

Detecção de *Outliers* e Observações Influentes em Modelos de Regressão Linear Simples

Zaudir Dal Cortivo¹
UNICENTRO, Irati, PR

Resumo. Diversos são os métodos estatísticos para detecção de *outliers* para a regressão linear. Determinar qual ou quais observações tem maior influência na estimativa dos parâmetros é fundamental para obtenção de um modelo confiável e não tendencioso. A detecção de *outliers* visa identificar quais são tais pontos, a fim de melhorar a análise dos dados e obtenção de um modelo com menor erro residual. Normalmente são utilizados mais de um método para identificá-los, como métodos gráficos, distância de Cook, DFFITS, DFBETAS e \hat{h}_{ii} , entre outros métodos. Neste trabalho, avaliou-se a utilização da razão dos coeficientes de determinação combinado com a análise gráfica para identificar e classificar os pontos discrepantes e influentes. Alguns conjuntos de dados para regressão linear simples foram utilizados para a análise dos dados e comparação dos resultados. O método mostrou-se ser bastante eficiente na identificação destes pontos, além de identificar os *outliers*, verifica quais são as observações influentes.

Palavra-chave: Regressão linear simples, *outliers*, pontos influentes.

1 Introdução

Na regressão linear simples, valores discrepantes podem comprometer a confiabilidade do modelo obtido e desta forma é importante que sejam identificados no decurso da construção do modelo. Estas observações podem ocorrer devido a um erro de registro, um erro no procedimento experimental ou pode representar um caso raro, entre outras possibilidades [5]. Os pontos discrepantes são comumente denominados como outliers e distinguimos em dois tipos. Um outlier é um ponto do conjunto de observações cuja resposta y não segue a tendência geral do restante dos dados. Por outro lado, se uma observação que tem valores discrepantes do preditor x é denominada ponto de alavancagem, que pode ser um valor particularmente alto ou baixo de x . As observações classificadas como alavancagem não necessariamente são outliers. Um ponto de alavancagem é ruim, se também for um outlier, ou seja, o valor de y não segue o padrão definido pelos demais pontos de dados. Um ponto de alavancagem é bom, se não for um valor influente para o modelo [1]. Uma observação é influente se influencia indevidamente qualquer parte da análise de regressão, na determinação dos coeficientes ou a análise diagnóstica da regressão. Valores atípicos e pontos de alta alavancagem têm o potencial de serem influentes, mas geralmente temos que investigar se tal observação pode ser classificada como influente.

Diversos métodos tem sido propostos para detectar valores discrepantes na regressão linear, mas classificar uma ou mais observações como outlier pode ser uma tarefa difícil, especialmente quando há vários valores discrepantes no conjunto de dados [3] e [5]. O mascaramento e swamping podem dificultar a detecção destes pontos. O mascaramento ocorre quando um valor discrepante obscurece a existência de outro, enquanto o swamping ocorre quando um valor não discrepante é incluído

¹zaudir@unicentro.br

erroneamente em um grupo de observações consideradas como outliers. Os outliers poderiam ser identificados observando-se apenas os resíduos de mínimos quadrados, mas quando estes pontos são de alavancagem, os pontos influentes podem permanecer ocultos, pois não aparecem no gráfico residual.

Os Métodos gráficos podem ser bastante úteis quando o número é limitado de observações suspeitas. Pode-se utilizar simplesmente o gráfico de dispersão, gráfico de resíduos padronizados vs índice, qqplot, resíduos padronizados vs valores ajustados, entre outros. No software R, é possível utilizar vários tipos de gráficos com os pacotes ggplot e ggplot2. Os métodos utilizados para detecção de outliers, destacamos os mais conhecidos: Resíduos estudantizados, distância robusta, distância de Cook, \hat{h}_{ii} , DFFITS, DFBETAS, razão de covariância, entre outros. Para medir se uma observação é influente ou não, pode-se utilizar o coeficiente de determinação (R^2). Esta métrica representa a proporção total da amostra explicada pelo modelo e quantifica o quão bem a reta de regressão se ajusta aos dados [2]. Os valores de R^2 pertencem ao intervalo $[0, 1]$ e quanto mais próximo de 1, mais explicativo é o modelo. Neste trabalho, utilizamos a razão do coeficiente de determinação conjunta com a análise gráfica, para determinar se uma observação um *outlier* e se é influente para o modelo de regressão linear simples.

2 Revisão de Métodos

O modelo de regressão linear simples consiste na função média e na função variância

$$E(Y|X = x) = \beta_0 + \beta_1 x \text{ e } Var(Y|X = x) = \sigma^2. \quad (1)$$

Os parâmetros são o intercepto β_0 , e β_1 , que é a taxa de variação da função $E(Y|X = x)$. A relação entre X e Y é linear, os valores de X são fixos (ou controlados) e a função de variância é assumidamente constante, com valor positivo para σ^2 , pois geralmente é desconhecido. O erro estatístico ou resíduo (e_i) é a diferença entre o valor observado da i -ésima resposta y_i e o valor esperado. Os erros e_i dependem de parâmetros desconhecidos na função média e portanto não são quantidades observáveis. Pressupomos que a média do erro é nula $E(e_i|x_i) = 0$, que o erro de uma observação é independente do erro da outra observação, isto é, $Cov(e_i, e_j) = 0, i \neq j$ e os erros têm distribuição normal.

2.1 Alavancagem

A matriz chapéu tem vários usos na regressão e uma delas é que fornece uma medida de alavancagem. É útil para investigar se uma ou mais observações são influenciadas pelos valores de X e, portanto, podem estar influenciando excessivamente os resultados da regressão [1]. Para identificar os pontos de alavancagem pode-se utilizar a diagonal da matriz chapéu, dado por:

$$H = X(X'X)^{-1}X'. \quad (2)$$

A alavancagem \hat{h}_{ii} , que tem valores no intervalo $[0, 1]$, quantifica a influência que a resposta observada sobre o valor previsto e é uma medida da distância entre o valor x e a média dos valores x para todos os n pontos de dados. Se \hat{h}_{ii} for grande, a resposta observada desempenha relevância no valor da variável resposta prevista. Se o valor de x tem grande distância, a alavancagem será grande; caso contrário não. Geralmente, um ponto com $\hat{h}_{ii} > 3k/n$ deve ser cuidadosamente examinado, onde k é o número de preditores e n o número de observações.

2.2 Distância de Cook

A distância de Cook para a i -ésima observação é baseada nas diferenças entre as respostas previstas do modelo construído a partir de todos os dados e as respostas previstas do modelo construído, deixando de lado a i -ésima observação. A distância de Cook para a observação i é:

$$CD_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{k \cdot MSE} = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{(k\sigma_{j=1}^n e_j^2/n)} = \frac{e_i^2}{k \cdot MSE} \cdot \frac{\hat{h}_{ii}}{(1 - \hat{h}_{ii})^2}, \quad (3)$$

onde \hat{y}_j é o j -ésimo valor de resposta ajustado, $\hat{y}_{j(i)}$ é o j -ésimo valor da resposta ajustada, onde o ajuste não inclui a observação i , MSE é o erro quadrático médio, e_i é o i -ésimo resíduo e \hat{h}_{ii} é o i -ésimo valor de alavancagem e k é o número de coeficientes no modelo de regressão. Dois critérios para o ponto de corte, em um nível de significância α , são: $CD_i \leq F(1 - \alpha; k, n - k)$ e $CD_i \geq 4/(n - (k + 1))$ que indica que a i -ésima observação é influente.

2.3 DFFITS

A medida DFFITS é semelhante a distâncias de Cook. O numerador mede a diferença nas respostas previstas com e sem a observação i . A diferença nos ajustes para observação i , é definida como:

$$DFFITS_i = \frac{\hat{y}_j - \hat{y}_{(i)}}{\sqrt{MSE_{(i)} \cdot \hat{h}_{ii}}}, \quad (4)$$

onde \hat{y}_j e $\hat{y}_{(i)}$ são as previsões para o ponto i com e sem o ponto i incluído na regressão. Uma observação é considerada influente se o valor absoluto do seu valor DFFITS for maior que $2\sqrt{k/n}$, onde n é o número de observações e k o número de parâmetros incluindo o intercepto.

2.4 DFBETAS

A DFBETA é uma métrica útil para avaliar a influência de uma determinada observação e mede a diferença na estimativa de cada parâmetro com e sem o ponto influente. A padronização dos valores de DFBETA divide-se cada DFBETA pelo erro padrão, denominada DFBETAS. Uma medida escalonada da mudança pode ser definida como

$$DFBETA_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} \cdot C_{ij}}}, \quad (5)$$

onde $\hat{\beta}_j$ é o coeficiente da regressão calculado usando todos os dados e $\hat{\beta}_{j(i)}$ é o coeficiente da regressão calculado sem a observação, $MSE_{(i)}$ é o erro quadrático médio da regressão calculada sem a i -ésima observação e C_{ij} é o j -ésimo elemento diagonal da matriz da regressão $(X'X)^{-1}$, calculada com todas as observações. Valor de corte é $\pm \frac{2}{\sqrt{n}}$.

2.5 Razão de Covariância

A razão de covariância é definida por:

$$CVR_i = \left(\frac{\hat{\sigma}_{(i)}^2}{\hat{\sigma}^2}\right)^k \cdot \frac{1}{1 - \hat{h}_{ii}}, \quad (6)$$

onde $\hat{\sigma}_{(i)}^2$ é a variância estimada sem a observação i e $\hat{\sigma}^2$ é a variância estimada com as n observações. O desvio da unidade indica que a i -ésima observação é potencialmente influente. Uma observação é considerada influente se $|CVR_i - 1| > 3k/n$.

2.6 Razão do Coeficiente de Determinação

A razão do coeficiente de determinação (CDR) é uma métrica proposta por [4]. A medida é determinada pelo quociente do coeficiente de determinação com todos os elementos do conjunto, com o coeficiente de determinação calculado sem a observação j . Dado que SST e SSR são respectivamente a soma dos quadrados totais e que SSR é a soma dos quadrados dos resíduos, tem-se que:

$$CDR_i = \frac{R^2_{(i)}}{R^2} = \frac{SST}{SST_{(i)}} \cdot \frac{SSR_{(i)}}{SSR}. \quad (7)$$

Uma observação é considerada influente se o valor do CDR_i for significativamente diferente dos demais. Para [4] o ponto de corte é dado pelo quantil superior da distribuição Beta em um nível de significância α [6].

3 Aplicação para Conjunto de Dados

Nesta seção, ilustramos o uso da medida CDR para detecção de outliers. Para isto, foram utilizados 5 conjunto de dados, denominados Dados 1, Cars, Forbes, Trees e Davis. Os conjuntos Cars, Forbes, Trees e Davis estão disponíveis no software R. O conjunto Dados 1 foi subdividido em mais 3 conjuntos: Dados 1a um conjunto sem observações influentes, Dados 1b com um outlier mas não influente, Dados 1c com um único outlier, observação influente e Dados 1d com 4 outliers, os 4 influentes. Em todos os conjuntos foram verificados os pressupostos da regressão linear simples: teste de normalidade, homogeneidade, independência e que a média dos resíduos é nula. Os resultados foram obtidos através do software R, para a visualização, os modelos são ajustados e destacados nos gráficos: *Residuals* \times *Fitted*, gráfico QQ Normal, *Scale* \times *Location* e Resíduos \times Alavancagem. Alguns pontos discrepantes podem ser observados nestes gráficos. O primeiro passo será avaliar o gráfico de cada conjunto, para observar os pontos candidatos a pontos discrepantes. Na figura 1, temos os gráficos dos conjuntos Dados 1a e 1b. Para o conjuntos 1a pode-se observar que as observações 4 e 14 com maiores resíduos, indicando ser possíveis valores discrepantes. Para o conjunto Dados 1b, indicam que a observação 21 é suspeita de ser um ponto discrepante.

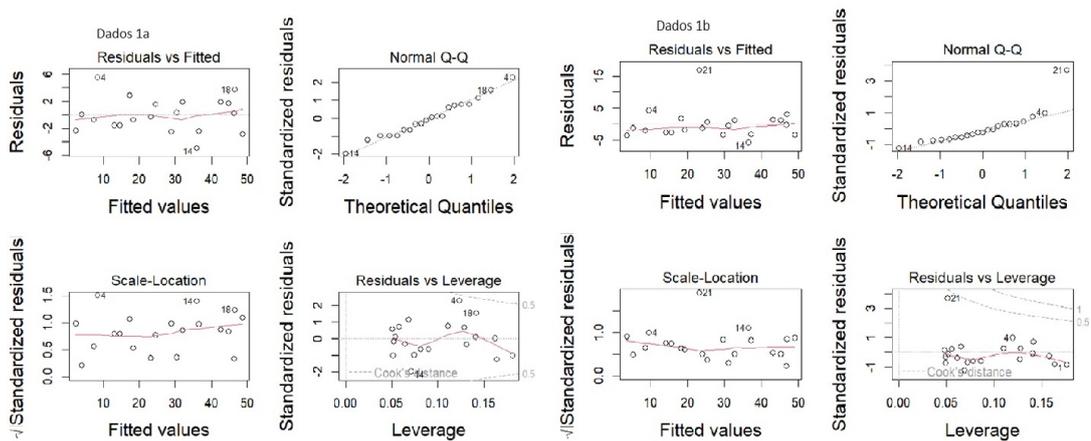


Figura 1: Gráficos dos conjuntos Dados 1a e 1b. Fonte: O autor (2024).

Os gráficos apresentados na figura 2, são para as observações dos conjuntos Dados 1c e 1d. Para o conjuntos Dados 1c, observa-se que a observação 21 tem alto resíduo e pode ser um um ponto

influyente. Utilizando a medida distância de Cook, está observação é indicada como outlier. Para o conjunto dados 1d, observamos que há um conjunto de ponto candidatos a serem classificados como outliers, os pontos 21, 22, 23. No gráfico *Fitted values* × *Residuals*, estes pontos "puxam" a linha de ajusta para baixo, também sugerindo que temos 4 pontos com altos resíduos e que alteram a direção da curva de ajuste.

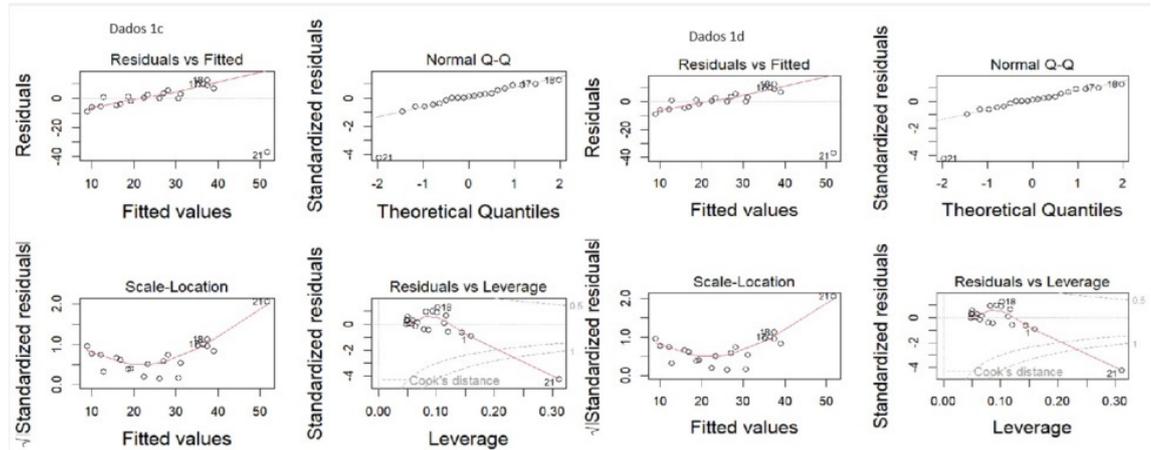


Figura 2: Gráficos dos conjuntos Dados 1c e 1d. Fonte: O autor (2024).

Na figura 3, são apresentados os gráficos dos conjuntos Cars e Forbes. Para o conjunto Cars, observa-se que as observações 23, 35 e 49, são os pontos indicados como possíveis outliers. No gráfico *Residuals* × *Leverage* é indicado apenas o 39. No conjunto Forbes, apenas a observação 12 é suspeita de ser um outlier.

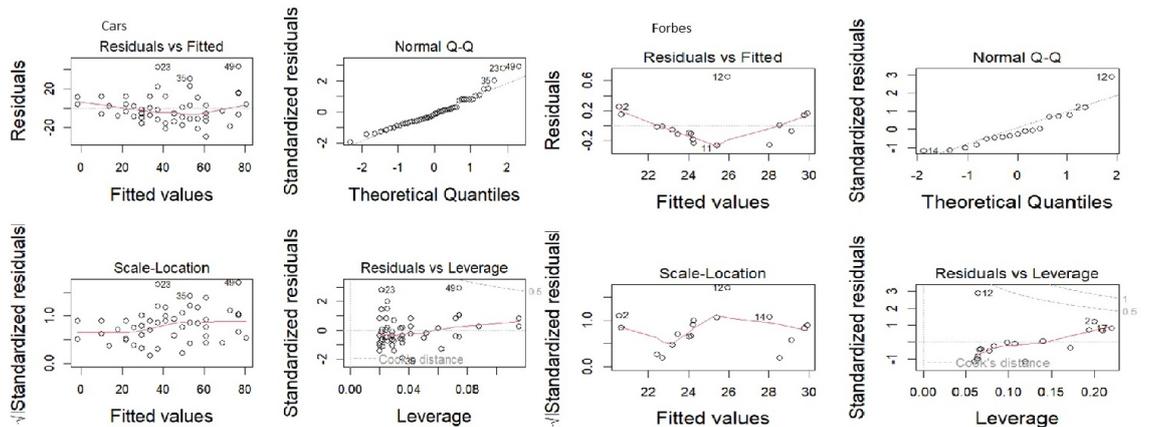


Figura 3: Gráficos dos conjuntos Cars e Forbes. Fonte: O autor (2024).

Na figura 4, são apresentados os gráficos dos conjuntos Trees e Davis. Para o conjunto Trees, as observações 18 e 20 aparecem mais distantes dos demais pontos, embora a análise destes gráficos, não trazem esta afirmação com contudência. No conjunto Davis, a observação 12 é a discrepante em relação aos demais dados. Retirando a observação 12, poderemos verificar se há mais pontos

candidatos a *outlier*. Mas pode-se afirmar que o ponto 12 afeta a precisão do ajuste da reta, assim uma ou mais medida de corte.

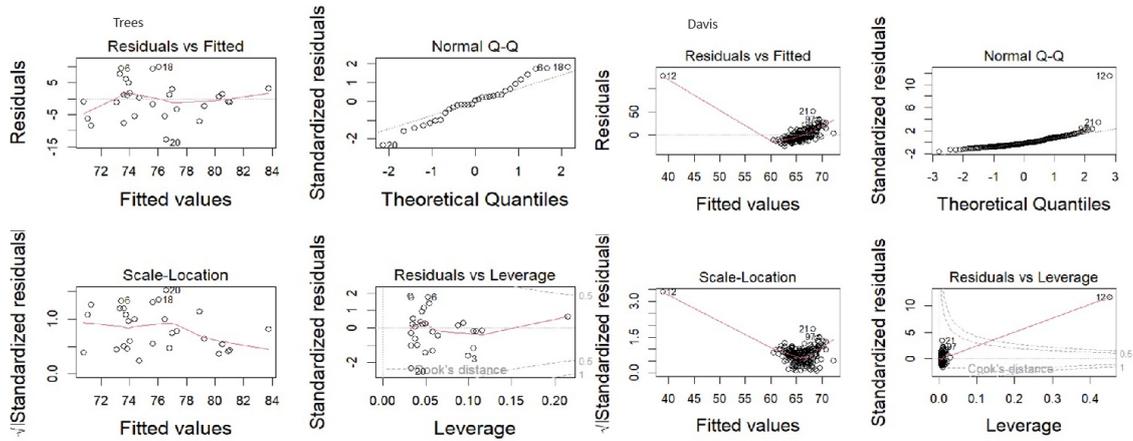


Figura 4: Gráficos dos conjuntos Trees e Davis. Fonte: O autor (2024).

Na tabela 1, temos os resultados para o conjunto Dados 1, utilizando as métricas CDR, DFBE-TAS, DFFIT, COVR, Cook e \hat{h}_{ii} . Comparados os resultados obtidos, temos que o CDR apresentou melhor desempenho que as demais. Apenas para o conjunto Dados 1c, os resultados foram iguais. Para o conjunto Dados 1a foram identificados 4 observações como outliers, porém nenhuma destas observações é influente. Já para Dados 1b, apenas uma observação foi indicada como outlier, mas o CDR indica que não. Todos os valores do CDR, para estes conjuntos foram semelhantes. Para o conjunto Dados 1d, houve o mascaramento. A observação 21, escondeu a influência das observações 22, 23 e 24, pois estes 4 pontos tiveram CDR semelhantes e significativamente diferente dos demais. Em comparação com que foi observado nos gráficos, pode-se concluir que a análise gráfica pode ajudar na indicação dos pontos que precisam ser classificados, mas sempre há a necessidade de usar uma métrica para a classificação. Deve-se ter o cuidado com a remoção de uma ou mais observação, pois estas observações podem conter informações relevantes para o modelo.

Tabela 1: Resultados 1. Fonte: O autor (2024).

1	Dados 1 a	Dados 1 b	Dados 1 c	Dados 1 d
n	20	21	21	24
DFBETAS, DFFIT, COVR, Cook e \hat{h}_{ii}	1, 2, 4 e 18	21	21	21
CDR	Nenhuma	Nenhuma	21	21, 22, 23, 24

Na tabela 2, para o conjunto Cars, o CDR teve valores no intervalo $[0,96; 1,11]$, isto é, os valores são semelhantes. Foram retirada as observações 23, 49, 23, 35, 50, 23, 35, 49 e 01, 02, 23, 49, 50, mas os valores do CDR permaneceram no intervalo $[0,96; 1,11]$, o que indica que estes pontos não são influentes. Para os conjuntos Forbes, todos os pontos obtiveram valores CDR próximos de 1, inclusive o $CDR_{12} = 1,003$, logo este ponto não é influente. Para o conjunto Trees, também os CDRs obtiveram valores semelhantes próximos a 1. Nem a retirada dos pontos 1, 20 e 21, indicados pelas outras medidas, alterou o resultado. Para este conjunto foi retirado os pontos 6, 18 e 20. por possuírem os maiores CDRs, e o resultado foi significativo, com grande aumento do Coeficiente de determinação R^2 . Para o conjunto Davis, apenas os CDRs das observações 12 e

21 foram significativamente diferentes, com valores $CDR_{12} = 16,51$ e $CDR_{12} = 17,10$, enquanto os demais estão próximos de 1. Assim podemos afirmar que apenas as observações 12 e 21 são influentes.

Tabela 2: Resultados 2. Fonte: O autor (2024).

2	Cars	Forbes	Trees	Davis
n	50	17	31	200
DFBETAS, ...	1, 2, 23, 49 e 50	12	1, 20 e 31	12, 19 e 21
CDR	Nenhuma	Nenhuma	21	12 e 21

3.1 Discussão dos Resultados

Os resultados da implementação do CDR para 8 conjuntos de dados são apresentados. Avaliou-se o desempenho da medida CDR na detecção de outliers para ao modelo de regressão linear simples, comparando com as medidas DFBETAS, DFFITS, COVR, Cook e hii. Os resultados mostram que a medida CDR foi bem-sucedida em identificar valores discrepantes e influentes nos conjuntos de dados. Enquanto que, demais medidas não foram tão eficientes, principalmente para indicar se a observação é influente. Para estudos futuros, pode-se fazer a avaliação das características de cada medida com mais profundidade e comparar com o CDR, também o uso da medida de corte proposta por [6].

Referências

- [1] D. Blatná. **Site da Semantic Scholar**. Online. Acessado em 08/01/2024, <https://www.semanticscholar.org/paper/OUTLIERS-IN-REGRESSION>.
- [2] J. A. Cornell e D. Berger. “Factors that Influence the Value of the Coefficient of Determination in Simple Linear and Nonlinear Regression Models”. Em: **The American Phytopathological Society** 77 (1987), pp. 63–70.
- [3] N. Dervilis et al. “Robust methods for outlier detection e regression for SHM applications”. Em: **Int. J. Sustainable Materials and Structural Systems** 1 e 2 (2015).
- [4] M. O. Oyeyemi G. M. Oluwaseun O. B. e Adeleke. “Comparisons of Some Outlier Detection Methods in Linear Regression Model”. Em: **Ilorin Journal of Science** 1 (2017), pp. 130–138.
- [5] A. Zakaria, N. K. Howard e B. K. Nkansah. “On the detection of influential outliers in linear regression analysis”. Em: **American Journal of Theoretical and Applied Statistics** 4 (2014), pp. 100–106. DOI: 10.11648/j.ajtas.20140304.14.
- [6] A. Zakaria e B. K. Nkansah. “On the Coefficient of Determination Ratio for Detecting Influential Outliers in Linear Regression Analysis”. Em: **American Journal of Theoretical and Applied Statistics** 11 (2022), pp. 27–35. DOI: 10.11648/j.ajtas.20221101.14.