="N"])

  Coefficients:
  (Intercept)   cpa[T=="N"]
         3.70           0.76

> plot(cpa[T=="N"], ptsd[T=="N"])    # scatterplot 'cpa' vs. 'ptsd'
> abline(rslt_N, col="blue")         # regression line 'NotAbused'
```

Notice that the scales of the variables in the two plots are not in the same range, so it is difficult to see how similar the slopes of the regression lines are. The estimated slopes of 0.45 and 0.76 are not quite the same, for sure. (A statistical test for parallel regression lines will be introduced in Section 7.) Nevertheless, one plot showing both regression lines seems more appropriate. To that end, we first reset the plotting device to one graphical display.

```
> par(mfrow=c(1,1))                 # resetting plotting device
> plot(cpa, ptsd, pch=as.character(T))
> abline(rslt_A, col="red")         # regression line 'Abused'
> abline(rslt_N, col="blue")        # regression line 'NotAbused'
> abline(rslt_t, col="green")       # regression line total group
> identify(cpa, ptsd, n=3)          # identifying influential points
```

We can obtain more detailed statistical information on the results of linear model estimation by the general command `summary(res)`, as will be shown below.

6. Treatment effect: Difference in population means

We first perform a null hypothesis significance test to decide whether the population means of `ptsd` in the two populations, `Abused` and `NotAbused`, differ. We use a two-sample Student t-test, where the samples (`A` and `N`) are considered as two independent samples from two different populations. The null and the alternative hypothesis of this test can be described as $H_0: \ \mu_A = \mu_N$ and $H_1: \ \mu_A \neq \mu_N$, where μ_A is the population mean of the `Abused`, μ_N that of the `NotAbused`.

By default of the function `t.test`, it is assumed that the variances of `ptsd` in the two populations are unequal. In that case, a so-called Welch two-sample t-test is performed, which makes a correction to the number of degrees of freedom ($n_A + n_N - 2 = 74$) of Student's two-sample t-test; see, for example Hogg and Tanis (2001, p. 459ff.).

```
> t.test(ptsd[T=="A"], ptsd[T=="N"])

         Welch Two Sample t-test

  data:  ptsd[T=="A"] and ptsd[T=="N"]
  t = 8.9, df = 63.7, p-value = 8.803e-13
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   5.62 8.87
  sample estimates:
  mean of x mean of y
       11.9       4.7
```

Notice that the estimated 95% confidence interval for the (unadjusted) difference in population means does not cover the value of zero. Hence, the null hypothesis of a zero population difference is rejected, as is also obvious from the extremely small p-value of the test statistic.

We could, of course, have checked in advance how plausible it is to assume homogeneous variance of `ptsd` in the two populations by simply inspecting the variance or the standard deviation of `ptsd` in these two samples under study.

```
> sd(ptsd[T=="A"])     # standard deviation 'ptsd' for the 'Abused'
  [1] 3.44

> sd(ptsd[T=="N"])     # sd 'ptsd' for the 'NotAbused'
  [1] 3.52
```

From this result, the assumption of homogeneous variances seems plausible. The appropriate Student t-test could then be performed as follows:

```
> t.test(ptsd[T=="A"], ptsd[T=="N"], var.equal=TRUE)

        Two Sample t-test

data:  ptsd[T=="A"] and ptsd[T=="N"]
t = 8.94, df = 74, p-value = 2.172e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.63 8.86
sample estimates:
mean of x mean of y
     11.9       4.7
```

The substantive conclusions are the same for each of the two null hypothesis tests. The estimated 95% confidence interval for the (unadjusted) difference in means is just slightly narrower when variances of `ptsd` are assumed to be equal (an interval width of 3.23 versus 3.25).

The unadjusted treatment effect of `csa` on `ptsd` is estimated as the difference of the estimated population means of `ptsd`: $11.9 - 4.7 = 7.2$.

7. Linear models for the response variable `ptsd`

In principle we consider three linear models when doing an *analysis of covariance* for the two groups under study. For a general model description below, let Y be the response variable, X an explanatory variable, T a dummy variable indicating group membership or treatment (a qualitative, or categorical explanatory variable), and ϵ an error term. Below, the index i denotes the respondents ($i = 1, 2, ..., n_j$), and the index j group membership ($j = 0, 1$), where $j = 0$ refers to the `Abused` group, and $j = 1$ to the `NotAbused` group (sic!), and the number of respondents in the total group is denoted as $n = n_0 + n_1$.

With increasing simplicity the three linear models (cf. Tatsuoka, 1971, Chapter 3) can be expressed as follows, along with `R`'s general model formulation.

Model 1. Separate regression lines for each group with *different slopes*.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 T_j + \beta_3 X_{ij} T_j + \epsilon_{ij} \quad , \quad i = 1, 2, ..., n_j \; ; \quad j = 0, 1$$

```
> lm(Y ~ X + T + X:T)    # for short in R
```

Model 2. Separate regression lines for each group with the *same slope*.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 T_j + \epsilon_{ij} \quad , \quad i = 1, 2, ..., n_j \; ; \quad j = 0, 1$$

```
> lm(Y ~ X + T)          # for short in R
```

Model 3. The *same regression line* for both groups, i.e., for all respondents.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad , \quad i = 1, 2, ..., n$$

```
> lm(Y ~ X)              # for short in R
```

7.1. Model with interaction effect: Different slopes

As for our sample data, we first construct a linear model for the response variable `ptsd` with two main effects of `cpa` and `csa`, and an interaction effect of `cpa` with `csa`. Our Model 1 can then be specified as

$$\texttt{ptsd}_{ij} = \beta_0 + \beta_1 \texttt{cpa}_{ij} + \beta_2 \texttt{csa}_j + \beta_3 \texttt{cpa}_{ij} \texttt{csa}_j + \epsilon_{ij} \quad . \tag{1a}$$

Notice that $\texttt{csa}_j \equiv \texttt{T}_j$ is a dummy variable in this regression equation, indicating that a subject belongs either to the `Abused` group ($j = 0$) or to the `NotAbused` group ($j = 1$). In this regression model, we have separate regression lines for each group with different slopes and different intercepts. The interpretation of the effect of `csa` on `ptsd` then also depends on `cpa` (cf. Faraway, 2005, p. 168). Therefore, in such a model we cannot say that there is a homogeneous effect of `csa` on `ptsd`, *regardless* the degree of childhood physical abuse `cpa`.

For $j = 0$, the `Abused`, the dummy variable `csa` is coded as `0`, hence for this group Equation 1a reads

$$\texttt{ptsd}_{i0} = \beta_0 + \beta_1 \texttt{cpa}_{i0} + \epsilon_{i0} \quad , \quad i = 1, 2, \ldots, n_0 \quad . \tag{1b}$$

For $j = 1$, the `NotAbused`, `csa` is coded as `1`, and Equation 1a can be written and rearranged as

$$\begin{aligned} \texttt{ptsd}_{i1} &= \beta_0 + \beta_1 \texttt{cpa}_{i1} + \beta_2 + \beta_3 \texttt{cpa}_{i1} + \epsilon_{i1} \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \texttt{cpa}_{i1} + \epsilon_{i1} \quad , \quad i = 1, 2, \ldots, n_1 \quad . \end{aligned} \tag{1c}$$

Compare Equations 1b and 1c, and notice that we have both different intercepts and different slopes for the two population groups in Model 1. The first objective of our present analysis now is to investigate how plausible it is to assume that the slopes in the two populations are the same. The reason for doing so is that unconditional causal inference, i.e., irrespective the values of the covariate `cpa`, is only feasible when the slopes are the same.

Model 1 with the interaction term is estimated as follows:

```
> m1 <- lm(ptsd ~ cpa + csa + cpa:csa, sexab)              # Model 1
> summary(m1)

  Call:
  lm(formula = ptsd ~ cpa + csa + cpa:csa, data = sexab)
  Residuals:
     Min      1Q  Median      3Q     Max
  -8.200 -2.531 -0.181   2.774   6.975

  Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
  (Intercept)        10.557      0.806   13.09  < 2e-16 ***
  cpa                 0.450      0.208    2.16    0.034 *
  csaNotAbused       -6.861      1.075   -6.38  1.5e-08 ***
  cpa:csaNotAbused    0.314      0.368    0.85    0.397
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 3.28 on 72 degrees of freedom
  Multiple R-Squared: 0.583,      Adjusted R-squared: 0.565
  F-statistic: 33.5 on 3 and 72 DF,  p-value: 1.13e-13
```

It turns out that the interaction effect of `cpa` with `csa` is statistically not significant. Hence, the assumption of different regression slopes in the two groups does not seem to be plausible. That is great, because a necessary assumption for the analysis of covariance (ANCOVA) is not rejected.

7.2. Coding of the dummy or treatment variable `csa`

Because the dichotomous variable `csa` is nonnumeric, `R` automatically treats it as a categorical variable and sets up its coding. This coding is shown by inspection of the so-called *design matrix* $\mathbf{X}$ in the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

```
> model.matrix(m1)

   (Intercept)   cpa  csaNotAbused cpa:csaNotAbused
1            1 2.048            0            0.000
2            1 0.839            0            0.000

44           1 1.353            0            0.000
45           1 5.119            0            0.000
46           1 1.492            1            1.492
47           1 0.610            1            0.610

75           1 2.853            1            2.853
76           1 0.811            1            0.811
```

Notice from the design matrix extract that the `Abused` category is coded as `0`, and the `NotAbused` as `1`. The default choice is made alphabetically. This means that the `Abused` group is the *reference level* and that the model parameters (the regression coefficients) represent the *difference* between this reference level, the `Abused`, and the `NotAbused`. (In Section 9 we discuss how to change the reference level to the `NotAbused` group, if that would be convenient.)

The interaction term, `cpa:csaNotAbused`, is represented in the fourth column of the design matrix as the product of the numbers in columns 3 and 4, representing the basic terms of the interaction product values.

7.3. Model without an interaction effect: Parallel slopes

Since the interaction parameter of the linear model is not significant, we can simplify the model to one with main effects only. The interpretation of treatment effects is much simpler when we eliminate the slope interaction term because, from a modeling point of view, the effect of `csa` on the *expected difference* between the group means of `ptsd` is then no longer conditional on the value of the covariate `cpa`. The general linear equation of Model 2 reads

$$\texttt{ptsd}_{ij} = \beta_0 + \beta_1 \texttt{cpa}_{ij} + \beta_2 \texttt{csa}_j + \epsilon_{ij} \quad . \tag{2a}$$

For $j = 0$, the `Abused`, `csa` is coded as `0`, hence for this group Equation 2a simplifies to

$$\texttt{ptsd}_{i0} = \beta_0 + \beta_1 \texttt{cpa}_{i0} + \epsilon_{i0} \quad , \quad i = 1, 2, \ldots, n_0 \quad . \tag{2b}$$

For $j = 1$, the `Abused`, `csa` is coded as `1`, and Equation 2a can be written and rearranged as

$$\begin{aligned} \texttt{ptsd}_{i1} &= \beta_0 + \beta_1 \texttt{cpa}_{i1} + \beta_2 + \epsilon_{i1} \\ &= (\beta_0 + \beta_2) + \beta_1 \texttt{cpa}_{i1} + \epsilon_{i1} \quad , \quad i = 1, 2, \ldots, n_1 \quad . \end{aligned} \tag{2c}$$

Compare Equations 2b and 2c, and notice that the intercepts are different but the slopes are the same.

Model 2 is estimated by using the following `R` command:

```
> m2 <- lm(ptsd ~ cpa + csa, sexab)                          # Model 2
> summary(m2)

  Call:
  lm(formula = ptsd ~ cpa + csa, data = sexab)

  Residuals:
     Min      1Q  Median      3Q     Max
  -8.157 -2.364 -0.153   2.147   7.142

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept)    10.248      0.719   14.26  < 2e-16 ***
  cpa             0.551      0.172    3.21    0.002 **
  csaNotAbused   -6.273      0.822   -7.63  6.9e-11 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 3.27 on 73 degrees of freedom
  Multiple R-Squared: 0.579,      Adjusted R-squared: 0.567
  F-statistic: 50.1 on 2 and 73 DF,  p-value: 2e-14
```

We do not consider a further simplification of the linear model, by assuming the same regression line (slopes and intercepts) for each group. The reason for not considering Model 3: $\texttt{ptsd}_i = \beta_0 + \beta_1 \texttt{cpa}_i + \epsilon_i$, is that the parameters of the remaining explanatory variables in Model 2 are statistically significant. We conclude that there is a significant treatment effect of `csa` on `ptsd`, after having controlled for `cpa`. The question whether the effect is also practically significant,

i.e., substantively important, also needs an answer. What is the effect size of `csa` on `ptsd`, and is it substantial?

The difference of the intercepts between the groups is estimated as $b_2 \equiv \hat{\beta}_1 = -6.273$ [cf. Equation 2c], a substantial difference. That is, the intercept for `ptsd` in the `NotAbused` group is 6.273 *lower* than that in the `Abused` group, the one that served as the reference level. Hence the intercepts are 3.975 and 10.248, respectively. And the results for Model 2 in the table above show that from a statistically point of view this is a significant difference.

We conclude, therefore, that we observe both a significant and substantial treatment effect of `csa` on `ptsd`.

7.4. Plotting parallel regression lines

We can put two parallel regression lines (the same estimated slope $b_1 = 0.551$) in the plot of `ptsd` against `cpa`. The intercept in the `Abused` group equals 10.248, that in the `NotAbused` group $10.248 - 6.273 = 3.975$, as we just learned. We thus have two parallel regression lines with intercept parameter $b_0 = 10.248$ and slope $b_1 = 0.551$ in the `Abused` group, and intercept $b_0 = 3.975$ and slope $b_1 = 0.551$ in the `NotAbused` group.

```
> plot(ptsd~cpa, pch=as.character(csa))       # scatterplot
> abline(10.248, 0.551, col="red")            # regr. line 'Abused'
> abline(10.248-6.273, 0.551, col="blue")     # 'NotAbused'
> abline(rslt_t, col="green")                 # total group
```

Notice that the estimated *common regression slope*, usually denoted as b_w, is not equal to the estimated slope for the total group, b_t, which was estimated as $b_1 = 1.03$. The common slope is a *within-group estimate* of the regression slope: the slope parameter β_1 is estimated by calculations involving deviations from within-group sample means instead of deviations from overall sample means; see Equations 22.38 and 22.40 in Neter and Wasserman (1974, p. 708f.).

7.5. Adjusted differences in means

The estimate of the *unadjusted* (i.e., without controlling for the linear effect of `cpa` on `ptsd`) difference between the means (`Abused` minus `NotAbused`) from Student's t-test was $11.9 - 4.7 = 7.2$. Here, from the results of Model 2, we infer that the estimate for the *adjusted* (i.e., after controlling for the linear effect of `cpa` on `ptsd`) difference in means equals the rounded value of 6.3 (`Abused` − `NotAbused`: $10.248 - 3.975 = 6.273$).

Hence, after adjusting for the effect of childhood physical abuse `cpa`, the effect of childhood sexual abuse `csa` on the score of the post-traumatic disorder syndrome `ptsd` is slightly reduced.

As for the adjusted sample means of `ptsd` in each of the two groups, it should be noted that usually a general formula is used (see, for example, Cochran, 1983; Neter & Wasserman, 1974), in our case for $j = 0, 1$ specified as

$$\texttt{mean(ptsd}_j(adj)\texttt{)} = \texttt{mean(ptsd}_j\texttt{)} - b_w\texttt{(mean(cpa}_j\texttt{)} - \texttt{mean(cpa}_t\texttt{))} \ , \qquad (3a)$$

where $\texttt{mean(ptsd}_j\texttt{)}$ is the unadjusted sample mean of `ptsd` in group j, b_w is the estimated *common within-group slope*, and $\texttt{mean(cpa}_j\texttt{)}$ and $\texttt{mean(cpa}_t\texttt{)}$ are the sample means of `cpa` in group j and in the total group t (both groups combined), respectively.

Notice, that for getting the estimate of the adjusted treatment effect (the difference of two adjusted means in our case) we could also simply use

$$\texttt{mean(ptsd}_j(adj)\texttt{)} = \texttt{mean(ptsd}_j\texttt{)} - b_w\texttt{mean(cpa}_j\texttt{)} \ , \ j = 0, 1 \ , \qquad (3b)$$

as an estimator of the adjusted group means instead, leaving the common group term $b_w\,\texttt{mean(cpa}_t\texttt{)}$ out. The estimated adjusted treatment effect (again, the difference between the two adjusted means) would remain the same. This was implicitly the approach in the foregoing analysis.

In order to check this, we could make the following calculations, using Equation 3a first.

```
> coef(m2)                          # regr. coefficients of Model 2

   (Intercept)         cpa  csaNotAbused
        10.248       0.551        -6.273

> bw <- coef(m2)[2]                           # the common slope b_w

> bw
    cpa
  0.551

> ptsd_adj_0 <- mean(ptsd[T=="A"])-bw*(mean(cpa[T=="A"])-mean(cpa))
> ptsd_adj_1 <- mean(ptsd[T=="N"])-bw*(mean(cpa[T=="N"])-mean(cpa))
```

The adjusted sample means and their difference are then shown by

```
> c(ptsd_adj_0, ptsd_adj_1, (ptsd_adj_0 - ptsd_adj_1))

  11.54   5.27   6.27
```

And similarly, when using Equation 3b, we get

```
> ptsd_adj_0 <- mean(ptsd[T=="A"]) - bw*(mean(cpa[T=="A"]))
> ptsd_adj_1 <- mean(ptsd[T=="N"]) - bw*(mean(cpa[T=="N"]))

> c(ptsd_adj_0,  ptsd_adj_1, (ptsd_adj_0 - ptsd_adj_1))

  10.25   3.98   6.27

> bw*mean(cpa)               # the difference between (3a) and (3b)
    cpa
   1.30
```

Clearly, due to rounding $11.54 - 10.25 \approx 5.27 - 3.98 \approx 1.30$, and it was easily checked that this equals the value of b_w `mean(cpa`$_t$`)`.

7.6. Confidence interval for treatment effect

We can also estimate, for example, a 95% confidence interval for the *adjusted* causal effect of `csa` on `ptsd`.

```
> confint(m2)[3,]    # adjusted conf. interval for 'csaNotAbused'
                     # parameter in row 3 of the coefficient matrix

    2.5 % 97.5 %
  -7.91  -4.63
```

In the previous command line, `(m2)[3,]` refers to row 3 of the coefficient matrix, `summary(m2)`, which contains information on parameter `csaNotAbused`, estimated as -6.273.

We can compare this interval with the 95% confidence interval for the *unadjusted* causal effect of `csa` on `ptsd`, which was [5.63, 8.86], and after recoding [−8.86, −5.63]. In this particular case, the estimated confidence intervals have about the same width: 3.28 (adjusted) and 3.23 (unadjusted). In other cases, adjusting for a covariate can increase the precision of a causal effect estimate.

8. Regression diagnostics

The usual diagnostics of the regression model could and should be used to check the viability of regression model assumptions. Before jumping to any hasty substantive conclusion, it is worth inspecting, for example, whether there are residual differences related to the categorical variable `csa` regarding Model 2.

```
> plot(fitted(m2), residuals(m2), pch=as.character(csa),
       xlab="Fitted", ylab="Residuals")     # residual plot Model 2

> identify(fitted(m2), residuals(m2), n=3)  # identifying outliers
```

We can observe that there are no clear signs of heteroscedasticity of residuals. Since the two groups happen to be separated — more or less — we can also infer that the variation in the two groups is about the same. If this were not the case, we would need to make some adjustments to the analysis, possibly by using some weights.

The command `plot(m2)` produces a sequence of residual plots useful for diagnostic evaluations of Model 2, among which Cook's distance plot.

```
> plot(m2)                                  # residual diagnostic plots
```

9. Change of reference level

For convenience — ease of interpretation mainly — we could change the reference level, so that the `NotAbused` is the reference group ($j = 0$), and the `Abused` group is indexed as $j = 1$. This can be accomplished as follows:

```
> level(sexab$csa)
> sexab$csa <- relevel(sexab$csa, ref="NotAbused")
> level(sexab$csa)

> m2r <- lm(ptsd ~ cpa + csa, sexab)                      # Model 2r

> summary(m2r)

  Call:
  lm(formula = ptsd ~ cpa + csa, data = sexab)

  Residuals:
     Min      1Q  Median      3Q     Max
  -8.157 -2.364 -0.153   2.147   7.142
```

```
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)    3.975      0.629    6.32  1.9e-08 ***
  cpa            0.551      0.172    3.21    0.002 **
  csaAbused      6.273      0.822    7.63  6.9e-11 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 3.27 on 73 degrees of freedom
  Multiple R-Squared: 0.579,      Adjusted R-squared: 0.567
  F-statistic: 50.1 on 2 and 73 DF,  p-value: 2e-14

> confint(m2r)[3,]        # 95% confidence interval for 'csaAbused'

  2.5 % 97.5 %
   4.63   7.91
```

Although some of the coefficients have different numerical values, this coding leads to the same substantive conclusions as before.

10. Discussion

Faraday (2005) points out that childhood physical abuse might not be the only factor that is relevant to assessing the effects of childhood sexual abuse. It is quite possible, he writes, that the two groups differ according to other variables such as socio-economic status and age, and refers to Rodriguez et al. (1997). That could very well be: causal inference in observational studies is hardly ever without dispute.

References

Cochran, W.G. (1983). *Planning and analysis of observational studies* (L.E. Moses & F. Mosteller, Eds.). New York: Wiley.

Cohen, Y., & Cohen, J.Y. (2008). *Statistics and Data with R: An applied approach through examples*. Chichester: Wiley.

Field, A. (2006). *Discovering statistics using SPSS* (2nd ed). London: Sage.

Faraway, J.J. (2002). *Practical regression and anova using R*. Unpublished manuscript. Retrieved May 20, 2009, from `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`. [Data sets and scripts are also directly available from `http://www.maths.bath.ac.uk/~jjf23/book/`.]

Faraway, J.J. (2005). *Linear models with R.* Boca Raton, FL: Chapman & Hall/ CRC.

Faraway, J.J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models.* Boca Raton, FL: Chapman & Hall/CRC.

Hogg, R.V., & Tanis, E.A. (2001). *Probability and statistical inference* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Moore, D.S. & McCabe, G.P. (2004). *Introduction to the practice of statistics* (5th ed.). New York: Freeman.

Neter, J., & Wasserman, W. (1974). *Applied linear statistical models: Regression, analysis of variance, and experimental designs.* Homewood, IL: Irwin.

Rodriguez, N., Ryan, S.W., Vande Kemp, H., & Foy, D.W. (1997). Posttraumatic stress disorder in adult female survivors of childhood sexual abuse: A comparison study. *Journal of Consulting and Clinical Psychology*, **65**, 53–59.

Rosenbaum, P.R. (2004). Matching in observational studies. In A. Gelman & X-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family* (pp. 15–24). Chichester: Wiley.

Tatsuoka, M.M. (1971). *Multivariate analysis: Techniques for educational and psychological research.* New York: Wiley.

Verzani, J. (2005). *Using R for introductory statistics.* London: Chapman & Hall/CRC.