Relatório Final

Meari Caldeira & Thaís G. P. Faria

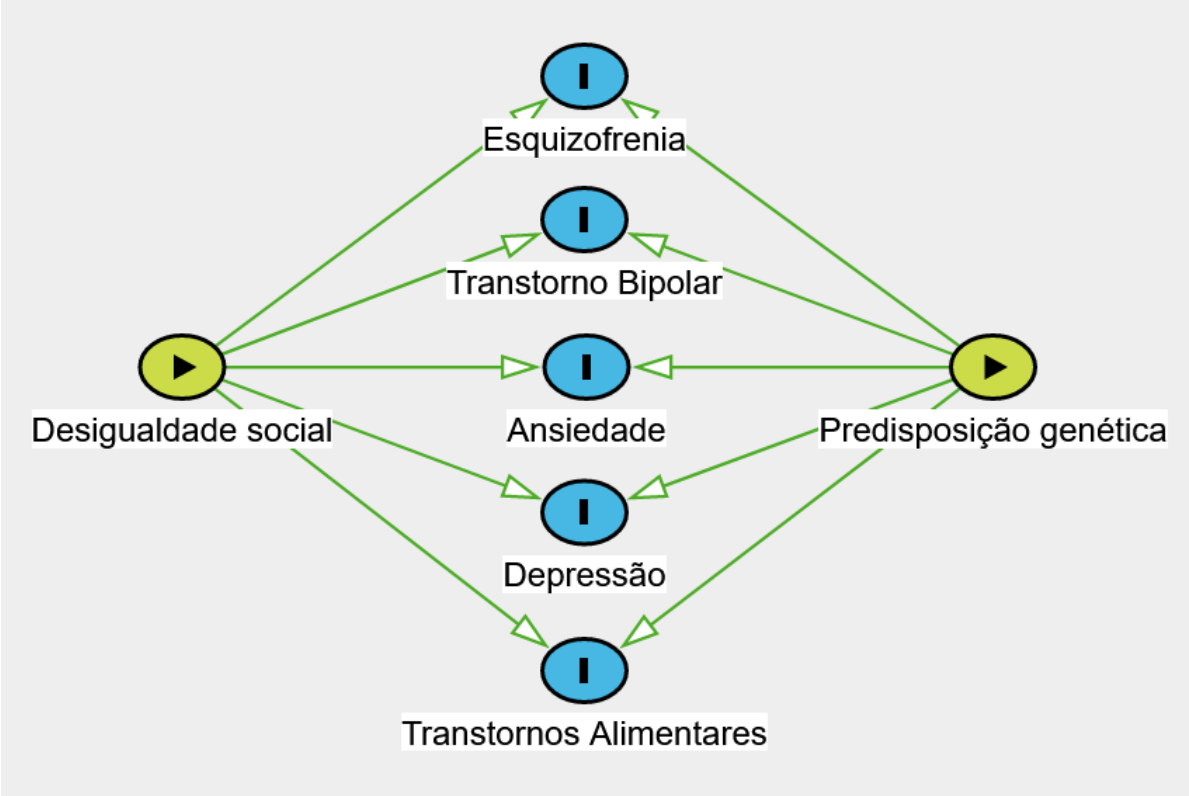
1 Introdução

Transtornos mentais impactam negativamente a vida de centenas de milhões de pessoas no mundo todo. Muito embora exista uma vasta literatura, principalmente em psiquiatria, sobre as origens e comorbidades dos diferentes transtornos mentais, não há consenso sobre as causas da maioria deles. Sabemos, no entanto (e inclusive por experiência própria), que fatores externos ao indivíduo acometido podem acarretar no desenvolvimento de transtornos mentais, por exemplo, péssimas condições de trabalho e insegurança social, que estão correlacionadas a ansiedade e depressão (Prins et al. (2015)), bem como fatores internos, como predisposição genética (Vereczkei and Mirnics (2011), Gordovez and McMahon (2020)).

Nos perguntamos, então, se transtornos mentais são causados por desigualdade social e/ou predisposição genética. Para responder a essa pergunta, selecionamos cinco transtornos mentais de grande prevalência na população mundial: esquizofrenia, transtorno bipolar, transtornos alimentares, ansiedade e depressão. Esperamos encontrar relação causal entre predisposição genética e todos os transtornos, no entanto, esperamos que ela seja mais forte para esquizofrenia, transtorno bipolar e transtornos alimentares, segundo aponta a literatura (Vereczkei and Mirnics (2011), Gordovez and McMahon (2020), @doi:10.1177/2045125318814734, @https://doi.org/10.1002/ajmg.c.30171), e também porque, como já citado, esperamos que depressão e ansiedade sejam causadas também por fatores socioeconômicos (Prins et al. (2015)). A nível populacional, esperamos também que exista alguma relação entre os transtornos, porque uma maior incidência de transtornos mentais na população pode tanto ser um sintoma de uma má situação social no país, gerando mais transtornos, ou causar problemas a nível individual que podem se traduzir em uma maior incidência de transtornos como ansiedade e depressão a nível populacional.

Usamos como indicador socioeconômico o índice de Gini, que mede a desigualdade social dos países, segundo dados do World Inequality Database, e como indicador de predisposição genética a população de cada país individualmente, embora conheçamos os riscos dessa escolha (veja a discussão mais à frente). Nossa hipótese é de que as relações de causalidade sejam independentes, ou seja, que a desigualdade social e a predisposição genética causem transtornos mentais independentemente uns dos outros, e não levamos em consideração outros fatores que

também podem causar esses transtornos a nível populacional, como criminalidade e drogadição. Nosso grafo acíclico direcionado (DAG) de causalidade para o cenário que imaginamos, então, é o seguinte:



Uma vez que os dados que temos são de prevalência de transtornos mentais em populações, julgamos que os modelos mais apropriados para a situação seriam Poisson e binomial, então vamos testá-los. Também testamos um modelo linear, que é costumeiramente empregado para fazer análises exploratórias.

2 Modelagem

2.1 Pacotes necessários

O pacote `here` (Müller 2020) é usado para facilitar a reprodutibilidade do script em diferentes sistemas operacionais fazendo uso do arquivo .RProj na raiz do diretório do projeto. O pacote `dplyr` (Wickham et al. 2023) faz parte do framework Tidyverse (Wickham et al. 2019), usado para algumas operações de limpeza e combinação de dados em conjunto com funções nativas da linguagem de programação R (2024). Já para a visualização dos dados, são usados os

pacotes `ggplot2` (Wickham 2016), `scales` (Wickham, Pedersen, and Seidel 2023), `cowplot` (Wilke 2024) e `patchwork` (Pedersen 2024).

```
# Carregando pacotes
library(here)
library(ggplot2)
library(scales)
library(cowplot)
library(patchwork)
```

2.2 Dados

2.2.1 Incidência de Transtornos Mentais

Os dados principais sobre incidência de transtornos mentais na população foram baixados da plataforma Kaggle, que disponibiliza conjuntos de dados para projetos de ciência de dados, e podem ser baixados através deste link. Os dados baixados do Kaggle foram compilados da literatura pelo Institute for Health Metrics and Evaluation (IHME)(n.d.), associado à University of Washington School of Medicine, e limpos e processados pela equipe do website Our World In Data Dattani et al. (2023) . Ele apresenta dados de prevalência de transtornos mentais (ansiedade, depressão, transtornos alimentares, esquizofrenia e transtorno bipolar) em 214 países entre 1990 a 2019.

Vamos primeiro carregar os dados:

```
# Leitura dos dados de prevalência
prevalence <- read.csv(here("data",
                            "raw",
                            "1-mental-illnesses-prevalence.csv"),
                       header=T, sep=",")
```

A tabela de dados brutos inclui alguns dados atribuídos a conjuntos de países, como "European Union" ou "High-income countries". Vamos filtrar a tabela para considerar apenas os países individuais:

```
countries <- unique(prevalence$Entity)
countries <- countries[-c(2,5,12,66,67,85,109,112,205,211)]

prevalence_by_country <- prevalence[which(prevalence$Entity %in% countries==T),]
```

Vamos combinar esses dados com o Índice de Gini a seguir.

2.2.2 Índice de Gini

O Índice de Gini foi proposto por Corrado Gini em 1912 no trabalho "Variabilità e Mutabilità" (Ceriani and Verme 2012) e é atualmente utilizado para quantificar o nível de desigualdade de renda de um país. Ele expressa a distância esperada ao acaso entre a renda de duas pessoas de uma população relativa à média de renda do país, de forma que valores mais próximos de 0 indicam maior igualdade de renda, e valores próximos de 1 indicam maior desigualdade. Os dados utilizados neste projeto foram obtidos da base World Inequality Database e processado pelo website Our World In Data, podendo ser acessados através deste link.

Primeiro, vamos carregar os dados:

```
gini <- read.csv(here("data",
                      "raw",
                      "economic-inequality-gini-index.csv"),
                 header=T, sep=",")
```

Agora vamos criar uma tabela combinando os dados de prevalência de transtornos mentais e o Índice de Gini. Vamos primeiro encontrar os países que constam tanto na tabela de prevalência quanto na tabela do Gini e criar uma base para a tabela `prevalence_with_gini`:

```
#
prevalence_with_gini <- data.frame(
  country=prevalence_by_country$Entity,
  year=prevalence_by_country[,3],
  schizophrenia=prevalence_by_country[,4],
  depression=prevalence_by_country[,5],
  anxiety=prevalence_by_country[,6],
  bipolar=prevalence_by_country[,7],
  ed=prevalence_by_country[,8],
  gini=rep(NA,length(prevalence_by_country$Entity))
)

countries_gini <- unique(gini$Entity)

prevalence_with_gini <- prevalence_with_gini[which(prevalence_with_gini$country %in% countrie

countries_prevalence <- unique(prevalence_with_gini$country)
```

Agora vamos combinar as tabelas com um `for loop` que itera entre cada país, filtrando os anos que apresentam tanto dados de prevalência quanto dados de Índice de Gini e combinando-os em uma tabela única (`prevalence_gini_intersection`):

```r
prevalence_gini_intersection <- data.frame(
  country=NA,
  year=NA,
  schizophrenia=NA,
  depression=NA,
  anxiety=NA,
  bipolar=NA,
  ed=NA,
  gini=NA
)

for(i in 1:length(countries_prevalence)){

  b <- gini[which(gini$Entity==countries_prevalence[i]),]

  a <- prevalence_with_gini[which(prevalence_with_gini$country==countries_prevalence[i]),]

  a <- a[which((a$year) %in% gini$Year[which(gini$Entity==countries_prevalence[i])]),]

  b <- b[b$Year %in% a$year,]

  a$gini <- b[which(b$Entity==countries_prevalence[i]),4]

  prevalence_gini_intersection <- rbind(prevalence_gini_intersection, a)

}

prevalence_gini_intersection <- prevalence_gini_intersection[-1,]
```

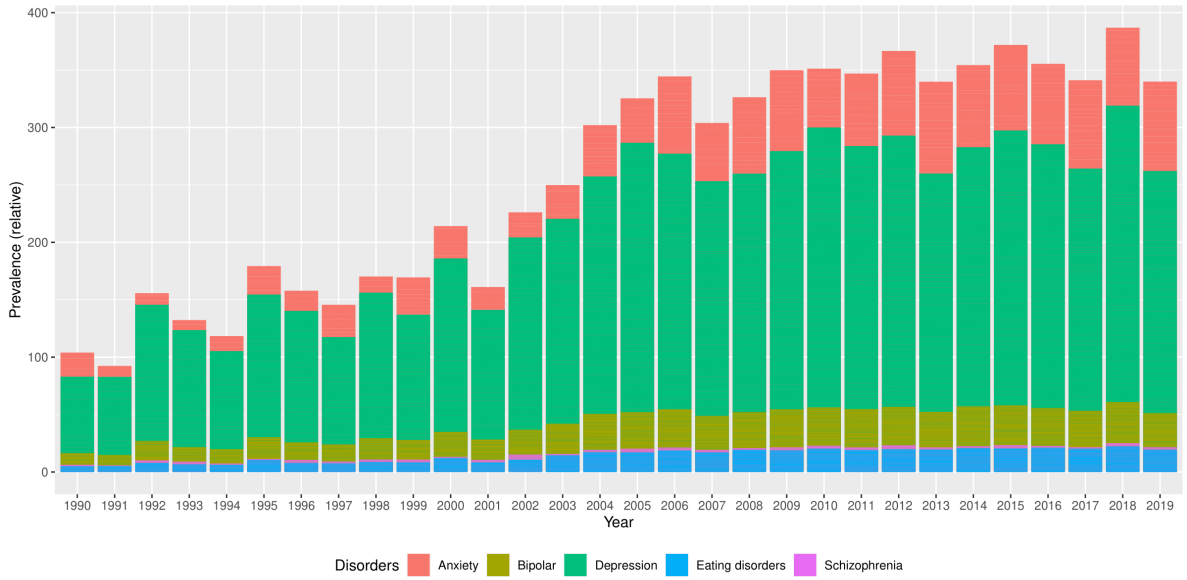
O código a seguir pode ser usado para salvar a tabela construída até o momento, e para carregá-la novamente quando necessário:

```r
write.table(prevalence_gini_intersection,
            file = here("data",
                        "processed",
                        "relative_world_prevalence_with_gini.txt"),
            sep="\t",
            row.names=F)

relative_gini <- read.table(here("data",
                                 "processed",
                                 "relative_world_prevalence_with_gini.txt"),
```

```
                                sep="\t",
                                header=T)
```

A tabela combinada é composta pelos dados de prevalência dos transtornos mentais por país, por ano, em porcentagem, mas é mais fácil ter uma ideia inicial dos padrões dos dados através de gráficos, então vamos fazer um combinando os dados de todos os países:

```
relative_gini$year <- as.factor(relative_gini$year)

ggplot(relative_gini) +
        geom_bar(aes(x=year, y=anxiety, fill="Anxiety"), stat="identity") +
        geom_bar(aes(x=year, y=depression, fill="Depression"), stat="identity") +
        geom_bar(aes(x=year, y=bipolar, fill="Bipolar"), stat="identity") +
        geom_bar(aes(x=year, y=schizophrenia, fill="Schizophrenia"), stat="identity") +
        geom_bar(aes(x=year, y=ed, fill="Eating disorders"), stat="identity") +
        labs(x="Year", y="Prevalence (relative)", fill="Disorders") +
        theme(legend.position="bottom")
```



Considerando os dados de todos os países juntos por ano, podemos observar que depressão é o transtorno mental com a maior prevalência durante todo o período analisado, enquanto ansiedade é o segundo transtorno com maior prevalência pelo menos desde 2006. Por outro lado, esquizofrenia e transtornos alimentares consistentemente têm as menores prevalências. No entanto, os modelos estatísticos utilizados nas próximas análises requerem dados de contagem absoluta da população, então vamos utilizá-los nas próximas etapas da construção dos dados finais.

2.2.3 Dados de população absoluta

Os dados foram obtidos da base de dados DataBank, mantida pelo Banco Mundial com dados de censos de de diversas fontes, obtidos principalmente através de programas da ONU. Ele podem ser acessados e baixados através deste link.

Vamos primeiro ler os dados e filtrar os países para os quais temos o Índice de Gini:

```
worldbank_1 <- read.csv(here("data",
                             "raw",
                             "worldbank",
                             "WDICSV.csv"),
                        sep=",",
                        header=T)

# Coluna com população total
pop <- worldbank_1[which(worldbank_1$Indicator.Name=="Population, total"),]

# Seleção dos países com Índice de GINI
pop <- pop[which(pop$Country.Name %in% relative_gini$country),]
```

Agora vamos transformar os dados de prevalência em porcentagem de cada transtorno em dados de contagem populacional para permitir a modelagem estatística com modelos de contagem:

```
# Criando uma nova tabela para os dados absolutos
absolute_gini <- relative_gini

# For Loop
for(j in 1:length(pop$Country.Name)){
  a <- relative_gini[(grep(pop$Country.Name[j], relative_gini$country)),]
  b <- pop[j,]
  year <- a$year
  c <- colnames(b)
  c <- sub("X", "", c)
  colnames(b) <- c
  b <- b[,c(1,which(c %in% year))]

  for(i in 2:length(colnames(b))){
    a$schizophrenia[which(a$year == colnames(b)[i])] <- as.numeric(a$schizophrenia[which(a$y
    a$depression[which(a$year == colnames(b)[i])] <- as.numeric(a$depression[which(a$year ==
    a$anxiety[which(a$year == colnames(b)[i])] <- as.numeric(a$anxiety[which(a$year == colnar
    a$bipolar[which(a$year == colnames(b)[i])] <- as.numeric(a$bipolar[which(a$year == colnar
    a$ed[which(a$year == colnames(b)[i])] <- as.numeric(a$ed[which(a$year == colnames(b)[i])]
```

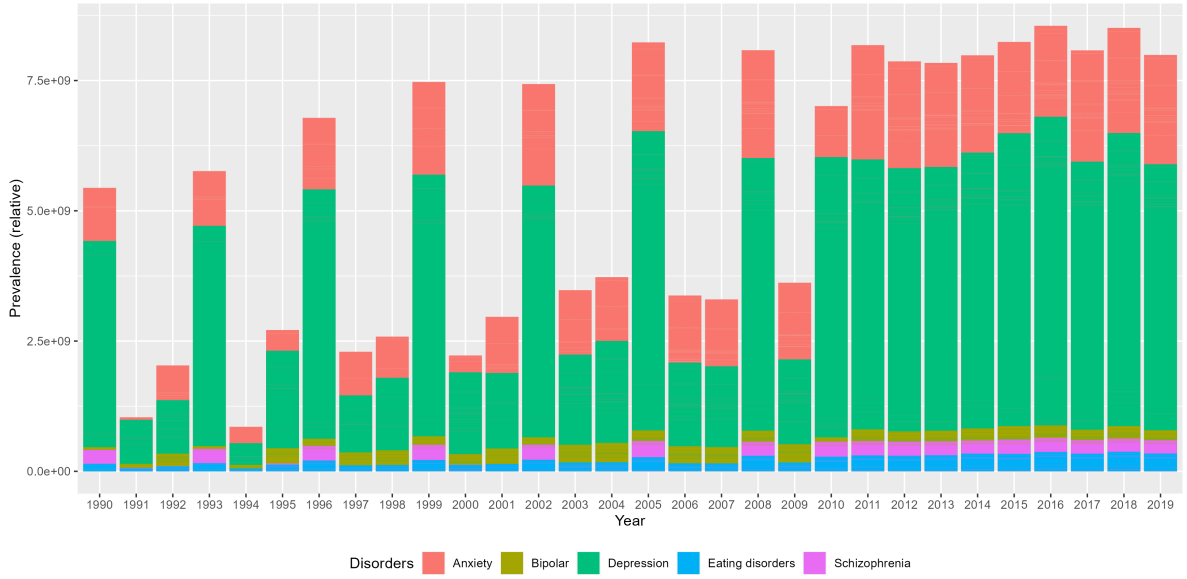
```
  }

  absolute_gini[(grep(pop$Country.Name[j], relative_gini$country)),] <- a

}

# Removendo linhas sem dados populacionais
absolute_gini <- absolute_gini[-which(absolute_gini$schizophrenia < 1),]
```

O código a seguir pode ser usado para salvar a tabela com os dados de população absoluta:

```
write.table(absolute_gini,
            file=here("data",
                      "processed",
                      "absolute_world_prevalence_with_gini.txt"),
            sep="\t",
            row.names=F)

absolute_gini <- read.table(here("data",
                                 "processed",
                                 "absolute_world_prevalence_with_gini.txt"),
                            header=T,
                            sep="\t")
```

Vamos agora fazer um novo gráfico para comparar os dados relativos com os dados absolutos, considerando todos os países com dados disponíveis juntos:

```
absolute_gini$year <- as.factor(absolute_gini$year)

ggplot(absolute_gini) +
        geom_bar(aes(x=year, y=anxiety, fill="Anxiety"), stat="identity") +
        geom_bar(aes(x=year, y=depression, fill="Depression"), stat="identity") +
        geom_bar(aes(x=year, y=bipolar, fill="Bipolar"), stat="identity") +
        geom_bar(aes(x=year, y=schizophrenia, fill="Schizophrenia"), stat="identity") +
        geom_bar(aes(x=year, y=ed, fill="Eating disorders"), stat="identity") +
        labs(x="Year", y="Prevalence (relative)", fill="Disorders") +
        theme(legend.position="bottom")
```



O gráfico apresenta algumas colunas maiores que outras, pois nem todos os anos possuem dados de população para todos os países. Por exemplo, não há dados para países populosos, como Bangladesh, no ano de 1994. Essa discrepância não afetará as análises, pois elas serão feitas com a média por país por década. Em relação ao

2.2.4 Separação dos dados em décadas

Para simplificar nossa análise, que atualmente é composta por 28 pontos por país para cada transtorno (um por ano entre 1991 e 2019), vamos separar nossos dados por década e extrair a média da prevalência de cada transtorno e do índice de Gini dentro da década. Dessa forma, teremos no máximo 3 pontos por país para cada transtorno.

Vamos começar criando três objetos, `decade1`, `decade2` e `decade3`, para guardar nossos dados de cada década, e três objetos análogos para guardar os dados de médias:

```
decade1 <- absolute_gini[which(absolute_gini$year < 2000),]
decade2 <- absolute_gini[which(absolute_gini$year > 1999 & absolute_gini$year < 2010),]
decade3 <- absolute_gini[which(absolute_gini$year > 2009),]

countries <- unique(decade1$country)

decade1_mean <- data.frame(country=rep(NA, length(countries)),
                           schizophrenia=rep(NA, length(countries)),
                           depression=rep(NA, length(countries)),
                           anxiety=rep(NA, length(countries)),
```

```
                            bipolar=rep(NA, length(countries)),
                            ed=rep(NA, length(countries)),
                            gini=rep(NA, length(countries))
)


decade2_mean <- data.frame(country=rep(NA, length(countries)),
                            schizophrenia=rep(NA, length(countries)),
                            depression=rep(NA, length(countries)),
                            anxiety=rep(NA, length(countries)),
                            bipolar=rep(NA, length(countries)),
                            ed=rep(NA, length(countries)),
                            gini=rep(NA, length(countries))
)


decade3_mean <- data.frame(country=rep(NA, length(countries)),
                            schizophrenia=rep(NA, length(countries)),
                            depression=rep(NA, length(countries)),
                            anxiety=rep(NA, length(countries)),
                            bipolar=rep(NA, length(countries)),
                            ed=rep(NA, length(countries)),
                            gini=rep(NA, length(countries))
)
```

Agora, vamos criar for loops para preencher esses objetos de médias com as médias por década:

```
for(i in 1:length(countries)){

  decade1_mean$country[i] <- countries[i]
  decade1_mean$schizophrenia[i] <- mean(decade1$schizophrenia[which(decade1$country==countrie
  decade1_mean$depression[i] <- mean(decade1$depression[which(decade1$country==countries[i])]
  decade1_mean$anxiety[i] <- mean(decade1$anxiety[which(decade1$country==countries[i])])
  decade1_mean$bipolar[i] <- mean(decade1$bipolar[which(decade1$country==countries[i])])
  decade1_mean$ed[i] <- mean(decade1$ed[which(decade1$country==countries[i])])
  decade1_mean$gini[i] <- mean(decade1$gini[which(decade1$country==countries[i])])

}

decade1_mean <- na.exclude(decade1_mean)
```

```
for(i in 1:length(countries)){

  decade2_mean$country[i] <- countries[i]
  decade2_mean$schizophrenia[i] <- mean(decade2$schizophrenia[which(decade2$country==countrie
  decade2_mean$depression[i] <- mean(decade2$depression[which(decade2$country==countries[i])]
  decade2_mean$anxiety[i] <- mean(decade2$anxiety[which(decade2$country==countries[i])])
  decade2_mean$bipolar[i] <- mean(decade2$bipolar[which(decade2$country==countries[i])])
  decade2_mean$ed[i] <- mean(decade2$ed[which(decade2$country==countries[i])])
  decade2_mean$gini[i] <- mean(decade2$gini[which(decade2$country==countries[i])])

}

decade2_mean <- na.exclude(decade2_mean)
```

```
for(i in 1:length(countries)){

  decade3_mean$country[i] <- countries[i]
  decade3_mean$schizophrenia[i] <- mean(decade3$schizophrenia[which(decade3$country==countrie
  decade3_mean$depression[i] <- mean(decade3$depression[which(decade3$country==countries[i])]
  decade3_mean$anxiety[i] <- mean(decade3$anxiety[which(decade3$country==countries[i])])
  decade3_mean$bipolar[i] <- mean(decade3$bipolar[which(decade3$country==countries[i])])
  decade3_mean$ed[i] <- mean(decade3$ed[which(decade3$country==countries[i])])
  decade3_mean$gini[i] <- mean(decade3$gini[which(decade3$country==countries[i])])

}

decade3_mean <- na.exclude(decade3_mean)
```

E vamos resumir tudo isso em um só objeto, `decades`:

```
decade1_mean$decade <- "1990-1999"
decade2_mean$decade <- "2000-2009"
decade3_mean$decade <- "2010-2019"

decades <- rbind(decade1_mean, decade2_mean, decade3_mean)

decades <- decades[order(decades$country),]
```

Acabamos não precisando disso, mas categorizamos os países em continentes também. Não vamos incluir o código para fazer isso aqui, porque não é necessário para a análise, mas você vai encontrar essa informação na tabela final, se a abrir.

Por fim, vamos salvar essa tabela, para podermos acessá-la facilmente quando quisermos:

```
write.table(decades, file=here("data", "processed", "decade_means_absolute.txt"), sep="\t", r
```

Com isso, nossos dados estão prontos para fazermos as análises de modelagem!

2.3 Ajuste de modelos e diagnóstico

2.3.1 Correlação linear entre transtornos

Como não estamos avaliando causalidade entre transtornos, somente a correlação, usamos uma correlação linear simples para testar se os transtornos mentais estão relacionados entre si. Vamos começar fazendo essa avaliação carregando a tabela de médias por década com dados absolutos populacionais, e guardá-la no objeto `decades`:

```
decades <- read.table(here("data","processed","decade_means_absolute.txt"), header=T)
```

Agora, como não estamos interessados na correlação entre transtornos ao longo do tempo, vamos fazer a média de prevalência dos transtornos por país, de forma a obter apenas um valor de prevalência na população por país:

```
decades$country <- as.factor(decades$country)

schizophrenia <- aggregate(decades$schizophrenia~factor(decades$country), FUN=mean)
depression <- aggregate(decades$depression~factor(decades$country), FUN=mean)
anxiety <- aggregate(decades$anxiety~factor(decades$country), FUN=mean)
bipolar <- aggregate(decades$bipolar~factor(decades$country), FUN=mean)
ed <- aggregate(decades$ed~factor(decades$country), FUN=mean)
gini <- aggregate(decades$gini~factor(decades$country), FUN=mean)
```

Por fim, vamos fazer os modelos lineares, guardando-os em objetos cujo nome é `lm_inicial do primeiro transtorno_inicial do segundo transtorno`:

```
lm_sd <- lm(schizophrenia$`decades$schizophrenia`~depression$`decades$depression`)
lm_sa <- lm(schizophrenia$`decades$schizophrenia`~anxiety$`decades$anxiety`)
lm_sb <- lm(schizophrenia$`decades$schizophrenia`~bipolar$`decades$bipolar`)
lm_se <- lm(schizophrenia$`decades$schizophrenia`~ed$`decades$ed`)
lm_da <- lm(depression$`decades$depression`~anxiety$`decades$anxiety`)
lm_db <- lm(depression$`decades$depression`~bipolar$`decades$bipolar`)
lm_de <- lm(depression$`decades$depression`~ed$`decades$ed`)
lm_ab <- lm(anxiety$`decades$anxiety`~bipolar$`decades$bipolar`)
lm_ae <- lm(anxiety$`decades$anxiety`~ed$`decades$ed`)
lm_be <- lm(bipolar$`decades$bipolar`~ed$`decades$ed`)
```

Aplicando a função **summary** às correlações, obtemos os R^2 ajustados e a significância da correlação, resumidos na tabela abaixo:

```
summary(lm_sd)
summary(lm_sa)
summary(lm_sb)
summary(lm_se)
summary(lm_da)
summary(lm_db)
summary(lm_de)
summary(lm_ab)
summary(lm_ae)
summary(lm_be)
```

	Esquizofrenia	Depressão	Ansiedade	Bipolar	Alimentares
Esquizofrenia	-	0.9527 ***	0.9737 ***	0.7666 ***	0.7666 ***
Depressão	-	-	0.929 ***	0.8587 ***	0.7547 ***
Ansiedade	-	-	-	0.8328 ***	0.8095 ***
Bipolar	-	-	-	-	0.8251 ***

Três asteriscos indicam correlações significativas (p-valor $< 0{,}001$).

Dessa forma, todos os transtornos parecem estar significativamente correlacionados entre si, sendo as correlações entre esquizofrenia e ansiedade, esquizofrenia e depressão e depressão e ansiedade são as mais fortes. Isso não necessariamente indica causalidade, porque o que estamos avaliando aqui, novamente, não é uma relação causal, e sim, uma correlação. Também não indica que indivíduos esquizofrênicos costumam ter mais depressão e ansiedade, ou que indivíduos deprimidos são mais ansiosos, pois é um efeito populacional. O que esse valor nos diz é que populações mais esquizofrênicas também tendem a ser mais ansiosas e deprimidas, e populações mais deprimidas tendem a ser mais ansiosas e vice-versa.

2.3.2 Teste de causalidade entre transtornos, desigualdade social e predisposição genética

Para testarmos a causalidade entre nossos transtornos, a desigualdade social e a predisposição genética, vamos precisar de mais alguns pacotes:

```
library(bbmle)
library(sads)
library(DHARMa)
```

Também vamos precisar dos dados absolutos, e de transformar as décadas em fatores numéricos. Isso é importante porque o texto separado por hífen pode nos causar problemas em diante:

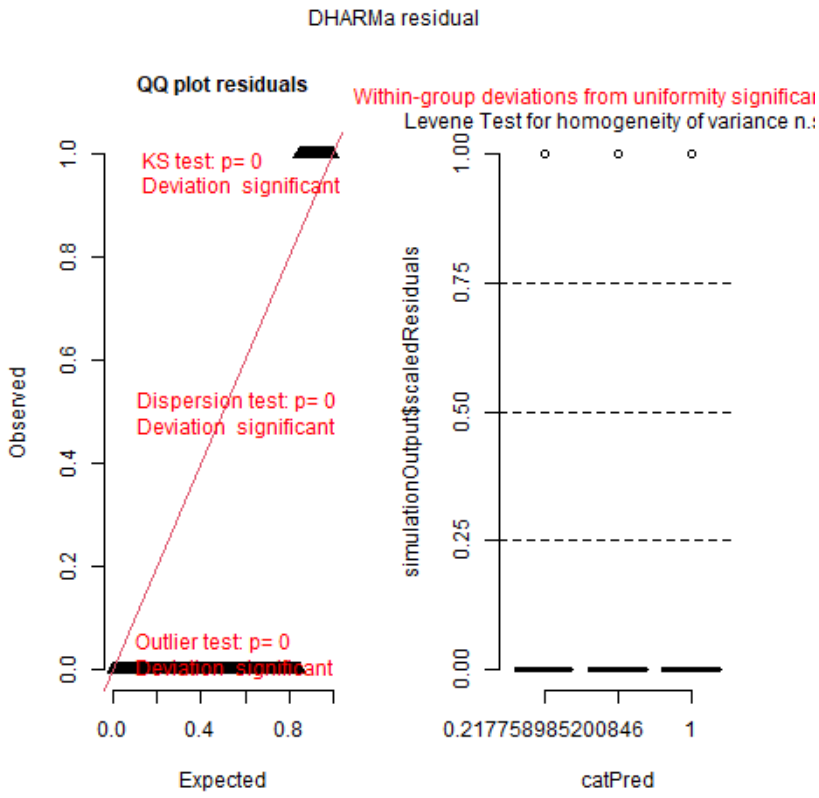
```
abs_decades <- read.table(here("data","processed","decade_means_absolute.txt"), header=T)

abs_decades$decade[which(abs_decades$decade=="1990-1999")] <- 1
abs_decades$decade[which(abs_decades$decade=="2000-2009")] <- 2
abs_decades$decade[which(abs_decades$decade=="2010-2019")] <- 3
abs_decades$decade <- as.factor(abs_decades$decade)
```

Vamos colocar na tabela de dados absolutos também os dados de população como uma coluna nova:

```
worldbank_1 <- read.csv(here("data","raw","worldbank","WDICSV.csv"), sep=",", header=T)
pop <- worldbank_1[which(worldbank_1$Indicator.Name=="Population, total"),] # only populatio
pop <- pop[which(pop$Country.Name %in% abs_decades$country),]
pop <- t(pop)

colnames(pop) <- pop[1,]
pop <- as.data.frame(pop)
pop <- pop[5:68,]
pop[65,] <- colnames(pop)
pop$year <- rownames(pop)
pop$year <- sub("X", "", pop$year)

abs_decades$pop <- rep(NA, length(abs_decades$country))

countries <- as.character(pop[65,])

pop <- pop[1:64,]

pop <- apply(pop, FUN=as.numeric, MARGIN=c(1,2))
```

```
pop <- data.frame(pop)

for(i in 1:length(abs_decades$country)){

  abs_decades$pop[which(abs_decades$country==countries[i] & abs_decades$decade==1)] <- mean(p

  abs_decades$pop[which(abs_decades$country==countries[i] & abs_decades$decade==2)] <- mean(p

  abs_decades$pop[which(abs_decades$country==countries[i] & abs_decades$decade==3)] <- mean(p

}
```

IMPORTANTE! Note que, daqui em diante, precisamos logaritmizar os dados de prevalência de transtornos e de população e também arredondar os logaritmos. Não gostamos dessa solução, mas, quando tentamos fazer as análises sem logaritmos, obtivemos gráficos um tanto quanto bizarros, como é possível ver no gráfico dos resíduos do DHARMa da distribuição Poisson sem a transformação logarítmica e arredondamento dos dados:



Concluímos que isso se devia ao fato de que alguns países têm populações enormes, como a Índia, e outros têm populações minúsculas, como Vanuatu. O logaritmo transforma todas as populações para a mesma escala, permitindo uma análise mais pareada. O arredondamento se deveu ao fato de que os modelos binomiais não rodavam com casas após a vírgula, e os modelos Poisson ficavam com vários NAs. Não sabemos muito bem o que mais poderíamos ter feito nessa situação, mas gostaríamos de deixar claro que não gostamos dessa solução.

Caso você queira ver os gráficos gerados com os dados sem transformações, pode encontrá-los em `images -> models`, com nomes que começam com `NOLOG_NOROUND`.

Vamos fazer a transformação citada acima: logaritmizar os dados de prevalência e população e arredondá-los.

```
abs_decades$schizophrenia <- log(abs_decades$schizophrenia)
abs_decades$depression <- log(abs_decades$depression)
abs_decades$anxiety <- log(abs_decades$anxiety)
abs_decades$bipolar <- log(abs_decades$bipolar)
abs_decades$ed <- log(abs_decades$ed)
abs_decades$pop <- log(abs_decades$pop)

abs_decades$schizophrenia <- round(abs_decades$schizophrenia)
abs_decades$depression <- round(abs_decades$depression)
abs_decades$anxiety <- round(abs_decades$anxiety)
abs_decades$bipolar <- round(abs_decades$bipolar)
abs_decades$ed <- round(abs_decades$ed)
abs_decades$pop <- round(abs_decades$pop)
```

O próximo passo é a modelagem. Essa parte é extensa e será demonstrada somente para a esquizofrenia, mas você pode encontrar o código para todos os transtornos no arquivo `glm.R`.

Vamos começar fazendo modelos linear, Poisson e binomial para a esquizofrenia como variável resposta ao tempo:

```
glm.linear.year <- lm(schizophrenia ~ decade,
                      data=abs_decades)

glm.pois.year <- glm(schizophrenia ~ decade,
                     data=abs_decades,
                     family=poisson(link="log"))

glm.bin.year <- glm(cbind(schizophrenia,pop) ~ decade,
```

```
                data=abs_decades,
                family=binomial(link="logit"))
```

Vamos plotar os resíduos desses modelos usando o DHARMa:

```
sim_res_lm <- simulateResiduals(fittedModel = glm.linear.year)
plot(sim_res_lm)
```

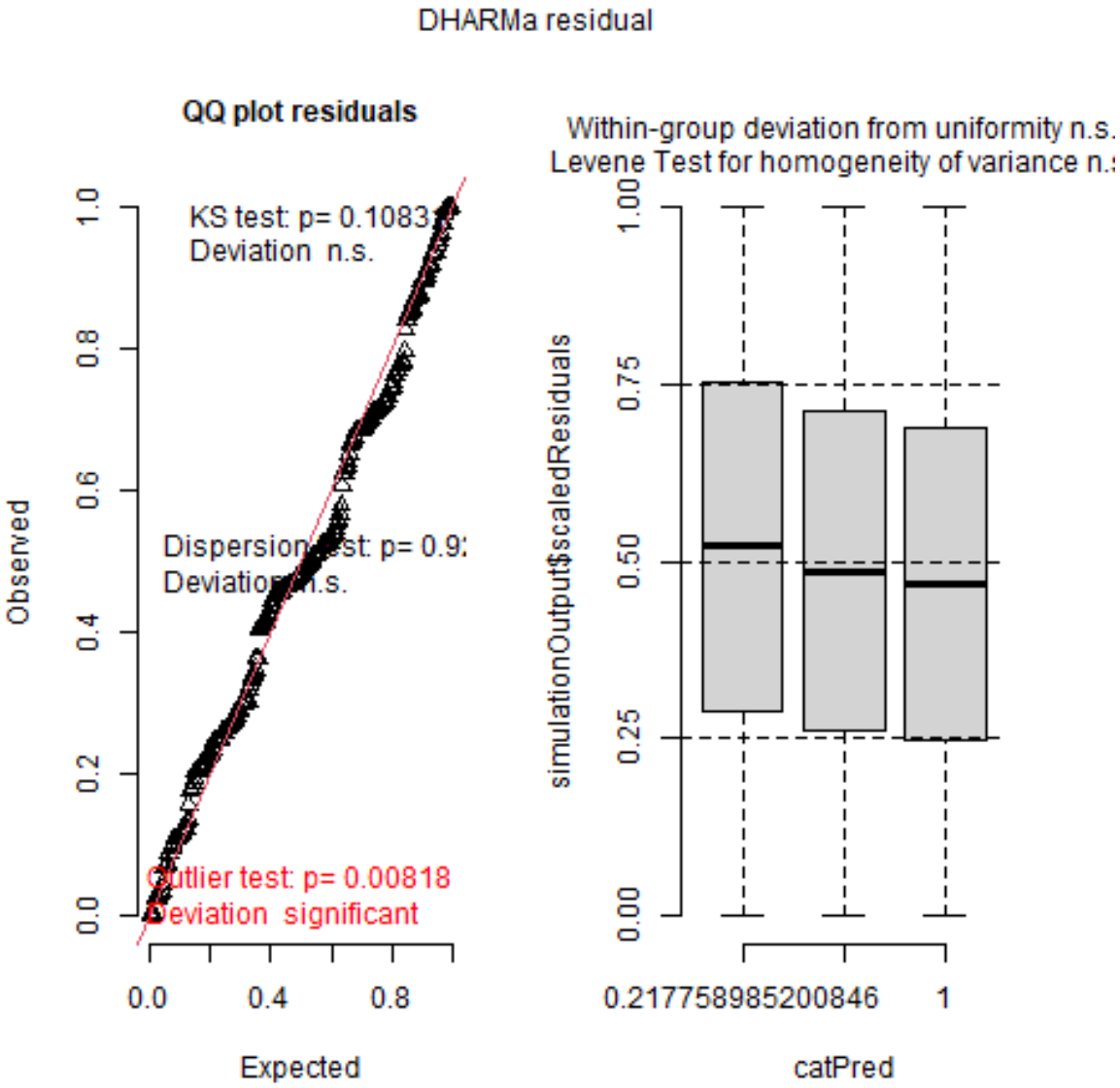


Figure 1: Esquizofrenia, Linear, Tempo

```
sim_res_pois <- simulateResiduals(fittedModel = glm.pois.year)
plot(sim_res_pois)
```

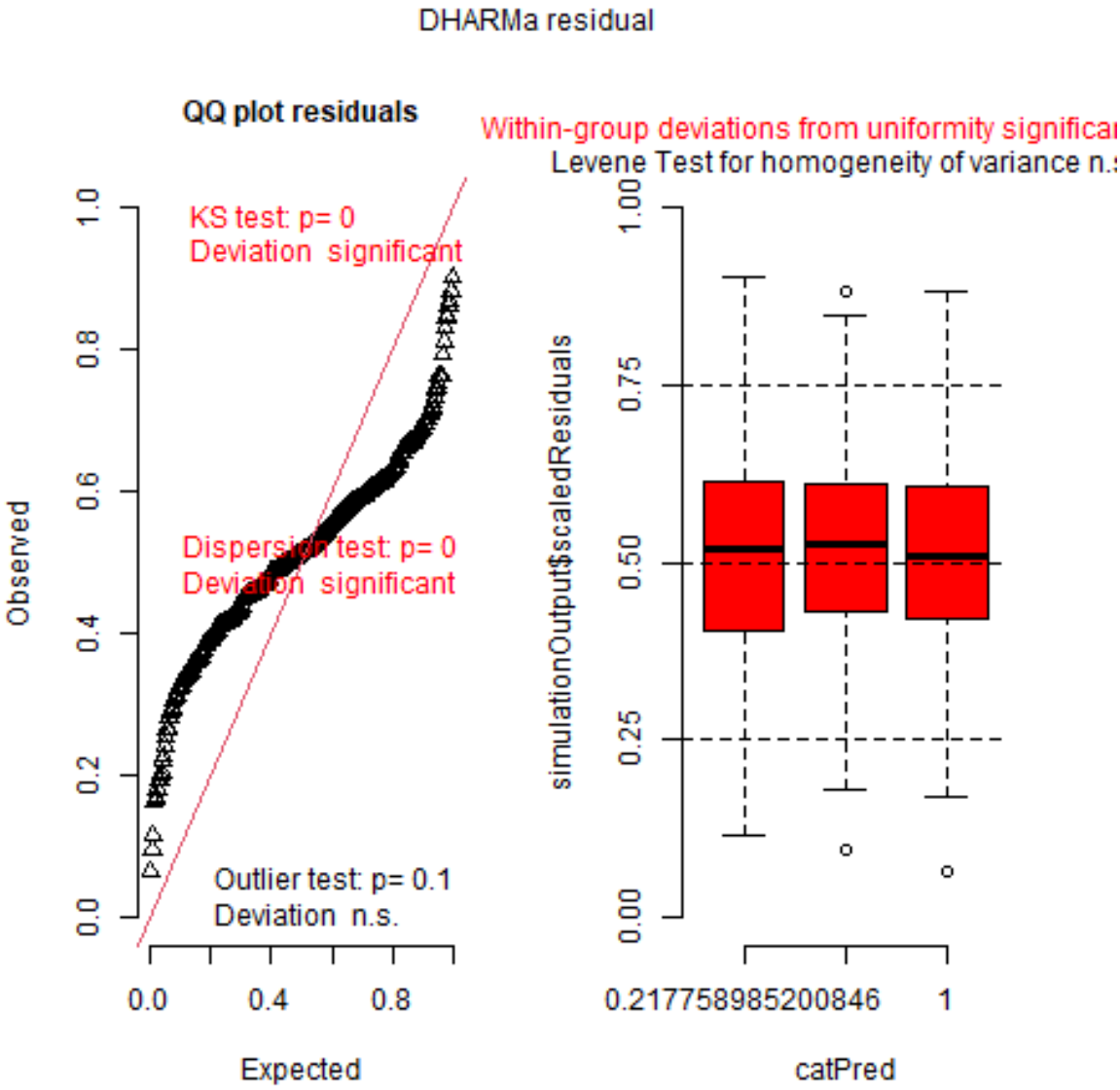


Figure 2: Esquizofrenia, Poisson, Tempo

```
sim_res_bin <- simulateResiduals(fittedModel = glm.bin.year)
plot(sim_res_bin)
```

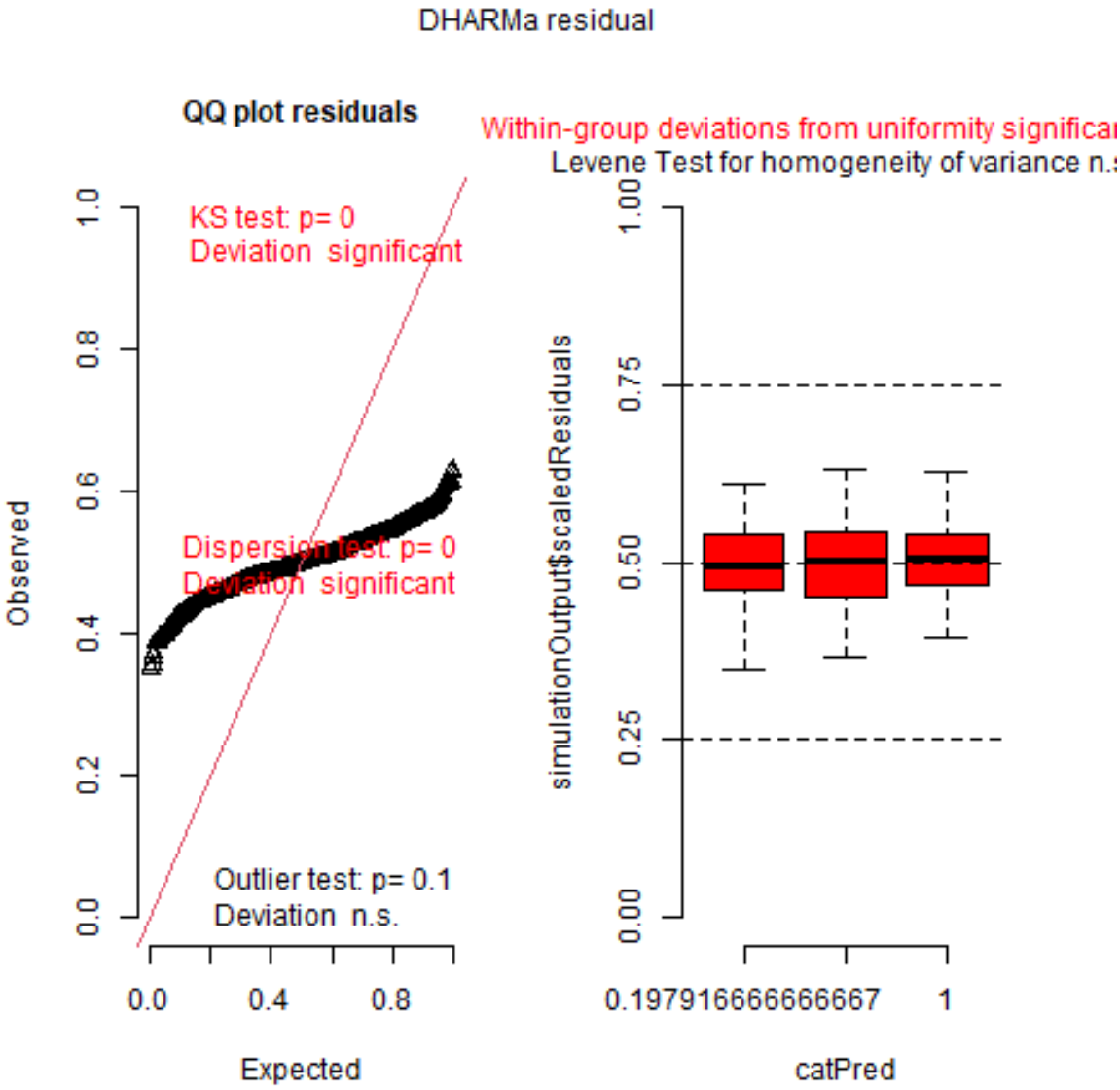


Figure 3: Esquizofrenia, Binomial, Tempo

O modelo linear parece ser o que melhor se ajusta aos nossos dados, segundo o DHARMa, porque os resíduos estão mais alinhados com a bissetriz (em vermelho) no QQ plot. Isso se confirma nos modelos subsequentes, tanto na análise visual quanto pelo AIC e log-verossimilhança (que serão apresentados a seguir). Como nós geramos as imagens do DHARMa e também dos resíduos dos modelos de forma clássica, temos muitas imagens (mais de 100!), por isso, vamos apresentar, daqui em diante, somente as imagens dos modelos lineares. Você pode ver as

imagens dos modelos Poisson e binomiais na pasta `images -> models`.

Agora, vamos testar um modelo em que a esquizofrenia é uma variável resposta do tempo e também do índice de Gini, testando, assim, a causalidade entre a desigualdade social e a prevalência de esquizofrenia. Estamos procurando por um desvio de resíduos aqui: a diferença entre os resíduos do modelo anterior e desse corresponde à causalidade entre a desigualdade social e a esquizofrenia, portanto, procuramos pelo modelo que tenha resíduos mais diferentes do anterior.

```
glm.linear.gini <- lm(schizophrenia ~ decade + gini,
                      data=abs_decades)

glm.pois.gini <- glm(schizophrenia ~ decade + gini,
                     data=abs_decades,
                     family=poisson(link="log"))

glm.bin.gini <- glm(cbind(schizophrenia,pop) ~ decade + gini,
                    data=abs_decades,
                    family=binomial(link="logit"))
```

Vamos visualizar esses modelos. A partir de agora, vamos apresentar os gráficos do DHARMa para os modelos lineares logo após o código de plotagem, e a plotagem dos demais modelos será apresentada somente em código.

```
sim_res_lm <- simulateResiduals(fittedModel = glm.linear.gini)
plot(sim_res_lm)
sim_res_pois <- simulateResiduals(fittedModel = glm.pois.gini)
plot(sim_res_pois)
sim_res_bin <- simulateResiduals(fittedModel = glm.bin.gini)
plot(sim_res_bin)
```

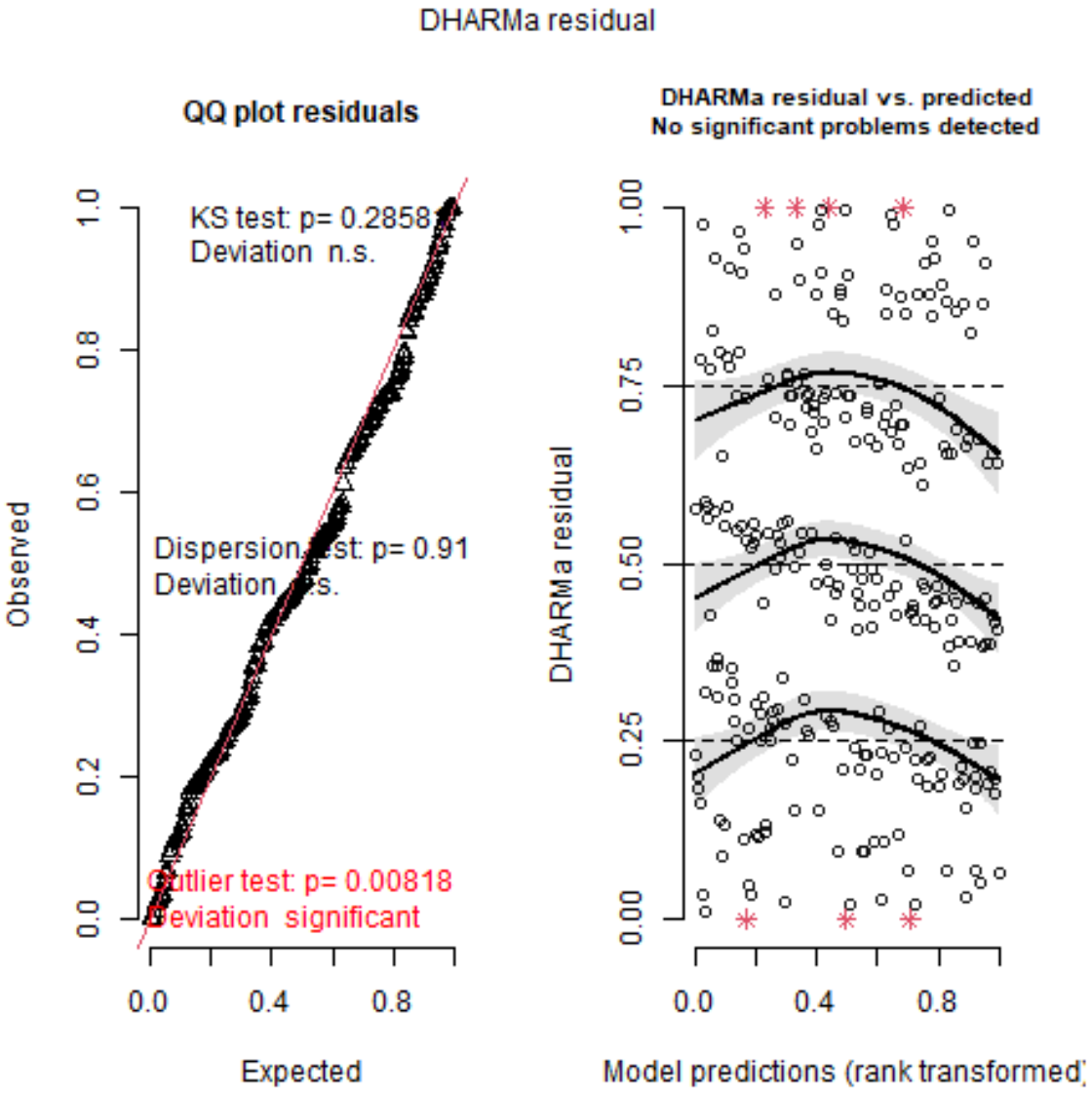


Figure 4: Esquizofrenia, Linear, Tempo + Gini

O ajuste melhorou, mas, além disso, os resíduos não parecem ter mudado muito. Talvez a predisposição genética, representada aqui por cada país, nos diga algo sobre isso:

```
glm.linear.country <- lm(schizophrenia ~ decade + country,
                         data=abs_decades)
```

```r
glm.pois.country <- glm(schizophrenia ~ decade + country,
                        data=abs_decades,
                        family=poisson(link="log"))

glm.bin.country <- glm(cbind(schizophrenia,pop) ~ decade + country,
                       data=abs_decades,
                       family=binomial(link="logit"))

sim_res_lm <- simulateResiduals(fittedModel = glm.linear.country)
plot(sim_res_lm)
sim_res_pois <- simulateResiduals(fittedModel = glm.pois.country)
plot(sim_res_pois)
sim_res_bin <- simulateResiduals(fittedModel = glm.bin.country)
plot(sim_res_bin)
```

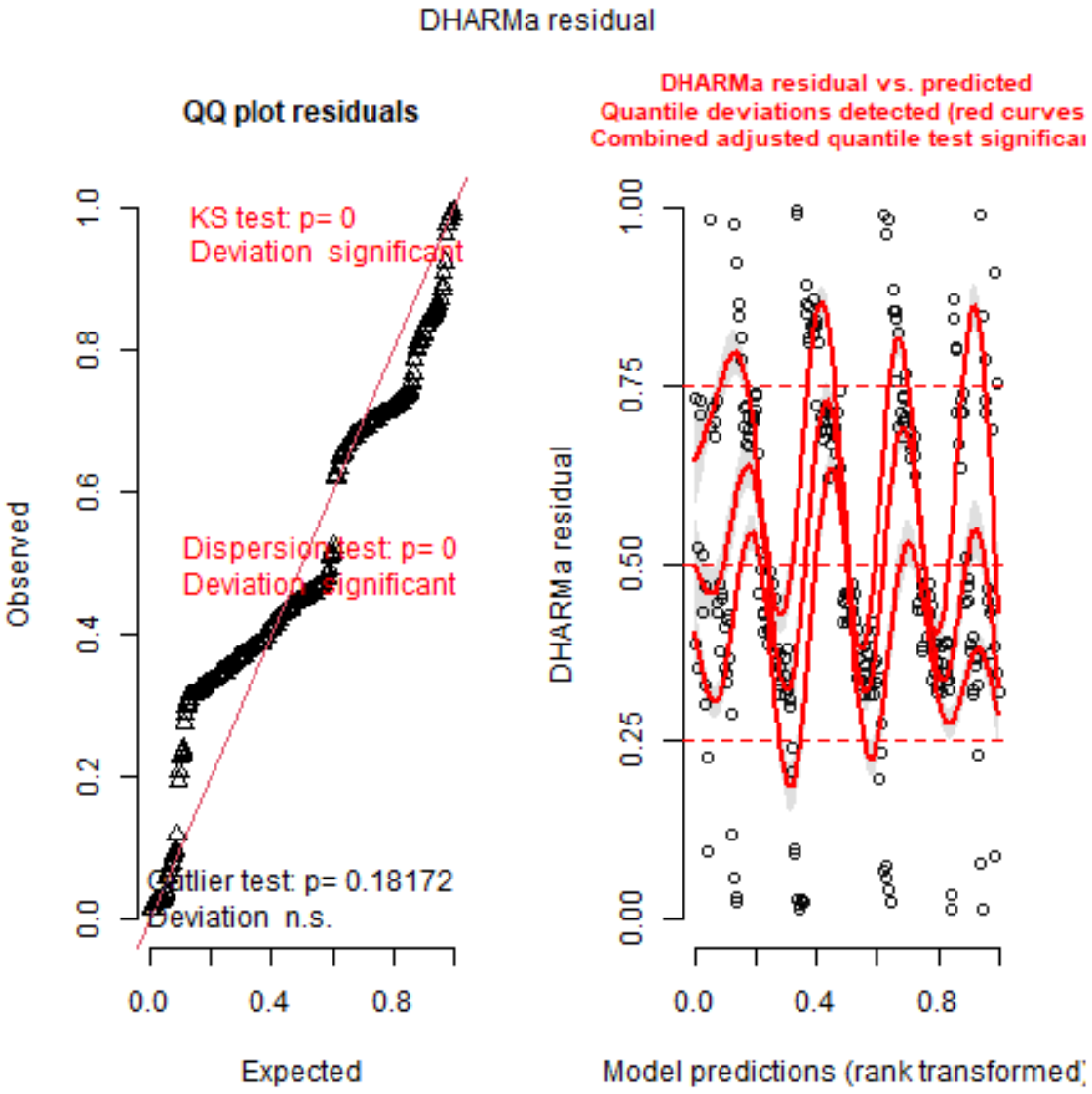


Figure 5: Esquizofrenia, Linear, Tempo + país

Os resíduos estão bem diferentes agora! Parece que o país onde uma população está tem uma relação de causalidade com a prevalência de esquizofrenia nela.

Não estava na nossa hipótese inicial, mas ficamos curiosos para saber se a desigualdade e o país onde a população está podem estar causando transtornos conjuntamente, e não de forma independente. Por isso, fizemos também uma quarta categoria de modelos, na qual juntamos os dois anteriores:

```r
glm.linear.country.gini <- lm(schizophrenia ~ decade + country + gini,
                          data=abs_decades)

glm.pois.country.gini <- glm(schizophrenia ~ decade + country + gini,
                         data=abs_decades,
                         family=poisson(link="log"))

glm.bin.country.gini <- glm(cbind(schizophrenia,pop) ~ decade + country + gini,
                        data=abs_decades,
                        family=binomial(link="logit"))

sim_res_lm <- simulateResiduals(fittedModel = glm.linear.country.gini)
plot(sim_res_lm)
sim_res_pois <- simulateResiduals(fittedModel = glm.pois.country.gini)
plot(sim_res_pois)
sim_res_bin <- simulateResiduals(fittedModel = glm.bin.country.gini)
plot(sim_res_bin)
```

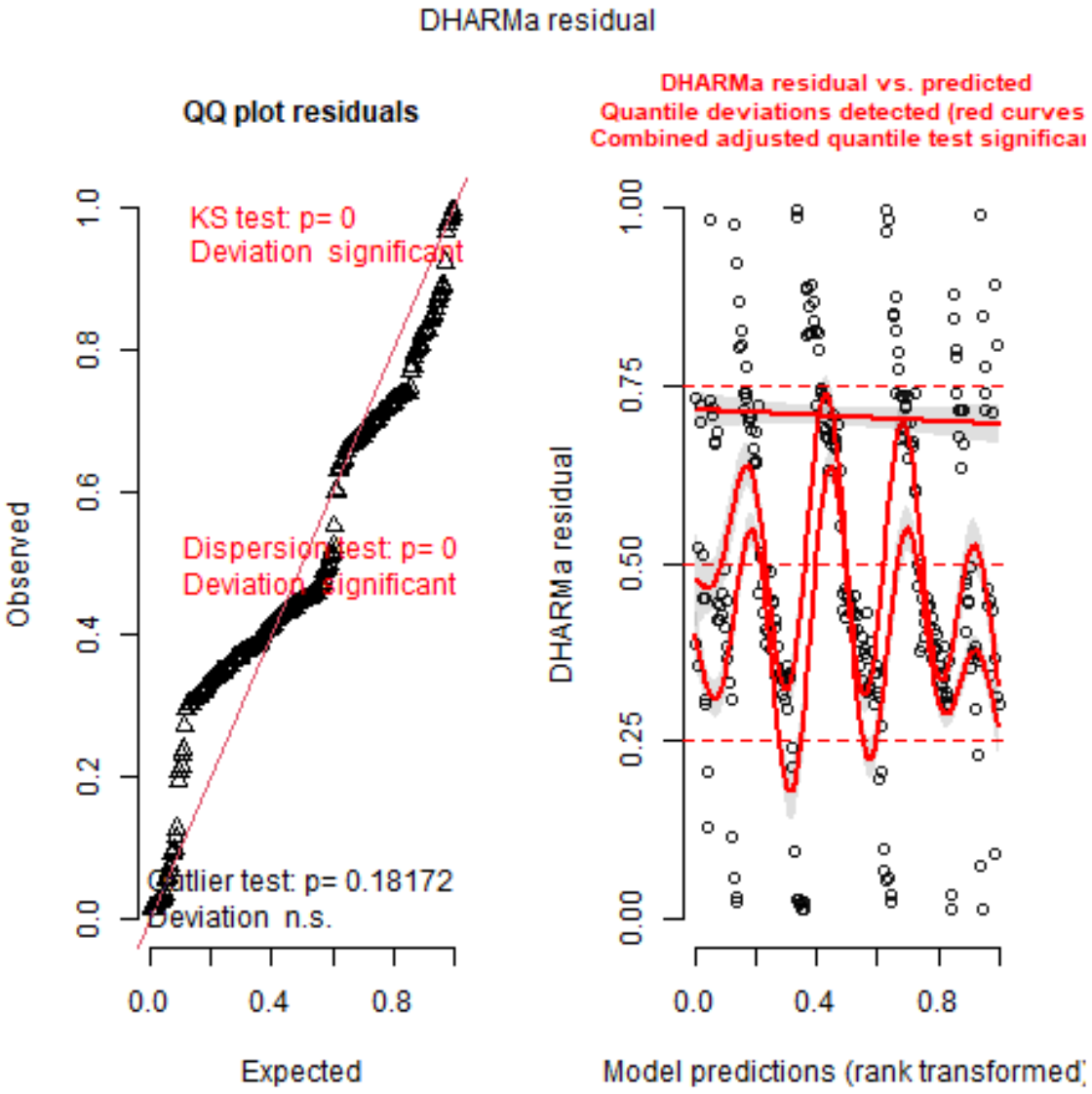


Figure 6: Esquizofrenia, Linear, Tempo + país + Gini

Não parece ter mudado muito, então provavelmente a causalidade, ou pelo menos a maior parte dela, vem do país onde a população está, e não da desigualdade social para a esquizofrenia.

Extraímos também o AIC e a log-verossimilhança para todos os modelos de causalidade da esquizofrenia:

```
logLik(glm.linear.year)
logLik(glm.pois.year)
logLik(glm.bin.year)

AIC(glm.linear.year)
AIC(glm.pois.year)
AIC(glm.bin.year)

logLik(glm.linear.gini)
logLik(glm.pois.gini)
logLik(glm.bin.gini)

AIC(glm.linear.gini)
AIC(glm.pois.gini)
AIC(glm.bin.gini)

logLik(glm.linear.country)
logLik(glm.pois.country)
logLik(glm.bin.country)

AIC(glm.linear.country)
AIC(glm.pois.country)
AIC(glm.bin.country)

logLik(glm.linear.country.gini)
logLik(glm.pois.country.gini)
logLik(glm.bin.country.gini)

AIC(glm.linear.country.gini)
AIC(glm.pois.country.gini)
AIC(glm.bin.country.gini)
```

O que obtivemos foi o seguinte:

!!!!! TABELA BONITINHA QUE SÓ PODE IR NO FINAL !!!!!

Repetimos todo esse procedimento para os outros quatro transtornos, como você pode verificar no arquivo `glm.R`.

Para a depressão, o modelo cujos resíduos foram mais distantes dos resíduos do modelo somente com o tempo foi o de tempo + país:

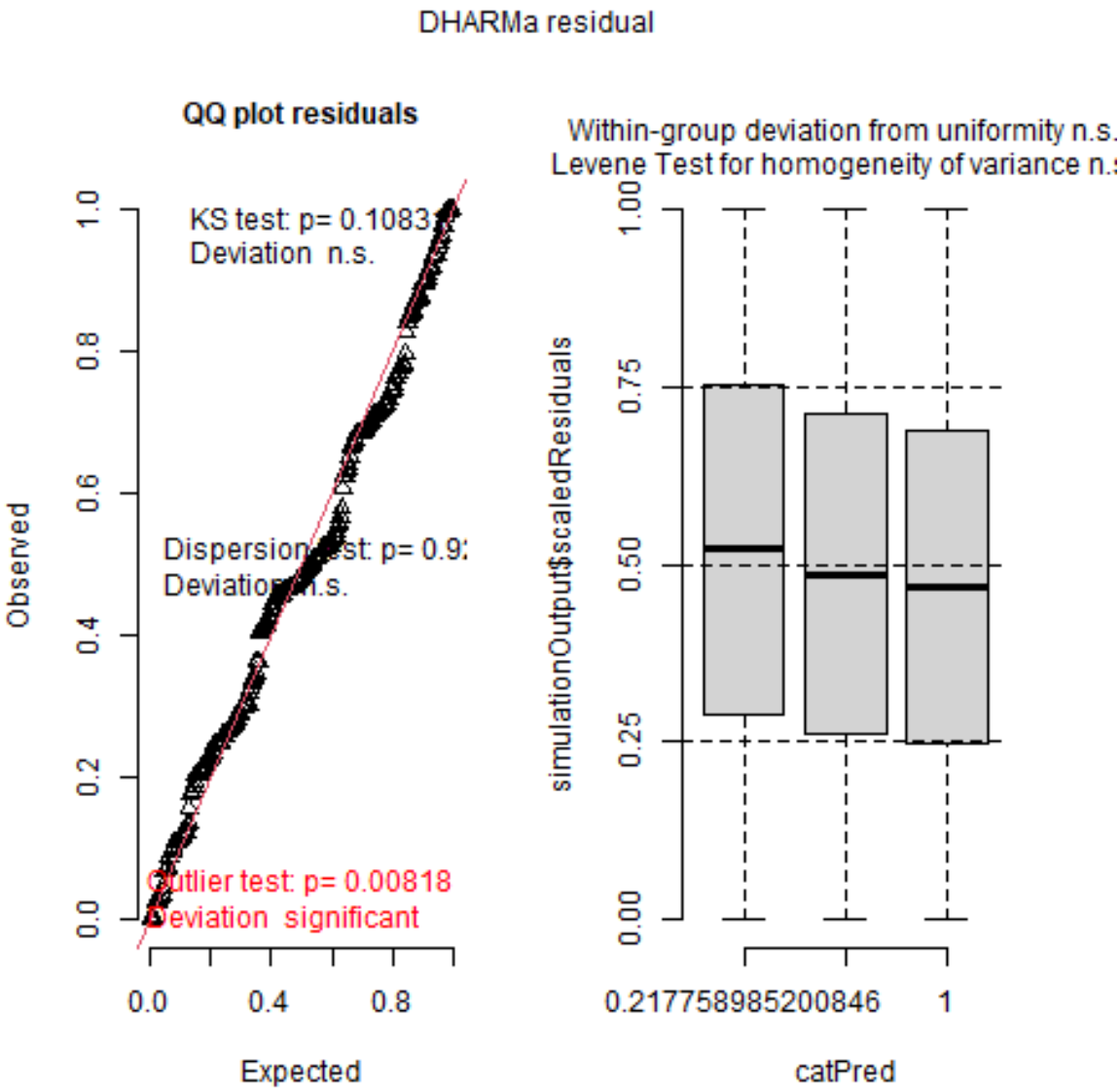


Figure 7: Depressão, Tempo

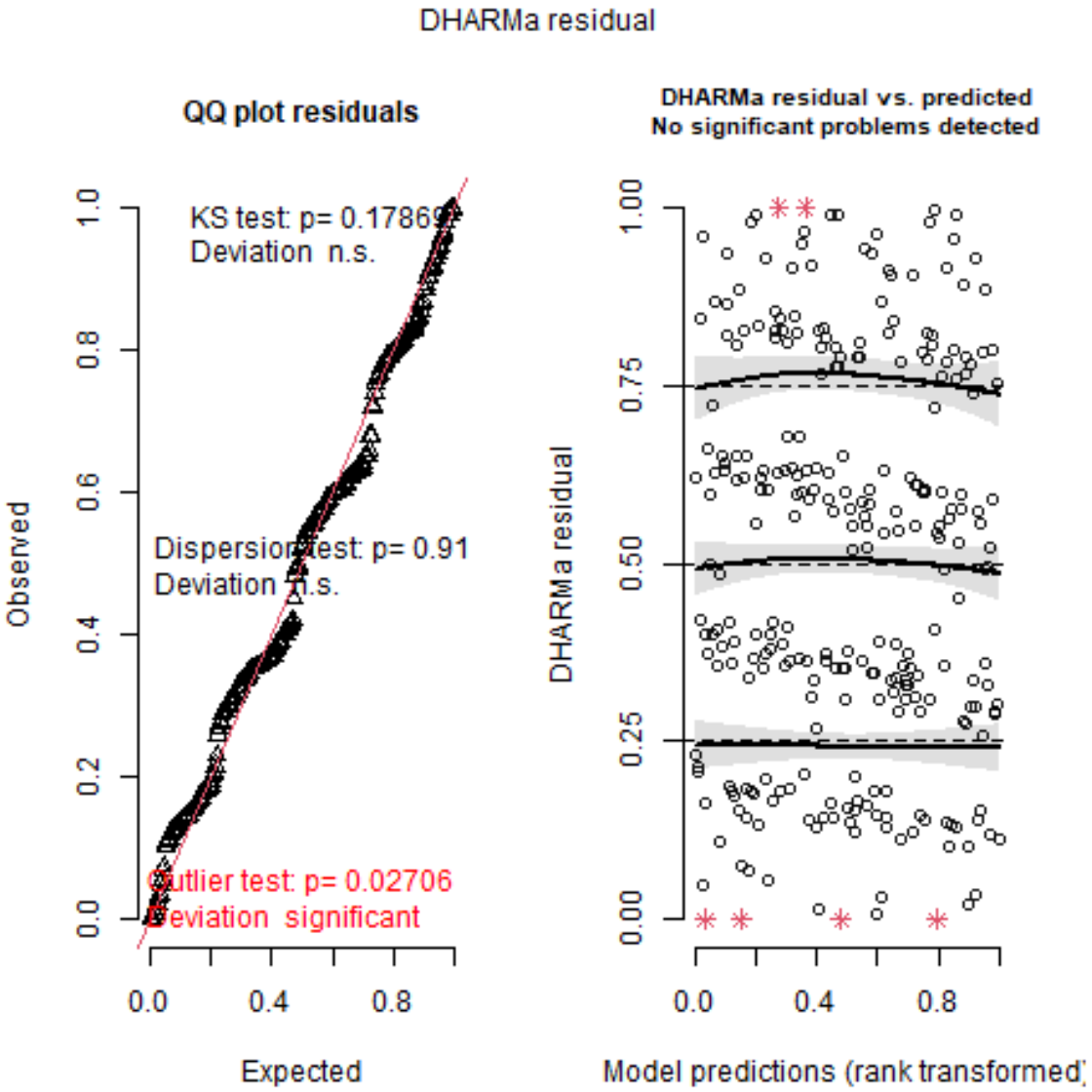


Figure 8: Depressão, Tempo + Gini

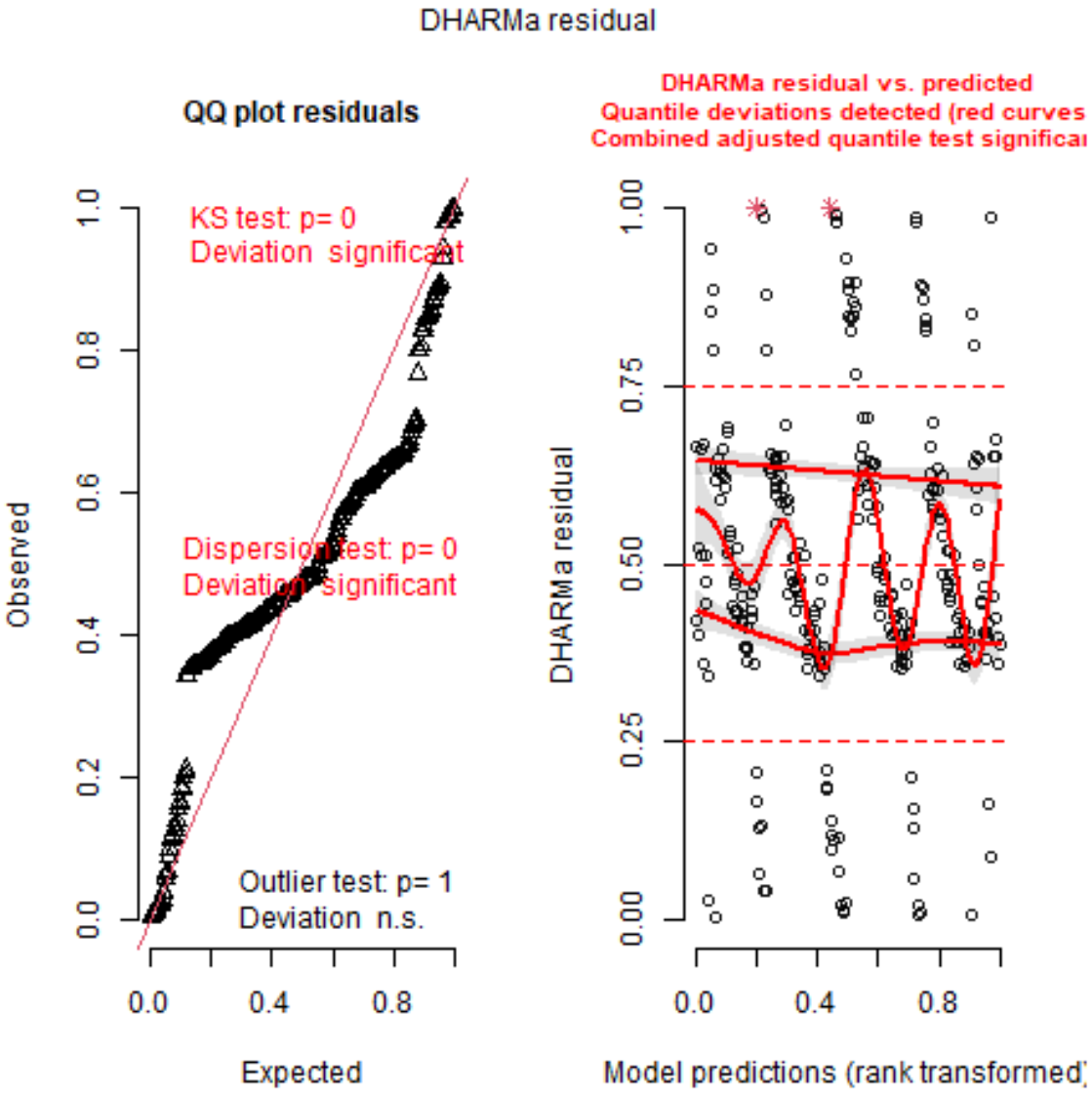


Figure 9: Depressão, Tempo + país

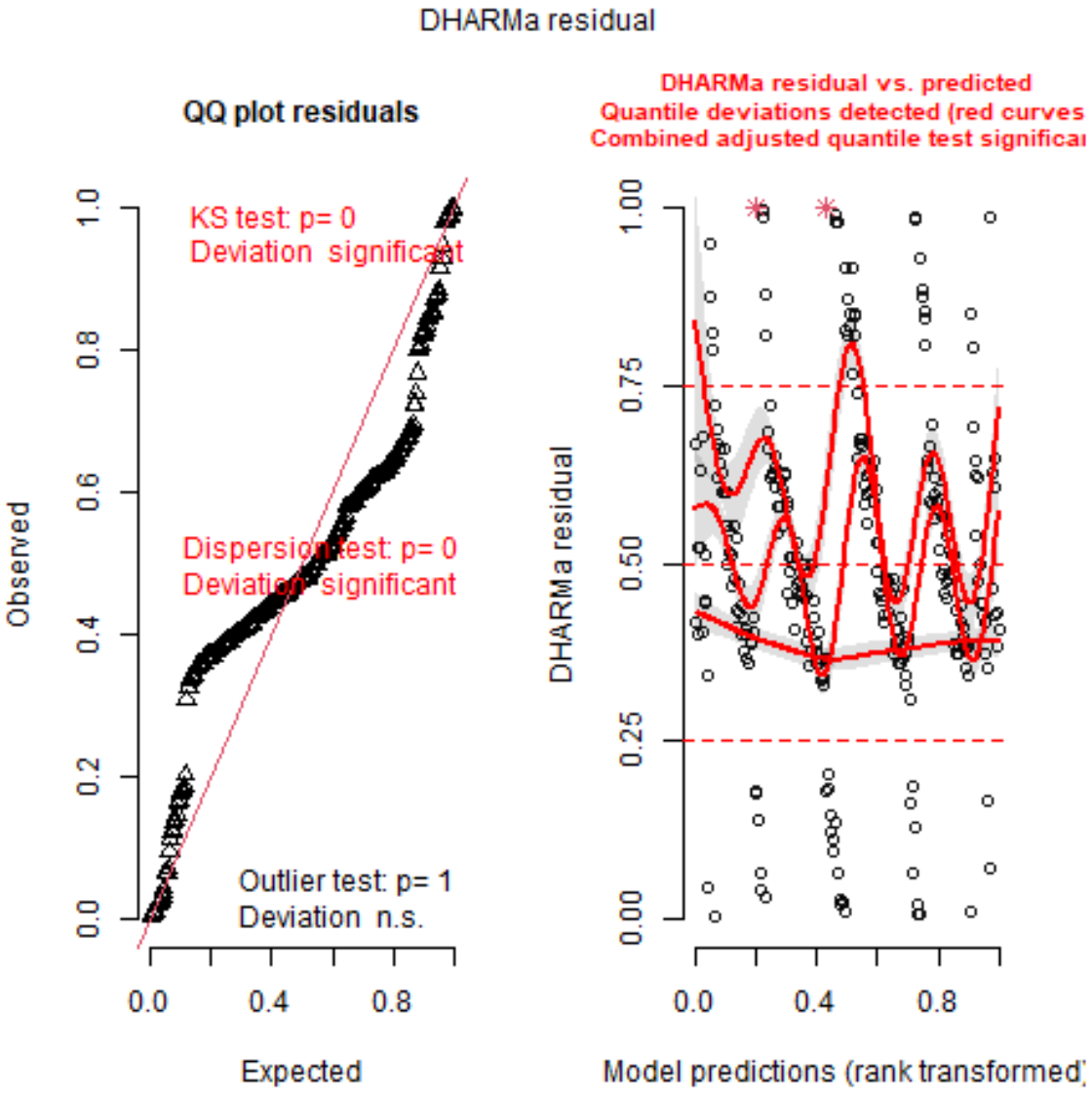


Figure 10: Depressão, Tempo + Gini + país

Também extraímos AICs e log-verossimilhanças para a depressão, obtendo os seguintes dados:

!!!!! TABELA BONITINHA QUE SÓ PODE IR NO FINAL !!!!!

Para a ansiedade:

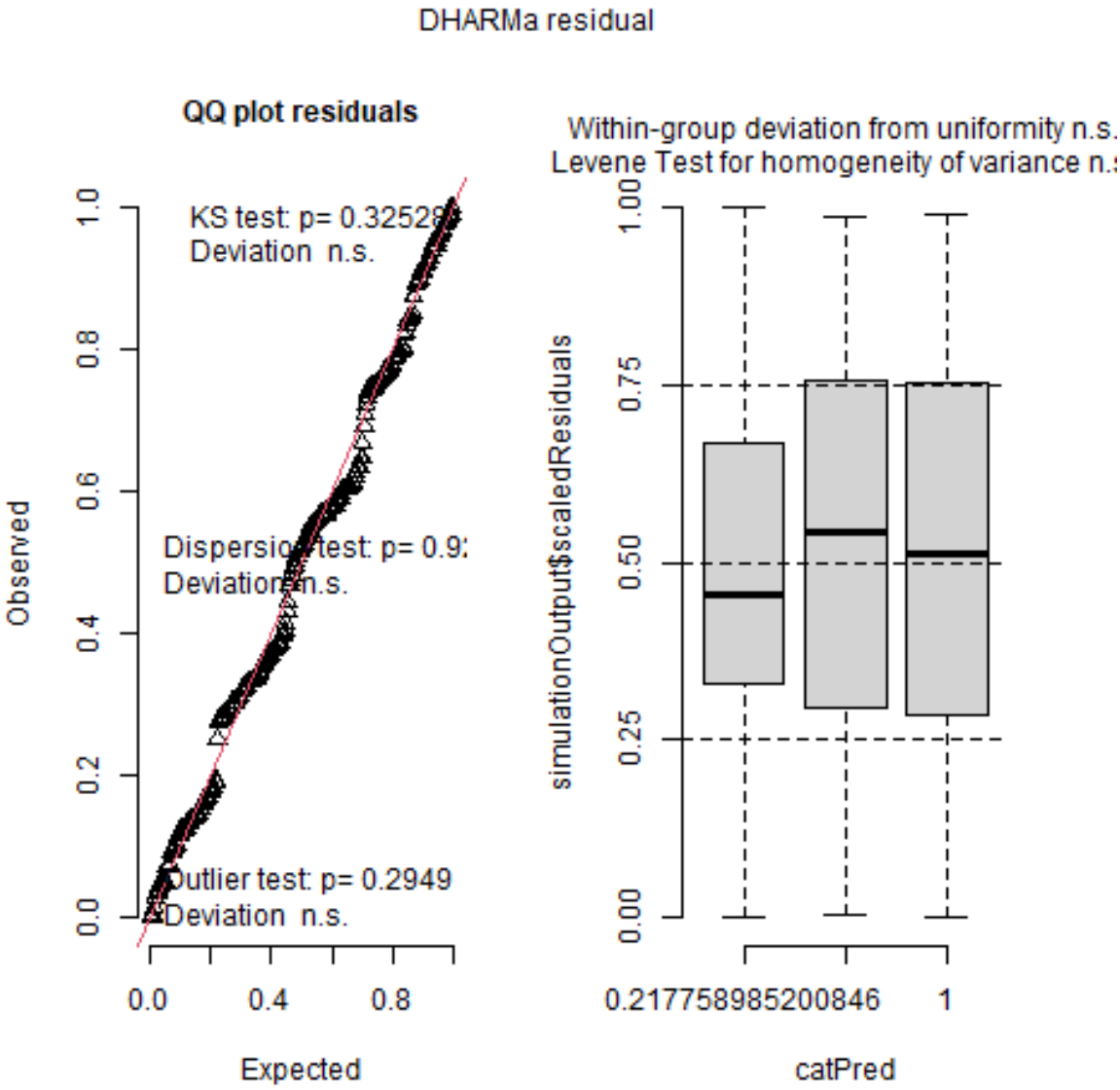


Figure 11: Ansiedade, Tempo

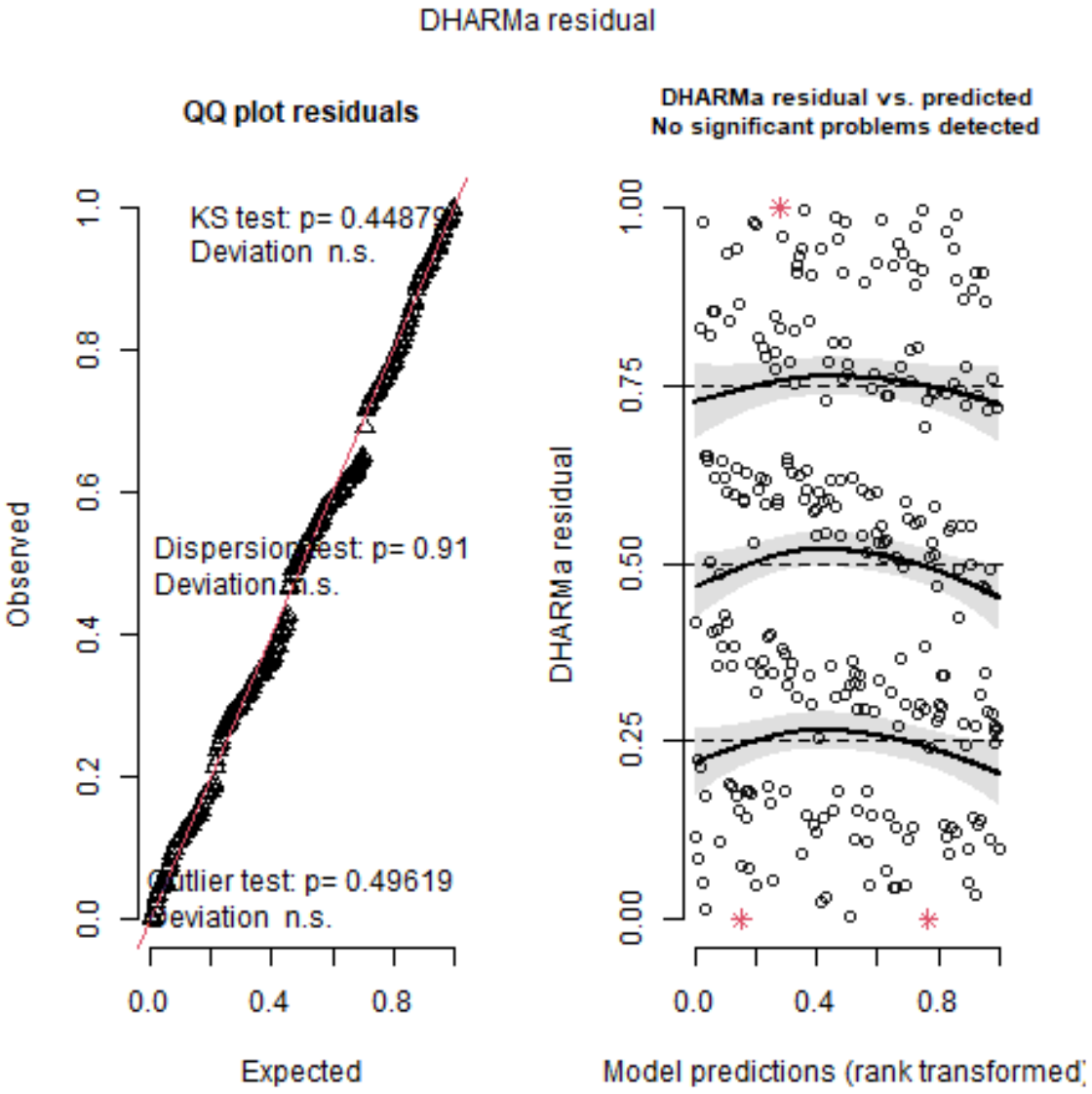


Figure 12: Ansiedade, Tempo + Gini

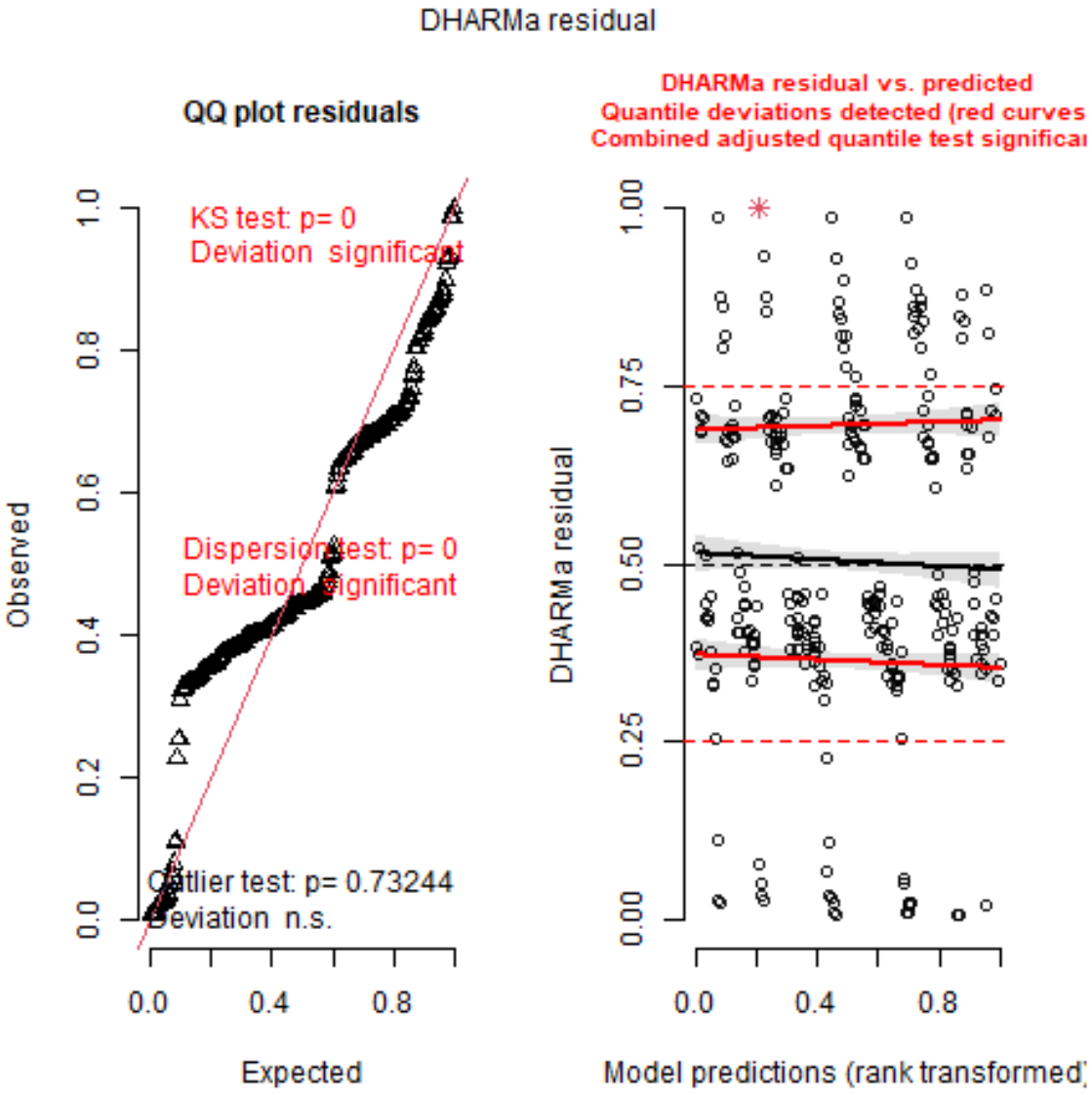


Figure 13: Ansiedade, Tempo + país

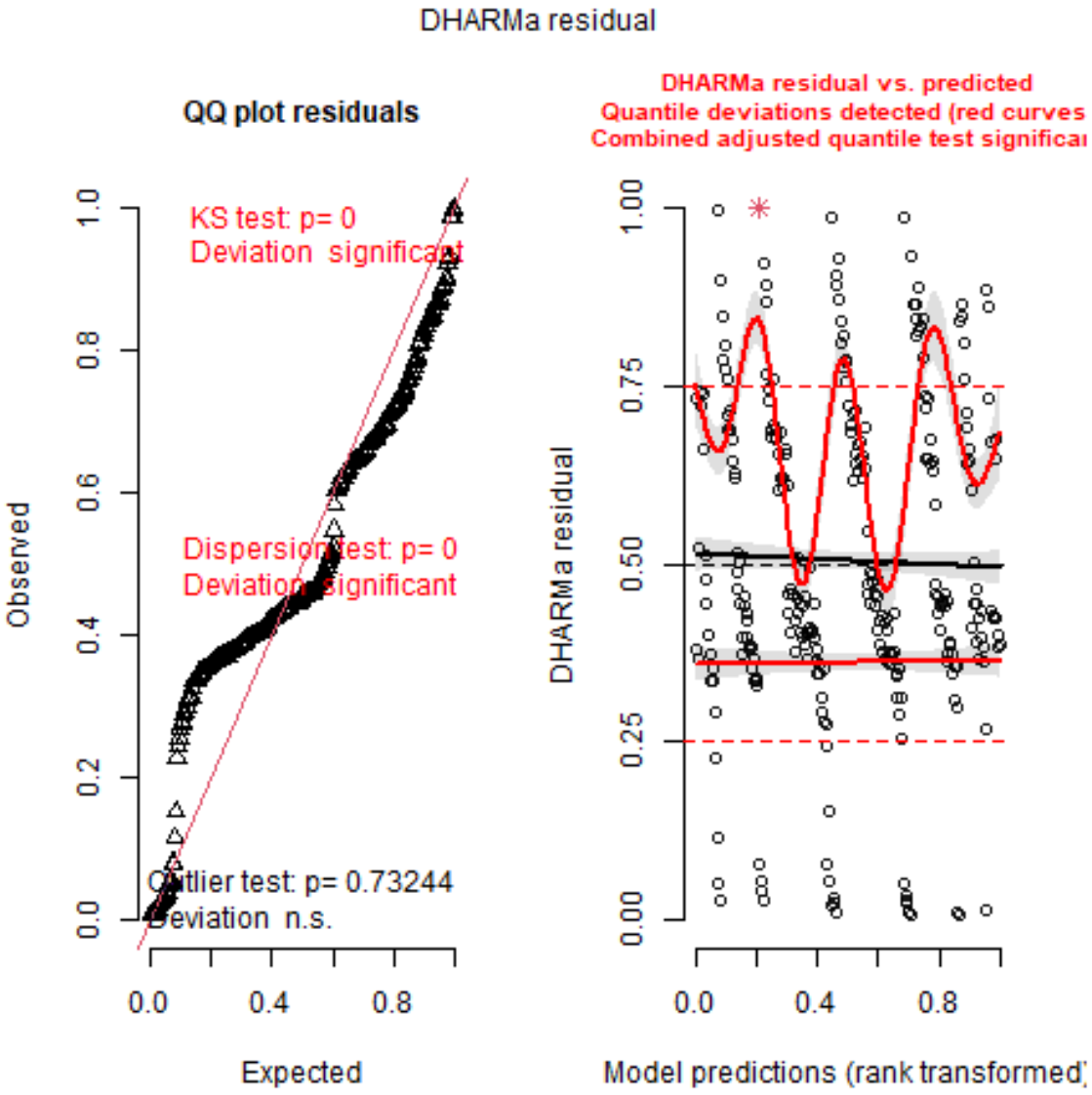


Figure 14: Ansiedade, Tempo + país + Gini

!!!!! TABELA BONITINHA QUE SÓ PODE IR NO FINAL !!!!!

Para o transtorno bipolar:

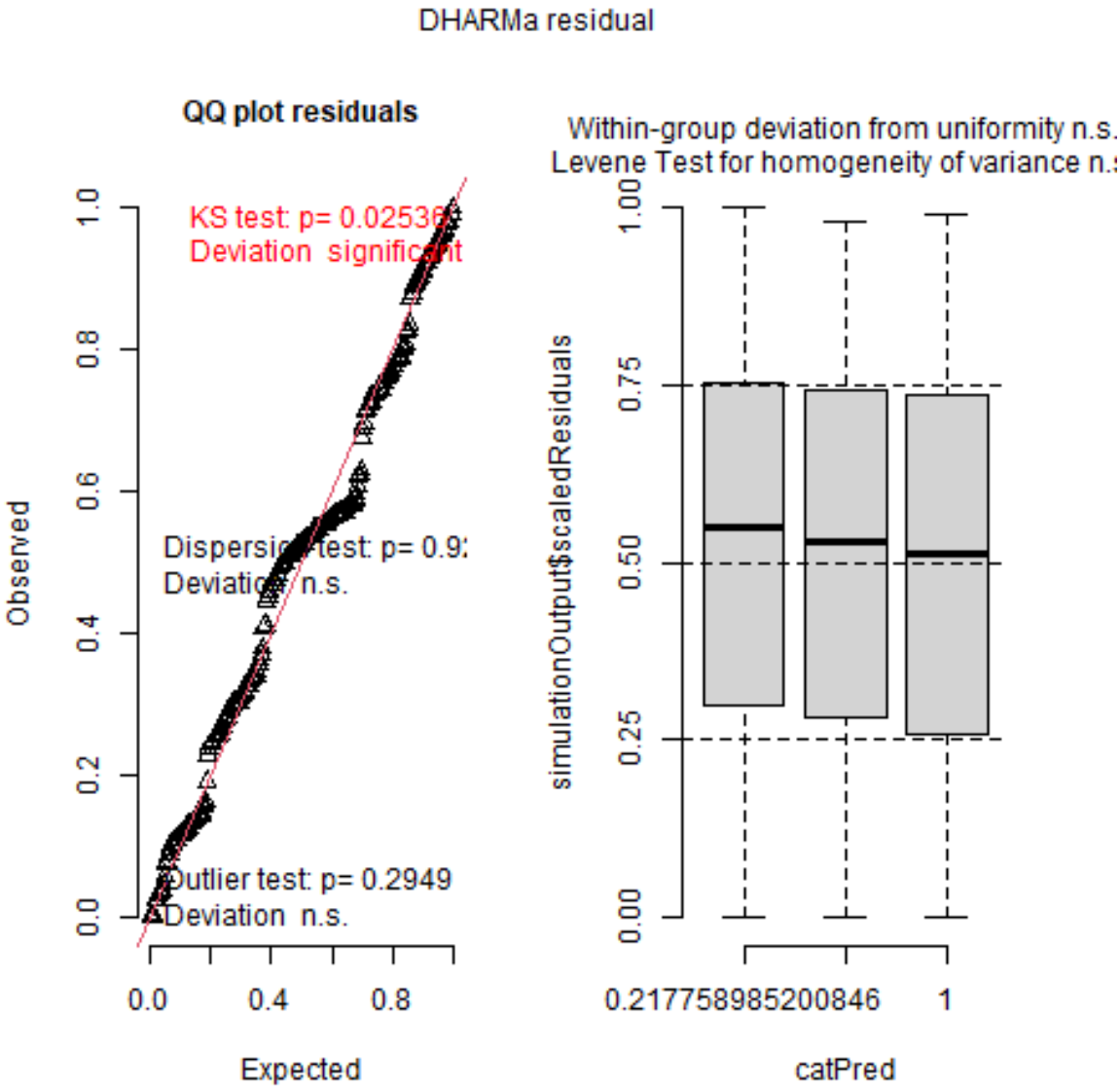


Figure 15: Transtorno bipolar, Tempo

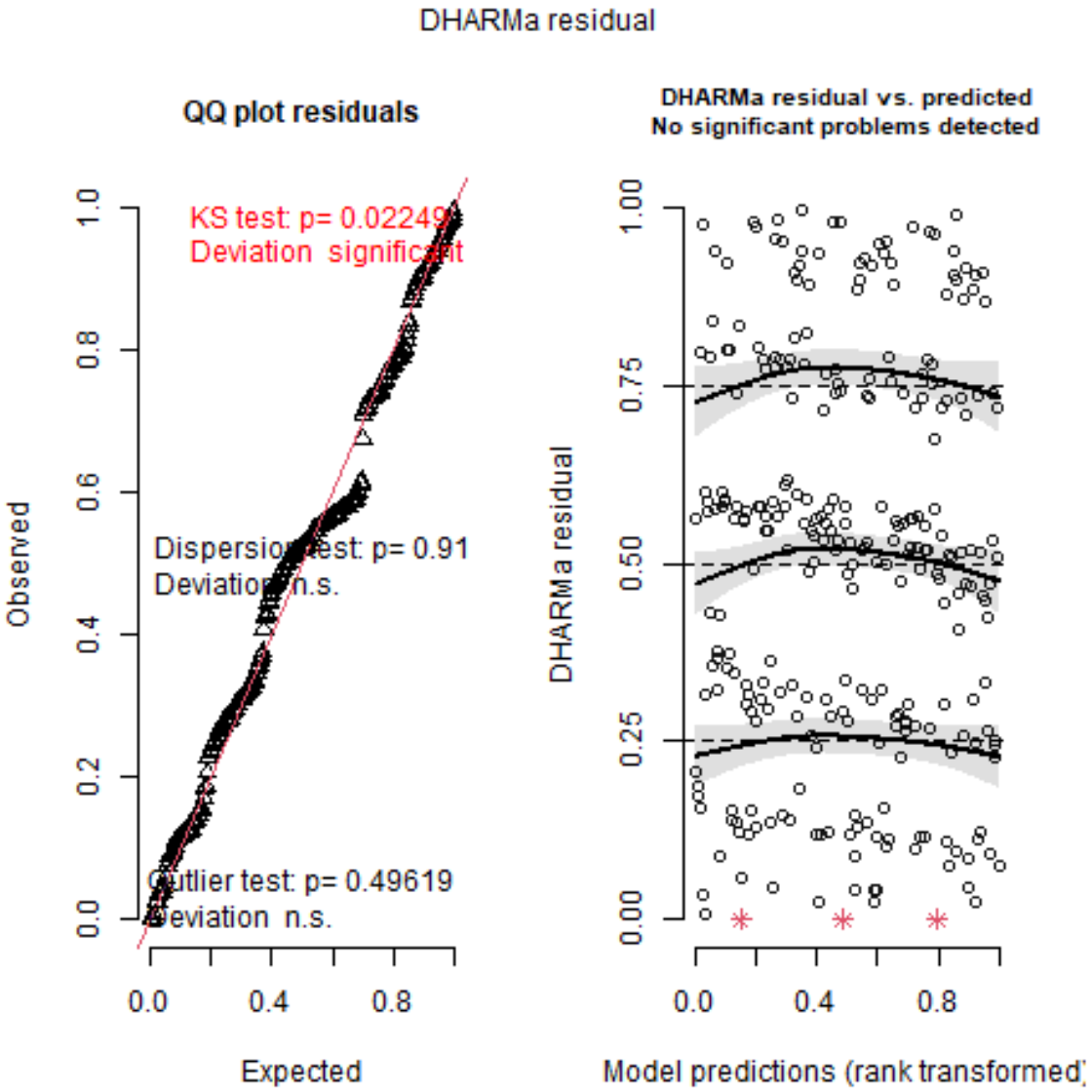


Figure 16: Transtorno bipolar, Tempo + Gini

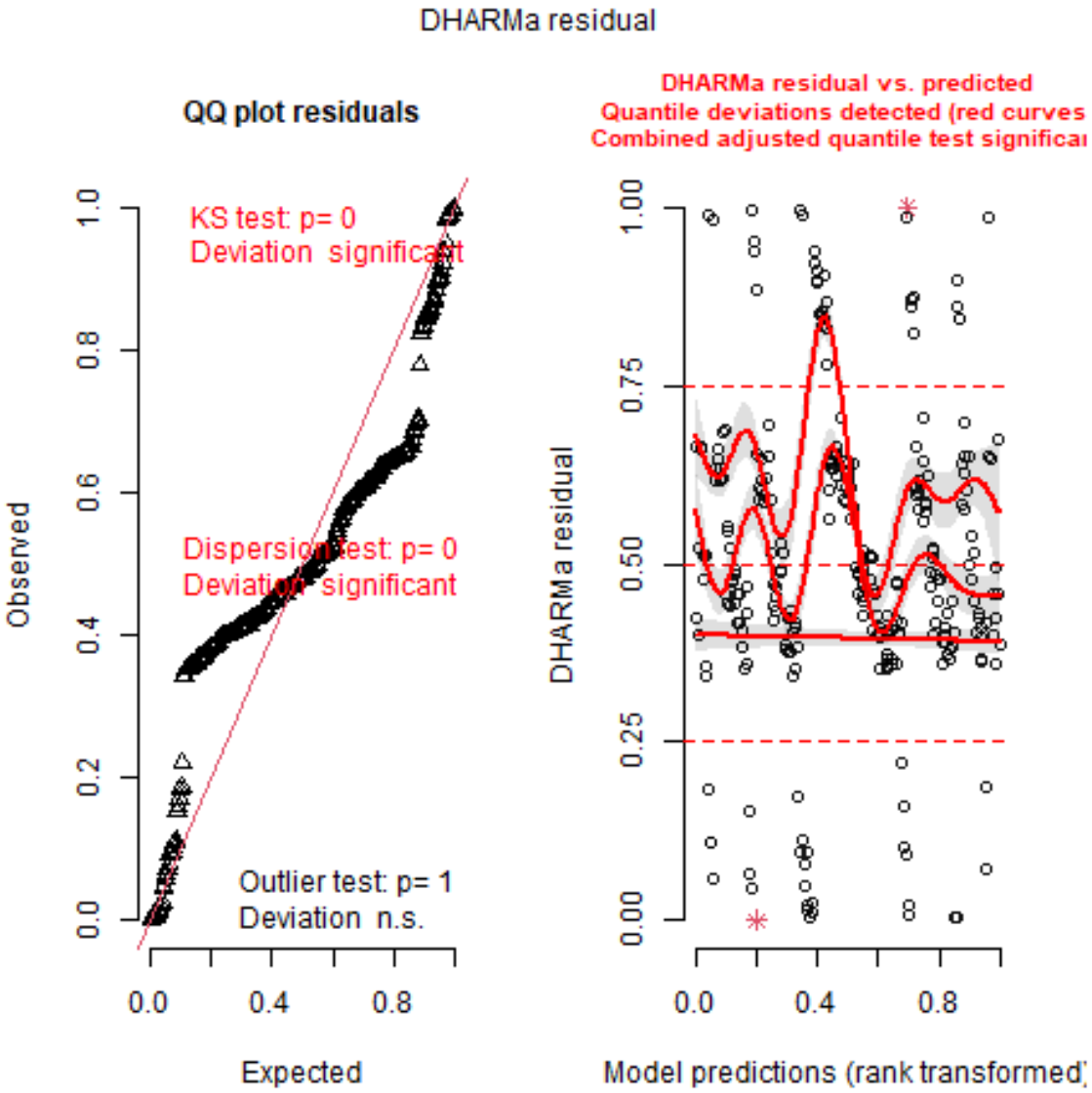


Figure 17: Transtorno bipolar, Tempo + país

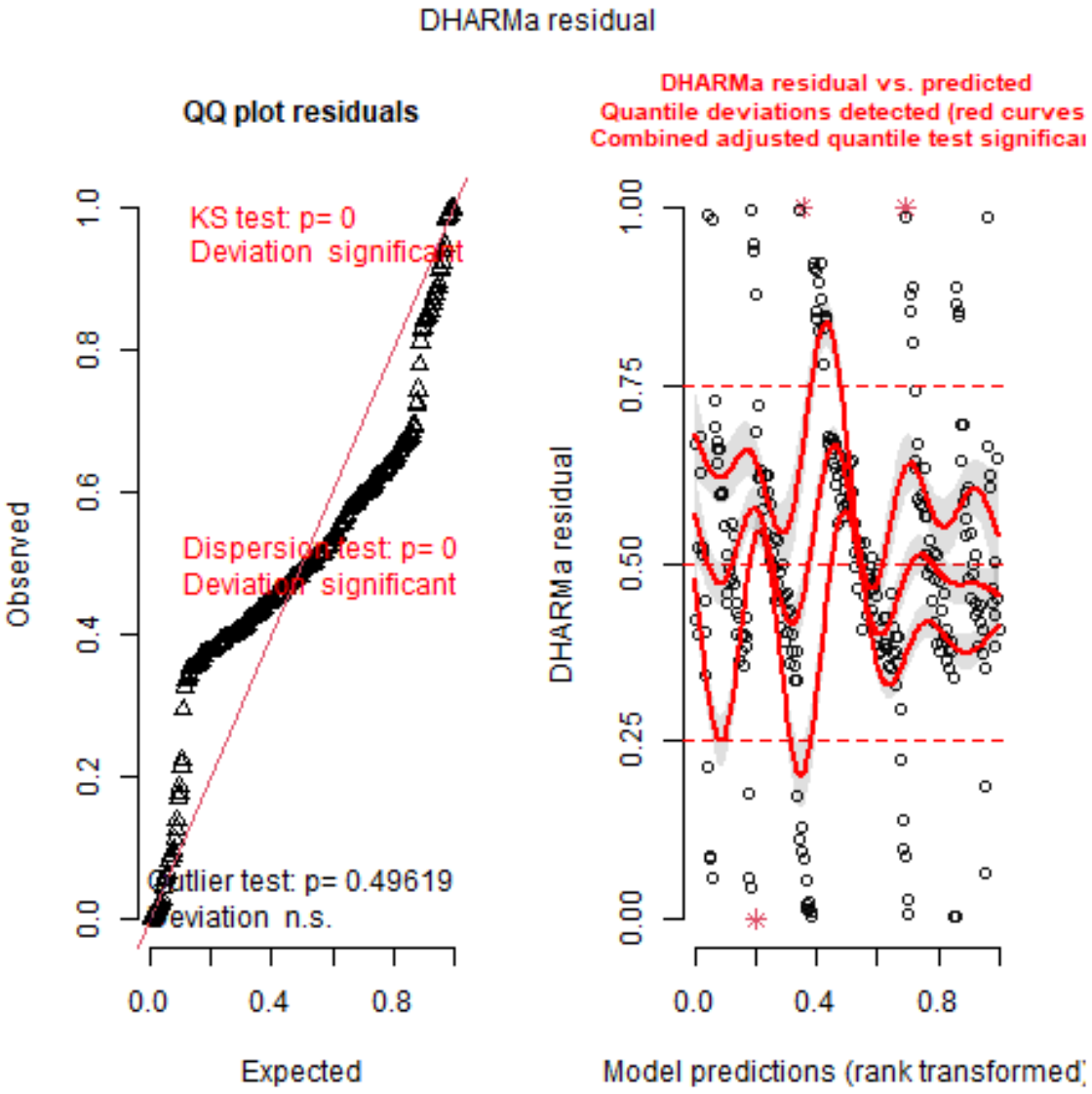


Figure 18: Transtorno bipolar, Tempo + país + Gini

!!!!! TABELA BONITINHA QUE SÓ PODE IR NO FINAL !!!!!

E para os transtornos alimentares:

!!!!! TABELA BONITINHA QUE SÓ PODE IR NO FINAL !!!!!

Obtivemos padrões parecidos, mesmo para os transtornos que imaginávamos terem causalidade forte com a desigualdade social, como a depressão e a ansiedade. Isso pode se dever tanto ao

fato de que realmente esses transtornos mentais não são causados pela desigualdade social, ou por problemas na nossa análise, como os que discutiremos a seguir. Lembre-se, você pode ver os gráficos dos resíduos clássicos para todos os modelos e dos resíduos do DHARMa para os modelos Poisson e binomial na pasta `images -> models`.

O índice de Gini é uma medida de desigualdade social, isto é, mede apenas a diferença de renda entre as pessoas de um mesmo país, não indicando, contudo, a renda delas. Isso significa que um país em que a maioria das pessoas é muito rica terá um bom índice de Gini, mas isso também acontecerá num país em que a maioria das pessoas é muito pobre. Dessa forma, reconhecemos um viés na nossa análise. Como já apontado, certos transtornos mentais podem estar correlacionados, ou serem causados, pelo menos a nível populacional, por pressões relacionadas às condições socioeconômicas e de trabalho da população. Assim, é possível que nossas relações de causalidade estejam enviesadas pela baixa sensibilidade do índice de Gini às condições socioeconômicas gerais, de trabalho e políticas da população.

Supor que uma população de um país tem uma homogeneidade genética tal que nos permita usar o país como indicador de predisposição genética apresenta um problema análogo. Países diferentes têm situações socioeconômicas diferentes, que podem causar transtornos mentais. Adicionalmente, os efeitos da entrada e saída de pessoas nos países, principalmente nos de baixa população, não foi estudado, mas comprometem a homogeneidade genética da população, fazendo o uso do país como indicador de predisposição genética ficar ainda mais frágil.

Reconhecemos, portanto, que nossa análise pode estar enviesada, considerando todas essas informações, e recomendamos o uso do PIB *per capita* e do índice de Gini conjuntamente para avaliar a situação socioeconômica da população. Para a predisposição genética, talvez indicadores moleculares sejam adequados, contanto que a pessoa interessada em estudar essa causalidade esteja ciente das possíveis implicações raciais desse tipo de estudo.

3 Conclusões

3.0.1 Lógica da causalidade

O modelo só com tempo não leva em consideração nenhuma causalidade. Ele terá resíduo de tamanho x. O modelo com gini, por exemplo, leva em consideração uma variável a mais, que pode ser causal. Ele tem resíduo de tamanho x + a. Essa diferença a entre os resíduos do modelo sem gini e com gini simboliza a causalidade de gini sobre os dados, então, quanto maior a diferença entre os resíduos do modelo sem a variável e com ela, maior a causalidade.

n.d. *Institute for Health Metrics and Evaluation.* Seattle, WA: IHME, University of Washington, 2024: Institute for Health Metrics; Evaluation. https://vizhub.healthdata.org/gbd-results/.

Ceriani, Lidia, and Paolo Verme. 2012. "The Origins of the Gini Index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini." *The Journal of Economic Inequality* 10 (3): 421–43. https://doi.org/10.1007/s10888-011-9188-x.

Dattani, Saloni, Lucas Rodés-Guirao, Hannah Ritchie, and Max Roser. 2023. "Mental Health." *Our World in Data*.

Gordovez, Francis James A., and Francis J. McMahon. 2020. "The Genetics of Bipolar Disorder." *Molecular Psychiatry* 25 (3): 544–59. https://doi.org/10.1038/s41380-019-0634-7.

Müller, Kirill. 2020. "Here: A Simpler Way to Find Your Files." https://CRAN.R-project.org/package=here.

Pedersen, Thomas Lin. 2024. "Patchwork: The Composer of Plots." https://CRAN.R-project.org/package=patchwork.

Prins, Seth J., Lisa M. Bates, Katherine M. Keyes, and Carles Muntaner. 2015. "Anxious? Depressed? You Might Be Suffering from Capitalism: Contradictory Class Locations and the Prevalence of Depression and Anxiety in the USA." *Sociology of Health & Illness* 37 (8): 1352–72. https://doi.org/10.1111/1467-9566.12315.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

Vereczkei, Andrea, and Karoly Mirnics. 2011. "Genetic Predisposition to Schizophrenia: What Did We Learn and What Does the Future Hold?" *Neuropsychopharmacologia Hungarica* 13 (4): 205–10. https://doi.org/10.5706/nph201112003.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. "Scales: Scale Functions for Visualization." https://CRAN.R-project.org/package=scales.

Wilke, Claus O. 2024. "Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'." https://CRAN.R-project.org/package=cowplot.