From: Patrick Hall [mailto:phall@0xdata.com]
Sent: Wednesday, February 01, 2017 1:27 PM

To: Wang, Youcai [GCB-CARDS]
Cc: Amy Wang; Kerry O'Shea (kerry@h2o.ai); Li, Peng4 [GCB-CARDS]
Subject: Re: variable inflation and Collinearity

Hi Youcai,

We are making good progress on how to replicate these features in h2o. We are meeting with the h2o.glm developer later today and will get back to you after that.

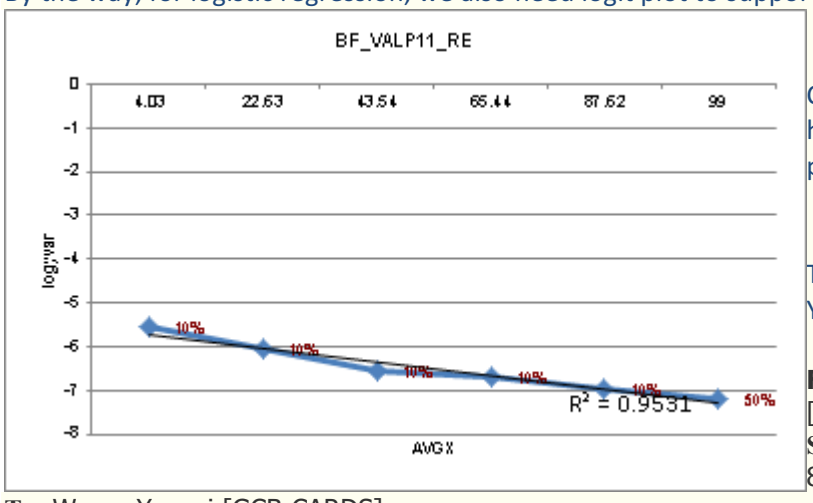
Thanks for patience.
p

--
Patrick Hall
phall@h2o.ai
(336) 693 4481

On Tue, Jan 31, 2017 at 9:39 PM, Wang, Youcai <youcai.wang@citi.com> wrote:
Hi Patrick,

I will provide your reasons when I need to answer model governance's ask, but do you have the new approach to measure the Collinearity?
You know prog reg is not for logistic regression process.
Proc reg just give us VIF and collinearity and which variable are main drivers.

By the way, for logistic regression, we also need logit plot to support the logistic regression as below:



Can you please check whether H2O have this functionality or not? If yes, please provide the application.

Thanks,
Youcai

From: Patrick Hall [mailto:phall@0xdata.com]
Sent: Tuesday, January 31, 2017 8:03 PM
To: Wang, Youcai [GCB-CARDS]
Cc: Amy Wang; Kerry O'Shea (kerry@h2o.ai); Li, Peng4 [GCB-CARDS]

Subject: Re: variable inflation and Collinearity

Hi All,

I've been looking at this issue tonight and there is a fundamental difference between the SAS and h2o approaches.

SAS PROC REG is one of the oldest SAS procedures and uses a traditional, non-penalized approach to fit regression parameters. This approach suffers from numeric instability during the training process if input variables are highly correlated. The COLLIN and COLLINOINT options on the PROC REG MODEL statement use methods described in a reference from 1980 to diagnose these potential numerical problems *after* the training process. Here is the SAS documentation for those options:
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_reg_details24.htm.

H2o uses a more contemporary approach that addresses correlation *during* training using elastic net penalization of regression parameters. Elastic net was introduced in the early 2000s, some 20 years after the methods implemented in the COLLIN and COLLINOINT
options: http://users.stat.umn.edu/~zouxx019/Papers/elasticnet.pdf.

I am not saying newer is necessarily better, but I am trying to communicate that the methods implemented in h2o.glm were specifically designed to address collinearity issues during training so that the type of analysis provided by the COLLIN and COLLINOINT options would not be necessary after training.

Of course Amy and I will continue to look into this, but I wanted to share these thoughts in case they are helpful.

Thanks,
p

--
Patrick Hall
phall@h2o.ai
(336) 693 4481

On Tue, Jan 31, 2017 at 6:35 PM, Wang, Youcai <youcai.wang@citi.com> wrote:

Hi Amy,
It is the goal to make VIF less than 2 and Collinearity less than 16.
These two goals are correlated, but not completely correlated (for some variables, the vifs are very small just close equals to 1, but they can make Collinearity very big).
So it is need to get the output of the values then we can show that it meet the requirement.

Thanks,
Youcai

From: Amy Wang [mailto:amy@0xdata.com]
Sent: Tuesday, January 31, 2017 6:30 PM
To: Wang, Youcai [GCB-CARDS]
Cc: Kerry O'Shea (kerry@h2o.ai); Li, Peng4 [GCB-CARDS]; Patrick Hall

Subject: Re: variable inflation and Collinearity

Hey Youcai,

So the VIF isn't currently exposed to the front end, is the goal to set VIF to be less than 2 and remove the collinear columns or is the goal to calculate the VIF for each variable and output it for the user?

I'm not familiar with SAS's proc reg so I'm cc-ed my colleague, Patrick Hall, who can shed more light on what the output **ColliNoInt** actually is.

Sincerely,
Amy Wang

On Tue, Jan 31, 2017 at 3:20 PM, Wang, Youcai <youcai.wang@citi.com> wrote:
Hi Amy,

I do not have R code,
But if I run

```
proc reg data=may2016ty_EXP_TU;;
model resp =
&tyet
/vif collin corrb tol ColliNoInt;

run;
```

in SAS, I can get all the results.

Thanks,
Youcai

From: Amy Wang [mailto:amy@0xdata.com]
Sent: Tuesday, January 31, 2017 6:17 PM
To: Wang, Youcai [GCB-CARDS]
Cc: Kerry O'Shea (kerry@h2o.ai); Li, Peng4 [GCB-CARDS]
Subject: Re: variable inflation and Collinearity

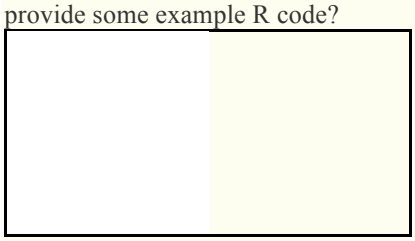
Hey Youcai,

So there is a **remove_collinear_columns** in the **h2o.glm** function. Which you can invoke by running:

```
h2o.glm(training_frame = data, x = ..., y = ..., family = "binomial", remove_collinear_columns = T, ...)
```

And the way that it is done is also by calculating the VIF, in our version of removing collinear columns we set the threshold to R2 = 1e-7, which means VIF = 1/(1-R2) is going to be smaller than 1. I'm linking the source code in github: Collinearity: Source Code Documentation.

I'm not sure what you mean to have collinearity less than 16 though. If this is an additional requirement can you provide some example R code?



Sincerely,
Amy Wang

On Tue, Jan 31, 2017 at 1:58 PM, Wang, Youcai <youcai.wang@citi.com> wrote:
Hi Amy,

We just had a meeting with our model governance, they told us it is required to build a logistic model as benchmark model on the same dataset which we are building the gbm model for model performance checking.
For gbm models, we almost get there.
But for logistic regression models, we need to reduce the variable inflation less than 2 and collinearity should be less than 16.
Do you have R packages in H2o to calculate the variable inflation as well as collinearity?
If yes, how can we leverage these packages to merge into logistic regression model building process in H2O environment?

Thanks,
Youcai