From: Ben Maslen
To: Peter Yates
Subject: Re: Statistics discussion
Date: Friday, 10 May 2019 2:59:03 PM
Attachments: image001.png

Hi Peter

Thanks for sending your notes through. I had a bit more of a think about it and stand by my suggestions to keep the variables in. I would only be really concerned of multicollinearity if you have two variables that are measuring very similar things or are highly related (like r>0.9). I guess you are interested in model selection anyway, so the problem of variance inflation with individual multicollinear terms, doesn't really impact you too greatly anyway. It is mainly a concern for when estimating collinear regression coefficients and their variance, as these coefficients are not reliable and the variance gets inflated. Thinking about this has also made me do a talk on these ideas which will be advertised soon if you are interested in coming!

A good article to aslo consider is below:
http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-621/confounding.pdf

> Basic Ideas - Faculty of Medicine, McGill University
>
> 4 collinearity poses a huge problem if the objective of the study is to estimate the individual effects of each independent variable. Strictly speaking, "collinear" implies an exact linear relationship between variables.
>
> www.medicine.mcgill.ca

Cheers

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

Authorship Guidelines when Consulting with a Statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Wednesday, 8 May 2019 2:34 PM
To: Ben Maslen
Subject: RE: Statistics discussion

Hi Ben, thanks again for the chat.
Attached is that document!

From: Peter Yates
Sent: Wednesday, 8 May 2019 9:33 AM
To: Ben Maslen <b.maslen@unsw.edu.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Great! See you at the lifts on level 2?

From: Ben Maslen <b.maslen@unsw.edu.au>
Sent: Wednesday, 8 May 2019 9:29 AM
To: Peter Yates <Peter.Yates@sims.org.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: Re: Statistics discussion

Hi Peter

I will see oyu at 1:30pm today then! I am not sure if Peter GS is in today, but if he is I will see if he can come along as well.

Cheers

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

Authorship Guidelines when Consulting with a Statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Wednesday, 8 May 2019 9:28 AM
To: Ben Maslen; Peter Geelan-Small
Subject: RE: Statistics discussion

Thanks Ben,

This arv after 1:30 works perfect! I'm flexible but earlier the better for me.

Cheers

From: Ben Maslen <b.maslen@unsw.edu.au>
Sent: Wednesday, 8 May 2019 9:25 AM
To: Peter Yates <Peter.Yates@sims.org.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: Re: Statistics discussion

Hi Peter

I am free all day tomorrow or today after 1:30pm for a chat if you would like to meet up?

Kind regards

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

Authorship Guidelines when Consulting with a Statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Tuesday, 7 May 2019 10:05 AM
To: Ben Maslen; Peter Geelan-Small
Subject: RE: Statistics discussion

Hi Ben and Peter,

I'm wondering if I could meet with either of you to show some GAM outputs and ask some questions? I think the corGaus bit is working so my questions are fairly general.

I won't be able to come to UNSW on the following:
*Friday 10th
*Monday 13th
*Tuesday mornings

Otherwise I am fairly flexible.

Thanks,
Peter

From: Peter Yates
Sent: Monday, 6 May 2019 11:31 AM
To: Ben Maslen <b.maslen@unsw.edu.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Thanks Ben,
Very much appreciated!
Regarding mice, I discovered that I might be able to keep things a bit simpler (for me at least!). I've omitted a bunch of candidate explanatory variables based on collinearity, and for which I had the strongest hypothesis for. Most of the shortlisted variables now have few or no NAs, so I'm hoping I can exclude rows with NAs and proceed with a more simple AIC approach.

Introducing the corGaus-style correlation structure is causing me some error message trouble. I'll do some more reading but if I'm still stuck in a few days, perhaps we could meet again for a chat?

Cheers,

Peter

From: Ben Maslen <b.maslen@unsw.edu.au>
Sent: Friday, 3 May 2019 4:34 PM
To: Peter Yates <Peter.Yates@sims.org.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: Re: Statistics discussion

Hi Peter

Glad it was a useful example. Multicollinearity is definitely also a good thing to check for and is where an effect is shared accross multiple predictors, which can be detected by comparing the correlations between the predictors. Multicollinearity can be problematic in inference that depends on the change to likelihood, as the change in likelihood of the overall effect will be shared amongst the two predictors. That is, you might have an overall effect, but becuase you have two terms in your model that share this overall effect, it might not be able to be detected in your inference.

I had a good example the other day were someone was investigating which parameters make a muffin 'tasty'. In their model they had two parameters that measure sweetness in some manner (which were highly correlated together). When including both sweetness predictors, neither one came out as important predictors. That's becuase the change in likelihood of the model from each predictor on it's own was small. When one of these predictors however were taken out, all of a sudden the other becomes significant as the effect of sweetness is now focused only on the one predictor and the change in likelihood becomes a lot larger. Multicollinearity is harder to picture with DAGs and has got me thinking on how to do so.... Another method that could have also been used is to include a term that represents both sweetness parameters, that is some sort of 'total sweetness'. This could be done perhaps with factor analysis or PCA as dimension reduction techniques that try and find an unobserved 'latent' variable that explains the two observed variables together.

So whilst going on this side tangent, yes I believe that to be an appropriate approach

Cheers

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

---

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Friday, 3 May 2019 4:05 PM

To: Ben Maslen; Peter Geelan-Small
Subject: RE: Statistics discussion

Hi Ben,

No worries at all, and thanks very much for the info. I'll retain Dbin, Vbin and Site and see how it goes!

The schoolkids example is a helpful one for me now…I'm currently at that stage of deciding which variables to keep and which to exclude, in part based on collinearity between them. My current approach, in the schoolkids analogy, would be to (1) notice that there is collinearity between age and height, (2) decide that age is more likely than height to influence maths scores, (3) omit height from the analyses because including it may reduce the ability of the model to detect a significant age effect (because of the collinearity). Is that approach ok?

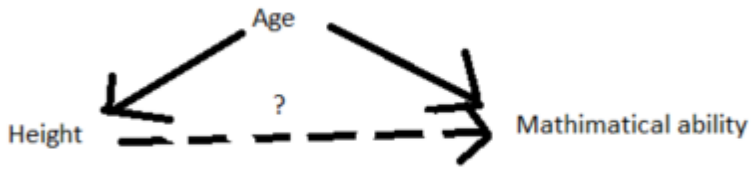
Thanks heaps,
Peter

From: Ben Maslen <b.maslen@unsw.edu.au>
Sent: Friday, 3 May 2019 3:25 PM
To: Peter Yates <Peter.Yates@sims.org.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: Re: Statistics discussion

Hi Peter

Sorry for the super late reply, I have been a bit swamped with emails lately. With these three factors you can't deterimine all the levels of one factor by using the other two so including them would probably be okto leave it as it is. If the model gives you some error messages, then maybe it is something to look into a bit more. Sometimes when 1 lvel of a factor can be determined based on another

Goig back to your first email where you metioned confounding: Thinking about confounding comes more from a biological perspective, on the direction you believe the relationships to be.
For instance, If you were interested in testing if a school kids height was related to their performance in a math test, you would find a relationship. But that is becuase you haven't accounted for a students age. That is, the 'real' relationship is that older kids have learnt more maths and therefore perform better. So once we account for a childs age, we no longer find a relationship between height and math test score, because the relationship between height and meth test score is confounding with age. I think the best way to visuallise or think about confounding is by drawing some 'directed acyclic graphs' or DAGs for short, which just point out the direction of relationships you have with your variables. For instance in the school kids example we have the following DAG:



Where the dotted line above represents the relationship we are interested in testing for. We can see from the above DAG that Age is a confounder variable as it affects both Height and Mathematical ability. So when we test for a relationship between Height and Mathematical ability we would need to include Age, otherwise we could draw false conlusions.

Hope this is helpful. How are you going with the mice package?

Cheers

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

Authorship Guidelines when Consulting with a Statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Monday, 29 April 2019 4:04 PM
To: Ben Maslen; Peter Geelan-Small
Subject: RE: Statistics discussion

Hi again,

Following on from our emails below, can I get your advice on the distribution of obs across some other pairwise combos?

What if only 1 level of a factor can be determined based on another factor (e.g. Dbin 5, in the first two tables). Do I need to exclude rows with Dbin =5 in order to include the factor Dbin?

Similarly in the third table for CapeByron and Weak_NS.

Thanks heaps

```
> table(df$Dbin, df$Vbin)

  Average_S Strong_S Weak_NS
1         3       13       0
2        44      171      27
3         0       97      32
4         0       81      20
5         0       23       0
> table(df$Dbin, df$site)

  CapeByron DiamondHead EvansHead NorthSolitary
1        65          71        81           106
2        65          67        60           107
3        44          50        22            91
4        32          22        10            78
5         0           0         0            32
> table(df$Vbin, df$site)

           CapeByron DiamondHead EvansHead NorthSolitary
Average_S          0          32         2            13
Strong_S         102           0        57           226
Weak_NS            0          79         0             0
```

From: Peter Yates
Sent: Monday, 29 April 2019 8:40 AM
To: Ben Maslen <b.maslen@unsw.edu.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Thanks heaps Ben,

Makes sense. Yep one of the sites contains all of the 178 Average_W, and nothing else (table below). Therefore I'll remove Ubin. Thanks also for the papers, will get to them this week. Writing some lectures today :/

Will keep you posted on how I go!

Cheers
Peter

```
table(df$Ubin, df$site)
```

```
##
##            CapeByron DiamondHead EvansHead NorthSolitary
##   Average_W         0           0         0           178
##   Weak_EW         102         111        59            62
```

From: Ben Maslen <b.maslen@unsw.edu.au>
Sent: Friday, 26 April 2019 2:01 PM
To: Peter Yates <Peter.Yates@sims.org.au>; Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: Re: Statistics discussion

HI Peter

Does site 1 only have the 178 observations that are Average_W or does it also have other observations that are Weak_EW? For me it is more of question of whether the levels of one predictor can be determined by the levels of the others. If that is the case then there is no need to have that extra predictor in the model. If for instance all 178 observations that are Average_W come from site 1, AND site 1 has no more observations, that would mean we could determing the levels of Ubin via site (as site 1= Average_W and site 2-4 = Weak_EW) and we wouldn't need to include Ubin. If not, then perhaps you could include it.

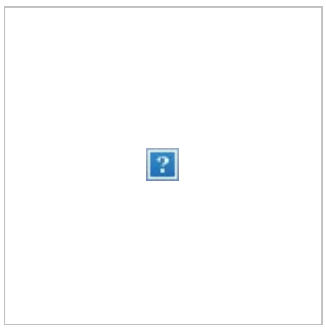
I also found some good spatial autocorellation articles if you are interested:
http://eco-stats.blogspot.com/2014/10/r-lab-inference-with-spatially.html

> ## Eco-Stats Research Blog: R-lab: Inference with Spatially correlated data with nlme
>
> This is as far as nlme will take you, I believe. The packages I have found that are able to do more complex spatial modelling are all Bayesian (so no p-values) but they can do things like spatial prediction (drawing a spatial map), and allow more than one observation per location, which nlme does not seem to do.
>
> eco-stats.blogspot.com

https://www.slu.se/globalassets/ew/org/centrb/statisticsslu/workshops/2016/sven-adler.pdf

> ## ODYI ǀUHNRPVW - SLU - Sveriges lantbruksuniversitet
>
> A habitat is an ecological or environmental area that is inhabited by a particular species of animal, plant, or other type of organism. The term typically refers to the zone in which the organism lives and where it can find food, shelter, protection
>
> www.slu.se

https://stats.stackexchange.com/questions/35510/why-does-including-latitude-and-longitude-in-a-gam-account-for-spatial-autocorre

> 
>
> ## Why does including latitude and longitude in a GAM account for spatial autocorrelation? - Cross Validated - Stack Exchange
>
> I have produced generalized additive models for deforestation. To account for spatial-autocorrelation, I have included latitude and longitude as a smoothed, interaction term (i.e. s(x,y)). I've ba...

stats.stackexchange.com

The first article goes through spatial autocorrelation for mixed effect models, however the methods can be similarly applied to a gam. The second two articles go through breifly a couple of techniques for spatial autocorrelation in a gam setting (between smoothers and correlation structures).

Cheers

Ben

Statistician | Stats Central | Mark Wainwright Analytical Centre | UNSW Sydney NSW 2052 | m: +61 (2) 9385 8220 | www.statscentral.unsw.edu.au
Location: Lvl 2 E26
I work as a statistician Wednesday - Friday and do not check emails on Mondays and Tuesdays

Authorship Guidelines when Consulting with a Statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Sent: Friday, 26 April 2019 12:41 PM
To: Peter Geelan-Small; Ben Maslen
Subject: RE: Statistics discussion

Hi Peter and Ben,

Thanks very much for the meeting today. Very helpful. An additional question if I may:

Re confounding between factors (?). The table below is the number of observations in each factor-level combo for two variables. The two factors are Ubin and Vbin; indicative of east-west and north-south current strengths, respectively. 'Average_W' (a level of Ubin) only occurs along with 'Stong_S' (a level of Vbin), and furthermore all of those 178 obs are from a single site. I think site should definitely be included in which case do I need to drop Ubin, because it is confounded with both Site and Vbin?

```
table(df$Vbin, df$Ubin)
```

```
##
##             Average_W Weak_EW
##   Average_S         0      47
##   Strong_S        178     208
##   Weak_NS           0      79
```

Many thanks!
Have a great weekend,

Peter

From: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Sent: Tuesday, 9 April 2019 12:21 PM
To: Peter Yates <Peter.Yates@sims.org.au>
Subject: Re: Statistics discussion

Great, Peter, see you then.

Have a pleasant Easter long weekend,

Peter

Peter Geelan-Small | Statistical Consultant | Stats Central | Mark Wainwright Analytical Centre
UNSW Sydney NSW 2052 | m: 0435 579 163 | Work days: Tuesday to Friday |
www.statscentral.unsw.edu.au

Stats Central Events

Open Seminars - Details here
Topic: How do you deal with count data? (11 April, 2019)

Short Courses - Details here
Next courses: Introduction to R (7 May, 2019) Introductory Statistics for Researchers (8-9 May, 2019)

Subscribe to Stats Central e-bulletin

Authorship guidelines when consulting with a statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Date: Tuesday, 9 April 2019 at 12:16
To: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Thanks Peter.,
Friday 26th at 9:30 is great!

From: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Sent: Tuesday, 9 April 2019 11:22 AM
To: Peter Yates <Peter.Yates@sims.org.au>
Subject: Re: Statistics discussion

G'day, Peter.

Postponing our meeting is fine. I hope the data processing goes well!

After Easter would be a better time to meet. How about Tuesday, 23rd April, sometime between 9.30 am and 3.30 pm or Friday, 26th April, from 9.30 am to 10.30 am?

See you,

Peter

Peter Geelan-Small | Statistical Consultant | Stats Central | Mark Wainwright Analytical Centre
UNSW Sydney NSW 2052 | m: 0435 579 163 | Work days: Tuesday to Friday |
www.statscentral.unsw.edu.au

Stats Central Events

Open Seminars - Details here
Topic: How do you deal with count data? (11 April, 2019)

Short Courses - Details here
Next courses: Introduction to R (7 May, 2019) Introductory Statistics for Researchers (8-9 May, 2019)

Subscribe to Stats Central e-bulletin

Authorship guidelines when consulting with a statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Date: Tuesday, 9 April 2019 at 11:09
To: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Hi Peter,

We have an appointment this Thursday 11th. Sorry to be a bother but I've had to repeat a large amount of data processing and won't be ready to talk about the analyses this week. Can we please postpone to late next week, or after Easter?

Apologies again,
Peter

From: Peter Yates
Sent: Friday, 29 March 2019 10:37 AM
To: 'Peter Geelan-Small' <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

nd

Thanks Peter, I'll meet you at the lifts on 2 floor
See you then!

From: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Sent: Friday, 29 March 2019 10:34 AM
To: Peter Yates <Peter.Yates@sims.org.au>
Subject: Re: Statistics discussion

G'day, Peter.

Thursday, 11th April, at 11.00 am sounds good.

We could meet in our building or yours, as you prefer. If you'd like to meet here, here are some directions. Could you please come to the Biological Sciences South (E26) building, 2nd floor, and I'll meet you in the open sitting area near the lifts. Then we can go to a meeting room.

The construction work around our building makes it a bit hard to find us. If you're not familiar with this part of the campus, there's a map attached with directions on how to get to our floor. If all else fails, ring me!

Please make sure you do not use the first set of lifts you come to. Keep walking past the lab with the animal skeletons and use the second set of lifts.

See you then,

Peter

Peter Geelan-Small | Statistical Consultant | Stats Central | Mark Wainwright Analytical Centre
UNSW Sydney NSW 2052 | m: 0435 579 163 | Work days: Tuesday to Friday |
www.statscentral.unsw.edu.au

Stats Central Events

Open Seminars - Details here
Topic: How do you deal with count data? (11 April, 2019)

Short Courses - Details here
Next courses: Introduction to R (7 May, 2019) Introductory Statistics for Researchers (8-9 May, 2019)

Subscribe to Stats Central e-bulletin

Authorship guidelines when consulting with a statistician

From: Peter Yates <Peter.Yates@sims.org.au>

Date: Friday, 29 March 2019 at 10:24
To: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Thanks Peter,

Any time on Thursday 11th would be great, thanks. Shall we say 11am?

Cheers,
Peter

From: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Sent: Friday, 29 March 2019 10:18 AM
To: Peter Yates <Peter.Yates@sims.org.au>
Subject: Re: Statistics discussion

G'day, Peter.

Thanks for getting back to me.

We could meet that week sometime in these time slots:

Tuesday, 9th April 10.00 am to 3.30 pm
Thursday, 11th April 10.00 am to 2.00 pm

Is any time there convenient for you?

See you,

Peter

Peter Geelan-Small | Statistical Consultant | Stats Central | Mark Wainwright Analytical Centre
UNSW Sydney NSW 2052 | m: 0435 579 163 | Work days: Tuesday to Friday |
www.statscentral.unsw.edu.au

Stats Central Events

Open Seminars - Details here
Topic: How do you deal with count data? (11 April, 2019)

Short Courses - Details here
Next courses: Introduction to R (7 May, 2019) Introductory Statistics for Researchers (8-9 May, 2019)

Subscribe to Stats Central e-bulletin

Authorship guidelines when consulting with a statistician

From: Peter Yates <Peter.Yates@sims.org.au>
Date: Friday, 29 March 2019 at 10:10
To: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Subject: RE: Statistics discussion

Hi Peter,

Thanks for your email and offer to meet. That would be great. Can we please meet some time in the following week, starting 8th April?

Cheers,
Peter

From: Peter Geelan-Small <p.geelan-small@unsw.edu.au>
Sent: Friday, 29 March 2019 9:30 AM
To: Peter Yates <Peter.Yates@sims.org.au>
Subject: Statistics discussion

Hi, Peter.

I'm Peter, one of the consultants at Stats Central. Thank you for contacting us about methods for working with your data. Apologies for the delay in replying to your request!!

It would be good to arrange a meeting to talk about your queries. I am currently free at these times to meet you for about 1 hour:

Today, (Friday, 29th March): 2.30 pm to 3.30 pm
or next week Tuesday 2nd or Thursday 4th April between 10.00 am and 3.30 pm.

Could you please let me know what time would suit you?

Regards,
Peter

Peter Geelan-Small | Statistical Consultant | Stats Central | Mark Wainwright Analytical Centre
UNSW Sydney NSW 2052 | m: 0435 579 163 | Work days: Tuesday to Friday | www.statscentral.unsw.edu.au

Stats Central Events

Open Seminars - Details here

Short Courses - Details here
Next courses: Introduction to R (7 May, 2019) Introductory Statistics for Researchers (8-9 May, 2019)

Subscribe to Stats Central e-bulletin

Authorship guidelines when consulting with a statistician