TREC & KREC of twins: Decomposing the covariance matrix

Henrique Aparecido Laureano* Wagner Hugo Bonat†
Stéfanne Maria Jeha Bortoletto‡
Carolina Cardoso de Mello Prando§

August, 2022

Abstract

Keywords:

Introduction

Methods

Data

Statistical analysis

The statistical analysis was performed through the R (R Core Team 2022) language and environment for statistical computing.

> Modelos multivariados de regressão feitos para lidar com as interrelações genéticas e ambientais de dados de gêmeos (Bonat and Hjelmborg 2022) foram aplicados para entender a dinâmica das medidas de TREc e KREC.

*Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Paraná, Brazil
†Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná, Curitiba, Paraná, Brazil
‡Faculdades Pequeno Príncipe & Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Paraná, Brazil
§Faculdades Pequeno Príncipe & Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Paraná, Brazil

The main used R packages are: `{dplyr}` (Wickham et al. 2022), `{stringr}` (Wickham 2019), `{ggplot2}` (Wickham 2016), `{tidyr}` (Wickham and Girlich 2022), `{broom}` (Robinson, Hayes, and Couch 2022) and `{mglm4twin}` (Bonat 2022, 2018; Bonat and Hjelmborg 2022; Bonat and Jorgensen 2016).

Results

Modelamos as medidas de TREC e KREC conjuntamente em duas frentes: na média e na variância. Como temos mais de uma variável e vamos também modelar a correlação entre elas, chamamos de covariância.

Para estudar a herdabilidade e o relacionamento genético e ambiental das medidas, usamos um modelo chamado de ACE. Basicamente, decompomos a matriz de variância-covariância (a partir daqui chamaremos simplesmente de matriz de covariância) em três matrizes:

- A: efeito/componente genético/herdabilidade;
- C: efeito/componente mútuo do ambiente (common environment);
- E: efeito/componente único do ambiente (unique environment).

Além de considerarmos essa decomposição da covariância podemos também colocar covariáveis nela.

Tanto na média quanto na covariância usamos/testamos o efeito de seis covariáveis:

- Peso;
- Idade gestacional;
- Tipo de parto;
- Sexo;
- Zigocidade;
- Gêmeo (1 ou 2, para ver se realmente existe uma aleatoriedade na disposição sdos dados).

A zigocidade é um termo chave dessa modelagem, dado que precisamos informar quantos pares de gêmeos são monozigotos e dizigotos. Portanto, o par que não tem essa informação foi descartado. Ficamos/usamos 198 gêmeos.

Começamos com um modelo bivariado ACE, modelando TREC e KREC conjuntamento. Contudo, observamos (p)-valores muito altos para o componente C. Tal fato indica que tal componente não é necessário. Portanto, ajustamos um modelo AE e vemos se é melhor ficar com ele. Abaixo temos algumas medidas de qualidade do ajuste dos dois modelos.

	Estimates	std.error	z value	Pr(>|z|)
TREC				
(Intercept)	-97.1912	127.8171	-0.7604	**0.4470**
id_gest	8.0203	3.5627	2.2512	**0.0244**
sexoMasculino	-34.7133	15.0805	-2.3019	**0.0213**
KREC				
(Intercept)	-160.7853	126.7882	-1.2681	**0.2047**
id_gest	7.7597	3.5385	2.1929	**0.0283**

```
  Model  plogLik Df    pAIC    pKLIC     pBIC
1   ACE -2353.64 23 4753.28 5405.218 4844.853
2    AE -2369.63 20 4779.26 5290.002 4858.888
```

A primeira medida, plogLik, é do tipo maior-melhor, as demais são do tipo menor-melhor. De modo geral, o modelo ACE apresenta melhores medidas. Contudo ele tem três parâmetros a mais. A diferença das medidas entre os modelos não é grande o suficiente para justificar permanecermos com um modelo maior sendo que o componente C não é estatisticamente significativo. Assim, ficamos com o modelo bivariado AE.

Agora, a seleção de variáveis. Começamos pela média. Das seis variáveis ficamos com apenas duas, ou seja, das seis apenas duas se mostram significativas. As medidas do modelo inicial e do final são apresentadas abaixo.

```
       Model  plogLik Df    pAIC    pKLIC     pBIC
1 AE initial -2372.17 11 4766.34 5441.763 4810.136
2   AE final -2372.17 11 4766.34 5441.763 4810.136
```

Vemos que as medidas dos modelos são bem similares, apesar do modelo final ter praticamente metade dos parâmetros do modelo inicial. No modelo final ficamos com duas covariáveis como significativas para o TREC e com uma variável significativa para o KREC.

Para TREC as variáveis significativas são a idade gestacional e o sexo, i.e. a idade gestacional e o sexo do gêmeo tem uma associação significativa com os valor de TREC. Para o KREC, apenas a idade gestacional é significativa, o sexo é irrelevante (sem diferença significativa de um sexo para outro).

Em todo modelo de regressão/estatístico temos um intercepto, quando falamos da interpretação dos coeficientes estimados. O intercepto é o nível de referência. Para o TREC, o intercepto corresponde a um gêmeo de idade gestacional média e do sexo feminino. Conforme aumentamos a idade gestacional aumentamos seu TREC

	Estimates	std.error	Percentage	z value	Pr(>|z|)
Environment component (E)					
TREC	3712.4482	2751.5400	27.1019	1.3492	**0.1773**
KREC	1391.8790	4100.7340	10.0955	0.3394	**0.7343**
TREC & KREC	862.7034	490.8413	19.6407	1.7576	**0.0788**
Genetic component (A)					
TREC	8703.5756	3370.8550	63.5385	2.5820	**0.0098**
KREC	9990.1531	5750.4968	72.4603	1.7373	**0.0823**
TREC & KREC	2633.3467	1090.5257	59.9521	2.4147	**0.0157**

	Estimates	std.error	z value	Pr(>|z|)
Environmentability				
TREC	0.2990	0.2096	1.4267	**0.1537**
KREC	0.1223	0.3634	0.3365	**0.7365**
Heritability				
TREC	0.7010	0.2096	3.3449	**0.0008**
KREC	0.8777	0.3634	2.4155	**0.0157**

no valor de seu coeficiente estimado (8.0203). Se o gêmeo é do sexo masculino, estimamos que seu TREC diminua 34.7133, em relação ao intercepto. Mesma ideia para o KREC e a idade gestacional.

Agora, a covariância. Decompomos a variância do TREC (1.3698107×10^4), a variância do KREC (1.3787069×10^4) e a covariância entre os dois (4392.4173) da seguinte maneira.

Temos as estimativas divididas pelos componentes ambiental e genético. Em outras palavras, quanto da variabilidade dos dados é explicada pelos tais componentes. Como podemos ver pela coluna percentage, a soma das estimativas (do TREC, por exemplo) não dá 1. Isso quer dizer que nem toda a variabilidade observada é explicada pelo componente ambiental ou genético.

Vemos que o componente genético é muito mais significativo, principalmente no TREC. Continunando no TREC, a fim de didática, 27% (não significativo, (p)-valor de 0.177) da variabilidade observada é explicada pelo componente ambiental e 63% (significativo, (p)-valor de 0.0098) pelo componente genético.

Abaixo temos as medidas de ambientalidade e herdabilidade. Vemos que apenas a herdabilidade é significativa, para ambos TREC e KREC.

	Estimates	std.error	z value	Pr(>|z|)
Bivariate environmentability	0.2468	0.1580	1.5619	**0.1183**
Environment correlation	0.3795	0.5943	0.6386	**0.5231**
Bivariate heritability	0.7532	0.1580	4.7676	**0.0000**
Genetic correlation	0.2824	0.1397	2.0209	**0.0433**

	Estimates	Std.error	Percentage
Environment component (E)	3534.162	2773.739	25.8004
Genetic component (A)			
partoCesariana	9851.598	3647.523	71.9194
partoNormal	-10976.948	3140.646	-80.1348

A única correlação significativa é a genética, a ambiental não é. Seguindo a mesma ideia, a herdabilidade bivariada/conjunta é significativa. A ambiental não é.

Tentamos verificar o efeito das covariáveis também na covariância, i.e. "quebrar" as medidas apresentadas acima pelas covariáveis, mas não conseguimos. Estamos lidando com um modelo bem complexo e nosso tamanho amostral não é grande. Portanto, não conseguimos convergência nos modelos bivariados com covariáveis na covariância. Solução? Ajustar modelos univariados. Os componentes de média se mantém exatamente os mesmos.

TREC

Partimos do modelo AE com idade gestacional e sexo como covariáveis na média e fizemos a seleção de covariáveis na covariância. Das seis covariáveis nenhuma é significativa no componente E, ambiental. No componente A, genético, a variável parto é significativa.

26% da variabilidade observada é explicada/atribuída pelo componente ambiental. O tipo de parto é significativo no componente genético. Com um gêmeo que teve parto do tipo cesaria, o componente genético explica 72% da variabilidade. Ou seja, os componentes ambiental e genético explicam cerca de 98% de toda a variabilidade da variável TREC. Agora, num gêmeo que teve parto normal, (o poder de) explicação cai 80%. Ou seja, num gêmeo com parto normal o componente genético não explica nada da variabilidade. Isso é justificado pela (baixa) representatividade nos dados, 7% apenas dos gêmeos tiveram parto normal (ou seja, a incerteza inerente é muito alta).

	Estimates	Std.error	Percentage
Environment component (E)			
partoCesariana.zigocidadeDZ	-8562.138	3940.7251	-62.1027
partoNormal	9185.825	4003.8748	66.6264
zigocidadeMZ	9829.275	609.2463	71.2934
Genetic component (A)			
partoCesariana	22781.346	3823.2523	165.2371
partoNormal	-22636.617	3909.9002	-164.1873

KREC

Partimos do modelo AE com idade gestacional como covariável na média e fizemos a seleção de covariáveis na covariância. Das seis covariáveis, tipo de parto e zigocidade são significativas no componente E, ambiental. No componente A, genético, a variável parto é significativa.

Num gêmeo com parto tipo cesaria e dizigoto, nada da variabilidade do KREC é explicada pelo componente ambiental (-62%). Contudo, se mudamos para um gêmeo de parto normal a explicação cresce 66%, ou seja, 4% (-62 + 66) da variabilidade é explicada pelo componente ambiental. Num gêmeo monozigoto o percentual explicado é 9% (-62 + 71).

Quando olhamos para o componente genético a diferença entre tipos de parto é abismal. Num gëmeo dizigoto de parto tipo cesaria temos 103% (-62 + 165) da variabilidade unicamente explicada pelo componente genético. Tal efeito é tão grande mas ao mesmo incerto, que o percentual estoura. No caso de parto normal temos exatamente o contrário. O desbalancemanto das frequências de tipo de parto na base de dados acaba gerando esses resultados "estranhos".

References

Bonat, Wagner Hugo. 2018. "Multiple Response Variables Regression Models in R: The mcglm Package." *Journal of Statistical Software* 84 (4): 1–30. https://doi.org/10.18637/jss.v084.i04.

———. 2022. *mglm4twin: Multivariate Generalized Linear Models for Twin Data*. R package version 0.3.0. https://github.com/wbonat/mglm4twin.

Bonat, Wagner Hugo, and Jacob V. B. Hjelmborg. 2022. "Multivariate Generalized Linear Models for Twin and Family Data." *Behavior Genetics* 52 (2): 123–40. https://doi.org/10.1007/s10519-021-10095-3.

Bonat, Wagner Hugo, and Bent Jorgensen. 2016. "Multivariate Covariance Generalized Linear Models." *Journal of Royal Statistical Society - Series C* 65: 649–75.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *broom: Convert Statistical Objects into Tidy Tibbles*. R package version 0.8.0. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

———. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Maximilian Girlich. 2022. *tidyr: Tidy Messy Data*. R package version 1.2.0. https://CRAN.R-project.org/package=tidyr.