UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”

Instituto de Biociências de Botucatu

Curso de Física Médica

MODELAGEM EM REGRESSÃO LOGÍSTICA NA ÁREA DA SAÚDE: ESTUDO DE CASOS DA LITERATURA E SIMULAÇÃO COMPUTACIONAL

Luíza Manso

Orientador(a): Profª. Drª. Miriam Harumi Tsunemi

Botucatu – SP
2025

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
Instituto de Biociências de Botucatu

Curso de Física Médica

MODELAGEM EM REGRESSÃO LOGÍSTICA NA ÁREA DA SAÚDE: ESTUDO DE CASOS DA LITERATURA E SIMULAÇÃO COMPUTACIONAL

Luíza Manso

Trabalho de Conclusão de Curso apresentado à Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Instituto de Biociências de Botucatu, como parte dos requisitos para obtenção do título de Bacharel em Física Médica.

Orientador: Profª. Drª. Miriam Harumi Tsunemi

Botucatu – SP
2025

M289m

Manso, Luíza
Modelagem em regressão logística na área da saúde: : estudo de casos da literatura e simulação computacional / Luíza Manso. -- Botucatu, 2025
56 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Física Médica) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu
Orientadora: Miriam Harumi Tsunemi

1. Análise de regressão logística. 2. Análise de regressão. 3. Simulação de dados. 4. Física médica. 5. Estatística. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

LUÍZA MANSO

Modelagem em Regressão Logística na Área da Saúde: Estudo de Casos da Literatura e Simulação Computacional

Trabalho de Conclusão de Curso, apresentado a Universidade Estadual Paulista, como parte das exigências para a obtenção do título de Bacharel, do curso de Graduação em Física Médica.

Botucatu, 17 de novembro de 2025.

BANCA EXAMINADORA

Profa. Miriam Harumi Tsunemi
Instituto de Biociências de Botucatu

Profa. Daniela Renata Cantane
Instituto de Biociências de Botucatu

Prof. Rogerio Antônio de Oliveira
Instituto de Biociências de Botucatu

Resumo

A análise de informações amostrais é de extrema importância na tomada de decisão nas diversas áreas do conhecimento. Na área da saúde, auxilia, de forma sistemática, na avaliação de processos, intervenções e desenvolvimento de produtos para a melhoria da qualidade de vida da população. Nesse contexto, a Física Médica tem se destacado como área estratégica para a análise quantitativa de dados em saúde, especialmente em cenários nos quais ainda são escassos estudos que explorem sistematicamente as informações geradas e extraiam conclusões a partir delas. Uma das maneiras de analisar esses dados é através dos modelos de regressão, que têm como objetivo avaliar a relação entre duas variáveis quantitativas, por exemplo, a taxa de vacinação com a mortalidade de uma população, entre outros. Por outro lado, se a variável dependente é binária, deve-se utilizar o modelo de regressão logística com foco na probabilidade de uma determinada característica de interesse. Devido à sua natureza de resposta binária, o modelo de regressão logística tem se destacado na área de diagnóstico por imagem, pois permite, a partir dos parâmetros extraídos das imagens, gerar predições que auxiliam os médicos na tomada de decisão clínica. Além disso, o desempenho de modelos estatísticos usualmente é avaliado por meio de dados simulados. As simulações levam em consideração a dinâmica do processo biológico e permite um estudo aprofundado do problema com menor custo. Programas computacionais são utilizados para estas simulações tais como o programa estatístico R. Este programa é disponível gratuitamente no site https://www.r-project.org/ e possui a colaboração voluntária de pesquisadores de todo o mundo. Dessa forma, este programa encontra-se em constante desenvolvimento, sendo incorporados os métodos de análise de dados mais recentes, por meio de bibliotecas e pacotes. O objetivo deste projeto é capacitação na modelagem de informações experimentais de pesquisas envolvendo modelos de regressão logística por meio do estudo de dados da literatura e simulação realizados no programa estatístico R, com foco em demonstrar como os modelos de regressão logística podem ser utilizados para gerar um diagnóstico a partir da simulação de parâmetros de imagens médicas.

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todos que, de alguma forma, contribuíram para a realização deste relatório. Em primeiro lugar, gostaria de agradecer à Universidade Estadual Paulista "Júlio de Mesquita Filho"pelo ambiente acadêmico e pelos recursos disponibilizados para o desenvolvimento desse estudo. Agradecer também ao Instituto de Biociências de Botucatu pela infraestrutura durante a realização do trabalho e de toda minha graduação.

Em especial, agradeço o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro proveniente do programa PIBIC, que possibilitou a dedicação necessária para o estudo e a pesquisa aqui apresentados.

Expresso minha imensa gratidão à minha orientadora, Professora Doutora Miriam Harumi Tsunemi, que durante esse período juntas foi uma excelente orientadora, professora e sempre me tratou com tanto carinho. Tudo isso foi fundamental em todas as etapas do trabalho, me inspirando e motivando o meu crescimento acadêmico e pessoal.

Agradeço ao meu pai, que é meu maior exemplo de dedicação e sacrifício, sempre colocando a família em primeiro lugar e me ensinando que com esforço e dedicação, eu sou capaz de aprender qualquer coisa. À minha mãe, sem ela, não saberia como aproveitar a vida da melhor forma e, independentemente de tudo, sempre olhar para a vida com um sorriso no rosto.

Agradeço muito a toda minha família pelo apoio, carinho e incentivo constantes, que foram fundamentais para a realização deste trabalho e para o meu crescimento pessoal. Em especial, a minha vó Regina, que nos deixou esse ano; esse trabalho é para você.

Sou grata aos meus amigos pelo companheirismo, suporte e palavras de incentivo que tornaram esse percurso mais leve e prazeroso. Em especial à minha amiga Laura, que durante esses 5 anos de graduação, foi a minha maior dupla durante todo esse processo, estando sempre presente nos momentos bons e ruins.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, seja com ideias, críticas construtivas ou apoio moral. Tudo isso fez com que este relatório pudesse ser finalizado com qualidade e satisfação.

Lista de Figuras

Lista de Tabelas

Sumário

Introdução

A análise de informações amostrais é de extrema importância na tomada de decisão nas diversas áreas do conhecimento. Na área da saúde, auxilia, de forma sistemática, na avaliação de processos, intervenções e desenvolvimento de produtos para a melhoria da qualidade de vida da população. Nesse contexto, a Física Médica tem se consolidado como área estratégica para a análise quantitativa de dados em saúde, sobretudo no diagnóstico por imagem [1], em que parâmetros extraídos das imagens podem ser utilizados em modelos estatísticos para apoiar a decisão clínica. Mais especificamente, os modelos de regressão têm como objetivo avaliar a relação entre duas variáveis quantitativas, por exemplo, a taxa de vacinação com a mortalidade de uma população, entre outros. Uma delas é denominada de variável resposta (ou dependente) e a outra covariável (independente). Para o ajuste deste tipo de modelo de regressão, há a necessidade de atendimento de algumas suposições tais como a variável resposta ser quantitativa, normalidade e homoscedasticidade dos resíduos, os quais limitam a sua aplicabilidade.

Quando a suposição de normalidade dos resíduos não é atendida pode-se utilizar uma classe de modelos de regressão mais abrangente denominada modelos lineares generalizados (MLG). Neste caso a variável resposta pode assumir outras distribuições tais como Binomial, Poisson, Gama, entre outros. A distribuição Poisson pode ser utilizada em dados de contagem, por exemplo, quando a variável resposta é o número de casos de dengue em um município em um ano. No caso de tempo até reação de um agente químico, a distribuição Gama pode ser uma candidata por ser uma variável estritamente positiva. Por fim, se a variável resposta é binária, deve-se utilizar o a distribuição Binomial através do modelo de regressão logístico com foco na probabilidade de uma determinada característica de interesse [2]. Em diagnóstico por imagem, por exemplo, a regressão logística permite modelar a probabilidade de um tumor ser maligno a partir de parâmetros quantitativos obtidos das imagens, contribuindo para o diagnóstico médico e, consequentemente, reduzindo procedimentos invasivos nos pacientes.

A escolha do modelo de regressão mais adequado depende também da análise do diagnóstico do modelo ajustado. Esta análise permite a identificação de pontos suspeitos que afetam de forma desproporcional nas estimativas dos parâmetros.

Além disso, o desempenho de modelos estatísticos usualmente é avaliado por meio de dados simulados. As simulações levam em consideração a dinâmica do processo biológico e permitem um estudo aprofundado do problema com menor custo. No contexto deste trabalho, foram realizadas simulações nas áreas da Biologia e Física Médica, incluindo dose absorvida em relação à intensidade de radiação incidente, marcador biológico binário e espessura do tecido; internação dependendo da idade e marcador biológico; e principalmente parâmetros extraídos de imagens médicas (como vascularidade tumoral) com desfecho binário (benigno/maligno), simulando cenários reais de diagnóstico.

Programas computacionais são utilizados para estas simulações tais como o programa estatístico R. Este programa é disponível gratuitamente no site https://www.r-project.org/ e possui a colaboração voluntária de pesquisadores de todo o mundo. Dessa forma, este programa encontra-se em constante desenvolvimento, sendo incorporados os métodos de análise de dados

mais recentes, por meio de bibliotecas e pacotes.

O presente estudo tem como foco a análise e aplicação de modelos de regressão, com ênfase na regressão logística, em problemas da área da saúde, especialmente em cenários de diagnóstico por imagem em Física Médica, utilizando dados da literatura e simulações computacionais no programa estatístico R. Busca-se compreender os contextos em que os modelos de regressão linear e logístico são mais adequados, explorando seus pressupostos teóricos e o processo de inferência dos parâmetros. Além disso, pretende-se realizar simulações de dados para investigar o comportamento dos ajustes desses modelos em diferentes situações experimentais. Dessa forma, o projeto visa à capacitação na modelagem de informações experimentais e à consolidação do conhecimento sobre o uso da regressão logística para apoiar o diagnóstico a partir de parâmetros extraídos de imagens médicas.

Desenvolvimento

Neste capítulo serão apresentados modelos de regressão. Primeiramente, será abordado um modelo clássico, a Regressão Múltipla. Em seguida, serão apresentados os Modelos Lineares Generalizados, com ênfase no Modelo Logístico.

2.1 Regressão Múltipla

Antes de compreender os conceitos da regressão múltipla, é essencial entender os princípios da regressão linear simples. Este modelo serve de base para modelos mais elaborados e é crucial pra entender conceitos mais avançados.

A regressão linear simples busca estudar a relação linear entre duas variáveis: uma variável dependente (resposta) e uma variável independente (explicativa). Com isso, assume-se que a variável dependente varia linearmente com a variável independente, com a adição de um termo de erro.

O modelo é representado pela equação [3]:

$$y = \alpha + \beta x + \varepsilon, \tag{1}$$

em que y e x representam as variáveis dependente e independente, respectivamente. O coeficiente α representa o intercepto (valor de $\hat{y}$ quando $x = 0$), o coeficiente β representa a inclinação da reta (mudança em y para uma unidade de mudança em x) e ε representa o termo de erro.

Os parâmetros α e β são estimados utilizando dados amostrais do método dos mínimos quadrados, que busca minimizar a soma dos quadrados dos resíduos. Esses parâmetros serão mais explorados no decorrer do capítulo.

Para que as estimativas dos parâmetros sejam válidas e as inferências estatísticas sejam confiáveis, a regressão linear simples assume alguns pressupostos importantes [3]:

- Linearidade: a relação entre a variável dependente e a independente é linear;
- Independência: os erros são independentes entre si;
- Homocedasticidade: a variância dos erros é constante para todos os valores da variável independente;
- Normalidade: os erros seguem uma distribuição normal.

Esses pressupostos também valem para a regressão múltipla e serão retomados adiante.

A qualidade do ajuste do modelo pode ser avaliada por meio de medidas como o coeficiente de determinação (R^2), dado por [4]:

$$R^2 = 1 - \frac{\text{Soma Quadrática do Erro}}{\text{Soma Quadrática Total}}, \tag{2}$$

em que a *Soma Quadrádica do Erro* (SQE) é dada por $SQE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ e *Soma Quadrática Total* (SQE) é dada por $SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$, onde n refere-se ao número total de observações na amostra. Assim, R^2 indica a proporção da variância da variável dependente explicada pelo modelo. Para complementar essa análise, aplica-se o teste F global, para verificar, por meio de teste de hipóteses, a significância estatística do modelo como um todo. O teste dado por [4]:

$$F = \frac{\text{Média Quadrática da Regressão}}{\text{Média Quadrática do Erro}} \tag{3}$$

em que a *Média Quadrática da Regressão* (MQR) é dada por $MQR = \frac{SQR}{q}$, com a Soma Quadrática de Regressão (SQR) $SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SQT - SQE$, e *Média Quadrática do Erro* (MQE) é dada por $MQE = \frac{SQE}{n-q-1}$, onde q representa o número de variáveis explicativas do modelo.

Concluindo, a regressão linear simples serve como base para a regressão múltipla, que estende o modelo para incluir múltiplas variáveis independentes.

2.1.1 O Modelo

Para estimar a equação da população na técnica de regressão múltipla, temos [3]:

$$\mu_{y|x_1,x_2,...x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q, \tag{4}$$

sendo que $x_1, x_2, ...$ e x_q são as q variáveis explicativas distintas e $\mu_{y|x_1,x_2,...x_q}$ é a média dos valores de y para um valor específico da variável explicativa. Da mesma forma que na regressão linear, α e β são os coeficientes da equação, com o primeiro sendo o intercepto, que é o valor médio da resposta y quando todas as variáveis explicativas assumem o valor 0, e o segundo sendo a inclinação, que é a variação do valor médio de y que indica o aumento de uma unidade de x_i para $i = 1, \ldots, q$, considerando que todas as outras variáveis explicativas se mantenham constantes.

Com isso, ajustamos a (4) para um modelo da forma

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon, \tag{5}$$

com ε sendo o erro. Para estimar os coeficientes, é utilizada uma amostra aleatória representada por $(x_{1i}, x_{2i}, \ldots, x_{qi}, y_i)$ para $i = 1, \ldots, n$. Do mesmo modo que foram feitas suposições para o modelo de regressão linear simples, também são feitas suposições para o modelo de regressão múltipla:

1. Para valores específicos de $x_1, x_2, ..., x_q$, todos considerados medidos sem erro, a distribuição dos valores y é normal com média $\mu_{y|x_1,x_2,...,x_q}$ e desvio padrão $\sigma_{y|x_1,x_2,...,x_q}$;

2. A relação entre $\mu_{y|x_1,x_2,...,x_q}$ e $x_1, x_2, ..., x_q$ é representada pela equação

$$\mu_{y|x_1,x_2,...x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q; \tag{6}$$

3. Para qualquer conjunto de valores $x_1, x_2, ..., x_q$, $\sigma_{y|x_1,x_2,...,x_q}$ é constante. Como na regressão linear simples, essa característica é conhecida como homocedasticidade;

4. Os resultados y são independentes.

A Equação da Regressão de Mínimos Quadrados

Para estimar a equação de regressão da população (4) é utilizado o método de mínimos quadrados para ajustar o modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon. \tag{7}$$

Como está sendo usada essa técnica de cálculo numérico, temos

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_q x_{qi})^2. \tag{8}$$

Sendo que y_i é o resultado observado da resposta $Y = \left[y_1, \ldots, y_q\right]^T$ para valores específicos $x_{1i}, x_{2i}, ..., x_{qi}$ e o $\hat{y}_i$ é o valor da equação ajustada. Os estimadores, dados por $\hat{\alpha}$ e $\hat{\beta}$ de α e β_i com $i = 1, \cdots, q$ são obtidos pela solução do sistema de equações (8).

Agora, como está sendo trabalhado com duas variáveis explicativas, o modelo representa um plano no espaço tridimensional, consequentemente, quando há três ou mais variáveis, o modelo representará um hiperplano em um espaço dimensional mais elevado.

Inferência para os Coeficientes da Regressão

Como os coeficientes α e $\beta_1, \ldots, \beta_q$ são estimados em relação a uma amostra de dados, os erros padrão dos estimadores são necessários para realizar o cálculo da inferência em relação aos parâmetros verdadeiros.

Para determinar essa inferência estatística, será usado o teste de hipótese para as inclinações

$$\begin{cases} H_0 : \beta_i = \beta_{i0} \\ H_A : \beta_i \neq \beta_{i0} \end{cases} \tag{9}$$

para $i = 1, \ldots, q$ (normalmente o caso $\beta_{i0} = 0$ e $\beta_i \neq 0$). Tal que a estatística do teste é dada por

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{ep}(\hat{\beta}_i)}. \tag{10}$$

Supondo que a hipótese nula seja verdadeira, calcula-se a estatística (10), que tem uma distribuição t de Student com $n - q - 1$ graus de liberdade, com q sendo o número de variáveis explicativas no modelo.

Além dos testes de hipótese, é possível calcular o intervalo de confiança para os coeficientes da regressão da população, construindo assim um intervalo de confiança para o valor médio de Y previsto, e também é possível calcular um intervalo de predição para valores específicos de y que correspondem a um conjunto de valores das variáveis explicativas.

Vale ressaltar que, para ambos os casos, os passos para calculá-los são iguais aos passos em relação a uma única variável explicativa.

Avaliação do Modelo

Para a validação do modelo, assim como no modelo de uma variável, existem duas maneiras de avaliá-lo: de forma gráfica, por um gráfico de dispersão, e a partir do valor de R^2 [3].

Em relação ao R^2, para determinar se o modelo é bem ajustado, o valor que ele assume para uma variável explicativa é comparado com o valor obtido ao considerar de mais de uma variável explicativa. Caso o valor de R^2 para a regressão múltipla seja maior do que o valor da

regressão linear simples, pode-se sugerir que ao adicionar mais uma variável explicativa, houve uma melhora no modelo de previsão do caso estudado.

Entretanto, deve-se ser cauteloso ao comparar esses dois valores, já que está sendo comparado os coeficientes de determinação de dois modelos diferentes. Mesmo incluindo mais uma variável, o valor de R^2 de uma variável não pode ser menor que esse novo valor de R^2. Assim, para que a análise não seja prejudicada, ajusta-se o valor de R^2 ($R^2 ajustado$).

Sendo possível comparar com mais precisão: quando $R^2 ajustado$ for maior que R^2, pode-se afirmar que o modelo de regressão múltipla explica melhor o estudo do que o modelo de apenas uma variável.

Em relação ao gráfico de dispersão (em relação ao $R^2 ajustado$), quando é possível notar um padrão de leque no gráfico, fica evidente que um dos pressupostos citados anteriormente, a homocedasticidade, não é respeitado. A Figura 2.1, mostra um gráfico de dispersão em que a suposição de homocedasticidade é respeitada.

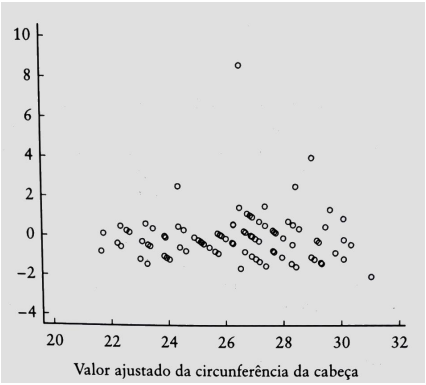

Figura 2.1: Resíduos *versus* valores ajustados. [Fonte: PAGANO, M; 2004, p. 401]

Variáveis Indicadoras

Além de variáveis explicativas contínuas, a análise de regressão pode ser feita por variáveis explicativas discretas ou nominais.

As variáveis nominais não tem nenhum valor quantitativo, sendo essas variáveis conhecidas por *variável indicadora* ou *variável dummy*, ou seja, esses números não representam qualquer medida real.

$$\begin{cases} H_0 : \beta_i = 0 \\ H_A : \beta_i \neq 0 \end{cases} \quad \text{para } i = 1, 2, \ldots, q. \tag{11}$$

Para determinar se a variável nominal é relevante para o modelo, é feito o teste de hipótese (11) e a estatística do teste t. Caso a hipótese nula seja rejeitada, pode-se concluir que a *variável dummy* influencia no modelo, ou seja, ou o valor de β_i pode influenciar ou não no valor específico de y_i.

Vale ressaltar que, a representação gráfica é feita uma linha de regressão em relação aos valores em que x_i é 1 e outra quando x_i for 0 e as linhas terão as mesmas inclinações.

Termos de Interação

Em modelos de regressão múltipla, assume-se inicialmente que o efeito de cada variável explicativa sobre a variável resposta é independente das demais variáveis presentes no modelo. Entretanto, em algumas situações práticas, essa suposição pode não ser válida. Uma **interação**

ocorre quando o efeito de uma variável explicativa x_i sobre a variável resposta y depende do valor de outra variável explicativa x_j. Dessa forma, a relação entre x_i e y é modificada por x_j, e o modelo aditivo simples não é suficiente para representar adequadamente o fenômeno estudado. Para contornar essa situação, é necessário incluir um termo de interação $x_i x_j$ no modelo, permitindo capturar esse efeito conjunto entre as variáveis [3].

Para descobrir se é preciso a utilização do termo, é feito um teste de hipótese:

$$\begin{cases} H_0 : \beta_{ij} = 0 \\ H_A : \beta_{ij} \neq 0 \end{cases} \quad \text{para } i = 1, 2, \ldots, q \text{ com } i \neq j. \tag{12}$$

Caso a hipótese nula seja rejeitada, é necessário utilizar o *termo de interação*. Portanto, caso a hipótese nula não possa ser rejeitada, não é possível usar o termo.

2.1.2 Seleção do Modelo

Como a regressão múltipla é utilizada para lidar com mais de uma variável, o seu real desafio é definir qual modelo utilizar, já que existem diversas opções. Para escolher o modelo ideal, leva-se em consideração a interpretação da base de dados estudada e considerações estatísticas e não estatísticas.

No Quadro 2.1 serão mostrados alguns modelos possíveis para esse tipo de regressão. Os modelos 1 a 4 indicam modelos de regressão linear possíveis, já os demais, são modelos de regressão múltipla.

Vale ressaltar que os modelos que apresentam termos como x_i^2 ou y_i^2 são chamados de modelos polinomiais de segunda ordem. A inclusão desses termos quadráticos permite capturar relações não lineares entre as variáveis, possibilitando modelar curvaturas nos dados, como pontos de máximo ou mínimo, que não seriam detectados por um modelo puramente linear [5].

Com isso, um dos maiores desafios da utilização dessa técnica é descobrir qual modelo usar de acordo com o banco de dados que será analisado. Como há diversas possibilidades de modelos, o processo de seleção costuma ser demorado e demanda um certo esforço computacional. Para isso, existem três tipos de processo: a seleção *forward*, a seleção *backward* e a junção de ambas, a seleção *stepwise*.

Modelos	Funções (notação algébrica)
1	$z_i = a + bx_i + \varepsilon$
2	$z_i = a + bx_i + cx_i^2 + \varepsilon$
3	$z_i = a + by_i + \varepsilon$
4	$z_i = a + by_i + cy_i^2 + \varepsilon$
5	$z_i = a + bx_i + cy_i + \varepsilon$
6	$z_i = a + bx_i + cx_i^2 + dy_i + \varepsilon$
7	$z_i = a + bx_i + cy_i + dy_i^2 + \varepsilon$
8	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + \varepsilon$
9	$z_i = a + bx_i + cy_i + dx_i y_i + \varepsilon$
10	$z_i = a + bx_i + cx_i^2 + dy_i + fx_i y_i + \varepsilon$

Quadro 2.1: Exemplos de Modelos de Regressão Múltipla

Seleção *Forward*

Esse processo é feito em etapas e, em cada uma delas, são adicionadas segundo o critério de maior significância variáveis ao modelo que, em seguida, é avaliado. Caso o critério estatístico seja atingido, encerra-se o processo; caso não seja atingido, serão adicionadas mais variáveis.

Primeiramente, se adiciona a variável explicativa que tem o maior coeficiente de determinação, ou seja, a variável que melhor explica a variação que ocorre na variável y. Em seguida, é incluída pelo usuário a variável que mais aumenta o R^2, sabendo que a primeira variável adicionada irá continuar no modelo e que o aumento de R^2 seja significativo estatisticamente.

Esse processo de adição de variáveis continuará até que nenhuma variável restante explique, de maneira significativa, a variável y.

Seleção *Backward*

Ao contrário do processo de seleção *Forward*, todas as variáveis são incluídas no modelo. Em seguida, elas serão retiradas uma de cada vez, começando com aquelas que resultem em menor impacto no valor de R^2. Depois, é analisada se essa remoção é significativa ou não para o modelo: se não for, essa variável é retirada definitivamente. Esse processo é feito até que as variáveis que restarem no modelo expliquem o comportamento da variação observada na variável y.

Seleção *Stepwise*

Esse processo é a junção dos processos *Forward* e *Backward*, ou seja, ambos são usados em conjunto. No começo, é feito exatamente igual à seleção *foward*: é adicionada uma variável de cada vez ao modelo. Quando essa nova variável é adicionada, as anteriores são verificadas para garantir a significância estatística, igual na seleção *backward*. Assim, uma variável que foi adicionada ao modelo na primeira etapa pode ser retirada na segunda etapa, caso não tenha significância.

Com isso, conclui-se que essa é a melhor seleção para descobrir qual será a expressão do modelo.

Colinearidade

O conceito de colinearidade é quando duas ou mais variáveis explicativas são semelhantes a ponto de terem a mesma informação sobre a variação observada em y. Uma consequência disso é a instabilidade dos valores dos coeficientes estimados e de seus erros-padrão, tornando-os muito grandes. Esse aumento nos valores indica uma grande variabilidade amostral nos coeficientes estimados.

Assim, independente do processo de seleção que for utilizado, é necessário verificar se há colinearidade ou não.

Para identificar a colinearidade, é utilizado indicadores como o Fator de Inflação da Variância (VIF). Para existem várias formas de solucionar esse problema, dentre elas, pode-se destacar a remoção ou combinação de variáveis altamente correlacionadas, a reavaliar a inclusão dos preditores no modelo ou a transformação das variáveis. Essas abordagens reduzem a instabilidade dos coeficientes estimados e melhoram a interpretação do modelo.

2.2 Modelos Lineares Generalizados

Para facilitar a compreensão da mudança de notação entre a regressão múltipla e os modelos lineares generalizados (MLGs), será demonstrada a equivalência entre a forma matricial e a forma de combinação linear (5).

Na regressão múltipla, o modelo é expresso como uma combinação linear de variáveis e essa equação pode ser reescrita de forma matricial como:

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \varepsilon, \tag{13}$$

em que $\mathbf{Y}$ é um vetor coluna contendo os valores da variável resposta $(y_1, y_2, \ldots, y_n)$, $\mathbf{X}$ é uma matriz contendo os valores das variáveis explicativas, $\vec{\beta}$ é um vetor coluna contendo os coeficientes de regressão $(\beta_0, \beta_1, \ldots, \beta_p)$ e ε é um vetor coluna contendo os termos de erro $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$.

A equivalência entre as duas formas fica clara ao expandir a equação matricial. Cada linha da equação matricial $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ representa a equação de combinação linear para uma observação específica. A forma matricial simplifica a notação e facilita a implementação computacional, especialmente quando lidamos com um grande número de variáveis explicativas e observações. Por esse motivo, passamos a utilizar a notação matricial nessa seção.

De maneira geral, pode-se representar um modelo linear da seguinte forma:

$$E(\mathbf{Y_i}) = \mu_i = \mathbf{x}_i^T \vec{\beta}, \tag{14}$$

com $\mathbf{Y}_i$ sendo as variáveis aleatórias independentes, que são a base para a maioria das análises de dados contínuos. Esse modelo é uma forma matricial de representar regressões lineares. Esse formato também é usado para relacionar uma resposta contínua com várias variáveis explicativas e compará-las com mais de uma média.

A evolução das teorias estatísticas e da tecnologia permitiu o desenvolvimento de modelos lineares mais gerais. Isso ocorreu já que agora variáveis respostas tem outras distribuições além da Distribuição Normal e a relação entre a variável resposta e explicativa não precisa ser representada de uma forma linear simples.

Um dos avanços mais significativos foi o conhecimento de que muitas propriedades importantes da Normal são compartilhadas com uma classe de distribuições: as distribuições da família exponencial.

Outro avanço importante foi o desenvolvimento dos métodos numéricos para estimar o parâmetro β do modelo (14) em situações em que uma função não linear de $E(\mathbf{Y}_i) = \mu_i$ está relacionada com uma componente linear de $\mathbf{x}_i^T \vec{\beta}$. Isso sendo representado da seguinte forma:

$$g(\mu_i) = \mathbf{x}_i^T \vec{\beta}. \tag{15}$$

A função g é chamada de **função de ligação** e na maioria dos modelos lineares generalizados g é uma simples função matemática. Para estimar essa função, teoricamente é um processo bem direto. Entretanto, ele pode demandar um grande esforço computacional, já que essa estimação é feita através da otimização numérica de funções não lineares.

2.2.1 Distribuições da Família Exponencial

Considere uma única variável aleatória Y com sua distribuição dependendo de um único parâmetro θ. A distribuição pertence à família exponencial se for escrita da seguinte maneira [6]:

$$f(y;\theta)) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{16}$$

onde a, b, s e t são funções conhecidas não negativas. A Equação (16) é reescrita da seguinte maneira:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)], \tag{17}$$

em que $s(y) = \exp d(y)$ e $t(\theta) = \exp c(\theta)$.

Se $a(y) = y$, a distribuição é chamada de **forma canônica** e $b(\theta)$ pode ser chamado de parâmetro natural da distribuição. Se existir em outros parâmetros além do θ, eles são chamados de **parâmetros de perturbação** e podem fazer parte das funções a, b, c e d.

Muitas distribuições conhecidas fazem parte da família exponencial. Por exemplo, as distribuições de Poisson, Normal - como citada anteriormente, e a Binomial. Com isso, pode-se escrevê-las de forma canônica.

Distribuição	Parâmetro Natural	c	d
Poisson	$\log\theta$	$-\theta$	-log$y$!
Normal	$\frac{\mu}{\sigma^2}$	-$\frac{\mu}{2\sigma^2}$ - $\frac{1}{2}log(2\pi\sigma^2)$	- $\frac{y^2}{2\sigma^2}$
Binomial	$\log(\frac{\pi}{1-\pi})$	$n\log(1-\pi)$	$\log\binom{n}{y}$

Tabela 2.1: Distribuição de Poisson, Normal, Binomial como parte da família exponencial. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 51]

2.2.2 Propriedades das Distribuições da Família Exponencial

Usualmente é necessário encontrar uma expressão para o valor esperado e a variância de $a(Y)$. Para encontrar esses valores, são utilizadas as propriedades das funções de probabilidade de densidade. Da definição da função de densidade de probabilidade, tem-se que [6]:

$$\int f(y;\theta)dy = 1, \tag{18}$$

em que a integral é em relação a todos os possíveis valores de y. Em casos em que os valores da variável aleatória Y são discretos, substituímos a integral por uma somatória.

Se em ambos os lados da Equação (18) são diferenciáveis em relação a θ, tem-se que:

$$\frac{d}{d\theta}\int f(y;\theta)dy = \frac{d}{d\theta}1 = 0. \tag{19}$$

Isso pode ser reescrito da seguinte forma:

$$\int \frac{df(y;\theta)}{d\theta}dy = 0. \tag{20}$$

Considerando também que a Equação (18) é diferenciável uma segunda vez em relação a θ, tem-se:

$$\int \frac{d^2f(y;\theta)}{d\theta^2}dy = 0. \tag{21}$$

A partir da Equação (16) tem-se

$$\frac{df(y;\theta)}{d\theta} = \exp\big[a(y)b'(\theta) + c'(\theta)\big]f(y;\theta), \tag{22}$$

e substituindo na Equação (21)

$$\int [a(y)b'(\theta) + c'(\theta)]f(y;\theta)dy = 0. \tag{23}$$

$$b'(\theta)E[a(y)] + c'(\theta) = 0. \tag{24}$$

Considerando a definição do valor esperado $\int a(y)f(y;\theta)dy = E[a(y)]$ e a partir a Equação (18): $\int c'(\theta)f(y;\theta)dy = c'(\theta)$ temos que

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}. \tag{25}$$

Esse raciocínio também é feito para obter a var$[a(Y)]$:

$$\frac{d^2 f(y;\theta)}{d\theta^2} = [a(y)b''(\theta) = c''(\theta)]f(y;\theta) + [a(y)b'(\theta) + c'(\theta)]^2 f(y;\theta) \tag{26}$$

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = b''(\theta)E[a(Y)] + c''(\theta) + [b'(\theta)]^2 var[a(Y)] = 0. \tag{27}$$

Portanto, tem-se que

$$var[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}. \tag{28}$$

As expressões (25) e (28) são usadas para obter o valor específico e a variância para as distribuições da família exponencial.

Além dessas expressões, são necessárias outras para determinar o valor esperado e a variância das derivadas da função log-verossimilhança. A partir da equação (17), a função de log-verossimilhança para a família exponencial é dada por

$$l(\theta;y) = a(y)b(\theta) + c(\theta) + d(y). \tag{29}$$

Derivando $l(\theta;y)$ em relação a θ

$$U(\theta : y) = \frac{dl(\theta;y)}{d\theta} = a(y)b'(\theta)c'(\theta). \tag{30}$$

Essa função U é chamada de **função escore** e como depende de y, pode ser reescrita

$$U = a(Y)b'(\theta) + c'(\theta). \tag{31}$$

Sendo seu valor esperado

$$E(U) = b'(\theta)E[a(Y)] + c'(\theta) = b'(\theta)[-\frac{c'(\theta)}{b'(\theta)}] + c'(\theta) = 0. \tag{32}$$

A variância de U é chamada de informação e é denotada por Im é dado por

$$\text{Im} = var(U) = [b'(\theta)^2]var[a(Y)] = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta). \tag{33}$$

A função escore U é usado na inferência dos parâmetros nos modelos lineares generalizados. Outra propriedade muito importante

$$var(U) = E(U^2) = -E(U'). \tag{34}$$

Obtendo a diferencial de U em relação a θ a partir de (31)

$$E(U') = b''(\theta)E[a(Y)] + c''(\theta) = b''(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta) = -var(U) = -\text{Im}. \tag{35}$$

2.2.3 Modelos Lineares Generalizados

Os modelos lineares generalizados (MLG) são definidos por um conjunto de variáveis independentes aleatórias $Y_1, ..., Y_n$, cada modelo tendo sua distribuição vinda da família exponencial e carregando as seguintes propriedades [6]:

1. A distribuição de cada Y_i tem sua forma canônica e depende apenas do parâmetro θ_i, então

$$f(y_i;\theta_i) = \exp[y_i \mathbf{b}_i(\theta_i) + c_i(\theta_i) + d(y_i)] \quad \text{para} \quad i = 1, \ldots, n; \tag{36}$$

2. A distribuição de todos os Y_i são iguais. Com isso, a função de probabilidade de densidade

$$\begin{aligned} f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) &= \sum_{i=1}^{n} \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] \\ &= \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right] \end{aligned} \tag{37}$$

Geralmente, o parâmetro θ_i não é de grande interesse, já que ele pode ter um valor para cada observação. Para especificar o modelo, analisar os parâmetros $\beta_1, ..., \beta_q$ (com $q < n$) é mais relevante. Supondo que a Equação (14), com μ_i sendo uma função de θ_i. Para os MLGs há uma transformação de μ_i dada pela Equação (15).

Essa Equação 15 em que g é a **função ligação**, que pode ser linear, crescente ou decrescente em relação a μ_i. Mas não pode ser crescente para alguns valores e decrescente para outros valores de μ_i. Também em relação a Equação (15), o vetor x_i é um vetor das variáveis explicativas com dimensões $p \times 1$, representado por

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \longrightarrow \mathbf{x}_i = \begin{bmatrix} x_{i1} & \cdots & x_{ip} \end{bmatrix} \text{ para } i = 1, \ldots, p, \tag{38}$$

e β é um vetor, também de dimensões $p \times 1$, sendo escrito da seguinte maneira

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}. \tag{39}$$

Assim, os MLGs têm três componentes:

1. Variáveis respostas $Y_1, ..., Y_n$ que compartilham a mesma distribuição da família exponencial.

2. O conjunto dos parâmetros β e as variáveis explicativas é definido por:

$$\mathbf{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{bmatrix} \tag{40}$$

3. Uma função ligação g:

$$g(\mu_i) = \mathbf{x}_i^T \vec{\beta}, \tag{41}$$

onde

$$\mu_i = E(Y_i). \tag{42}$$

2.2.4 Estimação dos parâmetros

Para obter uma estimativa pontual e intervalar de parâmetros para os modelos lineares generalizados, são utilizados dois métodos matemáticos: métodos numéricos iterativos baseados no algoritmo de Newton-Raphson e a máxima verossimilhança. Para compreender melhor esses processos, a teoria será desenvolvida a partir de um exemplo [6].

Os dados da Tabela 2.2 representam o total de horas que as válvulas de pressão *Kevlar Epoxy* a 70% do nível de estresse.

1051	4921	7886	10861	13520
1337	5445	8108	11026	13670
1389	5620	8546	11214	14110
1921	5817	8666	11362	14496
1942	5905	8831	11604	15395
2322	5956	9106	11608	16179
3629	6068	9711	11745	17092
4006	6121	9806	11762	17568
4012	6473	10205	11895	17568
4063	7501	10396	12044	

Tabela 2.2: Tempo de falha em horas de $n = 49$ válvulas de pressão.[Fonte: Andrews, D. F. e Herzberg, A. M.; 1985, p. 29.1]

Para esses casos de falha, um modelo comumente usado é a **distribuição Weibull**. Ela é representada pela função de densidade de probabilidade a seguir:

$$f(y;\lambda;\theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y}{\theta}\right)^{\lambda}\right], \tag{43}$$

onde $y > 0$ é o tempo de falha, λ é o parâmetro que determina a forma da distribuição, θ o parâmetro que determina a escala $\lambda > 0$ e $\theta > 0$. A Figura 2.2 apresenta o histograma dos dados. A Figura 2.3 mostra o gráfico da probabilidade dos dados comparados com a distribuição Weibull com $\lambda = 2$. A partir dela, é possível avaliar se a distribuição é um modelo adequado para representar o tempo de falha.

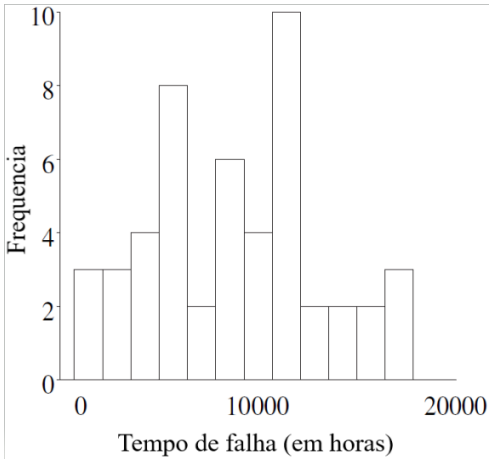

Figura 2.2: Histograma do tempo de falha. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 66]

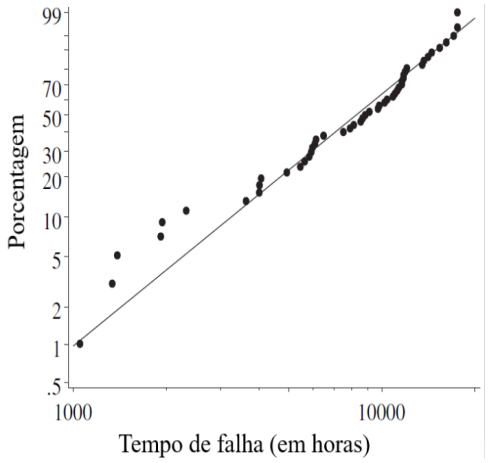

Figura 2.3: Gráfico da probabilidade dos dados de tempo de falha comparados com a distribuição Weibull com $\lambda = 2$. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 66]

Reescrevendo a expressão (43),

$$f(y;\theta) = \exp\left[\log\lambda + (\lambda-1)\log y - \lambda\log\theta - \left(\frac{y}{\theta}\right)^{\lambda}\right]. \tag{44}$$

Considerando (16), sabe-se que a expressão acima pertence à família exponencial, pois

$$a(y) = y, \quad b(\theta) = -\theta^{-\lambda}, \quad c(\theta) = \log\lambda - \lambda\log\theta \quad e \quad d(y) = (\lambda-1)\log y. \tag{45}$$

Seja $Y_1, ..., Y_n$ os dados da amostra com $n = 49$ e que Y_i' são variáveis aleatórias independentes. Com a distribuição Weibull e o parâmetro λ

$$f(y_1, ..., y_n; \theta, \lambda) = \prod_{i=1}^{n} \frac{\lambda y_i^{\lambda-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y_i}{\theta}\right)^{\lambda}\right]. \tag{46}$$

A função log-verossimilhança é dada por

$$l = l(\theta; y_1, ..., y_n, \lambda) = \sum_{i=1}^{n}\left[(\lambda-1)\log y_i + \log\lambda - \lambda\log\theta - \left(\frac{y_i}{\theta}\right)^{\lambda}\right]. \tag{47}$$

Para maximizar a função, é necessário o uso da derivada de θ, que é similar a função escore:

$$\frac{dl}{d\theta} = U = \sum U_i = \sum_{i=1}^{n}\left[\frac{-\lambda}{\theta} + \frac{\lambda y_i^{\lambda}}{\theta^{\lambda+1}}\right]. \tag{48}$$

O estimador da máxima verossimilhança $\hat{\theta}$ é a solução da equação $U(\theta) = 0$. Para obter-se a solução numérica do estimador, será utilizado o método de aproximação de Newton-Raphson.

A lógica do método é encontrar o valor de x em que a função t cruza o eixo x, ou seja, quando $t(x) = 0$, representado graficamente na Figura 2.4. A iteração m de t para um valor de x^{m-1} é dado por

$$\left[\frac{dt}{dx}\right]_{x=x^{m-1}} = t'(x^{(m-1)}) = \frac{t(x^m) - t(x^{(m-1)})}{x^m - x^{(m-1)}}, \tag{49}$$

em que a distância de $x^{(m)} - x^{(m-1)}$ é pequena. Se $x^{(m)}$ for a solução de $t(x^m) = 0$, reescreve-se a Expressão (49) como

$$x^{(m)} = x^{(m-1)} - \frac{t(x^{(m-1)})}{t'(x^{(m-1)})}. \tag{50}$$

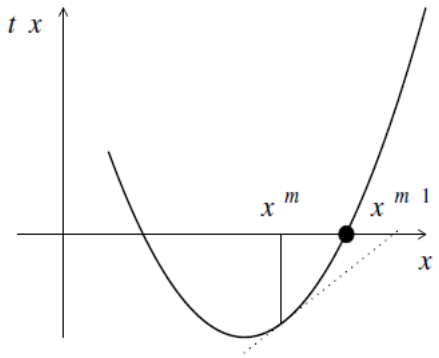

Figura 2.4: Ilustração da primeira iteração do método Newton-Ralphson para encontrar a solução, representada pelo circulo (•), da equação $t(x) = 0$. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 67]

A Equação (50) é usada para resolver $t(x) = 0$ utilizando o método Newton-Raphson. Assim como em outros métodos de iteração, tem-se uma solução inicial de $x^{(0)}$ que será alterada pelas aproximações em cada iteração até a convergência.

Para a estimação da máxima verossimilhança para um parâmetro θ utilizando a função escore, tem-se que

$$\theta^{(m)} = \theta^{(m-1)} - \frac{U^{(m-1)}}{U'^{(m-1)}}. \tag{51}$$

Por (48) e em relação à distribuição Weibull com $\lambda = 2$, obtém-se a seguinte expressão para U e U'

$$U = -\frac{2\times n}{\theta} + \frac{2\times\sum y_i^2}{\theta^3} \longrightarrow \frac{dU}{d\theta} = U' = \sum_{i=1}^{n}\left[\frac{\lambda}{\theta^2} - \frac{\lambda(\lambda+1)y_i^{\lambda}}{\theta^{\lambda+2}}\right] = \frac{2\times n}{\theta^2} - \frac{2\times 3\times\sum y_i^2}{\theta^4}. \tag{52}$$

Para essa estimação, é comum ser feita a aproximação de U' a partir de seu valor esperado $E(U')$ e, em específico, para a distribuição da família exponencial, isso é obtido através da relação (33). Assim, a informação Im é dada por

$$\text{Im} = \sum_{i=1}^{n}\left[\frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)\right] = \frac{\lambda^2 n}{\theta^2}, \tag{53}$$

em que U_i é o score para Y_i e as expressões de b e c são dadas por (45). Portanto,

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\text{Im}^{(m-1)}}. \tag{54}$$

A primeira linha da Tabela 2.3 mostra os resultados numéricos das iterações do método de Newton-Raphson considerando a Equação (51) e a média dos dados $\bar{y} = 8805,7$ sendo a solução inicial $[\theta^{(1)}]$. Já a segunda linha indica os resultados avaliados em relação a Equação (52) em $\theta^{(m)}$ que aproximam-se rapidamente para 0. A terceira e a quarta linhas tem valores similares; consequentemente, os valores das linhas cinco e seis também são similares. O método convergiu na quinta iteração, resultando em $\theta^{(5)} = 9892,1 - (-0,105) = 9892,2$. Esse valor é o estimador $\hat{\theta}$ da máxima verossimilhança. A partir desse valor, é possível calcular a log verossimilhança com (47) que é $l = -480,850$.

A Figura 2.5 mostra a função log-verossimilhança para os dados da Tabela 2.2 com distribuição Weibull com $\lambda = 2$.

Iteração	1	2	3	4
θ	8805,7	9633,9	9876,4	9892,1
$U \times 10^6$	2915,10	552,80	31,78	0,21
$U' \times 10^6$	-3,52	-2,28	-2,02	-2,00
$E(U') \times 10^6$	-2,53	-2,11	-2,01	-2,00
U/U'	-827,98	-242,46	-15,73	-0,105
$U/E(U')$	-1152,21	-261,99	-15,81	-0,105

Tabela 2.3: Resultados das iterações utilizando o método de Newton-Raphson para estimar o parâmetro escalar θ com $h = 2$ da máxima verossimilhança para distribuição Weibull para o tempo de falha da Tabela 2.2. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 69]

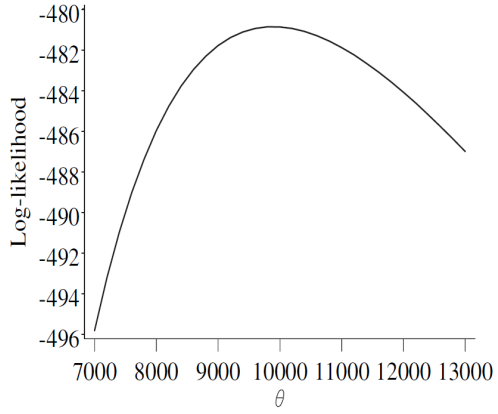

Figura 2.5: Função log verossimilhança em relação ao tempo de falha das válvulas *Kevlar epoxy*.[Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 70]

É importante estudar a curvatura do gráfico em relação à vizinhança do máximo para determinar a viabilidade do estimador $\hat{\theta}$. Essa curva é definida pela amplitude de mudança da derivada da função escore U'. Quando U' é menor que l, o gráfico é plano para que o U aproximasse a zero para a maioria dos valores de θ.

No exemplo, $\hat{\theta}$ não é um bom estimador e por isso o erro desse modelo é grande, sendo ele

$$\hat{ep}(\hat{\theta}) = \sqrt{\frac{1}{\text{Im}}} \longrightarrow \hat{\theta} = \frac{1}{\sqrt{0,000002}} = 707 \tag{55}$$

para $\hat{\theta} = 9892,2$ e $\text{Im} = -E(U') = 2,00 \times 10^{-6}$.

Considerando agora que os estimadores de máxima verossimilhança tem uma distribuição Normal com 95% de confiança, o intervalo do parâmetro θ é

$$IC(\theta) : 9892 \pm 1,96 \times 707 \longrightarrow [8506 \leq \theta \leq 11278]. \tag{56}$$

Esse raciocínio é chamado de **método escoring**.

Estimação da Máxima Verossimilhança

O processo feito no exemplo anterior é referente a uma forma não canônica. Assim, nessa seção, será desenvolvida uma forma para os modelos lineares generalizados (MLGs).

Considerando uma amostra de variáveis aleatórias independentes $Y_1, ..., Y_n$ que satisfaz as propriedades dos MLGs, se deseja estimar os parâmetros $\vec{\beta}$. Esses parâmetros são ligados aos Y_i devido à relação das Equações (14) e (15) [6].

Para cada valor de Y_i a função log-verossimilhança é

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i), \tag{57}$$

em que as funções b, c e d são definidas anteriormente na Equação (17). Elas também podem ser representadas da seguinte forma

$$E(Y_i) = \mu_i = -\frac{c'(\theta)}{b'(\theta)} \tag{58}$$

$$\text{var}(Y_i) = \frac{[b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)]}{[b'(\theta_i)]^3} \tag{59}$$

$$\text{e} \qquad g(\mu_i) = \mathbf{x}_i^T \vec{\beta} = \eta_i, \tag{60}$$

em que $\mathbf{x}_i$ sendo um vetor com os elementos x_{ij} com $j = 1, \ldots, \text{p}$.

Agora, escrevendo a função log-verossimilhança geral para todos os valores de Y_i

$$l = \sum_{i=1}^{n} l_i = \sum \text{y}_i b(\theta_i) + \sum c(\theta_i) + \sum d(\text{y}_i). \tag{61}$$

Para obter o estimador do parâmetro β_j (com $j = 1, \cdots, p.$) da máxima verossimilhança, é necessário utilizar a derivada da função l, que é encontrada através da regra da cadeia.

$$\frac{\partial l_i}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_j} \right] = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \theta_i} . \frac{\partial \theta_i}{\partial \mu_i} . \frac{\partial \mu_i}{\partial \beta_j} \right]. \tag{62}$$

Determinando o valor de cada derivada parcial do lado direito da expressão (62) e fazendo as manipulações essenciais, resulta-se no na função escore a seguir

$$U_j = \sum_{i=1}^{n} \left[\frac{(\text{y}_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right], \tag{63}$$

que tem sua matriz de variância-covariância de U_j dada por

$$\text{Im}_{jk} = E[U_j U_k], \tag{64}$$

também chamada de matriz de informação Im. Com base na expressão (63) e sabendo que $E\left[(Y_i - \mu_i)^2\right] = \text{var}(Y_i)$ e $E\left[(Y_i - \mu_i)(Y_l - \mu_l)\right] = 0$ para $i \neq l$ já que os valores de Y_i são independentes, tem-se

$$\text{Im}_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \tag{65}$$

Para os MLGs, a equação de estimação (54) agora é representada por

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left[\text{Im}^{(m-1)} \right]^{-1} \mathbf{U}^{(m-1)}, \tag{66}$$

onde $\mathbf{b}^{(m)}$ é o vetor das estimativas dos parâmetros $\beta_1, \ldots, \beta_p$ para a m-ésima iteração, $\left[\text{Im}^{(m-1)} \right]^{-1}$ é o inverso da matriz de informação com elementos Im_{jk} e $\mathbf{U}^{(m-1)}$ é o vetor da função escore avaliado em relação a $\mathbf{b}^{(m-1)}$.

Multiplicando ambos os lados de (66) por $\mathrm{Im}^{(m-1)}$, consiste em

$$\mathrm{Im}^{(m-1)}\,\mathbf{b}^{(m)} = \mathrm{Im}^{(m-1)}\,\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} \tag{67}$$

$$\mathrm{Im} = \mathbf{X}^T\mathbf{W}\mathbf{X}, \tag{68}$$

com $\mathbf{W}$ sendo a diagonal da matriz $N \times N$ com elementos calculados por

$$w_{ii} = \frac{1}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{69}$$

Levando em conta o lado direito de (68), pode-se escrever

$$\sum_{k=1}^{p}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mathrm{var}\,(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{n}\frac{(y_i-\mu_i)\,x_{ij}}{\mathrm{var}\,(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \longrightarrow \mathbf{X}^T\mathbf{W}\mathbf{z}, \tag{70}$$

com $\mathbf{z}$ tendo elementos formados por

$$z_i = \sum_{k=1}^{p} x_{ik}b_k^{(m-1)} + (\mathrm{y_i} - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right). \tag{71}$$

Portanto, conclui-se que

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}. \tag{72}$$

Essa forma é a mesma obtida pelo modelo de regressão linear simples pelo método de mínimos quadrados. A única diferença em relação a esse método é que ele é iterativo, já que $\mathbf{z}$ e $\mathbf{W}$ dependem de $\mathbf{b}$. Ou seja, para os MLGs, os estimadores da máxima verossimilhança são calculados pelos mínimos quadrados ponderados iterativos.

A maioria dos pacotes estatísticos que incluem procedimentos para ajustar os modelos lineares generalizados tem um algoritmo eficiente com base em (72). Eles começam usando alguma aproximação inicial $\mathbf{b}^{(0)}$ para calcular os primeiros valores de $\mathbf{z}$ e $\mathbf{W}$. Em seguida a Equação (72) é resolvida para encontrar $\mathbf{b}^{(1)}$, que por sua vez é usado para obter melhores aproximações para $\mathbf{z}$ e $\mathbf{W}$, e assim por diante, até que alcance a convergência. Quando a diferença entre as aproximações sucessivas $\mathbf{b}^{(m-1)}$ e $\mathbf{b}^{(m)}$ é suficientemente pequena, $\mathbf{b}^{(m)}$ é tomado como a estimativa de máxima verossimilhança.

2.2.5 Inferência em MLG

A inferência é uma ferramenta estatística utilizada para determinar se a amostra de uma determinada população estudada é adequada para estimar os parâmetros populacionais. Essa técnica usa duas ferramentas principais para realizar essa análise: os intervalos de confiança e os testes de hipóteses[6].

O primeiro tende a ser mais usado que o segundo, já que a amplitude do intervalo de confiança providencia uma precisão em relação aos valores estimados, além de ser mais simples de ser realizado comparado com um teste estatístico [6].

Bem diferente da abordagem do intervalo de confiança, os testes de hipóteses comparam a eficácia de dois modelos diferentes para o mesmo conjunto de dados. Nos MLGs, esses modelos tem a mesma distribuição de probabilidade e a mesma função de ligação, mas a componente linear de um deles tem mais parâmetros que a outra.

O modelo mais simples, representado por H_0, é um modelo que apresenta de maneira mais simples como os dados estudados podem ser analisados. Ao se realizar os testes de hipóteses,

caso H_0 se ajuste tão bem aos dados quanto o modelo mais geral, não se rejeita H_0. Se o oposto ocorrer, ou seja, o modelo geral tem um ajuste melhor aos dados, rejeita-se H_0 e se considera a hipótese H_1.

A "boa"qualidade do ajuste estatístico é baseada no valor máximo da função de verossimilhança, o valor máximo da função de log-verossimilhança, no menor valor da soma dos quadrados ou em uma estatística em relação aos resíduos. Todo esse processo pode ser resumido nos passos seguintes[6]:

1. Especifique um modelo M_0 correspondente a H_0. Especifique um modelo mais geral M_1 (com M_0 como um caso especial de M_1).

2. Ajuste M_0 e calcule a estatística de ajuste G_0. Ajuste M_1 e calcule a estatística de ajuste G_1.

3. Calcule a melhoria no ajuste, geralmente $G_1 - G_0$, mas $\frac{G_1}{G_0}$ é outra possibilidade

4. Use a distribuição amostral de $G_1 - G_0$ (ou alguma estatística relacionada) para testar a hipótese nula de que $G_1 = G_0$ contra a hipótese alternativa $G_1 \neq G_0$.

5. Se a hipótese de que $G_1 = G_0$ não for rejeitada, então H_0 não é rejeitada e M_0 é o modelo preferido. Se a hipótese $G_1 = G_0$ for rejeitada, então H_0 é rejeitada e M_1 é considerado o melhor modelo.

Tanto para calcular o intervalo de confiança quanto a análise dos testes de hipótese, é necessário encontrar, para o primeiro, a distribuição amostral do estimador e, do segundo, a distribuição amostral do melhor modelo. Com isso, para esses dois casos de inferência, é muito importante o entendimento dessas distribuições amostrais.

Para as observações independentes com distribuições das variáveis respostas que pertencem à família exponencial e, em específico, aos modelos lineares generalizados, é possível encontrar a distribuição amostral para ser utilizada no cálculo da inferência.

A ideia principal é que, em situações apropriadas, com S sendo a estatística de interesse,

$$\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{var}(S)}} \sim \mathrm{N}(0,1) \longrightarrow \frac{[S - \mathrm{E}(S)]^2}{\mathrm{var}(S)} \sim \chi^2(1), \tag{73}$$

considerando $E(S)$ a esperança, $\mathrm{var}(S)$ a variância e para $n \to \infty$.

Se o vetor da estatística de interesse é dado por $\begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ com a expectativa assintótica $E(S)$ e uma matriz variância-covariância também assintótica, tem-se aproximadamente

$$[\mathbf{S} - \mathrm{E}(\mathbf{S})]^T \mathbf{V}^{-1} [\mathbf{S} - \mathrm{E}(\mathbf{S})] \sim \chi^2(p), \tag{74}$$

com V sendo uma matriz não singular, portanto existe apenas uma única matriz inversa V^{-1}.

Distribuição da Amostral da Função Escore

A seguir será mostrado como é encontrada a distribuição amostral da função do escore estatístico que será muito importante no processo inferencial.

Supondo que $Y_1, ..., Y_n$ são variáveis independentes aleatórias em um modelo linear generalizado com parâmetros $\vec{\beta}$ com (14) e (60) e a função escore dado em (63).

Se houver apenas um parâmetro β, a função do score estatístico tem uma distribuição amostral assintótica dado por

$$\frac{U}{\sqrt{\text{Im}}} \sim \text{N}(0,1), \quad \text{ou equivalente a} \quad \frac{U^2}{\text{Im}} \sim \chi^2(1), \tag{75}$$

tal que U é dado em (63) e Im (64) temos um vetor dos parâmetros $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$, então o vetor da função escore $\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}$ tem uma distribuição normal multivariada $\mathbf{U} \sim \text{MVN}(0, \text{Im})$ e é assintótico. Então,

$$\mathbf{U}^T \, \text{Im}^{-1} \, \mathbf{U} \sim \chi^2(p) \tag{76}$$

é utilizado para amostras grandes [6].

Distribuição da Amostral para os Estimadores da Máxima Verossimilhança

Para obter a distribuição amostral assintótica para outras estatísticas, por exemplo, são utilizadas as aproximações por séries de Taylor. Uma função $f(x)$ com uma única variável x aproximada pela série a t com x sendo próximo de t, tem-se [6]

$$f(x) = f(t) + (x-t)\left[\frac{df}{dx}\right]_{x=t} + \frac{1}{2}(x-t)^2\left[\frac{d^2f}{dx^2}\right]_{x=t} + \cdots. \tag{77}$$

Tendo essa aproximação, para a função de log-verossimilhança de apenas um parâmetro β aproximando-se de b, onde $U(b) = \frac{dl}{d\beta}$ é a função escore em relação a b. Considerando também que $U' = \frac{d^2l}{d\beta^2}$ aproxima-se do valor esperado de $\text{E}(U') = -\text{Im}$ em que $\text{Im}(b)$ é a informação avaliada em b, a série de Taylor para a função log-verossimilhança para o vetor $\vec{\beta}$ é dada por

$$l(\vec{\beta}) = l(\mathbf{b}) + (\vec{\beta} - \mathbf{b})^T \mathbf{U}(\mathbf{b}) - \frac{1}{2}(\vec{\beta} - \mathbf{b})^T \mathfrak{I}(\mathbf{b})(\vec{\beta} - \mathbf{b}), \tag{78}$$

com U representando o vetor da função escore e Im a matriz informação.

Voltando para o caso em que a função escore tem apenas um parâmetro β e levando em conta novamente que $\text{E}(U') = -\text{Im}$, tem-se que a aproximação por série de Taylor para o parâmetro $\vec{\beta}$

$$\text{U}(\vec{\beta}) = \text{U}(\mathbf{b}) - \text{Im}(\mathbf{b})(\vec{\beta} - \mathbf{b}). \tag{79}$$

A aproximação por série de Taylor dada na Equação (79) é usada para obter a distribuição da máxima verossimilhança com estimador $\mathbf{b} = \hat{\vec{\beta}}$. Por definição, $\mathbf{b}$ é um estimador que maximiza $l(\mathbf{b})$ e, portanto, $\mathbf{U}(\mathbf{b}) = 0$ [6]. Assim,

$$\mathbf{U}(\vec{\beta}) = -\mathfrak{I}(\mathbf{b})(\vec{\beta} - \mathbf{b}) \quad \text{ou equivalente à} \quad (\mathbf{b} - \vec{\beta}) = \mathfrak{I}^{-1}\mathbf{U}, \tag{80}$$

relembrando que a matriz Im é não singular. Agora, se Im for tratada como constante, $E(\mathbf{b} - \beta) = 0$ já que $E(\mathbf{U}) = 0$. Consequentemente, $E(\mathbf{b}) = \beta$ e assintótico com $\mathbf{b}$ sendo um estimador consistente de $\vec{\beta}$. Dado então, a matriz de variância-covariância

$$\text{E}\left[(\mathbf{b} - \vec{\beta})(\mathbf{b} - \vec{\beta})^T\right] = \mathfrak{I}^{-1}\text{E}\left(\mathbf{U}\mathbf{U}^T\right)\mathfrak{I}^{-1} = \mathfrak{I}^{-1}, \tag{81}$$

já que $\mathrm{Im} = \mathrm{E}(\mathbf{UU})^T$ e $(\mathrm{Im}^{-1})^T = \mathrm{Im}^{-1}$ sabendo que a matriz é simétrica.

A distribuição assintótica para **b**, por (74)

$$(\mathbf{b} - \vec{\beta})^T \mathfrak{J}(\mathbf{b})(\mathbf{b} - \vec{\beta}) \sim \chi^2(p). \tag{82}$$

A expressão (82) é conhecida como **Estatística de Wald**, e para os casos de apenas um parâmetro a forma mais usada é

$$b \sim \mathrm{N}\left(\beta, \mathfrak{I}^{-1}\right). \tag{83}$$

Se as variáveis respostas do modelo linear generalizado forem uma distribuição normal, então sabe-se que tanto (82) quanto (83) terão resultados com solução analítica [6].

Razão de Verossimilhança

Como mencionada na introdução, uma das formas de determinar a adequação do modelo é comparando-o com outro mais geral, com maior número de parâmetros estimados. Esse modelo é chamado de modelo saturado, que é um MLG com a mesma distribuição e função de ligação do que o modelo de interesse.

Em uma situação em que há n observações Y_i, $i = 1, \ldots, n$, com todos os valores da componente linear, potencialmente diferentes, tendo a representação matricial $\mathbf{x}_i^T \vec{\beta}$, então um modelo saturado pode ser especificado com n parâmetros.

Mas, se algumas observações tiverem o mesmo componente linear ou padrão de covariância, esses valores repetidos são conhecidos como réplicas. Nesse caso, o valor máximo de parâmetros que serão estimados para o modelo saturado é igual ao número dos diferentes valores da componente linear, que pode ser menor que n.

Na maior parte, utilizamos k para denotar o número máximo de parâmetros estimados, $\vec{\beta}_{\max}$ é o vetor dos parâmetros para esse modelo saturado e $\mathbf{b}_{\max}$ o estimador da máxima verossimilhança de $\vec{\beta}_{\max}$.

A função de verossimilhança para o modelo saturado é avaliada em $\mathbf{b}_{\max}, L(\mathbf{b}_{\max}; \mathrm{y})$ e será maior do que qualquer outra função para essas observações, assumindo a mesma distribuição e função de ligação. Já que dessa forma, é proporciona uma melhor descrição dos dados.

O valor máximo para a função de verossimilhança é dado por $L(\mathbf{b}; \mathbf{y})$, assim a razão da verossimilhança é representada pela expressão abaixo. Sua principal função é indicar a qualidade do ajuste do modelo de interesse.

Na prática, o logaritmo da razão (84) é mais usado para indicar a qualidade do ajuste do modelo de interesse e é composto pela diferença entre as funções logarítmicas de verossimilhança - propriedades do logaritmo.

$$\lambda = \frac{L(\mathbf{b}_{\max}; \mathbf{y})}{L(\mathbf{b}; \mathbf{y})} \quad \longrightarrow \quad \log \lambda = l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y}) \tag{84}$$

Quando $\log\lambda$ assume valores grandes, isso indica que o modelo estudado é um ajuste ruim em relação aos dados. Para determinar a região crítica de $\log\lambda$, é preciso determinar sua distribuição amostral.

Distribuição Amostral do Desvio

Em calcular a distribuição da diferença da função de verossimilhança, o cálculo é feito em relação a $2\log\lambda$ por ter uma distribuição de qui-quadrado. Desse modo, tem-se a estatística da razão da log-verossimilhança, ou simplesmente o desvio.

$$D = 2\left[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})\right]. \tag{85}$$

A partir de (78), se $\mathbf{b}$ é o estimador da máxima verossimilhança do parâmetro $\vec{\beta}$ para que $\mathbf{U}(\mathbf{b}) = 0$, é dada por

$$l(\vec{\beta}) - l(\mathbf{b}) = -\frac{1}{2}(\vec{\beta} - \mathbf{b})^T \mathfrak{I}(\mathbf{b})(\vec{\beta} - \mathbf{b}). \tag{86}$$

Assim, a estatística é (87), na qual é expressa por uma distribuição qui-quadrado $\chi^2(p)$, onde p é o número de parâmetros de (74)

$$2[l(\mathbf{b};\mathbf{y}) - l(\vec{\beta};\mathbf{y})] = (\vec{\beta} - \mathbf{b})^T \mathfrak{I}(\mathbf{b})(\vec{\beta} - \mathbf{b}). \tag{87}$$

Desse resultado, a distribuição do desvio é dada por

$$\begin{aligned} D =& 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l(\mathbf{b};\mathbf{y})\right] \\ =& 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l\left(\vec{\beta}_{\max};\mathbf{y}\right)\right] - 2[l(\mathbf{b};\mathbf{y}) - l(\vec{\beta};\mathbf{y})] + 2\left[l\left(\vec{\beta}_{\max};\mathbf{y}\right) - l(\vec{\beta};\mathbf{y})\right]. \end{aligned} \tag{88}$$

O primeiro termo do parênteses tem distribuição em $\chi^2(m)$, com m sendo o número de parâmetros do modelo saturado. O segundo termo tem distribuição em $\chi^2(p)$, com p sendo o número de parâmetros do modelo de interesse. O terceiro termo $v = \left[l\left(\vec{\beta}_{\max};\mathbf{y}\right) - l(\vec{\beta};\mathbf{y})\right]$ é a constante positiva que tenderá a zero se o modelo de interesse se ajustar aos dados tão bem quanto o modelo saturado. Desse modo, a distribuição do desvio é aproximadamente

$$D \sim \chi^2(m - p, v), \tag{89}$$

em que v é um parâmetro de não centralidade. Entender a distribuição do desvio é essencial para compreender a maior parte dos testes de hipóteses feitos para os modelos lineares generalizados.

Se as variáveis respostas Y_i são normalmente distribuídas, então D tem uma distribuição qui-quadrado exata. Neste caso, no entanto, D depende de $\text{var}(Y_i) = \sigma^2$, que, na prática, é geralmente desconhecido. Isso significa que D não pode ser usado diretamente como uma estatística de qualidade de ajuste [6].

Para Y_i com outras distribuições, a distribuição amostral de D pode ser apenas aproximadamente qui-quadrado. No entanto, para as distribuições Binomial e Poisson, por exemplo, D pode ser calculado e usado diretamente como uma estatística de ajuste de qualidade de D [6].

Testes de Hipóteses

Os testes de hipóteses são feitos para estudar o vetor $\vec{\beta}$ com dimensão p. Ele é testado a partir da distribuição amostral da estatística de Wald dado na Equação (82) com $b = \hat{\beta}$. Ou, para alguns casos, é usado a função escore (76).

Considerando a hipótese nula H_0 correspondendo ao modelo M_0, que é o mais simples (normalmente, $\beta_0 = 0$),

$$H_0 : \vec{\beta} = \vec{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}, \tag{90}$$

e a hipótese mais geral H_1, agora em relação ao modelo M_1 (normalmente, $\beta_1 \neq 0$)

$$H_1 : \vec{\beta} = \vec{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \tag{91}$$

e tendo $q < p < n$. Assim, testa-se H_0 contra H_1 usando a diferença dos desvios estatísticos

$$\begin{aligned}\Delta D &= D_0 - D_1 = 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l\left(\mathbf{b}_0;\mathbf{y}\right)\right] - 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l\left(\mathbf{b}_1;\mathbf{y}\right)\right] \\ &= 2\left[l\left(\mathbf{b}_1;\mathbf{y}\right) - l\left(\mathbf{b}_0;\mathbf{y}\right)\right].\end{aligned} \tag{92}$$

Se ambos os modelos descreverem bem os dados, então $D_0 \sim \chi^2(N-q)$ e $D_1 \sim \chi^2(N-p)$ de modo que $\Delta D \sim \chi^2(p-q)$, desde que certas condições de regularidade sejam respeitadas. Se o valor de ΔD for consistente com a distribuição $\chi^2(p-q)$, geralmente é escolhido o modelo M_0 correspondente a H_0 por ser mais simples [6].

Nas situações em que M_0 não descreve bem os dados, D_0 será maior do que o valor esperado de $\chi^2(p-q)$. Isso pode até indicar que D_0 pode ser ajustado para uma distribuição não centralizada de qui quadrado. Essa tendo um grande valor esperado comparado com a distribuição centralizada. Portanto, se o M_1 descreve melhor os dados de maneira que $D_1 \sim \chi^2(N-p)$, então ΔD terá um valor maior que $\chi^2(p-q)$.

Este resultado é usado para testar a hipótese H_1 da seguinte forma: se o valor de ΔD estiver na região crítica (ou seja, maior que o ponto $100 \times \alpha\%$ da cauda superior da distribuição $\chi^2(p-q)$), então rejeita-se H_0 em favor de H_1 com base no fato de que o modelo M_1 fornece uma descrição significativamente melhor dos dados (mesmo que ele também possa não se ajustar particularmente bem aos dados) [6].

2.2.6 Regressão Logística

Na Regressão logística, considera-se que os modelos lineares generalizados tem como resposta variáveis medidas por uma escala binária. De uma maneira geral, as respostas são categorizadas como *sucesso* ou *fracasso*.

Assim, definindo as variáveis aleatórias binárias como

$$Z = \begin{cases} 1, \text{ se o resultado for um sucesso} \\ 0, \text{ se o resultado for uma falha} \end{cases}, \tag{93}$$

sendo as probabilidades dadas por $\Pr(Z=1) = \pi$ e $\Pr(Z=0) = 1-\pi$, que são as propriedades da distribuição de Bernoulli com parâmetro π. Levando em conta n variáveis aleatórias $Z_1, \ldots, Z_n$ independentes com $\Pr(Z_j = 1) = \pi_j$, a função de distribuição de probabilidade é

$$f(z|\pi) = \prod_{j=1}^{n} \pi_j^{z_j} \left(1-\pi_j\right)^{1-z_j} = \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log\left(1-\pi_j\right)\right] \tag{94}$$

que é uma distribuição da família exponencial (17).

Supondo agora que todos os valores de π_j são iguais, reescreve-se Y como

$$Y = \sum_{j=1}^{n} Z_j, \tag{95}$$

com Y sendo o número de sucessos em N tentativas. A variável aleatória Y tem a distribuição $\text{Bin}(n,\pi)$:

$$\Pr(Y=y) = \binom{n}{y} \pi^y (1-\pi)^{N-y}, \quad y = 0, 1, \ldots, n. \tag{96}$$

Conclui-se então que o caso geral para k variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_k$ correspondendo ao número de sucessos nos k subgrupos indicados na Tabela 2.4. Se nesse

caso, $Y_i \sim \text{Bin}(n_i, \pi_i)$, a função de log-verossimilhança é

$$l(\pi_1, \ldots, \pi_k; y_1, \ldots, y_k) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]. \tag{97}$$

Tabela 2.4: Frequências para distribuições Binomiais em N.

	Subgrupos			
	1	2	...	k
Sucessos	Y_1	Y_2	...	Y_k
Falhas	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_k - Y_k$
Totais	n_1	n_2	...	n_k

A partir desse modelo, é estudada a proporção de sucessos, $P_i = \frac{Y_i}{n_i}$ para cada subgrupo e as variáveis respostas que caracterizam esse subgrupo. Desse forma, a Equação (14) agora é definida como $E(Y_i) = n_i \pi_i$ e $E(P_i) = \pi_i$ e a função de ligação (15) é representada por

$$g(\pi_i) = \mathbf{x}_i^T \vec{\beta}, \tag{98}$$

com $\mathbf{x}_i$ sendo um vetor de variáveis explicativas, $\vec{\beta}$ é um vetor de parâmetros.

O MLG mais simples é o modelo linear

$$\pi = \mathbf{x}^T \vec{\beta}, \tag{99}$$

ele é utilizado para aplicações práticas, mas tem uma desvantagem: os valores que compõem $\mathbf{x}^T \mathbf{b}$ podem ser menores que zero ou maiores que um.

Para garantir que a probabilidade π esteja entre o intervalo $[0,1]$, é utilizada a distribuição de probabilidade cumulativa para ajustar o modelo

$$\pi = \int_{-\infty}^{t} f(s) ds. \tag{100}$$

onde $f(x) \geq 0$ e $\int_{-\infty}^{\infty} f(s) ds = 1$. A função $f(s)$ é conhecida como **função de tolerância**, do inglês *tolerance distribution*.

Modelos de Resposta de Dose

Uma das primeiras utilizações do modelo binomial de regressão foi para resultados de bioensaios [7]; esses dados eram proporções ou porcentagens de "sucessos". A finalidade do modelo era descrever a probabilidade de "sucesso", π, em relação a uma função de dose, x, com a função de ligação $g(\pi) = \beta_1 + \beta_2 x$, por exemplo.

Se a função de tolerância $f(s)$ é uma distribuição uniforme Figura 2.6 no intervalo $[c_1, c_2]$

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1}, & \text{se } c_1 \leqslant s \leqslant c_2 \\ 0, & \text{caso contrário} \end{cases}. \tag{101}$$

Lembrando que π é cumulativa

$$\pi = \int_{c_1}^{x} f(s) ds = \frac{x - c_1}{c_2 - c_1} \quad \text{para } c_1 \leqslant x \leqslant c_2. \tag{102}$$

Essa equação tem a forma $\pi = \beta_1 + \beta_2 x$ com

$$\beta_1 = \frac{-c_1}{c_2 - c_1} \text{ e } \beta_2 = \frac{1}{c_2 - c_1}. \tag{103}$$

Este **modelo linear** é equivalente a usar a função identidade como a função de ligação g e impor condições em x, β_1 e β_2 correspondentes a $c1 \leq x \leq c2$ [6]. Devido a essas condições, os métodos mais usuais para estimar os valores de β_1 e β_2 para os MLGs não podem ser aplicados diretamente. Consequentemente, esse modelo não é o mais utilizado.

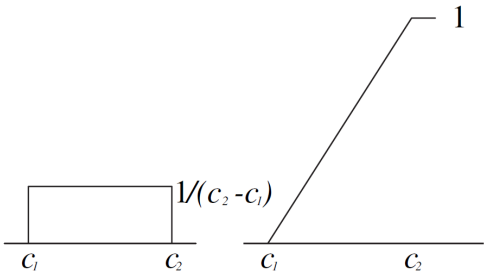

Figura 2.6: Distribuição Uniforme de $f(s)$ e π. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 151]

Assim, o modelo utilizado para estudar os bioensaios é o **modelo probito**. A função de tolerância é dada pela distribuição Normal descrita pela expressão a seguir

$$\begin{aligned} \pi &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right), \end{aligned} \tag{104}$$

com Φ representando a função de probabilidade cumulativa para uma distribuição Normal $\mathrm{N}(0,1)$ dada por:

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 x, \tag{105}$$

onde $\beta_1 = -\frac{\mu}{\sigma}$, $\beta_2 = \frac{1}{\sigma}$ e a função de ligação g é o inverso da função de probabilidade cumulativa da distribuição Normal Φ^{-1}.

Esse modelo probito é usado em diversas áreas da biologia e das ciências sociais. Uma das formas principais é para analisar a dose necessária para matar metade dos animais, ou seja, quando $x = \mu$ e isso é conhecido como Dose Letal Mediana LD(50) [6].

Considerando agora um modelo que demanda um esforço computacional menor e que entrega valores numéricos similares ao modelo probito, é o **modelo Logístico** ou **modelo logit**. A função de tolerância deste modelo é

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2}, \tag{106}$$

assim,

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}. \tag{107}$$

Com isso, a função da ligação é

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x. \tag{108}$$

O termo $\log\frac{\pi}{(1-\pi)}$ é chamado de função logística e indica a probabilidade logarítmica. Além disso, o modelo logístico é vastamente utilizado para dados binários e é encontrado em vários programas estatísticos.

Também existem outros modelos que estudam a resposta de doses, entre eles o modelo que faz uso da distribuição de valores extremos (DVE):

$$f(s) = \beta_2 \exp\left[(\beta_1 + \beta_2 s) - \exp(\beta_1 + \beta_2 s)\right]. \tag{109}$$

Então, tem-se que

$$\pi = 1 - \exp\left[-\exp(\beta_1 + \beta_2 x)\right], \tag{110}$$

consequentemente $\log[-\log(1-\pi)] = \beta_1 + \beta_2 x$, onde a parte da esquerda da expressão é chamada de função complementar log-log. Esse modelo é similar ao logístico e probito em relação aos valores de π quando ele assume valores próximos a $0,5$, mas quando π é próximo de 0 ou 1, o modelo tem um comportamento diferente dos dois modelos apresentados anteriormente.

Para a melhor compreensão desses modelos, será apresentado um exemplo a seguir.

Exemplo: Mortalidade de besouros

Considerando o número de besouros mortos depois de cinco horas de exposição a dissulfeto de carbono gasoso em várias concentrações indicados na Tabela 2.5 [8], e considerando o gráfico da proporção $\mathrm{p}_i = \frac{y_i}{n_i}$ pela dose x_i, que é o logaritmo da quantidade de dissulfeto de carbono Figura 2.7.

Dose, x_i ($\log_{10}\mathbf{CS_2mgL^{-1}}$)	**Número de besouros,** n_i	**Número de mortes,** y_i
1,6907	59	6
1,7242	60	13
1,7552	62	18
1,7842	56	28
1,8113	63	52
1,8369	59	53
1,8610	62	61
1,8839	60	60

Tabela 2.5: Dados da mortalidade de besouros. [Fonte: Bliss, C. I.; 1935]

A partir dessas informações, é ajustado o modelo logístico,

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}. \tag{111}$$

Portanto,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i, \tag{112}$$

e também,

$$\log(1-\pi_i) = -\log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right]. \tag{113}$$

Levando em conta a função de log-verossimilhança (97)

$$l = \sum_{i=1}^{N} \left[y_i(\beta_1 + \beta_2 x_i) - n_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\binom{n_i}{y_i} \right], \tag{114}$$

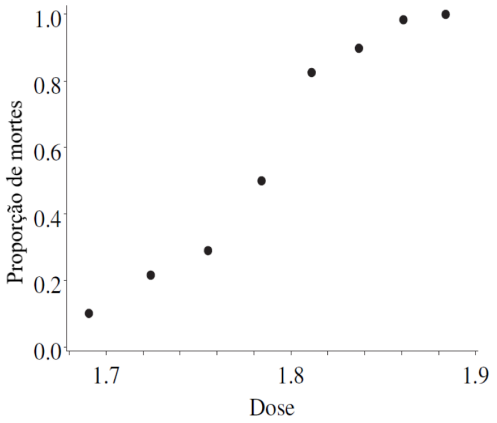

Figura 2.7: Dados da mortalidade de besouros da Tabela 2.5: proporção de mortes, $\mathrm{p}_i = \frac{y_i}{n_i}$, plotado contra dose, $x_i(\log_{10} CS_2 mgl^{-1})$. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 154]

e o score estatístico de β_1 e β_2 é

$$\begin{aligned} U_1 &= \frac{\partial l}{\partial \beta_1} = \sum \left\{ y_i - n_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\} = \sum (y_i - n_i \pi_i) \\ U_2 &= \frac{\partial l}{\partial \beta_2} = \sum \left\{ y_i x_i - n_i x_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\} = \sum x_i (y_i - n_i \pi_i). \end{aligned} \tag{115}$$

Do mesmo modo, a matriz informação Im é

$$\mathfrak{I} = \begin{bmatrix} \sum n_i \pi_i (1 - \pi_i) & \sum n_i x_i \pi_i (1 - \pi_i) \\ \sum n_i x_i \pi_i (1 - \pi_i) & \sum n_i x_i^2 \pi_i (1 - \pi_i) \end{bmatrix}. \tag{116}$$

É utilizado o processo de iteração para calcular o estimador de máxima verossimilhança. Este é representado por

$$\mathrm{Im}^{(m-1)} \mathbf{b}^m = \mathrm{Im}^{(m-1)} \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}, \tag{117}$$

onde (m) indica a m-ésima aproximação e $\mathbf{b}$ é o vetor dos estimadores. Os valores iniciais são $b_1^{(0)} = 0$ e $b_2^{(0)} = 0$, a partir deles são feitas as iterações apresentadas na Tabela 2.6. A tabela mostra também que os valores da função log-verossimilhança aumentam, omitindo a constante $\log \binom{n_i}{y_i}$. Os valores ajustados são calculados em cada iteração por $\hat{\mathrm{y}}_i = n_i \hat{\pi}_i$, que foi considerado inicialmente como $\hat{\pi}_i = 0,5$.

O valor estimado da matriz de variância-covariância para $\mathbf{b}$, $\left[\mathrm{Im}(\mathbf{b})^{-1}\right]$, isso em relação à última iteração, é indicado no final da Tabela 2.6 em conjunto com o desvio

$$D = 2 \sum_{i=1}^{N} \left[y_i \log \left(\frac{y_i}{\widehat{y_i}} \right) + (n_i - y_i) \log \left(\frac{n - y_i}{n - \widehat{y_i}} \right) \right]. \tag{118}$$

Os estimadores e seus erros são dados por

$$\begin{aligned} & b_1 = -60.72, \quad \text{erro } = \sqrt{26.840} = 5.18 \\ \text{e} \quad & b_2 = 34.27, \quad \text{erro } = \sqrt{8.481} = 2.91. \end{aligned} \tag{119}$$

Se o ajuste de modelo é bom em relação aos dados, o desvio deve ter uma distribuição $\chi^2(6)$, já que existem $N = 8$ padrões de covariáveis e $p = 2$ parâmetros [6]. Neste caso, o valor

	Estimativa	Aproximação			
	Inicial	Primeira	Segunda	$\cdots$	Sexta
β_1	0	-37.856	-53.853	$\cdots$	-60.717
β_2	0	21.337	30.384	$\cdots$	34.270
log-verossimilhança	-333.404	-200.010	-187.274	$\cdots$	-186.235
Observações			Valores Ajustados		
$y_1 = 6$	29.5	8.505	4.543	$\cdots$	3.458
$y_2 = 13$	30.0	15.366	11.254	$\cdots$	9.428
$y_3 = 18$	31.0	24.808	23.058	$\cdots$	22.451
$y_4 = 28$	32.5	34.949	35.036	$\cdots$	35.707
$y_5 = 52$	29.5	46.741	51.705	$\cdots$	52.941
$y_6 = 53$	29.0	45.939	58.061	$\cdots$	59.492
$y_7 = 61$	30.0	56.573	59.306	$\cdots$	59.978
$y_8 = 60$	30.0	54.734	58.036	$\cdots$	58.743

$$[\mathscr{J}(\mathbf{b})]^{-1} = \begin{bmatrix} 26.840 & -15.082 \\ -15.082 & 8.481 \end{bmatrix}, \quad D = 11.23$$

Tabela 2.6: Ajustando o modelo logístico em relação aos dados da mortalidade de besouros. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 156]

calculado de D é muito diferente do valor "esperado"de 6 que é $12,59$. Isso sugere que esse modelo não é o melhor ajuste para esses dados.

Em relação à mortalidade dos besouros, outros modelos têm um melhor ajuste para esses dados, dentre eles o modelo de valores extremos (DVE), cuja relação é descrita matematicamente pela Equação 120. Este modelo utiliza a função de ligação log-log complementar (Equação 121):

$$\pi_i = 1 - \exp(-\exp(\alpha + \beta x_i)) \tag{120}$$

$$\log(-\log(1 - \pi_i)) = \alpha + \beta x_i \tag{121}$$

O ajuste desses dados para outros modelos, incluindo o DVE, é apresentado na Tabela 2.7.

Valor observado de Y	Modelo Logístico	Modelo Probit	Modelo de Valores Extremos
6	3.46	3.36	5.59
13	9.84	10.72	11.28
18	22.45	23.48	20.95
28	33.90	33.82	30.37
52	50.10	49.62	47.78
53	50.18	50.08	54.14
61	59.22	59.66	61.11
60	58.74	59.23	59.95
D	11.23	10.12	3.45
b_1(s.e.)	-60.72(5.18)	-34.94(2.64)	-39.57(3.23)
b_2(s.e.)	34.27(2.91)	19.73(1.48)	22.04(1.79)

Tabela 2.7: Comparação do número de mortes com os valores ajustados a partir de vários modelos de resposta de dose para a mortalidade de besouros. As estatísticas dos desvios também são apresentadas.[Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 158]

Modelos Generalizados de Regressão Logística

O modelo usado na seção (2.2.6) é o Logístico Linear Simples $\log\left[\frac{\pi_i}{(1-\pi_i)}\right] = \beta_1 + \beta_2 x$ é um caso especial do modelo de regressão logística generalizada

$$\operatorname{logit}\pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\vec{\beta}, \tag{122}$$

com $\mathbf{x}_i$ sendo um vetor composto por covariáveis e variáveis *dummy* e $\vec{\beta}$ é o parâmetro desse vetor. Esse modelo é usado amplamente para analisar dados binários ou respostas binomiais, já que é uma técnica mais aprimorada comparada com a regressão múltipla.

Para obter a máxima verossimilhança dos valores estimados dos parâmetros $\vec{\beta}$ e suas probabilidades $\pi_i = g^{-1}(\mathbf{x}_i^T\vec{\beta})$ é feita a maximização da função log-verossimilhança.

$$l(\pi;\mathbf{y}) = \sum_{i=1}^{N}\left[y_i\log\pi_i + (n_i - y_i)\log(1-\pi_i) + \log\binom{n_i}{y_i}\right]. \tag{123}$$

O desvio é dado por

$$D = 2\sum_{i=1}^{N}\left[y_i\log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)\right]. \tag{124}$$

Simplificando,

$$D = 2\sum o\log\frac{o}{e}, \tag{125}$$

com o sendo os "sucessos"y_i e as "falhas"$(n_i - \mathrm{y}_i)$, isso, retirado da Tabela 2.4 e e corresponde à frequência esperada ou aos valores ajustados $\hat{\mathrm{y}}_i = n_i\hat{\pi}_i$ e $(n_i - \hat{\mathrm{y}}_i) = (n_i - n_i\hat{\pi}_i)$.

Importante ressaltar que D não tem relação com nenhum parâmetro de perturbação, então a qualidade do ajuste pode ser avaliada por um teste de hipótese utilizando a aproximação

$$D \sim \chi^2(N-p), \tag{126}$$

onde p é o número de parâmetros estimados e N o número de covariáveis.

Qualidade do Ajuste Estatístico

Para avaliar a qualidade do ajuste, ao invés de utilizar a máxima verossimilhança, é possível estimar os parâmetros pelo método dos mínimos quadrados, cuja principal vantagem é a simplicidade computacional: o método dos mínimos quadrados é de fácil implementação e geralmente resulta em soluções analíticas diretas, o que facilita a interpretação dos resultados e reduz o tempo de cálculo, especialmente em modelos lineares e com conjuntos de dados grandes [6]

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}, \tag{127}$$

já que $\mathrm{E}(Y_i) = n_i\pi_i$ e $(var)(Y_i) = n_i\pi_i(1-\pi_i)$. Isso é o mesmo processo que minimizar a estatística qui-quadrado de Pearson

$$\chi^2 = \sum \frac{(o-e)^2}{e}, \tag{128}$$

com o sendo a frequência observada e e representa as frequências esperadas e a soma é sobre todas as células $2 \times N$ da tabela. Reescrevendo a Equação (128)

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)} \\ &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_w. \end{aligned} \tag{129}$$

Quando χ^2 é avaliado em relação às frequências estimadas, a estatística é

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}, \tag{130}$$

como (129) é equivalente ao desvio (124)

$$D = 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i\widehat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i}\right) \right]. \tag{131}$$

A relação entre X^2 e D é dada a partir da expansão da série da Taylor de $s\log\left(\frac{s}{t}\right)$

$$s\log\frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \ldots \quad . \tag{132}$$

Assim,

$$\begin{aligned} D =& 2\sum_{i=1}^{N} \left\{ (y_i - n_i\widehat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i} + [(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)] + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)]^2}{n_i - n_i\widehat{\pi}_i} + \ldots \right\} \\ \cong& \sum_{i=1}^{N} \frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} = \chi^2. \end{aligned} \tag{133}$$

Sob a hipótese de modelo correto, a distribuição assintótica de D é aproximadamente qui-quadrado com $N-p$ graus de liberdade $(\chi^2(N-p))$. Embora D e χ^2 sejam usados como

medidas de ajuste, X^2 é frequentemente preferido, pois D é sensível a frequências muito pequenas [6]. Se as frequências esperadas forem baixas (< 1) ou as variáveis explicativas contínuas, esses testes não são as melhores opções para se realizar uma avaliação de modelo. Nesses casos, o método de Hosmer-Lemeshow é recomendado, já que as observações são agrupadas por probabilidades preditas (aprox. 10 grupos), e um teste qui-quadrado ($\chi^2 HL$) é aplicado à tabela de contingência resultante ($g \times 2$), com distribuição aproximada $\chi^2(g-2)$ [6].

Em algumas situações, a função de log-verossimilhança do modelo ajustado é comparada com a função de log-verossimilhança do modelo inicial, no qual os valores de π_i são iguais. Considerando, então, que o modelo inicial é $\tilde{\pi} = \frac{(\sum y_i)}{(\sum n_i)}$ e que $\hat{\pi}_i$ é a probabilidade estimada de Y_i em relação ao modelo de interesse, a estatística é definida por

$$C = 2[l(\widehat{\pi}; \mathbf{y}) - l(\widetilde{\pi}; \mathbf{y})], \tag{134}$$

em que l é a função de log-verossimilhança. Assim, reescrevemos a Equação (131) como

$$C = 2 \sum \left[y_i \log \left(\frac{\widehat{y}_i}{n_i \widetilde{\pi}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - \widehat{y}_i}{n_i - n_i \widetilde{\pi}_i} \right) \right], \tag{135}$$

em que C é conhecido como a razão de verossimilhança da estatística de qui-quadrado. Por analogia com R^2 para a regressão múltipla, outra estatística pode ser usada para fazer a análise

$$\text{pseudo } R^2 = \frac{l(\tilde{\pi}; \mathbf{y}) - l(\widehat{\pi}; \mathbf{y})}{l(\tilde{\pi}; \mathbf{y})}. \tag{136}$$

A relação acima representa a melhoria na função de log-verossimilhança a partir dos termos no modelo de interesse, em comparação com o modelo inicial.

Agora, em relação a regressão logística, R^2 assume valores bem pequenos mesmo quando outros meios de validar a qualidade do ajuste indicam que o modelo é adequado. O motivo pelo qual isso ocorre é que o pseudo R^2 é uma medida de predição de uma variável explicativa de Y_i, não uma medida de predição em relação a todas as variáveis explicativas [9].

O Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano ou Schwartz são outras maneiras de analisar a qualidade do ajuste baseada na função log-verossimilhança ajustada para os parâmetros estimados e para os dados em análise [6]. Isso é expresso por

$$\begin{aligned} AIC &= -2l(\widehat{\pi}; \mathbf{y}) + 2p \\ BIC &= -2l(\widehat{\pi}; \mathbf{y}) + p \times \ln(n). \end{aligned} \tag{137}$$

É importante notar que as estatísticas (exceto para pseudo R^2) discutidas nesta seção resumem o quão bem um modelo específico se ajusta aos dados. Portanto, um valor pequeno da estatística e, consequentemente, um valor p grande, indica que o modelo se ajusta bem [6].

Descrição	Valor
Log-verossimilhança sem variáveis	$l(\tilde{\pi}; \mathrm{y}) = -322,72$
Log-verossimilhança com variáveis	$l(\hat{\pi}; \mathrm{y}) = -186,2354$
Estatística C	$272,970$
Pseudo R^2	$0,4229$
AIC	$41,430$

Tabela 2.8: Resumo dos parâmetros do modelo logístico da mortalidade de besouros

Levando em conta o modelo logístico da mortalidade de besouros, a Tabela 2.8 indica os resultados do modelo. A estatística C com um grau de liberdade, mostra que o parâmetro β_1 é necessário para o ajuste do modelo. O valor do pseudo R^2 mostra que não é o melhor ajuste para os dados.

Resíduos

Para a regressão Logística existem duas maneiras de calcular os resíduos do ajuste indicados por D e χ^2, e para cada covariável m, será calculado m resíduos. Levando em consideração que Y_k é o número de sucessos, n_k o número de testes e $\hat{\pi}_k$ a previsão da probabilidade de sucesso para as k-ésimas covariáveis, o Pearson, ou quí-quadrado, ou resíduo é dado por

$$X_k = \frac{(y_k - n_k\widehat{\pi}_k)}{\sqrt{n_k\widehat{\pi}_k\,(1-\widehat{\pi}_k)}} \quad ,k = 1,\ldots,m. \tag{138}$$

A partir de (130), $\sum_{k=1}^{m} X_k^2 = X^2$, os resíduos padronizados de Pearson são

$$r_{Pk} = \frac{X_k}{\sqrt{1-h_k}}. \tag{139}$$

Assim, o desvio dos resíduos é expressado por

$$d_k = \text{sign}\,(y_k - n_k\widehat{\pi}_k)\left\{2\left[y_k \log\left(\frac{y_k}{n_k\widehat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\widehat{\pi}_k}\right)\right]\right\}^{1/2}. \tag{140}$$

Essas expressões para calcular os resíduos e os desvios desses resíduos podem ser usadas para checar se o modelo é adequado ou não. Se os dados de uma amostra forem binários, então alguns valores dos resíduos e o gráfico deles podem ser não uniformes. Nessa situação, será necessário analisar o ajuste pelo χ^2, D e outros diagnósticos que serão apresentados na próxima seção.

Outros diagnósticos

Para dados binomiais ou binários, é importante checar a escolha feita para a função de ligação. Assim, para analisar essa escolha, será considerada uma família mais geral das funções de ligação [10]

$$g(\pi,\alpha) = \log\left[\frac{(1-\pi)^{-\alpha} - 1}{\alpha}\right]. \tag{141}$$

Se $\alpha = 1$, então $g(\pi) = \log\left[\frac{\pi}{(1-\pi)}\right]$, que é a função de ligação logit. Se $\alpha \to 0$, então $\alpha \to \log[-\log(1-\pi)]$ que é a função de ligação complementar log-log. Sabendo que o valor de α é estimado a partir dos dados analisados.

Outro problema para ajustar o modelo para dados binomiais é a sobredispersão, que é quando a variabilidade observada em um conjunto de dados é maior do que a esperada com base em um modelo estatístico específico. Um dos indicadores desse problema é quando o valor de D é muito maior que o maior valor esperado de $N - p$. Para solucionar isso, é incluído um parâmetro extra ϕ no modelo para que a variância seja $\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$.

Exemplo: senilidade e WAIS

Os dados da Tabela 2.9 apresentam os dados de sintomas de selenilidade e o resultado do teste de escala de WAIS em uma amostra de pessoas idosas. Como observado, os dados dessa tabela são binários e é composto por $m = 17$ covariáveis. Considerando Y_i sendo o número de pessoas com sintomas dentro de n_i pessoas com i-ésima covariável. O modelo de regressão logística

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i; \quad Y_i \sim \text{Bin}\,(n_i,\pi_i) \quad i = 1,\ldots,m, \tag{142}$$

x	s	x	s	x	s	x	s	x
9	1	7	1	7	0	17	0	13
13	1	5	1	16	0	14	0	13
6	1	14	1	9	0	19	0	9
8	1	13	0	9	0	19	0	15
10	1	16	0	11	0	11	0	10
4	1	10	0	13	0	10	0	11
4	1	9	1	15	0	12	0	12
8	1	11	0	14	0	10	0	12
11	1	15	0	9	1	16	0	4
7	1	15	0	1	0	16	0	20
9	1	18	0	6	0	14	0	0

Tabela 2.9: Sintomas de selenidade ($s = 1$ se os existem os sintomas, caso contrário $s = 0$) e escala WAIS (x) para $N = 54$ pessoas. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 168]

com π_i é a probabilidade de uma pessoa apresentar sintomas. Dessa forma, a expressão é ajustado a partir dos seguintes resultados [6]:

$$\begin{aligned} &b_1 = 2,404;\ \text{erro}\ (b_1) = 1,192; \\ &b_2 = -0,3235;\ \text{erro}\ (b_2) = 0,1140; \\ &\chi^2 = \sum X_i^2 = 8,083\ \text{e}\ D = \sum d_i^2 = 9,419. \end{aligned} \tag{143}$$

A Figura 2.8 mostra que as frequências relativas $\frac{y_i}{n_i}$ para cada covariável e para o ajuste $\hat{\pi}_i$ plotado contra a escala de WAIS.

A Tabela 2.10 indica os valores de covariáveis, as estimativas de $\hat{\pi}_i$ e a distribuição de qui-quadrado e o desvio dos resíduos calculados.

Devido ao pequeno número de observações por valor de covariável, a avaliação dos resíduos foi complementada pela aplicação do teste de Hosmer-Lemeshow. Esta abordagem consiste em agrupar as observações em categorias com base nas probabilidades estimadas $(\hat{\pi})$, visando ter um número aproximadamente igual de observações por grupo. Para esta análise, foram criadas 3 categorias. As frequências observadas e esperadas foram comparadas, resultando em uma estatística de Hosmer-Lemeshow $X^2_{HL} = 1.15$. Comparada a uma distribuição $\chi^2(1)$, este valor não é significativo, sugerindo que o modelo proposto se ajusta adequadamente aos dados.

Para o modelo inicial, sem covariável x, o valor máximo da função de log-verossimilhança é $l(\tilde{\pi}, \mathrm{y}) = -30,9032$. Agora, em relação ao modelo com x, o valor é $l(\tilde{\pi}, \mathrm{y}) = -25,5087$. Em seguida, encontra-se o valor de $C = 10,789$, que é importante mostrando que o parâmetro de inclinação não é zero. Por fim para avaliar se o modelo apresenta o melhor ajuste para os dados, é calculado $R^2 = 0,17$. Esse valor sugere que esse modelo não apresenta um bom ajuste para os dados analisados.

x	y	n	$\hat{\pi}$	X	d
4	1	2	0.752	-0.826	-0.766
5	1	1	0.687	0.675	0.866
6	1	2	0.614	-0.330	-0.326
7	2	3	0.535	0.458	0.464
8	2	2	0.454	1.551	1.777
9	2	6	0.376	-0.214	-0.216
10	1	6	0.303	-0.728	-0.771
11	1	6	0.240	-0.419	-0.436
12	0	2	0.186	-0.675	-0.906
13	1	6	0.142	0.176	0.172
14	2	7	0.107	1.535	1.306
15	0	3	0.080	-0.509	-0.705
16	0	4	0.059	-0.500	-0.696
17	0	1	0.043	-0.213	-0.297
18	0	1	0.032	-0.181	-0.254
19	0	1	0.023	-0.154	-0.216
20	0	1	0.017	-0.131	-0.184
$\sum$	14	54		8.084*	9.418*

* As somas dos quadrados diferem ligeiramente das estatísticas de qualidade de ajuste X^2 e D mencionadas no texto devido a erros de arredondamento.

Tabela 2.10: Padrões de covariáveis e respostas, probabilidades estimadas ($\hat{\pi}$), resíduos de Pearson (X) e resíduos de desvio (d) para senilidade e WAIS. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 169]

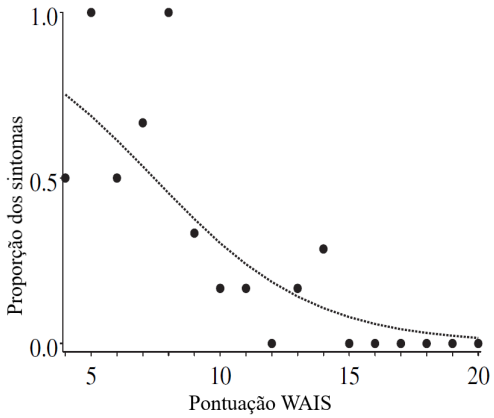

Figura 2.8: Associação entre a presença de sintomas e o escore WAIS a partir dos dados nas Tabelas 2.9 e 2.10; os pontos representam as proporções observadas e a linha pontilhada representa as probabilidades estimadas. [Fonte: Dobson, A. J. e Barnett, A. G.; 2018, p. 170]

Simulações de Dados

3.1 Regressão Múltipla

Para ilustrar na prática os conceitos apresentados de regressão linear múltipla, foi desenvolvido um exemplo baseado em uma situação comum na área da saúde e da física médica. Assim é possível compreender, a partir de dados simulados, como diferentes variáveis podem influenciar uma medida clínica biomédica, nesse caso, a dose absorvida em determinado tecido biológico.

A ideia da simulação é ajudar entender a relação entre a variável dependente (y), que corresponde à dose absorvida em $n = 100$ amostras de tecido, sendo influenciada pela intensidade de radiação incidente (x_1) que foi simulada pela distribuição normal, por um marcador biológico binário associado à amostra (x_2) que foi simulado pela distribuição binomial, e pela espessura do tecido (x_3) que foi simulada pela distribuição uniforme. Além disso, incluiu-se no modelo um termo de interação ($x_2 \times x_3$), com o intuito de explorar se a influência do marcador biológico depende também das características físicas do tecido.

Para a aplicação, o intercepto foi definido como $\alpha = 10$, e os parâmetros β assumem os valores $\beta_1 = 0,8$, $\beta_2 = -5$ e $\beta_3 = 0,5$. Além disso, o termo de erro ε foi simulado a partir de uma distribuição normal.

O modelo escolhido para essa análise é dado por

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_2 \times x_3) + \varepsilon, \tag{144}$$

em que

- Um coeficiente β_1 positivo indica que, quanto maior a intensidade da radiação, maior será a dose absorvida pelo tecido.
- O coeficiente β_2 negativo indica que a presença do marcador biológico está associada a uma redução na absorção média, ajustando-se para os demais fatores.
- Já a interação $\beta_3(x_2 \times x_3)$, permite avaliar se o efeito do marcador sobre a absorção é dependente da espessura, trazendo um componente importante de aplicabilidade direta à área de dosimetria.

Esse tipo de abordagem, ainda que baseada em dados simulados, reflete situações reais observadas tanto em pesquisas experimentais quanto em aplicações clínicas, reforçando a utilidade dos métodos estatísticos multivariados para investigar relações complexas no contexto das ciências da saúde.

Após a simulação dos dados e o ajuste do modelo de regressão múltipla, foi possível realizar uma análise dos resultados, consequentemente, avaliando os pressupostos do modelo e de sua capacidade preditiva a partir das variáveis escolhidas. Os comandos encontram-se no Anexo 1 (A).

As estimativas dos coeficientes do modelo ajustado, seus respectivos erros padrão, valores de t, valores p e intervalos (IC) de confiança de 95% encontram-se na Tabela 3.1. Observa-se que o coeficiente para a intensidade de radiação (x_1) tem um efeito positivo e altamente significativo. Já o marcador biológico (x_2) e termo de iteração entre o ele e a espessura do tecido x_3 não se mostraram significância estatística, mesmo que x_2 resultou em um coeficiente negativo.

O coeficiente de determinação ajustado foi $R^2_{ajust} = 0{,}66$, o desvio padrão residual ficou em 4,74, e o teste F global indicou significância ($F = 65{,}74$, $p < 2{,}2 \times 10^{-16}$), indicando que o modelo não é a melhor escolha para representar a situação proposta.

O teste de homocedasticidade (ncvTest) resultou em $p = 0{,}97$, reforçando a ausência de dependência da variância dos resíduos em relação aos valores ajustados. Os valores de VIF (Fator de Inflação da Variância, do inglês *Variance Inflation Factor*) abaixo de 5 para todos os preditores descartam colinearidade significativa. Dessa forma, os testes indicam que todos os pressupostos do modelo de regressão múltipla foram respeitados para os dados simulados.

Parâmetro	Estimativa	Erro Padrão	t	p-valor	IC 95%		VIF
Intercepto	13,94	2,71	5,15	<0,001	8,57	19,32	–
β_1	0,73	0,05	14,00	<0,001	0,63	0,84	1,00
β_2	-2,48	2,12	-1,17	0,245	-6,69	1,73	4,75
β_3	0,12	0,31	0,38	0,706	-0,49	0,73	4,75

Tabela 3.1: Estimativas dos parâmetros do modelo de regressão múltipla ajustado nos dados simulados com VIF dos preditores. [Fonte: Próprio autor]

Os principais gráficos diagnósticos do modelo ajustado estão apresentados na Figura 3.1.

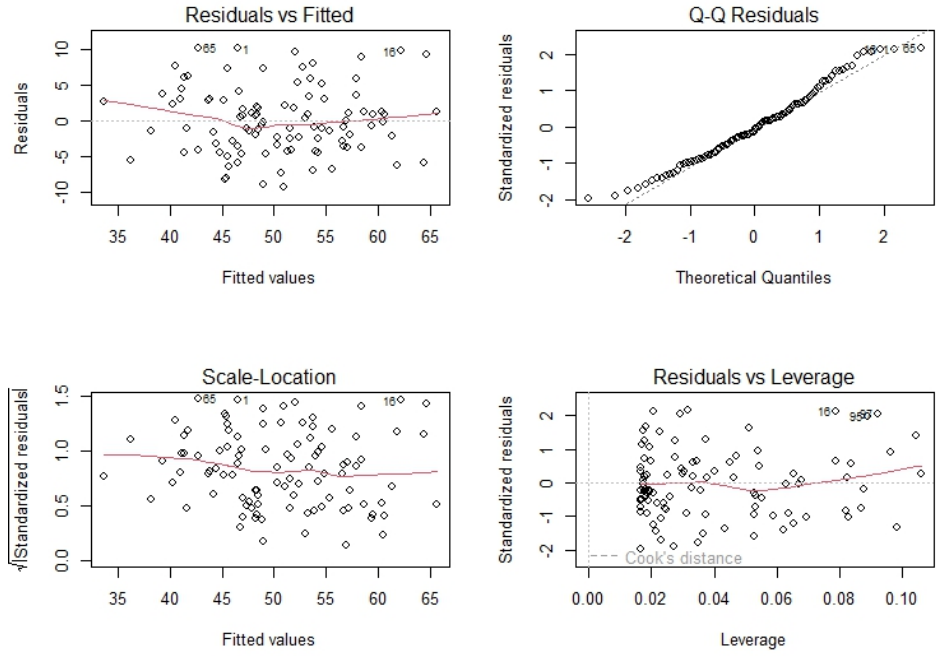

Figura 3.1: Diagnóstico dos pressupostos da regressão linear múltipla.[Fonte: Próprio autor]

- **Resíduos vs Ajustados (*Fitted*)**: Esse gráfico mostra que os resíduos estão distribuídos sem padrão sistemático em torno de zero, indicando a linearidade adequada do ajuste.

- **Gráfico Q-Q dos resíduos**: Os pontos seguem a linha de referência teórica de normalidade, reforçando o resultado do teste de Shapiro-Wilk ($p = 0{,}13$) e mostra que os resíduos tem uma distribuição normal.
- **Locação-escala (*Scale-Location*)**: Indica que a variância dos resíduos permanecem constante ao longo dos valores ajustados, não sendo observados padrões incomuns de dispersão, o que indica a homocedasticidade.
- **Resíduos vs *Leverage***: O gráfico mostra ausência de pontos influentes com alto *leverage* e resíduos extremos, indicando ausência de *outliers*.

A Figura 3.2 ilustra a relação entre as variáveis principais do exemplo.

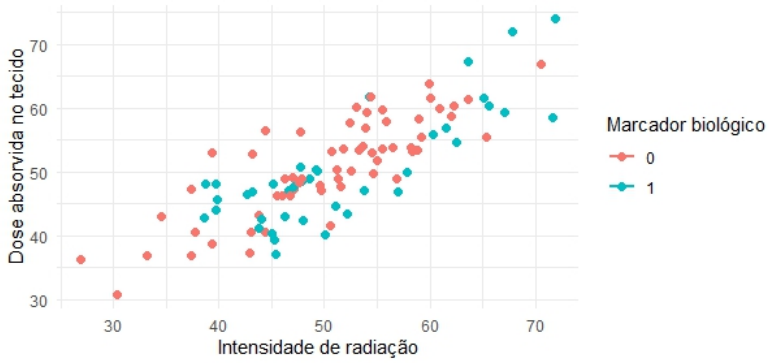

Figura 3.2: Dispersão dos dados simulados e linhas de regressão por grupo definido pelo marcador biológico.[Fonte: Próprio autor]

O gráfico evidencia a tendência linear positiva entre a intensidade de radiação (x_1) e a dose absorvida no tecido (y), para ambos os grupos definidos pelo marcador biológico (x_2). Embora o marcador biológico não tenha sido significativo na seleção final do modelo, é possível notar visualmente a separação entre os grupos, e que o comportamento dos pontos se mantém crescente em ambos, validando a coerência do ajuste.

Em resumo, o modelo ajustado mostra-se adequado para explicar os efeitos simulados, respeitando todos os pressupostos estatísticos da regressão múltipla. A partir dos valores fornecidos pelo R, a tabela de coeficientes e as visualizações gráficas indicam a qualidade do ajuste, ao mesmo tempo que mostram a importância de interpretar da significância de cada variável.

3.2 Modelos Lineares Generalizados de Poisson

O modelo de Poisson é amplamente utilizado em pesquisas na área da saúde para modelar eventos de contagem, como número de internações hospitalares em determinado período. Para ilustrar sua aplicação, foi realizada a simulação de dados representando internações em função da idade do paciente e da presença de um marcador biológico.

Foram simulados dados de $n = 150$ pacientes, com idade média de 50 anos (desvio padrão 12) e cerca de 40% apresentando um marcador biológico positivo. O código em R encontram-se no Anexo 2 (A). O número de internações foi gerado a partir de uma distribuição de Poisson com taxa λ parametrizada segundo o modelo:

$$\lambda_i = \exp(-1{,}2 + 0{,}03 \cdot \text{idade}_i + 0{,}7 \cdot \text{marcador}_i). \tag{145}$$

Parâmetro	Estimativa	Erro Padrão	z	p	IC 95%	
Intercepto	-1,12	0,28	-4,05	$5,1 \times 10^{-5}$	-1,67	-0,58
idade	0,029	0,0049	5,92	$3,3 \times 10^{-9}$	0,019	0,039
marcador	0,615	0,119	5,18	$2,2 \times 10^{-7}$	0,382	0,848

Tabela 3.2: Estimativas dos parâmetros do modelo de Poisson ajustado. [Fonte: Próprio autor]

O ajuste do modelo pelo procedimento glm com família Poisson no programa R resultou nas estimativas apresentadas na Tabela 3.2.

Ambas as variáveis explicativas mostraram-se estatisticamente significativas. O coeficiente estimado para idade (0,029) sugere que, para cada ano a mais, a taxa esperada de internações se eleva aproximadamente 3% (pois $e^{0,029} \approx 1,03$), mantendo fixo o marcador. Já pacientes positivos para o marcador biológico apresentam, em média, uma taxa de internação 1,85 vez maior do que aqueles sem o marcador ($e^{0,615} \approx 1,85$).

O intervalo de confiança de 95% para cada coeficiente confirma a robustez dos achados. O parâmetro de dispersão obtido pelos resíduos de Pearson ($\hat{\phi} = 0,96$) indica ausência de superdispersão no modelo, validando o uso da abordagem Poisson.

A Figura 3.3 ilustra a relação entre idade, marcador biológico e número previsto de internações, destacando a diferença de risco entre os grupos.

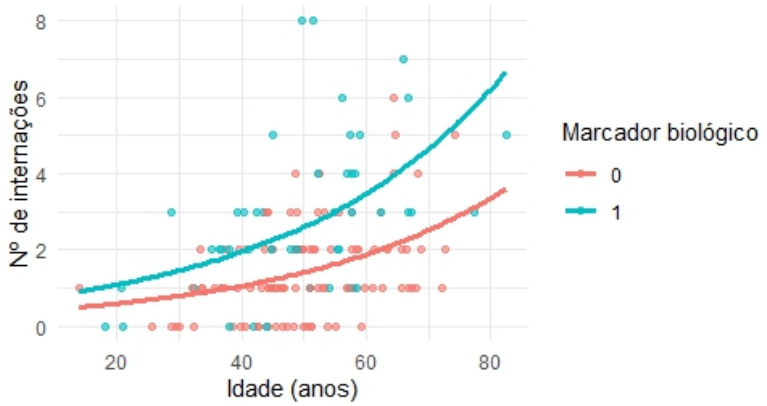

Figura 3.3: Relação entre idade, número de internações e marcador biológico no modelo de Poisson simulado. [Fonte: Próprio autor]

Esse exemplo evidencia a utilidade do modelo de Poisson em epidemiologia e saúde, permitindo estimar de forma quantitativa o efeito de fatores clínicos sobre eventos de contagem. O modelo, ajustado a partir de dados simulados, reforça a validade do raciocínio estatístico como ferramenta de apoio à tomada de decisão em contextos biomédicos.

3.3 Regressão Logística

O modelo de Regressão Logística (RL) tem sido amplamente utilizado em pesquisas na área médica, principalmente nas duas últimas décadas. Ele é usado sobretudo quando a amostra analisada tem uma resposta binária, sendo ela obtida a partir de uma ou mais variáveis independentes. Dentre os usos da RL, pode-se citar estudos dos fatores que preveem se haverá uma melhora ou não na saúde de pacientes após uma intervenção, explorar os efeitos e as relações

entre múltiplas variáveis independentes e para determinar se variáveis recém-exploradas aprimoram modelos já estabelecidos. Além disso, a RL também está sendo aplicada na área da física médica, realizando diagnósticos a partir da análise de imagens. Esse tipo de uso é muito relevante, já que pode evitar procedimentos invasivos nos pacientes e ainda auxiliar os médicos na análise das imagens. O objetivo desta etapa é analisar, a partir da simulação de dados, como o modelo de regressão logística relaciona um parâmetro retirado de uma imagem médica com o diagnóstico do paciente.

A Regressão Logística é um modelo estatístico útil para prever a probabilidade de ocorrência de uma variável dependente binária, como a probabilidade de um tumor ser maligno, a partir de variáveis independentes. Em um estudo recente, esse modelo foi usado para analisar imagens de ultrassom de tireoide [1]. Com base nisso, foi construído um código, presente no Anexo 3 (A), no RStudio utilizando o modelo dado por:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \tag{146}$$

em que π_i é a probabilidade de sucesso e x_i a variável categórica que representa as características vasculares de uma determinada região do corpo. Primeiro, foi feita a simulação dos dados, criando uma amostra binária, o diagnóstico (variável resposta), onde 0 são os tumores benignos e 1 os malignos, e uma amostra da vascularidade (variável explicativa) com três tipos de níveis: 0 indica baixa vascularidade, 1 é a média vascularidade e 2 é a alta vascularidade. Em seguida, foi feito a estimação dos parâmetros $\beta_0 = -2,5$ e $\beta_1 = 1,5$ pela função de verossimilhança, que é maximizada para encontrar os valores de β que melhor se adaptam aos dados simulados. Depois, o modelo de regressão logística é ilustrado pelo valor predito da probabilidade acompanhado do respectivo intervalo de confiança.

Tabela 3.3: Parâmetros estimados do modelo GLM para vascularidade e malignidade tumoral

	β	EP	p-valor
Baixa (intercepto)	$-2,2824$	0,3319	$< 0,001$
Média	1,1429	0,4389	0,0092
Alta	1,1429	0,3319	$< 0,001$

Depois da aplicação do modelo, foi utilizado o comando glm para estimar os parâmetros da vascularidade em relação à malignidade de um tumor. Os parâmetros indicados na Tabela 3.3, mostram que eles são estatisticamente relevante para determinar o diagnóstico de um tumor.

Pela Figura 3.6, também é possível afirmar que quanto mais baixa for a vascularidade, maior as chances de o tumor ser benigno; consequentemente, quanto maior a vascularidade, maior a probabilidade do tumor ser maligno.

A qualidade do ajuste do modelo foi avaliada com o teste de Hosmer-Lemeshow, em que resultado apresentou um valor de $X^2 = 2.05 \times 10^{-14}$ com $df = 0$ e p-valor $< 2.2 \times 10^{-16}$. Esse resultado é esperado em uma simulação com poucos grupos distintos e pode indicar que a divisão automática dos dados não permitiu formar o número desejado de grupos no teste, uma limitação comum quando a variável explicativa é categórica com poucos níveis. Entretanto, os outros resultados mostram um bom ajuste do modelo com os dados simulados devido ao desvio residual de 170.03 e a redução comparada ao desvio nulo (213.27), além do AIC informado (176.03), que indicam melhora do ajuste com a inclusão da variável vascularidade.

Em relação aos resíduos, foram analisados os resíduos de Pearson e deviance, apresentados nos gráficos (3.5) e (3.4), respectivamente. Os resíduos de Pearson mostraram distribuição próxima de zero, sem padrões visíveis de heterocedasticidade. Os resíduos deviance também

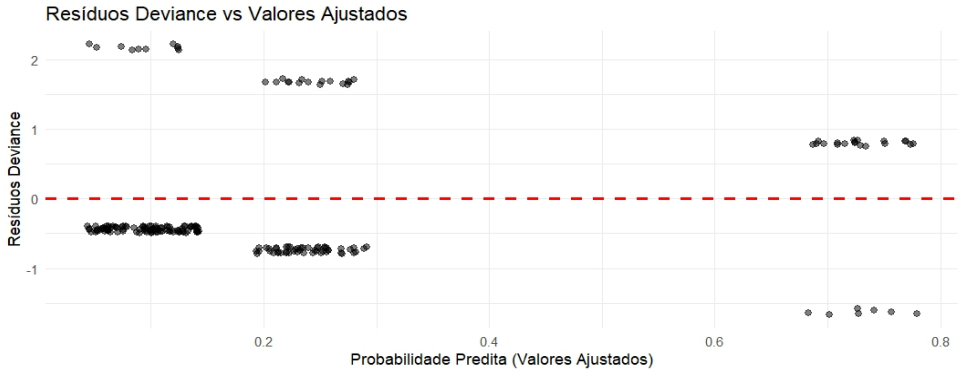

Figura 3.4: Gráfico do Resíduo *deviance* vs Valores ajustados.[Fonte: Próprio autor]

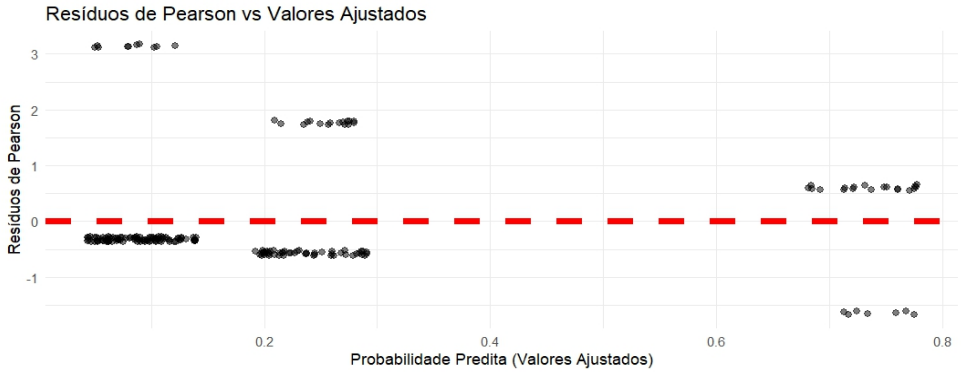

Figura 3.5: Gráfico do Resíduo de *Pearson* vs Valores ajustados.[Fonte: Próprio autor]

demonstraram dispersão semelhante, indicando que o modelo ajusta bem as probabilidades previstas aos dados observados, sem revelar a presença de pontos discrepantes. Estes resultados reforçam a adequação do modelo de regressão logística para os dados simulados.

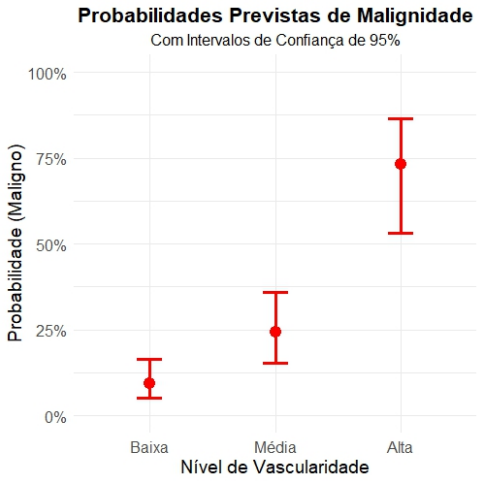

Figura 3.6: Gráfico de predição do tumor maligno para cada nível de vascularidade com intervalo de confiança de 95%. [Fonte: Próprio autor]

A Figura 3.6, por meio dos intervalos de confiança, mostra o efeito da vascularidade sobre a probabilidade de malignidade, conforme predito pelo modelo logístico, reiterando que tumores com menor vascularidade têm menor chance de serem malignos, enquanto alta vascularidade

aumenta significativamente esta probabilidade.

Dessa forma, o modelo de regressão logística é um procedimento eficiente para analisar a relação entre probabilidade de malignidade de um tumor e a sua vascularidade. A partir dos dados simulados, a análise possibilitou a ilustração do aumento da probabilidade de um tumor ser maligno conforme a sua vascularização, a partir da probabilidade predita. Dessa forma, a regressão logística se mostra uma ferramenta valiosa e promissora para auxiliar a interpretação de imagens médicas, fornecendo uma base quantitativa para o diagnóstico do paciente.

Conclusão

Ao longo deste projeto, foi possível compreender diferentes modelos estatísticos e suas diversas aplicações por meio da simulação de dados. Entre essas aplicações, destacou-se o uso desses modelos na área da saúde, especialmente no contexto da física médica.

Considerando as simulações realizadas no software R e a análise dos modelos ajustados, observou-se como diferentes fatores podem influenciar a variável dependente, além da avaliação dos ajustes por meio de diagnósticos estatísticos, como os testes de hipóteses.

Durante o desenvolvimento do estudo para a elaboração deste relatório, houve um aprofundamento dos conceitos sobre modelos lineares generalizados, com ênfase no modelo logístico. Além disso, foi possível compreender melhor a importância da simulação de dados e seu papel fundamental no auxílio à análise estatística.

A simulação de dados mostra-se especialmente relevante, pois facilita o desenvolvimento de estudos ao eliminar a necessidade de coleta de dados reais, um processo que geralmente demanda altos custos financeiros e requer aprovação em comitês de ética. Conforme demonstrado pelas simulações realizadas neste trabalho, os modelos gerados foram capazes de reproduzir com fidelidade padrões observados em cenários reais, validando a eficácia dessa abordagem. Dessa maneira, a simulação surge como uma alternativa viável, permitindo a geração de conjuntos de dados com características semelhantes aos reais, possibilitando a realização de pesquisas e testes metodológicos mesmo na ausência de dados originais.

A partir da análise realizada, pode-se afirmar que a regressão logística apresenta um futuro promissor na física médica, especialmente no diagnóstico por imagem, podendo contribuir para a melhoria dos fluxos e processos nos hospitais.

Dessa forma, a experiência proporcionou a aquisição de conhecimentos estatísticos e ampliou a visão sobre a aplicabilidade dos modelos de regressão logística na área da saúde, reforçando a importância da estatística para a pesquisa científica.

Bibliografia

[1] Gang Li et al. “The predictive models based on multimodality ultrasonography for the differential diagnosis of thyroid nodules smaller than 10 mm”. Em: *British Journal of Radiology* (2023). DOI: 10.1259/bjr.20221120. URL: https://doi.org/10.1259/bjr.20221120.

[2] G. A. Paula. *Modelos de Regressão com Apoio Computacional*. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2013. ISBN: 978-85-87023-39-2.

[3] M. Pagano e K. Gauvreau. *Princípios de Bioestatística*. 2ª ed. São Paulo: Pioneira Thomson Learning, 2004.

[4] Pedro A. Morettin e Wilton de O. Bussab. *Estatística básica*. 9ª ed. São Paulo: Saraiva, 2017. ISBN: 9788547220224.

[5] Douglas C. Montgomery, Elizabeth A. Peck e G. Geoffrey Vining. *Introduction to linear regression analysis*. 6ª ed. Hoboken, NJ: John Wiley & Sons, 2021. ISBN: 9781119578727.

[6] A. J. Dobson e A. G. Barnett. *An Introduction to Generalized Linear Models*. 4ª ed. Boca Raton, FL: CRC Press, 2018.

[7] D. J. Finney. *Statistical Methods in Bioassay*. 2nd. New York: Hafner, 1973.

[8] C. I. Bliss. “The Calculation of the Dose-Mortality Curve”. Em: *Annals of Applied Biology* 22 (1935), pp. 134–167.

[9] M. Mittlbock e H. Heinzl. “A note on R2 measures for Poisson and logistic regression models when both are applicable”. Em: *Journal of Clinical Epidemiology* 54 (2001), pp. 99–103.

[10] F. J. Aranda-Ordaz. “On two families of transformations to additivity for binary response data”. Em: *Biometrika* 68 (1981), pp. 357–363.

[11] P. McCullagh e J. A. Nelder. *Generalized Linear Models*. Second. London: Chapman e Hall/CRC, 1989. ISBN: 978-0412317606.

[12] D. F. Andrews e A. M. Herzberg. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. London: Springer, 1985.

[13] C. R. Padovani. *Bioestatística*. São Paulo: Cultura Acadêmica: Universidade Estadual Paulista, Pró-Reitoria de Graduação, 2012.

[14] O. A. Bakare, N. I. Okeke e G. O. Achumie. “Comprehensive review of logistic regression techniques in predicting health outcomes and trends”. Em: *World Journal of Advanced Pharmaceutical and Life Sciences* 7.2 (2024), pp. 16–26.

[15] E. Y. Boateng e D. A. Abaye. “A Review of the Logistic Regression Model with Emphasis on Medical Research”. Em: *Journal of Data Analysis and Information Processing* 7.4 (2019), pp. 190–207. DOI: 10.4236/jdaip.2019.74012.

[16] L. Wu et al. “Predictive value of magnetic resonance imaging parameters combined with tumor markers for rectal cancer recurrence risk after surgery”. Em: *World Journal of Gastrointestinal Surgery* 17.2 (2025), pp. 101–897. DOI: 10.4240/wjgs.v17.i2.101897.

Anexos

Anexo 1: Simulação de Dados com Regressão Múltipla

```
library(car)
library(GGally)
library(MASS)
library(leaps)
library(ggplot2)
library(dplyr)

set.seed(123)

dados <- data.frame(
x1 = rnorm(100, mean = 50, sd = 10),
x2 = rbinom(100, 1, 0.4),
x3 = runif(100, 2, 10))

dados$y <- 10 + 0.8*dados$x1 - 5*dados$x2 +
0.5*(dados$x2 * dados$x3) + rnorm(100, 0, 5)

modelo <- lm(y ~ x1 + x2 + I(x2*x3), data = dados)
summary(modelo)
confint(modelo)
anova(modelo)

x11(width=8, height=6)
par(mfrow = c(2,2))
plot(modelo)

shapiro.test(residuals(modelo))
ncvTest(modelo)
vif(modelo)

modelo_step <- stepAIC(modelo, direction = "both")
summary(modelo_step)

ggplot(dados, aes(x = x1, y = y, color = factor(x2))) +
geom_point(size=2) +
geom_smooth(method="lm", se=FALSE, formula=y~x1) +
```

```
labs(color = "Marcador␣biologico",
x = "Intensidade␣de␣radiacao",
y = "Dose␣absorvida␣no␣tecido") +
theme_minimal()
```

Anexo 2: Simulação de Dados com Modelo de Poisson

```
library(ggplot2)
library(dplyr)

set.seed(42)

n <- 150
idade <- rnorm(n, mean = 50, sd = 12)
marcador <- rbinom(n, 1, 0.4)

lambda <- exp( -1.2 + 0.03 * idade + 0.7 * marcador )
internacoes <- rpois(n, lambda)

dados <- data.frame(internacoes, idade, marcador)

modelo_pois <- glm(internacoes ~ idade + marcador,
family=poisson, data=dados)
summary(modelo_pois)
confint(modelo_pois)

disp <- sum(residuals(modelo_pois, type="pearson")^2) /
modelo_pois$df.residual
disp

pred_grid <- expand.grid(
idade = seq(min(idade), max(idade), length.out = 100),
marcador = c(0, 1))
pred_grid$pred <- predict(modelo_pois,
newdata=pred_grid, type='response')

ggplot(dados, aes(x=idade, y=internacoes,
color=factor(marcador)))
+ geom_point(alpha=0.6) +
geom_line(data=pred_grid, aes(x=idade, y=pred,
color=factor(marcador)),
size=1.2) + labs(x = "Idade␣(anos)",
y = "Numero␣de␣internacoes", color =
"Marcador␣biologico") + theme_minimal()
```

Anexo 3: Simulação de Dados com Regressão Logística

```r
library(ggplot2)
library(scales)

set.seed(42)

n <- 200

vascularidade <- sample(0:2, size = n, replace = TRUE,
prob = c(0.6, 0.3, 0.1))

beta0 <- -2.5
beta1 <- 1.5

prob_maligno <- exp(beta0 + beta1 * vascularidade) / (1 +
exp(beta0 + beta1 * vascularidade))

diagnostico <- rbinom(n, size = 1, prob = prob_maligno)

dados_ultrassom <- data.frame(vascularidade =
as.factor(vascularidade), diagnostico = diagnostico)

head(dados_ultrassom)

verossimilhanca_logistica <- function(theta, dados) {

        beta0 <- theta[1]
        beta1 <- theta[2]

        vascularidade_num <- as.numeric(
        as.character(dados$vascularidade))

        eta <- beta0 + beta1 * vascularidade_num
        prob <- exp(eta) / (1 + exp(eta))

        log_verossimilhanca <- sum(dados$diagnostico * log(prob) +
        (1 - dados$diagnostico) * log(1 - prob))

        return(-log_verossimilhanca)
}

parametros_iniciais <- c(0, 0)

ajuste_verossimilhanca <- optim(par = parametros_iniciais,
fn = verossimilhanca_logistica, dados = dados_ultrassom)

ajuste_verossimilhanca$par
```

```
modelo_logistico <- glm(diagnostico ~ vascularidade, data =
dados_ultrassom, family = binomial(link = "logit"))

summary(modelo_logistico)

coef(modelo_logistico)

fitted(modelo_logistico)

novos_dados_vascularidade <- data.frame(
vascularidade = factor(c(0, 1, 2)))

predicoes_log_odds <- predict(modelo_logistico, newdata =
novos_dados_vascularidade, se.fit = TRUE)

estimativa_log_odds <- predicoes_log_odds$fit
erro_padrao_log_odds <- predicoes_log_odds$se.fit

z_valor <- 1.96

ic_inferior_log_odds <- estimativa_log_odds -
z_valor * erro_padrao_log_odds
ic_superior_log_odds <- estimativa_log_odds +
z_valor * erro_padrao_log_odds

prob_prevista <- exp(estimativa_log_odds) /
(1 + exp(estimativa_log_odds))
prob_ic_inferior <- exp(ic_inferior_log_odds) /
(1 + exp(ic_inferior_log_odds))
prob_ic_superior <- exp(ic_superior_log_odds) /
(1 + exp(ic_superior_log_odds))

dados_grafico_predicao <- data.frame(
vascularidade_label = c("Baixa", "Media", "Alta"),
probabilidade = prob_prevista, ic_inferior = prob_ic_inferior,
ic_superior = prob_ic_superior)

dados_grafico_predicao$vascularidade_label <- factor(
dados_grafico_predicao$vascularidade_label,
levels = c("Baixa", "Media", "Alta"))

ggplot(dados_grafico_predicao, aes(x = vascularidade_label,
y = probabilidade)) + geom_point(size = 4, color = "red") +
geom_errorbar(aes(ymin = ic_inferior, ymax = ic_superior),
width = 0.2, color = "red", size = 1.2) +
labs(title = "Probabilidades␣Previstas␣de␣Malignidade",
subtitle = "Com␣Intervalos␣de␣Confianca␣de␣95%",
```

```
x = "Nivel de Vascularidade",
y = "Probabilidade (Maligno)") +
scale_y_continuous(labels = scales::percent_format(accuracy = 1),
limits = c(0, 1.0)) + theme_minimal() +
theme(
plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
plot.subtitle = element_text(hjust = 0.5, size = 12),
axis.title.x = element_text(size = 14),
axis.title.y = element_text(size = 14),
axis.text.x = element_text(size = 12),
axis.text.y = element_text(size = 12))

if (!require(ResourceSelection)) install.packages("ResourceSelection")
library(ResourceSelection)

residuos_pearson <- residuals(modelo_logistico, type = "pearson")
residuos_deviance <- residuals(modelo_logistico, type = "deviance")

hoslem_teste <- hoslem.test(modelo_logistico$y,
fitted(modelo_logistico), g=10)
print(hoslem_teste)

plot(fitted(modelo_logistico), residuos_pearson,
xlab = "Valores ajustados", ylab = "Residuos de Pearson",
main = "Residuos de Pearson vs Valores ajustados")
abline(h = 0, col = "red")

plot(fitted(modelo_logistico), residuos_deviance,
xlab = "Valores ajustados", ylab = "Residuos",
main = "Residuos vs Valores ajustados")
abline(h = 0, col = "red")
```