# Statistical Analysis

**Why we choose to do Confounder first?**

If we first add the interaction into the model, the model reflects the influence of A, B and A\*B separately. And the influence of A on y alone will be impacted by A\*B. Therefore, testing interaction before testing confounding will cover the confounding influence, thus leading to the result that confounders are ignored. Therefore, we do confounder first.

* Adding potential confounders one by one is to find out which covariate could be a confounder. If adding one confounder into the model with “gender” results in scale of coefficients of “gender” change over 10%, we consider this covariate confounder.
* “Rank” issue. Lower than 10%. However, lots of articles support the fact that “rank” is a confounder.

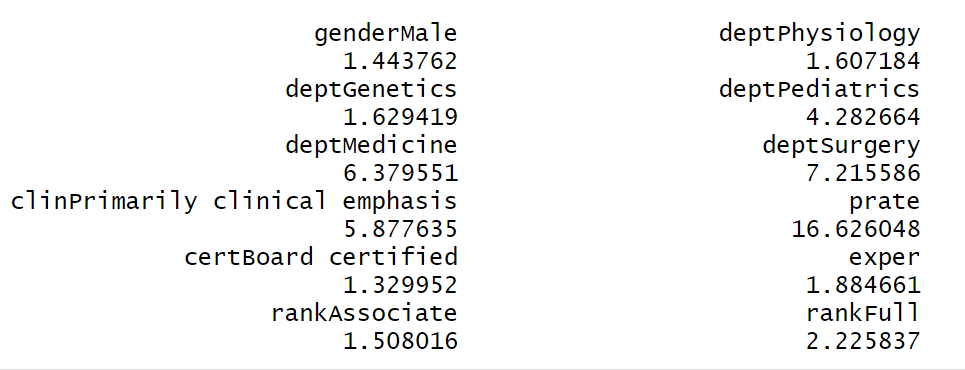
**Why we choose to put interaction one by one into the model, but not all in one time?**

Putting all the interactions between “gender” and other variables into the model once is not reasonable because coefficient of one interaction will be influenced by other interactions, therefore cover the “true” coefficient of one specific interaction.

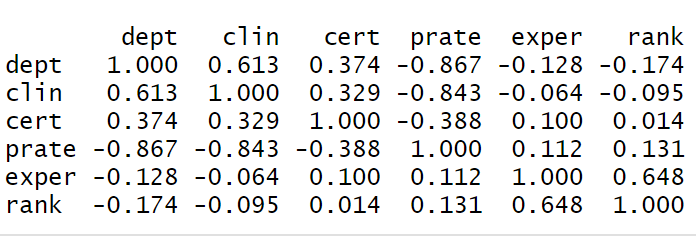
**What is the criteria for deciding multicollinearity?**

Do correlation coefficient first. Test if there is multicollinearity.

Do VIF:

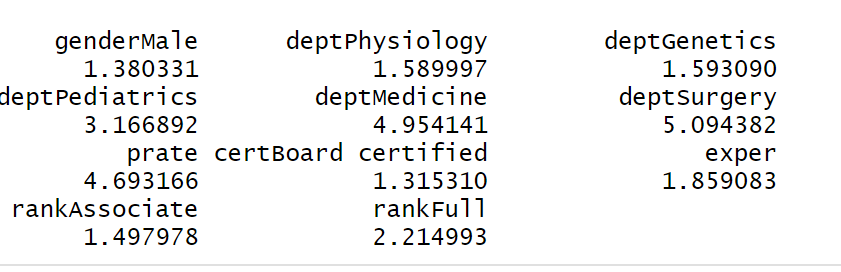


If there exists one VIF in the result that exceed 10, it indicates high collinearity in the model.

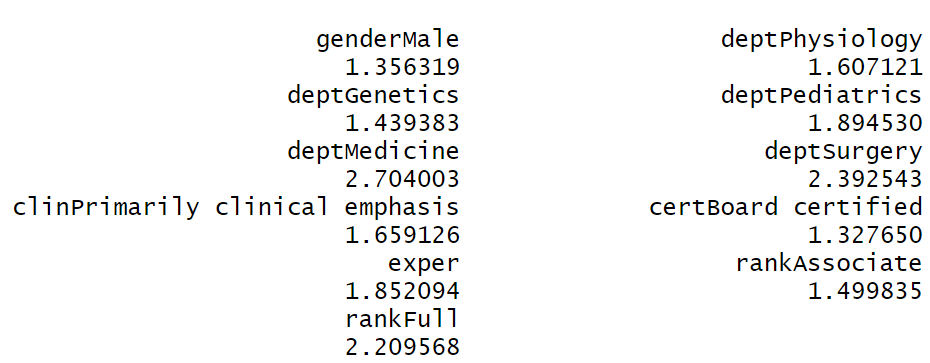


Prate\*dept, prate\*clin have high correlation.

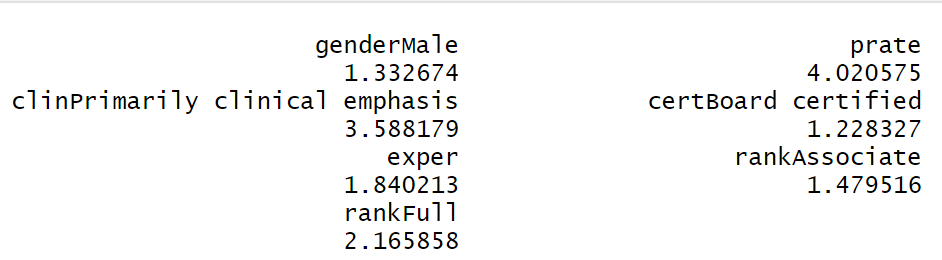
* Take prate out. Still exists one VIF larger than 5.



* Take clin out. No VIF larger than 5

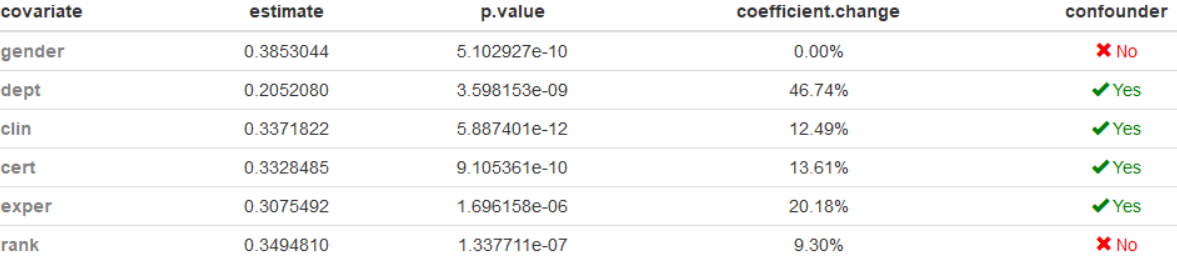


* Take dept out. No VIF larger than 5. However, The R.square decreases to 0.79.



\*\* Take Prate out for multicollinearity.

**What is the criteria for deciding confounder?**



* “Rank” is special. We **find lots of articles conclude that rank is a potential confounder.**

What is the criteria for interaction?



* Interpretation:……………..

The higher the rank, the difference between male and female in salary will get smaller.

Anova shows that lm(ln\_salavg ~ gender\*exper + dept + clin + cert + rank, data = data) is the most superior.