Inferência para CS
Tópico 12 - Suficiência e Famílias Exponenciais

Renato Martins Assunção

DCC - UFMG

2025

Acréscimos Futuros

- Explicar estimadores de Rao-Blackwell
- Seja $L(d(X), \theta)$ uma função de perda convexa qualquer e $R(d(X), \theta) = \mathbb{E}(L(d(X), \theta))$ o risco de estimar θ usando $d(X)$.
- Seja $T(X)$ estatística suficiente para θ e um estimador $\delta(X)$ qualquer.
- Então $\delta(X)$ é dominado por $g(T(X)) = \mathbb{E}(L(\delta(X)|T(X))$ no sentido de que $R(\delta(X), \theta) \geq R(g(T(X)), \theta)$ para todo θ with strict inequality unless $g(T(X)) = \delta(X)$ with probability one.
- Assim, é sempre melhor condicionar na estaística suficiente para estimar θ.
- The need for a non-trivial sufficient statistic obviously restricts the applicability of this theorem in mathematical statistics to exponential families.
- Note que δ não precisa ser não-viciado. Se for, então $g(T(X))$ também será.

Resumo

- Aprendemos que o MLE é um estimador assintoticamente eficiente.
- Se n não é pequeno e se escolhemos um modelo correto, o MLE é aproximadamente gaussiano, centrado no verdadeiro valore de $\boldsymbol{\theta}$ e com a "menor matriz" de variância possível para um estimador não-viciado (a matriz $I^{-1}(\boldsymbol{\theta})$)
- Vamos ver agora que o MLE é um estimador que extrai TODA a informação sobre $\boldsymbol{\theta}$ existente nos dados.
- Este é o conceito de suficiência.

Necessário e suficiente

- Existem várias situações em que usamos as expressões *necessário* ou *suficiente*.
- Em lógica de proposições, por exemplo, existe um sentido bem preciso para seu uso.
- Numa situação de análise de dados também temos uma definição precisa para estes termos e é esta situação que nos interessa.
- Vamos imaginar que temos interesse em conhecer o valor médio da pressão sistólica entre homens saudáveis de 40 anos de idade.

Pressão esperada θ_{40}

- Suponha que a pressão sistólica Y de um indivíduo saudável e de 40 anos escolhido ao acaso da população tenha distribuição $Y \sim N(\theta_{40}, 4)$.
- É claro que uma amostra i.i.d. de dados $Y_1, Y_2, \ldots, Y_n$ é a fonte natural de informação sobre θ_{40} que vamos buscar obter.
- Provavelmente iremos usar a média aritmética $\overline{Y}$ como estimador para θ_{40}.
- Estes dados são altamente informativos sobre θ_{40} porque a sua distribuição $N(\theta_{40}, 4)$ está diretamente controlada por θ_{40}.

Outros dados relevantes

- Na impossibilidade de obter uma amostra deste tipo, podemos talvez obter uma amostra de outras faixas etárias
- Por exemplo, suponha que seja conhecido que a pressão esperada aumenta com a idade mas não se sabe exatamente a que taxa.
- Assim, sabemos apenas que $\theta_{50} > \theta_{40}$.
- Se tivermos uma amostra i.i.d. de dados $Y_1, Y_2, \ldots, Y_n$ de homens saudáveis de 50 anos de idade escolhidos ao acaso, a média aritmética de seus valores trará uma informação viciada para estimar θ_{40} (superestimando o valor de θ_{40}).

Outros dados relevantes

- Podemos até querer corrigir este vício de alguma forma, talvez com informação suplementar.
- De qualquer modo, mesmo com a informação viciada, aprendemos algo sobre θ_{40}.
- A razão é que sabemos que $Y_1, Y_2, \ldots, Y_n$ são i.i.d. $N(\theta_{50}, 4)$ e também que $\theta_{50} > \theta_{40}$.
- Isto é, a distribuição dos dados está ligada de alguma forma a θ_{40}.
- Assim, podemos usar estes dados para extrair algum conhecimento sobre θ_{40}.
- Por exemplo, vamos induzir que θ_{40} deve ser um valor menor que $\overline{Y}$.

Dados irrelevantes

- Considere o mesmo problema de inferir o valor esperado θ_{40} da pressão sistólica entre homens saudáveis e de 40 anos.
- Suponha que agora sejam oferecidos milhares de dados selecionados ao acaso e independemente pelo computador e com distribuição uniforme no intervalo $(0, 1)$:
- `0.522, 0.287, 0.088, 0.413, 0.427, ...`
- É intuitivo que estes dados são de pouca valia para inferir o valor de θ_{40}.
- O que dados selecionados ao acaso do intervalo $(0, 1)$ podem informar sobre θ_{40}?
- Se a sua distribuição $U(0, 1)$ não tem nenhuma relação com θ_{40} fica difícil imaginar o que poderia ser feito com os dados dessa amostra, mesmo que os dados sejam em grande quantidade.

Dados irrelevantes ou relevantes?

- Obviamente, os milhares de dados SERIAM DE MUITA valia caso tivessem sido selecionados ao acaso do intervalo $(\theta_{40}, \theta_{40} + 1)$ ou talvez do intervalo $(\theta_{40} - 1, \theta_{40} + 1)$.
- Neste caso, a distribuição de Y depende de θ_{40}.
- Assim, os valores de uma amostra $Y_1, Y_2, \ldots, Y_n$ podem ser informativos sobre o parâmetro θ_{40}.
- Afinal de contas, agora θ_{40} governa ou influencia a distribuição dos dados.

Altura e espera

- CONCLUSÃO intuitiva: se uma amostra de dados aleatórios $Y_1, Y_2, \ldots, Y_n$ tiver uma distribuição de probabilidade que não envolve um parâmetro de interesse θ, não poderemos extrair informação sobre θ desses dados.
- Um outro exemplo: escolher um aluno ao acaso da UFMG e medir sua altura H.
- Interesse na altura esperada $\mathbb{E}(H) = \theta$.
- Não conseguiremos inferir sobre θ olhando dados $Y_1, Y_2, \ldots, Y_n$ do tempo de espera na fila do caixa do Banco do Brasil na Praça de Serviços.
- O tempo de espera na fila é uma variável aleatória Y.
- Sua distribuição de probabilidade não envolve de nenhuma maneira o parâmetro θ, que é a altura esperada de um aluno escolhidos ao acaso na UFMG.
- Como os dados de espera na fila poderiam ajudar a inferir sobre θ?
- Não podem. Isto parece óbvio, certo?

Concluindo...

- O conceito de suficiência é mais sutil que os exemplos anteriores mas tem um princípio similar.
- Resumimos os dados em alguns poucos números, as estatísticas suficientes.
- As estatísticas suficintes variam de problema para problemas mas são resumos dos dados como a média aritmética dos dados, por exemplo.
- Com estes resumos em mãos (as estatísticas suficientes), podemos dispensar os dados propriamente ditos, mesmo que milhares deles.
- Esses dados já não possuem mais nenhum valor adicional para estimar o parâmetro θ.
- Eles se tornam tão inúteis para estimar θ quanto os dados de espera na fila do banco são inúteis para estimar a altura dos alunos da UFMG.
- E a razão será a mesma: a sua distribuição já não terá nenhuma relação com θ depois de conhecermos os resumos.

Suficiência

- Suponha que $X_1, \ldots, X_n$ são i.i.d. com distribuição $N(\theta, 1)$.
- É óbvio que $\hat{\theta} = X_3$ é um estimador muito pobre de θ.
- Ele deixa muita informação não aproveitada sobre θ no resto do vetor

$$(X_1, X_2, X_4, \ldots X_n) .$$

- Que tal a média dos extremos (do mínimo e do máximo)

$$\hat{\theta}^* = \frac{X_{(1)} + X_{(n)}}{2}$$

- Existe alguma informação sobre θ não aproveitada no restante dos dados?
- Isto é, os valores intermediários $X_{(2)}, \ldots, X_{(n-1)}$ possuem alguma informação ADICIONAL sobre θ?
- Se eles possuirem informação adicional àquela contida em $\hat{\theta}^*$, ao descartá-los estaremos desperdiçando dados aproveitáveis.
- **Aparentemente** eles possuem informação adicional pois o estimador

$$\begin{aligned} \overline{X} &= \frac{\sum_i X_i}{n} = \frac{1}{n}(X_{(1)} + X_{(n)} + X_{(2)} + \ldots + X_{(n-1)}) \\ &= \frac{2\hat{\theta}^*}{n} + \frac{X_{(2)} + \cdots + X_{(n-1)}}{n} \end{aligned}$$

é o mais eficiente do que $\hat{\theta}^*$ para estimar θ pois $\overline{X}$ é não-viciado para estimar μ e atinge a cota de Cramér-Rao (o menor MSE possível dentre não-viciados).

- Quando podemos saber que **toda** informação sobre θ foi extraÃda dos dados $\mathbf{Y}$ e está concentrada num estimador $T(\mathbf{Y})$?
- Foi Fisher, em 1922, quem respondeu a esta pergunta.
- Ele introduziu a idéia de que certas estatÃsticas são **suficientes** para fazer inferência sobre θ.
- Isto é, com apenas alguns resumos estatísticos dos dados extraímos TODA A INFORMAÇÃO existente sobre θ nos dados.
- O que sobrar não-aproveitado nos dados, o que não está nesses resumos suficientes, é simplesmente lixo para estimar θ. Não serve para nada se o que queremos é estimar θ.
- Esse conceito é fundamental na teoria estatÃstica.

Exemplo:

- Sejam $X_1, \ldots, X_n$ i.i.d. *Bernoulli*(θ).
- Considere $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$, o número de sucessos nos n ensaios.
- Desta estatística $T(\mathbf{X})$ podemos obter o estimador usual de θ, a proporção amostral.

$$\hat{\theta} = \overline{X} = \frac{T(\mathbf{X})}{n}$$

- Dado o n^0 de sucessos $T(\mathbf{X})$,
  - o que resta de informação nos dados é a ordem dos 0's e 1's.
- *Se o modelo é verdadeiro (θ constante e independência)*, então
  - esta ordem parece irrelevante para conhecer melhor θ.
- Vamos formalizar esta intuição.

- Temos sequência X_i de sucessos e fracassos com probabilidade de sucesso θ.
- Nesta sequência, o que pode nos informar sobre θ.
- Com certeza, $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ é altamente informativo sobre θ.
- Se $T(\mathbf{X}) \approx n$ podemos induzir que $\theta \approx 1$.
- Se $T(\mathbf{X}) \approx 0$ podemos induzir que $\theta \approx 0$.
- Se $T(\mathbf{X}) \approx n/2$ podemos induzir que $\theta \approx 1/2$ e etc...
- OK, $T(\mathbf{X})$ é informativo sobre θ.

- Considere agora os dados CONDICIONADO NESTE CONHECIMENTO sobre $T(\mathbf{X})$
- Por exemplo, suponha que $T(\mathbf{X}) = 5$ e que faremos uso disso para inferir sobre θ.
- O que MAIS podemos inferir sobre θ a partir da sequência dos dados?
- Que outro aspecto da sequência (além do fato de que $T(\mathbf{X}) = 5$) podemos usar para, de alguma forma, melhorar nossa inferência sobre θ.
- Resposta: NADA. Depois de extrair a informação de que $T(\mathbf{X}) = 5$, não há mais nada nos dados que possa ser usado para melhorar nossa inferência sobre θ.
- Isto NÃO É UMA HEURÍSTICA, é um fato matemático!!

- Suponha que alguém informe apenas que $T(\mathbf{X}) = 5$ sem dizer qual é o vetor $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ realizado.
- As 2^n sequências possÃveis são todas aquelas com 5 1's e $n - 5$ 0's.
- SEM SABER QUE $T(\mathbf{X}) = 5$, cada uma das 2^n sequências possuem probabilidades diferentes.

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$$

- Veja que estas probabilidades DEPENDEM de θ.

- Temos
$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$$
- É por isto que saber **x** informa sobre θ.
- O raciocínio indutivo INVERTE o sentido do cálculo. Dependendo de θ, alguns **x** são os mais prováveis. Qual **x** aconteceu? Este? Então isto dá uma idéia de qual deve ser o θ que está por trás do mecanismo gerador de dados.
- SABENDO que $T(\mathbf{X}) = 5$, as 2^n sequencias passam a ter probabilidades diferentes da fórmula $\mathbb{P}(\mathbf{X} = \mathbf{x})$.
- Por exemplo, todas as sequencias com mais de 5 sucessos (ou com MENOS de 5 sucessos) tem agora chance ZERO.

- Vamos então considerar apenas as sequencias com 5 sucessos em n experimentos.
- Existem $n!/((n-5)!5!$ dessas sequencias.
- Condicionada em $T(\mathbf{X}) = 5$, cada uma dessas sequencias possui probabilidade, igual a

$$\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n) \,|\, T(\mathbf{X}) = 5\right) = \frac{\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n) \;\text{ e }\; T(\mathbf{X}) = 5\right)}{\mathbb{P}_\theta(T(\mathbf{X}) = 5)}$$

- Se o evento A está contido no evento B (isto é, $A \subset B$), nós temos $A \cap B = A$ e portanto $\mathbb{P}(A \cap B) = \mathbb{P}(A)$.
- Se a sequência $(x_1, \ldots, x_n)$ é uma daquelas com exatamente 5 sucessos, então o evento $[\mathbf{X} = (x_1, \ldots, x_n)]$ está contido no evento $[T(\mathbf{X}) = 5]$.
- Portanto, $\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n) \;\text{ e }\; T(\mathbf{X}) = 5\right) = \mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n)\right)$

- Portanto, temos

$$\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n) \,|\, T(\mathbf{X}) = 5\right) = \frac{\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n)\right)}{\mathbb{P}_\theta(T(\mathbf{X}) = 5)}$$

- Para o numerador:
  $\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n)\right) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^5(1-\theta)^{n-5}$
- Para o denominador: o número total de sucessos $T(\mathbf{X})$ tem distribuição binomial $Bin(n, \theta)$. Portanto,
  $\mathbb{P}_\theta(T(\mathbf{X}) = 5) = \binom{n}{5}\theta^5(1-\theta)^{n-5}$
- Em resumo, $\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_n) \,|\, T(\mathbf{X}) = 5\right)$ resulta em

$$\frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{5}\theta^5(1-\theta)^{n-5}} = \frac{\theta^5(1-\theta)^{n-5}}{\binom{n}{5}\theta^5(1-\theta)^{n-5}} = \frac{1}{\binom{n}{5}}\,.$$

- Isto é, dado o n^0 5 de sucessos em n ensaios
  - a probabilidade de qualquer sequência $\boldsymbol{x}$ com 5 1's e $n-5$ 0's é constante em θ;
  - As sequencias com cinco 1's são equiprováveis e sua probabilidade não depende de θ.
  - Qualquer uma das sequencias tem probabilidade $\frac{1}{\binom{n}{5}}$, não importa o valor de θ.
- Isso vale para qualquer θ, seja próximo de 0 ou de 1.

- Vamos considerar agora o caso genérico.
- Suponha que $T(\mathbf{X}) = t$,
  - onde t é inteiro $0 \leq t \leq n$.
- Então, por cálculos idênticos ao anterior, temos que se

$$(x_1, \ldots, x_n)$$

é uma sequência com t 1's então

$$P_\theta(\mathbf{X} = x_1, \ldots, x_n \,|\, T(\mathbf{X}) = t) = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

que não depende de θ.

- Por exemplo, suponha que $n = 200$ e que $\theta = 0.01$, um valor próximo de zero.
- Dentre as 2^{200} sequâncias possíveis, as mais prováveis serão aquelas com 10 sucessos no máximo pois, usando `pbinom(10, 200, 0.01)`, encontramos:

$$\mathbb{P}_{\theta=0.01}(T(\mathbf{X}) \leq 10) = 0.9999931$$

- SE SABEMOS que $T(\mathbf{X}) = 5$ ocorreu, apenas 5 sucessos em 200, então não existem mais 2^{200} sequências possíveis.
- Temos agora "apenas" $\binom{200}{5} = 2535650040$ sequências possíveis (basta escolher 5 das 200 posições para colocar os cinco 1's)
- Cada uma dessas 2535650040 tem probabilidade $1/2535650040$.
- Sem surpresas, aqui, certo?

- Mas suponha agora que $n = 200$ mas $\theta = 0.97$, um valor próximo de 1.
- Dentre as 2^{200} sequâncias possíveis, as mais prováveis serão aquelas com 180 sucessos no mínimo pois, usando `1- pbinom(179, 200, 0.97)` encontramos:

$$\mathbb{P}_{\theta=0.99}(T(\mathbf{X}) \geq 90) = 0.9999992$$

- SABEMOS que $T(\mathbf{X}) = 5$ ocorreu, apenas 5 sucessos em 200.
- Este é um evento de baixíssima probabilidade pois $\theta = 0.97$ e esperamos muitos 1's.
- Mas digamos foi isto que ocorreu. Afinal, probabilidade baixa não é o mesmo que probabilidade zero.

- Como antes, não existem mais 2^{200} sequências possíveis.
- Como antes, temos agora "apenas" $\binom{200}{5} = 2535650040$ sequências possíveis.
- Com $\theta = 0.97$, qualé a probabilidade de cada uma dessas 2535650040 sequências possíveis?
- Cada uma dessas 2535650040 sequencias tem probabilidade $1/2535650040$, a mesma probabilidade que obtivemos com $\theta = 0.01$.
- Razão: como mostramos antes, $\mathbb{P}_\theta\left(\mathbf{X} = (x_1, \ldots, x_{200}) \,|\, T(\mathbf{X}) = 5\right) = 1/\binom{200}{5}$
- Assim,

$$
\begin{aligned}
1/\binom{200}{5} &= \mathbb{P}_{\theta=0.01}\left(\mathbf{X} = (x_1, \ldots, x_{200}) \,|\, T(\mathbf{X}) = 5\right) \\
&= \mathbb{P}_{\theta=0.95}\left(\mathbf{X} = (x_1, \ldots, x_{200}) \,|\, T(\mathbf{X}) = 5\right) \\
&= \mathbb{P}_{\theta}\left(\mathbf{X} = (x_1, \ldots, x_{200}) \,|\, T(\mathbf{X}) = 5\right) \quad \forall\, \theta \in (0,1)
\end{aligned}
$$

- A conclusão genial de Fisher é que para estimar θ
  - é suficiente considerar $T(\mathbf{X})$=n^0 de uns.
- O que resta de aleatoriedade nos dados (sua ordem)
  - já não tem nenhuma relação com θ.

Definição

- *Sejam $X_1, \ldots, X_n$ v.a.'s com distribuição conjunta $p(\mathbf{x}; \boldsymbol{\theta})$.*
- *Uma estatÃstica é* **suficiente** *para $\boldsymbol{\theta}$ se a distribuição condicional de $\mathbf{X}$, dada o valor de $T(\mathbf{X})$,* **não depende de** $\boldsymbol{\theta}$.

Exemplo:

- Sejam $X_1, \ldots, X_n$ i.i.d. $Poisson(\theta)$ com função de probabilidade conjunta

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_{i=1}^n x_i!} .$$

- Seja $T(\mathbf{X}) = \sum_{i=1}^n X_i$.
- Temos que $T \sim Poisson(n\theta)$ (somas de Poisson indep é uma v.a. Poisson)
- Assim

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) &= \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ e } T(\mathbf{x}) = t)}{P(T(\mathbf{X}) = t)} \\ &= \frac{\frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_{i=1}^n x_i!}}{\frac{(n\theta)^t e^{-n\theta}}{t!}} \\ &= \frac{t!}{n^t \prod_{i=1}^n x_i!} \end{aligned}$$

que não depende de θ.

Exemplo:

- Vimos que

$$P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) = \frac{t!}{n^t \prod_{i=1}^n x_i!}$$

que não depende de θ.

- Note que, diferente do caso binomial, os eventos não são equiprováveis.
- Apesar de não depender de θ, as sequências $\mathbf{x}$ possíveis (aquelas compatíveis com $T(\mathbf{X}) = t$) não possuem probabilidades iguais.
- Por exemplo, com $n = 3$ e $T(\mathbf{X}) = 2$, temos as seguintes sequências possíveis para o vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)$: $(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), \ldots$.
- Temos as probabilidades:

$$P_\theta(\mathbf{X} = (2,0,0) | T(\mathbf{X}) = 2) = \frac{2!}{3^2\ 2!0!0!} = 1/9$$

enquanto que

$$P_\theta(\mathbf{X} = (1,1,0) | T(\mathbf{X}) = 2) = \frac{2!}{3^2\ 1!1!0!} = 2/9$$

O ponto crucial

- O ponto crucial na definição de uma estatística suficiente é que

$$P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t)$$

  NÃO DEPENDE DE θ.
- Não é importante que os eventos possíveis agora, após condicionarmos em $T(\mathbf{X}) = t$, sejam equiprováveis.
- O ponto é que sua distribuição não depende de θ.
- Assim, após sabermos o valor da estatística suficiente $T(\mathbf{X})$, a distribuição dos dados já não depende de θ, é tão livre do parâmetro θ quanto o tempo de espera na fila do banco é livre da altura esperada θ de um aluno escolhido ao acaso na UFMG.
- Dispensamos os dados como um todo ficando apenas com o resumo $T(\mathbf{X})$.

Como encontrar a estatística suficiente?

- Como encontrar uma estatística $T(\boldsymbol{Y})$ suficiente?
- Temos duas alternativas.
- A primeira é ter um insight, talvez genial, para selecionar um candidato $T(\boldsymbol{Y})$.
- A seguir, precisamos checar que a distribuição de $(\boldsymbol{Y} \,|\, T(\boldsymbol{Y}) = t)$ realmente não depende de $\boldsymbol{\theta}$.
- Por exemplo, num modelo de regressão logística, qual é a estatística suficiente?
- Se formos capazes de imaginar o que seria esta estatística suficiente, teríamos depois o trabalho de provar que $(\boldsymbol{Y} \,|\, T(\boldsymbol{Y}) = t)$ não depende de $\boldsymbol{\theta}$.
- Dureza, não?

Como encontrar a estatística suficiente?

- A segunda opção é usar o teorema da fatoração de Neyman-Fisher.
- O teorema da fatoração nos fornece uma maneira direta, automática, óbvia, de se encontrar a estatística suficiente em qualquer problema, por mais complexo que seja o modelo probabilístico envolvido.
- **Teorema da fatoracao:** Num modelo paramétrico, $T(\boldsymbol{Y})$ é suficiente para estimar $\boldsymbol{\theta}$ se, e somente se, a densidade conjunta $f(\boldsymbol{y}, \boldsymbol{\theta})$ puder ser escrita como $g(T(\boldsymbol{y}), \boldsymbol{\theta})h(\boldsymbol{y})$.
- Isto é:
  - a densidade conjunta (ou a verossimilhanca) é escrita como o produto de duas funçes: $g(T(\boldsymbol{y}), \boldsymbol{\theta})h(\boldsymbol{y})$
  - Somente uma delas envolve $\boldsymbol{\theta}$: a função $g(T(\boldsymbol{y}), \boldsymbol{\theta})$.
  - Nesta única função envolvendo $\boldsymbol{\theta}$, os dados $\boldsymbol{y}$ aparecem resumidos como $T(\boldsymbol{y})$.

Fatoração e a log-verossimilhança

- Às vezes, o teorema da fatoração é apresentado no contexto da log-verossimilhança.
- Neste caso, tomando o log da densidade conjunta, devemos ter

$$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y}, \boldsymbol{\theta}) = \log(g(T(\boldsymbol{y}), \boldsymbol{\theta})) + \log(h(\boldsymbol{y})) \,,$$

- Assim a log-verossimilhança $\ell(\boldsymbol{\theta})$ é expressa como uma soma (ao invés de produto) de funções com as características de g e h mencionadas antes.
- Isto é, $\ell(\boldsymbol{\theta})$ é uma soma de duas funções.
- Uma delas envolve apenas os dados $\boldsymbol{y}$.
- A outra função envolve o parâmetro $\boldsymbol{\theta}$ e os dados $\boldsymbol{y}$ mas os dados entram na função apenas através do valor da estatística $T(\boldsymbol{y})$.

Exemplo: Bernoulli i.i.d.

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Bernoulli(θ). Então

$$\begin{aligned} f(\boldsymbol{y}, \theta) &= \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i} = \theta^{\sum_i y_i}\,(1-\theta)^n\,(1-\theta)^{-\sum_i y_i} \\ &= \theta^{\sum_i y_i}\,(1-\theta)^n\,\left(\frac{1}{1-\theta}\right)^{-\sum_i y_i} \\ &= \left(\frac{\theta}{1-\theta}\right)^{\sum_i y_i}(1-\theta)^n \end{aligned}$$

- Seja $T(\boldsymbol{y}) = \sum_i y_i$. Então mostramos acima que

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\theta}) &= \left(\frac{\theta}{1-\theta}\right)^{T(\boldsymbol{y})}(1-\theta)^n \\ &= g(T(\boldsymbol{y}), \theta) \end{aligned}$$

- Neste modelo, tomamos $h(\boldsymbol{y}) = 1$.
- Assim, $T(\boldsymbol{Y}) = \sum_i Y_i$ é a estatística suficiente para estimar θ.

Exemplo: Poisson i.i.d.

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Poisson(θ). Então

$$\begin{aligned} f(\boldsymbol{y}, \theta) &= \prod_i \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \theta^{\sum_i y_i} e^{-n\theta} \frac{1}{\prod_i y_i!} \\ &= \theta^{T(\boldsymbol{y})} e^{-n\theta} \frac{1}{\prod_i y_i!} \\ &= g\left(T(\boldsymbol{y}), \theta\right) h(\boldsymbol{y}) \end{aligned}$$

  onde $T(\boldsymbol{y}) = \sum_i y_i$ e $h(\boldsymbol{y}) = 1/\left(\prod_i y_i!\right)$.
- Portanto, $T(\boldsymbol{Y}) = \sum_i Y_i$ é estatística suficiente para θ.

Caso multivariado

- Suponha que $Y_1, Y_2, \ldots, Y_n$ são v.a.'s com densidade $f(\boldsymbol{y}, \boldsymbol{\theta})$.
- Suponha que $\boldsymbol{\theta}$ é um VETOR k-dimensional.
- Geralmente, a estatística suficiente $\boldsymbol{T}(\boldsymbol{Y})$ será também um vetor k-dimensional.
- Cada elemento do vetor estatística suficiente $\boldsymbol{T}(\boldsymbol{Y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y}), \ldots, T_k(\boldsymbol{y}))$ será uma função dos dados.
- O teorema da fatoração continua válido:
- **Teorema da fatoracao:** O vetor $\boldsymbol{T}(\boldsymbol{Y}) = (T_1(\boldsymbol{Y}), \ldots, T_k(\boldsymbol{Y})$ é suficiente para estimar $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ se, e somente se,

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\theta})\ h(\boldsymbol{y})\ .$$

- Equivalentemente, se, e somente se, a log-verossimilhança é dada por

$$\ell(\boldsymbol{\theta}) = \log(f(\boldsymbol{y}, \boldsymbol{\theta})) = \log(g(\boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\theta})) + \log(h(\boldsymbol{y}))$$

Exemplo multivariado

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. $N(\mu, \sigma^2)$.
- Agora $\boldsymbol{\theta} = (\mu, \sigma^2)$ é um vetor.

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \right] \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\mu \sum_i y_i}{\sigma^2} \right) \exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \right] \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\mu \sum_i y_i}{\sigma^2} \right) \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \right] \exp\left(-\frac{T_1(\boldsymbol{y})}{2\sigma^2} + \frac{\mu T_2(\boldsymbol{y})}{\sigma^2} \right)
\end{aligned}
$$

- onde $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y})) = (\sum_i y_i^2, \sum_i y_i)$.

Exemplo multivariado

- Repetindo:

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \left[\left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\right] \; exp\left(-\frac{T_1(\boldsymbol{y})}{2\sigma^2} + \frac{\mu T_2(\boldsymbol{y})}{\sigma^2}\right)$$

  onde $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y})) = (\sum_i y_i^2, \sum_i y_i)$.
- Neste modelo, $f(\boldsymbol{y}, \boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\theta})\; h(\boldsymbol{y})$ com $h(\boldsymbol{y}) = 1$.
- O primeiro fator na expressão de $g(\boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\theta})$ envolve apenas $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- O segundo fator envolve $\boldsymbol{\theta}$ e os dados $\boldsymbol{y}$.
- Os dados aparecem na expressão apenas através de seus resumos $T_1(\boldsymbol{y}) = \sum_i y_i^2$ e $T_2(\boldsymbol{y}) = \sum_i y_i$.
- Assim, $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y})) = (\sum_i y_i^2, \sum_i y_i)$ é estatística suficiente para estimar $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$\boldsymbol{T}(\boldsymbol{y})$ e o MLE

- De fato, o MLE de $\boldsymbol{\theta} = (\mu, \sigma^2)$ é função direta da estatística suficiente $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{Y}), T_2(\boldsymbol{Y})) = (\sum_i Y_i^2, \sum_i Y_i)$.
- Temos $\widehat{\mu}_{MLE} = \overline{Y} = \sum_i Y_i/n = T_2(\boldsymbol{Y})/n$.
- E $\widehat{\sigma^2}_{MLE} = \sum_i (Y_i - \overline{Y})^2/n = \sum_i Y_i^2/n - (\overline{Y})^2 = T_1(\boldsymbol{Y})/n - T_2(\boldsymbol{Y})$
- A partir da estatística suficiente obtemos o MLE.

Regressão Linear

- $Y_1, Y_2, \ldots, Y_n$ são indep mas não são i.d.
- $Y_i \sim N(\mu_i, \sigma^2)$ onde

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ &= \beta_0 + \sum_j x_{ij}\beta_k \\ &= (1, x_{i1}, \ldots, x_{ip})' \, (\beta_0, \beta_1, \ldots, \beta_p) \\ &= \boldsymbol{x}_i' \boldsymbol{\beta} \end{aligned}$$

- Lembre-se que o vetor $(p+1) \times 1$ de features $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$ do i-ésimo item é composto por constantes (e não por v.a.'s).
- O vetor de parâmetros é $p+2$-dimensional: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.
- Temos a densidade bem parecida com o caso i.i.d., apenas fazendo o valor esperado μ_i de cada Y_i variar com o índice i ao invés de ser o valor constante μ.

Regressão Linear

- Temos

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right) \right] \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \mu_i)^2 \right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i (\mu_i y_i)}{\sigma^2} \right) \exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2} \right) \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2} \right) \right] \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i (\mu_i y_i)}{\sigma^2} \right)
\end{aligned}
$$

- A expressão acima está escrita em termos dos n valores $\mu_1, \mu_2, \ldots, \mu_n$, um para cada item da amostra.
- Queremos que apareçam os parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.
- Para isto, vamos substituir μ_i por $\boldsymbol{x}_i' \boldsymbol{\beta}$.

Regressão Linear

- Encontramos

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = k(\boldsymbol{\theta}) \ \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i (\mu_i y_i)}{\sigma^2}\right)$$

onde

$$k(\boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2}\right)$$

não envolve os dados $\boldsymbol{y}$.

- Substituindo $\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ na expressão para $f(\boldsymbol{y}, \boldsymbol{\theta})$, a única parte envolvendo $\boldsymbol{y}$ que será afetada é

$$\begin{aligned}
\sum_i (\mu_i y_i) &= \sum_i (y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \\
&= \sum_i (y_i \beta_0) + \sum_i (y_i \beta_1 x_{i1}) + \ldots + \sum_i (y_i \beta_p x_{ip}) \\
&= \beta_0 \sum_i y_i + \beta_1 \sum_i (y_i x_{i1}) + + \ldots + \beta_p \sum_i (y_i x_{ip})
\end{aligned}$$

Regressão Linear

- Portanto,

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= k(\boldsymbol{\theta}) \ \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\beta_0 \sum_i y_i + \beta_1 \sum_i (y_i x_{i1}) + + \ldots + \beta_p \sum_i (y_i x_{ip})}{\sigma^2} \right) \\
&= k(\boldsymbol{\theta}) \ \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\beta_0 T_0(\boldsymbol{y}) + \beta_1 T_1(\boldsymbol{y}) + + \ldots + \beta_p T_p(\boldsymbol{y})}{\sigma^2} \right)
\end{aligned}
$$

- onde $\boldsymbol{T}(\boldsymbol{y}) = (T_0(\boldsymbol{y}), T_1(\boldsymbol{y}), \ldots, T_p(\boldsymbol{y}), T_{p+1}(\boldsymbol{y})) = (\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip}), \sum_i y_i^2)$.
- Pelo Teorema da fatoração, a estatística $\boldsymbol{T}(\boldsymbol{y})$ é suficiente para estimar $\boldsymbol{\theta}$.

Regressão Linear

- Podemos escrever a estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$ do modelo de regressão linear de forma mais compacta usando matrizes.
- Esta notação é importante: ela vai aparecer em modelos de regressão generalizado.
- As primeiras entradas $(p+1)$ entradas $(T_0(\boldsymbol{y}), T_1(\boldsymbol{y}), \ldots, T_p(\boldsymbol{y}))$ da estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$ podem ser escritas como (VERIFIQUE!!!!)

$$\left(\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip}) \right) = \boldsymbol{X}^t \boldsymbol{y}$$

onde

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \text{e} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Regressão Linear

- A última entrada, $T_{p+1}(\boldsymbol{y}) = \sum_i y_i^2$ da estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$ é o comprimento ao quadrado do vetor $\boldsymbol{y}$.
- Isto é, $T_{p+1}(\boldsymbol{y}) = \sum_i y_i^2 = \boldsymbol{y}' \ \boldsymbol{y}$.
- Assim, a estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$ do modelo de regressão linear pode ser escrita como

$$\boldsymbol{T}(\boldsymbol{y}) = (\boldsymbol{X}^t \, \boldsymbol{y}, \boldsymbol{y}' \ \boldsymbol{y})$$

- Note que o MLE de $\boldsymbol{\beta}$ é baseado nesta estatística suficiente:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \, \boldsymbol{X})^{-1} \, \boldsymbol{X}^t \, \boldsymbol{Y}$$

- O MLE de σ^2 também é obtido diretamente da estatística suficiente.
- Veremos a seguir.

Regressão Linear

- Temos
$$\widehat{\sigma^2} = \frac{1}{n}\sum_i (y_i - \widehat{y_i})^2 = \frac{1}{n}|| \boldsymbol{Y} - \widehat{\boldsymbol{Y}}||^2$$
onde $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\beta}$, uma função da estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$.
- O vetor $\widehat{\boldsymbol{Y}}$ é a projeção ortogonal do vetor $\boldsymbol{Y}$ no espaço vetorial das combinações lineraes das colunas da matriz de desenho bX.
- Assim, $\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right) = \widehat{\boldsymbol{Y}} + \boldsymbol{r}$ com o vetor $\widehat{\boldsymbol{Y}}$ ortogonal ao vetor de resíduos $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$
- Por causa da ortogonalidade, temos
$$|| \boldsymbol{Y} ||^2 = ||\widehat{\boldsymbol{Y}}||^2 + || \boldsymbol{Y} - \widehat{\boldsymbol{Y}}||^2$$
- Em conclusão, o MLE $\widehat{\sigma^2}$ é função da estatística suficiente $\boldsymbol{T}(\boldsymbol{y})$ pois
$$\widehat{\sigma^2} = \frac{1}{n}|| \boldsymbol{Y} - \widehat{\boldsymbol{Y}}||^2 = \frac{1}{n}\left(|| \boldsymbol{Y} ||^2 - ||\widehat{\boldsymbol{Y}}||^2\right) = \frac{1}{n}\left(T_{p+1}(\boldsymbol{Y}) - \boldsymbol{X}\,\widehat{\beta}\right)$$

Regressão logística

- Lembre-se das definições: Observamos $Y_1, \ldots, Y_n$ v.a.'s binárias independentes mas nào são i.d.
- A probabilidade de sucesso $p_i = \mathbb{P}(Y_i = 1)$ varia de item para item (varia com i).
- Seja $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$ um vetor de atributos (features) medidos em cada exemplo.
- Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$ um vetor de parâmetros ou pesos desconhecidos.
- Assumimos que

$$\mathbb{P}(Y_i = 1) = p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}} = \frac{1}{1 + e^{-\boldsymbol{x}^t . \boldsymbol{\theta}}}$$

- Chamamos $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ de preditor linear do sucesso

Notação matricial na regressão logística

- Vamos lembrar

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Então

$$p_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip})}} = \frac{1}{1+e^{-\boldsymbol{x}_i^t \boldsymbol{\theta}}}$$

- onde $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^t$ é a i-ésima LINHA da matriz $\boldsymbol{X}$ visto como um vetor-coluna
- e $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor-coluna de parâmetros desconhecidos.

Verossimilhança

- Parâmetro que queremos estimar: $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$.
- Densidade conjunta:

$$\begin{aligned} p(\boldsymbol{y}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = 1)^{y_i} \; \mathbb{P}(Y_i = 0)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{p_i}{1-p_i} \right)^{y_i} \; (1-p_i) \end{aligned}$$

- Verifique, a partir da fórmula logística para p_i, que

$$\frac{p_i}{1-p_i} = \exp\left(\boldsymbol{x}_i^t \, \boldsymbol{\theta}\right)$$

- Assim

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left(e^{\boldsymbol{x}_i^t \boldsymbol{\theta}} \right)^{y_i} \; \prod_{i=1}^{n} \left(1 - \frac{1}{1 + e^{-\boldsymbol{x}_i^t \boldsymbol{\theta}}} \right)$$

Verossimilhança

- Repetindo:

$$
\begin{aligned}
p(\boldsymbol{y};\boldsymbol{\theta}) &= \prod_{i=1}^{n}\left(e^{\boldsymbol{x}_i^t\boldsymbol{\theta}}\right)^{y_i}\ \prod_{i=1}^{n}\left(1-\frac{1}{1+e^{-\boldsymbol{x}_i^t\boldsymbol{\theta}}}\right)\\
&= \prod_{i=1}^{n}e^{y_i\boldsymbol{x}_i^t\boldsymbol{\theta}}\ \prod_{i=1}^{n}\left(1-\frac{1}{1+e^{-\boldsymbol{x}_i^t\boldsymbol{\theta}}}\right)\\
&= \exp\left(\sum_i y_i\boldsymbol{x}_i^t\boldsymbol{\theta}\right)\ \prod_{i=1}^{n}\left(1-\frac{1}{1+e^{-\boldsymbol{x}_i^t\boldsymbol{\theta}}}\right)
\end{aligned}
$$

Verossimilhança

- Considerando apenas o expoente da exponencial, temos:

$$\begin{aligned}
\sum_i y_i \boldsymbol{x}_i^t \boldsymbol{\theta} &= \sum_i y_i \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right) \\
&= \beta_0 \sum_i y_i + \beta_1 \sum_i (y_i x_{i1}) + \ldots + \beta_p \sum_i (y_i x_{ip}) \\
&= \left(\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip})\right)' (\beta_0, \beta_1, \ldots, \beta_p) \\
&= \left(\boldsymbol{X}' \boldsymbol{Y}\right) \boldsymbol{\theta}
\end{aligned}$$

- Note que $\boldsymbol{X}' \boldsymbol{Y}$ é um vetor $(p+1) \times 1$.

Verossimilhança

- Em conclusão, a densidade conjunta é

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left(\left(\boldsymbol{X}'\,\boldsymbol{Y}\right)\boldsymbol{\theta}\right)\ \prod_{i=1}^{n}\left(1-\frac{1}{1+e^{-\boldsymbol{x}_i^t\boldsymbol{\theta}}}\right)$$

- O último fator é um termos que envolve apenas $\boldsymbol{\theta}$, não envolve os dados $\boldsymbol{Y}$.
- O primeiro termo, o fator exponencial, envolve $\boldsymbol{\theta}$ e os dados $\boldsymbol{Y}$.
- Os dados n-dimensionais aperecem apenas através do vetor $(p+1)$-dimensional

$$\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{X}'\,\boldsymbol{Y} = \left(\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip})\right)$$

Verossimilhança

- Obtivemos a estatística suficiente para estimar $\boldsymbol{\theta}$:

$$\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{X}'\,\boldsymbol{Y} = \left(\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip})\right)$$

- Lembre-se que $y_i = 1$ ou $y_i = 0$.
- Então, no vetor $\boldsymbol{T}(\boldsymbol{Y})$, a entrada j é a soma dos valores da covariável j apenas para aqueles itens em que $y_i = 1$.
- Isto é, $\boldsymbol{T}(\boldsymbol{Y})$ é um vetor de somas parciais das covariáveis, somando-se apenas os casos de sucesso.

Verossimilhança em notação matricial

- Revise os passos para obter o MLE de $\boldsymbol{\theta}$ na regressão logística (algumas aulas atrás.
- MLE é aquele vetor $\boldsymbol{\theta}$ tal que o vetor de probabilidades $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ induzido por este $\boldsymbol{\theta}$ é tal que satisfaz a equação de verossimilhança:

$$\boldsymbol{X}^t \boldsymbol{y} = \boldsymbol{X}^t \boldsymbol{p}$$

- Portanto, para obter o MLE precisamos apenas da matriz de constantes $\boldsymbol{X}$ e, no que concerne aos dados $\boldsymbol{Y}$, precisamos apenas da estatística suficiente $\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{X}' \boldsymbol{Y}$.
- Isto é, o MLE é função da estatística suficiente $\boldsymbol{T}(\boldsymbol{Y})$.

Prova do teorema da fatoração

- Prova omitida.

Neyman-Fisher Factorization Theorem (Multivariate Case)

- Suppose $Y_1, Y_2, \ldots, Y_n$ are random variables with joint density $f(\mathbf{y}; \boldsymbol{\theta})$.
- Let $\boldsymbol{\theta} \in \mathbb{R}^k$ be a k-dimensional parameter vector.
- In general, the sufficient statistic $\mathbf{T}(\mathbf{Y})$ is also a k-dimensional vector.
- Each element of $\mathbf{T}(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}), \ldots, T_k(\mathbf{Y}))$ is a function of the data.
- **Factorization Theorem:** $\mathbf{T}(\mathbf{Y})$ is sufficient for $\boldsymbol{\theta}$ if and only if:

$$f(\mathbf{y}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{y}), \boldsymbol{\theta}) \cdot h(\mathbf{y})$$

- Equivalently, in terms of the log-likelihood:

$$\ell(\boldsymbol{\theta}) = \log f(\mathbf{y}; \boldsymbol{\theta}) = \log g(\mathbf{T}(\mathbf{y}), \boldsymbol{\theta}) + \log h(\mathbf{y})$$

Sketch of the Proof: Step 1

Assume: $f(\mathbf{y};\boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{y}),\boldsymbol{\theta}) \cdot h(\mathbf{y})$

- Let $\mathbf{T}(\mathbf{Y})$ be the statistic defined in the factorization.
- For fixed $\mathbf{t} = \mathbf{T}(\mathbf{y})$, the conditional density of $\mathbf{Y}$ given $\mathbf{T}(\mathbf{Y}) = \mathbf{t}$ is:

$$f(\mathbf{y}|\mathbf{T}(\mathbf{Y}) = \mathbf{t};\boldsymbol{\theta}) \propto f(\mathbf{y};\boldsymbol{\theta}) = g(\mathbf{t},\boldsymbol{\theta}) \cdot h(\mathbf{y})$$

- Since $g(\mathbf{t},\boldsymbol{\theta})$ is constant over all $\mathbf{y}$ such that $\mathbf{T}(\mathbf{y}) = \mathbf{t}$, the conditional distribution depends only on $h(\mathbf{y})$, and not on $\boldsymbol{\theta}$.
- Hence, the conditional distribution of $\mathbf{Y}$ given $\mathbf{T}(\mathbf{Y})$ is independent of $\boldsymbol{\theta}$.

Sketch of the Proof: Step 2

Now assume: $\mathbf{T}(\mathbf{Y})$ is sufficient for $\boldsymbol{\theta}$.

- Then the conditional density $f(\mathbf{y}|\mathbf{T}(\mathbf{Y}) = \mathbf{t}; \boldsymbol{\theta})$ is free of $\boldsymbol{\theta}$.
- Thus, we can write:

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{T}(\mathbf{y}); \boldsymbol{\theta}) \cdot p(\mathbf{T}(\mathbf{y}); \boldsymbol{\theta})$$

- The first term is free of $\boldsymbol{\theta}$ and can be written as $h(\mathbf{y})$.
- The second term depends only on $\mathbf{T}(\mathbf{y})$ and $\boldsymbol{\theta}$: write it as $g(\mathbf{T}(\mathbf{y}), \boldsymbol{\theta})$.
- So:

$$f(\mathbf{y}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{y}), \boldsymbol{\theta}) \cdot h(\mathbf{y})$$

completing the factorization.

MLE é sempre função da estatística suficiente

- A estatística suficiente $T(\boldsymbol{Y})$ resume toda a informação sobre $\boldsymbol{\theta}$ existente nos dados.
- A distribuição dos dados $\boldsymbol{Y}$ condicionada no valor observado de $T(\boldsymbol{Y})$ não depende de $\boldsymbol{\theta}$.
- Qual a relação do MLE com a estatística suficiente?
- Temos um resultado simples: Se existir estatística suficiente $t(\boldsymbol{Y})$, então $\widehat{\boldsymbol{\theta}}_{MLE}$ é uma função de $T(\boldsymbol{Y})$.
- PROVA: Pelo Teorema da fatoração, $f(\boldsymbol{y}, \boldsymbol{\theta}) = g(T(\boldsymbol{y}), \boldsymbol{\theta})h(\boldsymbol{y})$. Se quisermos maximizar $f(\boldsymbol{y}, \boldsymbol{\theta})$ em $\boldsymbol{\theta}$, devemos maximizar $g(T(\boldsymbol{y}), \boldsymbol{\theta})$. Como os dados aparecem aqui apenas através do valor do resumo $T(\boldsymbol{y})$, a solução $\widehat{\boldsymbol{\theta}}_{MLE}$ que vai maximizar $g(T(\boldsymbol{y}), \boldsymbol{\theta})$ vai depender dos dados apenas através $T(\boldsymbol{y})$. Isto é, $\widehat{\boldsymbol{\theta}}_{MLE}$ é função de $T(\boldsymbol{Y})$.

MLE is Asymptotically Sufficient

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, $\theta \in \Theta \subset \mathbb{R}$.

A statistic $T(\mathbf{X})$ is sufficient to estimate θ if $f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$.

Equivalently:
$\log(f(\mathbf{x};\theta)) = \log(g(T(\mathbf{x});\theta)) + \log(h(\mathbf{x})) = G(T(\mathbf{x});\theta) + H(\mathbf{y})$.

If the sample size n is large, the MLE $\hat{\theta}_n$ is approximately sufficient to estimate θ. That is: $\log(f(\mathbf{x};\theta)) = G(\hat{\theta}_n;\theta) + H(\mathbf{x})$.

How to prove: In large samples, the likelihood becomes sharply peaked around $\hat{\theta}_n$, taking an approximately Gaussian shape. This will imply that the MLE is approximately sufficient to estimate θ.

Sketch of the Proof

We factorize the likelihood:

$$
\begin{aligned}
L_n(\theta) = f(\mathbf{x}, \theta) &= \prod_{i=1}^{n} f(x_i; \theta) \\
&= \exp\left(\log \prod_{i=1}^{n} f(x_i; \theta) \right) \\
&= \exp\left(\sum_{i=1}^{n} \log f(x_i; \theta) \right) \\
&= \exp(\ell_n(\theta))
\end{aligned}
$$

Sketch of the Proof

Using a second-order Taylor expansion around $\hat{\theta}_n$:

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)\ell'_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2\ell''_n(\hat{\theta}_n)$$

- Since $\hat{\theta}_n$ maximizes the log-likelihood, $\ell'_n(\hat{\theta}_n) = 0$ and then:

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2\ell''_n(\hat{\theta}_n)$$

Sketch of the Proof

- $\ell_n(\theta) = \log f(\mathbf{x}; \theta)$
- $\hat{\theta}_n$ is the MLE and therefore DO NOT INVOLVE the unknown parameter θ.
- Therefore, $\ell_n(\hat{\theta}_n) = \log f(\mathbf{x}; \hat{\theta}_n)$ is not a function of θ
- it is a function only of the data $\mathbf{x}$.
- That is: $\ell_n(\hat{\theta}_n) = H(\mathbf{x})$

$$\ell_n(\theta) \approx \overbrace{\ell_n(\hat{\theta}_n)}^{\text{constant in } \theta} + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n) \quad (1)$$

$$= H(\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n) \quad (2)$$

Sketch of the Proof

Repeating:

$$\log(f(\mathbf{x};\theta)) = \ell_n(\theta) \approx H(\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n)$$

Now, consider $(1/n)\ell_n''(\theta)$: By the Law of Large Numbers,

$$\frac{1}{n}\ell_n''(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log(f(x_i;\theta))}{\partial\theta^2} \to \mathbb{E}_\theta\left(\frac{\partial^2 \ell}{\partial\theta^2}\right) = -\mathbb{I}_1(\theta)$$

Sketch of the Proof

Therefore:

$$\begin{aligned}
\log(f(\mathbf{x};\theta)) &\approx H(\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n) \\
&= H(\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \frac{n}{n} \ell_n''(\hat{\theta}_n) \\
&\approx H(\mathbf{x}) - \frac{1}{2}(\theta - \hat{\theta}_n)^2 n\mathbb{I}_1(\hat{\theta}_n) \\
&= H(\mathbf{x}) + G(\hat{\theta}_n;\theta)
\end{aligned}$$

Exponentiating both sides of the approximation:

$$f(\mathbf{x};\theta)) \approx e^{H(\mathbf{x})} \cdot e^{G(\hat{\theta}_n,\theta)} = h(\mathbf{x})\ g(\hat{\theta}_n,\theta)$$

Conclusion: MLE $\hat{\theta}_n$ is asymptotically sufficient to estimate θ.

Teorema de Koopman, Pitman e Darmois

- Quando existe uma estatística suficiente? Sempre!
- O próprio vetor aleatório de dados $\boldsymbol{y}$ é uma estatística suficiente pois a distribuição de $(\boldsymbol{Y} \mid \boldsymbol{Y} = \boldsymbol{y})$ é massa pontual em $\boldsymbol{y}$.
- Esta estatística não é interessante: ela não resume os dados, ela cresce com o número de dados.
- Quando existe uma estatística suficiente $T(y)$ que **não escala** com os dados?
- Isto é, quando existe uma estatística suficiente $T(\boldsymbol{y})$ que seja um vetor de dimensão **fixa**, cuja dimensão não cresça com o número de dados n?
- Resposta: Se, e somente se, $\boldsymbol{Y}$ pertencer a família exponencial de distribuições.
- Este é o teorema de Koopman, Pitman e Darmois (provaram ao mesmo tempo mas independentemente, publicando em 1935 e 1936).

Família exponencial - caso univariado

- Considere um modelo de probabilidade que dependa de um parâmetro unidimensional θ.
- Uma família de distribuições paramétricas pertence à família exponencial com 1 parâmetro se

$$p(\boldsymbol{y}, \theta) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\eta(\theta) T(\boldsymbol{y})\right)$$

  e o suporte de $\boldsymbol{y}$ (conjunto de valores possíveis) não depende de θ.
- A expressão $\eta(\theta)$ é o parâmetro natural da família exponencial.
- $p(\boldsymbol{y}, \theta)$ envolve o produto de funções em que $\boldsymbol{\theta}$ e $\boldsymbol{y}$ aparecem separadamente: isto é, ela envolve $h(\boldsymbol{y})$. $g(\boldsymbol{\theta})$.
- Ela envolve também uma função $\exp\left(\eta(\theta) T(\boldsymbol{y})\right)$ em que os dados $\boldsymbol{y}$ e o parâmetro aparecem misturados.
- Não podemos separar esta função em dois pedaços, cada um envolvendo apenas $\boldsymbol{\theta}$ e apenas $\boldsymbol{y}$.
- Este é o ponto CRUCIAL na definição.

Família exponencial - caso univariado

- Família exponencial se

$$p(\boldsymbol{y}, \theta) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\eta(\theta) T(\boldsymbol{y})\right)$$

- O expoente do termo exponencial $\exp\left(\eta(\theta) T(\boldsymbol{y})\right)$ envolve um produto de uma função apenas dos dados e outra apenas de θ.
- É esta forma específica deste expoente do termo exponencial que fornece as boas propriedades dafamília exponencial para a estimação.

Exemplo - Bernoulli

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Bernoulli(θ).
- Seja $T(\boldsymbol{y}) = \sum_i y_i$. Então

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\theta}) &= \left(\frac{\theta}{1-\theta}\right)^{T(\boldsymbol{y})} (1-\theta)^n \\ &= \exp\left(\log\left(\frac{\theta}{1-\theta}\right)^{T(\boldsymbol{y})}\right)(1-\theta)^n \\ &= \exp\left(T(\boldsymbol{y})\log\left(\frac{\theta}{1-\theta}\right)\right)(1-\theta)^n \end{aligned}$$

Exemplo - Bernoulli

- Este modelo pertence à família exponencial com $h(\boldsymbol{y}) = 1$ e $g(\boldsymbol{\theta}) = (1-\theta)^n$.
- O parâmetro natural deste membro da família exponencial é a log-odds:

$$\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

- A estatística suficiente é $T(\boldsymbol{Y}) = \sum_i Y_i$.
- Note que as funções $h(\boldsymbol{y})$ e $g(\boldsymbol{\theta})$ podem ser constantes e iguais a 1.

Exemplo: Poisson i.i.d.

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Poisson(θ). Então

$$f(\boldsymbol{y}, \theta) = \theta^{T(\boldsymbol{y})} e^{-n\theta} \; \frac{1}{\prod_i y_i!}$$

onde $T(\boldsymbol{y}) = \sum_i y_i$.

- Tomando exp e log ao mesmo tempo:

$$f(\boldsymbol{y}, \theta) = \exp\left(T(\boldsymbol{Y}) \log(\theta)\right) \; e^{-n\theta} \; \frac{1}{\prod_i y_i!}$$

Exemplo: Poisson i.i.d.

- Assim, este modelo também pertence à família exponencial com

$$h(\boldsymbol{y}) = \frac{1}{\prod_i y_i!}$$

e

$$g(\boldsymbol{\theta}) = e^{-n\theta}$$

.

- O parâmetro natural deste membro da família exponencial é o log de θ:

$$\eta(\theta) = \log(\theta)$$

- A estatística suficiente é $T(\boldsymbol{Y}) = \sum_i Y_i$.

Exemplo: Pareto i.i.d.

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Pareto(α).
- É uma distribuição contínua muito importante para modelar dados de caudas pesadas, dados com valores extremos.
- Também chamada de power-law. A distribuição de Zipf é a versão discreta da Pareto.
- O suporte da distribuição (ou conjunto de valores possíveis) é (c, ∞) onde c é uma constante maior que zero.
- Densidade de $Y \sim$ Pareto(α) é

$$f(y;\alpha) = \begin{cases} 0, & \text{se } y \leq c \\ \frac{\alpha c^\alpha}{y^{\alpha+1}}, & \text{se } y > c \end{cases}$$

Exemplo: Pareto i.i.d.

- Densidade de $Y \sim \text{Pareto}(\alpha)$ é

$$f(y;\alpha) = \begin{cases} 0, & \text{se } y \leq c \\ \frac{\alpha c^\alpha}{y^{\alpha+1}}, & \text{se } y > c \end{cases}$$

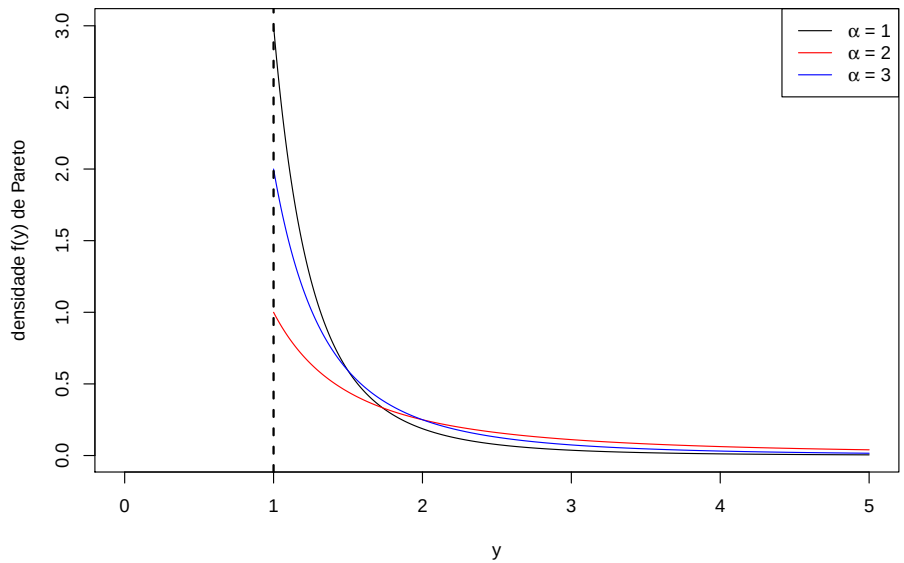
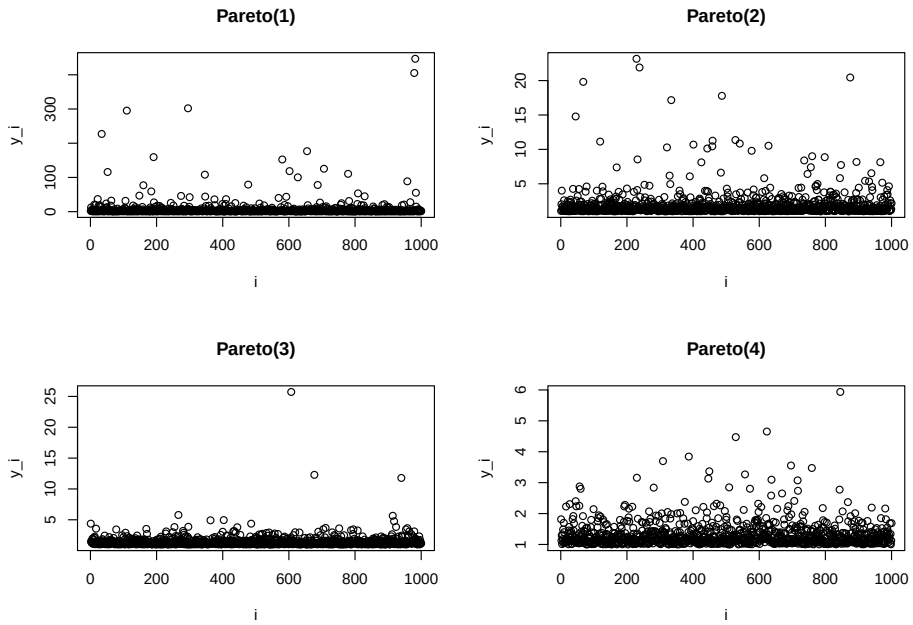


Figura: Densidade de Pareto com $c = 1$ e $\alpha = 1, 2, 3$.

Exemplo: Pareto i.i.d.

- Amostras de tamanho $n = 1000$ de Pareto(1), Pareto(2), Pareto(3), Pareto(4). Plots de y_i versus i para cada amostra



- Se Y_i é a perda financeira devido a incêndios, espera-se $\alpha \approx 1.5$. Para perdas em seguro completo de automóveis (incluindo contra terceiros), espera-se $\alpha \approx 2.5$.

Pareto i.i.d.

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com distribuição Pareto(α) e constante c. Então

$$\begin{aligned} f(\boldsymbol{y}, \alpha) &= \prod_{i=1}^{n} \frac{\alpha c^{\alpha}}{y_i^{\alpha+1}} = \frac{\alpha^n c^{n\alpha}}{(\prod_i y_i)^{\alpha+1}} \\ &= \alpha^n c^{n\alpha} \exp\left((\alpha+1) \log \left(\prod_i y_i \right)^{-1} \right) \\ &= \alpha^n c^{n\alpha} \exp\left((\alpha+1) \sum_i \log \frac{1}{y_i} \right) \end{aligned}$$

- Assim, esta distribuição pertence à classe da família exponencial 1-dim com $h(\boldsymbol{y}) = 1$, $g(\alpha) = \alpha^n c^{n\alpha}$.
- Temos $T(\boldsymbol{y}) = \sum_i \log(1/y_i)$ e $\eta(\alpha) = \alpha + 1$.

Outra expressão

- É comum alguns livros e papers apresentarem a família exponencial

$$p(\boldsymbol{y}, \theta) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\eta(\theta) T(\boldsymbol{y})\right)$$

da seguinte forma:

$$p(\boldsymbol{y}, \theta) = \exp\left(\eta(\theta) T(\boldsymbol{y}) + d(\boldsymbol{y}) + c(\boldsymbol{\theta})\right)$$

- Esta é apenas outra reexpressão pois

$$\begin{aligned} p(\boldsymbol{y}, \theta) &= \exp\left(\eta(\theta) T(\boldsymbol{y}) + d(\boldsymbol{y}) + c(\boldsymbol{\theta})\right) \\ &= \exp\left(\eta(\theta) T(\boldsymbol{y})\right) \exp(d(\boldsymbol{y})) \exp(c(\boldsymbol{\theta})) \end{aligned}$$

que cai na mesma expressão que usamos.

Família exponencial e suficiência

- Suponha que temos uma distribuição pertencente à família exponencial:

$$p(\boldsymbol{y}, \theta) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\eta(\theta) T(\boldsymbol{y})\right)$$

- A estatística $T(\boldsymbol{Y})$ de uma família exponencial é uma estatística suficiente. para estimar θ.
- PROVA: por definição, $T(\boldsymbol{Y})$ é suficiente se, e somente se,

$$p(\boldsymbol{y}, \theta) = h(\boldsymbol{y}) H(\boldsymbol{\theta}, T(\boldsymbol{y}))$$

- Se fizermos $g(\boldsymbol{\theta}) \exp\left(\eta(\theta) T(\boldsymbol{y})\right) \equiv H(\boldsymbol{\theta}, T(\boldsymbol{y}))$ vemos que $T(\boldsymbol{y})$ é estatística suficiente.

Família exponencial - caso multivariado

- Considere um modelo de probabilidade que dependa de um parâmetro k-dimensional $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ e onde o suporte de $\boldsymbol{y}$ (conjunto de valores possíveis) não depende de $\boldsymbol{\theta}$.
- Uma família de distribuições paramétricas pertence à família exponencial com k parâmetros se

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) T_j(\boldsymbol{y}) \right)$$

- Note que o termo no expoente da eponencial pode ser escrito em forma matricial:

$$\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) T_j(\boldsymbol{y}) = (\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta})) \begin{pmatrix} T_1(\boldsymbol{y}) \\ T_2(\boldsymbol{y}) \\ \vdots \\ T_k(\boldsymbol{y}) \end{pmatrix}$$

Estatística suficiente

- O vetor $\boldsymbol{T}(\boldsymbol{Y}) = (T_1(\boldsymbol{Y}), T_2(\boldsymbol{Y}), \ldots, T_k(\boldsymbol{Y}))$ é a estatística suficiente da família exponencial.
- O vetor $(\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))$ é chamado de parâmetro natural da família exponencial.

Exemplo: $N(\mu, \sigma^2)$

- $Y_1, Y_2, \ldots, Y_n$ v.a.'s i.i.d. com distribuição de probabilidade $N(\mu, \sigma^2)$.
- Densidade conjunta:
- Agora $\boldsymbol{\theta} = (\mu, \sigma^2)$ é um vetor.

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \right] \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \right] exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\mu \sum_i y_i}{\sigma^2} \right) \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \right] exp\left(-\frac{T_1(\boldsymbol{y})}{2\sigma^2} + \frac{\mu T_2(\boldsymbol{y})}{\sigma^2} \right)
\end{aligned}
$$

- onde $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y})) = (\sum_i y_i^2, \sum_i y_i)$ é a estatística suficiente
- e $(\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta})) = (-1/\sigma^2, \mu/\sigma^2)$.

Exemplo: $N(\mu, \sigma^2)$

- $Y_1, Y_2, \ldots, Y_n$ são indep mas não são i.d.
- $Y_i \sim N(\mu_i, \sigma^2)$ onde

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ &= \beta_0 + \sum_j x_{ij}\beta_k \\ &= (1, x_{i1}, \ldots, x_{ip})' \ (\beta_0, \beta_1, \ldots, \beta_p) \\ &= \boldsymbol{x}_i' \boldsymbol{\beta} \end{aligned}$$

- O vetor de parâmetros é $p+2$-dimensional: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.

Regressão Linear

- Temos

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right) \right] \\
&= \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2} \right) \right] \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i (\mu_i y_i)}{\sigma^2} \right)
\end{aligned}
$$

- A expressão acima está escrita em termos dos n valores $\mu_1, \mu_2, \ldots, \mu_n$, um para cada item da amostra.
- Queremos que apareçam os parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.
- Para isto, vamos substituir μ_i por $\boldsymbol{x}_i' \boldsymbol{\beta}$.

Regressão Linear

- Após alguma álgebra, encontramos

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= k(\boldsymbol{\theta}) \ \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\beta_0 \sum_i y_i + \beta_1 \sum_i (y_i x_{i1}) + + \ldots + \beta_p \sum_i (y_i x_{ip})}{\sigma^2} \right) \\
&= k(\boldsymbol{\theta}) \ \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\beta_0 T_0(\boldsymbol{y}) + \beta_1 T_1(\boldsymbol{y}) + + \ldots + \beta_p T_p(\boldsymbol{y})}{\sigma^2} \right)
\end{aligned}
$$

- onde $\boldsymbol{T}(\boldsymbol{y}) = (T_0(\boldsymbol{y}), T_1(\boldsymbol{y}), \ldots, T_p(\boldsymbol{y}), T_{p+1}(\boldsymbol{y})) = (\sum_i y_i, \sum_i (y_i x_{i1}), \ldots, \sum_i (y_i x_{ip}), \sum_i y_i^2)$.
- Já sabemos pelo Teorema da fatoração que a estatística $\boldsymbol{T}(\boldsymbol{y})$ é suficiente para estimar $\boldsymbol{\theta}$.

Outros modelos

- Também faz parte da família exponencial: modelo de regressão logística e muitos outros.
- É possível mostrar que, na família exponencial, estimadores baseados em estatísticas suficientes (tais como o MLE) possuem propriedades desejáveis ou ótimas.
- Podemos também deduzir fórmulas gerais para seu vício e variância: muito úteis para calcular o MSE (erro^2 médio de estimação)
- MAS ... não veremos mais a teoria de família exponencial.
- Passaremos agora a uma SUB-CLASSE DENTRO da família exponencial.
- Nesta SUB-CLASSE seremos capazes de ajustar modelos de regressão com MLE e estimá-los usando um único algoritmo: iterated reweighted least squares.
- Esta sub-classe é a dos modelos lineares generalizados: GLM

Teorema de Darmois-Pitman-Koopman

- Antes de GLM, um último comentário.
- Veja a direção da implicação.
- Se a distribuição de $\boldsymbol{Y}$ pertence à família exponencial

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) T_j(\boldsymbol{y}) \right)$$

  então $\boldsymbol{T}(\boldsymbol{Y}) = (T_1(\boldsymbol{Y}), T_2(\boldsymbol{Y}), \ldots, T_k(\boldsymbol{Y}))$ é estatística suficiente para estimar $\boldsymbol{\theta}$.
- A conversa NÃO É VÁLIDA:
- Se $T(\boldsymbol{y})$ é estatística suficiente NÃO IMPLICA que $\boldsymbol{Y}$ pertence à família exponencial.

Teorema de Darmois-Pitman-Koopman

- O Teorema de Darmois-Pitman-Koopman: com uma condição adicional a conversa é válida.
- A condição adicional é: $\boldsymbol{T}(\boldsymbol{y})$ deve ser um vetor de dimensão fixa, que não varia com o tamanho da amostra n.
- Se um vetor suficiente desta forma existir então a distribuição pertencerá à família exponencial.

Exemplo fora da família exponencial

- $Y_1, Y_2, \ldots, Y_n$ v.a.'s i.i.d. com distribuição de probabilidade Weibull.
- Para uma única v.a., temos a densidade

$$f(y;\boldsymbol{\theta}) = f(y;\alpha,\beta) = \begin{cases} \alpha\beta y^{\alpha-1} exp(-\beta y^{\alpha}), & \text{se } y > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

- Log-verossimilhança com n v.a.'s i.i.d.

$$\ell(\boldsymbol{\theta}) = n\log(\alpha\beta) + (\alpha - 1)\sum_i \log(y_i) - \beta\sum_i y_i^{\alpha}$$

Exemplo fora da família exponencial

- Usando o teorema da fatoração de Fisher-Neyman, conclímos que a única estatística suficiente é o próprio vetor completo das observações $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ pois não há como expressar a densidade conjunta com um "resumo" dos dados, uma função de dimensão menor.
- A dimensão da única estatística suficiente, o vetor $\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{Y}$, é igual a n, a dimensão do vetor de dados.
- Não existe estatística suficiente de dimensão menor que n.
- Não existe um pequeno (e fixo) número de funções *apenas dos dados* tais que, condicionada nestas estatísticas, a distribuição do que resta de aleatoriedade nos dados não dependa do parâmetro desconhecido $\boldsymbol{\theta}$.