**Regulation Techniques for Multicollinearty: LASSO, Ridge, and Elastic Net**

Questions & Answers from the Author (Deanna Schreiber-Gregory)

BASUG Virtual Meeting: 2020.07.22

* **Question**: Do we need a continuous outcome to get those multicollinearity diagnostics, or is there some way to do this with a binary outcome, as well?
	+ **Attendee Answer (Hanna):** PROC HPGENSELECT will run LASSO with a binary outcome
	+ **Author Answer**: You do not need a continuous outcome to run the multicollinearity diagnostics. In the diagnostic procedure, you are only concerned about the relationship between the predictors/covariates themselves, not the relationship between the predictors and the outcome variable.
* **Question**: Can you trust the estimated parameters on the other variables?
	+ **Author Answer**: If multicollinearity is present in the model, I would be suspicious of the entire model as a whole. Until multicollinearity is corrected, I would not trust the parameter estimates of the other variables. It is an assumption violation and should be treated as such.
* **Question**: Could you explain what "constraints" mean on the previous slide?
	+ **Author Answer**: Constraints are referring to the penalties applied by LASSO and Ridge Regression. In Elastic Net, an additional constraint is considered on the implementation of LASSO and Ridge to make sure that the individual penalties applied through these procedures are balanced.
* **Question**: Could you expand on the bias from Ridge or LASSO?
	+ **Author Answer**: Within Ridge, especially, since you are applying a theoretical method that limits how much a variable can influence the model, this, by definition, is introducing a bias. You are essentially manually impacting the model to say what you want it to say (in this case, you want it to hide the relationship between the different predictor variables). It is addressing the negative impact of collinearity, but you are introducing a necessary bias in order to do so.
* **Question**: Can we also use forward and backward selection regressions to deal with multicollinearity?
	+ **Attendee Answer (Gagnon)**: Collinearity with selection can have variables continually bouncing in and out in stepwise models
	+ **Author Answer**: If your model/question allows for an alternative selection technique (such as forward and backward selection), then it isn’t a bad option. Just keep in mind that the implementation of such a technique will be evaluating the fit of the models (and picking the “best” model) with unrestrained collinear variables. This could mean you might end up with a model that excludes some variables that would have otherwise been meaningful if you had approached the collinearity in another way (basically, you are considering that LASSO is a selection technique and are just choosing a different methodology).
* **Question**: I have some predictors working in Higher Ed such as GPA and class grades that are definitely highly correlated. Are there some tips for deciding which variable to drop (when possible)?
	+ **Author Answer**: This question would require a very context-specific and in-depth answer. As someone who also works with education data, I would strongly recommend making sure that all variables to be included in the model have a solid set of reasoning behind their inclusion. If not, a “fishing expedition” type of model selection is what will result, and this isn’t a very sound way of approaching a statistical problem. In some cases, GPA may be the best thing to include, in others, class grades may be the best. I would want to see the question and the data to make an informed decision. The model will try to approach it from a computational and data-centered standpoint, but that doesn’t mean that the result will make sense!
* **Question**: Are you able to specify a seed value to ensure the GLMSELECT proc gives the same results each time you run it?
	+ **Attendee Answer (Yemi)**: The cv=random method plots the dataset randomly each time so you're likely to get different estimates.
	+ **Author Answer**: Yes! Through seed= option.
* **Question**: Did you choose the order of effects or does SAS?
	+ **Author Answer**: SAS chose the order of effects based on the methodology of the GLMSELECT procedure and the identified options.
* **Question**: Is post-selection inference available in SAS? i.e. p-value for the final variables selected by the model. There are some implementations in R.
	+ **Author Answer**: I have never used this, but I would assume that this is the case. I strongly recommend checking out the SELECT procedures (GLMSELECT, HPGENSELECT, and QUANTSELECT) to see if they match what you are looking for!
* **Question**: Can these methods be used for repeated measures (and are there SAS procedures to do it)?
	+ **Author Answer**: Yes, just make sure your data and question fit the need for multicollinearity detection! There are several SAS books written for clinical trials, so I would recommend exploring them to see which procedures they recommend to use. SAS is well equipped for repeated measures!
* **Question**: How does LASSO behave if only associated (determined by univar analysis) variables are put into model?
	+ **Author Answer**: If all of the variables within the model are collinear with each other, LASSO will drop all variables except for one. It does not perform well with grouped selection. If you are interested in allowing the model to drop a variable, while keeping in as many as you can, I recommend using Elastic Net or another variable selection procedure to limit the number of variables going into your model.
* **Question**: Are you worried that using the coefficients as a basis for evaluating variables--without regard for their standard errors--might bias you to choose models with a bit too many non-significiant large effects and a bit too few significant small effects?
	+ **Author Answer**: I agree that this can be a concern. Multicollinearity is an assumption violation of the most frequently used regression models and needs to be controlled for. A model can not be used if its assumptions are violated, so avoidance of multicollinearity detection and correction is not really possible in many models. The procedures presented in this paper are common and effective approaches to multicollinearity correction. If this is a major concern for a specific model, and multicollinearity is present, I would recommend exploring another modeling technique that does not require a multicollinearity adjustment, but that still fits the assumptions and structure of your data and question.
* **Question**: The betas are standardized in the penalty, aren't they?
	+ **Author Answer**: Correct!
* **Question**: I have a subjective question. Do you know how fast SAS runs these computations? What is the dimension of the largest data you've ran through these algorithms and how long did it take? Do you remember?
	+ **Author Answer**: These computations ran fairly fast for me, but I have not had to run these applications through large datasets. For a standard sized dataset in my work (<2000 observations and <30 variables), the procedures took under a minute each to run.
* **Question**: What would you do for logistic regression for multicollinearity, I assume we can't use LASSO or Ridge, and we can only use PCA, or drop one variable?
	+ **Author Answer**: You can use LASSO, Ridge, Elastic Net, PCA, and Latent Structure Analyses with Logistic Regression. You are not limited. It’s just a little trickier with the set-up!
* **Question**: Did the table you just showed, have VIF values or the parameter estimates?
	+ **Author Answer**: Yes, the last table produced through Ridge Regression shows both the original and adjusted VIF values as well as the parameter estimates. If you specify the outseb option, you can also get the standard errors. Pay close attention to the \_Type\_ and \_Ridge\_ columns to identify which adjustment the particular VIF, parameter estimate, and standard error are connect to.
* **Question**: If we have to remove one of the variable versus other in logistic model, can we use Elastic Net in logistic model?
	+ **Author Answer**: You can absolutely use Elastic Net with a logistic model. Make sure you are employing this procedure carefully so as not to accidentally produce a naïve model.
* **Question**: When you say "adjust" Elastic Net, how would we do that?
	+ **Author Answer**: You are able to specify the different Elastic Net methodological applications through the different options available in GLMSELECT. You can also specify a specific “k” for Ridge if you have one in mind. The procedure itself is just as powerful and flexible as the Elastic Net methodology itself. I would make sure you are well versed in the assumptions and limitations of Elastic Net before its use.
* **Question**: Could you please share some documentation on different methods within Elastic Net? I can separately follow up on this.
	+ **Author Answer**: Elastic Net is considered a machine learning procedure, so there is a good number of reputable books and papers that cover its theoretical approach. I strongly recommend papers such as Zou and Hastie’s “Regularization and variable selection via Elastic Net”. Forums are also a good place to find information on different approaches to Elastic Net, but be careful as they are only as reliable as the individual giving the advice (much like this presentation!). I also strongly recommend pooling information from peer-reviewed papers and statistics books (especially those that look at regularization techniques on the whole, and not just Elastic Net). Powerful procedures can be tricky to use, but employed correctly, they yield strong results.
* **Great Note from an Attendee (Yemi)**: I think it is noteworthy that there are situations where it is of interest to select variables that are correlated to each other in your final model. For example, this is the case in analyses of large gene expression data. These algorithms are still valid and there are extensions to them that allows you to achieve this purpose. A quite popular extensions to these 3 algorithms is Adaptive Ridge, Adaptive LASSO, and Adaptive Elastic Net models.