Bivin,

I hope this email finds you well and that you've got classrooms full of hardworking students.

I'm at work, now, and have something exciting that I could use your help on. In my position, I'm doing a lot of data pulls and reporting currently (as the company is playing catchup and moving from Access to SAS (seriously)). But, someone in Risk reached out to me for some help with his MLR. Long story short, I ended up suggesting he try PCA, but he's doing it in excel and I'd much rather take a stab at all of this in SAS. Is there any way you'd send me your PCA code? I'd look at mine but I killed my computer from school.

And I hate to admit that I always meant to go back and understand that code better, but spent so much time trying to learn how to interpret the results that I never did that. :/

Have a great week!

Jessica!

So good to hear from you and that it looks like you are having a blast!  And ... that you are using SAS and PCA / PCR!  Attached are the resources I have for PCA / PCR (Principal Component Regression which is just regression using the PCs.)  Since your class, I have upgraded the slides and the functionality presented in them.  Please check them out and let me know if you have any questions or need more materials.

proc prinqual is for using principal components with qualitative variables ... you might leave that one for later after you have reviewed yourself on PCA with continuous variables ... but it is useful!

Bivin

PS.  Would you mind if I showed your email to the current class?  I think it will really motivate them!

Bivin,

Thank you so much for that incredibly speedy reply! Of course you can show them my email! I'll follow up soon too and give you some more insight into what my discussion with him was like...really applicable to them in terms of assuring them that what they're studying and what you touch on is absolutely not a waste of their time.

Thanks!

Jessica

Bivin,

I better write this while it's fresh on my mind.

First, like I said before, I'm not doing much high level stuff *right now.* The analytics portion of the marketing department is very new. Right before I came they had hired an outside company to run decision tree analysis and PCA for market segmentation. Of course I inwardly groaned because I'm like, "I could have done that!" I've spent much of my time these three months helping with production and, especially, transferring all the things they do into SAS. For example, a data pull that takes my colleague well over four hours to do in Access queries I ran in SAS in 17 minutes. But this has taken much work and rework to validate and ensure others that it's trustworthy.

Point is that my bosses are very excited to do more in the future, but it might be six months down the road. I say that for your students for two reasons: 1, job roles can change, and even for the better! (Albeit harder). 2, even if you are doing less crazy things, if you develop working relationships across departments you can get your hands on some other good work. On that note,   
I've actually been working with the VP (head of Risk) on a new initiative due to this; just asked to meet with him and pick his brain the first week I was here. Next thing I knew I was working with him to roll out a new product.

Anyway, another guy from Risk just asked me if I had experience reading residual plots; that he had some issues that he couldn't account for in his model. I'll just outline the major things we talked about. Note this guy definitely has a stats background; graduated probably 15 years ago.

1) He was looking at standardized residuals vs predicted values and seeing quite a few with values > 2 and < -2, which I said indicated we needed to look at them as outliers...but also that we should be looking at leverage since we really care about whether or not the observation has much *influence.* See all the terms, there?! Remind the students that they matter. Specifically we also looked at Cook's D. So here, if you don't know what you're looking for, this means nothing. He forgot what kind of Cook's D we were looking for. So do you go with close to 1? 4/n? 2p/n? And obviously that brings up more questions.

2) Do you just take those outliers out? If there seem to be many (which frighteningly here there were), do you model those and see if they, as a group, behave similar to each other but entirely different than the rest of the obs? A subpopulation that we can make decisions on differently?

3) Of course this is where I get clued in to the fact that no assumptions were checked and no scatterplots viewed...so I'm asking to look at some of those and seeing exactly what I expected-lots of variables highly correlated with each other. Can we check them all? No? So maybe PCA? We're talking demographic data here, so PCA could really produce some sensible components, here; ones we could really interpret for a board.

4) There are things that you, fresh out of school, can remind even your highly intelligent seasoned colleague. He was incredibly appreciative this morning and, even though I was thinking before I went, "What on earth can I add to what he already knows??", I was wrong. Know what you know and don't be afraid to put it on the table.

That covers most of it, I think. Better get back to work!!

-Jessica

\*Of course, feel free to share all.