D.3 - ANÁLISE DE DADOS MULTIVARIADOS I – REGRESSÃO (AVALIAÇÃO)

Professor: Reinaldo Soares de Camargo

Aluno: _____

Matricula: _____

1. Considerando os resultados da saída do software R abaixo, referente a um modelo econômico que busca explicar os fatores que contribuem para mortalidade infantil em municípios brasileiros, por meio da técnica de simulação por mínimos quadrados ordinários, responda as questões a seguir.

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.8311  -1.9372  -0.2724   1.5532  17.6192

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          3.048e+01  7.765e-01  39.256  < 2e-16 ***
dados3$renda_per_capita                             -1.880e-02  1.301e-03 -14.452  < 2e-16 ***
dados3$indice_gini                                  -3.239e+00  1.299e+00  -2.493  0.01270 *
dados3$salario_medio_mensal                          8.966e-05  8.239e-02   0.001  0.99913
dados3$perc_criancas_extrem_pobres                  -3.272e-02  1.141e-02  -2.866  0.00417 **
dados3$perc_criancas_pobres                          8.827e-02  1.077e-02   8.200 2.96e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados      3.581e-02  5.641e-03   6.348 2.36e-10 ***
dados3$perc_pessoas_dom_paredes_inadequadas          4.559e-02  6.818e-03   6.687 2.51e-11 ***
dados3$perc_pop_dom_com_coleta_lixo                 -1.306e-02  5.583e-03  -2.340  0.01934 *
dados3$perc_pop_rural                               -6.161e-01  2.862e-01  -2.152  0.03141 *
as.factor(dados3$Regiao)Centro-Oeste                -1.395e+01  7.037e-01 -19.829  < 2e-16 ***
as.factor(dados3$Regiao)Norte                       -7.670e+00  5.193e-01 -14.770  < 2e-16 ***
as.factor(dados3$Regiao)Sudeste                     -1.139e+01  4.538e-01 -25.110  < 2e-16 ***
as.factor(dados3$Regiao)Sul                         -1.393e+01  5.804e-01 -24.000  < 2e-16 ***
dados3$renda_per_capita:as.factor(dados3$Regiao)Centro-Oeste  1.937e-02  1.401e-03  13.828  < 2e-16 ***
dados3$renda_per_capita:as.factor(dados3$Regiao)Norte         1.026e-02  1.442e-03   7.114 1.27e-12 ***
dados3$renda_per_capita:as.factor(dados3$Regiao)Sudeste       1.464e-02  1.104e-03  13.262  < 2e-16 ***
dados3$renda_per_capita:as.factor(dados3$Regiao)Sul           1.610e-02  1.224e-03  13.150  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.426 on 5546 degrees of freedom
Multiple R-squared:  0.7703,    Adjusted R-squared:  0.7696
F-statistic:  1094 on 17 and 5546 DF,  p-value: < 2.2e-16
```

a. Qual a região base para análise das variáveis qualitativas (dummies)?

A região base para as análises é a região Nordeste.

b. Interprete os coeficientes que apresentam significância estatística com nível de significância de 5%.

O aumento de R$ 1,00 (uma unidade) na renda percapita reduz a mortalidade infantil em 0,0188 mortos para cada 1000 nascidos vivos, mantendo os demais fatores constantes (ceteris paribus).

A mortalidade infantil na região Centro-Oeste é menor do que na região Nordeste em 13,95 mortos para cada 1000 nascidos vivos, ceteris paribus.

Na região Sul o aumento de R$ 1,00 (uma unidade) na renda percapita elava a mortalidade infantil em 0,0161 morto para cada 1000 nascidos vivos em relação ao Nordeste, ceteris paribus.

c. Qual o percentual da variabilidade da mortalidade infantil que é explicada pelas variáveis explicativas?

77,03% da variabilidade da mortalidade infantil é explicada pelas variáveis explicativas.

2. Considerando a versão do modelo seguinte onde foram excluídas as dummies de interação entre renda per capita e região (modelo restrito), responda:

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.3942  -1.9967  -0.2933   1.6103  18.4841

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  2.696e+01  7.480e-01  36.039  < 2e-16 ***
dados3$renda_per_capita                     -1.990e-03  5.326e-04  -3.737 0.000188 ***
dados3$indice_gini                          -1.016e+01  1.194e+00  -8.510  < 2e-16 ***
dados3$salario_medio_mensal                 -1.016e-01  8.282e-02  -1.227 0.219889
dados3$perc_criancas_extrem_pobres           2.265e-02  1.075e-02   2.107 0.035138 *
dados3$perc_criancas_pobres                  1.026e-01  1.071e-02   9.579  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados 4.505e-02  5.605e-03   8.037 1.12e-15 ***
dados3$perc_pessoas_dom_paredes_inadequadas  5.500e-02  6.896e-03   7.976 1.83e-15 ***
dados3$perc_pop_dom_com_coleta_lixo         -1.490e-02  5.681e-03  -2.622 0.008762 **
dados3$perc_pop_rural                       -5.808e-01  2.912e-01  -1.995 0.046114 *
as.factor(dados3$Regiao)Centro-Oeste        -5.928e+00  2.356e-01 -25.156  < 2e-16 ***
as.factor(dados3$Regiao)Norte               -4.678e+00  2.099e-01 -22.290  < 2e-16 ***
as.factor(dados3$Regiao)Sudeste             -6.246e+00  1.827e-01 -34.195  < 2e-16 ***
as.factor(dados3$Regiao)Sul                 -7.875e+00  2.215e-01 -35.555  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.494 on 5550 degrees of freedom
Multiple R-squared:  0.7609,    Adjusted R-squared:  0.7604
F-statistic:  1359 on 13 and 5550 DF,  p-value: < 2.2e-16
```

a. Qual o modelo poderia ser selecionado segundo o R^2 ajustado, justifique sua resposta?

O modelo irrestrito (da questão 1), pois possui R^2 ajustado de 76,96% contra 76,04% do modelo restrito (questão 2).

b. Considerando os critérios de informação BIC e AIC, qual dos dois modelos poderia ser selecionado, justifique sua resposta?

```
> cbind(AIC(mod.ex), BIC(mod.ex))
          [,1]     [,2]
[1,] 29511.78 29637.64
>
```

```
> cbind(AIC(mod.ex.rest), BIC(mod.ex.rest))
         [,1]     [,2]
[1,] 29726.47 29825.83
>
```

O modelo irrestrito (da questão 1) seus AIC e BIC apresentam valores menores do que os do modelo restrito (questão 2).

c. Considerando o resultado abaixo de um teste anova, Qual a conclusão a partir dos resultados do teste de hipótese? Os coeficientes de interação entre renda per capita e regiões são significativos conjuntamente ou não, descreva a hipótese nula do teste anova e justifique sua resposta?

```
> anova(mod.ex.rest, mod.ex, test='LRT')
Analysis of Variance Table

Model 1: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
    dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +
    dados3$perc_pop_rural + as.factor(dados3$Regiao)
Model 2: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
    dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +
    dados3$perc_pop_rural + as.factor(dados3$Regiao) + as.factor(dados3$Regiao) *
    dados3$renda_per_capita
  Res.Df   RSS Df Sum of Sq  Pr(>Chi)
1   5550 67741
2   5546 65083  4    2657.7 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

A hipótese nula do teste ANOVA é que os coeficientes das variáveis excluídas do modelo sejam iguais a zero. O p-value < 5% rejeita-se a hipótese nula de os coeficientes de interação entre renda per capita e regiões modelos sejam iguais a Zero, portanto em conjunto esses coeficientes são significantes para o modelo estimado.

3. De acordo os resultados do teste de Breusch-Pagan abaixo, reponda:
   a. Qual hipótese nula do teste de Breusch-Pagan?

      A hipótese nula do teste BP é que os erros são homocedasticos, ou seja, os coeficientes estimados são iguais a zero.

b. É possível identificar a presença de heterocedasticidade nos resíduos do modelo estimado?

Como o p-value do teste BP é menor do 5%, rejeita-se a hipótese nula de erros homocedasticos, logo podemos concluir que há heterocedasticidade nos erros do modelo estimado.

c. Quais as consequências heterocedasticidade?

Erros heterocedásticos invalidam os resultados estimados para erro-padrão, estatística teste, p-valor e intervalos de confianças. Os coeficientes estimados não são invalidados

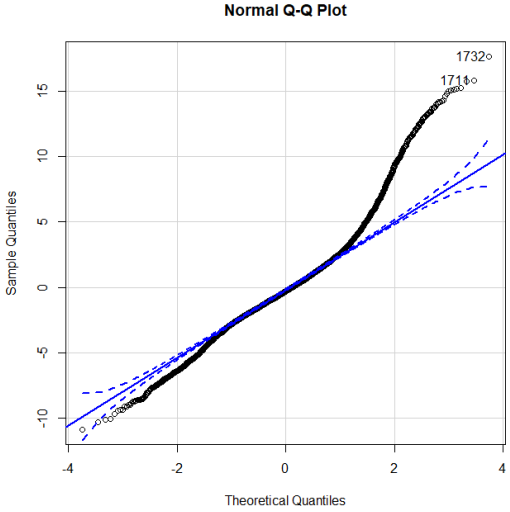
```
> bptest(dados3$mort_infantil ~ dados3$renda_per_capita
+         + dados3$indice_gini
+         + dados3$salario_medio_mensal
+         + dados3$perc_criancas_extrem_pobres
+         + dados3$perc_criancas_pobres
+         + dados3$perc_pessoas_dom_agua_estogo_inadequados
+         + dados3$perc_pessoas_dom_paredes_inadequadas
+         + dados3$perc_pop_dom_com_coleta_lixo
+         + dados3$perc_pop_rural
+         + as.factor(dados3$Regiao)
+         + as.factor(dados3$Regiao)*dados3$renda_per_capita)

        studentized Breusch-Pagan test

data:  dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +     dados3$salario_medio_mensal + dados3$pe
rc_criancas_extrem_pobres +     dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +     dado
s3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +     dados3$perc_pop_rural + as.factor(da
dos3$Regiao) + as.factor(dados3$Regiao) *     dados3$renda_per_capita
BP = 997.89, df = 17, p-value < 2.2e-16
```

4. De acordo com o gráfico qq-plot abaixo, é possível suspeitar de uma possível violação da hipótese de normalidade dos dados? Justifique sua resposta.

Nota-se que os quantis da amostra não se ajustam bem aos quantis da distribuição teórica (normal), nesse caso suspeita-se possível violação da hipótese de normalidade dos dados.


Normal Q-Q Plot
1732
1711
Sample Quantiles
15
10
5
0
-5
-10
-4
-2
0
2
4
Theoretical Quantiles