



INICIAÇÃO AO ESTUDO DE MODELAGEM ESTATÍSTICA

Mateus Vieira Saraiva (PIBIC/CNPq), Terezinha Aparecida Guedes (Orientadora), Ana Beatriz Tozzo Martins (Co-orientadora)
e-mail: taguedes@uem.br.
Universidade Estadual de Maringá / Centro de Ciências Exatas.

1.02.02.00-5 Estatística 1.02.02.06-4 Regressão e Correlação

Palavras-chave: Regressão linear; modelos mistos; aplicações em R.

Resumo:

A modelagem estatística vem sendo desenvolvida cada vez mais e empregada em análises de observações resultantes de diversas pesquisas científicas. Neste trabalho, iniciou-se o estudo da análise de regressão linear e de modelos mistos. A análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas (ou qualitativas) de tal forma que uma possa ser predita a partir da outra(s). O modelo misto é um modelo que incorpora parâmetros de efeitos fixos e efeitos aleatórios. A regressão linear simples foi empregada para ajustar medições (contagens em minutos) de radiação de fundo, em 39 locais com altitudes variando de 213,4 a 3237,7m. Utilizando o ambiente estatístico R para o ajuste, obteve-se que a média da contagem de radiações por minuto é de 8,553 e a cada metro de acréscimo na altitude, a contagem aumenta a uma taxa de 0,006. Observou-se que 76,46% das contagens estão sendo explicadas pela altitude.

Introdução

Francis Galton, no século 19, desenvolveu um estudo no qual foi analisado a altura dos filhos como dependente da altura dos pais, ficando conhecido por "regressão". Geralmente, define-se a relação entre duas variáveis da forma $Y=G(X)$, em que a variável X é denominada variável independente enquanto que Y é a variável dependente. Uma vez conhecida a expressão analítica da função G , obtêm-se um valor de Y dado um valor fixado de X . O objetivo da regressão é resumir dados observados da maneira mais simples e útil.





Materiais e métodos

Regressão Linear Simples

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas (ou qualitativas) de tal forma que uma variável pode ser predita a partir de outra(s).

Na análise de regressão os dados são representados através de um modelo linear aditivo, onde o modelo inclui um componente sistemático e pelo menos um aleatório. O modelo de regressão linear consiste de uma função linear $Y = f(X) + \varepsilon$, em que a função $f(\cdot)$ descreve a relação entre X e Y ; ε são os erros aleatórios; Y é variável resposta ou dependente e X é a variável independente, covariável ou variável preditora.

A Regressão Linear é uma equação para se estimar o valor esperado condicional de uma variável Y , dados os valores de algumas outras variáveis e o modelo pode ser escrito do seguinte modo: $E(Y | X = x_i) = \alpha + \beta x_i$.

Com esse modelo proposto pode-se inferir à cerca da característica em estudo na população com base na amostra observada. O modelo a ser utilizado será: $Y_i = E(Y | x_i) + \varepsilon_i = \alpha + \beta x_i + \varepsilon_i$ para $i = 1, \dots, n$, sendo n o número de observações. Na equação acima, α e β são denominados parâmetros do modelo e devem ser estimados e os erros podem ser calculados da seguinte forma: $\varepsilon_i = y_i - \alpha - \beta x_i$. A soma dos quadrados dos erros, para $i = 1, \dots, n$, é utilizada para a obtenção dos estimadores de mínimos quadrados dos parâmetros. Mas é necessário introduzir algumas suposições para as variáveis envolvidas. A primeira delas é que X deve ser uma variável supostamente controlada e não esteja sujeita à variações aleatórias. Segundo, para um dado valor de X , x_i , os erros ε devem ser distribuídos em torno da média $\alpha + \beta x_i$, de modo que $E(Y | x_i) = 0$. Terceiro, supõe-se que os erros possuam a mesma variabilidade em todos os níveis da variável X e sejam não correlacionados, ou seja, os dados devem ser homocedásticos, isto é, $V(\varepsilon | x) = \sigma_\varepsilon^2$ e $Cov(\varepsilon_i, \varepsilon_j) = 0$, para $i \neq j$. Além disso, os erros devem apresentar distribuição normal com média zero e variância σ^2 . Dado uma amostra de n indivíduos, têm-se n pares de valores (x_i, y_i) para a qual deverá ser estimado um modelo, ou seja, $y_i = \alpha + \beta x_i + \varepsilon_i$,





$i=1, \dots, n$. Assim, temos para cada $i=1, \dots, n$ uma equação e duas incógnitas (α, β). As derivadas da soma de quadrados dos erros ao serem igualadas a zero, fornecem os estimadores para as incógnitas ou parâmetros. Os parâmetros envolvidos neste modelo devem ser interpretados da seguinte maneira: α representa o ponto onde a reta corta o eixo Y e o β o coeficiente angular; ou ainda, o quanto varia a média de Y para o aumento de uma unidade da variável X.

Modelos Mistos

O termo modelos mistos é usado para descrever modelos que têm efeitos fixos e aleatórios. Estes modelos são empregados para modelar, por exemplo, dados coletados ao longo do tempo. Neste caso, os dados consistem, geralmente, de uma série relativamente curta, coletada ao fim de um período de tempo (esses dados são comumente referidos como dados longitudinais). Na notação matricial, os modelos lineares mistos podem ser representados como $Y = X\beta + Z\gamma + \varepsilon$, em que Y é um vetor de observações; β é um vetor de efeitos fixos; γ é um vetor de efeitos aleatórios; ε é um vetor de termos de erro aleatório; X é a matriz de delineamento para os efeitos fixos e Z é a matriz de delineamento para os efeitos aleatórios.

Resultados e Discussão

Uma aplicação de regressão linear simples consistiu no ajuste da medição de radiação de fundo, através de contagens Y num detector de Geiger por minutos, em 39 locais com altitudes X (em metros) variando de 213,4 a 3237,7. A variável resposta variou de 11 a 29 radiações por minutos com média de 17,48. Utilizando o ambiente estatístico R para o ajuste, obteve-se o seguinte modelo estimado: $\hat{y}_i = 8,553 + 0,006x_i$. Assim, a média da contagem de radiações por minuto é de 8,553 e a cada metro de acréscimo na altitude, a contagem aumenta a uma taxa de 0,006. A avaliação do ajuste foi verificada pela estatística R^2 que resultou igual a 0,7646, significando que 76,46% das contagens estão sendo explicadas pela altitude. Como se pode observar na Figura 1, os pontos tendem a uma reta, mas estão muito dispersos fazendo com que o modelo explique somente 76,46% da variabilidade dos dados.



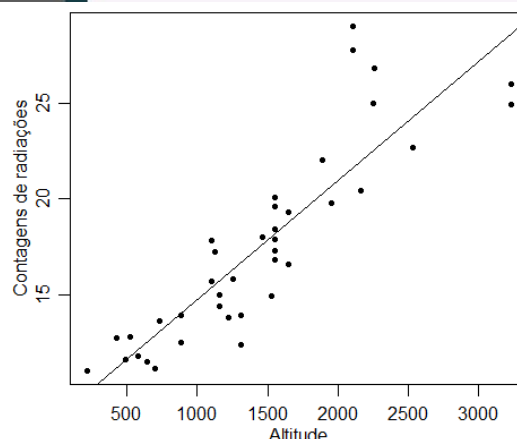


Figura 1 – Dispersão e ajuste das contagens de radiações segundo as altitudes, em metros.

Conclusões

A média da contagem de radiações por minuto foi de 8,553 e a cada metro de acréscimo na altitude, a contagem aumenta a uma taxa de 0,006. A avaliação do ajuste, verificada pela estatística R^2 , resultou igual a 0,7646, significando que 76,46% das contagens de medições de radiação estão sendo explicadas pela altitude. Pela análise gráfica, ficou visível a comparação da reta de regressão com a dispersão dos dados observados.

Agradecimentos

Ao Programa Institucional de Bolsas de Iniciação Científica - PIBIC/CNPq-FA-UEM pelo incentivo financeiro. Aos meus professores orientadores Terezinha Aparecida Guedes e Ana Beatriz Tozzo Martins pela paciência na orientação, incentivo, convívio, compreensão e amizade.

Referências

BUSSAB, W. O. **Métodos quantitativos: análise de variância e regressão**. 2. ed. São Paulo: Atual Editora, 2002.

BATES, D. Fitting linear mixed-effects models using lme4. **Journal of statistical software**. Madison: ResearchGate, Jun 2014. Disponível em: <<https://www.researchgate.net/publication/263352347>>. Acesso em: 15 maio 2016.

