Partial correlation in linear models

Brice Ozenne

July 8, 2022

1 Summary

This document starts by presenting how to extract from a (univariate) linear regression model partial correlation coefficients. It also precise what type of "partial" (i.e. adjusted on which covariate) we get. When having multiple measurements of pairs of variables, various technics to estimate (partial) correlations are being compared.

2 Illustration

For illustration we will use the following packages:

```
library(LMMstar);library(mvtnorm);library(ggplot2);library(nlme)
LMMstar.options(method.numDeriv = "Richardson",
  columns.confint = c("estimate","se","statistic","df","p.value"))
```

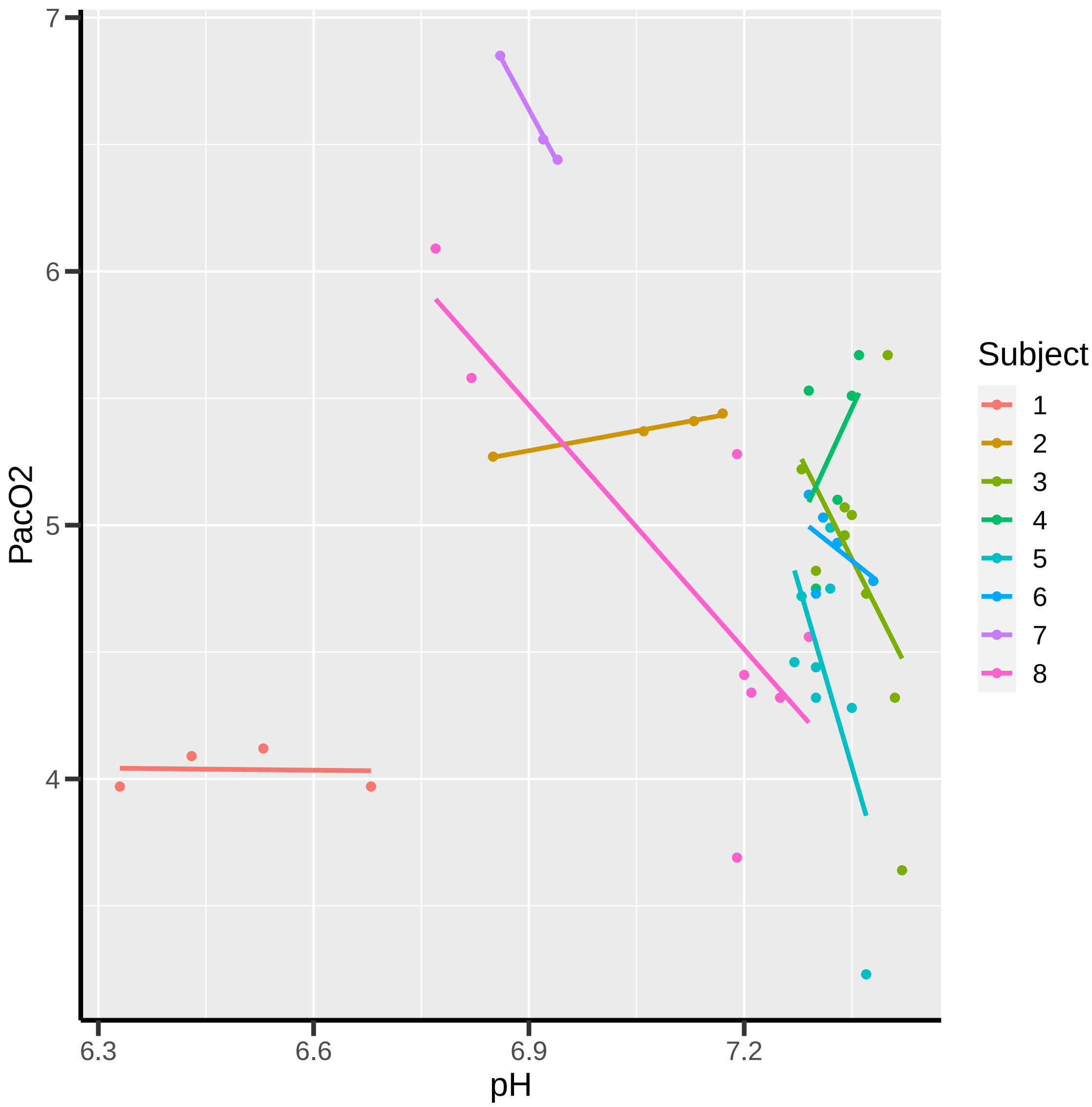
and dataset (Bland and Altman, 1995):

```
data("bland1995", package = "rmcorr")
bland1995$Subject <- as.factor(bland1995$Subject)
head(bland1995)
```

```
  Subject   pH PacO2
1       1 6.68  3.97
2       1 6.53  4.12
3       1 6.43  4.09
4       1 6.33  3.97
5       2 6.85  5.27
6       2 7.06  5.37
```

The aim is to relate intramural pH and PaCO2 using eight subjects:

```
gg <- ggplot(bland1995, aes(x = pH, y = PacO2,
       group = Subject, color = Subject))
gg <- gg + geom_point() + geom_smooth(method = "lm", se = FALSE)
gg
```



PacO2
7
6
5
4
6.3
6.6
6.9
7.2
pH
Subject
1
2
3
4
5
6
7
8

3 Partial partial in multiple linear regression

Consider the linear model:

```
e.lmm <- lmm(pH ~ Subject + PacO2, data = bland1995)
eTable.lmm <- confint(e.lmm)
eTable.lmm
```

```
             estimate     se statistic df  p.value
(Intercept)     6.930 0.1295     53.53 38 0.00e+00
Subject2        0.705 0.0774      9.11 38 4.28e-11
Subject3        0.950 0.0611     15.55 38 0.00e+00
Subject4        0.972 0.0735     13.22 38 8.88e-16
Subject5        0.860 0.0584     14.73 38 0.00e+00
Subject6        0.926 0.0660     14.04 38 0.00e+00
Subject7        0.692 0.1049      6.60 38 8.67e-08
Subject8        0.703 0.0616     11.42 38 7.46e-14
PacO2          -0.108 0.0299     -3.62 38 8.47e-04
```

The F-statistic testing the effect of each factor:

```
anova(e.lmm)
```

```
        || mean coefficients ||

 - Multivariate Wald test (global null hypothesis)
        F-statistic df.num df.denom    p.value
Subject      48.247      7       38          0 ***
PacO2        13.131      1       38 0.00084711 ***
---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

equal the Wald-statistic squared (divided by 1, the number of parameters)

```
Wald <- eTable.lmm["PacO2","statistic"]
Wald^2
```

```
[1] 13.13132
```

This F-statistic is also related to the sum of squares (ANOVA). Consider a model with a single regressor:

$$Y = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where we would have centered the outcome Y. Here we denote by X the design matrix, n the number of observations and $p = 1$ the number of coefficients, $H =$

$X(XX^\intercal)^{-1}X^\intercal$ the hat matrix and $\widehat{\beta} = (XX^\intercal)^{-1}X^\intercal Y$ the OLS estimator of the regression coefficients.

$$\begin{aligned}
\mathbb{V}ar(Y) = YY^\intercal =& YHY^\intercal + Y(1-H)Y^\intercal \\
SST =& SSR + SSE \\
=& \hat{\beta}(XX^\intercal)\hat{\beta}^\intercal + Y(1-H)Y^\intercal \\
=& \sigma^2(\hat{\beta}\Sigma_{\hat{\beta}}^{-1}\hat{\beta}^\intercal + n - p)
\end{aligned}$$

Introducing $MSSR = SSR/1$ and $MSSE = SSE/(n-p)$, we obtain that:

$$\frac{MSSR}{MSSE} = \frac{\hat{\beta}^2}{\Sigma_{\hat{\beta}}} = Wald^2$$

So the F-statistic equals the ratio of the residual sum of squared (normalized by their degrees of freedom). We can check that this extends to multiple regression using the usual anova table:

```
anova(lm(pH ~ Subject + PacO2, data = bland1995))
```

```
Analysis of Variance Table

Response: pH
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Subject    7 2.86484 0.40926  46.600 < 2.2e-16 ***
PacO2      1 0.11532 0.11532  13.131 0.0008471 ***
Residuals 38 0.33373 0.00878                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

⚠ Since **R** output type 1 anova only the last and second to last line are relevant. The first line (`Subject`) is for a model without `PacO2` so it should be expected that the F-value does not match with the one of `Subject` in a model with `PacO2`.

```
sigma2 <- as.double(sigma(e.lmm))
beta <- eTable.lmm["PacO2","estimate"]
sigma_beta <- eTable.lmm["PacO2","se"]
c(MSSE = sigma2, MSSR = sigma2 * beta^2 /sigma_beta^2)
```

```
       MSSE        MSSR 
0.008782435 0.115324959 
```

Finally the R^2 is defined as the proportion of variance explained, i.e.:

$$R^2 = \frac{SSR}{SSR + SSE}$$
$$= \frac{1}{1 + SSE/SSR}$$
$$= \frac{1}{1 + n - p/(\beta^2/\sigma^2_\beta)}$$
$$= \frac{Wald^2}{Wald^2 + n - p}$$

So the partial correlation coefficient is the square root of that quantity, with sign the sign of the test statistic:

```
df <- eTable.lmm["PacO2","df"]
sign(Wald)*sqrt(Wald^2/(Wald^2+df))
```

```
[1] -0.5067697
```

which matches exactly the partial correlation coefficient when **both** outcome are adjusted for `Subject`:

```
e.partialCor <- partialCor(list(pH ~ Subject, PacO2 ~ Subject), data =
    bland1995)
print(e.partialCor, digit = 5)
```

```
                estimate      se     df    lower    upper   p.value
rho(pH,PacO2) -0.50677 0.16013 25.669 -0.71027 -0.2251 0.0017756
Note: estimates and confidence intervals for rho have been back-transformed.
      standard errors are not back-transformed.
```

Similar values can be obtained using dedicated packages, e.g.:

```
library(rmcorr)
rmcorr(Subject, PacO2, pH, bland1995)$r
```

```
[1] -0.5067697
```

4 Partial partial with repeated measurements

Several references on the subject (Bland and Altman, 1995; Lipsitz et al., 2001; Bakdash and Marusich, 2017; Shan et al., 2020)

5 Reference

Bakdash, J. Z. and Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, 8:456.

Bland, J. M. and Altman, D. G. (1995). Calculating correlation coefficients with repeated observations: Part 2—correlation between subjects. *Bmj*, 310(6980):633.

Lipsitz, S. R., Leong, T., Ibrahim, J., and Lipshultz, S. (2001). A partial correlation coefficient and coefficient of determination for multivariate normal repeated measures data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1):87–95.

Shan, G., Zhang, H., and Jiang, T. (2020). Correlation coefficients for a study with repeated measures. *Computational and mathematical methods in medicine*, 2020.