

565_HW5

Matthew Stoebe

2025-04-25

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

one_way_anova <- function(dat) {
  dat$trt <- factor(dat$trt)
  aov_mod <- aov(y ~ trt, data = dat)
  print(summary(aov_mod))
  print(TukeyHSD(aov_mod))
  invisible(aov_mod)
}

covariate_analysis <- function(dat, p=.05) {
  dat$trt <- factor(dat$trt)

  m_add <- lm(y ~ trt + age, data = dat)
  m_full <- lm(y ~ trt * age, data = dat)

  full_sum <- summary(m_full)
  int_rows <- grep(":", rownames(full_sum$coefficients)) # rows with "trt:age"
  int_pvals <- full_sum$coefficients[int_rows, "Pr(>|t|)"]

  if (any(int_pvals < p)) {
    cat("Significant interaction(s) detected; using full model.\n")
    print(full_sum)
    best <- m_full
  } else {
    cat("No significant interactions; using additive model.\n")
    print(summary(m_add))
    best <- m_add
  }
}
```

```

}

invisible(best)
}

plot_results <- function(dat, model) {
  dat$trt <- factor(dat$trt)
  grid <- expand.grid(
    age = seq(min(dat$age), max(dat$age), length.out = 100),
    trt = levels(dat$trt)
  )
  grid$pred <- predict(model, grid)

  library(ggplot2)
  ggplot(dat, aes(x = age, y = y, color = trt)) +
    geom_point() +
    geom_line(data = grid, aes(y = pred)) +
    labs(x = "Age", y = "Reduction in CRP (mg/dL)", color = "Treatment") +
    theme_minimal()
}

files <- c("Data/inflam1.txt", "Data/inflam2.txt", "Data/inflam3.txt")

```

Dataset 1

a

```

dat1 <- read.table(files[1], header = TRUE)
aov_m <- one_way_anova(dat1)

##           Df Sum Sq Mean Sq F value Pr(>F)
## trt         2  142.4   71.21    2.73 0.0715 .
## Residuals   77 2008.6   26.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = y ~ trt, data = dat)
##
## $trt
##           diff           lwr           upr           p adj
## 2-1  3.1791935 -0.1394955  6.497883  0.0633021
## 3-1  0.8117935 -2.4692716  4.092859  0.8251951
## 3-2 -2.3674000 -5.8555416  1.120742  0.2425868

```

For Dataset 1, we get a p value of .0715 meaning that we fail to reject our null hypothesis that there is no significant evidence of treatment differences at a .05 level.

b

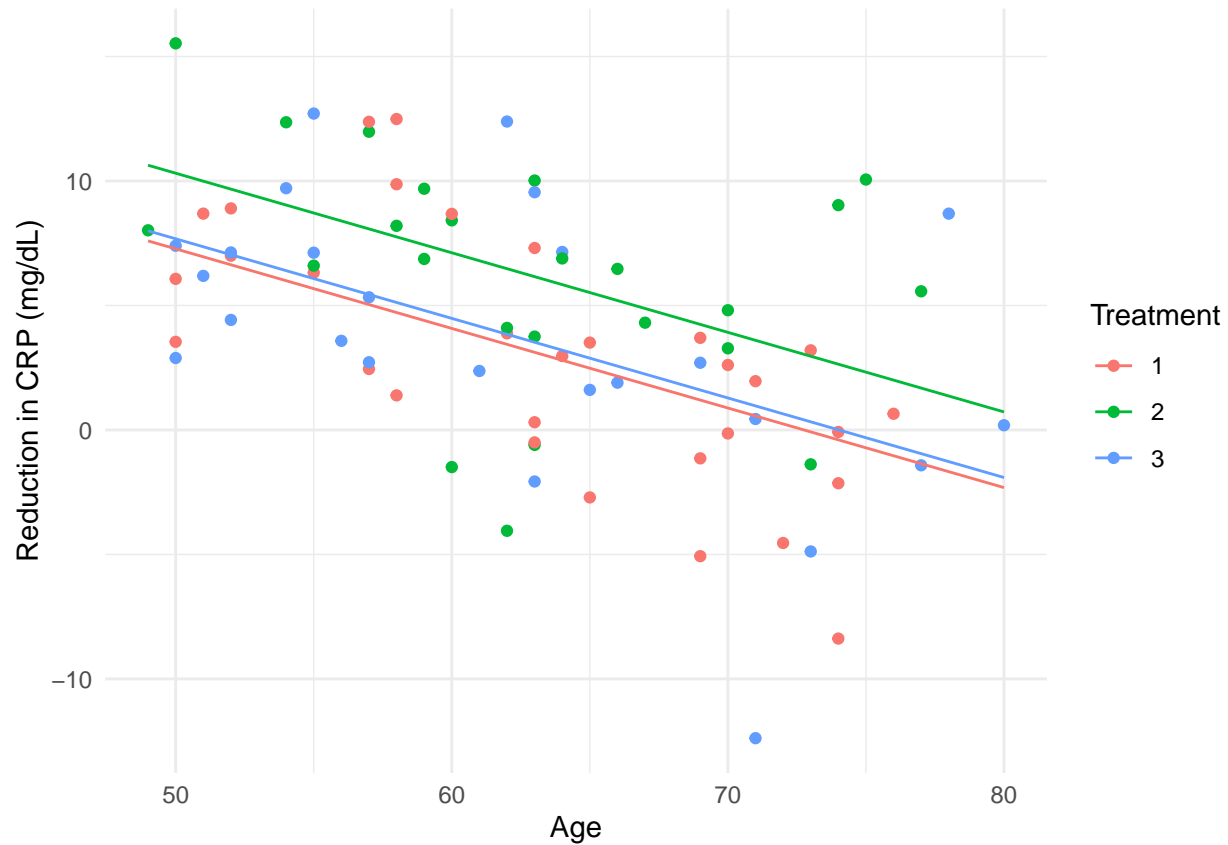
```
aov_m <- covariate_analysis(dat1)

## No significant interactions; using additive model.
##
## Call:
## lm(formula = y ~ trt + age, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3464  -2.4530   0.1304   2.3232   9.9611
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.25665    3.92040   5.932 8.30e-08 ***
## trt2         3.03914    1.19625   2.541  0.0131 *
## trt3         0.40430    1.18492   0.341  0.7339
## age        -0.31964    0.06061  -5.274 1.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.399 on 76 degrees of freedom
## Multiple R-squared:  0.3164, Adjusted R-squared:  0.2894
## F-statistic: 11.72 on 3 and 76 DF,  p-value: 2.144e-06
```

The interaction was not significant, So I relied on the additive model with treatment 1 as the reference. The age predictor is significant, and after adjusting for it, treatment 2 is the clear recommendable treatment as it has a positive estimate and is significant at a .05 level. treatment 3 is not different from treatment 1 after adjusting for age.

c

```
plot_results(dat1, aov_m)
```



d

Age is a strong negative predictor and is statistically significant. This tells us that older subjects on average experience less reduction in CRP. By adjusting for it, we improve the fit of our model, and protect against confounding in the case where age distributions vary across treatments.

Dataset 2

a

```
dat2 <- read.table(files[2], header = TRUE)
aov_m <- one_way_anova(dat2)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## trt        2  156.8    78.41   5.859 0.00428 **
## Residuals  77 1030.4    13.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = y ~ trt, data = dat)
```

```
##
## $trt
##      diff      lwr      upr    p adj
## 2-1 -2.8919195 -5.3960466 -0.3877924 0.0195504
## 3-1 -0.1036842 -2.7215955  2.5142271 0.9950732
## 3-2  2.7882353  0.5346278  5.0418428 0.0113712
```

For Dataset 2, we get a p value of .00428 meaning that we reject the null hypothesis that all treatment means are equal. When we do a pairwise comparison, it is clear that treatment 2-1 and 3-2 are significantly different with p values both below .05. Specifically, Treatment two performs significantly worse than both treatments 1 and 3 but treatments 1 and 3 do not differ.

b

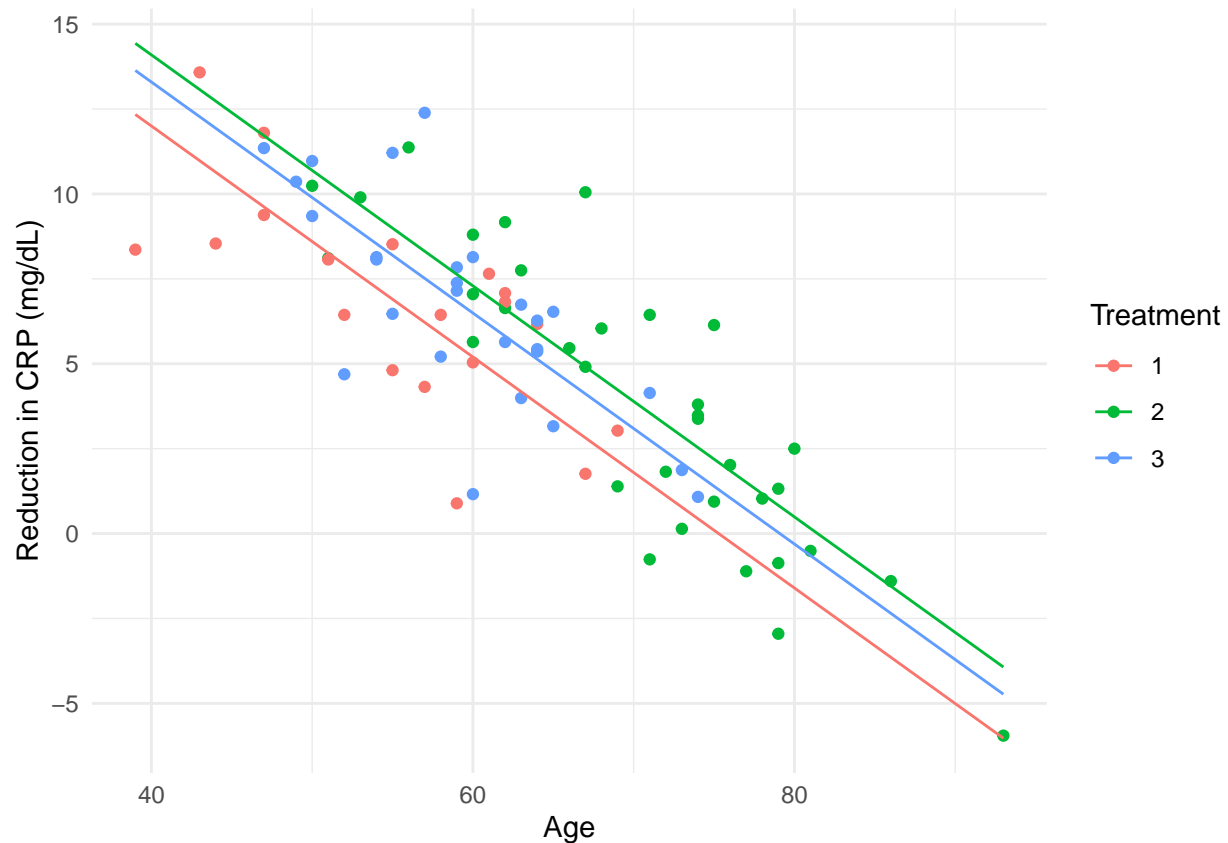
```
aov_m <- covariate_analysis(dat2)

## No significant interactions; using additive model.
##
## Call:
## lm(formula = y ~ trt + age, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.3337 -1.4803  0.1375  1.4075  5.1380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.60390     1.64080  15.605  < 2e-16 ***
## trt2         2.09413     0.74446   2.813  0.00624 **
## trt3         1.29512     0.65707   1.971  0.05236 .
## age        -0.34009     0.02825 -12.038  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.16 on 76 degrees of freedom
## Multiple R-squared:  0.7014, Adjusted R-squared:  0.6896
## F-statistic: 59.51 on 3 and 76 DF,  p-value: < 2.2e-16
```

There is no significant interaction so we use the additive model. Once again, age is significant, and after adjusting for age we see that Treatment 2 yields the largest and only significant boost in CRP reduction at the $p=.05$ level. As such, I would recommend treatment 2.

c

```
plot_results(dat2, aov_m)
```



d

In this model, age is once again added additive because it shows a significant relationship with CRP reduction. As such, we include it to improve the fit of our model and fix confounding. In this case, it actually flips the sign of the relationship removing the confounding

Dataset 3

a

```
dat3 <- read.table(files[3], header = TRUE)
aov_m <- one_way_anova(dat3)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trt         2   60.1   30.040    4.956 0.00945 **
## Residuals   77  466.7    6.061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = y ~ trt, data = dat)
```

```
##
## $trt
##          diff          lwr          upr          p adj
## 2-1 -1.8765184 -3.396521 -0.3565155 0.0115788
## 3-1 -1.6054839 -3.292961  0.0819932 0.0656248
## 3-2  0.2710345 -1.439110  1.9811785 0.9240512
```

For Dataset 3, we get a p value of 0.00945 meaning that there evidence against the null hypothesis that all treatment means are equal. When we do a pairwise comparison, it is clear that treatment 2 is significantly worse at reducing CRP than treatment 1 with a p value of .0116. treatments 1 - 3 and treatments 2-3 do not differ significantly although 1 - 3 is close with a p value of .0656

b

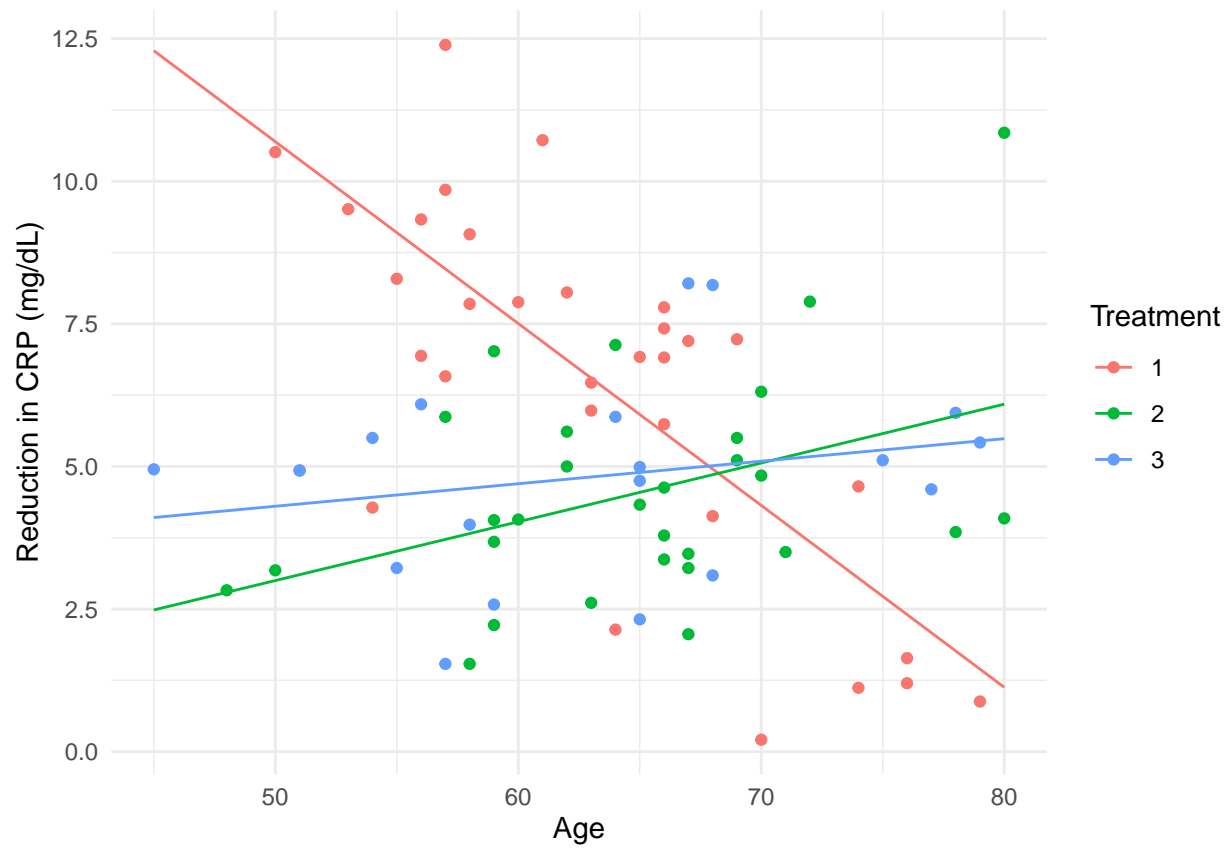
```
best_mod3 <- covariate_analysis(dat3)

## Significant interaction(s) detected; using full model.
##
## Call:
## lm(formula = y ~ trt * age, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.1390 -1.2794 -0.0234  1.0751  4.7577
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.63748    3.00549   8.863 3.02e-13 ***
## trt2         -28.79171    4.34280  -6.630 4.78e-09 ***
## trt3         -24.30898    4.18052  -5.815 1.44e-07 ***
## age          -0.31886    0.04707  -6.774 2.59e-09 ***
## trt2:age       0.42194    0.06720   6.279 2.10e-08 ***
## trt3:age       0.35834    0.06562   5.461 6.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.93 on 74 degrees of freedom
## Multiple R-squared:  0.4765, Adjusted R-squared:  0.4412
## F-statistic: 13.47 on 5 and 74 DF,  p-value: 2.424e-09
```

When looking at interactions, all interactions were significant at a .05 level so we keep the whole model. In this case, we choose our treatment based on the subjects age. This can be extracted from the coefficients but if you look at the plot below, younger subjects below the age of 70 do best receiving treatment 1, and older subjects over the age of 70 do best receiving treatment 2. This can also be derived by setting fit 1 equal to fit 3 and solving for age but manual inspection is sufficient.

c

```
plot_results(dat3, best_mod3)
```



d

there is a significant interaction meaning that treatment efficacy depends on age. without this interaction in the model, we would miss this relationship and fall victim to confounding. Including the interaction prevents masking this crossover and yeilds actionable age-specific guidance.