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

Modelos Lineares Generalizados

Anita Ferreira PG56093
Inês Gomes PG55575
Rui Alves PG55577

2 de junho de 2025

Motivação e Objetivos

- Crescimento do uso de bicicletas alugadas em contexto urbano.
- Analisar os fatores que influenciam o número de alugueres diários de bicicletas.
- Identificar a relação entre variáveis meteorológicas, sazonais e o uso do sistema.
- Ajustar modelos estatísticos adequados para dados de contagem com sobredispersão.

Descrição da Base de Dados

instant	Identificador sequencial
dteday	Data do registo
season	Estação do ano (1 = inverno, ..., 4 = outono)
yr	Ano (0 = 2011, 1 = 2012)
mnth	Mês do ano (1 a 12)
holiday	Feriado (1 = sim, 0=não)
weekday	Dia da semana (0 = dom, ..., 6 = sáb)
workingday	Dia útil (1 = sim, 0=não)
weathersit	Clima (1 = limpo, 2 = nevoeiro, 3 = chuva leve)
temp	Temperatura normalizada
atemp	Sensação térmica normalizada
hum	Humidade normalizada
windspeed	Velocidade do vento normalizada
casual	Alugueres por utilizadores não registados
registered	Alugueres por utilizadores registados
cnt	Total de bicicletas alugadas por dia (*casual* + *registered*)

Análise Exploratória

Análise Univariada – Variáveis Quantitativas

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

Estação	Frequência
Inverno	181
Primavera	184
Verão	188
Outono	178

Variável *season*

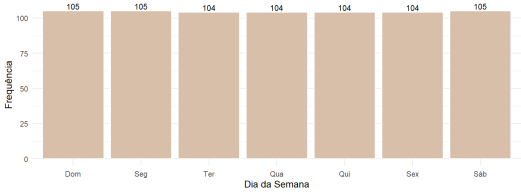
Condição Climática	Frequência
Céu limpo / Poucas nuvens	463
Nevoeiro / Nuvens dispersas	247
Chuva leve / Neve leve	21

Variável *weathersit*

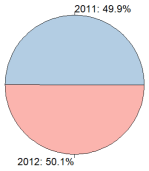
Análise Exploratória

Análise Univariada - Variáveis Qualitativas

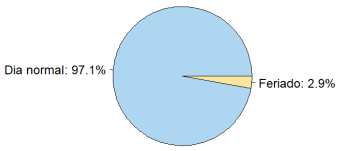
Variável *mnth*



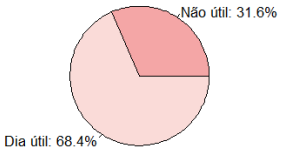
Variável *weekday*



Variável *yr*



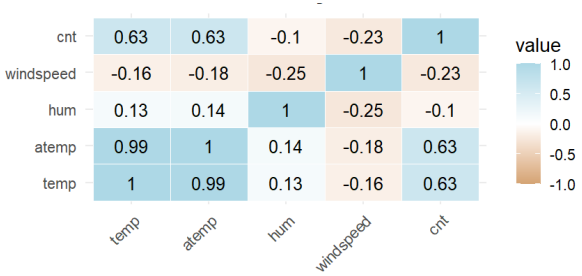
Variável *holiday*



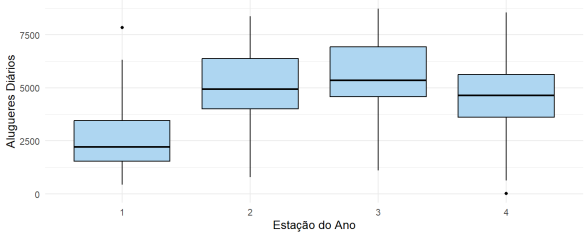
Variável *workingday*

Análise Exploratória

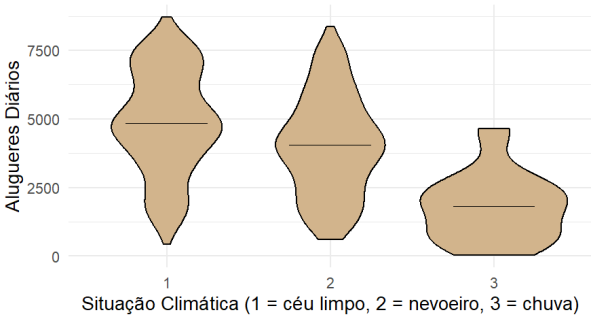
Análise Bivariada



- *temp*: forte associação positiva com o número de alugueres;
- *atemp*: comportamento muito semelhante ao da temperatura;
- *hum*: associação negativa fraca com o número de alugueres;
- *windspeed*: correlação negativa ligeiramente mais acentuada.



- A menor mediana de bicicletas alugadas ocorre no inverno, com valores significativamente mais baixos;
- Primavera e verão registam maior procura, com medianas perto dos 5000;
- No outono, a procura mantém-se elevada, embora ligeiramente inferior ao verão.



- Em dias com céu limpo, a maioria dos alugueres situa-se acima de 5000, com distribuição concentrada em valores altos;
- Em dias de nevoeiro, a distribuição é semelhante mas mais dispersa e com menos valores elevados;
- Em dias de chuva leve, a procura por bicicletas reduz-se drasticamente, com valores de aluguer bastante inferiores.

Testes de Hipóteses

Pergunta 1

As variáveis categóricas influenciam significativamente o número de bicicletas alugadas?

Hipóteses:

- 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

Teste: *Kruskal–Wallis*

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

Pergunta 2

A temperatura e os feriados afetam a procura pelas bicicletas?

Hipóteses:

- $H_0 : \beta_{\text{temp}} = \beta_{\text{holiday}} = 0$
- H_1 : Pelo menos um dos coeficientes é diferente de zero

Teste: *Wald*

Variável	Valor de prova
temp	$< 2 \times 10^{-16}$
holiday	0.0325

Sobredispersão em Modelos de Contagem

- A variância de uma variável de contagem Y pode ser aproximadamente proporcional à sua média:

$$\text{Var}(Y) = \phi\,\mathbb{E}[Y]$$

- Onde ϕ é o **parâmetro de dispersão**, tal que:
  - $\phi = 1$: modelo de Poisson ajustado.
  - $\phi > 1$: sobredispersão.
  - $\phi < 1$: subdispersão.

- **Por que a sobredispersão é um problema?**
  - Leva à **subestimação dos erros padrão** dos coeficientes.
  - Pode fazer com que *variáveis pareçam significativas quando não são*.
  - Compromete a validade de inferências estatísticas (teste z/t, ICs).

Avaliação da Sobredispersão

Foi aplicado o teste `check_overdispersion()` do pacote `performance` no R.

Resultados do teste:

$\hat{\phi} = 152{,}4$
Estatística de Pearson: 106832,42
p-valor $< 0{,}001 \quad \Rightarrow$ **Sobredispersão confirmada**

Para complementar, foi gerado um **QQ plot** dos resíduos simulados do modelo de Poisson. O gráfico revelou um afastamento acentuado da linha teórica, evidenciando **mau ajustamento**.

Q-Q normal dos Resíduos do Modelo de Poisson

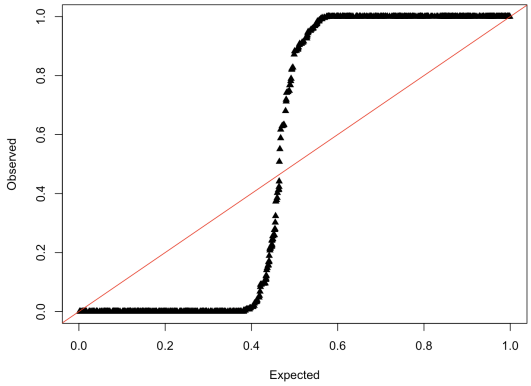


Figura: Gráfico Q-Q normal dos resíduos do modelo de Poisson

- Evidencia visualmente a inadequação do modelo de Poisson para os dados.
- Reflete a presença de **sobredispersão**, já confirmada estatisticamente.

Solução para a sobredispersão:modelo de regressão Binomial Negativa

- Como solução, foi ajustado um **modelo de regressão binomial negativa completo**, que inclui todas as variáveis explicativas disponíveis.
- Para obter um modelo mais parcimonioso e com boa capacidade preditiva, aplicaram-se três métodos automáticos de seleção de variáveis:
  - *Forward Selection*
  - *Backward Elimination*
  - *Stepwise Selection*
- Os modelos obtidos foram comparados com base nos seguintes critérios:

Modelo	AIC	BIC	RMSE
Forward	12155.03	12269.89	1975.45
Backward	12155.23	12293.06	1008.58
Stepwise	12155.23	12293.06	1008.58

Teste da Razão de Verossimilhança (LRT)

- Objetivo: verificar se os modelos reduzidos comprometem o ajustamento face ao modelo completo.
- Execução: `lrtest()` do pacote lmtest no R.
- Resultados:

Modelo Comparado	**Dif. DF**	χ^2	***p*-valor**
Forward vs. Completo	5	9,80	0,081
Backward vs. Completo	0	0	$<$ 2,2e-16
Stepwise vs. Completo	0	0	$<$ 2,2e-16

Conclusão:

- Apenas o modelo **forward** apresentou equivalência estatística ao modelo completo ($p = 0{,}081 > 0{,}05$).
- Os modelos **backward** e **stepwise** mantiveram praticamente a mesma estrutura do modelo completo o que resultou em uma diferença estatisticamente nula.

Seleção do Modelo Final

- O modelo obtido via **forward selection** foi escolhido como ponto de partida.

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

- Verificou-se que a variável `atemp` apresentava:
  - Ausência de significância estatística ($p = 0{,}110$)
  - Elevado fator de inflação da variância ($\text{GVIF}^{1/(2 \cdot df)} = 8{,}33$)
- Decidiu-se, por isso, remover `atemp` e reajustar o modelo.
- O modelo reestimado apresentou:
  - AIC semelhante (12155.57)
  - BIC mais baixo (12265.83)
  - Nenhum problema de multicolinearidade.

Conclusão: O modelo final selecionado foi o **modelo forward sem a variável** `atemp`, por aliar qualidade de ajustamento à parcimônia e estabilidade.

Expressão do Modelo Final

Seja Y_i o número de alugueres no dia i, assumimos:

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

O modelo final ajustado é:

$$\log(cnt) = \beta_0 + \beta_1 \cdot \texttt{yr} + \beta_2 \cdot \texttt{season} + \beta_3 \cdot \texttt{weathersit} + \beta_4 \cdot \texttt{mnth}$$

$$+\beta_5 \cdot \texttt{windspeed} + \beta_6 \cdot \texttt{ hum} + \beta_7 \cdot \texttt{holiday} + \beta_8 \cdot \texttt{temp} + \beta_9 \cdot \texttt{workingday}$$

Este modelo considera o efeito de variáveis sazonais, meteorológicas e contextuais no número esperado de alugueres de bicicletas.

Interpretação dos Coeficientes ($\exp(\beta)$)

- e^{β_j} representa o **fator multiplicativo** no número esperado de alugueres associado a um aumento unitário de x_j, mantendo as outras variáveis constantes.
- É de destacar, por exemplo:

`temp` $\exp(\hat{\beta}_{\texttt{temp}}) = \exp(1.53685) \approx 4{,}65$
$\Rightarrow$ Um aumento unitário na temperatura está associado a um aumento médio de **365%** no número de alugueres.

`yr1` $\exp(\hat{\beta}_{\texttt{yr1}}) = \exp(0.47793) \approx 1{,}61$
$\Rightarrow$ No segundo ano do estudo, o número de alugueres foi em média **61%** superior ao do primeiro ano.

`weathersit3` $\exp(\hat{\beta}_{\texttt{weathersit3}}) = \exp(-0.70790) \approx 0{,}49$
$\Rightarrow$ Em dias com mau tempo, a procura por bicicletas diminui em **51%** face a dias de bom tempo.

Interpretação dos Coeficientes ($\exp(\beta)$)

`season4` $\exp(\hat{\beta}_{\texttt{season4}}) = \exp(0.51343) \approx 1{,}67$
$\Rightarrow$ Durante o inverno, o número de alugueres foi em média **67%** superior em comparação com a primavera.

`windspeed` $\exp(\hat{\beta}_{\texttt{windspeed}}) = \exp(-0.76330) \approx 0{,}47$
$\Rightarrow$ O aumento da intensidade do vento está associado a uma redução de **53%** na procura por bicicletas.

`hum` $\exp(\hat{\beta}_{\texttt{hum}}) = \exp(-0.41339) \approx 0{,}66$
$\Rightarrow$ A humidade elevada está associada a uma diminuição média de **34%** nos alugueres.

`holiday1` $\exp(\hat{\beta}_{\texttt{holiday1}}) = \exp(-0.17325) \approx 0{,}84$
$\Rightarrow$ Em feriados, registou-se uma redução média de **16%** no número de alugueres.

Interpretação dos Coeficientes ($\exp(\beta)$) — Parte II

- As interpretações obtidas são consistentes com o comportamento esperado dos utilizadores:
  - Condições meteorológicas adversas reduzem a procura.
  - Dias úteis e temperaturas mais elevadas favorecem o uso da bicicleta.

As variáveis climáticas e sazonais exercem influência significativa sobre a procura de bicicletas, reforçando a utilidade do modelo como instrumento de apoio à tomada de decisão operacional.

Análise de Diagnóstico: Avaliação Global do Modelo

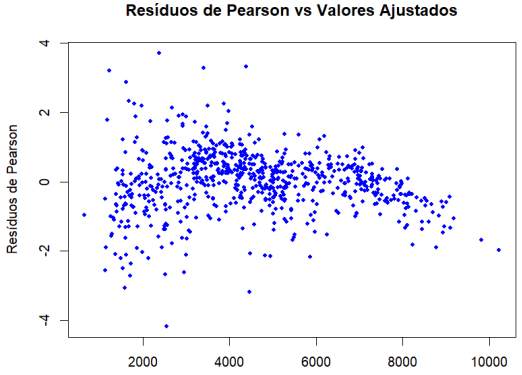
- A qualidade de ajustamento foi avaliada através da **deviance residual**, cujo valor foi de **747,6**, com **708 graus de liberdade**, resultando num p-valor de aproximadamente **0,15**.
- A **estatística de Pearson generalizada** revelou um p-valor muito próximo de **1**, indicando uma forte concordância entre os valores observados e os ajustados.
- Estes resultados sugerem que o modelo apresenta um **bom ajustamento global**, não sendo rejeitada a hipótese de adequação aos dados.
- A **percentagem de deviance explicada** foi de aproximadamente **1,84%**, calculada segundo:

$$\text{Dev. Explicada} = \frac{D_{\text{nulo}} - D_{\text{modelo}}}{D_{\text{nulo}}} \times 100$$

indicando que parte da variabilidade permanece por explicar.

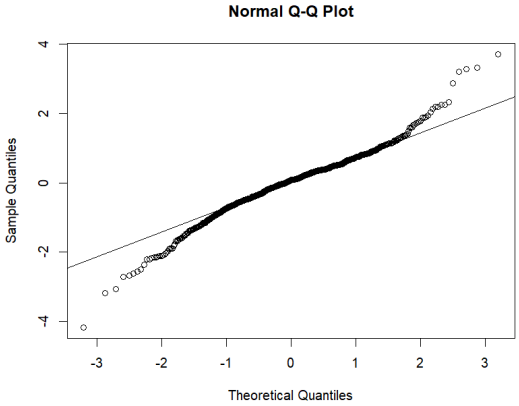
Análise Gráfica dos Resíduos

- O gráfico dos **resíduos de Pearson vs. valores ajustados** revelou uma tendência descendente.
- A maior parte dos resíduos encontra-se dentro do intervalo aceitável, embora o padrão sistemático indique que o modelo poderia beneficiar da inclusão de **termos não lineares** ou **interações**.



Avaliação da Normalidade dos Resíduos

- O **gráfico Q-Q dos resíduos de Pearson** evidenciou uma boa aproximação à normalidade na região central da distribuição.
- No entanto, foram observados desvios nas caudas, compatíveis com a natureza dos dados de contagem com sobredispersão.



Conclusão.