Universidade do Minho
Escola de Ciências

# Modelação Estatística de Alugueres Diários de Bicicletas: Uma Abordagem com Regressão Binomial Negativa

**Mestrado em Estatística para a Ciência de Dados**

**Modelos Lineares Generalizados**

**Anita Margarida Antunes Ferreira - PG56093**
**Inês Margarida Gonçalves Gomes - PG55575**
**Rui Miguel Pereira Alves - PG55577**

maio de 2025

# Índice



# Índice de figuras



# Índice de tabelas

# 1 Introdução

Com o crescente interesse por soluções de mobilidade urbana sustentáveis, os sistemas de aluguer de bicicletas têm-se afirmado como uma alternativa eficiente de transporte nas cidades. A análise de dados provenientes destes sistemas revela-se fundamental para compreender os padrões de utilização e os fatores que influenciam a procura. Neste contexto, o presente relatório tem como objetivo ajustar e validar modelos estatísticos para explicar o número diário de alugueres de bicicletas, com base em variáveis meteorológicas, sazonais e contextuais.

O trabalho teve início com uma análise exploratória dos dados, que permitiu identificar a distribuição das variáveis, tendências sazonais, valores extremos e possíveis relações entre as covariáveis e a variável de resposta. Esta etapa preliminar foi essencial para orientar as escolhas de modelação subsequentes.

Considerando que a variável resposta representa contagens, foi inicialmente ajustado um modelo de regressão de Poisson. No entanto, a análise da média e variância da variável de resposta, complementada por testes formais, revelou a presença de sobredispersão. Diante deste diagnóstico, optou-se pela utilização da regressão binomial negativa, mais adequada para lidar com variância superior à média.

Para selecionar as variáveis mais relevantes, foram aplicados métodos automáticos de seleção — *forward*, *backward* e *stepwise* — com base nos critérios de informação AIC e BIC. A qualidade do modelo foi avaliada através de testes estatísticos, da análise dos resíduos e da verificação de pressupostos como a ausência de colinearidade entre as covariáveis, analisada com recurso ao VIF.

Os resultados obtidos permitiram identificar os principais determinantes da procura por bicicletas, produzir um modelo estatisticamente sólido e interpretável, com potencial para apoiar decisões mais eficazes no planeamento urbano e na gestão de políticas de mobilidade.

## 2 Descrição da Base de Dados

A base de dados utilizada neste estudo é denominada *Bike Sharing Dataset*, proveniente do *UCI Machine Learning Repository*. Esta base contém dados diários sobre o número de bicicletas alugadas, bem como variáveis meteorológicas e temporais, ao longo dos anos de 2011 e 2012.

O conjunto de dados é composto por 731 observações e 16 variáveis, contemplando informações temporais, condições meteorológicas e indicadores de atividade. A variável dependente é o número total de bicicletas alugadas por dia, representado pela variável *cnt*. As variáveis presentes na base de dados são as seguintes:

- ***instant***: identificador sequencial de cada registo;
- ***dteday***: data do registo;
- ***season***: variável categórica indicando a estação do ano (1 = inverno, 2 = primavera, 3 = verão, 4 = outono);
- ***yr***: variável binária que indica o ano (0 = 2011, 1 = 2012);
- ***mnth***: mês do ano (1 a 12);
- ***holiday***: variável binária que indica se o dia é feriado (1 = feriado, 0 = não feriado);
- ***weekday***: dia da semana (0 = domingo, 1 = segunda-feira, 2 = terça-feira, 3 = quarta-feira, 4 = quinta-feira, 5= sexta-feira, 6 = sábado);
- ***workingday***: variável binária que assume o valor 1 se o dia é útil e 0 caso contrário;
- ***weathersit***: situação climática categorizada: 1 = Céu limpo, parcialmente nublado, 2 = Nevoeiro e nuvens dispersas, 3 = Chuva leve ou neve leve;
- ***temp***: temperatura média diária, normalizada com base na fórmula $(t - t_{\text{min}})/(t_{\text{max}} - t_{\text{min}})$, onde $t_{\text{min}} = -8°\text{C}$ e $t_{\text{max}} = 39°\text{C}$;
- ***atemp***: sensação térmica média diária, também normalizada segundo a fórmula $(t - t_{\text{min}})/(t_{\text{max}} - t_{\text{min}})$, com $t_{\text{min}} = -16°\text{C}$ e $t_{\text{max}} = 50°\text{C}$. Representa uma medida subjetiva de conforto térmico;
- ***hum***: humidade relativa média diária, expressa como uma proporção entre 0 e 1. Corresponde à divisão da humidade por 100;
- ***windspeed***: velocidade média do vento por dia, normalizada através da divisão pelo valor máximo teórico de 67;
- ***casual***: número de bicicletas alugadas por utilizadores ocasionais não registados;
- ***registered***: número de bicicletas alugadas por utilizadores registados;
- ***cnt***: total de bicicletas alugadas por dia (*cnt* = *casual* + *registered*).

A riqueza informacional desta base de dados permite não apenas analisar os padrões de utilização de bicicletas partilhadas em função de variáveis temporais e meteorológicas, como também investigar de que forma fatores como a estação do ano, o tipo de dia e as condições climáticas influenciam a procura por este serviço. Em particular, a modelação adequada da variável resposta, o número total de bicicletas alugadas por dia, justifica o uso de modelos de contagem, sendo neste estudo adotada a regressão de Binomial Negativa como abordagem principal.

# 3 Análise Exploratória

A análise exploratória tem como objetivo perceber o comportamento das variáveis em estudo antes de ajustar qualquer modelo. Nesta fase, identificam-se padrões, relações entre variáveis, valores atípicos e possíveis problemas nos dados. Esta análise é essencial para garantir a qualidade da informação e orientar a construção do modelo estatístico.

Importa referir que as variáveis *instant*, *dteday*, *casual* e *registered* não serão utilizadas como preditores. A variável *instant* corresponde apenas a um identificador sequencial e a *dteday* representa a data do registo, não fornecendo informação adicional útil para o modelo. As variáveis *casual* e *registered*, por sua vez, representam componentes do total de alugueres diários de bicicletas (*cnt*), a variável resposta do nosso estudo, sendo por isso excluídas para evitar redundância e violação do princípio de independência entre variáveis explicativas e a variável dependente. Consequentemente, estas quatro variáveis também não serão consideradas na análise exploratória, por não contribuírem para os objetivos preditivos do modelo estatístico a ser desenvolvido.

## 3.1 Análise Exploratória Univariada

Nesta fase, analisa-se cada variável isoladamente, com o objetivo de compreender a sua distribuição, dispersão e características principais. Para as variáveis quantitativas, apresentam-se na Tabela 3.1 alguns indicadores estatísticos essenciais, como a média, o desvio padrão e os valores extremos, permitindo uma primeira leitura do comportamento dos dados.

| Variável | Mínimo | 1.º Quartil | Mediana | Média | 3.º Quartil | Máximo |
|---|---|---|---|---|---|---|
| *temp* | -5.22 | 7.84 | 15.42 | 15.28 | 22.81 | 32.5 |
| *atemp* | -10.78 | 6.29 | 16.12 | 15.31 | 24.17 | 39.5 |
| *hum* | 0 | 52 | 62.67 | 62.79 | 73.02 | 97.25 |
| *windspeed* | 1.5 | 9.04 | 12.12 | 12.76 | 15.63 | 34 |
| *cnt* | 22 | 3152 | 4548 | 4504 | 5956 | 8714 |

Tabela 3.1: Estatísticas descritivas das variáveis quantitativas contínuas

Os valores observados para cada variável quantitativa situam-se nos seguintes intervalos:

- *temp*: assume valores entre $-5.22°\text{C}$ e $32.5°\text{C}$, representando a temperatura média diária desnormalizada;
- *atemp*: varia entre $-10.78°\text{C}$ e $39.5°\text{C}$, correspondendo à sensação térmica média diária, também desnormalizada;
- *hum*: assume valores no intervalo $[0\% \ ; \ 97.25\%]$, correspondendo à humidade relativa diária em percentagem;
- *windspeed*: varia entre 1.5 e 34 km/h, representando a velocidade média diária do vento desnormalizada;
- *cnt*: variável de contagem que representa o número total de bicicletas alugadas por dia, com valores inteiros entre 22 e 8714.

Em relação às variáveis qualitativas, a análise destas serão através da construção de gráficos de barras, com o objetivo de perceber a distribuição das categorias em cada variável.

| Estação | Frequência | Percentagem (%) |
|---|---|---|
| Inverno | 181 | 24,8 |
| Primavera | 184 | 25,2 |
| Verão | 188 | 25,7 |
| Outono | 178 | 24,4 |

Tabela 3.2: Distribuição da variável *season*

A Tabela 3.2 apresenta a distribuição das observações segundo a estação do ano. Verifica-se uma repartição relativamente equilibrada entre as quatro estações, o que é expectável dado que os dados abrangem dois anos completos. A estação com maior número de alugueres registados é o verão (25.7%), seguida da primavera (25.2%) e do inverno (24.8%). O outono apresenta a menor proporção (24,4%).

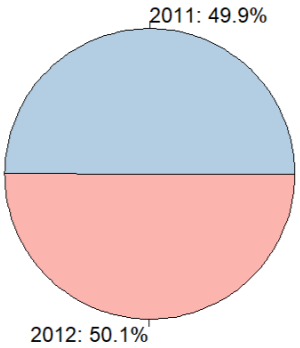


Figura 3.1: Distribuição das observações por ano

Verifica-se uma repartição praticamente equilibrada entre os dois anos considerados no estudo: cerca de 49,93% das observações referem-se ao ano de 2011 e 50,07% ao ano de 2012. Esta uniformidade na distribuição temporal é importante, pois assegura que os resultados obtidos não estão enviesados por uma eventual desproporção na amostragem entre os anos.

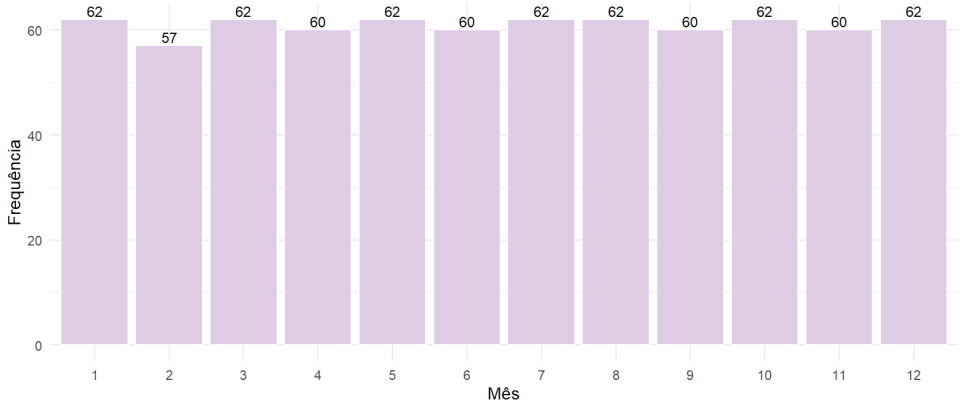


Figura 3.2: Distribuição das observações por mês

O histograma apresenta a distribuição das observações por mês do ano. Verifica-se uma distribuição relativamente uniforme, com cerca de 60 a 62 observações por mês, com exceção do mês de fevereiro, que apresenta uma frequência ligeiramente inferior. Esta regularidade na frequência mensal é consistente com o facto de os dados cobrirem dois anos completos e reflete uma amostragem equilibrada ao longo do tempo.

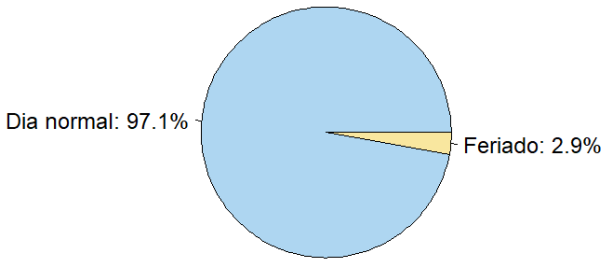


Figura 3.3: Distribuição das observações segundo o tipo de dia (feriado ou dia normal)

Observa-se que a esmagadora maioria dos registos corresponde a dias normais, enquanto que apenas 2.9% das observações ocorrem em feriados. Esta assimetria é esperada, dado o reduzido número de feriados ao longo do ano, e reflete-se ao longo dos dois anos cobertos pela base de dados (2011 e 2012).

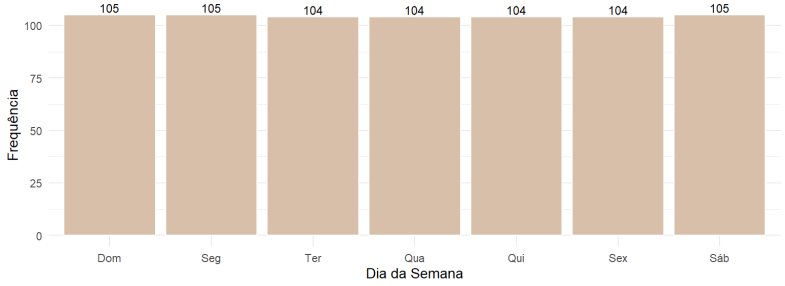


Figura 3.4: Distribuição das observações por dia da semana

Há uma repartição bastante uniforme, com aproximadamente 104 a 105 registos para cada dia, o que sugere que a recolha de dados foi feita de forma equilibrada ao longo dos diferentes dias da semana, sem sobrecarga ou omissão de nenhum deles.

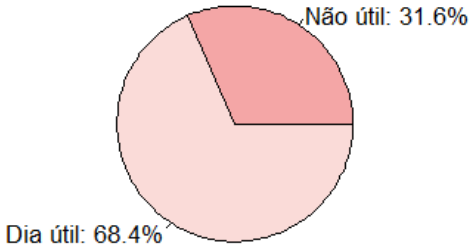


Figura 3.5: Distribuição das observações por tipo de dia (útil ou não)

A Figura 3.5 mostra a distribuição das observações segundo o tipo de dia, distinguindo entre dias úteis e não úteis. Verifica-se que 68.4% dos registos correspondem a dias úteis, enquanto os restantes 31.6% dizem respeito a fins de semana ou feriados. Esta assimetria é esperada, uma vez que, ao longo de um calendário anual, os dias úteis são naturalmente mais frequentes.

| Condição Climática | Frequência | Percentagem (%) |
|---|---|---|
| Céu limpo / Poucas nuvens | 463 | 63.4 |
| Nevoeiro / Nuvens dispersas | 247 | 33.8 |
| Chuva leve / Neve leve | 21 | 2.9 |

Tabela 3.3: Distribuição da variável *weathersit*

A Tabela 3.3 apresenta a distribuição das observações segundo a condição climática no momento do aluguer. A maioria dos registos (63.4%) corresponde a dias com céu limpo ou apenas parcialmente nublado, seguidos por situações de neblina ou nuvens dispersas (33.8%). Apenas uma pequena fração das observações (2.9%) ocorreu em dias com chuva leve ou neve leve. Esta distribuição é coerente com a expectativa de que os utilizadores tendem a alugar bicicletas principalmente em condições meteorológicas favoráveis.

## 3.2 Análise Exploratória Bivariada

Após a análise individual de cada variável, procede-se agora à análise exploratória bivariada, com o objetivo de identificar possíveis relações entre a variável resposta *cnt* e as restantes variáveis explicativas. Esta etapa permite observar padrões, dependências ou associações relevantes que poderão justificar a inclusão dessas variáveis no modelo estatístico. Serão utilizados diferentes tipos de representações gráficas, consoante a natureza das variáveis envolvidas, de forma a facilitar a interpretação visual dessas relações.

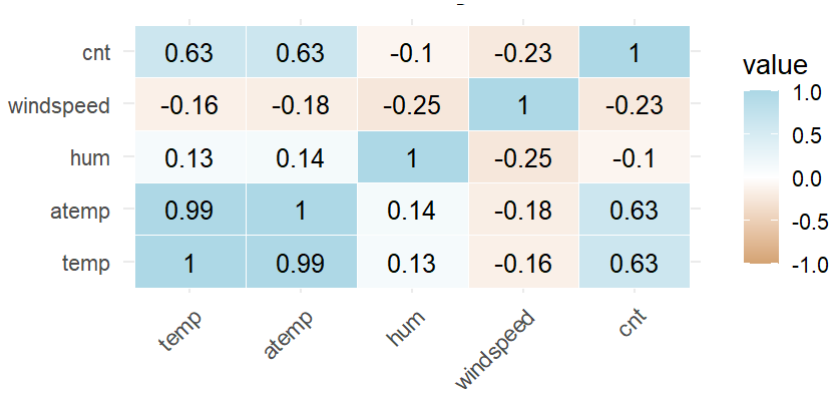


Figura 3.6: Matriz de correlação entre as variáveis contínuas e a variável *cnt*

A Figura 3.6 apresenta a matriz de correlação de *Spearman* entre a variável resposta *cnt* e as variáveis explicativas contínuas normalizadas. Observa-se uma forte correlação positiva de 0,63 entre a temperatura média diária (*temp*) e o número total de bicicletas alugadas, o que indica que a procura tende a aumentar em dias mais quentes. A sensação térmica (*atemp*), que representa o conforto térmico percebido, apresenta um padrão praticamente idêntico reforçando essa tendência. Por outro lado, a humidade (*hum*) exibe uma correlação negativa moderada com *cnt* com o valor de -0,10, o que sugere que níveis mais elevados de humidade estão ligeiramente associados a uma menor utilização do serviço. A velocidade do vento (*windspeed*) apresenta uma correlação ainda mais negativa (-0,23), embora fraca, o que poderá indicar algum desconforto ou desincentivo ao uso da bicicleta em dias com vento mais intenso.

Após a análise da relação entre as variáveis contínuas e a variável resposta *cnt*, procede-se agora à exploração das variáveis qualitativas mais relevantes.

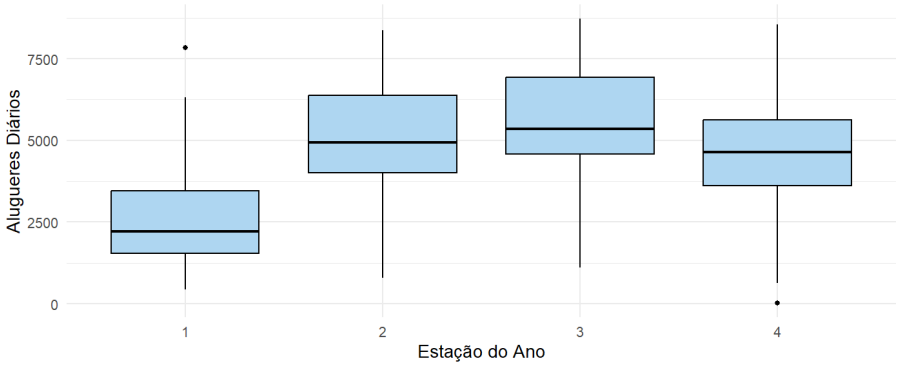


Figura 3.7: Distribuição dos alugueres diários (*cnt*) por estação do ano

A Figura 3.7 apresenta a distribuição do número de alugueres diários por estação do ano (*season*). Verifica-se que os valores mais baixos de aluguer ocorrem durante o inverno (1), com uma mediana inferior a 2500. Já na primavera (2) e no verão (3), observa-se um aumento expressivo no número de alugueres, com medianas superiores a 5000. No outono (4), os valores mantêm-se elevados, ainda que ligeiramente abaixo dos meses de verão. Estes resultados confirmam a forte influência da sazonalidade no comportamento dos utilizadores, com maior procura nos períodos de clima mais ameno.

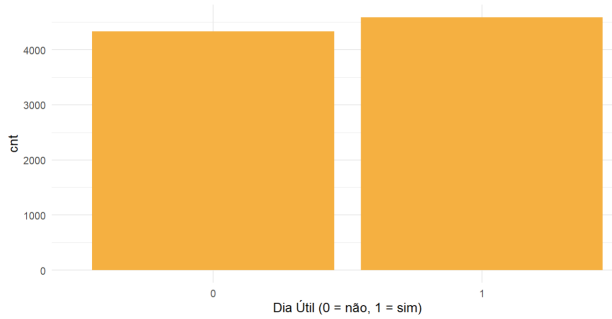


Figura 3.8: Média de alugueres diários (*cnt*) em dias úteis e não úteis

A Figura 3.8 apresenta a média do número de alugueres diários (*cnt*) em dias úteis (valor 1) e não úteis (valor 0), acompanhada do respetivo erro padrão. Verifica-se que, em média, os dias úteis registam um número ligeiramente superior de alugueres, o que pode estar associado ao uso da bicicleta como meio de transporte para deslocações regulares. Ainda assim, os dias não úteis também apresentam valores médios elevados, o que sugere uma utilização significativa para lazer.

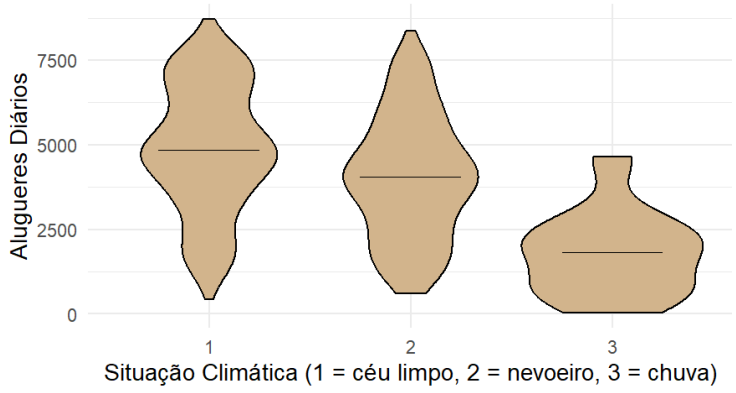


Figura 3.9: Distribuição de alugueres diários (*cnt*) por situação climática

Por fim a Figura 3.9 mostra a distribuição do número de alugueres diários (*cnt*) em função da situação climática, representada através de um *violin plot*. Observa-se que os dias com céu limpo (categoria 1) apresentam uma maior concentração de alugueres elevados, com a distribuição claramente deslocada para valores superiores a 5000. Em dias de nevoeiro, a distribuição mantém-se semelhante, embora ligeiramente mais dispersa e com menos alugueres muito altos. Já em dias de chuva leve, a procura por bicicletas diminui substancialmente, com a distribuição concentrada em valores bastante mais baixos. Estes resultados são consistentes com a expectativa de que o clima influencia diretamente o uso de bicicletas, sendo as condições adversas um fator inibidor da utilização do serviço.

## 3.3 Testes Estatísticos

Para aprofundar a análise e responder a questões específicas sobre os determinantes da procura por bicicletas, iremos testar, com suporte estatístico, hipóteses concretas sobre o efeito de determinadas variáveis explicativas.

**Pergunta 1** - As variáveis categóricas influenciam significativamente o número de bicicletas alugadas?

- $H_0$ : As distribuições do número de bicicletas alugadas são iguais entre os diferentes grupos de cada variável categórica analisada
- $H_1$ : Pelo menos uma das variáveis categóricas possui grupos com distribuições significativamente diferentes

| Variável | Valor de prova |
|---|---|
| *season* | $< 2 \times 10^{-16}$ |
| *mnth* | $< 2 \times 10^{-16}$ |
| *weekday* | $0.6311$ |
| *weathersit* | $2.589 \times 10^{-15}$ |
| *yr* | $< 2.2 \times 10^{-16}$ |
| *holiday* | $0.08314$ |
| *workingday* | $0.1186$ |

Tabela 3.4: Valores de prova do teste de *Kruskal–Wallis* para variáveis categóricas

Com base nos valores de prova apresentados na Tabela 3.4, observa-se que as variáveis *season*, *mnth*, *weathersit* e *yr* apresentam valores de prova significativamente inferiores ao nível de significância de 5%. Assim, rejeita-se a hipótese nula $H_0$ para estas variáveis, concluindo-se que pelo menos um dos grupos associados a cada uma delas apresenta uma distribuição distinta no número de bicicletas alugadas. Por outro lado, para as variáveis *weekday*, *holiday* e *workingday*, os valores de prova são superiores a 0.05, pelo que não se rejeita $H_0$. Isso indica que não existem evidências estatísticas suficientes para afirmar que estas variáveis categóricas influenciam significativamente a distribuição do número de alugueres.

**Pergunta 2** - A temperatura e a presença de feriados afetam a aderência diária pelas bicicletas?

- $H_0$ : Nenhuma das variáveis *temp* e *holiday* influencia significativamente o número de alugueres
- $H_1$ : Pelo menos uma das variáveis tem um efeito significativo sobre o número de alugueres

O teste de *Wald*, aplicado aos coeficientes estimados no modelo de Binomial Negativa, revelou que a variável *temp* tem um coeficiente estimado de $\hat{\beta}_{\text{temp}} = 1.7479$, com erro padrão de 0.0823 e estatística de teste $z = 21.24$. A variável *holiday*, tratada como categórica com o nível **0 (não feriado)** como referência, apresentou um coeficiente $\hat{\beta}_{\text{holiday=1}} = -0.1927$, com erro padrão de 0.0902 e estatística $z = -2.14$. Os valores de prova associados foram, respetivamente, $< 2 \times 10^{-16}$ e 0.0325, indicando que ambos os efeitos são estatisticamente significativos ao nível de 5%. Assim, rejeita-se $H_0$, concluindo-se que tanto a temperatura como a ocorrência de feriados influenciam significativamente a procura diária por bicicletas. Em particular, ser feriado está associado a uma redução esperada no número de alugueres, em comparação com dias úteis.

# 4 Análise do Modelo

## 4.1 Sobredispersão

Dados de contagem referem-se ao número de ocorrências de um determinado evento dentro de uma mesma unidade de observação, ao longo de um intervalo fixo de tempo ou espaço. Para a modelação estatística de dados de contagem, é frequentemente utilizado o modelo de regressão de Poisson, que assume que a variável resposta segue uma distribuição de Poisson.

Um fenómeno que ocorre com frequência nas aplicações é o fenómeno de sobredispersão. Sobredispersão surge quando a variância da variável resposta é superior ao valor da média. Por conseguinte, inicialmente, foi ajustado aos dados em estudo um modelo de regressão de Poisson considerando todas as variáveis explicativas disponíveis no conjunto de dados, tendo a variável *cnt* como variável resposta. Observou-se, contudo, que a variância da variável resposta era substancialmente superior ao valor da média, contrariando a principal suposição do modelo de Poisson, que exige que $\text{Var}(Y) = \mathbb{E}(Y)$.

A presença de sobredispersão foi formalmente avaliada por meio da função *check_overdispersion()*, do pacote *performance* no software R. Os resultados obtidos indicaram um índice de dispersão estimado em $\hat{\phi} \approx 152{,}4$, com um valor da estatística qui-quadrado de Pearson igual a $106832{,}42$ e um p-valor inferior a $0{,}001$, confirmando de forma estatisticamente significativa a existência de sobredispersão nos dados. Assim, o modelo de Poisson não foi considerado adequado.

Alternativamente, uma forma gráfica de avaliar se o modelo ajustado é adequado, e de identificar a possível presença de sobredispersão, é através do *envelope plot*. Este gráfico é uma extensão do gráfico quantil-quantil (Q-Q plot), no qual os resíduos obtidos do modelo são comparados com os resíduos esperados sob a distribuição normal.

Se os pontos do gráfico se afastarem de forma visível da linha reta, isso indica que os resíduos não seguem a distribuição esperada, o que pode sugerir que o modelo não está bem ajustado aos dados.

O *envelope plot* simula intervalos de confiança empíricos para determinar se os resíduos diferem significativamente da linha reta. Esses intervalos são obtidos a partir da simulação de várias amostras da variável resposta, com base no modelo ajustado e na distribuição assumida para a variável resposta. Se houver sobredispersão, a projeção dos resíduos cairá fora dos intervalos.

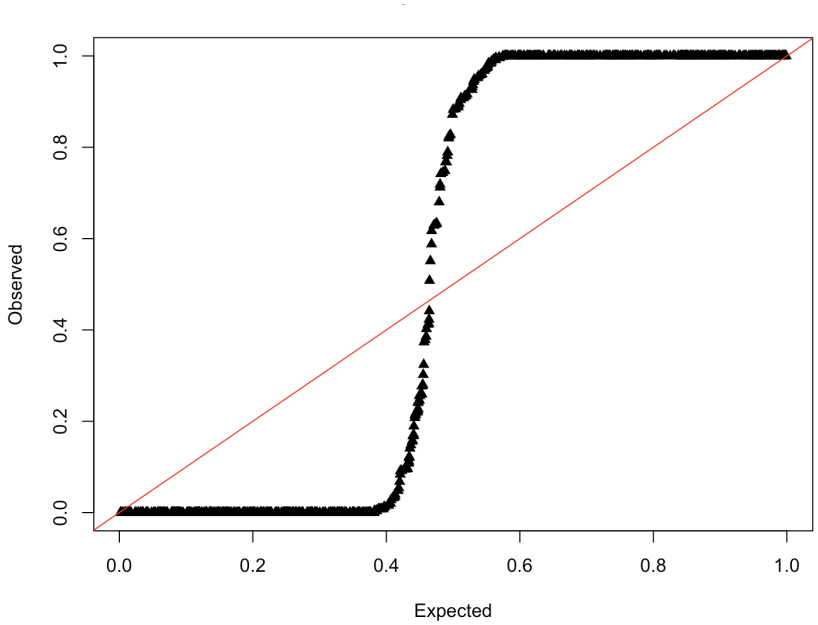


Figura 4.1: Q-Q normal dos resíduos do modelo de Poisson

O Gráfico 4.1 apresenta o Q-Q normal dos resíduos de Pearson do modelo de regressão de Poisson mencionado. Observa-se um desvio acentuado dos pontos em relação à linha reta esperada sob a hipótese de aderência à distribuição teórica. Este padrão indica que os resíduos não seguem a distribuição esperada, sendo uma evidência gráfica de que o modelo de Poisson não se ajusta adequadamente aos dados.

Esse comportamento, caracterizado por uma forte acumulação de resíduos nas extremidades e afastamento da diagonal, é típico da presença de sobredispersão, já previamente identificada por testes estatísticos. Assim, o QQ plot reforça a conclusão de que o modelo de Poisson não é apropriado, justificando a adoção de um modelo alternativo.

Deste modo, o modelo de Poisson não seria adequado para os dados em questão. O modelo de regressão Binomial Negativa, que é uma generalização do modelo de regressão de Poisson, permite resolver o problema da sobredispersão. Assim, como alternativa foi ajustado um modelo de regressão **Binomial Negativa**.

## 4.2 Seleção de variáveis

Após o ajustamento inicial do modelo com todas as variáveis explicativas disponíveis, procedeu-se à seleção de variáveis com o objetivo de obter um modelo mais parcimonioso, mantendo uma boa capacidade explicativa.

Para tal, foram aplicadas três abordagens clássicas de seleção automática de variáveis: *forward selection*, *backward elimination* e *stepwise selection*. Estas estratégias baseiam-se na comparação iterativa de modelos através do critério de informação de Akaike (AIC), adicionando ou removendo variáveis com base na sua contribuição para a melhoria do ajustamento do modelo.

Tabela 4.1: Comparação dos modelos segundo AIC, BIC e RMSE

| Modelo | AIC | BIC | RMSE |
|---|---|---|---|
| **Forward** | 12 155,03 | 12 269,89 | 1975,45 |
| **Backward** | 12 155,29 | 12 293,06 | 1008,58 |
| **Stepwise** | 12 155,29 | 12 293,06 | 1008,58 |

A Tabela 4.1 apresenta os valores do critério de informação de Akaike (AIC), do critério bayesiano de Schwarz (BIC) e do erro quadrático médio (RMSE) para cada modelo. O modelo obtido por *forward selection* apresentou os menores valores de AIC e BIC, sugerindo maior parcimônia. No entanto, esse modelo apresentou o maior RMSE entre os três, indicando uma menor precisão preditiva.

Observa-se ainda que os modelos *backward* e *stepwise* coincidiram em termos de estrutura, apresentando os mesmos valores de AIC, BIC e RMSE.

Para avaliar se a simplificação proposta pelo modelo obtido através do método de *forward selection* comprometeria a qualidade do ajustamento, procedeu-se à aplicação do teste da razão de verossimilhança (*Likelihood Ratio Test*) entre este modelo e o modelo completo. Para a realização do teste da razão de verossimilhança entre os modelos, foi utilizada a função *lrtest()* do pacote *lmtest*, disponível no ambiente estatístico R.

Tabela 4.2: Resultados do teste da razão de verossimilhança (LRT) entre os modelos reduzidos e o modelo completo

| Modelo Comparado | Diferença de Df | Estatística $\chi^2$ | $p$-valor |
|---|---|---|---|
| **Forward vs. Completo** | 5 | 9,80 | 0,081 |
| **Backward vs. Completo** | 0 | 0 | $< 2{,}2 \times 10^{-16}$ |
| **Stepwise vs. Completo** | 0 | 0 | $< 2{,}2 \times 10^{-16}$ |

A Tabela 4.2 apresenta os resultados do teste da razão de verossimilhança (*Likelihood Ratio Test*, LRT) para a comparação entre os modelos reduzidos e o modelo completo.

No caso do modelo *forward*, a estatística do teste foi $\chi^2 = 9{,}80$, com 5 graus de liberdade e um *p*-valor de $0{,}081$. Este valor, sendo superior ao nível de significância de $5\%$, indica que não há evidência estatística suficiente para rejeitar o modelo *forward* em favor do modelo completo. Em outras palavras, a simplificação proposta pelo modelo *forward* não compromete significativamente a qualidade do ajustamento.

Apesar de os métodos *backward* e *stepwise* terem retornado modelos com menor número de parâmetros, os testes de razão de verossimilhança apresentaram diferença nula de log-verossimilhança e de graus de liberdade em relação ao modelo completo, resultando em p-valores extremamente baixos ($p < 2{,}2 \times 10^{-16}$). No entanto, esses valores decorrem de limitações numéricas na implementação do teste no R. Como os modelos são estatisticamente equivalentes ao completo, e a exclusão de variáveis não produziu ganho de parcimônia nem melhoria interpretativa, optou-se por não adotar os modelos *backward* e *stepwise*.

Dessa forma, conclui-se que apenas o modelo obtido por *forward selection* pode ser estatisticamente comparado ao modelo completo, sendo considerado equivalente em termos de ajustamento, mas superior em termos de parcimônia. Por essa razão, o modelo *forward* foi adotado como modelo final.

## 4.3 Modelo final

Seja $Y$ uma variável aleatória, representando o número de alugueres diários de bicicletas com $n$ observações, $X = (X_1, \ldots, X_p)$ um vetor de covariáveis e

$$\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$$

uma observação do indivíduo $i$, e assume-se

$$Y_i \mid \mathbf{x}_i \sim \text{BN}(\mu(\mathbf{x}_i), \alpha)$$

onde $\mu_i = \mu(\mathbf{x}_i)$ é igual ao número médio de ocorrências de um dado acontecimento dada a observação $\mathbf{x}_i$.

O modelo de regressão binomial negativa, utilizando a seleção de *forward*, é definido por:

$$\begin{aligned}\log(\textit{cnt}) = \beta_0 &+ \beta_1 \cdot \text{atemp} + \beta_2 \cdot \textit{yr} + \beta_3 \cdot \textit{season} + \beta_4 \cdot \textit{weathersit} + \beta_5 \cdot \textit{mnth} \\ &+ \beta_6 \cdot \textit{windspeed} + \beta_7 \cdot \textit{hum} + \beta_8 \cdot \textit{holiday} + \beta_9 \cdot \textit{temp} + \beta_{10} \cdot \textit{workingday}\end{aligned}$$

Este modelo é adequado para dados de contagem com sobredispersão, como é o caso do presente conjunto de dados.

A Tabela 4.3 apresenta os coeficientes estimados do modelo de regressão binomial negativa ajustado por *forward selection*, juntamente com os respectivos erros padrão, estatísticas $z$ e $p$-valores.

Tabela 4.3: Coeficientes estimados do modelo binomial negativo ajustado por forward selection

| **Variável** | **Estimativa** | **Erro padrão** | $z$ | $p$**-valor** |
|---|---|---|---|---|
| (Intercept) | 7.25728 | 0.07273 | 99.790 | $<$ 2e-16 |
| atemp | 0.66256 | 0.41491 | 1.597 | 0.11029 |
| yr1 | 0.48097 | 0.01781 | 27.009 | $<$ 2e-16 |
| season2 | 0.26490 | 0.05380 | 4.924 | 8.49e-07 |
| season3 | 0.30240 | 0.06468 | 4.675 | 2.93e-06 |
| season4 | 0.48099 | 0.05593 | 8.600 | $<$ 2e-16 |
| weathersit2 | -0.12194 | 0.02324 | -5.247 | 1.55e-07 |
| weathersit3 | -0.60872 | 0.06137 | -9.918 | $<$ 2e-16 |
| mnth2 | 0.13189 | 0.04419 | 2.985 | 0.00284 |
| mnth3 | 0.22795 | 0.04988 | 4.570 | 4.88e-06 |
| mnth4 | 0.19059 | 0.07549 | 2.525 | 0.01158 |
| mnth5 | 0.23210 | 0.08266 | 2.808 | 0.00499 |
| mnth6 | 0.06857 | 0.08749 | 0.784 | 0.43317 |
| mnth7 | -0.07186 | 0.09674 | -0.743 | 0.45759 |
| mnth8 | 0.02428 | 0.09335 | 0.260 | 0.79478 |
| mnth9 | 0.18907 | 0.08144 | 2.321 | 0.02026 |
| mnth10 | 0.15354 | 0.07560 | 2.031 | 0.04228 |
| mnth11 | 0.11703 | 0.07131 | 1.641 | 0.10075 |
| mnth12 | 0.09876 | 0.05757 | 1.716 | 0.08623 |
| holiday1 | -0.31179 | 0.06279 | -4.966 | 6.84e-07 |
| windspeed | -0.65503 | 0.12569 | -5.211 | 1.87e-07 |
| hum | -0.35608 | 0.08691 | -4.097 | 4.19e-05 |
| weekday1 | 0.09336 | 0.03429 | 2.723 | 0.00648 |
| weekday2 | 0.09168 | 0.03224 | 2.844 | 0.00446 |
| weekday3 | 0.09153 | 0.03302 | 2.772 | 0.00557 |
| weekday4 | 0.11062 | 0.03306 | 3.346 | 0.00082 |
| weekday5 | 0.12719 | 0.03402 | 3.738 | 0.00019 |
| weekday6 | 0.10800 | 0.03238 | 3.335 | 0.00085 |
| temp | 0.90670 | 0.40199 | 2.256 | 0.02410 |

Observa-se que a maioria das variáveis incluídas no modelo é estatísticamente significativa ao nível de $5\%$, indicando que possuem efeito relevante na explicação do número de alugueres diários de bicicletas. Variáveis como *yr1*, *season3*, *season4*, *weathersit3*, *windspeed*, *hum* e diversas dummies de *weekday* destacam-se $p$-valores inferiores a $0{,}001$. Por outro lado, algumas variáveis como *atemp*, *mnth6*, *mnth7*, e *mnth8* não demonstram significância estatística, sugerindo que seu impacto sobre a variável resposta pode ser limitado neste modelo.

Tabela 4.4: Fatores de Inflação da Variância (VIF) para o modelo com `atemp`

| Variável | GVIF | Graus de liberdade | $\text{GVIF}^{1/(2\cdot df)}$ |
|---|---|---|---|
| atemp | 69.31 | 1 | **8.33** |
| yr | 1.05 | 1 | 1.02 |
| season | 169.37 | 3 | 2.35 |
| weathersit | 1.86 | 2 | 1.17 |
| mnth | 404.04 | 11 | 1.31 |
| windspeed | 1.27 | 1 | 1.13 |
| hum | 2.11 | 1 | 1.45 |
| holiday | 1.10 | 1 | 1.05 |
| temp | 79.96 | 1 | **8.94** |
| workingday | 1.09 | 1 | 1.04 |

A Tabela 4.4 apresenta os valores do Fator de Inflação da Variância (VIF) para as covariáveis incluídas no modelo final obtido via *forward selection*. Verifica-se que as variáveis *atemp* e *temp* apresentam valores de VIF elevados, com $\text{GVIF}^{1/(2\cdot df)}$ superiores a 5, sugerindo a presença de multicolinearidade.

Além disso, a variável *atemp* não apresentou significância estatística no modelo ($p-valor = 0{,}110$), o que, combinado com o alto VIF, justifica sua remoção. A variável *temp*, por outro lado, manteve significância estatística.

Assim, procedeu-se à exclusão de *atemp* do modelo e à reavaliação do ajustamento.

Tabela 4.5: Fatores de Inflação da Variância (VIF) após remover a varável atemp

| Variável | GVIF | Graus de liberdade | $\text{GVIF}^{1/(2\cdot df)}$ |
|---|---|---|---|
| yr | 1.05 | 1 | 1.02 |
| season | 169.16 | 3 | 2.35 |
| weathersit | 1.85 | 2 | 1.17 |
| mnth | 387.48 | 11 | 1.31 |
| windspeed | 1.22 | 1 | 1.10 |
| hum | 2.11 | 1 | 1.45 |
| holiday | 1.10 | 1 | 1.05 |
| temp | 6.97 | 1 | **2.64** |
| workingday | 1.09 | 1 | 1.04 |

Como se pode observar, os valores de $\text{GVIF}^{1/(2\times \text{Df})}$ encontram-se todos abaixo do limiar de 5, sendo a maioria próxima de 1, o que indica uma fraca correlação entre as variáveis independentes. A variável *temp* apresenta o valor mais elevado (2.68), ainda assim dentro de limites aceitáveis. As variáveis categóricas com mais níveis, como *season* e *mnth*, têm GVIFs elevados, mas os valores ajustados continuam abaixo de 2.5, o que não levanta preocupações. Assim, conclui-se que não existem indícios relevantes de multicolinearidade no modelo ajustado.

## 4.4 Modelo Final Selecionado

Após a aplicação do método de *forward selection*, foi ajustado um modelo de regressão binomial negativa para explicar a variável resposta *cnt*, correspondente ao número diário de alugueres de bicicletas. A avaliação dos coeficientes estimados revelou que a variável *atemp* não era estatisticamente significativa ($p$-valor = 0,110) e apresentava elevada multicolinearidade com outras covariáveis (GVIF corrigido = 7,76).

Consequentemente, procedeu-se à sua remoção, sendo o modelo reestimado. O modelo final inclui as seguintes variáveis explicativas: *yr*, *season*, *weathersit*, *mnth*, *holiday*, *windspeed*, *hum*, *weekday* e *temp*.

A equação do modelo é dada por:

$$\log(\textit{cnt}) = \beta_0 + \beta_1 \cdot \textit{yr} + \beta_2 \cdot \textit{season} + \beta_3 \cdot \textit{weathersit} + \beta_4 \cdot \textit{mnth} + \beta_5 \cdot \textit{windspeed} + \beta_6 \cdot \textit{hum} + \beta_7 \cdot \textit{holiday} + \beta_8 \cdot \textit{temp} + \beta_9 \cdot \textit{workingday}$$

Este modelo apresentou os seguintes critérios de qualidade de ajustamento:

- AIC = 12155.57
- BIC = 12265.83

Observa-se que o AIC sofreu um aumento insignificante (de 12155,03 para 12155,57), indicando uma perda mínima na qualidade do ajustamento. Por outro lado, o BIC apresentou uma ligeira redução (de 12269,89 para 12265,83), o que sugere uma melhoria em termos de parcimônia do modelo.

Considerando que a variável *atemp* apresentava elevada multicolinearidade ($\text{GVIF}^{1/(2\cdot df)} = 8{,}33$) e ausência de significância estatística no modelo original, optou-se pela sua exclusão. A pequena variação nos critérios de informação, aliada à melhoria na estabilidade e interpretabilidade do modelo, justifica plenamente a adoção do modelo final sem *atemp*.

A Tabela 4.6 apresenta os coeficientes estimados e respectivas estatísticas.

Tabela 4.6: Coeficientes estimados do modelo binomial negativo final

| **Variável** | **Estimativa** | **Erro padrão** | **Estat. $z$** | **$p$-valor** |
|---|---|---|---|---|
| (Intercept) | 7.39916 | 0.07023 | 105.355 | $<$ 2e-16 |
| yr1 | 0.47793 | 0.01797 | 26.600 | $<$ 2e-16 |
| season2 | 0.26981 | 0.05540 | 4.870 | 1.12e-06 |
| season3 | 0.30374 | 0.06577 | 4.619 | 3.86e-06 |
| season4 | 0.51343 | 0.05590 | 9.185 | $<$ 2e-16 |
| weathersit2 | -0.09167 | 0.02371 | -3.867 | 0.00011 |
| weathersit3 | -0.70790 | 0.06051 | -11.698 | $<$ 2e-16 |
| mnth2 | 0.12487 | 0.04444 | 2.810 | 0.00496 |
| mnth3 | 0.20723 | 0.05112 | 4.054 | 5.04e-05 |
| mnth4 | 0.15223 | 0.07641 | 1.992 | 0.04634 |
| mnth5 | 0.18610 | 0.08256 | 2.254 | 0.02418 |
| mnth6 | 0.05675 | 0.08689 | 0.653 | 0.51367 |
| mnth7 | -0.09735 | 0.09656 | -1.008 | 0.31341 |
| mnth8 | -0.02642 | 0.09292 | -0.284 | 0.77613 |
| mnth9 | 0.14502 | 0.08162 | 1.777 | 0.07561 |
| mnth10 | 0.07723 | 0.07459 | 1.035 | 0.30050 |
| mnth11 | 0.04495 | 0.07125 | 0.631 | 0.52810 |
| mnth12 | 0.03312 | 0.05630 | 0.588 | 0.55643 |
| windspeed | -0.76330 | 0.12531 | -6.091 | 1.12e-09 |
| hum | -0.41339 | 0.08963 | -4.612 | 3.98e-06 |
| holiday1 | -0.17325 | 0.05507 | -3.146 | 0.00166 |
| temp | 1.53685 | 0.12689 | 12.112 | $<$ 2e-16 |
| workingday1 | 0.04158 | 0.01970 | 2.111 | 0.03475 |

A análise dos coeficientes estimados revela que a maioria das covariáveis incluídas no modelo apresenta efeitos estatisticamente significativos sobre o número de alugueres diários de bicicletas. Embora algumas variáveis associadas aos meses (*mnth6* a *mnth12*) não tenham alcançado significância estatística individual, a sua inclusão contribui para capturar variações sazonais que podem ser relevantes em termos de tendência global do modelo.

## 4.5 Análise dos coeficientes

Com o objetivo de facilitar a interpretação dos coeficientes estimados, a Tabela 4.7 apresenta os fatores de variação esperada obtidos pela transformação exponencial dos coeficientes do modelo binomial negativo, bem como os respetivos intervalos de confiança a 95%.

Tabela 4.7: *exp*($\beta$) e respetivos intervalos de confiança a 95% para o modelo binomial negativo final.

| **Variável** | **exp**($\beta$) | **IC 95% (inf)** | **IC 95% (sup)** |
|---|---|---|---|
| (Intercept) | 1634,61 | 1419,10 | 1884,07 |
| yr1 | 1,61 | 1,56 | 1,67 |
| season2 | 1,31 | 1,17 | 1,46 |
| season3 | 1,35 | 1,19 | 1,55 |
| season4 | 1,67 | 1,49 | 1,87 |
| weathersit2 | 0,91 | 0,87 | 0,96 |
| weathersit3 | 0,49 | 0,44 | 0,56 |
| mnth2 | 1,13 | 1,04 | 1,24 |
| mnth3 | 1,23 | 1,11 | 1,36 |
| mnth4 | 1,16 | 1,00 | 1,36 |
| mnth5 | 1,20 | 1,02 | 1,42 |
| mnth6 | 1,06 | 0,89 | 1,26 |
| mnth7 | 0,91 | 0,75 | 1,10 |
| mnth8 | 0,97 | 0,81 | 1,17 |
| mnth9 | 1,16 | 0,98 | 1,36 |
| mnth10 | 1,08 | 0,93 | 1,25 |
| mnth11 | 1,05 | 0,91 | 1,21 |
| mnth12 | 1,03 | 0,92 | 1,16 |
| windspeed | 0,47 | 0,36 | 0,60 |
| hum | 0,66 | 0,55 | 0,79 |
| holiday1 | 0,84 | 0,76 | 0,94 |
| temp | 4,65 | 3,61 | 5,98 |
| workingday1 | 1,04 | 1,00 | 1,08 |

A interpretação dos coeficientes é realizada com base na sua exponenciação, representando o fator multiplicativo no número esperado de alugueres diários de bicicletas (*cnt*). Assim, temos que:

- **Intercept:** $e^{7,40} \approx 1634,6$ representa o número esperado de alugueres quando todas as covariáveis são nulas. Neste caso específico não tem interesse prático.
- **yr1:** $\exp(\beta) = 1,61$ indica que no segundo ano de operação houve, em média, um aumento de 61% nos alugueres diários em comparação com o primeiro ano. Este resultado é estatisticamente significativo ($p < 0,001$).
- **season:** Todas as estações, em relação ao inverno (categoria de referência), apresentam um efeito positivo significativo. Destaca-se:
  - Primavera (season2): aumento de 31%.
  - Verão (season3): aumento de 35%.
  - Outono (season4): aumento de 67%.

- **weathersit:**
  - weathersit2 (nevoeiro ou nuvens dispersas): $\exp(\beta) = 0{,}91$ indica uma redução de 9% nos alugueres em relação a dias limpos.
  - weathersit3 (chuva intensa ou neve): $\exp(\beta) = 0{,}49$ implica uma forte redução de 51%, estatisticamente significativa ($p < 0{,}001$).
- **mnth:** Algumas dummies mensais revelam efeitos significativos:
  - Fevereiro a Maio (mnth2–mnth5): aumentos entre 13% e 23%.
  - Setembro (mnth9): aumento de 16%.
  - Meses de Junho a Dezembro não apresentam significância estatística.
- **holiday1:** Dias feriados estão associados a uma redução média de 16% no número de alugueres ($p = 0{,}0017$).
- **workingday1:** Dias úteis apresentam um ligeiro aumento médio de 4% nos alugueres, com significância estatística ($p = 0{,}0347$).
- **windspeed:** $\exp(\beta) = 0{,}47$ indica que o aumento da velocidade do vento está associado a uma redução média de 53% nos alugueres ($p < 0{,}001$).
- **hum:** Aumentos na humidade relativa estão associados a uma redução de aproximadamente 34%, efeito significativo e consistente com o comportamento esperado dos utilizadores.
- **temp:** A variável com maior impacto. $\exp(\beta) = 4{,}65$ indica que um aumento unitário na temperatura (normalizada) está associado a um aumento médio de 365% nos alugueres. É altamente significativa ($p < 0{,}001$).

É de destacar que a variável *temp* tem o maior efeito positivo, com $\exp(\beta) = 4{,}65$, o que indica que, para cada unidade adicional na temperatura normalizada, o número esperado de alugueres diários aumenta, em média, cerca de 365%. Também o ano (*yr1*) tem um impacto positivo relevante, com um crescimento de cerca de 61% no segundo ano relativamente ao primeiro, evidenciando o aumento da popularidade do sistema ao longo do tempo.

Em contraste, variáveis como *weathersit3* (chuva intensa ou neve) e *windspeed* apresentam efeitos significativamente negativos, com reduções de 51% e 53%, respetivamente, no número esperado de alugueres. O efeito de *holiday1* (feriados) também é negativo, sugerindo uma diminuição de aproximadamente 16% nos alugueres.

A maior parte das categorias mensais não revela efeitos estatisticamente significativos. De modo geral, os resultados estão alinhados com a expectativa de que fatores meteorológicos, sazonais e contextuais influenciam de forma substancial o uso do sistema de bicicletas. O modelo ajustado revela-se útil para previsão da procura e apoio à gestão de sistemas de bicicletas partilhadas.

## 4.6 Análise de Diagnóstico do Modelo Final

A análise de diagnóstico permite avaliar a adequação do modelo binomial negativo ajustado à realidade dos dados observados. Para tal, foram utilizadas métricas estatísticas formais e ferramentas gráficas com o intuito de detetar eventuais falhas de especificação ou anomalias no ajustamento.

A *deviance residual* obtida para o modelo ajustado foi de 747,6, com 731 observações e 22 parâmetros estimados. Isto corresponde a 708 graus de liberdade residuais. A comparação deste valor com a distribuição qui-quadrado revelou um valor-p de aproximadamente 0,15. Este valor não é suficientemente pequeno para rejeitar a hipótese nula de que o modelo se ajusta bem aos dados, indicando, assim, uma discrepância aceitável entre os valores observados e os valores esperados. Este resultado sugere que, do ponto de vista global, o modelo fornece um ajustamento estatisticamente adequado.

Adicionalmente, foi avaliada a estatística de Pearson generalizada, cuja soma dos resíduos de Pearson ao quadrado resultou num valor associado a um valor-p de 0,9999. Este resultado reforça a conclusão anterior, evidenciando uma excelente concordância entre os dados e o modelo ajustado, do ponto de vista da variância dos resíduos.

No entanto, a análise revelou que o modelo apenas explica cerca de 1,84% da deviance total. Embora o teste de razão de verosimilhanças permita rejeitar o modelo nulo em favor do modelo ajustado, esta baixa percentagem de deviance explicada sugere que o modelo pode não captar devidamente toda a estrutura subjacente aos dados. Esta limitação pode dever-se à ausência de termos de interação ou não linearidades que não foram considerados no modelo atual.

Para complementar a análise formal, foram construídos gráficos de diagnóstico com base nos resíduos.

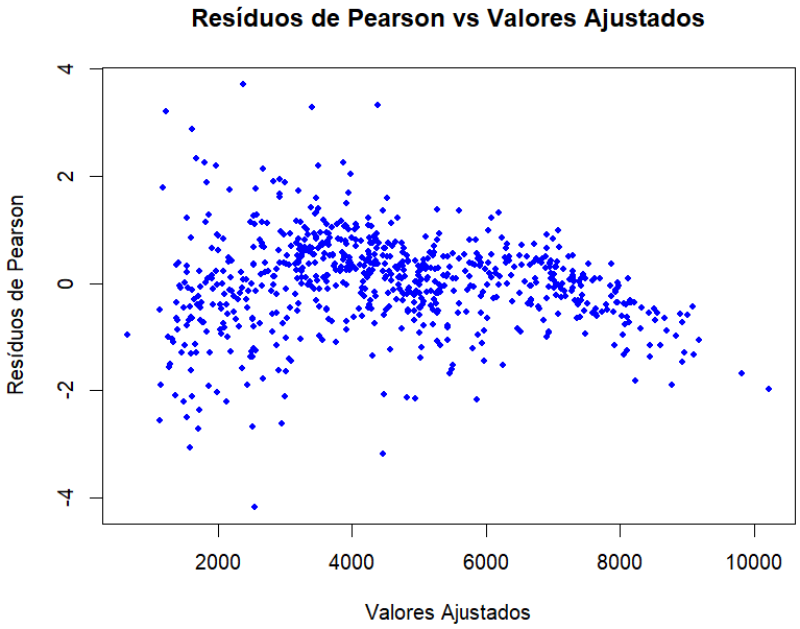


Figura 4.2: Resíduos de Pearson em função dos valores ajustados

O gráfico dos resíduos de Pearson em função dos valores ajustados revela um padrão decrescente. Em vez de uma dispersão aleatória em torno de zero, observa-se uma tendência sistemática, o que sugere a presença de heterocedasticidade ou alguma má especificação do modelo. Este padrão pode indicar que a relação entre as covariáveis e a variável resposta não é inteiramente linear, ou que alguns efeitos importantes não estão devidamente modelados. Apesar disso, a maioria dos resíduos permanece dentro do intervalo entre -2 e 2, o que é geralmente aceitável.

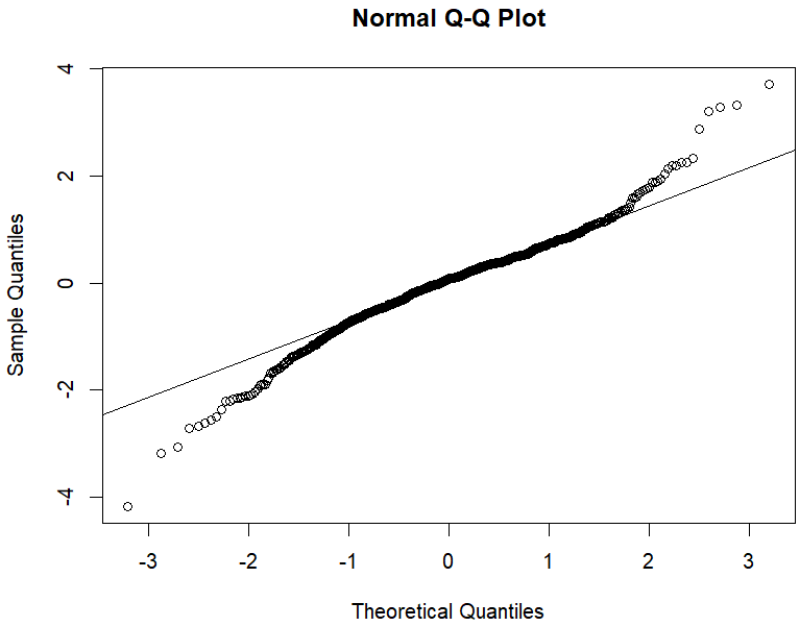


Figura 4.3: Gráfico Q-Q normal dos resíduos de Pearson

O gráfico Q-Q dos resíduos mostra um alinhamento satisfatório ao longo da linha teórica na zona central, embora existam desvios nas caudas. Estes desvios indicam a presença de valores extremos, o que é expectável em modelos de contagem com dispersão, como o binomial negativo. Apesar destas discrepâncias, o comportamento geral dos resíduos sugere que a aproximação assintótica à normalidade é razoável, e as inferências baseadas no modelo mantêm a sua validade.

Em suma, pode concluir-se que o modelo ajustado apresenta um bom ajustamento global aos dados, tanto com base em critérios estatísticos formais como em inspeção gráfica. Contudo, a baixa percentagem de deviance explicada e os padrões detetados nos resíduos apontam para a possibilidade de melhorias, nomeadamente através da introdução de transformações, termos de interação ou outras formas funcionais mais flexíveis no modelo.

# 5 Conclusão

Ao longo deste trabalho, foi realizada uma análise exploratória abrangente com o objetivo de compreender o comportamento das variáveis presentes no conjunto de dados e identificar padrões relevantes na procura por bicicletas. Foram analisadas tanto variáveis quantitativas, com recurso a estatísticas descritivas e matrizes de correlação, como variáveis qualitativas, através de tabelas de frequência e representações gráficas. Posteriormente, recorreu-se à aplicação de testes estatísticos para avaliar, de forma formal, a influência de diversas variáveis na variável resposta *cnt*. O teste de Kruskal–Wallis foi utilizado para variáveis categóricas, revelando que fatores como estação do ano, mês, situação climática e ano afetam significativamente a distribuição do número de aderência às bicicletas.

Foi desenvolvido um modelo estatístico para explicar o número de alugueres diários de bicicletas com base em diversas variáveis meteorológicas, sazonais e contextuais. Inicialmente, considerou-se a utilização da regressão de Poisson, modelo clássico para dados de contagem. No entanto, verificou-se que a variância da variável resposta excedia a média, caracterizando um caso claro de sobredispersão. Esta situação foi confirmada através de testes estatísticos específicos.

Dada a presença de sobredispersão, optou-se por modelar os dados com recurso à regressão binomial negativa, que permite acomodar variância superior à média. Para a seleção das variáveis explicativas mais relevantes, aplicaram-se três métodos automáticos: *forward*, *backward* e *stepwise selection*, com base nos critérios de AIC e BIC. O modelo selecionado por *forward selection* foi inicialmente considerado o mais adequado, apresentando um bom equilíbrio entre qualidade de ajustamento e parcimônia.

Contudo, a análise de colinearidade (através do VIF) revelou que as variáveis *temp* e *atemp* estavam fortemente correlacionadas, sendo que *atemp* apresentava um elevado VIF e não era estatisticamente significativa. Por esta razão, decidiu-se ajustar um novo modelo, eliminando *atemp*. A versão final do modelo, estimada sobre o conjunto completo de dados, demonstrou um ajustamento igualmente robusto, com menor colinearidade, e apresentou valores competitivos de AIC e BIC.

A adequação do modelo final foi avaliada através de uma análise de diagnóstico, que incluiu a observação de resíduos e a aplicação de testes formais. Os resultados indicaram que o modelo apresenta um ajustamento global adequado, não tendo sido detetadas discrepâncias estatisticamente significativas entre os valores observados e os valores esperados. A estatística de Pearson evidenciou uma forte concordância entre os dados e o modelo ajustado. Apesar da proporção da deviance explicada ser relativamente modesta, os resíduos mostraram-se globalmente bem comportados. Foi detetado um ligeiro padrão nos resíduos de Pearson, sugerindo possível heterocedasticidade ou lacunas na especificação do modelo, mas o gráfico Q-Q dos resíduos revelou uma aproximação razoável à normalidade na zona central. Estas observações, embora recomendem alguma cautela, não comprometem a validade inferencial nem a utilidade prática do modelo.

Os coeficientes do modelo final foram interpretados através da sua transformação exponencial, fornecendo uma leitura direta sobre o efeito multiplicativo de cada covariável na contagem esperada de alugueres. Conclui-se que fatores como a temperatura, o ano, as estações mais amenas e os dias úteis têm efeitos positivos significativos sobre o número de alugueres, enquanto variáveis como vento forte, humidade elevada e condições meteorológicas adversas reduzem substancialmente a procura.

De forma geral, o modelo binomial negativo ajustado mostrou-se estatisticamente sólido, interpretável e adequado aos dados, revelando-se uma ferramenta eficaz para entender e prever padrões de utilização do sistema de bicicletas alugadas.

# 6 Referências


1. **UCI Machine Learning Repository**. *Bike Sharing Dataset*, 2011. Disponível em: `https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset`. Acedido em maio de 2025.

2. **Faria, Susana**. *Slides das Aulas de Modelos de Regressão Binomial Negativa*. Universidade do Minho, 2025. Material de apoio da unidade curricular.

3. **Braga, Maria João**. *Modelos Lineares Generalizados Aplicados a Dados de Contagem*. Dissertação de Mestrado, Universidade do Minho, 2014. Disponível em: `https://hdl.handle.net/1822/29402`.

# A Anexo A: Código R

```r
##An lise Explorat ria
library(ggplot2)
library(readr)
library(dplyr)
library(ggcorrplot)
library(reshape2)
dados <- basededados
str(dados)
dados <- na.omit(dados)
sum(is.na(dados))
attach(dados)
#transformar os normalizados em reais para a analise univariada (summary)
desnormalizar_temp <- function(temp_norm) {
  tmin <- -8
  tmax <- 39
  temp_norm * (tmax - tmin) + tmin
}

desnormalizar_atemp <- function(atemp_norm) {
  atemp_norm * 66 - 16
}

desnormalizar_hum <- function(hum_norm) {
  hum_norm * 100
}

desnormalizar_windspeed <- function(wind_norm) {
  wind_norm * 67
}

dados$temp_real       <- desnormalizar_temp(dados$temp)
dados$atemp_real      <- desnormalizar_atemp(dados$atemp)
dados$hum_real        <- desnormalizar_hum(dados$hum)
dados$windspeed_real  <- desnormalizar_windspeed(dados$windspeed)

summary(dados[, c("temp_real", "atemp_real", "hum_real", "windspeed_real")
    ])
table(season)
table(weathersit)

#Analise Univariada
year_freq <- table(dados$yr)
year_labels <- c("2011", "2012")
year_pct <- round(100 * year_freq / sum(year_freq), 1)
labels <- paste0(year_labels, ": ", year_pct, "%")
cores <- c("#B3CDE3", "#FBB4AE")
pie(year_freq,
    labels = labels,
    col = cores,
    main = "Distribui  o das Observa  es por Ano")


mnthfator <- factor(dados$mnth, levels = 1:12)
frequencias <- as.data.frame(table(mnthfator))
colnames(frequencias) <- c("mnth", "freq")

```

```r
ggplot(frequencias, aes(x = mnth, y = freq)) +
  geom_bar(stat = "identity", fill = "#DECBE4", color = "white") +
  geom_text(aes(label = freq), vjust = -0.3, size = 4) +
  labs(
       x = "M s", y = "Frequ ncia") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold")
  )

holiday_freq <- table(dados$holiday)
holiday_labels <- c("Dia normal", "Feriado")
holiday_pct <- round(100 * holiday_freq / sum(holiday_freq), 1)
labels <- paste0(holiday_labels, ": ", holiday_pct, "%")
cores <- c("#AED6F1", "#F9E79F")
pie(holiday_freq,
    labels = labels,
    col = cores)




dados_week <- factor(dados$weekday, levels = 0:6,
                         labels = c("Dom", "Seg", "Ter", "Qua", "Qui", "Sex"
                            , "Sab"))
frequencias_weekday <- as.data.frame(table(dados_week))
colnames(frequencias_weekday) <- c("weekday", "freq")
ggplot(frequencias_weekday, aes(x = weekday, y = freq)) +
  geom_bar(stat = "identity", fill = "#D8BFAA", color = "white") +
  geom_text(aes(label = freq), vjust = -0.3, size = 4) +
  labs(x = "Dia da Semana", y = "Frequencia") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold")
  )



workingday_freq <- table(dados$workingday)
workingday_labels <- c("Nao  til ", "Dia util")
workingday_pct <- round(100 * workingday_freq / sum(workingday_freq), 1)
labels <- paste0(workingday_labels, ": ", workingday_pct, "%")
cores_working <- c("#F4A6A6", "#FADBD8")
pie(workingday_freq,
    labels = labels,
    col = cores_working,
    cex = 0.7)

table(dados$weathersit)


#Analise Bivariada
shapiro.test(dados$temp)
shapiro.test(dados$atemp)
shapiro.test(dados$hum)
shapiro.test(dados$windspeed)
shapiro.test(dados$cnt)
variaveis_quant <- dados[, c("temp", "atemp", "hum", "windspeed", "cnt")]
```

```r
mat_cor <- round(cor(vars, method = "spearman"), 2)
cor_data <- melt(mat_cor)
ggplot(cor_data, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = value), size = 4.5) +
  scale_fill_gradient2(low = "#D4A373", mid = "white", high = "lightblue",
      midpoint = 0, limit = c(-1, 1)) +
  labs(title = "Matriz de Correla  o com cnt", x = "", y = "") +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )

ggplot(dados, aes(x = factor(season), y = cnt)) +
  geom_boxplot(fill = "#AED6F1", color = "black") +
    x = "Esta  o do Ano", y = "Alugueres Di rios"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))


ggplot(dados, aes(x = factor(workingday), y = cnt)) +
  stat_summary(fun = mean, geom = "bar", fill = "#F5B041") +
  labs(
    title = "M dia de Alugueres por Tipo de Dia",
    x = "Dia  til  (0 = n o, 1 = sim)"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

ggplot(dados, aes(x = factor(weathersit), y = cnt)) +
  geom_violin(fill = "#D2B48C", color = "black") +
  stat_summary(fun = median, geom = "crossbar",
               width = 0.5,
               size = 3,
               color = "black",
               fatten = 0) +
  labs(
    x = "Situa  o Climtica (1 = c u limpo, 2 = nevoeiro, 3 = chuva)",
    y = "Alugueres Di rios"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))



#Testes estat sticos
basededados <- subset(basededados, select = -c(instant, dteday, casual,
    registered))
str(basededados)
summary(base)
basededados$season     <- factor(basededados$season)
basededados$mnth       <- factor(basededados$mnth)
basededados$weekday    <- factor(basededados$weekday)
basededados$weathersit <- factor(basededados$weathersit)
basededados$yr         <- factor(basededados$yr)
basededados$holiday    <- factor(basededados$holiday)
```

```r
basededados$workingday <- factor(basededados$workingday)
modelo_binom_neg <- glm.nb(cnt ~ atemp + yr + season + weathersit + mnth +
    windspeed + hum + holiday + temp + workingday,data=basededados)
modelo_temp_holiday <- glm.nb(cnt ~ basededados$temp + basededados$holiday,
                              data = basededados)
summary(modelo_temp_holiday)

kruskal.test(cnt ~ basededados$season, data = basededados)
kruskal.test(cnt ~ basededados$mnth, data = basededados)
kruskal.test(cnt ~ basededados$weekday, data = basededados)
kruskal.test(cnt ~ basededados$weathersit, data = basededados)
kruskal.test(cnt ~ basededados$yr, data = basededados)
kruskal.test(cnt ~ basededados$holiday, data = basededados)
kruskal.test(cnt ~ basededados$workingday, data = basededados)
#An lise do Modelo
day <- subset(day, select = -c(instant, dteday, casual, registered))
str(day)
summary(day)
# Transformar vari veis categ ricas (sem adicionar labels)
day$season     <- factor(day$season)
day$mnth       <- factor(day$mnth)
day$weekday    <- factor(day$weekday)
day$weathersit <- factor(day$weathersit)
# Manter vari veis bin rias como 0 e 1
day$yr         <- factor(day$yr)
day$holiday    <- factor(day$holiday)
day$workingday <- factor(day$workingday)
str(day)
summary(day)

attach(day)
mean(day$cnt)      # m dia da vari vel de contagem
var(day$cnt)       # vari ncia da vari vel de contagem

 #sobredispers o ELEVADA-> BINOMIAL NEGATIVA


#confirmar:
modelopoisson<-glm(cnt~., family=poisson, data = day)

library(performance)
check_overdispersion(modelopoisson) #confirma-se

#######
library(DHARMa)

# Simular os res duos (sem plot)
sim <- simulateResiduals(fittedModel = modelopoisson, plot = FALSE)

# Obter os res duos simulados
res <- recalculateResiduals(sim)

# Criar QQ plot manualmente (sem testes)
qqplot(x = sort(runif(length(res$scaledResiduals))),
       y = sort(res$scaledResiduals),
       xlab = "Expected", ylab = "Observed",
       main = "QQ plot residuals", pch = 17)
abline(0, 1, col = "red")
```

```r


#############

library(MASS)

# Modelo completo com todas as vari veis
modelo_completo <- glm.nb(cnt ~ ., data = day)



# Ajustar o modelo nulo
modelo_nulo <- glm.nb(cnt ~ 1, data = day)

# Definir escopo do modelo completo com todas as vari veis de 'day'
completo <- formula(glm.nb(cnt ~ ., data = day))

# Aplicar  s e l e    o  forward
modelo_forward <- step(modelo_nulo,
                       scope = completo,
                       direction = "forward")


# BACKWARD ELIMINATION
modelo_backward <- step(modelo_completo, direction = "backward")

# STEPWISE (both directions)
modelo_stepwise <- step(modelo_completo, direction = "both")

# Ver os resumos dos modelos finais
summary(modelo_forward)
summary(modelo_backward)
summary(modelo_stepwise)

#  Ver as f rmulas finais selecionadas
formula(modelo_forward)
formula(modelo_backward)
formula(modelo_stepwise)


# Obter AIC e BIC dos t r s modelos
aic_forward   <- AIC(modelo_forward)
bic_forward   <- BIC(modelo_forward)

aic_backward  <- AIC(modelo_backward)
bic_backward  <- BIC(modelo_backward)

aic_stepwise  <- AIC(modelo_stepwise)
bic_stepwise  <- BIC(modelo_stepwise)

# Criar data frame com os resultados
tabela_resultados <- data.frame(
  Modelo = c("Forward", "Backward", "Stepwise"),
  AIC    = c(aic_forward, aic_backward, aic_stepwise),
  BIC    = c(bic_forward, bic_backward, bic_stepwise)
)
```

```r

# Visualizar a tabela
print(tabela_resultados)

#Modelo      AIC      BIC
#1  Forward 12155.03 12269.89
#2 Backward 12155.23 12293.06
#3 Stepwise 12155.23 12293.06

formula(modelo_forward)
#cnt ~ atemp + yr + season + weathersit + mnth + windspeed + hum + holiday
    + temp + workingday









# Teste de raz o de verossimilhan a

library(lmtest)
lrtest(modelo_stepwise, modelo_completo)

lrtest(modelo_forward, modelo_completo) #-----> o teste valida a
    utiliza  o deste modelo

lrtest(modelo_backward, modelo_completo)



##########
vif(modelo_forward)

formula(modelo_forward)

# Reajustar o modelo removendo 'atemp'
modelo_ajustado <- glm.nb(cnt ~ yr + season + weathersit + mnth + windspeed
     + hum +
                              holiday + temp + workingday, data = day)

# Ver resumo do novo modelo
summary(modelo_ajustado)

# Calcular os novos VIFs
library(car)
vif(modelo_ajustado)

summary(modelo_ajustado)
AIC(modelo_ajustado)
BIC(modelo_ajustado)


formula(modelo_ajustado)

# Exponencial dos coeficientes
exp(coef(modelo_ajustado))

# Intervalos de confian a (95%) dos coeficientes
exp(confint(modelo_ajustado))
```

```r


##############

set.seed(3333)


# Dividir o dataset day em treino (80%) e teste (20%)

indices <- sample(seq_len(nrow(day)), size = 0.8 * nrow(day))
treino <- day[indices, ]
teste  <- day[-indices, ]

# -------------------------
# FORWARD SELECTION
# -------------------------

# Modelo nulo
modelo_nulo <- glm.nb(cnt ~ 1, data = treino)

# Modelo completo (f rmula apenas para escopo)
escopo <- formula(cnt ~ .)

# Aplicar forward
modelo_forward <- step(modelo_nulo,
                       scope = escopo,
                       direction = "forward",trace = 1)

# Previs es e RMSE
pred_forward <- predict(modelo_forward, newdata = teste, type = "response")
rmse_forward <- sqrt(mean((teste$cnt - pred_forward)^2))


# -------------------------
# BACKWARD SELECTION
# -------------------------

# Ajustar modelo backward com os dados de treino
modelo_backward <- step(glm.nb(cnt ~ ., data = treino), direction = "
    backward", trace = 1)

# Previs es e RMSE para backward
pred_backward <- predict(modelo_backward, newdata = teste, type = "response
    ")
rmse_backward <- sqrt(mean((teste$cnt - pred_backward)^2))


# -------------------------
# STEPWISE SELECTION
# -------------------------

# Ajustar modelo stepwise com os dados de treino
modelo_stepwise <- step(glm.nb(cnt ~ ., data = treino), direction = "both",
     trace = 1)

# Previs es e RMSE para stepwise
```

```r
pred_stepwise <- predict(modelo_stepwise, newdata = teste, type = "response")
rmse_stepwise <- sqrt(mean((teste$cnt - pred_stepwise)^2))


# -------------------------
# Tabela comparativa dos RMSEs
# -------------------------

tabela_rmse <- data.frame(
  Modelo = c("Forward", "Backward", "Stepwise"),
  RMSE   = c(rmse_forward, rmse_backward, rmse_stepwise)
)

print(tabela_rmse)

#Analise Diagnostico
summary(modelo_ajustado)


#O modelo ajusta-se aos dados?

731-23#n-p-1
deviance(modelo_ajustado)#deviance est  proximo do modelo ajustado, o que um bom inidicador de qualidade de ajustamento do modelo

1-pchisq(deviance(modelo_ajustado), df.residual(modelo_ajustado)) #p_value = 0.15, nao rejeitamos H0, o modelo ajusta-se bem aos dados => confirma o vusto anteriormente

#Comparar com modelo Nulo
anova(modelo_nulo, modelo_ajustado, test = "Chisq")

#Percentagem de Deviance Explicada
100 * (deviance(modelo_nulo)-deviance(modelo_ajustado))/deviance(modelo_nulo)
#1.84% da deviance explicada

#ETPG
ETPG <- sum(residuals(modelo_ajustado, type="pearson")^2)
1-pchisq(ETPG, 731-22) #p_value = 0.9999, o que indica um bom ajustameto dos dados

plot(
  fitted(modelo_ajustado),
  residuals(modelo_ajustado, type = "pearson"),
  main = "Res duos de Pearson vs Valores Ajustados",
  xlab = "Valores Ajustados",
  ylab = "Res duos de Pearson",
  pch = 20,              # pontos preenchidos
  col = "blue"
)
abline(h = 0, col = "red", lty = 2)  # linha horizontal a 0
#O gr fico dos res duos de Pearson em fun  o dos valores ajustados revela um padr o descendente, sugerindo a presen a de heterocedasticidade ou m  especifica o do modelo. Apesar de a maioria dos res duos estar compreendida entre -2 e 2, o padr o sistem tico indica que o modelo pode beneficiar da inclus o de termos
```

```
    n o lineares ou transforma es adicionais.
#
qqnorm(residuals(modelo_ajustado, type = "pearson"))
qqline(residuals(modelo_ajustado, type = "pearson"))

#O gr fico Q-Q mostra que os res duos seguem razoavelmente uma
    distribui o normal na zona central, mas apresentam desvios nas caudas
    , indicando poss veis outliers ou desvios da normalidade. Este
    comportamento esperado em modelos de contagem com vari ncia n o
    constante, como o binomial negativo. Ainda assim, a aproxima o
    assint tica parece aceit vel para efeitos de infer ncia.
```

Listing 1: Código completo em R para análise dos dados de aluguer de bicicletas.