Multivariable regression

Regression

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Consider the following simulated data

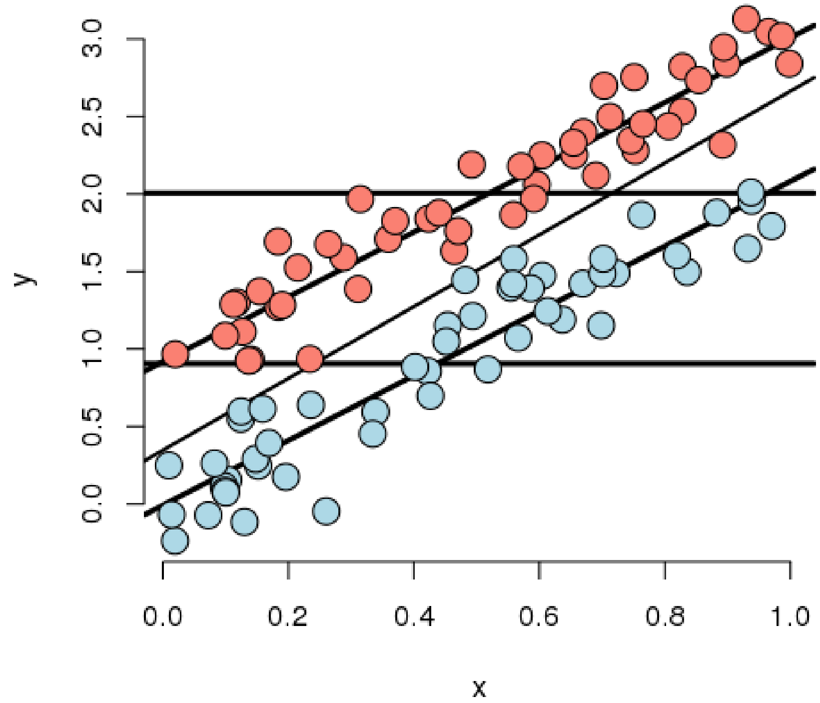
Code for the first plot, rest omitted (See the git repo for the rest of the code.)

t = [0, 0, 0, …(n/2 times), 1, 1, 1, … (n/2 times)]   << t for Treatment

```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```

mean of treatment = 0
mean of treatment = 1

Simulation 1




```
der Unterschied zwischen t=1
und t=0 ist ungefaehr gleich,
egal ob man x beruecksichtigt
oder nicht.

vgl Abstand dieser beiden
mean-Linien und obiger Abstand
der beiden Regressionslinien.
```

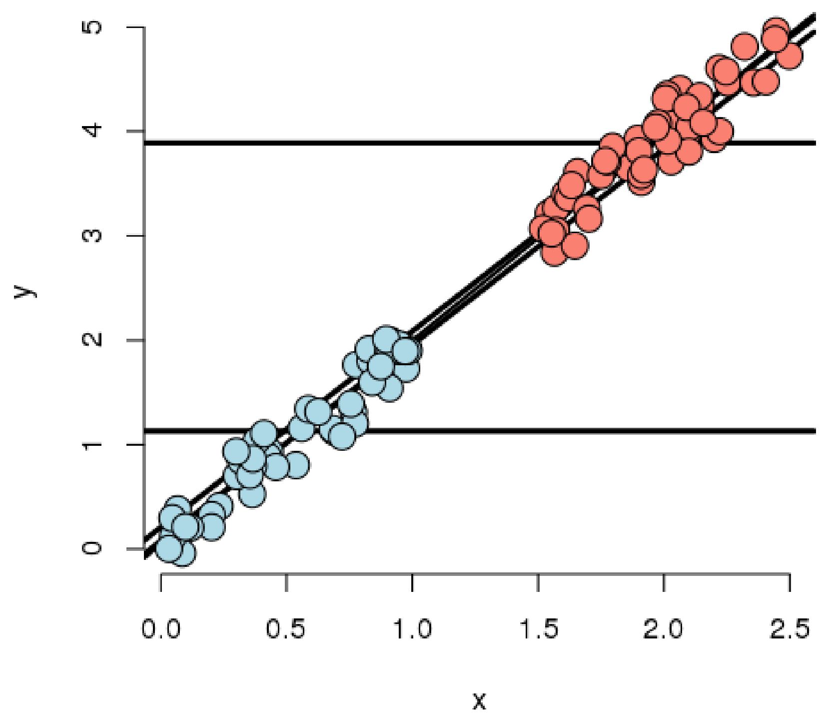
Discussion

Some things to note in this simulation

- The X variable is unrelated to group status dh. unabhaengig von X kommen etwa gleich viele t=0 und t=1 -Punkte vor (?)
- The X variable is related to Y, but the intercept depends on group status.
- The group variable is related to Y. t=0 (blau) hat niedrigeres Y als t=1 (rot)
  - The relationship between group status and Y is constant depending on X.
  - The relationship between group and Y disregarding X is about the same as holding X constant

    dh. der Abstand zwischen den Mean-Linien (die also den Einfluss von X nicht beruecksichtigen
    ist immer etwa gleich wie jener zwischen den beiden Regressionslinien
    (die also den Einfluss von X beruecksichtigen)

Simulation 2



Unterschied zw mean t=0 und mean t=1: massive

Aber wenn man den Einfluss von x beruecksichtigt, verschwindet er (Intercept und Slope fast gleich).

Discussion

Some things to note in this simulation

- The X variable is highly related to group status If I told you X, you'd know the group status (1 or 0)
- The X variable is related to Y, the intercept doesn't depend on the group variable.
  - The X variable remains related to Y holding group status constant
- The group variable is marginally related to Y disregarding X.
- The model would estimate no adjusted effect due to group.
  - There isn't any data to inform the relationship between group and Y.
  - This conclusion is entirely based on the model.

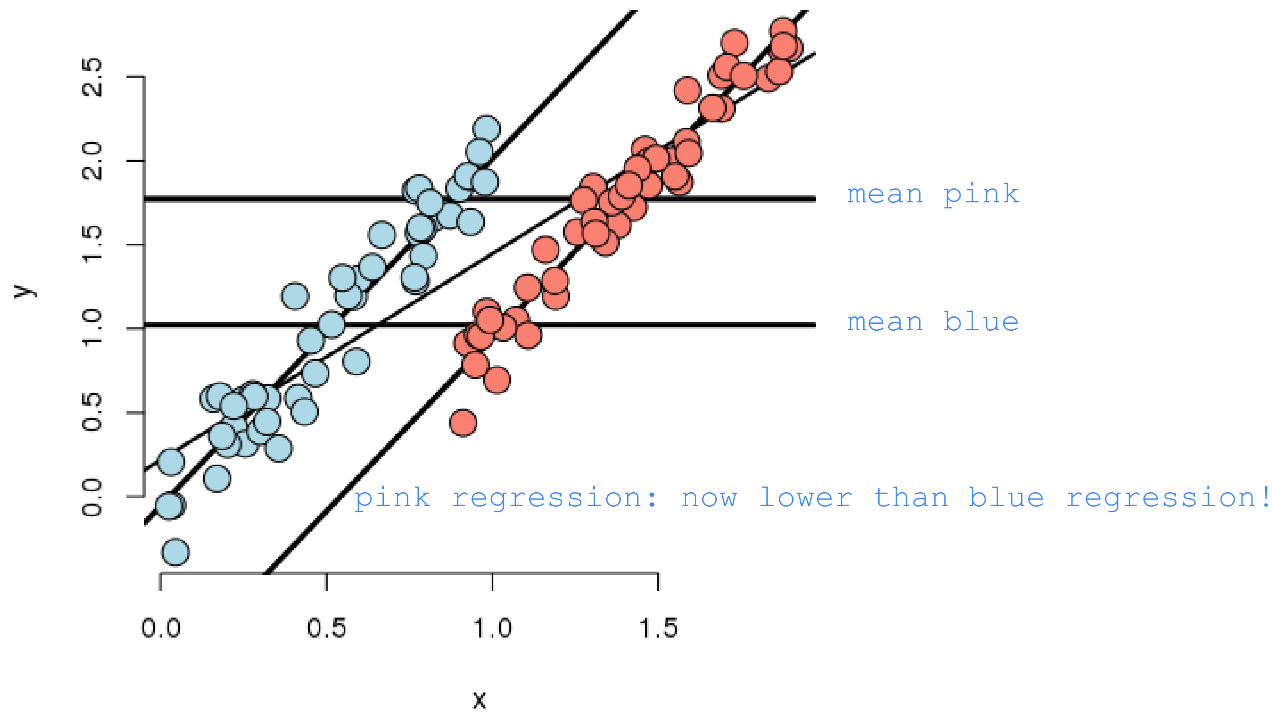
Allerdings haben wir fuer die rosa Gruppe keine Daten nahe beim Intercept, es haengt also voellig vom Modell ab.

So the group means are very different — if we ignore X!

The blue group seems to be linearly related to Y if you only look at it. Same for the pink group.

'adjusted' means considering X. Unadjusted (eg. only the group means): there is a huge effect.

Simulation 3



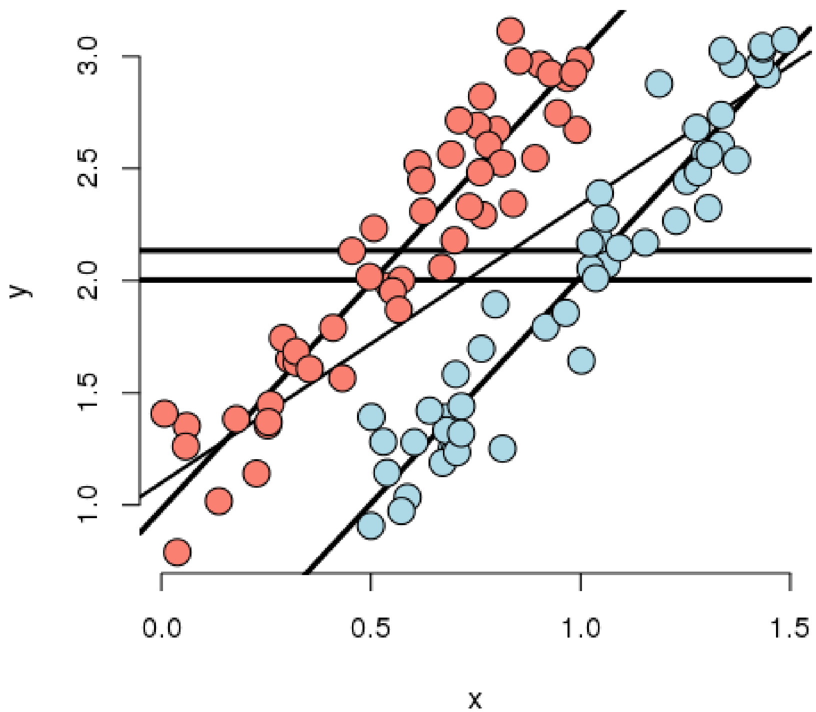
so: treatment effect reverses itself when you consider (“adjust for”) x!

Discussion

Some things to note in this simulation

- Marginal association has red group higher than blue.
- Adjusted relationship has blue group higher than red.
- Group status related to X.
- There is some direct evidence for comparing red and blue holding X fixed.

Simulation 4



hardly any difference between treatments when we don't adjust for x.

If we do: a clear difference.

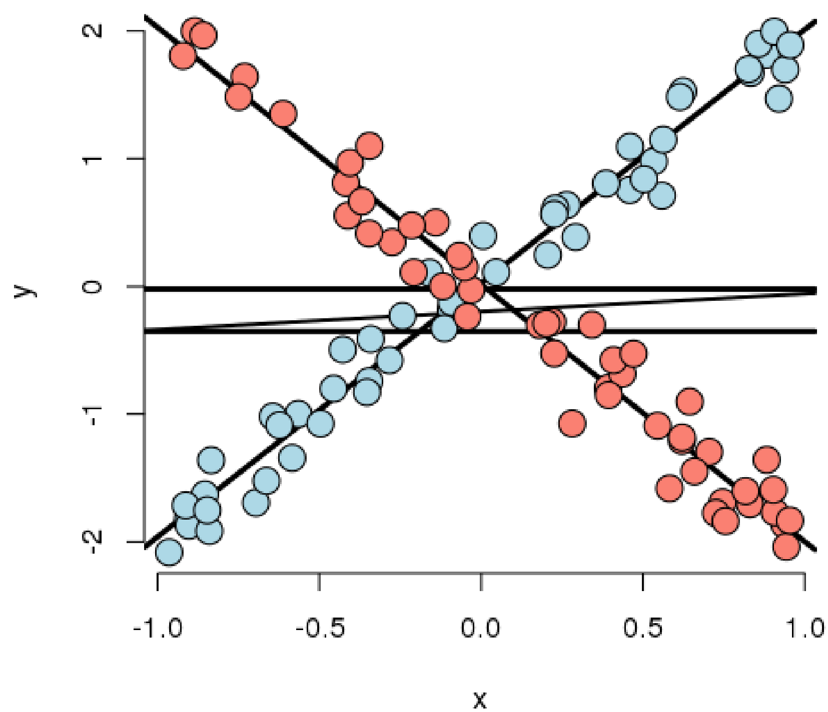
Discussion

Some things to note in this simulation

- No marginal association between group status and Y.
- Strong adjusted relationship.
- Group status not related to X.
- There is lots of direct evidence for comparing red and blue holding X fixed.

Simulation 5

An Interaction:



Interaction

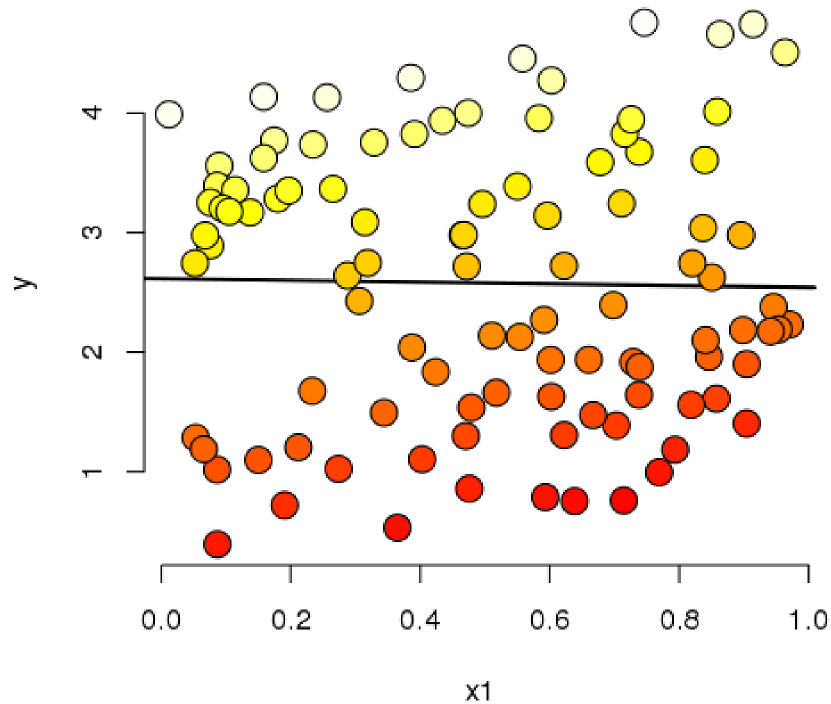
no treatment effect

Discussion

Some things to note from this simulation

- There is no such thing as a group effect here.
  - The impact of group reverses itself depending on X.
  - Both intercept and slope depends on group.
- Group status and X unrelated.
  - There's lots of information about group effects holding X fixed.

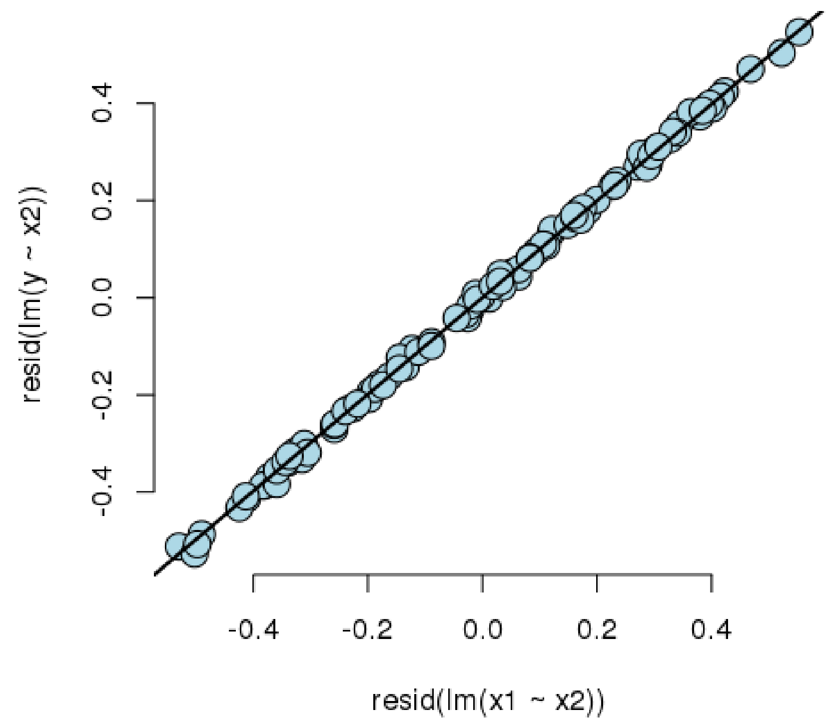
Simulation 6



Do this to investigate the bivariate relationship

```r
library(rgl)
plot3d(x1, x2, y)
```

Residual relationship



Discussion

Some things to note from this simulation

- X1 unrelated to X2
- X2 strongly related to Y

because

- Adjusted relationship between X1 and Y largely unchanged by considering X2.
    - Almost no residual variability after accounting for X2.

Some final thoughts

- Modeling multivariate relationships is difficult. `if you want to interpret them! Prediction alone is easier.`
- Play around with simulations to see how the inclusion or exclustion of another variable can change analyses.
- The results of these analyses deal with the impact of variables on associations.
  - Ascertaining mechanisms or cause are difficult subjects to be added on top of difficulty in understanding multivariate associations.

`causal interference is of course ‘difficult’`