POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Projeto de Métodos de Regressão e Previsão

Mestrado em Ciência de dados

Francisco Fernandes Nº 2230390

João Oliveira Nº 2230383

Leonardo Nogueira Nº 2230394

Leiria, junho de 2024

Índice

1. Avaliação de popularidade de filmes

Neste trabalho, nosso objetivo foi desenvolver um modelo de regressão logística para prever a popularidade de filmes utilizando a base de dados MoviesDB.RData. Foi usada uma amostra aleatória de 97 filmes, sendo que um filme é considerado popular se a variável popularity atingir uma pontuação mínima de 50 pontos. Para isso, criamos uma nova variável, *'evaluation'* que identifica se um filme é popular ou não.

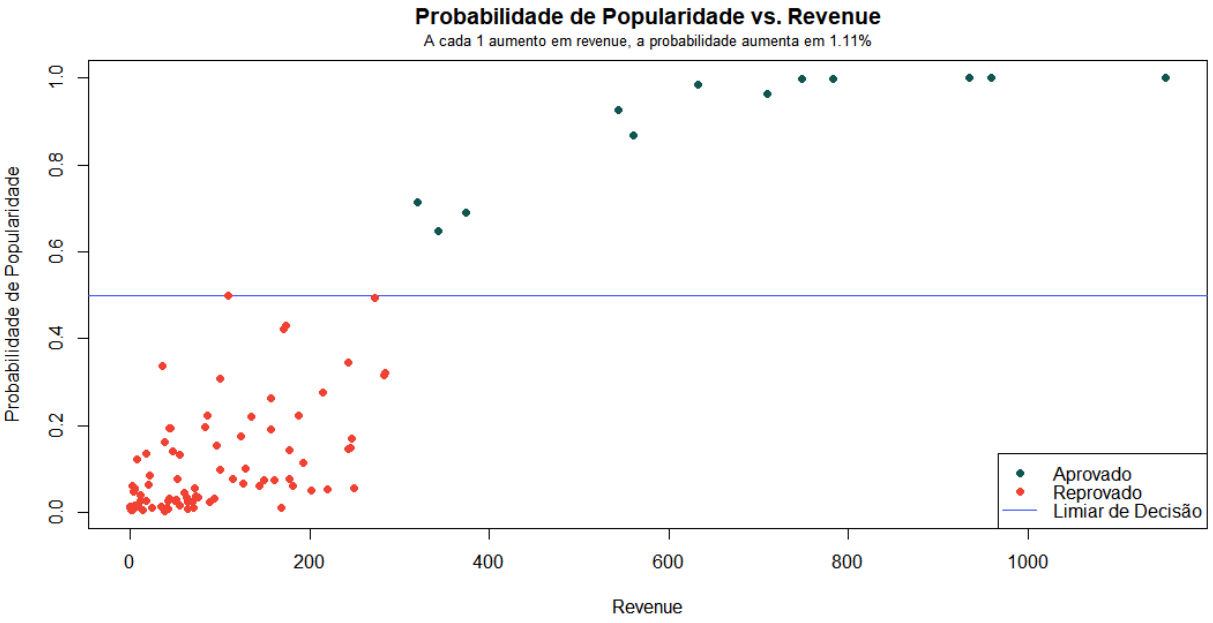
1.1. Modelo de regressão logística

Foram selecionadas duas variáveis numéricas, *'revenue'* e *'vote_average'*, para construir o modelo de regressão logística. A fórmula é o seguinte:

$$\widehat{Ev}(evaluation = 1) = \frac{1}{1 + e^{-(-12.501 + 0.011 * Rv + 1.411 * Va)}}$$

- Êv = Estimativa para previsão de 'evaluation'
- Rv: 'revenue'
- Va: 'vote_average'

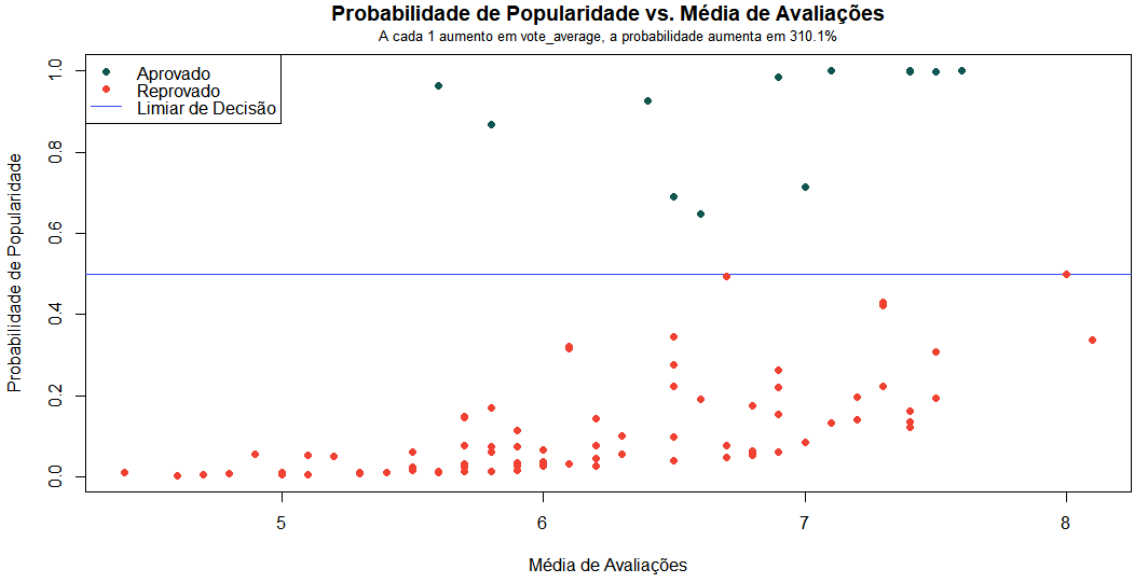
Ao avaliar a equação formulada, podemos ver que, quando *'revenue'* e *'vote_average'* são zero, os log-odds de um filme ser popular são -12.501.



O gráfico acima apresenta a relação entre a receita (*'revenue'*) e a probabilidade prevista de um filme ser popular. Observa-se que filmes com receitas mais altas tendem a ter uma maior probabilidade de serem classificados como populares.

A linha azul horizontal representa o limiar de decisão de 0.5, acima do qual os filmes são considerados populares (indicados em verde) e abaixo do qual são considerados não populares (indicados em vermelho), o mesmo se aplica ao gráfico a seguir.

A análise revela que *'revenue'* é um preditor significativo da popularidade de um filme, com a probabilidade de popularidade aumentando em 1.11% para cada aumento unitário em receita.



O gráfico logo acima ilustra a relação entre a média de avaliações (*'vote_average'*) e a probabilidade prevista de um filme ser popular. Observa-se que filmes com médias de avaliações mais altas tendem a ter uma maior probabilidade de serem classificados como populares.

A análise revela que *'vote_average'* é um preditor significativo da popularidade de um filme, com a probabilidade de popularidade aumentando em 310.1% para cada aumento unitário na média de avaliações.

1.1.1. Teste de Hipótese de Wald

Para validação da importância das variáveis selecionadas foi aplicado um teste de hipótese de Wald, no qual foi considerado o seguinte:

- Hipótese nula: $\beta_i = 0$
- Hipótese alternativa: $\beta_i \neq 0$

A partir dos resultados obtidos da função summary() do modelo logístico, observamos os seguintes valores de p:

- p value $(Rv) = 0.000334$
- p value $(Va) = 0.009404$

Considerando a convenção de um alpha = 0,05, temos que, para ambas as variáveis, há evidência estatística para rejeitarmos a hipótese nula. Portanto, podemos concluir que ambas as variáveis são significativas, influenciando a probabilidade de um filme ser popular ou não.

1.2. Comparativo com modelo nulo

Para avaliar a qualidade do modelo de regressão logística desenvolvido, foi construído um modelo nulo para comparação. O modelo nulo considera apenas a média geral da variável dependente (*'evaluation'*), sem levar em conta as variáveis preditoras (*'revenue'* e *'vote_average'*). A equação do modelo nulo é dada por:

$$\widehat{\pi} = \frac{1}{1 + e^{-(-1.348)}} = 0.2062$$

Isso significa que, segundo o modelo nulo, a probabilidade de um filme ser popular é de aproximadamente 20.62%.

Para testar se o modelo logístico é significativamente melhor que o modelo nulo, aplicamos o teste de razão de verossimilhança (likelihood ratio test). As hipóteses consideradas foram:

- Hipótese nula: $\beta_{Rv} = \beta_{Va} = 0$
- Hipótese alternativa: Algum $\beta \neq 0$

Como o p-value resultante foi de $1.11x10^{-10}$, que é significativamente menor que o nível de significância convencional ($\alpha = 0.05$), rejeitamos a hipótese nula, indicando que pelo menos um dos parâmetros do modelo não é zero. Portanto, o modelo logístico com as variáveis *'revenue'* e *'vote_average'* é preferível ao modelo nulo.

1.2.1. Avaliação de Deviances

Considerando o Deviance, que é uma medida estatística usada para avaliar a qualidade de um modelo de regressão logística, como uma medida da ausência de ajuste de um modelo aos dados, observamos que:

- Deviance do modelo nulo: 98.719 (com 96 graus de liberdade)
- Deviance residual do modelo logístico: 52.876 (com 94 graus de liberdade)

A deviance do modelo logístico é significativamente menor que a do modelo nulo, o que indica um melhor ajuste aos dados.

1.2.2. Critério de Akaike

O Critério de Informação de Akaike (AIC) é uma medida usada para comparar a qualidade de diferentes modelos estatísticos, penalizando a complexidade do modelo para evitar o overfitting. Comparando os modelos usando este critério, temos:

- AIC do modelo logístico: 58.8757
- AIC do modelo nulo: 100.7186

O modelo logístico apresenta um AIC consideravelmente menor, reforçando a ideia de que o modelo ajustado é preferível ao modelo nulo. Isso implica que as variáveis *'revenue'* e *'vote_average'* são importantes preditores da popularidade dos filmes, contribuindo para um melhor ajuste do modelo.

1.3. Cálculo para Pseudo R^2

A fim de avaliar a adequação do modelo de regressão logística na explicação da variabilidade na popularidade dos filmes, foram calculados os valores de Pseudo R^2. Esses valores oferecem insights sobre o quão bem o modelo se ajusta aos dados, fornecendo uma medida de sua capacidade preditiva. Neste contexto, os valores de Cox & Snell, Nagelkerke e McFadden foram calculados e analisados, oferecendo uma visão abrangente do desempenho do modelo em relação às variáveis *'revenue'* e *'vote_average'*

- Cox & Snell: Este valor de pseudo R-quadrado é uma medida de ajuste do modelo que varia de 0 a 1. Quanto mais próximo de 1, melhor o ajuste do modelo aos dados. Ele indica que cerca de 37.7% da variabilidade na popularidade dos filmes é explicada pelo modelo logístico.
- Nagelkerke: O pseudo R-quadrado de Nagelkerke é uma versão ajustada do Cox & Snell, e também varia de 0 a 1. Ele é uma estimativa melhorada do verdadeiro R-quadrado e pode fornecer uma indicação mais precisa da qualidade do modelo. Neste caso, ele indica que aproximadamente 58.98% da variabilidade na popularidade dos filmes é explicada pelo modelo.
- McFadden: O pseudo R-quadrado de McFadden é uma medida mais conservadora e pode ser interpretado de forma um pouco diferente dos outros dois. Ele não é uma medida de ajuste absoluto, mas sim uma medida relativa, onde valores maiores indicam melhor ajuste. Neste caso, o valor de 46.44% sugere que o modelo logístico tem um ajuste relativamente bom aos dados.

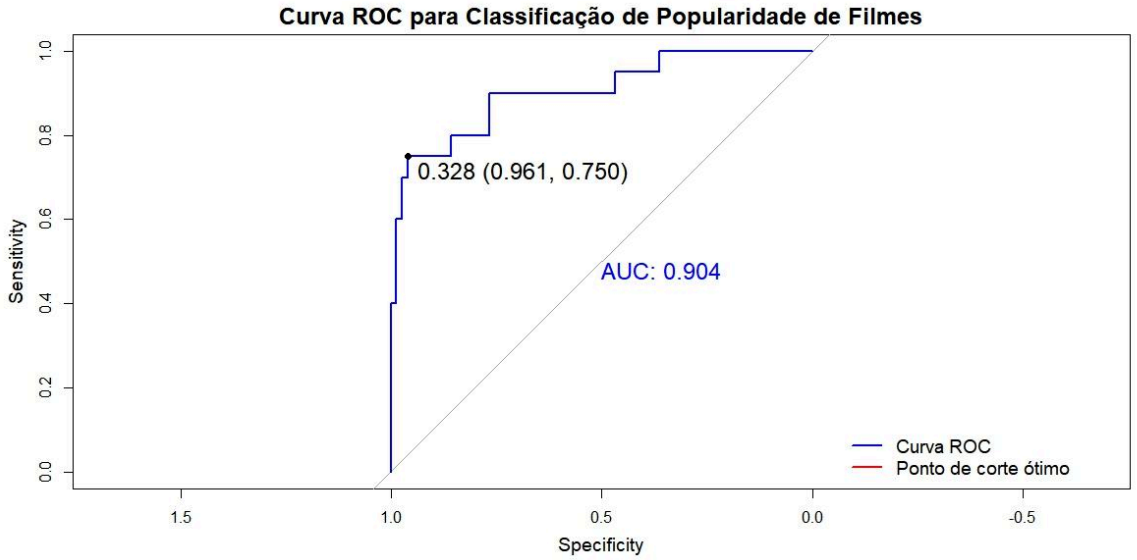
Esses resultados indicam que o modelo logístico é eficaz em explicar a variabilidade na popularidade dos filmes com base nas variáveis *'revenue'* e *'vote_average'*, fornecendo uma boa capacidade preditiva em comparação com o modelo nulo.

1.4. Curva ROC

A curva ROC (Receiver Operating Characteristic) é uma ferramenta fundamental na avaliação do desempenho de modelos de classificação. No contexto deste estudo, a curva ROC foi utilizada para determinar o ponto de corte ótimo para o modelo de regressão logística desenvolvido para prever a popularidade dos filmes com base nas variáveis *'revenue'* e *'vote_average'*.

Primeiramente, a curva ROC foi calculada usando a biblioteca pROC em R. Essa curva representa a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade) para diferentes valores de corte. A área sob a curva ROC (AUC) fornece uma medida da capacidade discriminativa do modelo, onde valores mais próximos de 1 indicam um melhor desempenho.

A área sob a curva ROC obtida foi de 0.9039, indicando um bom poder discriminativo do modelo em distinguir entre filmes populares e não populares.

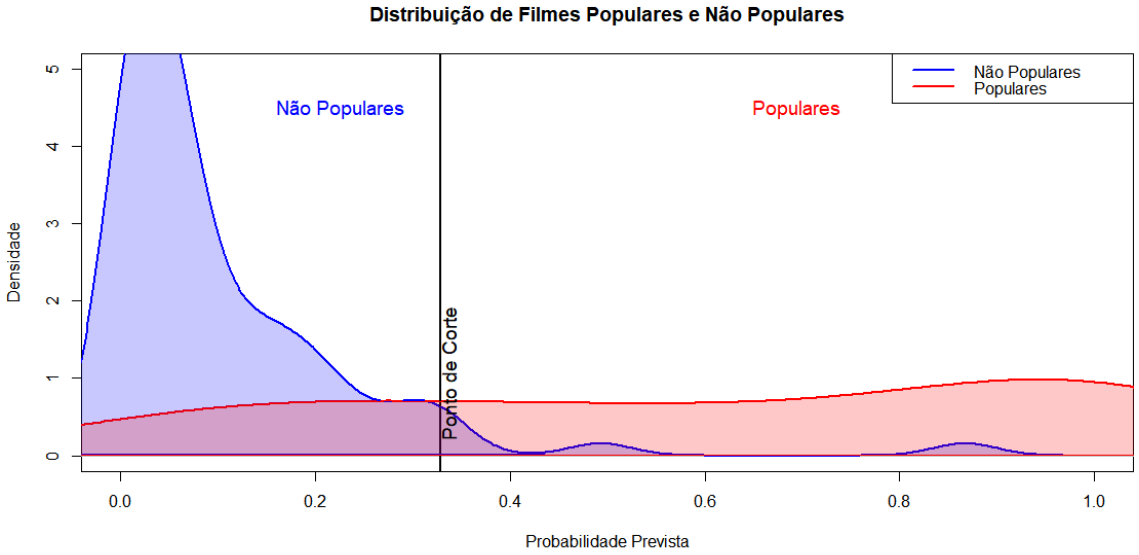


1.4.1. Determinação do Ponto de Corte Ótimo

A escolha do ponto de corte ótimo em um modelo de classificação é crucial para equilibrar a sensibilidade (taxa de verdadeiros positivos) e a especificidade (taxa de verdadeiros negativos). Neste estudo, o ponto de corte ótimo foi determinado utilizando o método que maximiza simultaneamente a sensibilidade e a especificidade.

O ponto de corte ótimo foi calculado a partir da curva ROC utilizando a função coords do pacote pROC em R. O método "best" foi especificado para selecionar o ponto que maximiza simultaneamente a sensibilidade e a especificidade. O ponto de corte ótimo encontrado foi de 0.3282472.

O ponto de corte ótimo identificado na curva ROC oferece um equilíbrio ideal entre sensibilidade e especificidade, permitindo maximizar a taxa de verdadeiros positivos e minimizar a taxa de falsos positivos. Esse ponto de corte é crucial para classificar corretamente os filmes como populares ou não populares com base no modelo de regressão logística desenvolvido.



Além disso, a análise dos valores de falsos positivos e falsos negativos confirma a eficácia do ponto de corte selecionado. A matriz de confusão gerada pelo modelo mostra os seguintes resultados:

- Verdadeiros Negativos (TN): 74
- Falsos Positivos (FP): 3
- Falsos Negativos (FN): 5
- Verdadeiros Positivos (TP): 15

As proporções correspondentes são:

- TN: 76.29%
- FP: 3.09%
- FN: 5.15%
- TP: 15.46%

Esses valores refletem o desempenho do modelo em termos de classificação correta e incorreta dos filmes. A baixa taxa de falsos positivos (3.09%) e falsos negativos (5.15%) indica que o ponto de corte escolhido é eficaz para o propósito do modelo.

2. Densidade populacional e criminalidade nos EUA

2.1 Descrição Sumária do Dataset

O dataset utilizado na presente tarefa, a qual consiste na criação de modelos aplicados a dados em painel, é constituído por dados relativos à prática de crimes em 90 dos 100 condados do estado norte americano da Carolina do Norte (ver Figura abaixo) entre os anos de 1981 e 1987. O conjunto de dados inclui 630 linhas de registos (dados de 7 anos relativos a multiplicar pelos 90 condados) e 60 colunas, sendo que as 3 primeiras dizem respeito ao índice do registo, condado e ano. O significado de cada variável pode ser consultado em várias referências como por exemplo o seguinte link: https://rdrr.io/cran/plm/man/Crime.html.

Mapa do Estado Da Carolina do Norte

[Fonte: https://www.britannica.com/place/North-Carolina-state#/media/1/419058/61582]

2.2 Pré-Processamento do Dataset

No que respeita ao pré-processamento do dataset, este consistiu em 2 etapas:

- Reduzir o conjunto de dados aos condados a analisar;
- Eliminação de colunas.

Em relação ao primeiro ponto, sendo o presente trabalho elaborado pelo grupo 3, os dados a utilizar serão os referentes aos condados resultantes da aplicação da seguinte fórmula:

Condado (2N + 1) até ao condado (2N + 11)=>Condado 7 ao condado 17

Posto isto eliminaram-se do ficheiro csv os dados referentes aos restantes condados.

Em relação ao segundo ponto, removeram-se:

- 20 colunas (as cujo nome começa com a letra “l”) que representam apenas logaritmos de dados de outras colunas;
- 8 colunas (as cujo nome começa com as letra “cl”) que continham valores nulos e representam apenas a diferença entre os algoritmos e respetivas variáveis.
- 6 colunas (as cujo nome é constituído pela letra “d” seguido do ano) que correspondem a variáveis dummy de 6 dos 7 anos em análise.
- 3 colunas que representam variáveis binárias referentes à localização do condado no estado.
- 1 coluna, a primeira, que inclui apenas os índices dos registos.

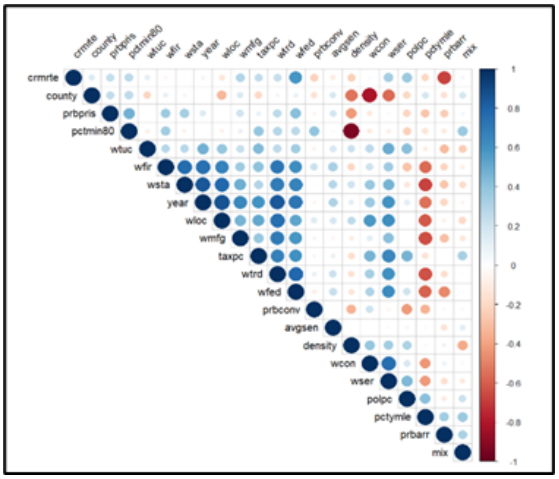
O ficheiro tratado com as 22 colunas e 42 linhas resultantes foi depois guardado no formato “xlsx” e importado no RStudio. Este conjunto de dados é então balanceado uma vez que cada condado tem o mesmo número de observações.

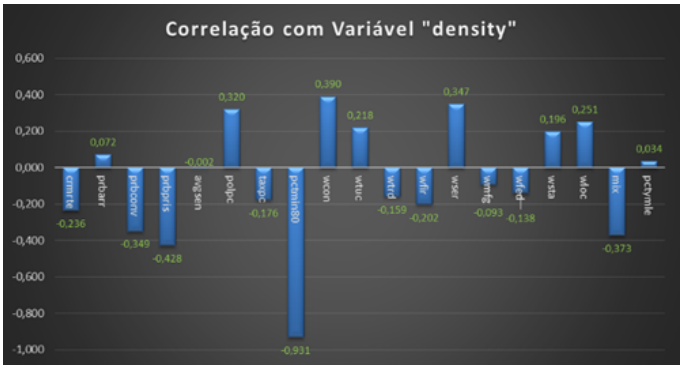
2.3 Escolha da Variável a incluir nos modelos

O enunciado do presente trabalho refere que se devem construir modelos com vista a prever a densidade populacional com base em duas variáveis, sendo uma delas a criminalidade per capita e a outra uma variável à escolha.

Como o primeiro modelo que será criado será o “pooled model”, em tudo idêntico à regressão linear múltipla, optou-se por construir uma matriz de correlação para averiguar qual a variável a utilizar.

Utilizando o comando “corrplot” criou-se o gráfico seguinte (à esquerda), que permite visualizar a correlação entre as diferentes variáveis. A tabela com os valores da correlação foi depois extraída para um ficheiro excel com o comando “xlsx” da biblioteca “writexl”. No excel construiu-se o gráfico seguinte (à direita) que apresenta os valores de correlação com a variável a prever. Pela análise do mesmo conclui-se que a variável “pctmin80”, a qual representa a percentagem minoritária em 1980, apresenta um elevado valor de correlação inversa, pelo que se optou por incluir esta variável nos modelos.





2.4 Modelo 1 - Pooled Model / Modelo dos Mínimos Quadrados Ordinários (MQO)

De seguida apresenta-se a equação do modelo dos mínimos quadrados ordinários, o qual "empilha os dados", não tendo por isso em conta a estrutura em painel dos dados, o que sigfica que é equivalente à regressão linear múltipla.

- **Eq. Pooled Model:** D^= 0.50712332 + 4.72695522*C -0.00445334*M
- $\mathbf{R^2} = 0.77081$

Relativamente à interpretação dos coeficientes anteriores pode referir-se o seguinte:

- **0.50712332:** É a densidade populacional estimada pelo modelo quando todas as variáveis são nulas
- **4.72695522:** É o aumento marginal previsto para a densidade populacional quando o a criminalidade aumenta 1 unidade e as restantes variáveis se mantêm inalteradas.
- **0.00445334:** É o aumento marginal previsto para a densidade populacional quando as minorias em 1980 aumentam 1 unidade e as restantes variáveis se mantêm inalteradas.

Já em relação ao valor coeficiente de determinação é um valor elevado que indica que o modelo explica cerca de 77% da variação da densidade populacional no entanto isto não significa necessariamente que estejamos perante o modelo desejável uma vez que o modelo de dados empilhados não permite distinguir os diferentes condados.

2.4.1. Teste F – Avaliação de Significância do MQO

Para avaliar a significância do modelo anterior recorreu-se a um teste de hipóteses F a um nível de significância de 5% e com a formulação das seguintes hipóteses:

Para avaliar a significância do modelo anterior recorreu-se a um teste de hipóteses F a um nível de significância de 5% e com a formulação das seguintes hipóteses:

- **Hipótese nula:** Todos os parâmetros são iguais a 0
- **Hipótese alternativa:** Algum parâmetro é diferente de 0

O p-value obtido foi de 3.341e-13. Assim sendo, como um este valor é menor que 0.05, deve rejeitar-se a hipótese nula, pelo que é possível concluir que existe evidência estatística para afirmar que não existem parâmetros não nulos e por conseguinte o modelo de dados empilhados é preferível ao modelo constante (que é simplesmente igual ao valor médio).

2.5 Modelo 2 - Modelo de Efeitos Fixos (MEF)

As equações obtidas com o modelo de efeitos fixos para cada condado com base nas mesmas variáveis, bem como o coeficiente de determinação, são:

#D-densidade populacional; C-criminalidade; M-minorias

- **Eq. Condado 7:** D^_7 = 0.48745 + 0.10831c + 0M
- **Eq. Condado 9:** D^_9 = 0.54035 + 0.10831c + 0M
- **Eq. Condado 11**: D^_11 = 0.60080 + 0.10831c + 0M
- **Eq. Condado 13:** D^_13 = 0.50859 + 0.10831c + 0M
- **Eq. Condado 15:** D^_15 = 0.30113 + 0.10831c + 0M
- **Eq. Condado 17:** D^_17 = 0.34748 + 0.10831c + 0M
- $\mathbf{R^2}$ = 0.0030141

Das equações anteriores salta à vista o facto de a variável minorias ser excluída pelo modelo. Tal deve-se ao facto dessa variável, apesar de variar de condado para condado, ser invariante no tempo. Como essas são precisamente as características das constantes de efeitos fixos a variável é então descartada pelos modelos.

Em relação ao coeficiente de determinação é importante realçar que este acaba por não ter muito interesse pois não diz respeito a cada um dos modelos dos diferentes condados.

2.6 Modelo 3 - Modelo de Efeitos Aleatórios (MEA)

Por fim foi criado o modelo de efeitos aleatórios, cuja equação e R^2 é apresentada de seguida:

- **Eq. MEA:** D^= 0.57991629 + 0.43776198*C - 0.00397076*M
- **R^2** = 0.58218

2.7. Comparação de Modelos

Para averiguar qual o modelo mais fiável foram feitos três testes, todos a um nível de significância de 5%: teste F, teste de Breusch – Pagan e o teste de Hausman, cujos resultados e significados serão alvo de discussão seguidamente.

2.7.1. Teste F - Existência de efeitos fixos

- **Hipótese nula:** Todos os efeitos fixos são iguais
- **Hipótese alternativa:** Existe pelo menos um efeito fixo diferente dos restantes.

O p-value obtido foi de 2.2e-16, menor que 0.05, pelo que se rejeitou a hipótese nula, o que significa que existe evidência estatística para concluir que os efeitos fixos não são todos iguais e, por conseguinte, o modelo de efeitos fixos é preferível ao modelo de dados empilhados.

2.7.2. Teste de Breusch – Pagan – Existência de Efeitos Aleatórios

- **Hipótese nula:** Não existem efeitos aleatórios
- **Hipótese alternativa:** Existem efeitos aleatórios.

O p-value obtido foi de <2.2e-16, menor que 0.05, pelo que se rejeitou a hipótese nula, o que significa que existe evidência estatística para concluir que existem diferentes efeitos aleatórios. Posto isto, assume-se que existem efeitos aleatórios e, portanto, o modelo de efeitos aleatórios é preferível ao de dados empilhados.

2.7.3 . Teste de Hausman – Consistência de MEF e MEA

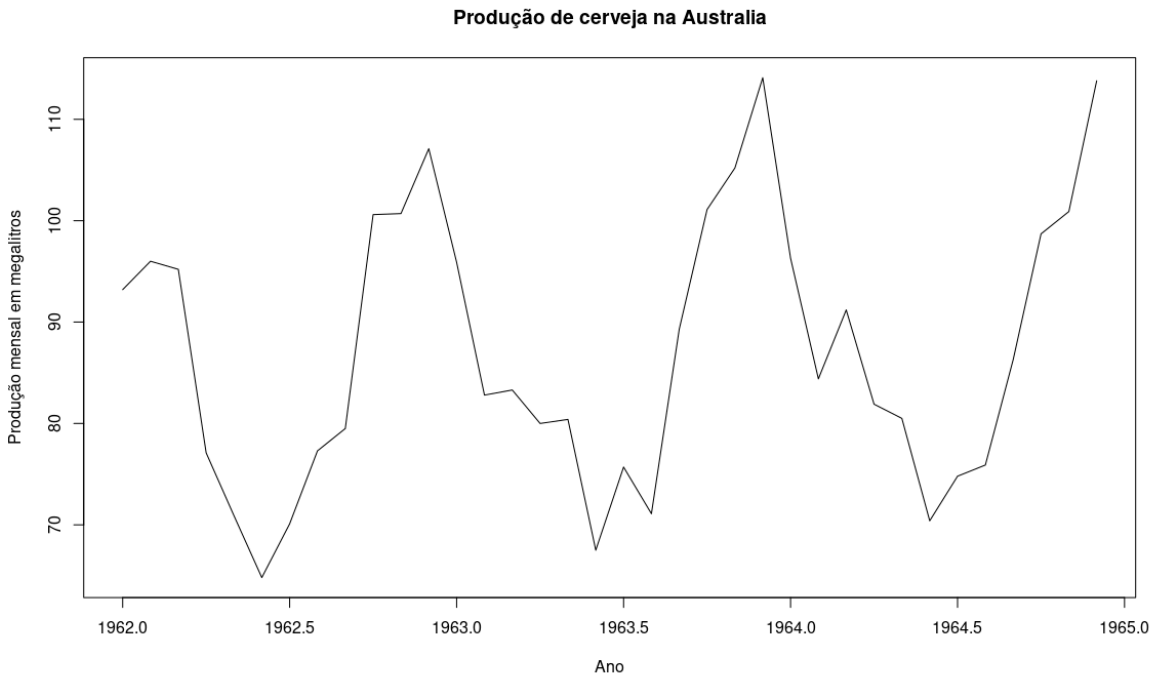
Como se concluiu no teste anterior que existem efeitos aleatórios e fixos, ou seja, que quer o MEF quer o MEA são preferíveis ao *Pooled Model,* fez-se de seguida teste de Hausman para averiguar se MEA é confiável.

- **Hipótese nula:** MEF e MEA são consistentes
- **Hipótese alternativa:** MEF é consistente e MEA é inconsistente

O p-value obtido foi de 0.3553, maior que 0.05, pelo que não se rejeitou a hipótese nula, o que significa que não existe evidência estatística para afirmar que MEA é um modelo inconsistente pelo que, à falta de prova em contrário, assume-se que este modelo é consistente e que quer o MEF quer o MEA são modelos fiáveis.

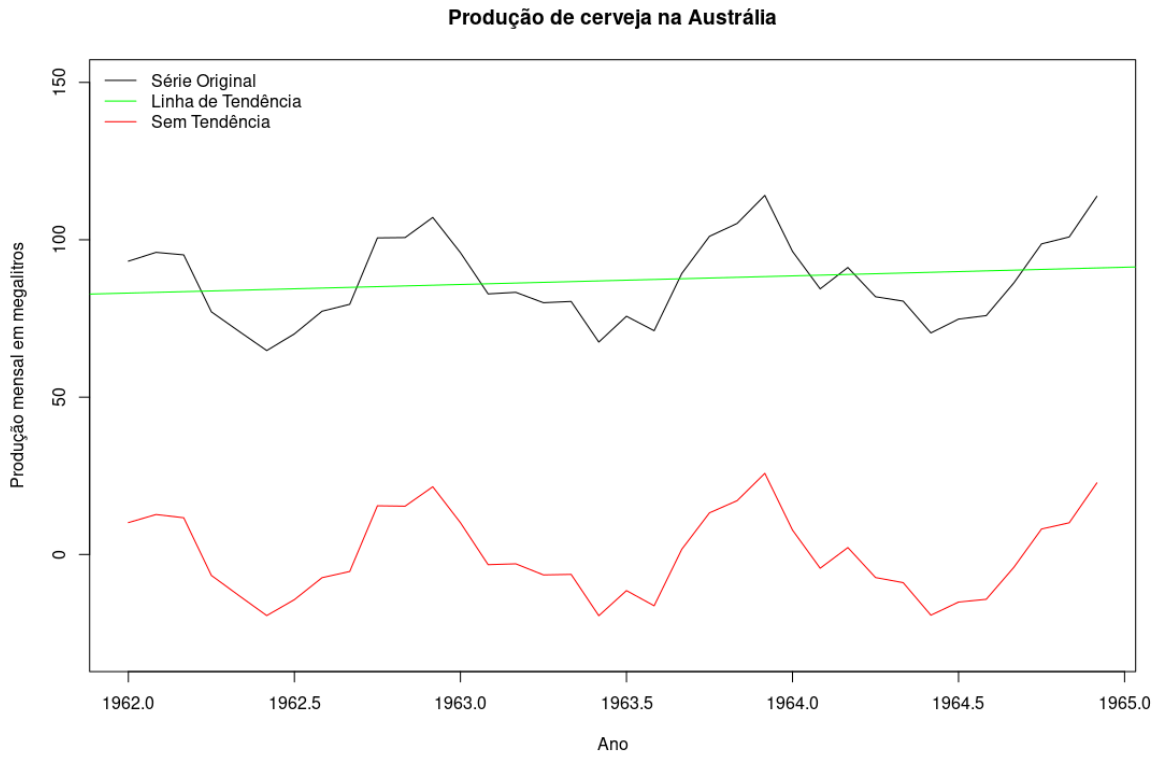
3. Produção de cerveja na Austrália

De seguida apresenta-se a produção de cerveja na Austrália, mensalmente e em megalitros, entre os períodos de Janeiro de 1962 e Dezembro de 1964.



3.1. Características de uma Série Temporal

3.1.1. Tendência



Na representação gráfica acima temos a linha preta que representa a produção mensal de cerveja ao longo dos anos em megalitros com flutuações visíveis ao longo do tempo que indicam variações na produção mensal.

Depois na linha verde, a linha de regressão que mostra a tendência geral da produção de cerveja ao longo do tempo. A equação da tendência é dada por $T_t = -5283.874792 + 2.735444 \times t$, onde t representa o tempo.

Finalmente, a vermelho, a série temporal com tendência removida. Destaca-se as variações sazonais e flutuações que não são explicadas pela tendência linear.

Com estes dados conseguimos ver que a produção de cerveja na Austrália teve variações mensais significativas entre 1962 e 1965, com uma tendência geral de aumento. Removendo essa tendência, podemos observar melhor as flutuações sazonais e outras variações a curto prazo na produção.

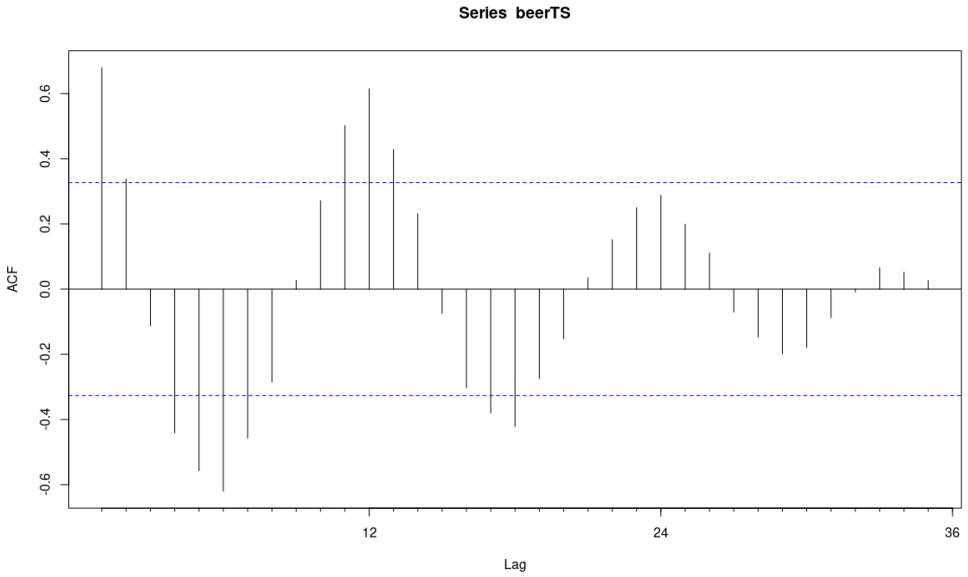
3.1.2. Sazonalidade

3.1.2.1. ACF

Os picos que ultrapassam as linhas tracejadas azuis são estatisticamente significativos, indicando correlações importantes entre os valores da série temporal em diferentes lags.

O primeiro lag mostra uma auto-correlação positiva significativa, sugerindo que os valores de um mês estão fortemente correlacionados com os valores do mês seguinte. Além disso, há um pico significativo em torno do lag 12, indicando uma auto-correlação sazonal anual. Este padrão sazonal é reforçado por picos repetidos em intervalos de 12 lags (por exemplo, nos lags 12, 24, etc.), sugerindo que a produção de cerveja segue um ciclo repetitivo a cada 12 meses. Isso evidencia a presença de uma variação sazonal clara na produção.

Entre os lags significativos, observa-se um padrão de descida, que pode indicar a presença de componentes regressivos na série temporal. Além disso, alguns picos negativos apontam para uma correlação inversa em certos lags, o que significa que altos valores em um ponto no tempo estão correlacionados com baixos valores após determinados períodos.

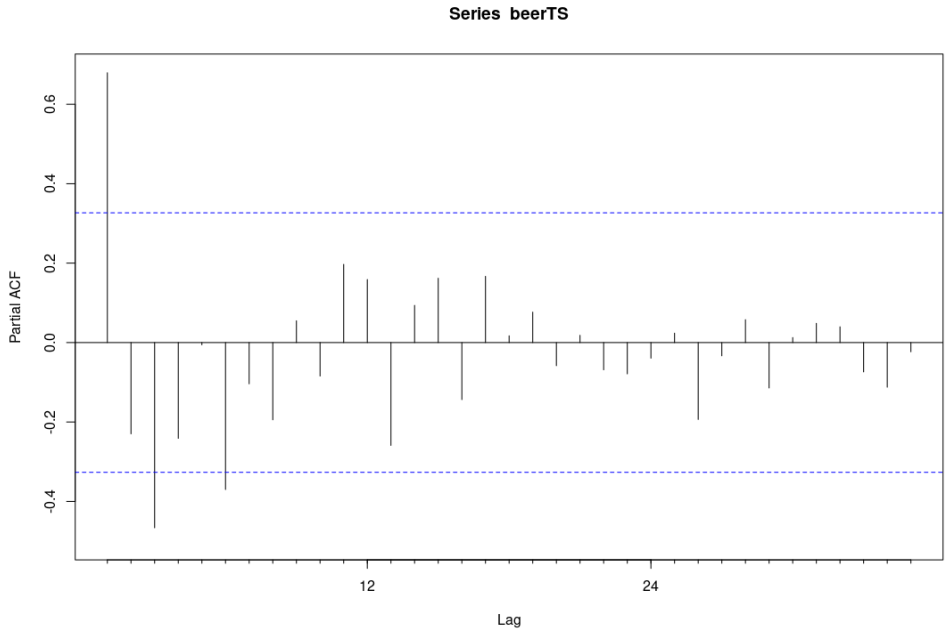


3.1.2.2. PACF

No gráfico de PACF, os picos que ultrapassam as linhas tracejadas azuis são estatisticamente significativos, indicando uma correlação direta significativa em certos lags.

O primeiro lag mostra uma correlação parcial positiva significativa, confirmando a forte correlação entre os valores de um mês e os valores do mês seguinte, mesmo após remover a influência dos lags intermediários. Ao contrário do gráfico de ACF, não há picos significativos claros nos lags múltiplos de 12 (por exemplo, 12, 24, etc.), sugerindo que a correlação sazonal não é tão forte quando os efeitos dos lags intermediários são controlados. No entanto, ainda pode haver uma indicação de padrões sazonais mais sutis.

Em resumo, a produção de cerveja tem uma forte correlação direta a curto prazo (lag 1), mesmo após remover a influência dos lags intermediários. A ausência de picos significativos em múltiplos de 12 lags indica que a sazonalidade não é tão pronunciada na PACF quanto na ACF, ou que a sazonalidade é capturada por uma combinação de lags.



3.1.3. Componente cíclica

Uma vez que não existem quantidades de cerveja produzida num período de tempo considerável que sejam bastante diferentes das temperaturas típicas não é visível qualquer componente cíclica.

3.1.4. Componente irregular

Nesta série específica, embora a componente irregular seja mais suave comparada a outras séries, ainda é possível observar pequenas variações aleatórias. Estas variações são evidentes em pequenas subidas nas épocas onde geralmente se esperaria uma descida na produção, ou vice-versa. Estas flutuações não seguem uma tendência específica ou um padrão sazonal definido, mas são resultado de fatores aleatórios que influenciam a produção de cerveja, como mudanças súbitas na demanda, interrupções na produção, ou outros eventos imprevistos.

A presença desta componente irregular é crucial para entender a totalidade da série temporal, pois ela representa a parte imprevisível e não sistemática dos dados. Apesar de ser mais suave nesta série, a componente irregular contribui para a complexidade da análise e para a necessidade de modelos robustos que possam acomodar essas variações aleatórias.

3.2. Estimativa do modelo

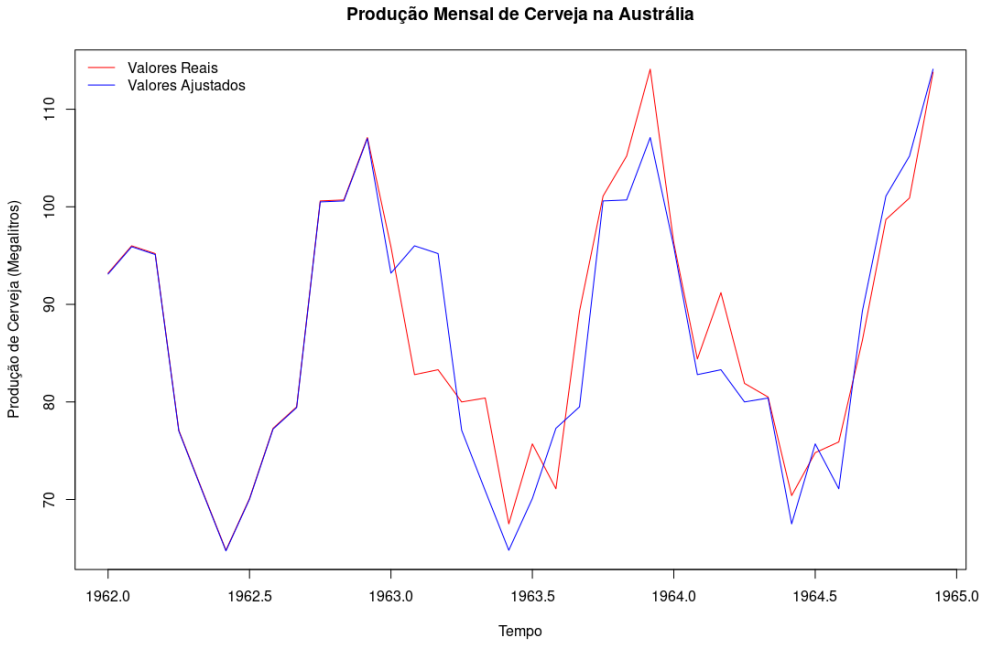
Uma vez que a série temporal apresenta sazonalidade, foi estimado um modelo SARIMA.

Modelo SARIMA obtido: ARIMA(0,0,0)(0,1,0)[12] with drift

O RMSE de 4.71 e o MAE de 3.00 indicam que, em média, as previsões do modelo estão cerca de 4.71 e 3.00 unidades longe dos valores reais, respectivamente. Embora esses valores forneçam uma medida geral da precisão, é importante considerar o contexto da série temporal.

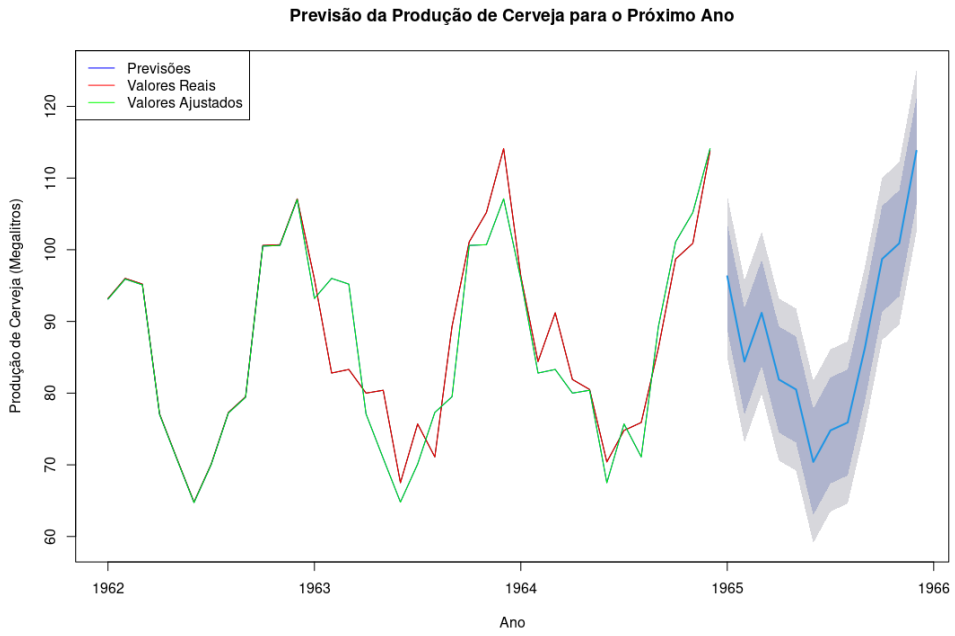
O MAPE de 3.53% indica que, em média, as previsões do modelo diferem dos valores reais em cerca de 3.53%, o que pode ser considerado um bom resultado dependendo do contexto da série temporal.

O ME de 0.656 sugere uma pequena tendência do modelo em subestimar os valores reais.



Representando graficamente a série ajustada pelo modelo SARIMA é possível visualizar, que a qualidade do Modelo SARIMA é alta apesar de apresentar algumas irregularidades.

3.3. Previsão do modelo



De acordo com as previsões realizadas pelo modelo, no ano de 1965, a produção atingirá 113.8 megalitros de cerveja, tendo uma queda para 70.4 megalitros a meio do ano.