Introdução à Análise de Dados em FAE
(05/10/2024)

Exercícios de Estatística para Análise de Dados em HEP

Professores:
Eliza Melo, Dilson Damião e Mauricio Thiel

Nome:
Matheus da Costa Geraldes

ATIVIDADE 1

Ao realizar um ajuste linear na forma $y(x) = ax + b$ para um conjunto de N pontos (x_i, y_i), onde cada valor de y possui incertezas diferentes, empregamos uma função de soma dos quadrados ponderada $S(a, b)$ que deve ser minimizada. Esta função é expressa como:

$$S(a,b) = \sum_{i=1}^{N} \frac{(y_i - (ax_i + b))^2}{\sigma_i^2} \tag{1}$$

Ao expandir a expressão quadrática, podemos reescrever S:

$$S(a,b) = \sum_{i=1}^{N} \left(\frac{y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2}{\sigma_i^2} \right) \tag{2}$$

Essa expansão permite que separemos os componentes da soma:

$$S(a,b) = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - 2a \sum_{i=1}^{N} \frac{y_i x_i}{\sigma_i^2} - 2b \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} + a^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} + 2ab \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + b^2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{3}$$

Para determinar os parâmetros a e b, calculamos as derivadas parciais de S em relação a esses parâmetros e as igualamos a zero:

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{N} \frac{x_i(y_i - ax_i - b)}{\sigma_i^2} = 0 \tag{4}$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{N} \frac{(y_i - ax_i - b)}{\sigma_i^2} = 0 \tag{5}$$

A primeira equação pode ser reescrita como:

$$\sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} - a \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - b \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = 0 \tag{6}$$

E a segunda como:

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - a \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} - b \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = 0 \tag{7}$$

Definimos agora as somas ponderadas para simplificar a notação:

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \quad \bar{y} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \quad \bar{x^2} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}, \quad \bar{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}, \quad \bar{1} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

Com essas definições, substituímos nas equações anteriores, resultando no seguinte sistema:

$$\begin{cases} \bar{x^2}a + \bar{x}b = \bar{xy} \\ \bar{x}a + \bar{1}b = \bar{y} \end{cases} \tag{8}$$

Agora, isolamos uma das variáveis, digamos b, na primeira equação:

$$b = \frac{\bar{xy} - \bar{x^2}a}{\bar{x}} \tag{9}$$

Substituímos b na segunda equação:

$$\bar{x}a + \bar{1}\left(\frac{\bar{xy} - \bar{x^2}a}{\bar{x}}\right) = \bar{y} \tag{10}$$

Multiplicando ambos os lados por $\bar{x}$ para eliminar o denominador, obtemos:

$$\bar{x}^2 a + \bar{1}(\bar{xy} - \bar{x^2}a) = \bar{y}\bar{x} \tag{11}$$

Distribuindo os termos:

$$\bar{x}^2 a + \bar{1}\bar{xy} - \bar{1}\bar{x^2}a = \bar{y}\bar{x} \tag{12}$$

Reorganizando, encontramos uma expressão para a:

$$a(\bar{x}^2 - \bar{1}\bar{x^2}) = \bar{y}\bar{x} - \bar{1}\bar{xy} \tag{13}$$

Assim, podemos isolar a:

$$a = \frac{\bar{y}\bar{x} - \bar{1}\bar{xy}}{\bar{x}^2 - \bar{1}\bar{x^2}} \tag{14}$$

Finalmente, utilizando o valor de a obtido, substituímos de volta na expressão para b:

$$b = \bar{y} - a\bar{x} \tag{15}$$

Dessa forma, temos os valores finais de a e b:

$$\begin{cases} a = \frac{\bar{xy}\bar{1} - \bar{x}\bar{y}}{\bar{x^2}\bar{1} - \bar{x}^2} \\ b = \bar{y} - a\bar{x} \end{cases} \tag{16}$$

Assim, encontramos os parâmetros ideais para o ajuste linear, levando em consideração as incertezas específicas de cada ponto.

ATIVIDADE 2

A seção de choque, σ, pode ser calculada pela seguinte fórmula:

$$\sigma = \frac{N_{Total} - N_{Background}}{\mathcal{L}} \tag{17}$$

Os dados disponíveis são:

1. $N_{Total} = 2567$ (total de eventos observados);
2. $N_{Background} = 1223.5$ (número de eventos de fundo);
3. $\mathcal{L} = 25\, fb^{-1}$ (luminosidade integrada).

Primeiro, calculamos a diferença entre os eventos totais e os eventos de fundo:

$$N_{Signal} = N_{Total} - N_{Background} = 2567 - 1223.5 = 1343.5$$

Agora, substituímos esse valor na fórmula da seção de choque:

$$\sigma = \frac{N_{Signal}}{\mathcal{L}} = \frac{1343.5}{25}$$

Realizando a divisão:

$$\sigma = 53.74\, fb$$

—

Para calcular a incerteza estatística, utilizamos a distribuição de Poisson. A fórmula para a variância da incerteza estatística é:

$$\sigma_{stat}^2 = \left(\frac{\sqrt{N_{Total}}}{\mathcal{L}}\right)^2 + \left(\frac{\sqrt{N_{Background}}}{\mathcal{L}}\right)^2 \tag{18}$$

Calculamos cada termo separadamente:
1. **Para o primeiro termo**:

$$\frac{\sqrt{N_{Total}}}{\mathcal{L}} = \frac{\sqrt{2567}}{25} \approx \frac{50.67}{25} \approx 2.0268$$

2. **Para o segundo termo**:

$$\frac{\sqrt{N_{Background}}}{\mathcal{L}} = \frac{\sqrt{1223.5}}{25} \approx \frac{34.96}{25} \approx 1.3984$$

Agora, elevamos ao quadrado e somamos:

$$\sigma_{stat}^2 = (2.0268)^2 + (1.3984)^2 \approx 4.111 + 1.955 \approx 6.066$$

Em seguida, calculamos a incerteza estatística:

$$\sigma_{stat} = \sqrt{6.066} \approx 2.46\, fb$$

—

A incerteza sistemática é dada como 10% da seção de choque. Portanto:

$$\sigma_{sist} = 0.10 \times \sigma = 0.10 \times 53.74 = 5.374\, fb$$

Assim, a seção de choque total e suas incertezas:

$$\sigma = 53.74 \pm 2.46_{stat} \pm 5.37_{sist}$$

—

ATIVIDADE 3

Para determinar o número esperado de eventos após os cortes, utilizamos a distribuição de Poisson e buscamos um nível de confiança de 95% (C.L.). A fórmula é:

$$P(n; x) = \frac{e^{-x} x^n}{n!} \tag{19}$$

Para o caso em que $n = 0$, a equação simplifica-se para:

$$P(0; x) = e^{-x}$$

Queremos que essa probabilidade atenda ao critério:

$$P(0; x) \geq 0.05 \quad \Rightarrow \quad e^{-x} \geq 0.05$$

Tomando o logaritmo natural de ambos os lados, obtemos:

$$-x \geq \ln(0.05) \quad \Rightarrow \quad x \leq -\ln(0.05)$$

Calculamos o logaritmo:

$$-\ln(0.05) \approx 2.9957$$

Portanto, para um nível de confiança de 95%, concluímos que:

$$x \geq 2.9957$$

Como estamos interessados em um número inteiro de eventos, podemos afirmar que o número esperado de eventos é 3. Assim, temos:

$$N_{esperados} = 3$$

ATIVIDADE 4

A função χ^2 é uma ferramenta essencial para a análise de ajuste de dados e é definida pela seguinte expressão:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2} \tag{20}$$

Aqui, y_i representa os dados observados, $f(x_i)$ é o valor previsto pelo modelo para cada ponto x_i, e σ_i denota a incerteza associada a cada y_i.

O número de graus de liberdade, frequentemente denotado como ndf, é calculado pela fórmula:

$$ndf = N - p \tag{21}$$

onde N é o total de observações e p é o número de parâmetros que estão sendo ajustados. Para uma análise mais profunda, podemos considerar a média e a variância da distribuição χ^2:

$$Med[\chi^2] = ndf, \tag{22}$$
$$Var(\chi^2) = 2\,ndf. \tag{23}$$

Um ajuste ideal é caracterizado pela condição:

$$Mdf\left[\frac{\chi^2}{ndf}\right] = \frac{Med[\chi^2]}{ndf} = \frac{ndf}{ndf} = 1. \tag{24}$$

Esse resultado indica que, para um ajuste perfeito, devemos esperar que $\chi^2/ndf \to 1$.

Se $\chi^2/ndf > 1$: Isso significa que o modelo não está se ajustando bem aos dados. Pode ser que o modelo seja muito simples ou que não tenhamos calculado as incertezas corretamente. - Se $\chi^2/ndf < 1$: Nesse caso, o modelo pode estar superajustado, ou seja, é mais complexo do que os dados precisam, ou ainda que as incertezas foram superestimadas.

O valor de χ^2/ndf é muito importante para entender se o modelo que escolhemos é bom. Um valor perto de 1 significa que o modelo está se ajustando bem, enquanto valores muito diferentes disso indicam que precisamos repensar o que estamos fazendo.

Com isso, fica evidente que o uso da estatística χ^2 é fundamental para garantir que o modelo escolhido se adeque adequadamente aos dados observados, permitindo assim uma análise mais robusta e confiável.