Análise de Regressão

Trabalho Prático:
Sono em Mamíferos

Componentes:

Carolina Penido Barcellos

Gabrielly Xavier dos Santos

Matheus Soares dos Santos de Freitas

Novembro 2025

Sumário

1 Introdução

O sono é um processo biológico essencial e apresenta grande variação entre os mamíferos, tanto na quantidade total dormida quanto nas fases REM e não REM. Essas diferenças estão relacionadas a fatores fisiológicos, ecológicos e comportamentais das espécies. Nesse contexto, a análise estatística permite investigar como variáveis como massa corporal, tempo de gestação, longevidade e pressão de predação influenciam o tempo total de sono. Modelos de regressão linear, por exemplo, ajudam a quantificar esses efeitos, embora sujeitos a limitações como amostras pequenas e possível multicolinearidade.

Este relatório, portanto, tem como objetivo analisar os determinantes do sono em mamíferos por meio de exploração dos dados, ajuste de modelos, verificação de pressupostos e validação cruzada. A abordagem busca compreender como características biológicas e ecológicas se relacionam ao tempo total de sono.

1.1 Estrutura do Relatório

Este documento está organizado da seguinte forma:

- Na Seção 2, detalhamos as fontes de dados e a metodologia utilizada no trabalho.
- Na Seção 3, discutimos os principais resultados empíricos obtidos.
- Na Seção 4, interpretamos esses resultados e suas possíveis causas.
- Na Seção 5, sintetizamos as conclusões do relatório.

2 Dados e Metodologia

A metodologia adotada neste estudo foi estruturada em etapas que combinaram análise exploratória, ajuste e diagnóstico de modelos, seleção de variáveis, regularização e validação cruzada. A seguir, apresentam-se as subseções correspondentes.

2.1 Base de Dados

Foram utilizados dados provenientes da base *Sleep in Mammals* (Kaggle). O dataset contempla 39 espécies de animais distribuídas em 13 ordens. No total, constituem-se 62 observações das seguintes variáveis:

Variável Resposta

- **total_sleep**: Horas totais de sono por dia.

Variáveis Explicativas

- **species** – espécie do mamífero.
- **body_wt** – peso corporal em kg.
- **brain_wt** – peso do cérebro em kg.
- **life_span** – expectativa de vida em anos.
- **gestation** – duração média da gestação em dias.
- **predation** – índice de probabilidade de ser predado (1 a 5).
- **exposure** – exposição do mamífero ao dormir (1 a 5).
- **danger** – índice geral de perigo (1 a 5).

2.1.1 Padronização e Pré-Processamento

Com o intuito de trabalhar com observações dos dados em sua completude, procedeu-se à inspeção de valores ausentes, seguidos de remoção de casos inviáveis.

2.2 Análise Exploratória

Realizou-se inicialmente uma análise exploratória detalhada, incluindo:

- estatísticas descritivas das variáveis numéricas;
- inspeção das distribuições (histogramas e boxplots);
- matriz de correlação para avaliar relações lineares;
- identificação de possíveis outliers.

Essa etapa permitiu mapear padrões iniciais e orientar a especificação dos modelos de regressão.

2.3 Ajuste do Modelo de Regressão Linear

O primeiro modelo ajustado foi um modelo linear múltiplo contendo todos os preditores relevantes. Variáveis perfeitamente correlacionadas com a resposta foram removidas para evitar violações das condições do modelo.

A avaliação do modelo considerou aspectos fundamentais da regressão, como a significância dos coeficientes, a normalidade dos resíduos, a homocedasticidade e a independência dos erros.

2.4 Diagnóstico das Suposições

2.4.1 Teste de Shapiro–Wilk

A normalidade dos resíduos foi avaliada utilizando o teste de Shapiro–Wilk, cujo objetivo é verificar se uma amostra provém de uma distribuição normal. A estatística do teste é definida por:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \, x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

onde $x_{(i)}$ representa a i-ésima ordem da amostra, $\bar{x}$ é a média amostral e os coeficientes a_i são constantes dependentes de n, calculados a partir das médias e covariâncias dos order statistics da normal padrão.

Sob a hipótese nula H_0 (normalidade dos resíduos), valores de W próximos de 1 indicam forte aderência à distribuição normal. O p-valor associado é obtido por aproximações específicas do teste, e rejeita-se H_0 quando o p-valor é inferior ao nível de significância adotado.

2.4.2 Teste de Breusch–Pagan

A presença de heterocedasticidade foi examinada utilizando o teste de Breusch–Pagan, baseado na estatística:

$$BP = nR^2_{\text{aux}},$$

onde R^2_{aux} é o coeficiente de determinação do modelo auxiliar que regressa os resíduos ao quadrado sobre os preditores originais. Sob a hipótese nula H_0 (homocedasticidade), vale:

$$BP \sim \chi^2_k,$$

com k sendo o número de preditores do modelo auxiliar.

2.4.3 Transformação Box–Cox

Em casos onde a variância dos resíduos indicou heterocedasticidade, avaliou-se a aplicação da transformação Box–Cox:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln(y), & \lambda = 0, \end{cases}$$

buscando estabilizar a variância e melhorar linearidade.

2.5 Seleção de Variáveis

Para reduzir redundâncias e multicolinearidade observadas via VIF, empregou-se:

- seleção Stepwise baseada no critério AIC;
- regularização do tipo LASSO.

2.5.1 Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike (AIC) é amplamente utilizado para comparar modelos e selecionar aquele que oferece o melhor equilíbrio entre qualidade de ajuste e complexidade. O AIC é definido como:

$$\text{AIC} = -2\ln(\hat{L}) + 2k,$$

onde $\hat{L}$ é o valor da verossimilhança maximizada do modelo e k representa o número de parâmetros estimados. Quanto menor o AIC, melhor o compromisso entre ajuste e parcimônia.

A seleção de variáveis baseada em AIC, frequentemente implementada por métodos Stepwise (forward, backward ou both), busca identificar o subconjunto de preditores que minimiza esse critério. Essa abordagem é especialmente útil quando há múltiplas combinações possíveis de variáveis, permitindo uma escolha eficiente sem recorrer à avaliação exaustiva de todos os modelos.

2.5.2 Regularização LASSO

O método LASSO (Least Absolute Shrinkage and Selection Operator) minimiza:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

produzindo coeficientes esparsos e facilitando a seleção automática de variáveis.

2.6 Validação Cruzada

O desempenho preditivo do modelo final foi avaliado por diferentes técnicas, tais como:

- validação cruzada k-fold ($k = 10$);
- validação leave-one-out (LOOCV).

As métricas analisadas foram:

- RMSE (Root Mean Squared Error);

- MAE (Mean Absolute Error);
- R^2 de validação;

Esses procedimentos permitiram comparar diferentes técnicas para avaliação do modelo final.

2.7 Síntese

A combinação dessas etapas forneceu uma abordagem robusta, integrando rigor estatístico, diagnóstico cuidadoso das suposições, técnicas modernas de regularização e avaliação preditiva confiável.

3 Resultados

Estatísticas Descritivas

Foram calculados resumos estatísticos das variáveis para compreender padrões gerais:

Tabela 1: Resumo estatístico das variáveis do conjunto `mammals`.

Variável	**Min**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máx**
body_wt (kg)	0.005	0.600	3.342	198.790	48.203	6654.000
brain_wt (g)	0.14	4.25	17.25	283.13	166.00	5712.00
non_dreaming (h)	2.10	6.25	8.35	8.673	11.00	17.90
dreaming (h)	0.00	0.90	1.80	1.972	2.55	6.60
total_sleep (h)	2.60	8.05	10.45	10.53	13.20	19.90
life_span (anos)	2.00	6.63	15.10	19.878	27.75	100.00
gestation (dias)	12.00	35.75	79.00	142.35	207.50	645.00
predation	1.00	2.00	3.00	2.871	4.00	5.00
exposure	1.00	1.00	2.00	2.419	4.00	5.00
danger	1.00	1.00	2.00	2.613	4.00	5.00

Obs.: Algumas variáveis possuem valores ausentes: non_dreaming (14), dreaming (12), total_sleep (4), life_span (4), gestation (4).

A normalidade da variável resposta foi verificada via inspeção visual:

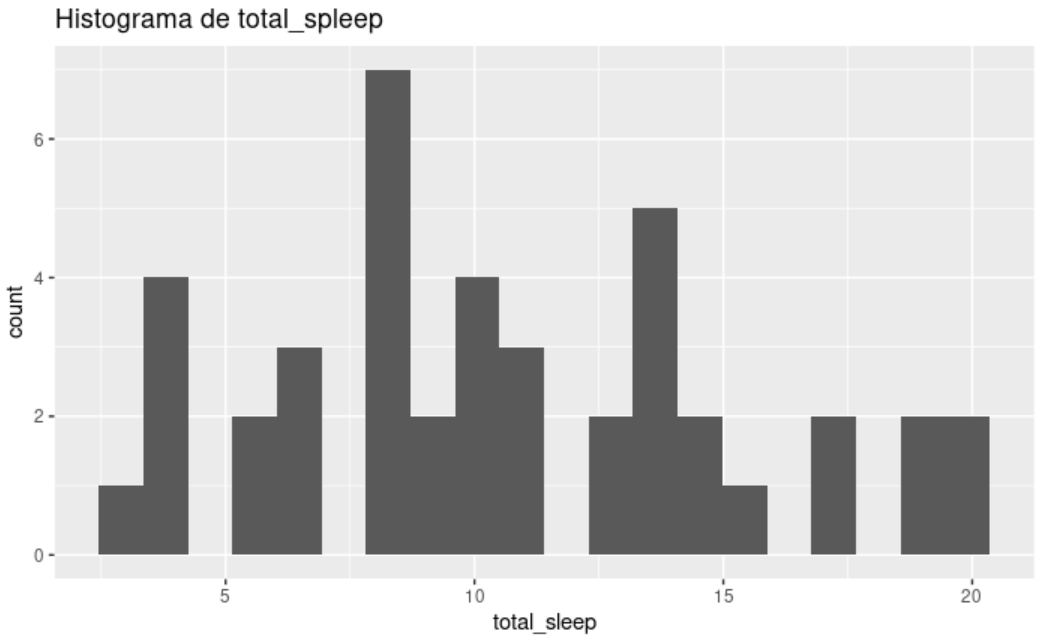


Figura 1: Histograma da variável `total_sleep`.

Conforme nota-se da figura, os dados não demonstram uma distribuição Normal ("curva de sino") e parecem possuir vários picos (multimodal), sugerindo que há subgrupos diferentes de mamíferos.

A princípio, tal fato não se caracteriza como problemático, mas reforça a ideia de que, após a criação do modelo, seria fundamental a realização da checagem da normalidade dos resíduos (erros) - suposição principal da regressão.

Correlação entre Variáveis Explicativas

Foram estimadas correlações entre variáveis numéricas e construída uma matriz de correlação:

Tabela 2: Correlação de Pearson entre variáveis do conjunto *mammals*.

Par de Variáveis	**r**	**Valor-p**	**IC 95%**
body_wt vs brain_wt	0.956	< 2.2e-16	[0.919 ; 0.976]
exposure vs danger	0.790	5.0e-10	[0.639 ; 0.882]
dreaming vs non_dreaming	0.518	0.00044	[0.254 ; 0.710]
predation vs exposure	0.626	9.38e-06	[0.397 ; 0.781]
predation vs danger	0.927	< 2.2e-16	[0.868 ; 0.961]
total_sleep vs danger	-0.604	2.26e-05	[-0.767 ; -0.368]
total_sleep vs exposure	-0.621	1.13e-05	[-0.778 ; -0.391]
total_sleep vs dreaming	0.717	9.1e-08	[0.528 ; 0.838]
total_sleep vs non_dreaming	0.968	< 2.2e-16	[0.940 ; 0.983]
total_sleep vs body_wt	-0.343	0.026	[-0.586 ; -0.043]
total_sleep vs brain_wt	-0.337	0.029	[-0.581 ; -0.037]
total_sleep vs predation	-0.405	0.0078	[-0.631 ; -0.115]
total_sleep vs life_span	-0.382	0.012	[-0.615 ; -0.089]
total_sleep vs gestation	-0.614	1.50e-05	[-0.774 ; -0.382]

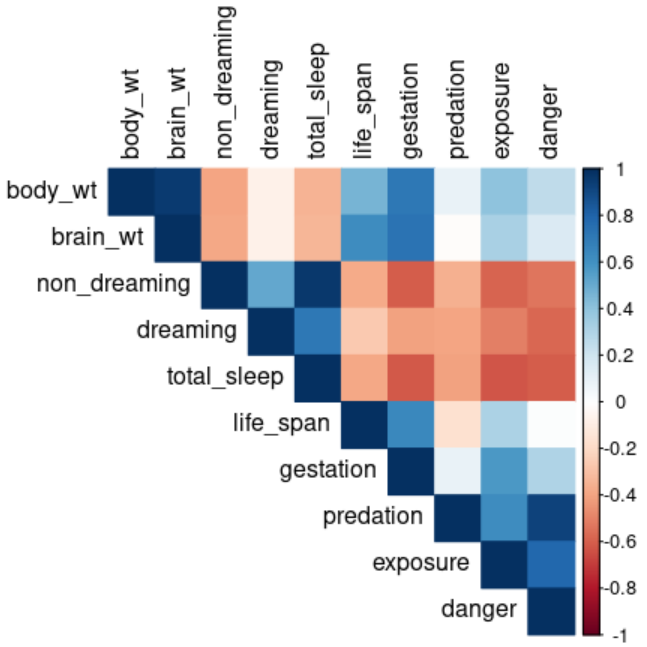


Figura 2: Heatmap de correlação das variáveis numéricas.

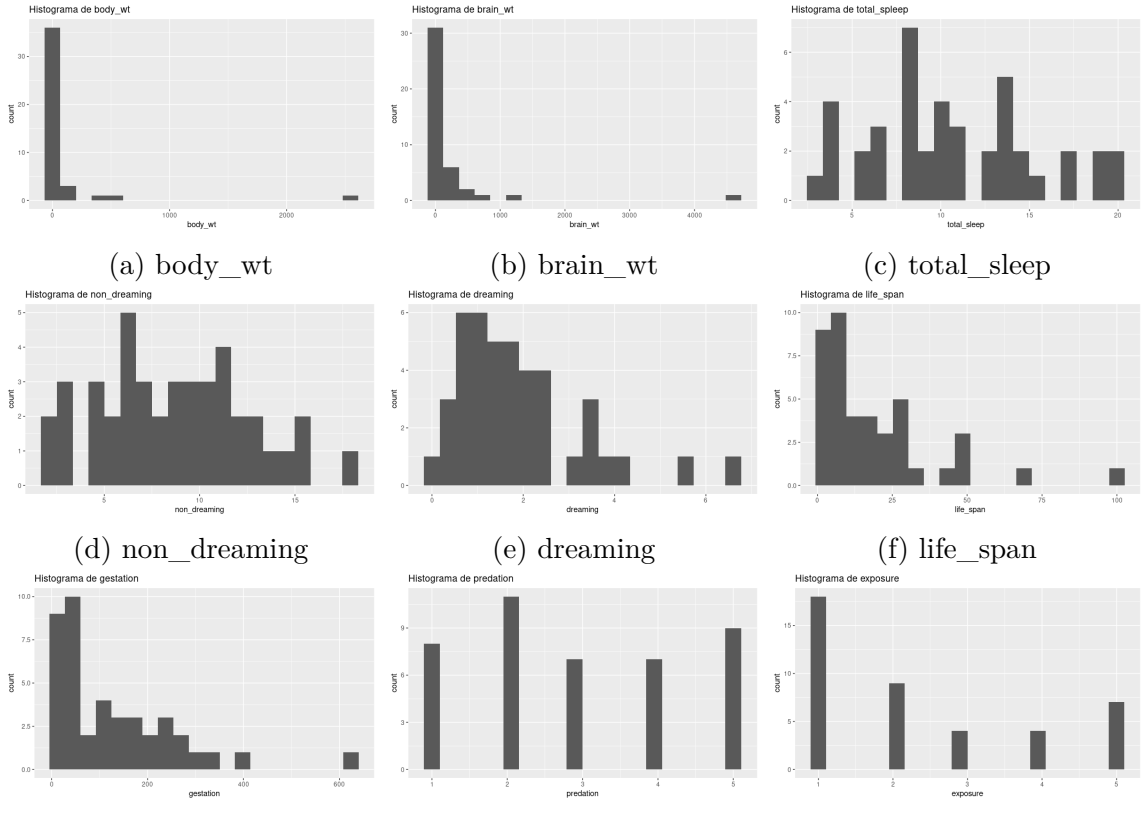
As correlações de Pearson indicam diversos padrões relevantes. Primeiro, observou-se uma correlação muito forte entre massa corporal e massa cerebral (r = 0.96), o que era esperado dado o crescimento conjunto dessas variáveis em mamíferos. Variáveis ecológicas, como exposição e perigo, também apresentaram correlação alta (r = 0.79 e r = 0.93, respectivamente), sugerindo consistência entre as classificações ambientais.

Com relação ao sono, identificou-se uma forte correlação positiva entre tempo total de sono e sono não-REM (r = 0.97), além de uma correlação significativa com o sono REM (r = 0.72). Variáveis relacionadas à vulnerabilidade predatória (predation, danger, exposure) mostraram correlações negativas moderadas com o tempo total de sono (entre r = -0.60 e -0.62), indicando que espécies mais vulneráveis tendem a dormir menos.

Também foram observadas correlações negativas significativas entre total de sono e características fisiológicas como massa corporal, massa cerebral, longevidade e tempo de gestação, todas entre -0.34 e -0.61.

Detecção de Outliers

Histogramas e boxplots permitiram inspeção das distribuições para detecção de possíveis outliers:

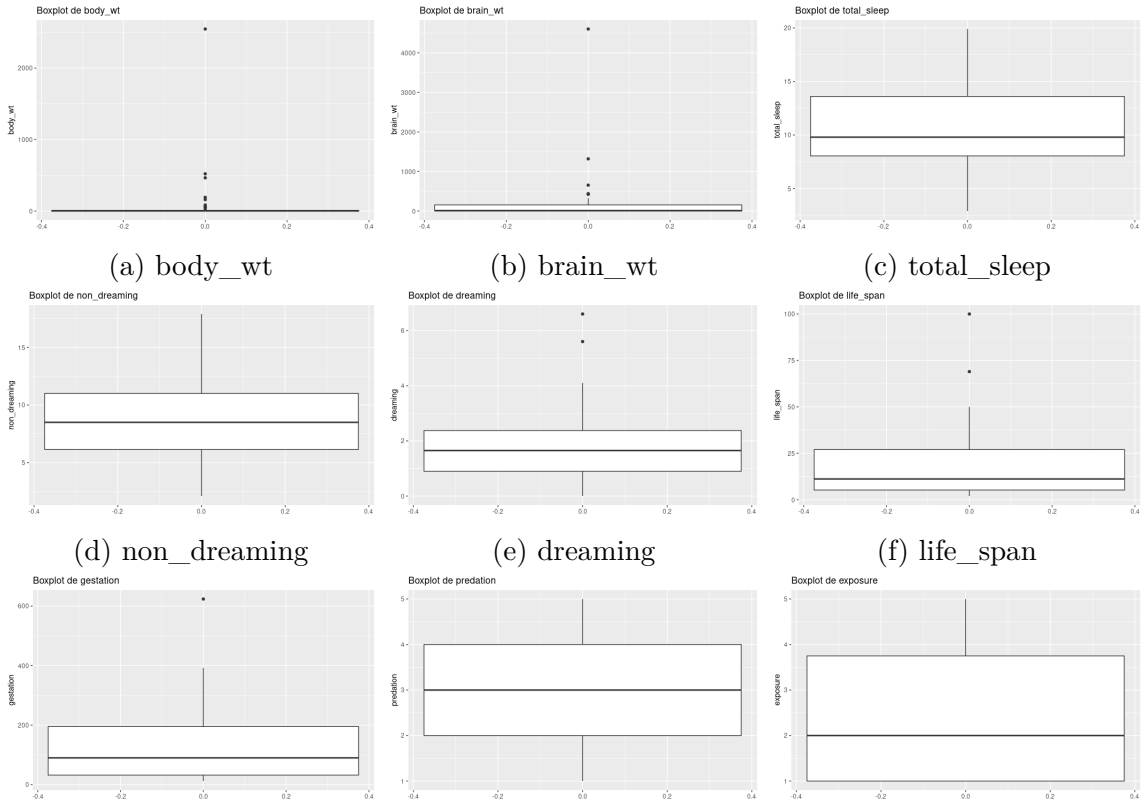


(a) body_wt (b) brain_wt (c) total_sleep

(d) non_dreaming (e) dreaming (f) life_span

(g) gestation (h) predation (i) exposure

Figura 3: Histogramas das variáveis numéricas do conjunto *mammals*.



(a) body_wt (b) brain_wt (c) total_sleep

(d) non_dreaming (e) dreaming (f) life_span

(g) gestation (h) predation (i) exposure

Figura 4: Boxplots das variáveis numéricas do conjunto *mammals*.

As variáveis relativas a peso corporal e à massa cerebral se destacam como possuindo distribuições extremamente assimétricas.

3.1 Regressão Linear Múltipla

Variáveis `non_dreaming` e `dreaming` foram removidas por serem preditoras perfeitas da variável resposta. A seguir, estão indicados os coeficientes do modelo completo ajustado:

Tabela 3: Resultados do modelo linear para *total_sleep*.

Variável	**Estimativa**	**Erro Padrão**	**t valor**	**Pr(> |t|)**
Intercept	17.1091	1.3364	12.803	1.47×10^{-14}
body_wt	0.00470	0.00592	0.794	0.43266
brain_wt	-0.000998	0.00355	-0.281	0.78059
life_span	-0.01458	0.04628	-0.315	0.75471
gestation	-0.01881	0.00698	-2.695	0.01086*
predation	2.31514	1.09269	2.119	0.04150*
exposure	0.58444	0.68458	0.854	0.39924
danger	-4.53757	1.35676	-3.344	0.00202**
Residual standard error: 3.039 (df = 34)				
Multiple R-squared: 0.6548		*Adjusted R-squared*: 0.5837		
F-statistic: 9.213 (df = 7, 34),		*p-value*: 2.398×10^{-6}		

Multicolinearidade (VIF)

A multicolinearidade é um problema conhecido em modelos de regressão, pelo fato de inflar as variâncias dos coeficientes ajustados do modelo, degradando a qualidade das estimativas dos coeficientes e dificultando a interpretação dos resultados obtidos. Desse modo, avaliamos os valores do Fator de Inflação da Variância (VIF) para cada um dos coeficientes ajustados no modelo exposto na tabela anterior. Abaixo, estão dispostos os resultados obtidos:

Tabela 4: Fator de Inflação da Variância (VIF) dos coeficientes da regressão.

Variável	**Valor**
body_wt	25.176774
brain_wt	30.126322
life_span	3.902487
gestation	3.535876
predation	11.109379
exposure	4.854897
danger	15.746424

Utilizando a regra prática de se considerar como sinais de multicolinearidade problemática valores de VIF superiores a 10 - um limite mais relaxado -, destacamos variâncias infladas para coeficientes de quatro variáveis, a saber: `danger`, `predation`, `brain_wt` e `body_wt`.

3.2 Modelos Ajustados

Dada a detecção de uma provável multicolinearidade prejudicial ao modelo, o número de variáveis explicativas foi progressivamente reduzido. Para isso, removemos do modelo a variável associada ao maior Fator de Inflação da Variância (VIF), seguindo-se a isto um novo ajuste. Esse processo foi repetido até que todos os VIFs calculados estivessem abaixo do limite prático estabelecido. A seguir, estão os VIFs calculados ao final do procedimento supracitado:

Tabela 5: Fator de Inflação da Variância (VIF) ao final do procedimento.

Variável	**Valor**
body_wt	2.055404
life_span	1.956265
gestation	3.472368
predation	2.26399
exposure	3.058928

Tabela 6: Resultados do modelo linear para *total_sleep* após procedimento.

Variável	**Estimativa**	**Erro Padrão**	**t valor**	**Pr($>\|t\|$)**
Intercept	17.60552	1.48557	11.851	5.52×10^{-14}***
body_wt	0.00252	0.00190	1.329	0.19219
life_span	-0.02484	0.03671	-0.676	0.50305
gestation	-0.02141	0.00775	-2.762	0.00899**
predation	-0.94610	0.55271	-1.712	0.09555.
exposure	-0.49240	0.60887	-0.809	0.42400
Residual standard error: 3.405 (df = 36)				
Multiple R-squared: 0.5411		*Adjusted R-squared*: 0.4774		
F-statistic: 8.49 (df = 5, 36),		*p-value*: 2.2×10^{-5}		

Note que o novo modelo sofreu uma redução em sua explicabilidade - R^2 regrediu de 0.58 para 0.47, o que não é desejável, dada a busca por um modelo explicativo do sono em mamíferos. Assim, com intuito de minimizar a degradação da explicabilidade do modelo, optou-se por métodos alternativos para se lidar com a multicolinearidade, mais especificamente: Stepwise e LASSO.

3.3 Seleção de Variáveis: Stepwise

Para tratar da multicolinearidade de forma a mitigar o impacto na explicabilidade do modelo, foi feito uso de métodos mais robustos, sendo a primeira tentativa com

Stepwise. O Stepwise se utiliza do Critério de Informação de Akaike (AIC) como aspecto de decisão pela manutenção ou remoção de variáveis no modelo. Nesta lógica, se a inserção ou remoção de uma explicativa reduz esta métrica no modelo, a ação em questão será tomada. O algoritmo é interrompido quando não há nenhuma ação de remoção ou adição que possa minimizar o AIC. Como resultado, o Stepwise otimizou o AIC (Critério de Informação de Akaike), mas não eliminou a multicolinearidade. Os VIFs das variáveis independentes podem ser consultados a seguir:

Tabela 7: Fatores de Inflação da Variância (VIF) do modelo reduzido.

Variável	**VIF**
body_wt	2.053251
gestation	2.514358
predation	10.128898
danger	11.104151

Ainda que o R^2 ajustado do modelo tenha melhorado, o método em questao não tratou a multicolinearidade, visto que ele foi contruído para se "preocupar"com com AIC do modelo, e não com o VIF.

O modelo final está exibido na tabela abaixo:

Tabela 8: Resultados do modelo linear para *total_sleep* após execução do Stepwise.

Variável	**Estimativa**	**Erro Padrão**	**t valor**	**Pr($>\|t\|$)**
Intercept	16.69277	1.19276	13.995	2.35×10^{-16}***
body_wt	0.002903	0.001655	1.754	0.08767.
gestation	-0.018581	0.005761	-3.225	0.00263**
predation	2.184876	1.021129	2.140	0.03904*
danger	-3.857624	1.115072	-3.460	0.00138**
Residual standard error: 2.974 (df = 37)				
Multiple R-squared: 0.6402		*Adjusted R-squared*: 0.6013		
F-statistic: 16.46 (df = 4, 37),		*p-value*: 7.88×10^{-8}		

3.4 Seleção de Variáveis: LASSO

O LASSO produziu o modelo final:

$$\text{total\_sleep} \sim \text{gestation} + \text{danger} + \text{life\_span}$$

que é muito mais simples - aspecto desejável - e robusto. Os coeficientes tanto de `danger` quanto de `gestation` e `life_span` são estatisticamente significantes. Estes

resultados podem ser lidos do dados do método *summary()* tabelados a seguir:

Tabela 9: Resultados do modelo linear para *total_sleep* após execução do LASSO.

Variável	**Estimativa**	**Erro Padrão**	**t valor**	**Pr(> |t|)**
Intercept	17.480006	1.180942	14.802	$< 2 \times 10^{-16}$***
gestation	-0.014134	0.005526	-2.558	0.014651*
danger	-1.645508	0.388417	-4.236	0.000139***
life_span	-0.029601	0.033202	-0.892	0.378243
Residual standard error: 3.18 (df = 38)				
Multiple R-squared: 0.5775 *Adjusted R-squared*: 0.5441				
F-statistic: 17.31 (df = 3, 38), *p-value*: 3.016×10^{-7}				

3.5 Análise de Resíduos

A análise de resíduos é essencial para verificar se o modelo linear ajustado atende aos pressupostos clássicos da regressão: normalidade dos erros, homocedasticidade, independência e ausência de padrões sistemáticos.

3.5.1 Normalidade dos erros

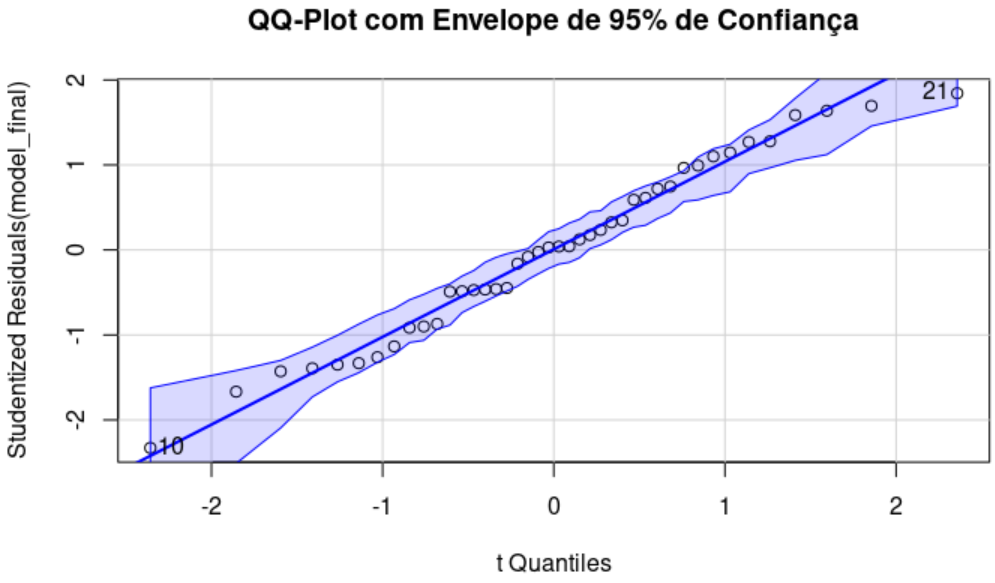


Figura 5: QQ-plot dos reísduos.

O Teste Shapiro-Wilk, cujos resultados são exibidos abaixo, confirmou a aparente normalidade visível no gráfico qq-plot, a 5

Tabela 10: Teste de Normalidade Shapiro-Wilk para os Resíduos

Estatística	Valor
W	0.97699
p-value	0.5479

3.5.2 Homocedasticidade

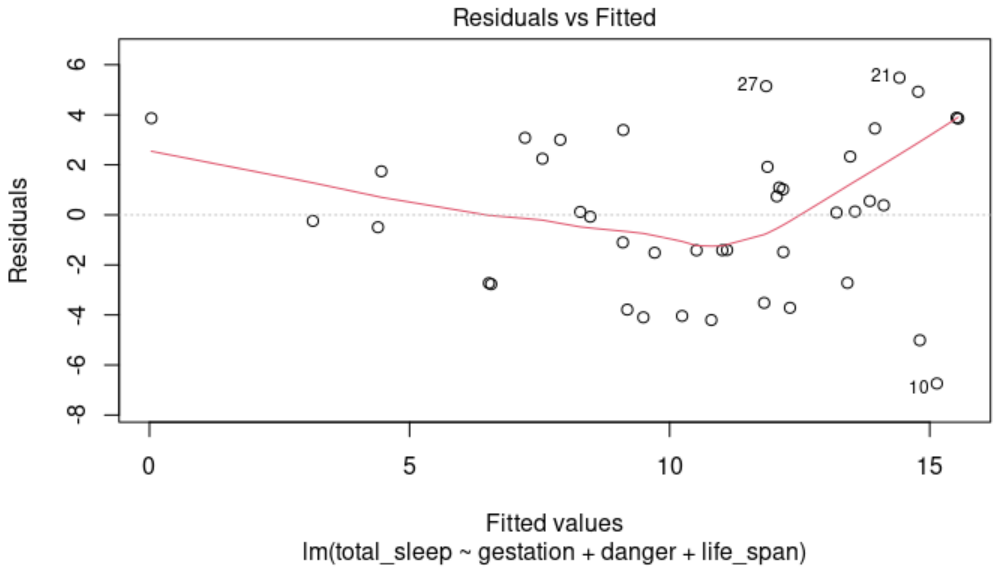


Figura 6: Resíduos vs. valores ajustados.

A linha forma uma curva, indicando que existe não-linearidade. Assim, o modelo linear atual não parece capturar bem a relação. Note, também, que a largura dos pontos varia ao longo da linha horizontal em y=0, configurando indício de heterocedasticidade.

Através do Teste Breusch–Pagan, conforme pode-se verificar na saída do R tabelada abaixo, foi detectada heterocedasticidade significativa.

Tabela 11: Teste de Breusch-Pagan (studentized) para heterocedasticidade.

Estatística	Valor	df	p-value
BP	10.873	3	0.01243

Como o p-valor é menor do que 0.05 , então a um nível de 5rejeitamos H_0 e consideramos que o modelo apresenta heterocedasticidade.

3.6 Transformação Box–Cox

Com os resultados apresentados na subseção anterior, buscou-se fazer uma transformação de variáveis. Uma medida corretiva comum para a não-normalidade e/ou variância não-constante é transformar a variável resposta Y, criando uma nova variável Y'. A transformação Box–Cox reduziu heterocedasticidade e melhorou normalidade dos resíduos:

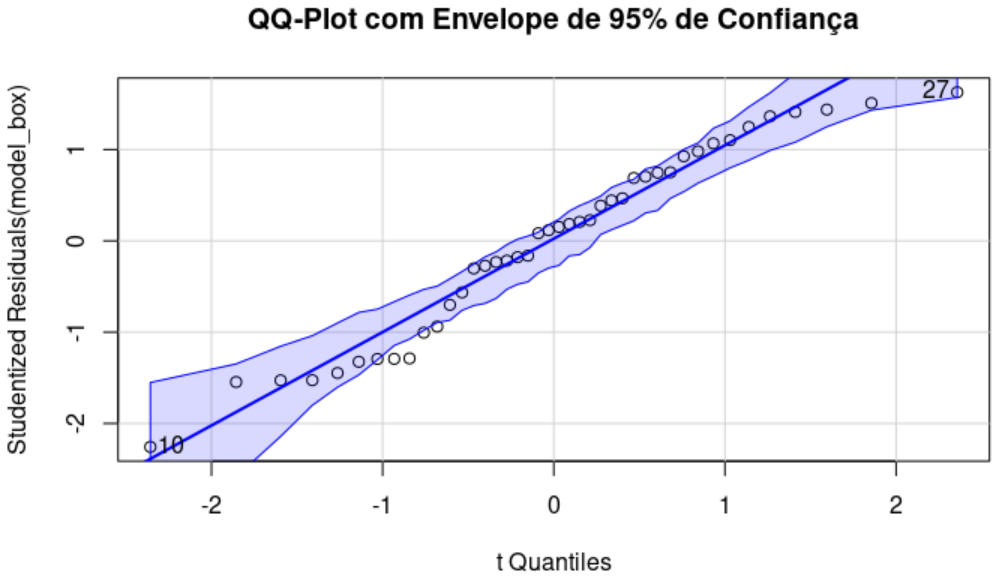


Figura 7: Modelos transformados via Box–Cox.

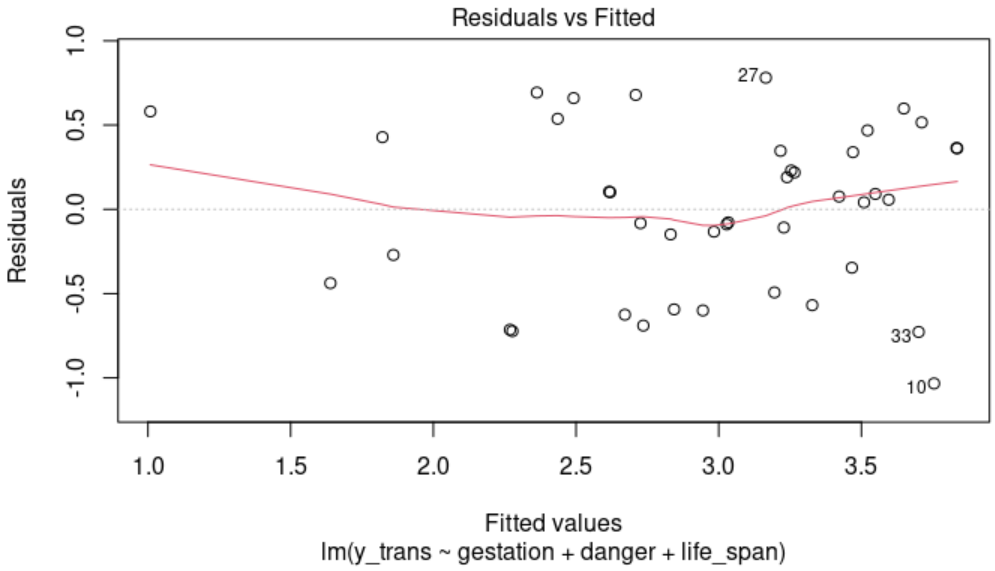
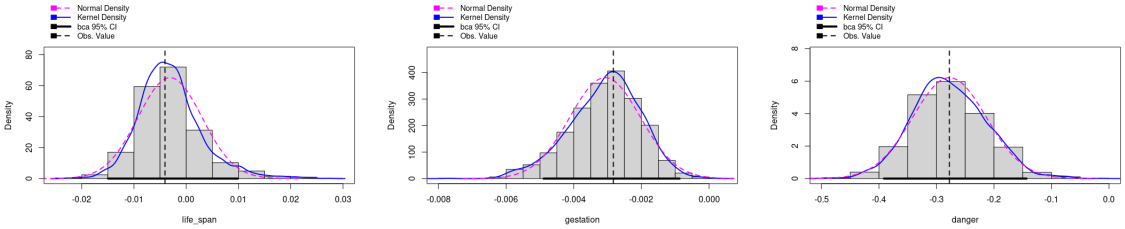


Figura 8: Modelos transformados via Box–Cox.

No que se refere à heterocedasticidade, houve melhora significativa, sendo resolvida em grande parte. Não parece haver padrão no gráfico Residuos vs Valores Preditos. Quanto à Normalidade, os erros continuam seguindo distribuição Normal. O modelo foi o que melhor se adequou aos dados.

3.7 Bootstrap dos Coeficientes

Com os problemas de multicolinearidade inexata e heterocedasticidade tratadas, buscamos estimar a incerteza ao redor dos coeficientes ajustados da regressão. Abaixo, indicamos os resultados para os coeficientes:



(a) boots_life$_s$*pan* (b) boots_gestation (c) boots_danger

Figura 9: Histogramas das variáveis numéricas do conjunto *mammals*.

A seguir, apresentamos o resultado para o intercepto:

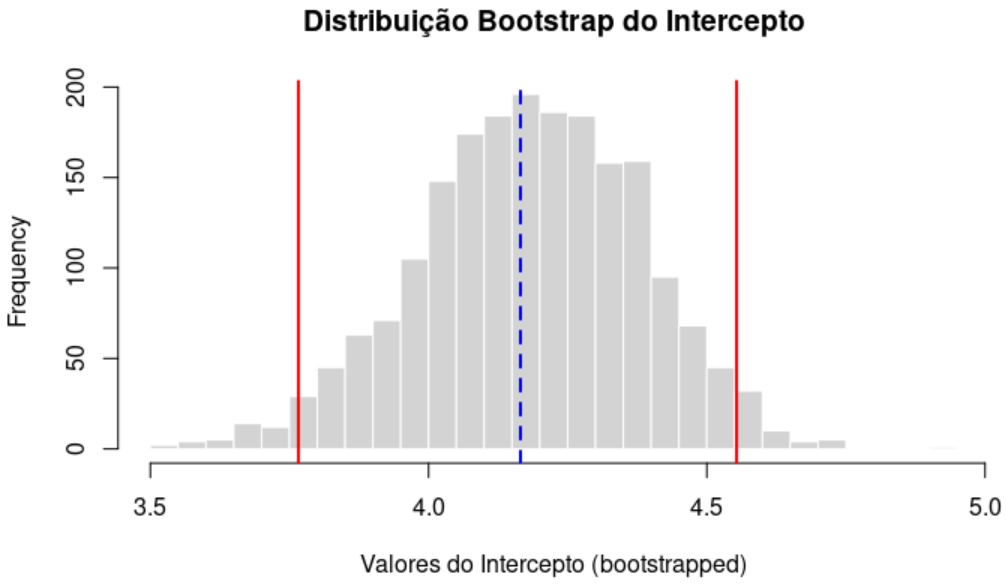


Figura 10: Distribuição bootstrap do intercepto do modelo.

3.8 Validação Cruzada

Após desenvolvermos o modelo explicativo, surgiu a necessidade de elaborar-se um modelo preditivo, avaliado de acordo com as seguintes etapas:

1. Separação dos dados em conjuntos de treino e teste.
2. Treinamento com o conjunto de treino e aplicação de validação cruzada para avaliar a performance.
3. Avaliação da performance final com o conjunto de teste.

Anteriormente, nosso interesse estava voltado para compreender a relação entre a variável resposta **total_sleep** (horas totais de sono por dia) e as variáveis explicativas. Nesse contexto, buscávamos aprimorar métricas como o R^2, avaliar a significância dos coeficientes e aumentar a interpretabilidade do modelo de regressão. Agora, entretanto, nosso objetivo é diferente: desejamos utilizar as variáveis independentes para *predizer* novos valores da variável dependente. Em outras palavras, dado que seja incluída uma nova observação — isto é, um novo mamífero com seus respectivos atributos coletados — queremos estimar qual será sua quantidade média de horas de sono. Os dados resultantes da validação cruzada com 10 folds podem ser lidos a seguir:

Tabela 12: Resultados da Validação Cruzada (Regressão Linear)

Métrica	**RMSE**	R^2	**MAE**
Valor	3.415415	0.7267677	2.824077

Amostras: 42
Preditores: 7
Pré-processamento: nenhum
Validação: 10-fold Cross Validation
Parâmetro: intercepto mantido em TRUE

Não é possível tirar conclusões apenas observando isoladamente as métricas retornadas (RMSE, R^2, MAE). Por isso, consideramos útil compará-las com uma alternativa adequada do método k-folds. Dado que nossa amostra é relativamente pequena (n=42), optamos por utilizar também o método LOOCV, que consiste em gerar n folds — cada um deixando exatamente uma observação para teste e utilizando as demais n1 para treino. Os resultados comparativos são apresentados na tabela a seguir:

Tabela 13: Resultados da Validação Cruzada (Regressão Linear — LOOCV)

Métrica	**RMSE**	R^2	**MAE**
Valor	4.243069	0.4131189	2.958826

Amostras: 42
Preditores: 7
Pré-processamento: nenhum
Validação: Leave-One-Out Cross Validation (LOOCV)
Parâmetro: intercepto mantido em TRUE

Observe que há aumento no erro quadrático médio (RMSE) e no erro médio absoluto (MAE), mas redução no R^2 - o que indica menor explicabilidade pelo modelo. Este último não é um grande problema, dado que estamos interessados no caráter preditivo do modelo. Com apenas 42 observações, o método LOOCV gera alta variância nas estimativas, pois cada treino é feito com quase toda a base e cada observação tem grande influência no erro final. Isso tende a aumentar o RMSE e o MAE e reduzir o R^2. Já o k-fold (k=10) reduz a variância ao usar partições maiores, produzindo estimativas mais estáveis. Por isso, no nosso caso, o 10-fold apresentou erros menores e R^2 maior, indicando melhor equilíbrio entre viés e variância para esse tamanho de amostra. Assim sendo, como nosso foco é a capacidade de previsão do modelo, o método k-fold parece ser o mais adequado, o que vai de encontro com a ideia de sua posição como 'padrão ouro' da validação cruzada.

4 Discussão

Os resultados obtidos ao longo da análise mostram que o tempo total de sono entre mamíferos é determinado por uma combinação complexa de fatores fisiológicos e ecológicos. O conjunto de evidências aponta que espécies mais vulneráveis dormem menos, conforme sugerido pelas correlações negativas entre predation, danger e exposure e o total de sono. Esse padrão é consistente com interpretações evolutivas: animais expostos a maior risco tendem a adotar comportamentos que reduzam períodos de inatividade prolongada.

Do ponto de vista fisiológico, características como massa corporal e tempo de gestação também influenciam o comportamento de sono. Mamíferos de maior porte, por exemplo, tendem a dormir menos — tendência amplamente descrita na literatura evolutiva e observada nos dados. Da mesma forma, a associação negativa entre gestation e total_sleep sugere possíveis mecanismos bioenergéticos relacionados ao investimento reprodutivo.

Durante o ajuste dos modelos, verificou-se uma forte multicolinearidade, especialmente entre body_wt e brain_wt, que apresentaram comportamento quase redundante. Essa característica estrutural dos dados demandou a redução de variáveis e o uso de abordagens robustas para seleção de preditores. O método Stepwise baseado em AIC mostrou-se eficiente para identificar modelos mais parcimoniosos, enquanto o LASSO destacou-se por penalizar variáveis altamente correlacionadas e estabilizar os coeficientes, reforçando a importância de preditores como gestation e predation.

A avaliação dos resíduos indicou violação moderada da homocedasticidade, posteriormente corrigida por meio da transformação Box–Cox, que contribuiu para estabilizar a variância e melhorar a adequação do modelo final. O teste de Shapiro–Wilk indicou que os resíduos transformados permanecem próximos de uma distribuição normal, sustentando a validade da inferência.

No que diz respeito ao desempenho preditivo, observou-se discrepância entre LOOCV e 10-fold cross-validation. Para a amostra pequena disponível (n=42), o LOOCV apresentou maior variabilidade e maior erro preditivo — comportamento esperado dada sua maior sensibilidade a observações influentes. Em contraste, o 10-fold exibiu menor RMSE e maior R^2, sugerindo maior estabilidade e melhor capacidade de generalização.

Em síntese, os dados indicam que pressões ecológicas, características fisiológicas e restrições evolutivas moldam o sono dos mamíferos. A modelagem estatística, quando aliada a técnicas de seleção de variáveis e transformações adequadas, foi capaz de capturar parte importante dessa complexidade, produzindo um modelo final coerente com as hipóteses biológicas.

5 Conclusão

A análise realizada permite concluir que o tempo total de sono em mamíferos está associado a um conjunto de fatores ecológicos e fisiológicos que interagem de forma não trivial. Variáveis como gestation, predation e danger se mostraram particularmente relevantes em diferentes etapas da modelagem, sugerindo que tanto pressões ambientais quanto demandas bioenergéticas têm papel fundamental na determinação do comportamento de sono.

A presença de multicolinearidade estrutural, ocasionada sobretudo pela redundância entre body_wt e brain_wt, reforçou a necessidade de métodos mais robustos para seleção e estabilização dos modelos. O uso conjunto do Stepwise (AIC), do LASSO e da transformação Box–Cox demonstrou-se essencial para obter um conjunto parcimonioso de preditores, melhorar a qualidade dos resíduos e aumentar a interpretabilidade.

As validações cruzadas demonstraram que, embora o modelo apresente desempe-

nho preditivo moderado, sua capacidade de generalização depende sensivelmente do método utilizado. O 10-fold mostrou melhor desempenho que o LOOCV, evidenciando menor variância e maior estabilidade — característica desejável para aplicações preditivas com amostras pequenas.

Em síntese, após diversos procedimentos estatísticos

- A multicolinearidade foi reduzida com LASSO.
- Heterocedasticidade foi sanada com transformação Box–Cox.
- O modelo final apresenta bom desempenho, interpretação simples e resíduos adequados.

Assim, o melhor modelo encontrado — consistente com as análises de AIC, LASSO e validação cruzada — foi:

$$\text{total\_sleep}^{(\lambda)} = \beta_0 + \beta_1 \cdot \text{gestation} + \beta_2 \cdot \text{danger}$$

o que evidencia que tanto fatores reprodutivos quanto fatores ecológicos associados ao risco desempenham papel central na explicação da variação no tempo de sono entre mamíferos.

6 Referências

Referências

[1] Slides da disciplina DIG EST035 - Análise de Regressão, 2025/2.

[2] Kaggle. *Sleep in Mammals*. Disponível em: `https://www.kaggle.com/datasets/volkandl/sleep-in-mammals`. Acesso em: XX.