Gabarito - Lista 11

Regressão: Especificação e Diagnóstico

Gabarito

Exercício 1 [10 pontos]. Um colega de departamento está rodando sua primeira regressão e, como não tem muita experiência, achou melhor incluir todas as variáveis do banco como VIs. Analise os resultados abaixo e sugira maneiras de melhorar o modelo. Não se esqueça de comentar os seguintes pontos:

- Seleção de variáveis
- Transformações necessárias
- Colinearidade
- Comportamento dos resíduos
- *Outliers* influentes

Cada sugestão de melhoria deve ser acompanhada por suas respectivas justificativas. Ao fim, use o banco *world95* para rodar um modelo final que, na sua opinião, melhore o máximo possível os resultados do seu colega.

```
. reg babymort lifeexpf populatn density urban pop_incr gdp_cap region lit_fema p_polity2 calories lifeexpm

      Source |       SS       df       MS              Number of obs =      55
-------------+------------------------------           F( 11,    43) =   57.56
       Model |  71690.0265    11  6517.27514            Prob > F      =  0.0000
    Residual |  4868.95904    43  113.231605            R-squared     =  0.9364
-------------+------------------------------           Adj R-squared =  0.9201
       Total |  76558.9856    54  1417.75899            Root MSE      =  10.641

------------------------------------------------------------------------------
   babymort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lifeexpf |  -3.223495   1.125526    -2.86   0.006    -5.493335   -.9536544
   populatn |  -1.73e-06   8.70e-06    -0.20   0.843    -.0000193    .0000158
    density |  -.0043279   .0027347    -1.58   0.121    -.0098429    .0011872
      urban |   .1092234   .1220651     0.89   0.376    -.1369444    .3553912
   pop_incr |  -6.313744   2.814103    -2.24   0.030    -11.98892   -.638564
    gdp_cap |   .0000605   .0006805     0.09   0.930    -.0013118    .0014329
     region |    1.47115   1.855233     0.79   0.432    -2.270284    5.212584
   lit_fema |  -.4507339   .1145148    -3.94   0.000     -.681675   -.2197927
  p_polity2 |  -.4253521   .3095957    -1.37   0.177    -1.049711    .1990072
   calories |  -.0063174   .0059007    -1.07   0.290    -.0182174    .0055826
   lifeexpm |   .9437096   1.173016     0.80   0.426    -1.421902    3.309321
      _cons |   260.2079   16.73908    15.54   0.000     226.4503    293.9655
------------------------------------------------------------------------------
```

As sugestões deste gabarito não são exaustivas, e o modelo final proposto aqui tampouco é definitivo. Em particular, vale notar que não está sendo levada em consideração aqui a literatura sobre mortalidade infantil, que não é minha especialidade. Respostas que incluírem justificativas teóricas na argumentação serão avaliadas com bons olhos – recorrer à literatura é um ótimo motivo para incluir ou excluir variáveis em um modelo. Por questões práticas, a seleção de variáveis neste gabarito será feita (quase) puramente com base na análise dos dados. De modo geral, as respostas serão julgadas pelo repertório de sugestões feitas por cada aluno e pela coerência das justificativas apresentadas.

A primeira sugestão é especificar corretamente a sintaxe para variáveis independentes categóricas. A variável *region* está sendo tratada no modelo como uma variável contínua e, consequentemente, gerando um coeficiente de pouca utilidade. Se rodarmos a regressão com a especificação `i.region`, obteremos estimativas mais interessantes do ponto de vista substantivo:

```
reg babymort lifeexpf lifeexpm populatn density urban pop_incr gdp_cap
lit_fema p_polity2 calories i.region
```

Source	SS	df	MS		
Model	71958.1692	14	5139.86923		
Residual	4600.81635	40	115.020409		
Total	76558.9856	54	1417.75899		

Number of obs	=	55		
F(14, 40)	=	44.69		
Prob > F	=	0.0000		
R-squared	=	0.9399		
Adj R-squared	=	0.9189		
Root MSE	=	10.725		

babymort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeexpf	-2.981309	1.207835	-2.47	0.018	-5.422434	-.5401832
lifeexpm	.5160415	1.261269	0.41	0.685	-2.033079	3.065162
populatn	-3.62e-06	9.60e-06	-0.38	0.708	-.000023	.0000158
density	-.004415	.0038151	-1.16	0.254	-.0121256	.0032957
urban	.1121682	.1247137	0.90	0.374	-.1398876	.3642241
pop_incr	-6.004794	3.106959	-1.93	0.060	-12.28419	.2746046
gdp_cap	.0000899	.0008417	0.11	0.916	-.0016112	.0017909
lit_fema	-.4707455	.1188598	-3.96	0.000	-.7109702	-.2305209
p_polity2	-.4442919	.3450071	-1.29	0.205	-1.141577	.2529934
calories	-.0085573	.0071882	-1.19	0.241	-.0230853	.0059706
region						
Pacific/Asia	.1306996	14.8085	0.01	0.993	-29.79839	30.05979
Africa	-6.433918	14.88198	-0.43	0.668	-36.51151	23.64368
Middle East	3.973346	14.16951	0.28	0.781	-24.6643	32.61099
Latn America	3.053253	14.17152	0.22	0.831	-25.58845	31.69496
_cons	283.4409	25.24914	11.23	0.000	232.4105	334.4713

Mas o modelo ainda está ruim, e não precisamos de todas essas variáveis. Para avaliar se existe colinearidade entre elas, vamos usar o comando `vif`

Variable	VIF	1/VIF
lifeexpf	88.17	0.011341
lifeexpm	78.15	0.012796
populatn	1.76	0.567258
density	2.48	0.403562
urban	4.79	0.208739
pop_incr	3.76	0.265995
gdp_cap	8.06	0.124057
lit_fema	4.85	0.206373
p_polity2	2.18	0.459379
calories	6.16	0.162416
region		
3	14.35	0.069681
4	21.85	0.045775
5	10.66	0.093775
6	21.14	0.047297

Existe muita colinearidade ente a expectativa de vida feminina e a expectativa de vida masculina, portanto vamos eliminar uma delas. Como a mortalidade infantil está mais associada à saúde da mulher do que à do homem, vamos eliminar a última.

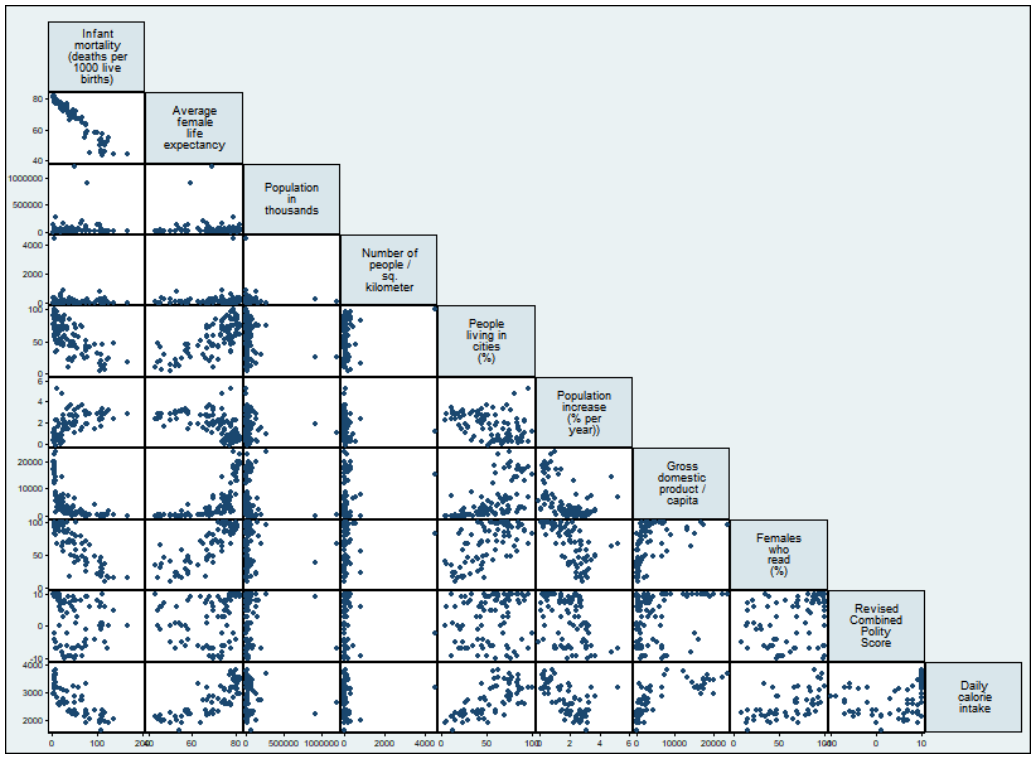
Em seguida, vamos avaliar se precisamos fazer a transformação logarítmica em alguma variável. Decidimos isso a partir da inspeção visual das relações de todas as variáveis contínuas com a variável dependente.

```
graph matrix babymort lifeexpf populatn density urban pop_incr gdp_cap
lit_fema p_polity2 calories, half
```

A matriz de dispersão se encontra na próxima página. A imagem fica pequena na página. Se tiver dificuldade de ver os dados, rode o código acima no seu próprio computador.

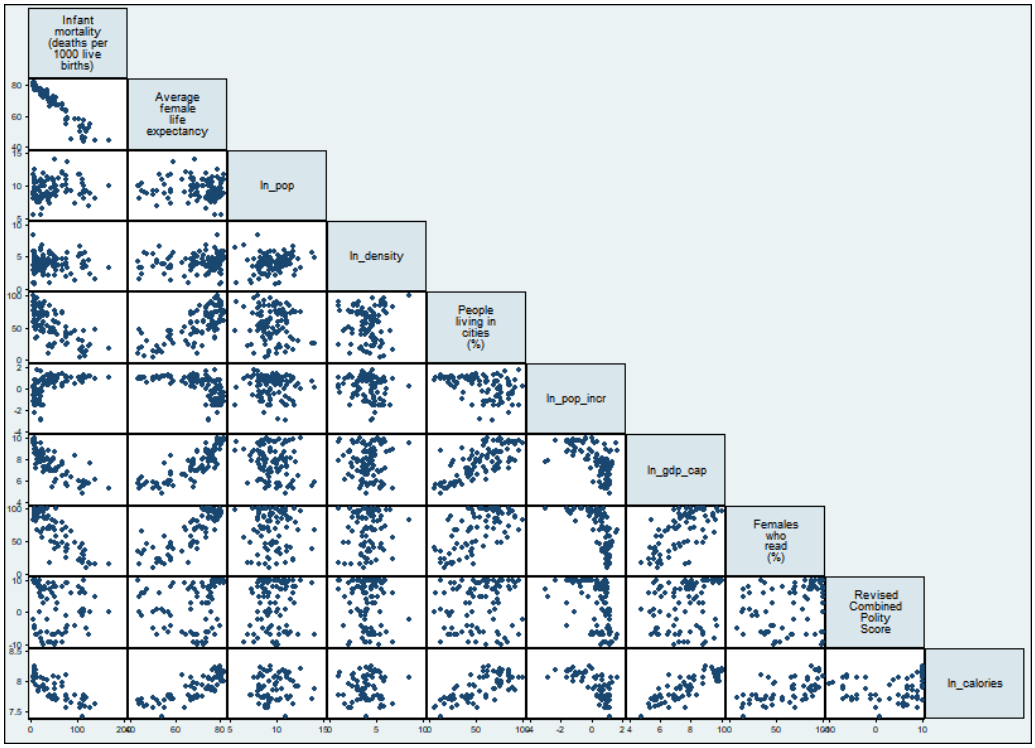
Como queremos saber a associação das variáveis com a VD, estamos preocupados apenas com a primeira coluna da matriz. Ela indica que a mortalidade infantil tem uma relação não linear com população, densidade, crescimento populacional, pib *per capita*, e ingestão de calorias. Vamos calcular o log natural dessas variáveis.

```
generate ln_pop = ln(populatn)
generate ln_density = ln(density)
generate ln_pop_incr = ln(pop_incr)
generate ln_gdp_cap = ln(gdp_cap)
generate ln_calories = ln(calories)
```



Agora podemos ver como fica a associação entre as variáveis depois das transformações:

```
graph matrix babymort lifeexpf ln_pop ln_density urban ln_pop_incr
ln_gdp_cap lit_fema p_polity2 ln_calories, half
```



Melhorou bastante, mas ainda não está perfeito. A maior preocupação é com a variável de crescimento populacional, que parece continuar tendo uma relação não linear com a mortalidade infantil mesmo depois da transformação logarítmica. Vamos ficar de olho nessa variável.

Agora que fizemos as transformações, vamos rodar o modelo com as novas variáveis:

```
reg babymort lifeexpf ln_pop ln_density urban ln_pop_incr ln_gdp_cap
lit_fema p_polity2 ln_calories i.region
```

Source	SS	df	MS	Number of obs =	55
				F(13, 41) =	49.06
Model	71934.7213	13	5533.4401	Prob > F =	0.0000
Residual	4624.26423	41	112.786932	R-squared =	0.9396
				Adj R-squared =	0.9204
Total	76558.9856	54	1417.75899	Root MSE =	10.62

babymort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeexpf	-2.552029	.3449715	-7.40	0.000	-3.248713	-1.855345
ln_pop	1.429491	1.2603	1.13	0.263	-1.115736	3.974718
ln_density	-1.190186	1.302994	-0.91	0.366	-3.821635	1.441264
urban	.0202929	.1210114	0.17	0.868	-.2240945	.2646803
ln_pop_incr	-6.315595	4.592306	-1.38	0.177	-15.58994	2.958755
ln_gdp_cap	1.127088	2.956902	0.38	0.705	-4.844496	7.098672
lit_fema	-.4251499	.1149124	-3.70	0.001	-.6572202	-.1930796
p_polity2	-.378786	.3471694	-1.09	0.282	-1.079909	.3223369
ln_calories	-25.34312	18.37976	-1.38	0.175	-62.4618	11.77556
region						
Pacific/Asia	-1.205402	9.406022	-0.13	0.899	-20.20125	17.79044
Africa	-6.725782	10.97771	-0.61	0.543	-28.89572	15.44416
Middle East	5.132051	10.82192	0.47	0.638	-16.72326	26.98736
Latn America	5.702975	8.782112	0.65	0.520	-12.03286	23.43881
_cons	436.9075	133.2016	3.28	0.002	167.9015	705.9136

Seleção de variáveis

Será que precisamos manter todas essas variáveis no modelo? Vamos seguir as sugestões de Gelman e Hill para decidir se podemos eliminar alguma delas sem perder informação que seja relevante para a análise:

- Se uma variável não é significante e tem o sinal esperado, deixe-a no modelo.

- Se uma variável não é significante e **não** tem o sinal esperado, considere removê-la do modelo.
- Se uma variável é significante e **não** tem o sinal esperado, repense seu modelo e considere se há alguma variável importante não sendo controlada.
- Se uma variável for significativa e tem o sinal esperado, não a tire de jeito nenhum.

Partindo desses critérios, vamos construir uma tabela com os sinais que esperamos para cada variável, os sinais observados, a significância e a decisão se ela fica ou sai.

	Sinal esperado	Sinal observado	Significante?	Decisão
lifeexpf	-	-	Sim	Fica
ln_pop	+	+	Não	Fica
ln_density	+	-	Não	Sai
urban	-	+	Não	Sai
ln_pop_incr	+	-	Não	Sai
ln_gdp_cap	-	+	Não	Sai (?)
lit_fema	-	-	Sim	Fica
p_polity2	-	-	Não	Fica
ln_calories	-	-	Não	Fica (?)

Como foi enfatizado em aula, esses pontos não são regras a serem adotados cegamente. Eles indicam, por exemplo, que a variável *per capita* deveria sair do modelo, mas eu a considero um controle importante demais para não ter. Mesmo que não seja significativa, minha sugestão, neste caso, seria deixá-la. Outra coisa que vale notar é que a variável de crescimento populacional, cuja relação com a variável dependente permaneceu não linear mesmo depois da transformação logarítmica, de fato não vai ficar no modelo.

Uma última preocupação: note nos modelos acima que nossa regressão só tem 55 observações, enquanto o banco de dados tem 102 países. Por que estamos perdendo tanta informação?

```
sum lifeexpf ln_pop ln_gdp_cap lit_fema p_polity2 ln_calories
```

As estatísticas descritivas mostram que a variável de consumo de calorias tem mais de 30 *missings*. O tratamento de *missing data* não é trivial e está fora do escopo deste curso. Para este exercício, vamos eliminar a variável, levando em consideração que ela não é significativa e que, sozinha, nos faz perder 30% do banco.

Como ficamos então?

```
reg babymort lifeexpf ln_pop ln_gdp_cap lit_fema p_polity2 i.region
```

Source	SS	df	MS	Number of obs =	78
				F(10, 67) =	91.98
Model	109994.118	10	10999.4118	Prob > F =	0.0000
Residual	8012.30933	67	119.586706	R-squared =	0.9321
				Adj R-squared =	0.9220
Total	118006.428	77	1532.55101	Root MSE =	10.936

babymort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeexpf	-2.942202	.2933173	-10.03	0.000	-3.527666	-2.356738
ln_pop	.5316802	.9971411	0.53	0.596	-1.458622	2.521982
ln_gdp_cap	-.8248663	2.060072	-0.40	0.690	-4.936787	3.287055
lit_fema	-.3162535	.0858205	-3.69	0.000	-.487552	-.1449551
p_polity2	-.1905034	.2573553	-0.74	0.462	-.7041867	.3231799
region						
East Europe	1.344061	6.492907	0.21	0.837	-11.61584	14.30396
Pacific/Asia	-2.814329	7.309935	-0.39	0.701	-17.40502	11.77636
Africa	-11.73324	8.427354	-1.39	0.168	-28.55431	5.087834
Middle East	-1.332346	7.099383	-0.19	0.852	-15.50277	12.83808
Latn America	1.065685	6.529567	0.16	0.871	-11.96739	14.09876
_cons	276.3686	22.66607	12.19	0.000	231.1269	321.6103

Outliers

Agora que corrigimos, na medida do possível, a especificação do modelo, podemos testar se nossos resultados são sensíveis a *outliers* influentes.

```
reg babymort lifeexpf ln_pop ln_gdp_cap lit_fema p_polity2 i.region, ro-
bust
```

A regressão robusta identifica os *outliers* influentes e os elimina da regressão. É uma das maneiras de certificarmos que estamos avaliando a relação entre os fenômenos propriamente ditos, sem depender demais de alguns casos extremos.

Os resultados (na próxima página) mostram que não foi eliminada nenhuma observação, e as estimativas permanecem as mesmas. Não é preciso mudar quaisquer decisões que tenhamos tomado sobre as variáveis que têm ou não efeito sobre mortalidade infantil.

Linear regression

					Number of obs =	78
					F(10, 67) =	111.73
					Prob > F =	0.0000
					R-squared =	0.9321
					Root MSE =	10.936

babymort	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeexpf	-2.942202	.4348992	-6.77	0.000	-3.810265	-2.07414
ln_pop	.5316802	.9629802	0.55	0.583	-1.390436	2.453797
ln_gdp_cap	-.8248663	2.432809	-0.34	0.736	-5.680773	4.03104
lit_fema	-.3162535	.1286957	-2.46	0.017	-.5731311	-.0593759
p_polity2	-.1905034	.2550612	-0.75	0.458	-.6996078	.318601
region						
East Europe	1.344061	4.008505	0.34	0.738	-6.656949	9.345071
Pacific/Asia	-2.814329	6.041884	-0.47	0.643	-14.87398	9.245321
Africa	-11.73324	9.910603	-1.18	0.241	-31.51488	8.048411
Middle East	-1.332346	6.276606	-0.21	0.833	-13.86051	11.19581
Latn America	1.065685	5.489206	0.19	0.847	-9.890816	12.02219
_cons	276.3686	29.32029	9.43	0.000	217.8451	334.8921

Resíduos

Por fim, vamos analisar como estão os resíduos do nosso modelo. Primeiramente, vamos gerar uma variável com os resíduos de cada observação em relação ao valor predito, e rodar um histograma para ver se os resíduos têm distribuição normal:

```
predict e, resid
histogram e, normal
```

O histograma (primeira imagem na próxima página) mostra que a distribuição dos erros é razoavelmente normal, o que é uma boa notícia. O gráfico de resíduos vs. valores previstos também se comportará da maneira esperada?

```
rvfplot
```

Mesmo após o aperfeiçoamento do modelo, a associação entre os resíduos do modelo e os valores preditos ainda mostra sinais de heterocedasticidade.. Chegamos a essa conclusão pela inspeção visual do gráfico de dispersão na próxima página: há mais variação residual nos países com maiores taxas preditas de mortalidade infantil do que nos países com taxas pequenas. Idealmente, gostaríamos de ver uma variância homogênea dos resíduos,

independentemente do valor de $\hat{y}$

Neste exercício não há muito o que fazer para corrigir isso. A sugestão que poderíamos fazer ao nosso colega seria pensar em alguma variável omitida que ainda está no termo de erro, coletá-la e incluí-la na análise. O modelo final é melhor do que aquele com o qual começamos, mas ainda é problemático e deve ser visto com suspeição. Quem disse que fazer pesquisa é fácil?