Biostatistics 201a – Lecture 18

Contents

- Mediation
- Multicollinearity + associated problems
- VIFs and multic detection
- Multic fixes

pages = 16

Note: Exams returned today

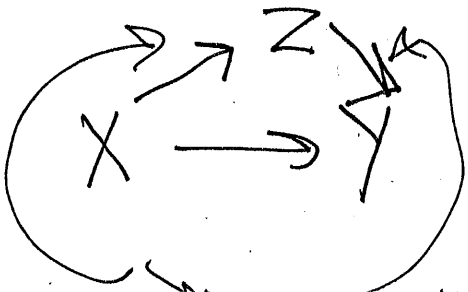
Mean ≈ 76
S.D. ≈ 13
Median ≈ 78

11/4/11
Lecture 18

More on Mediation + Multicollinearity

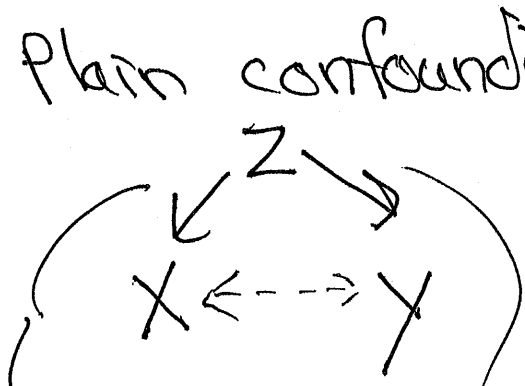
Mediation: X is related to Y through its effect on a "mediator" Z which is the direct cause of Y. Mediation is not exactly the same as plain confounding because there is a causal path implied

Mediation



indirect path from X to Y

Plain confounding



Z causes both X and Y or other possibilities

How do we check whether the data are consistent with mediation?

(1) Need X to be related to Y.
Check this w/ a SLR

(2) Need X to be related to Z.
Otherwise X's influence on Y can't go through Z
Check this w/ a SLR

(3) Need the relationship between X and Y to go away ("full mediation") or become less strong ("partial mediation") if Z is included in the regression of Y on X AND Z has to be significant in this model.

- What I've outlined here is called the "Baron-Kenney" approach to mediation.
- There's a formal test of the significance of the indirect path (X→Z→Y) with "Sobel's test"

Note: A successful test of mediation as I just described doesn't by itself prove the "causal" part of the picture. Doesn't account for feedback loops among variables either.

Confounding, adjustment + mediation are all special cases of the situation where two or more of our predictors are related. Idea was to figure out what the direction of causal paths that underlay those relationships might do to those relationships. But it's possible just to have 2 X variables with overlapping information about Y, neither of which causes the other or uniquely causes Y,

In general if two or more predictors are related to each other we say we have multicollinearity. The most common case is that X_1 and X_2 are linearly related but you can more generally have a whole group of inter-related X's. Multicollinearity can cause lots of problems:

(1) Interpretations of regression coefficients become complicated
(2) "Useful" variables may look insignificant
(3) Estimates of regression coefficients may become unstable, producing values that are unrealistic in either magnitude or sign.
(4) The effect of X_1 on Y may appear to change dramatically depending on whether the other X's are in the model
(5) Stand errors of the coefficients, the s_{b_j}'s - can become very

large, reflecting the uncertainty about which X's should get credit for explanatory power about Y (leads to #2)

The problems show up mostly in the X variables that are related to each other which can help spot the culprits

① Interpretation of coefficients

Y = heart rate after exercising

X_1 = how far you run

X_2 = how long you ran

Suppose people were on a treadmill set to run 3 mph

β_1 = slope for distance

= change in Y → rate associated with running an extra mile assuming you ran for the same amount of time (X_2 fixed) - uh oh. Can't fix X_2 and change X_1 at least not w/out changing speed.

(2) Significance of coefficients:

Individually we expect that ~~both~~ all of distance, time and speed would be good predictors of exercising O-rate But if we put them all in model we get the following

H_0: $\beta_1 = 0$ - distance explains nothing about O rate that isn't already accounted for by time, speed

H_A: $\beta_1 \neq 0$ - distance has unique explanatory power.

If I know time + speed I know distance run so it can't explain anything new. ⇒ Fail reject. Same holds for testing β_{time}, β_{speed}. All variables look useless

But the overall F test will be significant - the group of variables does provide info about O rate.

(3/5) Estimates and s.e. of the coefficients may become very strange.

Suppose that a "correct" version of the 0 rate model is the following.

$$Y = 75 + 5X_1$$

0 rate ↑ X_1 ↑ distance

Well if I'm making everyone go 3 mph and X_2 = time run

$X_1 = 3X_2$ = ~~0~~ speed × time

so $Y = 75 + 5(3X_2) = 75 + 15X_2$

is an equivalent model w/ $\beta_1 = 0$ and $\beta_2 = 15$

so is ~~$Y = 75 + 10X_1 - 15X_2$~~

$$Y = 75 + 5X_1 + (3X_2 - X_1)$$

$$= 75 + 4X_1 + 3X_2$$ ($3X_2 - X_1 = 0$)

or $Y = 75 + 5X_1 + (30X_2 - 10X_1)$

$= 75 - 5X_1 + 30X_2$

$\beta_1 = -5$ and $\beta_2 = 30$

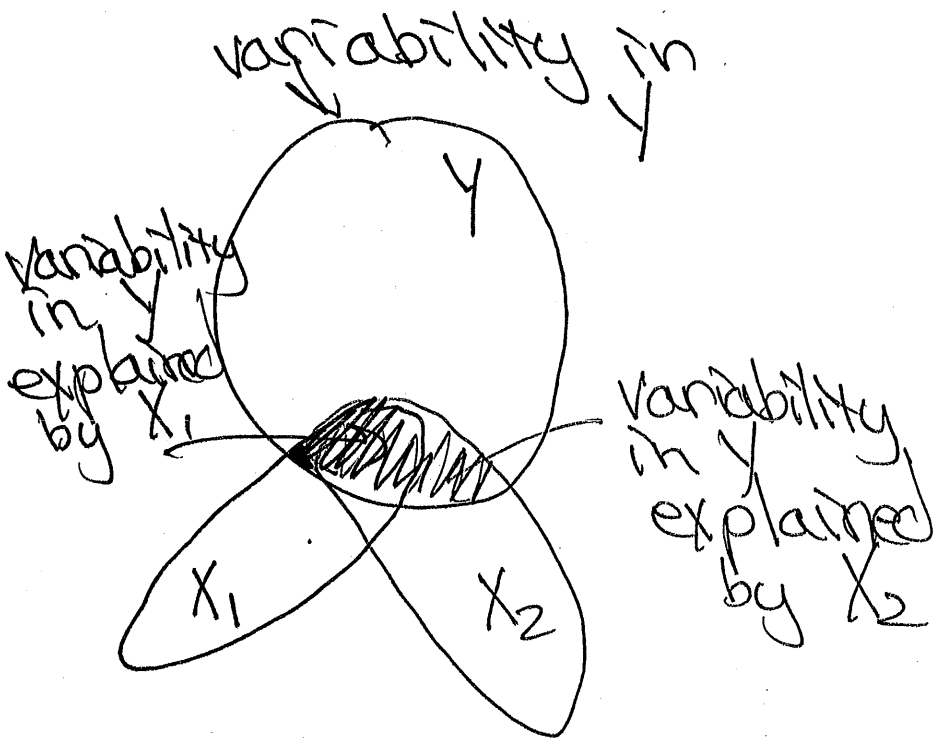
There are an infinite # of to equivalent models - in some you get only X_1 and X_2 is useless. In others X_2 is useful and X_1 is not. Worse in others you get negative coefficients ⇒ R² rate goes down as you run more. And magnitudes can get crazy. These sorts of situation provide good flags that multicollinearity has occurred. Problem from computer's point of view is it doesn't know which version to pick. If you put in perfectly related X's it will (a) choke or (b) remove one of the X's

of course you'll rarely put in things that are perfectly related but if it's close which of the many equivalent models you get is pure luck depending on random noise in your data.

How do we detect multic.?

(1) Look for any of the above problems - works pretty well especially if only 2 X's are related but a bit ad hoc and harder to figure out if multiple/complex relationships.

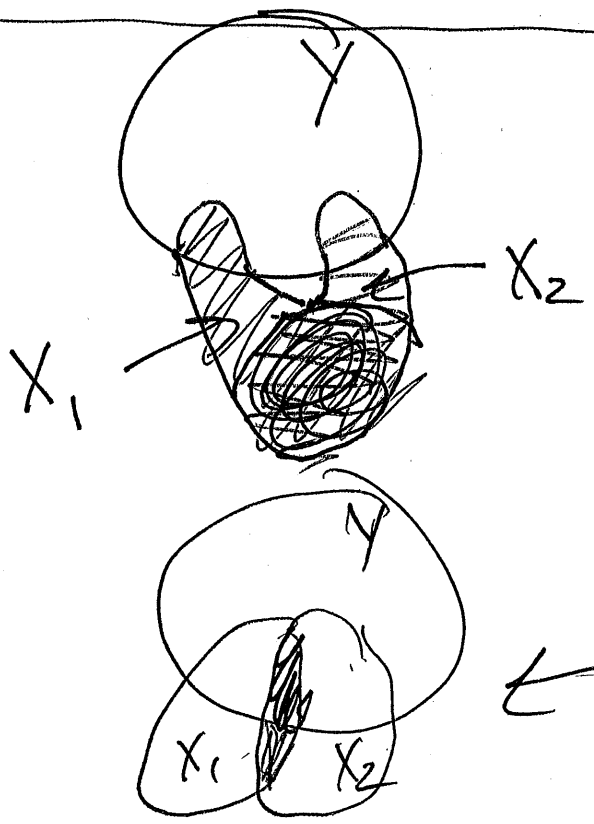
(2) Look at correlations among the X variables - can do this even before fitting M.LR. There are a few problems
- Only deals w/ pairwise relationships among X's.
- Not clear how big the correlations need to be for you to have a multicollinearity problem.



Here overlap in what X_1 and X_2 explain about Y is almost complete

Even though X_1 and X_2 in total don't overlap that much

Correlation between X_1 and X_2 won't be huge but we do have a multic. problem



Here, X_1 and X_2 explain completely separate things about Y even though their correlation is high.

← more standard

Variance Inflation Factors - Less adhoc check for multic.

Idea: If your X variables were completely uncorrelated then their coefficients and s.e.'s would not depend on the presence of other variables in the model. and there wouldn't be any uncertainty about which X should get credit for explaining a particular part of Y. On the other hand if the X's are related the standard errors will go up because there is uncertainty about the contributions. We can quantify this. For the estimated slope b_j of variable X_j, the standard error is

$$s_{b_j} = \sqrt{\frac{MSE}{\sum (X_{ij} - \bar{X}_j)^2} \cdot \frac{1}{1 - R_j^2}}$$

$\sum (X_{ij} - \bar{X}_j)^2$ → SSX_j variability in variable X_j

$\frac{1}{1 - R_j^2}$ ↑ this piece has to do w/ X relationships

Specifically R^2_j is the R^2 value you would get if you regress X_j on all the other X variables $(X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots X_p)$

The piece $\frac{1}{1-R^2_j}$ is called the <u>variance inflation factor</u> (VIF_j)

Because as there is more multic, $R^2_j \uparrow$ and so $1-R^2_j \downarrow$ and $VIF_j \uparrow$. If X_j is perfectly explained by the other variables $R^2_j = 1$ and the VIF is ∞.

What counts as a big VIF? There's no precise answer but your book suggests that values above 4 are bad and values above 10 are really bad. The important point is that the minimum VIF is 1 and higher is generally worse.

Aside: In SLR $\sqrt{R^2}$ is just the correlation between X and Y

It in MLR $\sqrt{R^2}$ called the "multiple correlation coefficient" and it turns out that it represents the correlation between the observed Y_i's and $\hat{Y}_i$'s, the values predicted by the model. The VIF is just the multiple correlation coefficient between X_j and the other predictors.

The whole VIF process gives you idea both whether there's a problem and which variables are causing it. If we fit what are called the "auxilliary" regressions of X_j on the rest of the X's and look for which are the significant predictors of X_j we get likely multic. families

How to fix multicollinearity:

(1) Remove some of your correlated X variables. Which ones?
- Base the choice on context - which variables are more interesting or important.
- Pick one(s) that produce the "best" model ($R^2 \uparrow$, RMSE $\downarrow$, stable, significant coefficients, ...)

(2) Rewrite the model in a new form - this works if you have what's called a structural multicollinearity caused by the way you defined special X variables - we'll see more of this later after we talk about curvilinear models and interactions ⇒ leads to centering of variables.

(3) Get additional data to reduce correlations - this works if multic. was a result of experimental design - e.g. in jogging example try to force more combos of speed/distance.

(4) Impose constraints on the model you fit: e.g.

- Ridge regression – which forces the b_j's to be closer to 0 not allowing them to get too inflated. Cost - introduces some bias; no longer have MLE
- principal components regression