Texto adicional do Experimento 3

Regressão linear e teste de compatibilidade por χ^2

Regressão linear e teste de compatibilidade por χ^2

Em teoria, diversas grandezas Físicas se relacionam através de expressões matemáticas bem definidas. Por exemplo, no movimento unidimensional uniformemente acelerado, a posição $s(t)$, a velocidade $v(t)$ e a aceleração $a(t)$ de um corpo se relacionam através das fórmulas $a(t) = a(\text{cte})$, $v(t) = v_0 + at$, $s(t) = s_0 + v_0 t + at^2/2$ e $v^2(t) = v_0^2 + 2a[s(t) - s_0]$, em cada instante de tempo t, onde s_0 é a posição inicial e v_0, a velocidade inicial do corpo. Uma das questões abordadas em Teoria dos Erros é a especificação experimental das relações das grandezas Físicas, bem como critérios para validar uma dada função específica entre as grandezas.

Considere duas grandezas físicas y e x que, em teoria, possuam uma dependência bem definida e especificada por uma fórmula matemática $y = f(x)$. Ou seja, para cada valor possível de x, a outra grandeza possui um valor determinado que vale $y(x)$. De maneira mais objetiva, consideramos que a grandeza x é dada com incerteza desprezível (ou que a incerteza dessa variável possa ser adicionada no valor da incerteza da outra variável) enquanto o valor de $y(x)$ é estimado experimentalmente com uma incerteza $\sigma(x)$. Os valores da incerteza da variável $y(x)$ podem ser diferentes entre si em função da variável x. Considere que um experimento permita calcular a seguinte tripla de valores $[x_k; y_k = y(x_k), \sigma_k = \sigma(x_k)]$, para n valores diferentes da variável x ($k = 1, 2, \cdots, n$). O índice k é utilizado para especificar uma realização específica do experimento. Vamos considerar que cada um dos k resultados experimentais é obtido com um valor diferente da variável x. Por exemplo, a variável x pode representar o tempo e o experimento consiste em uma única observação da grandeza monitorada em instantes de tempo sucessivos.

A relação matemática $y = f(x)$ que relaciona as grandezas y e x podem ser colocadas em funções de parâmetros constantes. Por exemplo, para um relação linear, a função mais geral entre y e x é dada em termos de duas constantes a_0 (coeficiente linear) e a_1 (coeficiente angular): $y(x) = a_1 x + a_0$. Para uma relação quadrática, a função mais geral entre y e x é dada em termos de três constantes a_0, a_1 e a_2, através da igualdade $y(x) = a_2 x^2 + a_1 x + a_0$. O ajuste de uma dada função a um conjunto de dados experimentais busca determinar os melhores valores para os parâmetros que especificam uma dada função, de acordo com algum critério de

qualidade. A regressão linear, portanto, busca determinar os valores das constantes a_0 e a_1, as quais especificam a função linear entre y e x.

Os n resultados experimentais, referentes a valores determinados da variável x, podem ser agrupados da forma

$$(x_1; y_1 \pm \sigma_1)\,,\ (x_2; y_2 \pm \sigma_2)\,,\ \cdots\,,(x_k; y_k \pm \sigma_k)\,,\ \cdots\,,(x_n; y_n \pm \sigma_n) \tag{1}$$

Um método de ajuste de função é o Método dos Mínimos Quadrados. Em função dos dados coletados, os valores dos parâmetros que especificam uma dada relação entre y e x são determinados através da distância entre os dados experimentais e a função teórica, em unidades da incerteza da medida,

$$D[\{x_k\},\{y_k\},\{\sigma_k\}] = \sum_{k=1}^{n} \frac{\left[y_k - f(x_k)\right]^2}{\sigma_k^2}. \tag{2}$$

A distância $D[\{x_k\},\{y_k\},\{\sigma_k\}]$ é função de todos os dados do experimento e também de todas as constantes que definem a função $f(x)$. A regressão linear se aplica para funções lineares e é definida pela determinação das duas constantes que definem a reta ($f(x) = a_1 x + a_0$). A distância $D[\{x_k\},\{y_k\},\{\sigma_k\}]$ é uma função quadrática dos parâmetros que definem a função $f(x)$.

Os parâmetros da função que se ajustam aos dados experimentais são os valores pra os quais minimiza o valor de $D[\{x_k\},\{y_k\},\{\sigma_k\}]$. A distância $D[\{x_k\},\{y_k\},\{\sigma_k\}]$ é uma quantidade sem unidades físicas. De forma geral, o Método dos Mínimos Quadrados, utiliza todos os dados experimentais e os pontos mais relevantes são os de menor incerteza. A divisão pela incerteza σ_k de cada resultado experimental permite que pontos mais incertos (ruidosos) estejam mais afastados do melhor ajuste (o que é de se esperar para pontos de grande incerteza) e os pontos mais precisos, mais próximos da curva ajustada.

O ajuste linear pelo Método de pode ser feito analiticamente em função dos dados experimentais. Para um conjunto de n resultados experimentais, escritos da forma $(x_k; y_k \pm \sigma_k)$ $(k = 1, 2, \cdots, n)$ os dois parâmetros que devem ser determinados são a_1 e a_0. Em função desses dois parâmetros, a distância $D[\{x_k\},\{y_k\},\{\sigma_k\}]$ se escreve

$$D[\{x_k\},\{y_k\},\{\sigma_k\}] = \sum_{k=1}^{n} \frac{\left[y_k - (a_1 x_k + a_0)\right]^2}{\sigma_k^2}, \tag{3}$$

a qual pode ser escrita em função de seis termos

$$\begin{aligned} D[\{x_k\},\{y_k\},\{\sigma_k\}] &= \left(\sum_{k=1}^{n}\frac{y_k^2}{\sigma_k^2}\right) - 2\left(\sum_{k=1}^{n}\frac{x_k y_k}{\sigma_k^2}\right)a_1 - 2\left(\sum_{k=1}^{n}\frac{y_k}{\sigma_k^2}\right)a_0 \\ &+ 2\left(\sum_{k=1}^{n}\frac{x_k}{\sigma_k^2}\right)a_1 a_0 + \left(\sum_{k=1}^{n}\frac{x_k^2}{\sigma_k^2}\right)a_1^2 + \left(\sum_{k=1}^{n}\frac{1}{\sigma_k^2}\right)a_0^2 . \end{aligned} \tag{4}$$

Os seis termos entre parênteses são determinados pelos dados experimentais e a função anterior permite determinar os valores de a_1 e a_0. Os valores de a_1 e a_0 que minimizam a expressão anterior representam o melhor ajuste, segundo o Método dos Mínimos Quadrados. A minimização de $D[\{x_k\},\{y_k\},\{\sigma_k\}]$ em função a_1 e a_0 pode ser feita de maneira usual (o procedimento é similar com o de encontrar o ponto que minimiza uma parábola de concavidade positiva) e esses valores são dados por

$$a_1 = \frac{\left(\sum_{k=1}^{n}\frac{1}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{x_k y_k}{\sigma_k^2}\right) - \left(\sum_{k=1}^{n}\frac{y_k}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{x_k}{\sigma_k^2}\right)}{\left(\sum_{k=1}^{n}\frac{1}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{x_k^2}{\sigma_k^2}\right) - \left(\sum_{k=1}^{n}\frac{x_k}{\sigma_k^2}\right)^2}, \tag{5}$$

$$a_0 = \frac{\left(\sum_{k=1}^{n}\frac{x_k^2}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{y_k}{\sigma_k^2}\right) - \left(\sum_{k=1}^{n}\frac{x_k y_k}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{x_k}{\sigma_k^2}\right)}{\left(\sum_{k=1}^{n}\frac{1}{\sigma_k^2}\right)\left(\sum_{k=1}^{n}\frac{x_k^2}{\sigma_k^2}\right) - \left(\sum_{k=1}^{n}\frac{x_k}{\sigma_k^2}\right)^2}. \tag{6}$$

Apesar de uma aparência complexa, os valores de a_1 e a_0 podem ser determinados em função daqueles termos entre os parênteses, os quais envolvem uma única soma com todos os valores encontrados no experimento. Para facilitar os cálculos, cada soma pode ser realizada de forma

independente. Para facilitar a notação, definimos

$$S_\sigma = \sum_{k=1}^{n} \frac{1}{\sigma_k^2}\,, \tag{7}$$

$$S_x = \sum_{k=1}^{n} \frac{x_k}{\sigma_k^2}\,, \tag{8}$$

$$S_y = \sum_{k=1}^{n} \frac{y_k}{\sigma_k^2}\,, \tag{9}$$

$$S_{xy} = \sum_{k=1}^{n} \frac{x_k y_k}{\sigma_k^2}\,, \tag{10}$$

$$S_{xx} = \sum_{k=1}^{n} \frac{x_k^2}{\sigma_k^2}\,, \tag{11}$$

$$\Delta = S_\sigma S_{xx} - S_x^2\,, \tag{12}$$

de tal forma que os parâmetros do melhor ajuste se escrevem

$$a_1 = \frac{S_\sigma S_{xy} - S_x S_y}{\Delta}\,, \tag{13}$$

$$a_0 = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}\,. \tag{14}$$

Portanto, o melhor ajuste é uma função dos dados experimentais $[\{x_k\}, \{y_k\}, \{\sigma_k\}]$ e essa dependência é linear com a variável ruidosa $y_k \pm \sigma_k$. De fato, as fórmulas de a_1 e a_0 mostram que os valores de y_k aparecem apenas no numerador da fração, e sempre de forma linear.

As incertezas dos parâmetros a_1 e a_0 podem ser encontradas via propagação de incerteza. Por exemplo, a contribuição da variável ruidosa y_k para as incertezas de cada um dos parâmetros a_1 e a_0 é dada por

$$\delta a_{1_{y_k}} = \frac{1}{\Delta}\left|\frac{S_\sigma x_k - S_x}{\sigma_k}\right|\,, \tag{15}$$

$$\delta a_{0_{y_k}} = \frac{1}{\Delta}\left|\frac{S_{xx} - x_k S_x}{\sigma_k}\right|\,. \tag{16}$$

Logo, as incertezas desses parâmetros são (após algumas simplificações)

$$\delta a_1 = \sqrt{(\delta a_{1y_1})^2 + (\delta a_{1y_2})^2 + \cdots + (\delta a_{1y_n})^2} = \sqrt{\frac{S_\sigma}{\Delta}}\,, \tag{17}$$

$$\delta a_0 = \sqrt{(\delta a_{0y_1})^2 + (\delta a_{0y_2})^2 + \cdots + (\delta a_{0y_n})^2} = \sqrt{\frac{S_{xx}}{\Delta}}\,. \tag{18}$$

$$\tag{19}$$

Os coeficientes da reta que melhor se ajustam aos dados experimentais podem ser encontrados pelas expressões

$$a_1 \;=\; \left(\frac{S_\sigma S_{xy} - S_x S_y}{\Delta}\right) \pm \sqrt{\frac{S_\sigma}{\Delta}}\,, \tag{20}$$

$$a_0 \;=\; \left(\frac{S_{xx} S_y - S_x S_{xy}}{\Delta}\right) \pm \sqrt{\frac{S_{xx}}{\Delta}}\,. \tag{21}$$

nas quais as incertezas são dadas com apenas um algarismo significativo.

As fórmulas anteriores sempre permitem encontrar os coeficientes dessa melhor reta (desde que $\Delta \neq 0$), independente de a grandeza y possuir uma dependência linear como função de x. Portanto, é necessário realizar uma análise de qualidade do ajuste. Nos casos em que exista um dependência linear entre y e x, a relação entre essas duas grandezas pode ser dada por $y(x) = a_{1\mathrm{ver}}x + a_{0\mathrm{ver}} + \delta y(x)$, onde $a_{1\mathrm{ver}}$, $a_{0\mathrm{ver}}$ e $\delta y(x)$ são os valores verdadeiros dos coeficientes angular e linear e do erro (devido a algum processo ruidoso). Para os modelos de ruídos simétricos em torno do valor nulo, o valor médio para cada valor de y resulta $\overline{y(x)} = a_{1\mathrm{ver}}x + a_{0\mathrm{ver}}$.

Em média, as estimativas para os coeficientes angular e linear da reta podem ser calculadas e comparadas com os valores verdadeiros dos coeficientes da reta desconhecida. Uma vez que as médias das somas que contém as variáveis y são dadas por $\overline{S_{xy}} = a_{1\mathrm{ver}}S_{xx} + a_{0\mathrm{ver}}S_x$ e

$$\overline{S_y} \;=\; \sum_{k=1}^{n} \frac{\bar{y}_k}{\sigma_k^2} = a_{1\mathrm{ver}} \sum_{k=1}^{n} \frac{x_k}{\sigma_k^2} + a_{0\mathrm{ver}} \sum_{k=1}^{n} \frac{1}{\sigma_k^2} = a_{1\mathrm{ver}}S_x + a_{0\mathrm{ver}}S_\sigma\,, \tag{22}$$

$$\overline{S_{xy}} \;=\; \sum_{k=1}^{n} \frac{\bar{y}_k x_k}{\sigma_k^2} = a_{1\mathrm{ver}} \sum_{k=1}^{n} \frac{x_k^2}{\sigma_k^2} + a_{0\mathrm{ver}} \sum_{k=1}^{n} \frac{x_k}{\sigma_k^2} = a_{1\mathrm{ver}}S_{xx} + a_{0\mathrm{ver}}S_x\,, \tag{23}$$

os valores médios dos coeficientes encontrados pelo Método dos Mínimos Quadrados valem

$$\overline{a_1} = \frac{S_\sigma \left(a_{1\text{ver}} S_{xx} + a_{0\text{ver}} S_x\right) - S_x \left(a_{1\text{ver}} S_x + a_{0\text{ver}} S_\sigma\right)}{\Delta} = a_{1\text{ver}}\,, \tag{24}$$

$$\overline{a_0} = \frac{S_{xx} \left(a_{1\text{ver}} S_x + a_{0\text{ver}} S_\sigma\right) - S_x \left(a_{1\text{ver}} S_{xx} + a_{0\text{ver}} S_x\right)}{\Delta} = a_{0\text{ver}}\,. \tag{25}$$

Ou seja, as estimativas fornecem os valores corretos em média. Quando a relação entre as duas grandezas y e x for linear, com um ruído nulo em média, o Método dos Mínimos Quadrados fornece estimativas consistentes com os valores a serem determinados.

Porém, mesmo para relações não lineares entre y e x, é possível determinar a melhor reta que se ajusta aos dados. É possível fazer uma análise da qualidade do ajuste em função do valor do menor valor da distância $D[\{x_k\}, \{y_k\}, \{\sigma_k\}]$, em função dos parâmetros livres que descrevem uma dada relação entre y e x. Definimos a função chi-quadrado (χ^2) da amostra de dados pelo valor mínimo de $D[\{x_k\}, \{y_k\}, \{\sigma_k\}]$:

$$\chi^2 = \text{Min} \sum_{k=1}^{n} \frac{\left[y_k - f(x_k)\right]^2}{\sigma_k^2}. \tag{26}$$

Ou seja, definimos χ^2 em termos dos valores encontrados via o Método dos Mínimos Quadrados. Com as expressões anteriores, é possível simplificar o valor do χ^2 e escrevê-lo apenas em função dos dados do experimento. Por simplicidade, essa expressão será omitida para o caso linear. O χ^2 da amostra de dados é também uma grandeza aleatória, uma vez que depende dos valores de y. O valor médio dessa grandeza dá o valor (sem unidades físicas) $\overline{\chi^2} = n - p$, na qual n é o número de pontos obtidos no experimento e p é a quantidade de parâmetros livres usado para o ajuste da função. No caso linear, $p = 2$ pois existem apenas dois parâmetros livres. O valor médio do chi-quadrado para a regressão linear é dado, portanto, por $\overline{\chi^2} = n - 2$. Esse é o valor esperado para um experimento com n medidas.

Por causa das flutuações e incertezas estatísticas, em uma dada realização experimental, o valor encontrado do χ^2 é diferente do valor médio. Através da distribuição de probabilidade para o χ^2, pode-se determinar um intervalo de 98% de confiança com os possíveis valores para o χ^2. Esse intervalo não é simétrico em relação ao valor médio e pode ser determinado com o auxílio de algumas tabelas, em função do número de dados e da quantidade de parâmetros livres do modelo. Por exemplo, com $n = 10$ dados experimentais, se a relação entre y e x for linear ($n - p = 10 - 2 = 8$), o valor calculado do χ^2 deve ficar compreendido entre $0,21 < (\chi^2/8) < 2,51$ com uma margem de confiança de 98%. Outros dois exemplos, se a quantidade de dados experimentais coletados for de $n = 50$ ($n - p = 50 - 2 = 48$), o intervalo

de confiança fica $0,59 < (\chi^2/48) < 1,54$ enquanto para $n = 100$ ($n - p = 100 - 2 = 98$), o intervalo é dado por $0,70 < (\chi^2/98) < 1,36$, ambos os exemplos possuem uma margem de confiança de 98%.

O intervalo de confiança possui um valor mínimo e outro valor máximo. Se uma função possui muitos parâmetros livres, esses parâmetros ajustados podem diminuir o valor do χ^2 quando os dados forem descritos por uma função menos complexa (com uma quantidade menor de parâmetros livres ajustáveis). Por exemplo, com n dados experimentais, existe um polinômio de grau $n - 1$ capaz de descrever completamente os valores observados sem nenhuma incerteza (todos os dados se encaixam perfeitamente na curva do polinômio). Em uma situação com incertezas e erros instrumentais, o valor do χ^2 nulo ou muito pequeno não é compatível com os dados do experimento. Por outro lado, se a função utilizada não possui a complexidade necessária para descrever a relação existente entre y e x, mesmo com o ajuste dos parâmetros livres da função, existirá um erro apreciável. Esses erros tornam o valor do χ^2 grande, uma vez que diversos pontos experimentais estarão distantes da curva, em várias ordens de grandeza em relação à incerteza de cada medida.