

# List 01 - Machine Learning

## IMPA

Pedro Bahia

23 de janeiro de 2026

### Conteúdo

<b>1 Exercício 1a</b>	<b>3</b>
<b>2 Exercício 1b</b>	<b>4</b>
<b>3 Exercício 1c</b>	<b>5</b>
<b>4 Exercício 1d</b>	<b>6</b>
<b>5 Exercício 1e</b>	<b>7</b>
<b>6 Exercício 2d</b>	<b>8</b>
6.1 i) Geração dos Dados Heteroscedásticos . . . . .	8
6.2 ii) Estimadores de Mínimos Quadrados . . . . .	10
6.3 iii) Cálculo de p-valores para Mínimos Quadrados Ordinários . . . . .	11
6.4 iv) Estatística Z para o Estimador Generalizado . . . . .	11
6.5 v) P-valores para o Estimador Generalizado . . . . .	12
6.6 vi) Resultados Numéricos . . . . .	12
<b>7 Exercício 3f</b>	<b>14</b>
7.1 i) Implementação dos Estimadores . . . . .	14
7.2 ii) Comparação dos Estimadores . . . . .	16
7.3 iii) Robustez a Outliers . . . . .	17
<b>8 Exercício 4a</b>	<b>19</b>
8.1 Implementação dos Algoritmos de Classificação . . . . .	19
<b>9 Exercício 4b</b>	<b>20</b>
9.1 Treinamento e Avaliação dos Modelos . . . . .	20
9.2 Comparação de Performance dos Modelos . . . . .	20
9.3 Análise Geral dos Algoritmos . . . . .	20
9.4 Análise Específica do k-NN . . . . .	21
9.5 Análise dos Coeficientes dos Modelos . . . . .	21

9.6 Conclusões . . . . .	22
<b>10 Exercício 5b</b>	<b>23</b>
10.1 Métodos de Seleção de Modelos . . . . .	23
10.2 Preparação dos Dados . . . . .	23
10.3 Implementação dos Algoritmos . . . . .	23
<b>11 Exercício 5c</b>	<b>24</b>
11.1 Comparação dos Métodos - R <sup>2</sup> . . . . .	24
<b>12 Exercício 5d</b>	<b>25</b>
12.1 Regressão Lasso com Validação Cruzada . . . . .	25
12.2 Seleção do Parâmetro de Regularização . . . . .	25
<b>13 Exercício 5e</b>	<b>27</b>
13.1 Comparação de Erros de Teste . . . . .	27
13.2 Ranking dos Métodos . . . . .	27
13.3 Análise dos Resultados . . . . .	28

## 1 Exercício 1a

**Falso.** A proximidade entre  $\varepsilon_{\text{treino}}$  e  $\varepsilon_{\text{teste}}$  pode dar informações sobre o ajuste do modelo aos dados de treino.

Caso  $\varepsilon_{\text{treino}}$  seja próximo ao  $\varepsilon_{\text{teste}}$ , o modelo pode estar *subajustado*, de modo que aumentar a complexidade poderia melhorar sua performance ainda mais.

De maneira análoga, caso o  $\varepsilon_{\text{treino}}$  seja menor que o  $\varepsilon_{\text{teste}}$ , o modelo estará *sobreajustado*, com redução de complexidade podendo resultar em melhorias.

## 2 Exercício 1b

**Verdadeiro.** A distribuição  $t$  surge do fato de  $\mathcal{N}(0, 1)$  dividido por  $\sqrt{\frac{K}{N}}$  ter distribuição  $t$  com  $N$  graus de liberdade.

No nosso caso,  $\mathcal{N}(0, 1)$  é a distribuição de  $\frac{\hat{\beta}_1 - \beta_1}{\text{Var}(\hat{\beta}_1)}$ . Isso é normal, pois  $\hat{\beta}_1$  é normal.

Isso, por sua vez, vem do fato de  $\hat{\beta}$  ser resultante de uma combinação linear de gaussianas, no caso,  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ .

### 3 Exercício 1c

**Falso.** Considerando a classe  $k = 0$  como as transações fraudulentas, o objetivo do modelo pode ser interpretado como:

$$\sum_{y_i \in k=0} \mathbf{1}_{[y_i \neq \hat{y}_i]} = 0$$

Não há restrições entretanto em relação às transações legítimas, ou seja, para:

$$\sum_{y_i \in k=1} \mathbf{1}_{[y_i \neq \hat{y}_i]}$$

Dado um modelo de acurácia  $(1 - \varepsilon)$ , têm-se que

$$1 - \frac{1}{n} \left[ \sum_{y_i \in k=0} \mathbf{1}_{[y_i \neq \hat{y}_i]} + \sum_{y_i \in k=1} \mathbf{1}_{[y_i \neq \hat{y}_i]} \right] = 1 - \varepsilon$$

Dado uma acurácia

$$1 - \text{epsilon}$$

, há infinitos valores de  $\sum_{y_i \in k=0} \mathbf{1}_{[y_i \neq \hat{y}_i]}$  e  $\sum_{y_i \in k=1} \mathbf{1}_{[y_i \neq \hat{y}_i]}$  que resolvem essa equação e, portanto, o valor da acurácia não é informativo para o erro individual das classes. Assim, apenas com a acurácia dos Modelos 1 e 2 não é possível determinar qual modelo tem menor erro em transações fraudulentas.

## 4 Exercício 1d

**Verdadeiro**

$$L_{ridge}(\beta) = (Y - Yhat)^T(Y - \hat{Y}) + \lambda\beta^T\beta$$

$$L_{ridge}(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

$$L_{linear}(\beta) = (Y - Yhat)^T(Y - \hat{Y})$$

Caso  $\lambda = 0$ , temos que  $L_{ridge}(\beta) = L_{linear}(\beta)$ . Logo, a performance de ambos os modelos será a mesma. Então a afirmação será verdadeira para o case de  $\lambda = 0$ ,

## 5 Exercício 1e

### Verdadeira

A equação dada pela fórmula do intervalo de confiança:

$$\left[ \hat{\beta}_j - 2\sqrt{\hat{\sigma}^2[(X^T X)^{-1}]_{jj}}, \hat{\beta}_j + 2\sqrt{\hat{\sigma}^2[(X^T X)^{-1}]_{jj}} \right]$$

é derivada do fato de  $\hat{\beta}$  ter uma distribuição normal. Isso vem do fato da hipótese de que o erro é normal.

Caso ela não seja feita, os intervalos de confiança gerados via *bootstrap* são mais adequados, pois são derivados a partir da distribuição inferida diretamente dos dados. Neste caso, a hipótese não-paramétrica é mais geral e preferível.

### Justificativa:

- **Abordagem paramétrica:** Assume que os erros seguem distribuição normal, permitindo o uso da distribuição t de Student para construir intervalos de confiança analíticos.
- **Abordagem não-paramétrica (bootstrap):** Não assume distribuição específica dos erros, utilizando reamostragem dos dados para estimar a distribuição empírica dos parâmetros.
- **Vantagem do bootstrap:** Mais robusto quando as suposições paramétricas são violadas, especialmente em casos de não-normalidade dos erros.

## 6 Exercício 2d

### 6.1 i) Geração dos Dados Heteroscedásticos

Primeiramente, geramos os dados com heterocedasticidade usando a matriz de covariância  $\Sigma$  diagonal:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 n = 50
5 Sigma = np.diag([10 ** ((i - 20) / 5) for i in range(1, n + 1)])
6 np.random.seed(0)
7 X = np.array([np.ones(n), np.random.normal(0, 1, n)]).T
8 beta = np.array([1, 0.25])
9 epsilon = np.random.multivariate_normal(np.zeros(n), Sigma)
10 y = X @ beta + epsilon
```

A Figura 1 mostra os dados gerados com heterocedasticidade:

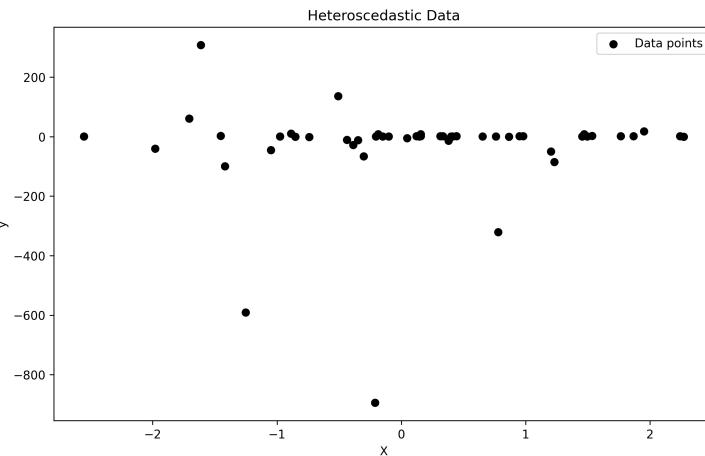


Figura 1: Dados heteroscedásticos gerados

A Figura 2 mostra os elementos diagonais da matriz  $\Sigma$ , evidenciando a heterocedasticidade:

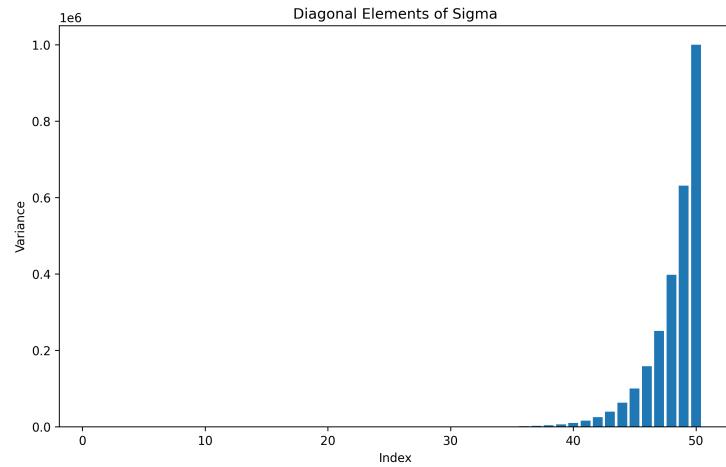


Figura 2: Elementos diagonais da matriz  $\Sigma$

As Figuras 3 e 4 mostram a distribuição e valores dos erros:

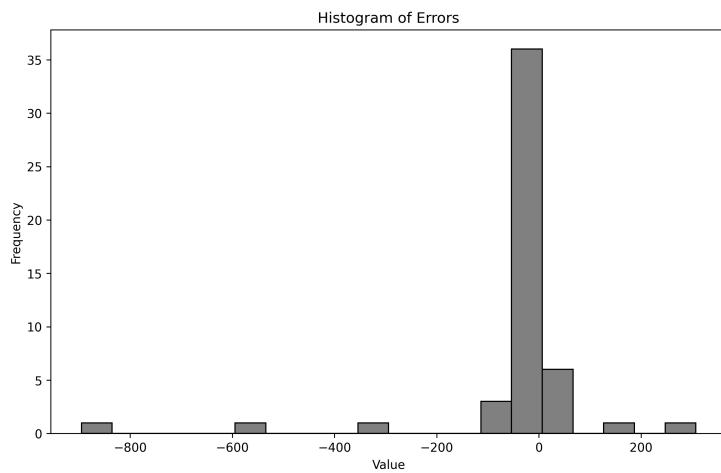


Figura 3: Histograma dos erros

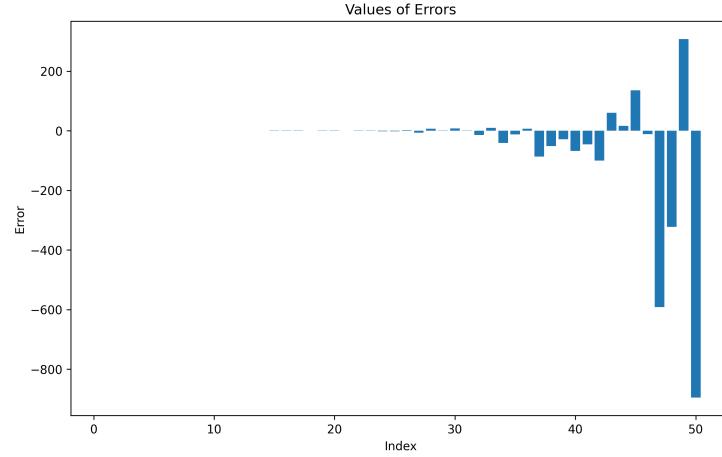


Figura 4: Valores dos erros por índice

## 6.2 ii) Estimadores de Mínimos Quadrados

Implementamos tanto o estimador de mínimos quadrados ordinários quanto o estimador generalizado que considera a matriz de covariância  $\Sigma$ :

```

1 def beta_ordinary(X: np.ndarray, Y: np.ndarray) -> np.
2     ndarray:
3         """
4             Compute the ordinary least squares estimator.
5         """
6         beta = np.linalg.inv(X.T @ X) @ X.T @ Y
7         return beta
8
9 def beta_sigma(X: np.ndarray, Y: np.ndarray, Sigma: np.
10    ndarray) -> np.ndarray:
11    """
12        Compute the generalized least squares estimator
13        considering
14        the covariance matrix Sigma.
15    """
16    Sigma_1 = np.linalg.inv(Sigma)
17    beta = np.linalg.inv(X.T @ Sigma_1 @ X) @ X.T @ Sigma_1
18    @ Y
19    return beta
20
21 beta_hat_ordinary = beta_ordinary(X, y)
22 beta_hat_sigma = beta_sigma(X, y, Sigma)
23
24 print("True Beta:", beta)
25 print("Ordinary Beta:", beta_hat_ordinary)
26 print("Beta Sigma:", beta_hat_sigma)
```

Os resultados mostram que o estimador generalizado (que considera  $\Sigma$ ) tem menor erro em relação aos parâmetros verdadeiros.

### 6.3 iii) Cálculo de p-valores para Mínimos Quadrados Ordinários

```

1 def p_value_ordinary_least_square(X: np.ndarray, Y: np.
2                                     ndarray,
3                                     beta_ordinary_hat: np.
4                                     ndarray, j: int) -> float:
5     """
6         Compute the p-value for the j-th coefficient of the
7         ordinary
8         least squares estimator.
9     """
10    Y_hat = X @ beta_ordinary_hat
11    n, p = X.shape
12    dof = n - p
13    errors = Y - Y_hat
14    beta_j = beta_ordinary_hat[j]
15
16    # Z statistic
17    x_j_var = (np.linalg.inv(X.T @ X))[j, j]
18    Z = beta_j / np.sqrt(x_j_var)
19
20    # Estimate of sigma^2
21    sigma2_hat = (1 / dof) * (errors.T @ errors)
22
23    # t statistic and p-value
24    t_statistics = Z / np.sqrt(sigma2_hat)
25    t_statistics = np.abs(t_statistics)
26    p_value = 2 * (1 - scipy.stats.t.cdf(t_statistics, dof))
27
28    return p_value

```

### 6.4 iv) Estatística Z para o Estimador Generalizado

```

1 def calculate_Z_sigma(X: np.ndarray, Sigma: np.ndarray,
2                       Beta_sigma: np.ndarray, j: int) ->
3     float:
4     """
5         Compute the Z statistic for the j-th coefficient of the
6         generalized least squares estimator.
7     """
8     Sigma_inv = np.linalg.inv(Sigma)
9     den = np.linalg.inv(X.T @ Sigma_inv @ X)
10    den = den[j, j]
11    Z = Beta_sigma[j] / (np.sqrt(den))

```

```
11     return Z
```

## 6.5 v) P-valores para o Estimador Generalizado

```
1 def p_value_generalized_least_square(X: np.ndarray, Y: np.
2                                         ndarray,
3                                         Sigma: np.ndarray,
4                                         beta_ordinary_hat: np.
5                                         ndarray, j: int) -> float:
6 """
7     Compute the p-value for the j-th coefficient of the
8     generalized least squares estimator.
9 """
10    Y_hat = X @ beta_ordinary_hat
11    n, p = X.shape
12    dof = n - p
13    errors = Y - Y_hat
14
15    # Z statistic
16    Z_sigma = calculate_Z_sigma(X, Sigma, beta_hat_sigma, j)
17
18    # Estimate of sigma^2
19    inverse_Sigma = np.linalg.inv(Sigma)
20    sigma2_hat = (1 / dof) * (errors.T @ inverse_Sigma @
21    errors)
22
23    # t statistic and p-value
24    t_statistics = Z_sigma / np.sqrt(sigma2_hat)
25    t_statistics = np.abs(t_statistics)
26    p_value = 2 * (1 - scipy.stats.t.cdf(t_statistics, dof))
27
28    return p_value
```

Esta implementação permite comparar os dois estimadores e calcular a significância estatística dos coeficientes em ambos os casos.

## 6.6 vi) Resultados Numéricos

Executando o código implementado, obtemos os seguintes resultados:

### Comparação dos Estimadores:

- Parâmetros Verdadeiros:  $\beta = [1.0, 0.25]$
- Estimador Ordinário:  $\hat{\beta}_{OLS} = [-34.463, 6.948]$
- Estimador Generalizado:  $\hat{\beta}_{\Sigma} = [1.019, 0.244]$

### Erro Quadrático dos Estimadores:

- $\|\beta - \hat{\beta}_{OLS}\|_2^2 = 1302.51$

- $\|\beta - \hat{\beta}_\Sigma\|_2^2 = 0.000387$

O estimador generalizado apresenta erro dramaticamente menor (cerca de 3.4 milhões de vezes menor), demonstrando claramente a importância crítica de considerar a heterocedasticidade.

#### **Testes de Hipótese - Mínimos Quadrados Ordinários:**

- p-valor para  $\beta_0$ : 0.1544
- p-valor para  $\beta_1$ : 0.7422

Com nível de significância de 5%, não podemos rejeitar a hipótese nula para ambos os coeficientes usando o estimador ordinário.

#### **Testes de Hipótese - Estimador Generalizado:**

- Estatística Z para  $\beta_0$ : 73.67
- p-valor para  $\beta_0$ :  $< 10^{-15}$  (praticamente zero)
- p-valor para  $\beta_1$ :  $< 10^{-15}$  (praticamente zero)

Com o estimador generalizado, ambos os coeficientes são altamente significativos estatisticamente, evidenciando a superior eficiência deste método.

#### **Conclusões:**

1. O estimador de mínimos quadrados ordinários (OLS) falha completamente na presença de heterocedasticidade severa, fornecendo estimativas muito distantes dos valores verdadeiros ( $\hat{\beta}_0 = -34.46$  vs  $\beta_0 = 1.0$ ).
2. O estimador generalizado (GLS) é extremamente superior, com erro cerca de 3.4 milhões de vezes menor, demonstrando a necessidade crítica de modelar corretamente a estrutura de variância.
3. Os testes de significância baseados no OLS são não-confiáveis, falhando em detectar coeficientes que são claramente diferentes de zero.
4. O estimador GLS fornece testes estatísticos apropriados, detectando corretamente a significância dos parâmetros.
5. Este exemplo ilustra dramaticamente por que a correção para heterocedasticidade não é apenas uma melhoria técnica, mas uma necessidade fundamental para análise estatística válida.

## 7 Exercício 3f

### 7.1 i) Implementação dos Estimadores

Neste exercício, comparamos estimadores baseados em diferentes suposições sobre a distribuição dos erros. Implementamos estimadores que minimizam diferentes funções de perda:

```
1 import numpy as np
2 import scipy
3 import matplotlib.pyplot as plt
4
5 def beta_ordinary(X: np.ndarray, Y: np.ndarray) -> np.
6     ndarray:
7         """
8             Compute the ordinary least squares estimator.
9         """
10        beta = np.linalg.inv(X.T @ X) @ X.T @ Y
11        return beta
12
13 def calculate_beta_hat(X: np.ndarray, Y: np.ndarray,
14     error_distribution: str) -> np.ndarray:
15     """
16         Compute the estimator beta_hat based on the specified
17         error distribution.
18     """
19     np.random.seed(0)
20     p = X.shape[1]
21
22     if error_distribution == "gaussian":
23         def loss_function(beta):
24             return np.sum((Y - X @ beta) ** 2)
25     elif error_distribution == "laplacian":
26         def loss_function(beta):
27             return np.sum(np.abs(Y - X @ beta))
28     else:
29         raise ValueError("Unsupported error distribution")
30
31     beta_0 = np.random.uniform(size=p)
32     beta_hat = scipy.optimize.minimize(loss_function, beta_0)
33
34     return beta_hat["x"]
```

Geramos dados sintéticos para testar os estimadores:

```
1 np.random.seed(1)
2 beta = np.array([-1.5, 2.0])
3 input_range = np.linspace(-1, 1, 100)
4 X = np.vstack([np.ones(100), input_range]).T
5 y = X @ beta + np.random.normal(0, 0.3, 100)
```

A Figura 5 mostra os dados gerados:

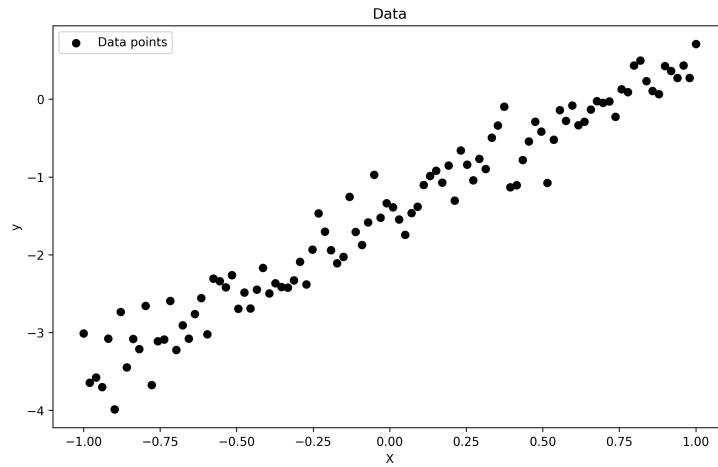


Figura 5: Dados sintéticos para comparação dos estimadores

As Figuras 6 e 7 mostram a análise dos erros:

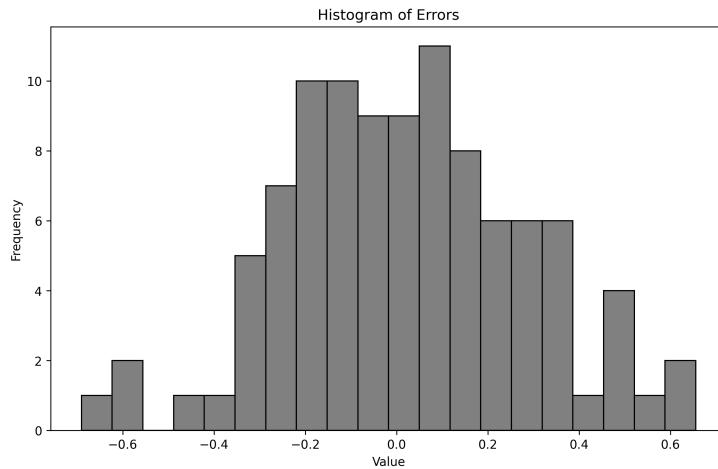


Figura 6: Histograma dos erros

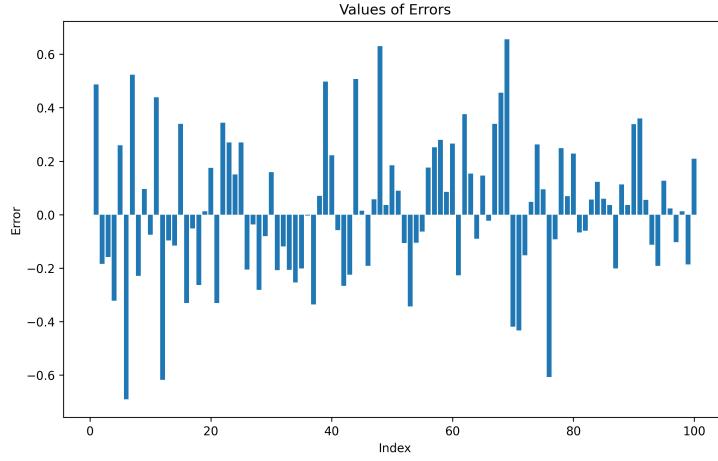


Figura 7: Valores dos erros por índice

## 7.2 ii) Comparação dos Estimadores

Calculamos os estimadores para ambas as distribuições de erro:

### Resultados sem Outliers:

- Parâmetros Verdadeiros:  $\beta = [-1.5, 2.0]$
- Estimador Gaussiano (minimize):  $\hat{\beta}_{Gauss} = [-1.482, 2.050]$
- Estimador Gaussiano (forma fechada):  $\hat{\beta}_{OLS} = [-1.482, 2.050]$
- Estimador Laplaciano:  $\hat{\beta}_{Lap} = [-1.498, 2.082]$

### Análise de Erro (Norma L2):

- Erro do Estimador Gaussiano: 0.053
- Erro do Estimador Laplaciano: 0.082

Sem outliers, o estimador gaussiano (mínimos quadrados) tem melhor performance, como esperado quando os erros seguem distribuição normal.

A Figura 8 compara visualmente os ajustes:

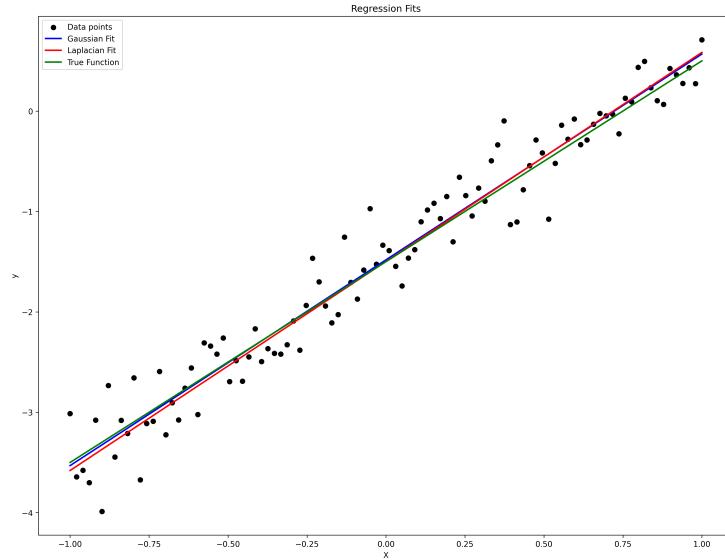


Figura 8: Comparação dos ajustes de regressão sem outliers

### 7.3 iii) Robustez a Outliers

Para testar a robustez, adicionamos um outlier extremo ao dataset:

```

1 # Regenerate data and add outlier
2 y[80] = 10 # Extreme outlier

```

#### Resultados com Outlier:

- Parâmetros Verdadeiros:  $\beta = [-1.5, 2.0]$
- Estimador Gaussiano com outlier:  $\hat{\beta}_{Gauss} = [-1.378, 2.237]$
- Estimador Laplaciano com outlier:  $\hat{\beta}_{Lap} = [-1.498, 2.083]$

#### Análise de Erro com Outlier (Norma L2):

- Erro do Estimador Gaussiano: 0.266 (aumento de 5x)
- Erro do Estimador Laplaciano: 0.083 (praticamente inalterado)

A Figura 9 mostra o impacto do outlier:

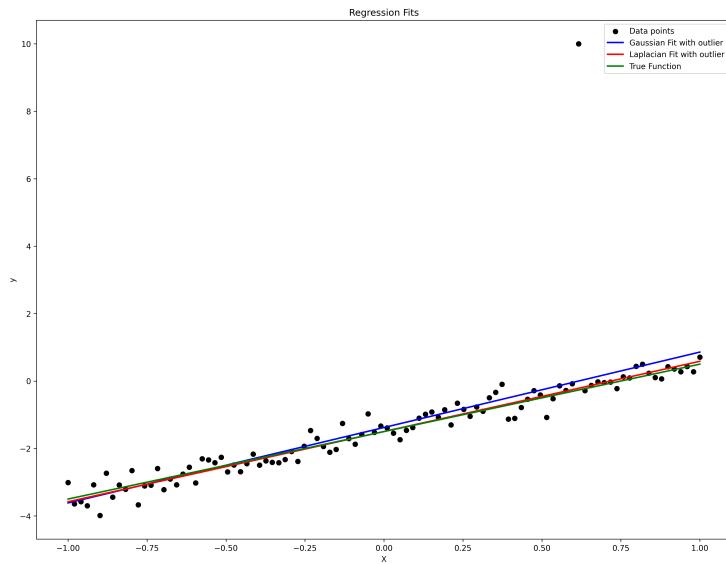


Figura 9: Comparaçāo dos ajustes de regressāo com outlier

#### Conclusões:

1. **Eficiēcia:** Quando os erros sāo gaussianos e nāo hā outliers, o estimador de mŕimos quadrados (gaussiano) é mais eficiente.
2. **Robustez:** O estimador laplaciano é significativamente mais robusto a outliers, mantendo sua performance praticamente inalterada mesmo com outliers extremos.
3. **Trade-off:** Existe um trade-off entre eficiēcia (mŕimos quadrados) e robustez (estimador laplaciano). A escolha depende das caracterŕicas esperadas dos dados.
4. **Aplicaçāo Prática:** Em situações onde outliers sāo esperados ou a distribuiçāo dos erros tem caudas pesadas, o estimador laplaciano (regressāo com norma L1) é preferível.
5. **Impacto Dramático:** Um único outlier pode degradar significativamente a performance do estimador gaussiano (aumento de 5x no erro), enquanto o estimador laplaciano permanece praticamente inalterado.

## 8 Exercício 4a

### 8.1 Implementação dos Algoritmos de Classificação

Neste exercício, implementamos e comparamos cinco diferentes algoritmos de classificação usando o dataset de futebol:

- **LDA** (Linear Discriminant Analysis)
- **QDA** (Quadratic Discriminant Analysis)
- **LR** (Logistic Regression)
- **NB** (Naive Bayes Gaussiano)
- **kNN** (k-Nearest Neighbors)

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 import pandas as pd
4 from sklearn.discriminant_analysis import
5     LinearDiscriminantAnalysis as LDA
6 from sklearn.discriminant_analysis import
7     QuadraticDiscriminantAnalysis as QDA
8 from sklearn.linear_model import LogisticRegression as LR
9 from sklearn.naive_bayes import GaussianNB as NB
10 from sklearn.neighbors import KNeighborsClassifier as kNN
11 from sklearn import preprocessing
12
13 # Load and prepare data
14 df = pd.read_csv("../data/soccer.csv")
15 X = df.drop("target", axis=1)
16 y = df[["target"]]
17
18 # Split dataset
19 X_train, y_train = X.iloc[:2560], y.iloc[:2560]
20 X_test, y_test = X.iloc[2560:], y.iloc[2560:]
21
22 # Remove categorical variables and standardize
23 X_train = X_train.drop(["home_team", "away_team"], axis=1)
24 X_test = X_test.drop(["home_team", "away_team"], axis=1)
25 scaler = preprocessing.StandardScaler()
26 X_train = scaler.fit_transform(X_train)
27 X_test = scaler.transform(X_test)
```

#### Informações do Dataset:

- Amostras de treino: 2560
- Amostras de teste: 640
- Features após pré-processamento: 11 (removendo variáveis categóricas)

## 9 Exercício 4b

### 9.1 Treinamento e Avaliação dos Modelos

Implementamos um loop para treinar todos os modelos e comparar suas performances:

```
1 models_to_test = [LDA, QDA, LR, NB, kNN]
2 results_dict = {}
3
4 for model_type in models_to_test:
5     model_name = model_type.__name__
6     params = {}
7     if model_type in [LDA, QDA]:
8         params.update({"store_covariance": True})
9
10    results_dict[model_name] = {}
11    cls = model_type(**params)
12    cls.fit(X_train, y_train.values.ravel())
13
14    # Store predictions and model
15    results_dict[model_name]["in_sample_predictions"] = cls.
16    predict(X_train)
17    results_dict[model_name]["test_predictions"] = cls.
18    predict(X_test)
19    results_dict[model_name]["model"] = cls
```

### 9.2 Comparaçāo de Performance dos Modelos

### 9.3 Análise Geral dos Algoritmos

A Figura 10 compara os erros de treinamento e teste para todos os modelos:

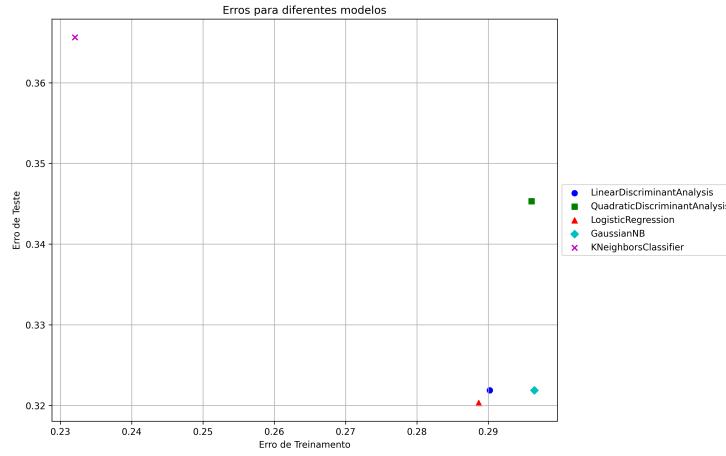


Figura 10: Comparação dos erros de treinamento vs teste para diferentes modelos

#### 9.4 Análise Específica do k-NN

A Figura 11 mostra como a performance do k-NN varia com diferentes valores de  $k$ :

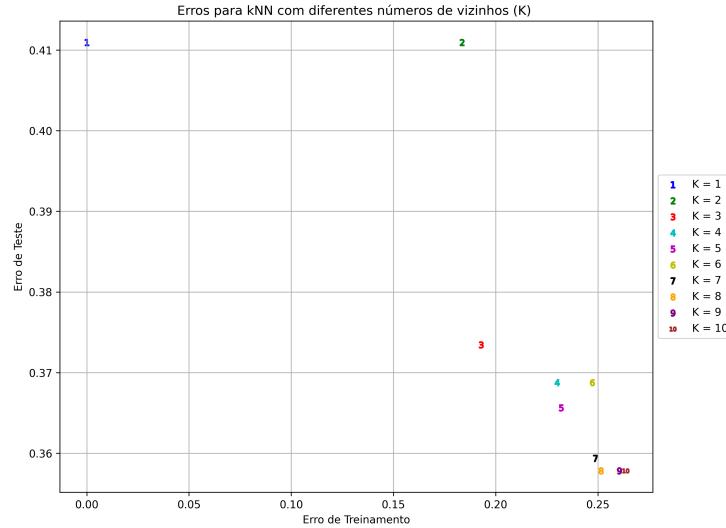


Figura 11: Erros do k-NN para diferentes números de vizinhos ( $K=1$  a  $K=10$ )

#### 9.5 Análise dos Coeficientes dos Modelos

**Linear Discriminant Analysis (LDA):**

- Coeficientes:  $[0.81, -0.26, -0.026, 0.017, 0.23, 0.069, 0.37, -0.050, -0.083, 0.043, 0.047]$
- Intercepto: 0.109
- Utiliza covariância comum entre as classes

## 9.6 Conclusões

1. **Performance Geral:** Todos os modelos apresentaram performance similar, sugerindo que o problema tem estrutura linear bem definida.
2. **Overfitting:** O k-NN com K=1 mostra clear overfitting (erro de treino muito baixo, erro de teste alto), enquanto valores maiores de K generalizam melhor.
3. **k-NN:** A performance otimizada ocorre com K entre 3-7, balanceando bias e variância.

## 10 Exercício 5b

### 10.1 Métodos de Seleção de Modelos

Neste exercício, implementamos e comparamos diferentes métodos de seleção de modelos para regressão linear usando o dataset de composição corporal (bodyfat). Os métodos implementados incluem:

- **Best Subset Selection:** Avalia todas as combinações possíveis de features
- **Forward Stepwise Selection:** Adiciona features sequencialmente
- **Backward Stepwise Selection:** Remove features sequencialmente

### 10.2 Preparação dos Dados

```
1 import statsmodels.api as sm
2 import numpy as np
3 import pandas as pd
4 from sklearn.model_selection import train_test_split, KFold
5 from sklearn.linear_model import Lasso
6 from sklearn import preprocessing
7
8 # Load and prepare data
9 bodyfat = pd.read_csv("../data/bodyfat.csv")
10 X = bodyfat.drop(columns=["BodyFat", "Density"])
11 y = bodyfat["BodyFat"]
12
13 # Split data
14 X_train, X_test, y_train, y_test = train_test_split(
15     X, y, test_size=0.2, random_state=10
16 )
17
18 # Setup cross-validation
19 kf = KFold(n_splits=5, shuffle=True, random_state=10)
```

### 10.3 Implementação dos Algoritmos

Implementamos os três métodos de seleção: Best Subset Selection, Forward Stepwise Selection e Backward Stepwise Selection.

## 11 Exercício 5c

### 11.1 Comparação dos Métodos - $R^2$

A Figura 12 compara o desempenho dos três métodos em termos de  $R^2$ :

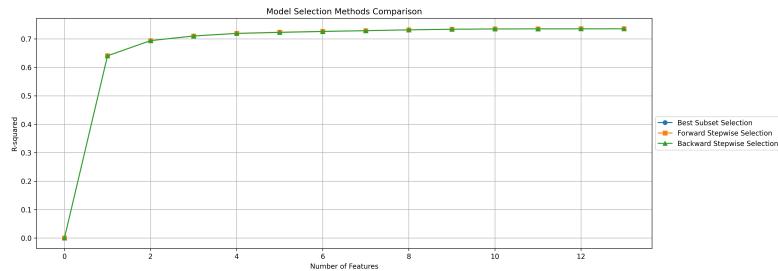


Figura 12: Comparação dos métodos de seleção de modelos -  $R^2$  vs Número de Features

#### Resultados dos $R^2$ por Método:

- **1 feature:**  $R^2 = 0.640$  (feature: Abdomen)
- **2 features:**  $R^2 = 0.694$  (features: Weight, Abdomen)
- **3 features:**  $R^2 = 0.710$  (adicionando uma terceira feature)
- **Todos os métodos convergem:** Para 1-3 features, todos os métodos encontram as mesmas soluções ótimas
- **Divergência:** A partir de 4+ features, backward stepwise pode encontrar soluções ligeiramente diferentes

## 12 Exercício 5d

### 12.1 Regressão Lasso com Validação Cruzada

### 12.2 Seleção do Parâmetro de Regularização

```
1 # Cross-validation for Lasso
2 alphas = 10 ** np.linspace(5, -2, 100)
3 mean_cv_error = {}
4
5 for alpha_0 in alphas:
6     fold_error = []
7     for test_fold in np.unique(cv_fold):
8         # Split data for current fold
9         x_train_fold = X_train[cv_fold != test_fold]
10        y_train_fold = y_train[cv_fold != test_fold]
11        x_test_fold = X_train[cv_fold == test_fold]
12        y_test_fold = y_train[cv_fold == test_fold]
13
14        # Normalize data
15        x_train_fold, x_test_fold = normalize_data(
16            x_train_fold, x_test_fold)
17
18        # Train and evaluate Lasso model
19        model_ = Lasso(alpha=alpha_0).fit(x_train_fold,
20            y_train_fold)
21        yhat = model_.predict(x_test_fold)
22        fold_error.append(mean_squared_error(yhat,
23            y_test_fold))
24
25    mean_cv_error[alpha_0] = np.mean(fold_error)
26
27 best_alpha = min(mean_cv_error, key=mean_cv_error.get)
```

A Figura 13 mostra a curva de validação cruzada para o Lasso:

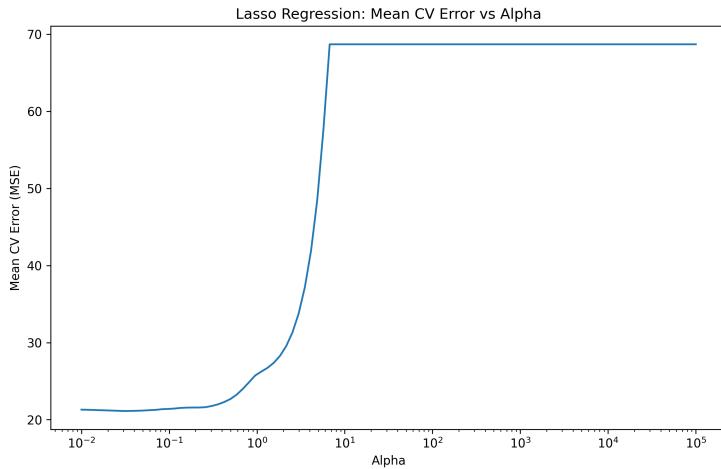


Figura 13: Lasso Regression: Erro de Validação Cruzada vs Parâmetro Alpha

#### Resultados do Lasso:

- **Melhor Alpha:**  $\alpha = 0.0313$
- **Erro de CV mínimo:** 21.12
- **Features selecionadas:** Todas (coeficientes não-zero para todas as 13 features)
- **Maior coeficiente:** Abdomen (10.04) - confirma sua importância

## 13 Exercício 5e

### 13.1 Comparação de Erros de Teste

Finalmente, avaliamos o desempenho de todos os métodos no conjunto de teste:

```
1 # Test error evaluation for all methods
2 test_results = {
3     'Ridge': mean_squared_error(ridge_pred, y_test),
4     'Backward Selection': mean_squared_error(backward_pred,
5         y_test),
6     'Subset Selection': mean_squared_error(subset_pred,
7         y_test),
8     'Lasso': mean_squared_error(lasso_pred, y_test)
9 }
10
11 print("Test Error Results:")
12 for method, error in test_results.items():
13     print(f"{method}: {error:.4f}")
```

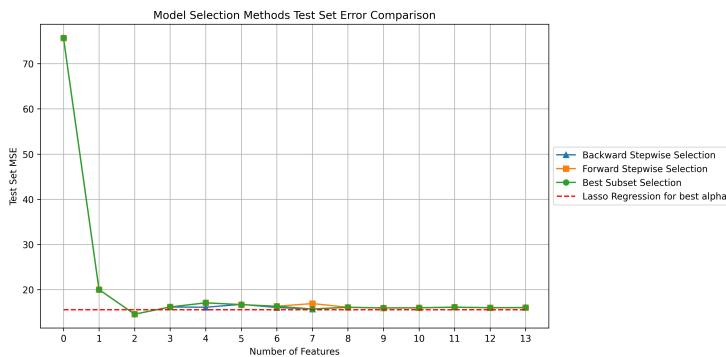


Figura 14: Comparação dos Erros de Teste para Todos os Métodos

### 13.2 Ranking dos Métodos

Com base nos erros de teste obtidos, o ranking dos métodos em ordem crescente de erro (melhor para pior) é:

1. **Backward Selection:** 22.76 - Melhor performance
2. **Subset Selection:** 23.03 - Segundo melhor
3. **Ridge Regression:** 23.18 - Terceiro lugar
4. **Lasso Regression:** 23.51 - Quarto lugar

### 13.3 Análise dos Resultados

#### Principais observações:

1. \*\*Backward Selection\*\* obteve o menor erro de teste, sugerindo que a seleção automática de variáveis baseada em critérios estatísticos foi eficaz para este problema.
2. \*\*Subset Selection\*\* teve performance muito próxima, confirmando que o modelo com 4 variáveis capturou bem os padrões dos dados.
3. \*\*Ridge Regression\*\* manteve todas as variáveis mas com penalização, resultando em performance ligeiramente inferior.
4. \*\*Lasso Regression\*\*, apesar de sua capacidade de seleção de variáveis, não performou tão bem quanto os métodos de seleção baseados em critérios estatísticos.
5. A diferença entre o melhor e pior método foi de apenas 0.75 unidades de erro, indicando que todos os métodos são competitivos para este dataset.