Universidade Federal de Juiz de Fora
Departamento de Estatística
Curso de Estatística

Victor Basílio Faria

Estimação de Máxima Verossimilhança via Algoritmo EM

Juiz de Fora
2011

Victor Basílio Faria

Estimação de Máxima Verossimilhança via Algoritmo EM

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Orientador: Clécio da Silva Ferreira

Doutor em Estatística - Universidade de São Paulo

Juiz de Fora

2011

Faria, Victor

Estimação de Máxima Verossimilhança via Algoritmo EM / Victor Faria - 2011

44.p

1. Estimador de Máxima Verossimilhança 2. Algoritmo EM 3. Modelos Hierárquicos 4.Modelo t-Student. I.Título.

CDU N/A

Victor Basílio Faria

Estimação de Máxima Verossimilhança via Algoritmo EM

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Aprovado em 1° de Julho de 2011

BANCA EXAMINADORA

Clécio da Silva Ferreira

Doutor em Estatística - Universidade de São Paulo

Camila Borelli Zeller

Doutora em Estatística - Universidade Estadual de Campinas

Tufi Machado Soares

Pós-Doutor em Estatística - Universidade Federal do Rio de Janeiro

À minha família.

Resumo

O presente trabalho tem como objetivo a apresentação do algoritmo EM, que é uma ferramenta computacional utilizada para o cálculo do estimador de máxima verossimilhança (EMV) de forma iterativa, principalmente em problemas envolvendo dados incompletos. Para isso precisamos obter o conjunto dos dados completos, que é o conjunto dos dados observados aumentado com o conjunto dos dados faltantes e a partir daí obter a função log-verossimilhança associada aos dados completos.

Sabemos que este algoritmo seguramente converge para o EMV e tem como base a ideia de substituir uma difícil maximização por uma sequência de maximizações mais fáceis, envolvendo dois passos, o passo "E"(esperança) que calcula o valor esperado do logaritmo da verossimilhança completa; e o passo "M", que encontra seu máximo. Os passos são repetidos até se atingir a convergência.

Antes de exemplificarmos o cálculo do EMV via algoritmo EM, definiremos os modelos hierárquicos que possuem como vantagem modelar processos complicados através de uma sequência de modelos relativamente simples, colocados em uma hierarquia. Além disso, lidar com hierarquia não é mais difícil que lidar com distribuições condicionais ou marginais.

Como exemplo de cálculo do EMV via algoritmo EM, utilizaremos modelos de regressão, onde os erros assumem distribuição t-student, que além de serem simétricos, possuem caudas mais pesadas, sendo o processo de estimação mais robusto (no sentido de acomodar valores extremos). Apresentaremos também aplicações para essa distribuição.

Palavras-chave: Estimação de Máxima Verossimilhança, Algoritmo EM, Modelos Hierárquicos, Modelo t-Student.

Abstract

This paper aims to present the EM algorithm, which is a software tool used for calculating the maximum likelihood estimator (MLE) in an iterative manner, especially in problems involving missing data. For this we need to get the complete data set which is the set of observed data augmented with the set of missing data, from there get the log-likelihood function associated with complete data.

We know that this algorithm converges surely to MLE and is based on the idea of replacing a difficult maximization by a sequence of maximizations easier, involving two steps, the step "E"(Expectation) calculates the expected value of the complete log likelihood, and the step "M"(Maximization), which finds its fullest. The steps are repeated until convergence is achieved.

Before exemplify the calculation of MLE via EM algorithm, define the hierarchical models that have the advantage of modeling complicated processes through a sequence of relatively simple models, placed in a hierarchy. In addition, dealing with the hierarchy is no more difficult to deal with marginal or conditional distributions.

As an example of calculating MLE via EM algorithm, we use regression models, where the errors take on student-t distribution, which in addition to being symmetrical, have heavier tails (robust to accommodate extreme values). We will also present applications for these distribution.

Keywords: Maximum Likelihood Estimation, EM Algorithm, Hierarchical Models, Student's t-model.

Agradecimentos

Em primeiro lugar tenho que agradecer a Deus por sempre me guiar e me colocar no caminho certo. Agradeço a todos que sempre estiveram do meu lado desde que eu era um pequeno garoto fascinado por números, até hoje quando ainda o sou.

Em particular tenho que agradecer a algumas pessoas que contribuíram direta e indiretamente para o desenvolvimento e finalização deste trabalho:

À minha mãe e ao meu pai que sempre estiveram ao meu lado me apoiando e fazendo de tudo pra que eu pudesse chegar até aqui. Aos meus irmãos e irmã pelos conselhos e pelo apoio que sempre me deram.

À turma da faculdade, à Laura que sempre me ajudou e esteve bem perto de mim neste longo período.

Ao Thiago pela grande amizade que nos acompanha desde o ensino médio, ao Samuel, ao Thales e ao Lu pelas caronas concedidas. À Carol, à Priscila, ao Marcos, ao Jarbas, à Sarah, à Raquel, à Leiliane, ao Bruno, ao Luís Gustavo e ao Iago por me apoiarem, por me ajudarem e pelo companheirismo.

Aos professores do Departamento de Estatística pelos ensinamentos concedidos, em especial, ao professor Clécio pela excelente orientação neste trabalho e pelas ótimas aulas de inferência estatística ministradas. Ao professor Tufi e à professora Camila por participarem da banca de avaliação.

"O conhecimento de estatística é como o de uma língua estrangeira ou álgebra; ele poderá ser útil a qualquer tempo ou circunstância".

A. L. Bowley

Sumário

Lista de Figuras **7**

Lista de Tabelas **8**

1 Introdução **9**

2 Algoritmo EM e Modelos Hierárquicos **11**

2.1 Algoritmo EM 11

2.2 Modelos Hierárquicos 12

3 Modelo t-Student **19**

3.1 Modelo de Regressão Normal 19

3.2 Distribuição t-Student 19

3.3 Distribuição t-Student de Locação-Escala 21

3.4 Estimação de Máxima Verossimilhança via Algoritmo EM 21

3.5 Matriz de Informação Observada de Fisher 23

3.6 Aplicação em dados simulados 25

3.7 Análise de Resíduos em Modelos Simétricos 28

3.8 Aplicação em dados reais 29

4 Conclusão **35**

5 Apêndice - Derivadas Matriciais - Modelo t-Student **36**

Referências Bibliográficas **43**

Lista de Figuras

3.1 Gráfico da distribuição t-Student com diferentes valores para ν 20

3.2 Valores de $\ell(\widehat{\sigma^2}, \widehat{\nu})$ próximos (**platô**), associada a cada par $(\widehat{\sigma^2}, \widehat{\nu})$ e $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ constante. 28

3.3 Comportamento da Luminosidade dos produtos ao longo das semanas. . . 30

3.4 Gráfico (a) referente a suposição de homocedasticidade dos erros; Gráfico (b) referente a presença de observações aberrantes. 31

3.5 Gráficos normais de probabilidades com envelope para o resíduo do modelo definido em (3.16) - sob erros normais (a); sob erros t-Student com $\nu = 4,45$ (b). 32

3.6 Gráfico (a) referente a suposição de homocedasticidade dos erros; Gráfico (b) referente a presença de observações aberrantes. 33

3.7 Gráfico de $\widehat{u}_i \times$ número da observação, para os modelos assumindo erros normais e erros t-Student. 34

Lista de Tabelas

3.1 Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=200). 26

3.2 Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=500). 26

3.3 Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=1000). 27

3.4 Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=5000). 27

3.5 EMV dos parâmetros do modelo (3.16) ajustado aos dados de luminosidade sob erros normais, com e sem pontos aberrantes (obs. 7,8 e 9). 31

3.6 EMV dos parâmetros do modelo (3.16) ajustado aos dados de luminosidade sob erros t-Student, com e sem pontos aberrantes (obs. 7,8 e 9). 33

1 Introdução

A distribuição normal e os modelos de regressão sob a suposição de normalidade para os erros aleatórios são amplamente utilizados no contexto de variáveis aleatórias contínuas, havendo um grande desenvolvimento em diversas áreas da estatística. Sabemos, porém, que existem casos onde ambos não se aplicam, por exemplo quando a distribuição é sensível a presença de observações aberrantes (outliers). Surge então a necessidade de se desenvolver metologias robustas contra essas observações, neste sentido destacam-se os chamados modelos robustos (Cysneiros, Paula e Galea, 2005).

Na linha de modelos robustos, alternativas à suposição de erros aleatórios normais têm sido propostas na literatura. Uma dessas alternativas é assumir para os erros aleatórios, distribuições com caudas mais pesadas do que a normal, a fim de tentar reduzir a influência de pontos aberrantes nas estimativas dos coeficientes. Neste contexto, citamos Lange, Little e Taylor (1989) que propõem o modelo t-Student com ν graus de liberdade. Possíveis alternativas são o uso de outras distribuições simétricas, tais como logística, exponencial potência, normal contaminada e Cauchy.

Discutiremos com mais ênfase os modelos t-Student, porém ao se utilizar tais modelos torna-se necessário a obtenção de estimadores para esses, o que não é tão simples quanto para o modelo normal. A obtenção de soluções analíticas através do estimador de máxima verossimilhança (EMV) pode ser trabalhosa, sugerimos então que se adote uma solução iterativa visto que os recursos computacionais são cada vez mais utilizados.

Nesse sentido, o principal objetivo desta monografia é calcular de forma iterativa através do algoritmo EM (Dempster et al. 1977) estimadores e erros-padrão (através da matriz de informação observada de Fisher) referentes ao modelo t-Student, para que possam ser utilizados nos casos em que a suposição usual de normalidade não é atingida e este se adequa.

Estruturamos esta monografia da seguinte forma, no Capítulo 2, apresentamos o algoritmo EM, suas vantagens, porquê e como utilizá-lo, e sua estruturação, além de acharmos necessário a introdução dos chamados modelos hierárquicos através de exemplos práticos, pois julgamos que a compreensão destes facilitará no entendimento do

algoritmo EM. Em seguida, no Capítulo 3, descreveremos a obtenção do EMV via algoritmo EM, utilizando modelos de regressão, onde os erros aleatórios terão distribuição t-student. Neste Capítulo, através de simulações vamos validar o algoritmo EM proposto e mostraremos de forma empírica que este realmente fornece um EMV e logo após tem-se uma aplicação prática. No Capítulo 4, serão apresentadas as conclusões do trabalho e por fim o Capítulo 5 trata-se de um apêndice que contém toda a álgebra matricial utilizada no Capítulo 3.

2 Algoritmo EM e Modelos Hierárquicos

Neste capítulo, vamos introduzir o Algoritmo EM e mostrar que o conhecimento de modelos hierárquicos facilita a sua estruturação.

2.1 Algoritmo EM

O algoritmo EM (Dempster et al. 1977) é uma ferramenta computacional utilizada para o cálculo do estimador de máxima verossimilhança (EMV) de forma iterativa, e é principalmente utilizado em problemas envolvendo dados incompletos.

Há duas principais aplicações do algoritmo EM, primeiro quando os dados realmente são incompletos, devido a problemas ou limitação do processo de observação (*missings* ou dados faltantes) e segundo, quando a maximização da função de verossimilhança é analiticamente problemática, mas a função de probabilidade pode ser simplificada, admitindo a existência de valores adicionais (utilização de modelos hierárquicos). A segunda aplicação é mais comum e mais simples computacionalmente, o trabalho desenvolvido aqui é focado em tal situação.

Segundo Casella e Berger (2010), o EM é um algoritmo que seguramente converge para o EMV e tem como base a ideia de substituir uma difícil maximização da verossimilhança por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original. A demonstração original da convergência do EM realizada por Dempster, Laird e Rubin (1977) tinha uma falha, mas provas válidas de convergência foram apresentadas posteriormente por Boyles (1983) e Wu (1983).

Seja $\mathbf{y}$ o conjunto de dados observados e $\mathbf{s}$ denotando o conjunto de dados faltantes. O dado completo $\mathbf{y}_c = (\mathbf{y}, \mathbf{s})$ é $\mathbf{y}$ aumentado com $\mathbf{s}$ e sua função de densidade é $p(\mathbf{y}_c|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Denota-se por $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)$, a função log-verossimilhança dos dados completos e por $\ell(\boldsymbol{\theta}|\mathbf{y})$ a função log-verossimilhança dos dados observados. Segundo Zhu e Lee (2001), na maioria das aplicações estatísticas, a função log-verossimilhança dos dados completos geralmente tem forma mais simples que a log-verossimilhança dos dados observados. Cada iteração do algoritmo EM envolve dois passos, um passo E (esperança)

e um passo M (maximização), definidos como:

- Passo E: Calcule $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}, \boldsymbol{\theta}^{(k)}]$, onde a esperança é tomada com respeito a distribuição condicional $p(\mathbf{y}|\mathbf{s}, \boldsymbol{\theta}^{(k)})$.
- Passo M: Encontre $\boldsymbol{\theta}^{(k+1)}$ que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.

Resumindo, o "passo E"do algoritmo calcula o valor esperado do logaritmo da verossimilhança, e o "passo M"encontra seu máximo.

Estes passos devem ser repetidos até se atingir uma convergência, pode ser adotado como critério de parada, por exemplo $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| < \epsilon$, onde ϵ é um valor determinado maior que zero e $||a||$, denota a norma do vetor a.

O seguinte teorema, criado por Wu (1983), assegura a convergência para um *ponto estacionário*, que pode ser um máximo local ou ponto de sela.

Teorema 2.1.1. *Se o log da verossimilhança esperada para os dados completos* $E[\log L(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}, \widehat{\boldsymbol{\theta}}]$ *é contínua em* $\boldsymbol{\theta}$ *e em* $\widehat{\boldsymbol{\theta}}$*, então todos os pontos-limite de uma sequência do algoritmo EM* $(\widehat{\boldsymbol{\theta}}^{(k)})$ *são pontos estacionários de* $L(\boldsymbol{\theta}|\mathbf{y}_c)$ *e* $L(\widehat{\boldsymbol{\theta}}^{(k)}|\mathbf{y}_c)$ *converge monotonicamente para* $L(\widehat{\boldsymbol{\theta}}|\mathbf{y}_c)$ *considerando algum ponto estacionário* $\widehat{\boldsymbol{\theta}}$*.*

Segundo Zeller (2009) quando o "passo M"do algoritmo EM é complicado, este pode ser amenizado realizando o processo de maximização condicional a alguma função dos parâmetros que estão sendo estimados. Este algoritmo EM generalizado, proposto por Meng e Rubin (1993), é denominado algoritmo de maximação condicional de esperança (ECM). Neste trabalho, consideraremos apenas situações onde a maximização é obtida no próprio algoritmo EM.

2.2 Modelos Hierárquicos

Observamos que em geral uma variável aleatória tem uma única distribuição, porém frequentemente é mais fácil modelar uma situação considerando uma hierarquia. A vantagem da hierarquia é que processos complicados podem ser modelados por uma sequência de modelos relativamente mais simples, colocados em uma hierarquia, ideia parecida com uma estruturação do algoritmo EM. Além disso, lidar com uma hierarquia não é mais difícil do que lidar com distribuições marginais.

Definiremos de forma mais clara o conceito de Modelos Hieráquicos, através de alguns exemplos.

Definição 2.2.1. Uma variável aleatória X tem distribuição binomial se sua função de probabilidade (fp) é dada por

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \quad 0 < p < 1,$$

onde n é o número de tentativas independentes, cada uma com probabilidade se sucesso p e probabilidade de fracasso $(1-p)$ e X = número de sucessos ocorridos nas n tentativas. Denotamos $X \sim binomial(n, p)$, com $E(X) = np$ e $Var(X) = np(1-p)$.

Definição 2.2.2. Uma variável aleatória Y tem distribuição de Poisson se sua fp é dada por

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots,$$

para qualquer $\lambda > 0$. Denotamos $Y \sim Poisson(\lambda)$, com $E(Y) = Var(Y) = \lambda$.

Exemplo 2.2.1. Hierarquia binomial-Poisson

Um inseto põe um grande número de ovos, cada um deles com uma probabilidade de sobrevivência p. Na média, quantos ovos sobreviverão?

O "grande"número de ovos colocados é uma variável aleatória, geralmente assumindo que seja Poisson(λ). Além disso, se assumirmos que a sobrevivência de cada ovo é independente, então, temos provas de Bernoulli. Portanto, se considerarmos X = número de sobreviventes e Y = número de ovos que foram colocados, temos um modelo hierárquico

$$\begin{aligned} X|Y &\sim binomial(Y, p), \\ Y &\sim Poisson(\lambda). \end{aligned}$$

A variável aleatória de interesse, X = número de sobreviventes, tem a distribuição dada por

$$\begin{aligned} P(X = x) &= \sum_{y=0}^{\infty} P(X = x, Y = y) \\ &= \sum_{y=0}^{\infty} P(X = x|Y = y)P(Y = y) \\ &= \sum_{y=x}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right] \left[\frac{e^{-\lambda}\lambda^y}{y!}\right], \end{aligned}$$

uma vez que $X|Y = y$ é binomial(y, p) e Y é Poisson(λ). Se agora simplificarmos esta última expressão e multiplicarmos por λ^x/λ^x, obtemos

$$\begin{aligned} P(X = x) &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{(t)!}, \quad (t = y - x) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p}, \end{aligned}$$

assim $X \sim$ Poisson(λp). Portanto, qualquer inferência marginal em X diz respeito a uma distribuição de Poisson(λp), e Y não representa nenhuma parte. A introdução de Y na hierarquia foi principalmente para ajudar o entendimento do modelo. Existe uma vantagem a mais ao se considerar que o parâmetro da distribuição de X é o produto de dois parâmetros, sendo cada um deles relativamente simples de se entender. Portanto, a resposta a pergunta original pode ser calculada através de

$$E(X) = \lambda p,$$

então, na média, λp ovos sobreviverão.

Algumas vezes, os cálculos podem ser ainda mais simplificados se usarmos propriedades de esperanças condicionais, como o teorema seguinte

Teorema 2.2.1. *Se X e Y forem duas variáveis aleatórias quaisquer, então*

$$E(X) = E(E(X|Y)),$$

desde que os valores esperados existam.

Prova 2.2.1. Seja $f(x, y)$ denotando a função densidade de probabilidade (fdp) conjunta de X e Y. Por definição, temos

$$E(X) = \int\int x f(x, y) dx dy = \int \left[\underbrace{\int x f(x|y) dx}_{} \right] f_Y(y) dy, \tag{2.1}$$

onde $f(x|y)$ e $f_Y(y)$ são a fdp condicional de X dado que $Y = y$ e a fdp marginal de Y, respectivamente. Mas agora observe que a integral interna em (2.1) é a esperança condicional $E(X|y)$, e temos

$$E(X) = \int E(X|y) f_Y(y) dy = E(E(X|Y)).$$

Se voltarmos ao Exemplo 2.2.1 , podemos facilmente calcular o número esperado de sobreviventes, através do teorema 2.2.1

$$\begin{aligned} E(X) &= E(E(X|Y)) \\ &= E(pY) \\ &= pE(Y) \\ &= \lambda p. \end{aligned}$$

Definição 2.2.3. Uma variável aleatória X tem distribuição Gama com parâmetros α e β, denotada por $Gama(\alpha, \beta)$, se sua fdp é dada por

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

onde $\Gamma(.)$ é a função Gama; para α e $\beta > 0$. Com $E(X) = \alpha/\beta$ e $Var(X) = \alpha/\beta^2$.

Se $X \sim Gama(1, \beta)$, ela é dita exponencial com parâmetro β, denotada por $exp(\beta)$.

Definição 2.2.4. Uma variável aleatória Y tem distribuição Binomial Negativa com parâmetros r e p, denotada por $BN(r, p)$, se sua fp é dada por

$$P(X = x) = \binom{x + r - 1}{x} p^r (1 - p)^x \quad x > 0,$$

com $E(X) = \frac{r(1-p)}{p}$ e $Var(X) = \frac{r(1-p)}{p^2}$.

Se $X \sim BN(1, p)$, ela é dita geométrica com parâmetro p, denotada por $Geo(p)$.

O Exemplo 2.2.1, mostra uma hierarquia em dois estágios, porém podem haver casos que existem vantagens na modelagem de um fenômeno através de uma hierarquia com mais de dois estágios, desde que este seja mais fácil de ser entendido.

Exemplo 2.2.2. Podemos generalizar o Exemplo 2.2.1, em que, em vez de um inseto-mãe, existe um grande número de mães, e uma delas é escolhida aleatoriamente. Ainda estamos interessados em conhecer o número médio de sobreviventes, mas agora não está mais claro que o número de ovos que foram colocados segue a mesma distribuição de Poisson para cada mãe. Temos então a seguinte hierarquia de três estágios que pode ser mais apropriada. Seja X = número de sobreviventes em uma ninhada, então X pode ser

expresso com o seguinte modelo hierárquico

$$\begin{aligned} X|Y &\sim binomial(Y,p), \\ Y|\Lambda &\sim Poisson(\Lambda), \\ \Lambda &\sim exp(1/\beta), \end{aligned}$$

onde o último estágio da hierarquia corresponde à variabilidade em relação a diferentes mães.

A média de X pode ser calculada como

$$\begin{aligned} E(X) &= E(E(X|Y)) \\ &= E(pY) \\ &= pE(E(Y|\Lambda)) \\ &= pE(\Lambda) \\ &= p\beta. \end{aligned} \tag{2.2}$$

Neste exemplo, utilizamos um tipo de modelo um pouco diferente do anterior, pois temos duas variáveis aleatórias discretas e uma contínua. Segundo Casella e Berger (2010), a utilização desses modelos não deve apresentar problemas. Podemos definir uma densidade conjunta, $f(x,y,\lambda)$; densidades condicionais, $f(x|y)$, $f(x|y,\lambda)$, etc.; e densidades marginais, $f(x)$, $f(x,y)$, etc., do mesmo modo. Basta notar que quando probabilidades ou esperanças são calculadas, variáveis discretas são somadas e variáveis contínuas são integradas.

O modelo de três estágios apresentado acima, também pode ser reescrito como uma hierarquia de dois estágios, combinando os dois últimos estágios.

Se $Y|\Lambda \sim Poisson(\Lambda)$ e $\Lambda \sim exp(1/\beta)$, então

$$\begin{aligned}
P(X=y) &= \int_0^\infty f(y,\lambda)d\lambda \\
&= \int_0^\infty f(y|\lambda)f(\lambda)d\lambda \\
&= \int_0^\infty \left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\frac{1}{\beta}e^{-\lambda/\beta}d\lambda \\
&= \frac{1}{\beta y!}\int_0^\infty \lambda^y e^{-\lambda(1+1/\beta)}d\lambda \\
&\quad \text{*núcleo da fdp}-Gama(y+1,1+1/\beta)) \\
&= \frac{1}{\beta y!}\int_0^\infty \overbrace{\lambda^{(y+1)-1}e^{-\lambda(1+1/\beta)}}d\lambda \\
&= \frac{1}{\beta y!}\Gamma(y+1)\left(\frac{1}{1+1/\beta}\right)^{y+1} \\
&= \frac{1}{(1+\beta)}\left(\frac{1}{1+1/\beta}\right)^{y} \\
&= \frac{1}{(1+\beta)}\left(\frac{\beta}{1+\beta}\right)^{y}. \qquad (2.3)
\end{aligned}$$

Portanto a função em (2.3) tem fp binomial negativa com parâmetros $r=1$ e $\theta = \frac{1}{1+\beta}$. E a hierarquia descrita em três estágios no Exemplo 2.2.2, é equivalente a hierarquia de dois estágios

$$\begin{aligned}
X|Y &\sim binomial(Y,p), \\
Y &\sim BN\left(r=1, p=\frac{1}{1+\beta}\right).
\end{aligned}$$

Vale ressaltar, que em termos de modelagem do problema, o modelo hierárquico de três estágios é de mais fácil entendimento.

Por fim, temos um último exemplo de modelo hierárquico.

Definição 2.2.5. Uma variável aleatória X tem distribuição Beta com parâmetros α e β, denotada por $Beta(\alpha,\beta)$, se sua fdp é dada por

$$f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad 0<x<1,$$

onde $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ é a função beta; para $\alpha>0$ e $\beta>0$, com $E(X) = \alpha/(\alpha+\beta)$ e $Var(X) = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

Se $\alpha=\beta=1$, $X \sim U(0,1)$.

Exemplo 2.2.3. Hierarquia beta-binomial

Uma generalização da distribuição binomial é permitir que a probabilidade de sucesso varie de acordo com uma distribuição. Um modelo padrão para este caso é

$$\begin{aligned} X|P &\sim binomial(n, P), \quad i = 1, \ldots, n, \\ P &\sim beta(\alpha, \beta). \end{aligned}$$

Para calcular a média de X, novamente usamos o teorema 2.2.1

$$E(X) = E(E(X|P)) = E(nP) = n\frac{\alpha}{\alpha + \beta}.$$

3 Modelo t-Student

No estudo da estatística, mais precisamente no contexto de análise de dados, a distribuição normal e os modelos de regressão normais são de grande ajuda e por isso utilizados amplamente, sabemos porém que existem casos onde ambos não se aplicam, por exemplo quando a distribuição possui caudas pesadas (pontos extremos em ambas as caudas da distribuição), que podem gerar formatos que não permitem a suposição de normalidade. Para que se consiga modelar tal situação sugerimos como alternativa a utilização de modelos robustos, em especial, a distribuição t-Student.

Os *softwares* utilizados nas aplicações deste capítulo foram o R versão 2.12.2 e o MatLab versão R2009b. Os resultados matriciais apresentados neste capítulo, possuem cálculos anexados no Capítulo 5 (Apêndice).

3.1 Modelo de Regressão Normal

Para facilitar uma posterior comparação entre modelos, vamos definir o modelo de regressão assumindo para os erros aleatórios distribuição normal, da seguinte forma:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim NM_n(\mathbf{0}; \sigma^2\mathbf{I}_n) \end{aligned} \tag{3.1}$$

onde $\mathbf{X}$ é a matriz do modelo de dimensão $n \times p$, $\boldsymbol{\beta}$ é o vetor coluna dos parâmetros $p \times 1$, $\mathbf{0}$ é o vetor nulo n e $\mathbf{I}_n$ é a matriz identidade $n \times n$ e NM_n denota a distribuição normal multivariada de dimensão n.

3.2 Distribuição t-Student

A distribuição t-Student é uma distribuição de probabilidade estatística definida por um único parâmetro ν (graus de liberdade) que define sua forma, e assim como a normal padrão, é simétrica, porém possuindo caudas mais pesadas (robusta no sentido de acomodar valores extremos). Temos que uma variável aleatória T tem distribuição

t-Student, denotada por t_ν, se

$$T = \frac{Z}{\sqrt{V/\nu}},$$

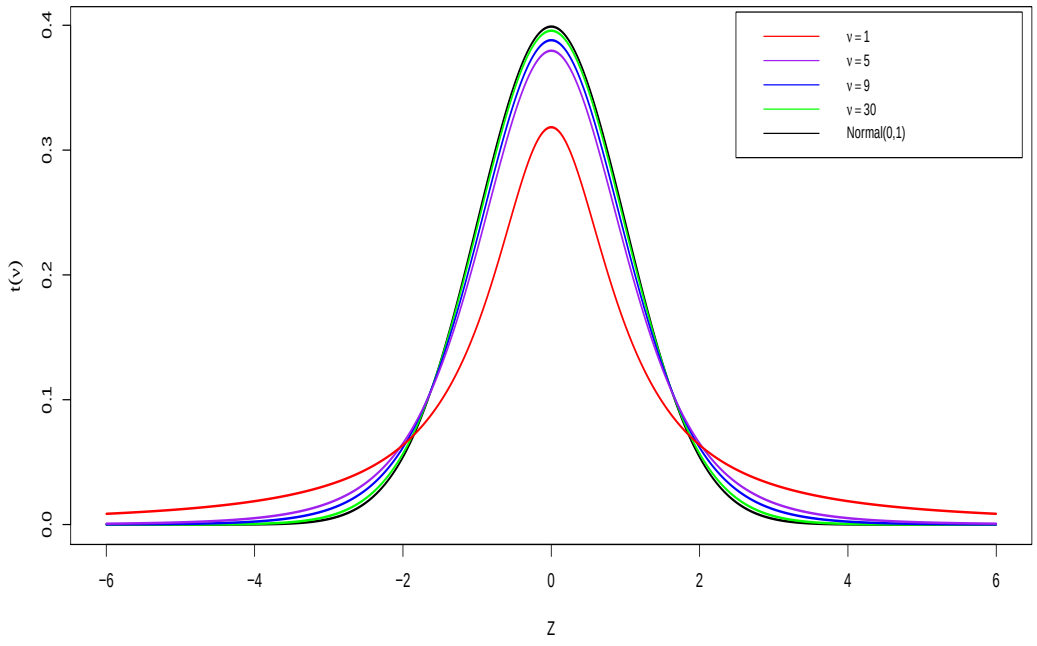
onde Z tem distribuição normal padrão e V tem distribuição Chi-quadrado com ν graus de liberdade, com Z e V independentes. Sua fdp é dada por

$$f_T(t) = c(\nu)\left(1 + \frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}, \quad \nu > 0, \tag{3.2}$$

onde $c(\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}$.

O seu valor esperado e sua variância, são respectivamente

Figura 3.1: Gráfico da distribuição t-Student com diferentes valores para ν



$$\begin{aligned} E(T) &= 0, \\ Var(T) &= \frac{\nu}{\nu-2}, \quad \nu > 2. \end{aligned}$$

Para valores de $\nu \leq 2$ a $Var(T)$ não está definida. Notamos que

$$\lim_{\nu\to\infty} Var(T) = \lim_{\nu\to\infty} \frac{\nu}{\nu-2} = 1 \tag{3.3}$$

O limite (3.3) mostra que a variância de uma variável aleatória $T \sim t_\nu$, tende a 1 quando ν cresce, o que é esperado, pois quando $\nu \to \infty$, $T \sim N(0,1)$.

3.3 Distribuição t-Student de Locação-Escala

O modelo em (3.2) pode ser escrito adicionando parâmetros de locação $\mu \in \mathbb{R}$ e de escala $\sigma > 0$, denotado por $Y \sim t_\nu(\mu, \sigma^2)$.

A distribuição t-Student com parâmetro de locação μ e de escala σ de uma variável aleatória Y é dada por

$$f_Y(y) = \frac{\mathrm{c}(\nu)}{\sigma}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{(\nu+1)}{2}}, \quad y \in \mathbb{R}. \tag{3.4}$$

Verificamos que se $T \sim t_\nu$ e $Y = \mu + \sigma T$, então $Y \sim t_\nu(\mu, \sigma^2)$ e o seu valor esperado e sua variância são dados, respectivamente, por

$$\begin{aligned} E(Y) &= \mu, \\ Var(Y) &= \sigma^2 \frac{\nu}{\nu-2}, \quad \nu > 2. \end{aligned}$$

3.4 Estimação de Máxima Verossimilhança via Algoritmo EM

Suponha que se tenha um conjunto de n observações independentes, denotadas por $Y_1, \ldots, Y_n,$ onde $Y_i \sim t_\nu(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. Associado com a observação i, considere um vetor $p \times 1$ de covariáveis $\mathbf{x}_i$, através do qual especifica-se o preditor linear $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, onde $\boldsymbol{\beta}$ é um vetor p-dimensional de coeficientes de regressão desconhecidos. Sob essas condições a função de log-verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ é dada por

$$\ell(\boldsymbol{\theta}) = n \log \mathrm{c}(\nu) - \frac{n}{2}\log\sigma^2 - \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{n} \log\left[1 + \frac{(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]. \tag{3.5}$$

A maximização analítica da função acima pode ser complicada, já que envolve derivada de um somatório-*log*. A saída discutida é realizar tal maximização através do algoritmo EM. Porém, vale ressaltar que não é possível a obtenção de um estimador analítico (uma equação) para o parâmetro ν através deste método. Sendo assim, adotaremos uma estimação perfilada (simultânea com o EM) maximizando diretamente a função (3.5), com $\boldsymbol{\theta}$ atualizado pelo EM.

Segundo Ferreira (2008), através de uma propriedade conhecida de uma variável aleatória t-Student, tem-se a seguinte representação hierárquica

$$\begin{aligned} Y_i|U=u_i &\sim N\left(\mathbf{x}_i^\top\boldsymbol{\beta}, \frac{\sigma^2}{u_i}\right), \\ U_i &\sim Gama(\nu/2,\nu/2)(u_i), \quad i,\ldots,n. \end{aligned} \tag{3.6}$$

e, temos que

$$\begin{aligned} f(\mathbf{y}) &= \prod_{i=1}^{n} t_\nu(y_i|\mathbf{x}_i^\top\boldsymbol{\beta},\sigma^2) \\ &= \prod_{i=1}^{n}\int_0^{+\infty}\phi(y_i|\mathbf{x}_i^\top\boldsymbol{\beta},\sigma^2/u_i)Gama(\nu/2,\nu/2)(u_i)du_i. \end{aligned}$$

Como $f(u_i)$ é uma função constante que depende apenas de ν e sendo $\mathbf{y} = (y_1,\ldots,y_n)^\top$ e $\mathbf{u} = (u_1,\ldots,u_n)^\top$, $\mathbf{u}$ tratado como dado faltante, temos que a função log-verossimilhança completa associada com $\mathbf{y}_c = (\mathbf{y}^\top,\mathbf{u}^\top)^\top$ é dada por

$$\begin{aligned} \ell_c(\boldsymbol{\theta}|\mathbf{y}_c) &= \sum_{i=1}^{n}\log f(y_i,u_i) \\ &\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}u_i(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{u})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \end{aligned} \tag{3.7}$$

onde $\mathbf{X} = (\mathbf{x_1},\ldots,\mathbf{x_n})^\top$ é a matriz do modelo de dimensão $n\times p$ e $\mathbf{D}(\mathbf{u}) = \text{Diag}(u_1,\ldots,u_n)$.

Proposição 3.4.1. *$U_i|y_i \sim Gama(\frac{\nu+1}{2}, \frac{\nu+d_i^2}{2})$, onde $d_i = \frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})}{\sigma}$.*

Prova 3.4.1.

$$f(u_i,y_i) = \frac{u_i^{\frac{1}{2}}}{\sqrt{2\pi}\sigma}e^{-\frac{u_i}{2\sigma^2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2} \times \frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}u_i^{\frac{\nu}{2}-1}e^{\frac{\nu u_i}{2}}.$$

Se olharmos a função acima apenas como função de u_i, e y_i como uma constante, temos que

$$\begin{aligned} f(u_i|y_i) &\propto u_i^{\frac{1}{2}}e^{-\frac{u_i}{2\sigma^2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}u_i^{\frac{\nu}{2}-1}e^{\frac{\nu u_i}{2}} \\ &= u_i^{\frac{\nu+1}{2}-1}e^{-\frac{1}{2}(\nu+d_i^2)u_i}. \end{aligned}$$

Logo $U_i|y_i \sim Gama(\frac{\nu+1}{2}, \frac{\nu+d_i^2}{2})$.

Como $\widehat{u}_i = \text{E}[U_i|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, y_i]$, tem-se que

$$\widehat{u}_i = \frac{(\nu+1)}{(\nu+d_i^2)}. \tag{3.8}$$

Denotamos por $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)^\top}, \sigma^{2(k)})^\top$ a estimativa de $\boldsymbol{\theta}$ para a k-ésima iteração. Segue que a esperança com respeito a $\mathbf{u}$, condicionada em $\mathbf{y}$, da função log-verossimilhança completa (Passo E), tem a forma

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= \mathrm{E}[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}] \\ &= -\frac{n}{2}\log\sigma^{2(k)} - \frac{1}{2\sigma^{2(k)}}\sum_{i=1}^{n}\widehat{u}_i^{(k)}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})^2 \\ &= -\frac{n}{2}\log\sigma^{2(k)} - \frac{1}{2\sigma^{2(k)}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top\mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)}). \end{aligned} \tag{3.9}$$

Temos, então o seguinte algoritmo EM:
Passo E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, calcule $\widehat{u}_i^{(k)}$, para $i, \ldots, n$, usando (3.8).
Passo M: Atualize $\widehat{\boldsymbol{\theta}}^{(k+1)}$ maximizando $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ em $\boldsymbol{\theta}$, que leva às seguintes soluções analíticas:

$$\widehat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^\top\mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{y} \tag{3.10}$$

e,

$$\widehat{\sigma^2}^{(k)} = \frac{1}{n}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top\mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})]. \tag{3.11}$$

Percebemos que se $u_i = 1$, para todo $i = 1, \ldots, n$, temos em (3.6) um modelo normal com parâmetros $\mathbf{x}_i^\top\boldsymbol{\beta}$ e σ^2 e ainda que $\mathbf{D}(\widehat{\mathbf{u}})$ se torna uma matriz identidade de ordem n, resultando em (3.10), $\widehat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, que é o EMV de $\boldsymbol{\beta}$ para tal modelo. Note que em (3.11) temos, $\widehat{\sigma^2}^{(k)} = \frac{1}{n}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})] = \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(k)})^2}{n}$, que também coincide com o EMV de σ^2 do modelo normal. Lembrando que um estimador para ν pode ser obtido de forma perfilada, ou seja, para cada etapa do algoritmo EM, encontramos $\hat{\nu}$ maximizando a função de verossimilhança (3.5) em ν, avaliada em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(k)}$ e $\sigma^2 = \widehat{\sigma^2}^{(k)}$.

3.5 Matriz de Informação Observada de Fisher

Sabemos que a Matriz de Informação Esperada de Fisher, denotada por $\mathbf{I}(\boldsymbol{\theta})_F$, é igual a $-E[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}]$ possui cálculos trabalhosos, porque não envolve somente derivadas, mas também esperanças relacionadas a estas. Uma opção é utilizar a Matriz de Informação Observada de Fisher denotada por $\mathbf{I}(\boldsymbol{\theta})_{Fobs}$, que é igual a $-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}$. Utilizando a propriedade

da invariância dos estimadores de máxima verossimilhança, $\mathbf{I}(\boldsymbol{\theta})_{Fobs}$ pode ser considerada como um EMV de $\mathbf{I}(\boldsymbol{\theta})_F$, denotado por $\widehat{\mathbf{I}(\boldsymbol{\theta})_F}$, ver (Ferreira, 2008).

O vetor de derivadas parciais (gradiente) de $\ell(\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$, conhecido como função escore é definido por

$$U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (U(\boldsymbol{\beta})^\top, U(\sigma^2), U(\nu))^\top, \tag{3.12}$$

que tem elementos

$$\begin{aligned}
U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{(\nu+1)}{\nu\sigma^2}\mathbf{X}^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\\
U(\sigma^2) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(\nu+1)}{2\nu\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\\
U(\nu) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \nu} &= n\frac{\partial}{\partial\nu}[\log c(\nu)] - \frac{1}{2}\sum_{i=1}^{n}\log d_i^{-1} + \frac{(\nu+1)}{2\nu^2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),
\end{aligned}$$

onde $\mathbf{D}(\mathbf{d}) = \text{diag}(d1, \ldots, d_n)$, com $d_i = \left[1 + \frac{(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}$;
$\frac{\partial}{\partial\nu}[\log c(\nu)] = \frac{1}{2}[\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right) - \left(\frac{1}{\nu}\right)]$;
Ψ é a função digama com $\Psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$.

Definiremos $\mathbf{I}(\boldsymbol{\theta})_F = -\mathbf{I}(\boldsymbol{\theta})_{Fobs}$, neste caso

$$\mathbf{I}(\boldsymbol{\theta})_F = \begin{pmatrix} I_{\boldsymbol{\beta}\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\sigma^2} & I_{\boldsymbol{\beta}\nu} \\ & I_{\sigma^2\sigma^2} & I_{\sigma^2\nu} \\ & & I_{\nu\nu} \end{pmatrix}, \tag{3.13}$$

com elementos

$$\begin{aligned}
I_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^\top\partial\boldsymbol{\beta}} &= \frac{(\nu+1)}{\nu\sigma^2}\mathbf{X}^\top\left[\frac{2}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right]\mathbf{X},\\
I_{\boldsymbol{\beta}\sigma^2} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\sigma^2\partial\boldsymbol{\beta}} &= \frac{(\nu+1)}{\nu\sigma^4}\mathbf{X}^\top\left[\frac{1}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\\
I_{\boldsymbol{\beta}\nu} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\nu\partial\boldsymbol{\beta}} &= \frac{1}{\nu^2\sigma^2}\mathbf{X}^\top\left[\frac{(\nu+1)}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\\
I_{\sigma^2\sigma^2} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\sigma^2\partial\sigma^2} &= \frac{(\nu+1)}{\nu\sigma^6}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{1}{2\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\
&\quad + \frac{n}{2\sigma^4},\\
I_{\sigma^2\nu} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\nu\partial\sigma^2} &= \frac{1}{2\nu^2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{(\nu+1)}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\\
I_{\nu\nu} = \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\nu\partial\nu} &= \frac{1}{\nu^3\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{(\nu+1)}{2\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\
&\quad + n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right],
\end{aligned}$$

onde $\frac{\partial^2}{\partial\nu\partial\nu}[\log c(\nu)] = \frac{1}{4}[\Psi'\left(\frac{\nu+1}{2}\right) - \Psi'\left(\frac{\nu}{2}\right) - \left(\frac{2}{\nu^2}\right)]$;
Ψ' é a função trigama com $\Psi'(x) = \frac{\partial^2}{\partial x\partial x}\log\Gamma(x)$.

3.6 Aplicação em dados simulados

Para validar o algoritmo EM proposto na seção 3.3, geramos $M = 200$ amostras de tamanhos n_i (200, 500, 1000 e 5000), através de regressões lineares da forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, onde $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ e $\epsilon_i \sim t_\nu(0, \sigma^2)$: $n \times 1$, gerados para cada amostra, $\boldsymbol{\beta}$ é o vetor coluna de parâmetros: $p \times 1$ e $\mathbf{X}$ é a matriz regressora: $n \times p$ (única para todas as M amostras).

Definimos $\boldsymbol{\beta} = (5, 2)^\top$, $\nu = 10$, $\sigma^2 = 1$ e $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ onde $\mathbf{1} = (1, 1, \ldots, 1)^\top$ e $\mathbf{x} : n \times 1$ foi gerado através de uma $U(0, 10)$, ambos com tamanho n. Desta forma, o modelo simulado foi:

$$y_i = 5 + 2x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

A partir do algoritmo EM, estimamos $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ para cada amostra utilizando (3.10) e (3.11) , obtendo o estimador de ν de forma perfilada. Guardando-se os erros-padrão de $\widehat{\boldsymbol{\theta}}(\widehat{EP(\widehat{\boldsymbol{\theta}})} = \sqrt{\mathbf{D}(\mathbf{I}^{-1}_{Fobs}(\widehat{\boldsymbol{\theta}}))})$, tem-se então uma matriz de estimadores $\widehat{\boldsymbol{\theta}}_k$ com seus respectivos $\widehat{EP(\widehat{\boldsymbol{\theta}}_k)}$, $k = 1, \ldots, M$ (para cada n), da forma

$$\begin{pmatrix} \widehat{\beta}_{0(1)} & \widehat{\beta}_{1(1)} & \widehat{\sigma^2}_{(1)} & \widehat{\nu}_{(1)} & \widehat{EP(\widehat{\beta}_{0(1)})} & \widehat{EP(\widehat{\beta}_{1(1)})} & \widehat{EP(\widehat{\sigma^2}_{(1)})} & \widehat{EP(\widehat{\nu}_{(1)})} \\ \widehat{\beta}_{0(2)} & \widehat{\beta}_{1(2)} & \widehat{\sigma^2}_{(2)} & \widehat{\nu}_{(2)} & \widehat{EP(\widehat{\beta}_{0(2)})} & \widehat{EP(\widehat{\beta}_{1(2)})} & \widehat{EP(\widehat{\sigma^2}_{(2)})} & \widehat{EP(\widehat{\nu}_{(2)})} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\beta}_{0(k)} & \widehat{\beta}_{1(k)} & \widehat{\sigma^2}_{(k)} & \widehat{\nu}_{(k)} & \widehat{EP(\widehat{\beta}_{0(k)})} & \widehat{EP(\widehat{\beta}_{1(k)})} & \widehat{EP(\widehat{\sigma^2}_{(k)})} & \widehat{EP(\widehat{\nu}_{(k)})} \end{pmatrix}. \quad (3.14)$$

Como o algoritmo EM calcula o EMV, esperamos que seja assintoticamente não viciado ($\lim_{n\to\infty} E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$) e assintoticamente eficiente ($Var(\widehat{\boldsymbol{\theta}})$ tende a zero a medida que n cresce), logo seria um estimador consistente $(\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta})$.

Na Tabela 3.1, para amostras de tamanho 200, observamos que as médias das estimativas dos parâmetros β_0, β_1 e σ^2 estão próximas de seus verdadeiros valores, ou seja, assintoticamente não viciados e a média de seus erros-padrão se aproximam do desvio-padrão amostral de cada estimativa. A estimativa de ν, obtida de forma simultânea com o algoritmo EM, apresentou valor inferior ao de seu parâmetro e a média dos erros-padrão não se aproximou tão bem de seu desvio-padrão amostral. Porém, esperamos que, com o aumento de n, essas aproximações melhorem.

Na Tabela 3.2, para amostras de tamanho 500, as análises para β_0, β_1 e σ^2 continuam as mesmas e vale ressaltar que a média de seus erros-padrão estão diminuindo.

Tabela 3.1: Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=200).

$\boldsymbol{\theta}_0$	Valor Real	Média($\widehat{\boldsymbol{\theta}}_{kj}$)	desvio-padrão($\widehat{\boldsymbol{\theta}}_j$)	Média($EP(\widehat{\boldsymbol{\theta}}_{kj})$)
β_0	5,0000	4,9966	0,1717	0,1705
β_1	2,0000	2,0011	0,0304	0,0302
σ^2	1,0000	0,8987	0,1634	0,1690
ν	10,0000	3,6696	1,3382	1,1836

A estimativa de ν continua inferior ao de seu parâmetro, mas a média dos erros-padrão está se aproximando do desvio-padrão amostral.

Tabela 3.2: Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=500).

$\boldsymbol{\theta}_0$	Valor Real	Média($\widehat{\boldsymbol{\theta}}_{kj}$)	desvio-padrão($\widehat{\boldsymbol{\theta}}_j$)	Média($EP(\widehat{\boldsymbol{\theta}}_{kj})$)
β_0	5,0000	4,9993	0,1013	0,1067
β_1	2,0000	2,0007	0,0172	0,0176
σ^2	1,0000	0,8986	0,1111	0,1060
ν	10,0000	3,4411	0,5737	0,6010

Na Tabela 3.3, para amostras de tamanho 1000, também provamos que β_0, β_1 e σ^2 são assintoticamente não viciados e que a média de seus erros-padrão, realmente tendem a zero a medida que n aumenta e neste último caso média dos erros-padrão de ν se aproximou de seu desvio-padrão amostral, lembrando que a estimativa de ν, continua inferior ao valor do parâmetro.

A Tabela 3.4, possui análises parecidas com as demais, e esta nos serviu para confirmar que a média dos erros-padrão de todas as estimativas estão realmente tendendo a zero.

Portanto provamos de forma empírica que o algoritmo EM proposto realmente é um EMV.

A estimativa do parâmetro ν ser inferior nos quatro tamanhos amostrais não foi

Tabela 3.3: Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=1000).

$\boldsymbol{\theta}_0$	Valor Real	Média($\widehat{\boldsymbol{\theta}}_{kj}$)	desvio-padrão($\widehat{\boldsymbol{\theta}}_j$)	Média($EP(\widehat{\boldsymbol{\theta}}_{kj})$)
β_0	5,0000	5,0022	0,0653	0,0627
β_1	2,0000	1,9999	0,0114	0,0125
σ^2	1,0000	0,9068	0,0785	0,0749
ν	10,0000	3,4241	0,4104	0,4135

Tabela 3.4: Modelo de regressão t-student com dados simulados: Média das estimativas dos parâmetros $\boldsymbol{\theta}_k$, com seus respectivos desvios-padrão e média dos erros-padrão, (M=200 amostras de tamanho n=5000).

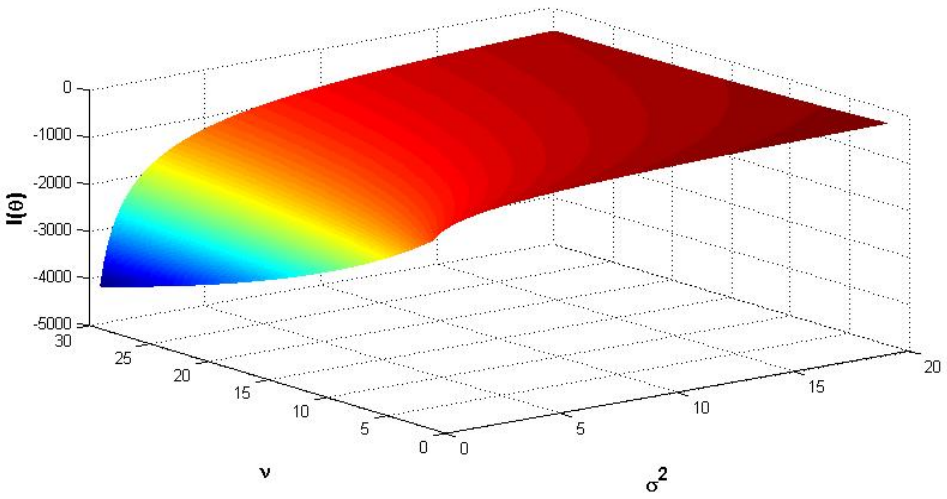
$\boldsymbol{\theta}_0$	Valor Real	Média($\widehat{\boldsymbol{\theta}}_{kj}$)	desvio-padrão($\widehat{\boldsymbol{\theta}}_j$)	Média($EP(\widehat{\boldsymbol{\theta}}_{kj})$)
β_0	5,0000	4,9987	0,0331	0,0328
β_1	2,0000	2,0006	0,0060	0,0056
σ^2	1,0000	0,9051	0,0336	0,0335
ν	10,0000	3,3743	0,1596	0,1585

considerado como um problema e pode ser explicado da seguinte forma: se considerarmos $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ constante, para cada par $(\widehat{\sigma^2}, \widehat{\nu})$ temos uma $\ell(\widehat{\sigma^2}, \widehat{\nu})$ associada. O que pode acontecer no caso da simulação é que para diferentes pares $(\widehat{\sigma^2}, \widehat{\nu})$ os valores de $\ell(\widehat{\sigma^2}, \widehat{\nu})$ são praticamente iguais, criando uma região (**platô**), onde essa não possui um ponto bem definido. Então o algoritmo EM "patina"nesses valores, podendo gerar uma estimativa exata para o parâmetro que foi estimado de forma perfilada, no caso ν. O fato é que a função do algoritmo EM é encontrar o valor estimado para cada parâmetro que maximiza a verossimilhança, e isto está sendo feito. Em situações práticas, não conhecemos os valores dos parâmetros para fazer a comparação com suas estimativas, nos baseamos apenas no valor da verossimilhança maximizada.

Podemos notar a existência de platôs através da Figura 3.2.

A grande região em vermelho, na Figura 3.2, indica os pontos que possuem maiores verossimilhanças, notamos que para diversos pares de $(\widehat{\sigma^2}, \widehat{\nu})$ essas são pratica-

Figura 3.2: Valores de $\ell(\widehat{\sigma^2}, \widehat{\nu})$ próximos (**platô**), associada a cada par $(\widehat{\sigma^2}, \widehat{\nu})$ e $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ constante.



mente iguais, portanto não existindo um ponto bem definido para maximizá-la. Uma sugestão para um próximo trabalho, seria mudar o critério de parada utilizado, aqui usamos $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| < \epsilon$ que poderia ser substituído por $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})|| < \epsilon$, pois este critério possivelmente evitaria a ocorrência do platô.

3.7 Análise de Resíduos em Modelos Simétricos

Após o ajuste do modelo proposto, nos interessa saber se este se ajusta bem aos dados, bem como saber a confiabilidade dos testes estatísticos e se existem evidências sobre possíveis violações nas suposições usuais do modelo (distribuição dos resíduos e homocedasticidade). Uma técnica que nos ajuda nessas questões é a análise dos resíduos.

A definição mais usual de resíduo é dada por $r_i = y_i - \widehat{\mu}_i$ (resíduo ordinário). Esses resíduos são, em geral, viesados e têm distribuição não normal, mesmo assintoticamente, dificultando a verificação da adequacidade dos modelos pelos métodos tradicionais.

Cysneiros, Paula e Galea (2005) definem um resíduo padronizado t_{ri} para modelos simétricos, dado por

$$t_{ri} = \frac{y_i - \widehat{y_i}}{(\xi\widehat{\sigma^2})(1 - (4d_g\xi)^{-1}\widehat{h_{ii}})^{1/2}}, \quad i = 1, \ldots, n, \tag{3.15}$$

onde $\xi = \frac{\nu}{\nu-2}$, $d_g = \frac{\nu+1}{4(\nu+2)}$, h_{ii} são os elementos da diagonal principal da matriz de projeção ortogonal de vetores do $\Re^n$ no subespaço gerado pelas colunas da matriz $\mathbf{X}$,

dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ e $\widehat{y}_i$ denota cada valor predito. Algumas propriedades da matriz $\mathbf{H}$ se encontram em Paula (2004).

Estudos desenvolvidos pelos autores acima indicam que o resíduo t_{ri} tem aproximadamente uma distribuição $N(0,1)$. Podemos assim gerar bandas empíricas de confiança através do modelo ajustado para t_{ri}, também conhecidos como envelope simulado (Atkison, 1981, 1985). Tais bandas podem auxiliar na qualidade do ajuste do modelo. Para detectarmos observações aberrantes e se existe homogeneidade de variância podemos utilizar o gráfico de $t_{ri} \times \widehat{y}_i$.

3.8 Aplicação em dados reais

Estudo da Luminosidade de um novo produto alimentício

Foi desenvolvido no Departamento de Nutrição da Faculdade de Saúde Pública da Universidade de São Paulo um produto do tipo "snack", que possui baixo teor de gordura saturada e ácidos graxos, em que substituiu-se, totalmente ou parcialmente, gordura vegetal hidrogenada por óleo de canola. Há interesse em comparar 5 novas formas desses novo produto: A (22% de gordura, 0% de óleo de canola), B (0% de gordura, 22% de óleo de canola), C (17% de gordura, 5% de óleo de canola), D (11% de gordura, 11% de óleo de canola), E (5% de gordura, 17% de óleo de canola). Um experimento foi conduzido durante 20 semanas em que nas semanas ímpares 3 embalagens de cada um dos produtos A, B, C, D e E foram analisadas em laboratório e observadas diversas variáveis, dentre as quais a luminosidade do produto (quanto maior o valor mais claro o produto) na escala de 0 a 100, cujos resultados serão discutidos a seguir.

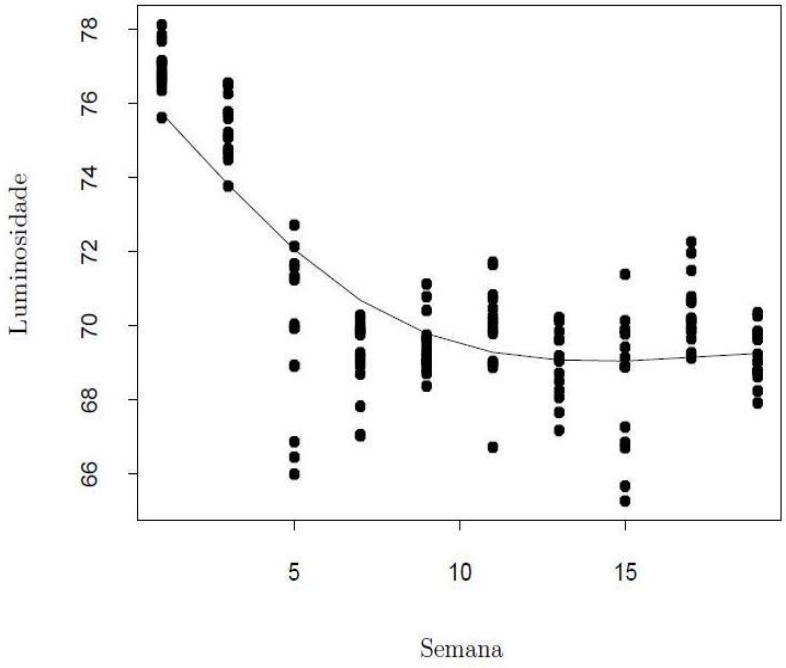
Na Figura 3.3, tem-se o comportamento da luminosidade (para todos os grupos) ao longo das 20 semanas. Notamos um decrescimento do grau de luminosidade ao longo do tempo havendo uma estabilidade a partir da 11ª semana. Queremos comparar os 5 grupos segundo a luminosidade média do produto e como o comportamento ao longo das semanas é muito similar entre os grupos, a variável tempo a princípio será tratada como uma covariável, assumindo a mesma tendência para os 5 grupos. Denotando y_{ijk} a luminosidade do $k-$ésimo produto do $i-$ésimo tipo na $j-$ésima semana, em que $i = 2(B), 3(C), 4(D), 5(E)$, sendo a forma A como referência; $j = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$

e $k = 1, 2, 3$, propomos o seguinte modelo:

$$y_{ijk} = \alpha + \beta_i + \gamma_1 x_j + \gamma_2 x_j^2 + \epsilon_{ijk}, \tag{3.16}$$

em que $\alpha + \beta_i$ é o efeito (controlado por semana) do $i-$ésimo grupo(no caso, $\beta_1 = 0$), x_j : $j-$ésima semana e ϵ_{ijk} são erros mutuamente independentes com distribuição primeiramente $N(0, \sigma^2)$ e depois $t_\nu(0, \sigma^2)$.

Figura 3.3: Comportamento da Luminosidade dos produtos ao longo das semanas.



- Análise sob erros normais

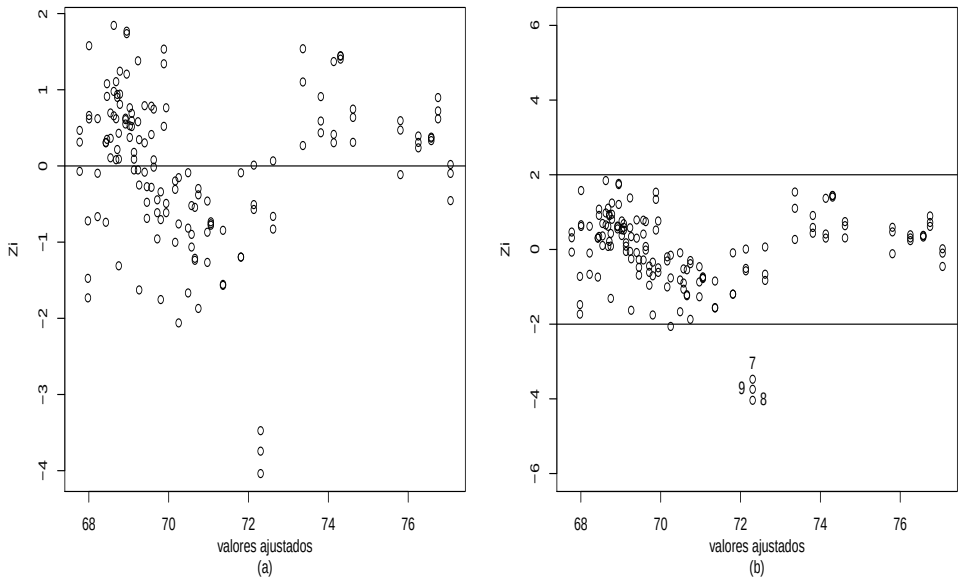
Analisando as estimativas da Tabela 3.5 sob enfoque de erros normais, nota-se pelos p-valores e ao nível de significância de 5% que apenas o grupo D parece ter uma luminosidade média inferior ao grupo A.

Na figura 3.4a (z_i, denota os resíduos padronizados do modelo normal, ver (Charnet et al. 1999)), podemos notar uma tendência não muito forte inicialmente de queda e depois de crescimento dos resíduos padronizados (sob normalidade), o que nos leva a desconfiar da homocedasticidade dos erros. Na Figura 3.4b, as observações 7,8 e 9 (referentes a 5ª semana do grupo A) aparecem como pontos aberrantes, já que se encontram fora do intervalo [-2, 2].

Tabela 3.5: EMV dos parâmetros do modelo (3.16) ajustado aos dados de luminosidade sob erros normais, com e sem pontos aberrantes (obs. 7,8 e 9).

Efeito	Com todos os pontos			Sem pontos aberrantes		
	Estimativa	t-valor	p-valor	Estimativa	t-valor	p-valor
Constante	78,13	164,25	0,00	78.91	193,81	0,00
Grupo B	0,31	0,75	0,44	-0.35	-0,99	0.31
Grupo C	-0,17	-0,41	0,68	-0.83	-2,36	0.02
Grupo D	-0,94	-2,28	0,02	-1.60	-4,56	0,00
Grupo E	-0,49	-1,19	0,23	-1.15	-3,28	0,00
γ_1	-1,44	-15,60	0,00	-1.43	-18,69	0,00
γ_2	0,05	12,32	0,00	0.05	14,45	0,00
σ^2	2,54	8,01	0,00	1,74	9,91	0,00

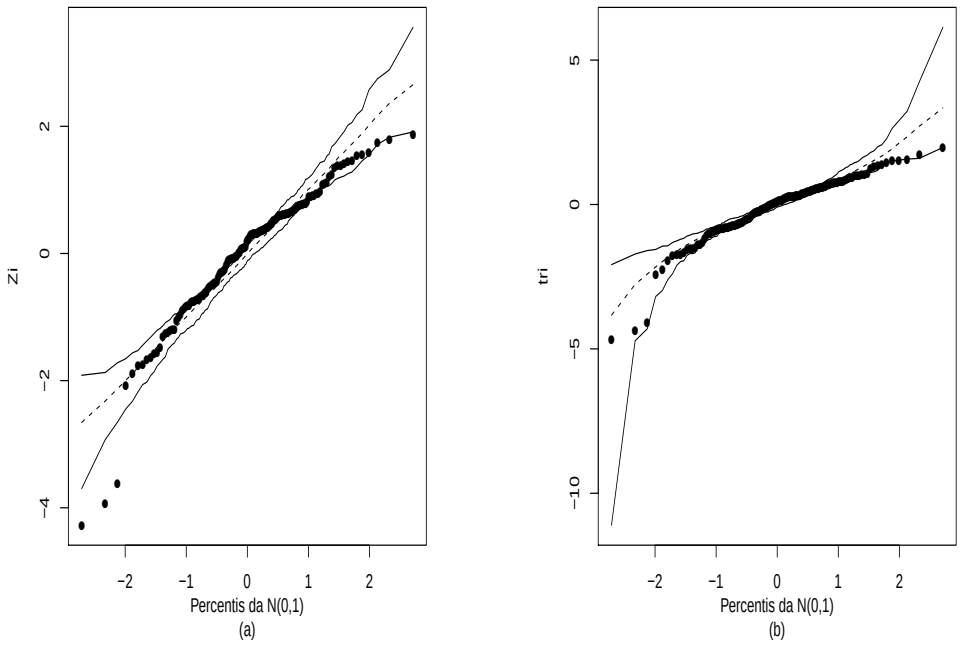
Figura 3.4: Gráfico (a) referente a suposição de homocedasticidade dos erros; Gráfico (b) referente a presença de observações aberrantes.



Na Figura 3.5, essas três observações aparecem fora da banda gerada. A eliminação desses pontos causam mudanças inferenciais, ao nível de significância de 5%, como se pode notar pela Tabela 3.5. Com a eliminação das três observações apenas o grupo B parece não diferir do grupo A, os demais grupos parecem ter um nível médio de

luminosidade menor. Nos resta saber qual modelo devemos considerar, com ou sem as três observações aberrantes. A fim de tentar reduzir a influência desses pontos nos resultados , assumiremos a seguir erros com caudas mais pesadas, no caso com distribuição t-Student.

Figura 3.5: Gráficos normais de probabilidades com envelope para o resíduo do modelo definido em (3.16) - sob erros normais (a); sob erros t-Student com $\nu = 4,45$ (b).



- Análise sob erros t-Student

Considerando erros t-Student com ν graus de liberdade para o modelo (3.16), obtemos, através do algoritmo EM, as estimativas de máxima verossimilhança apresentadas na Tabela 3.6.

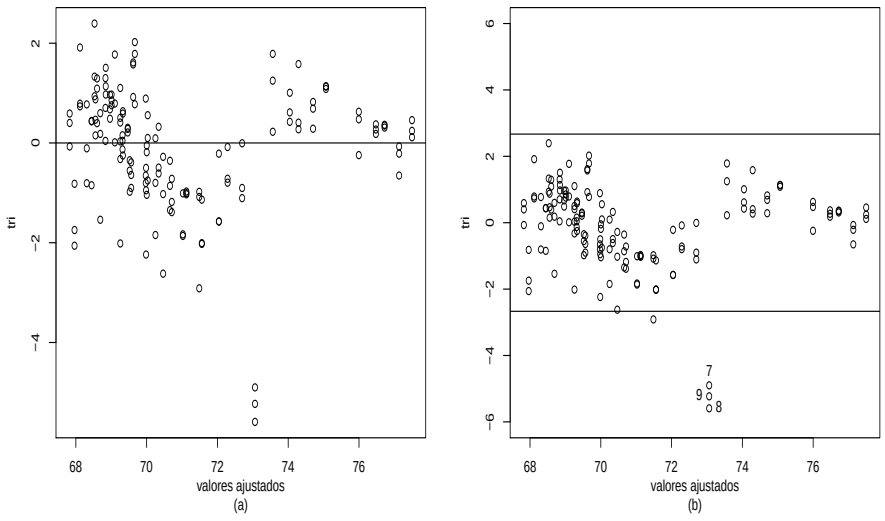
Na Figura 3.6a, podemos notar que os resíduos padronizados do modelo t-Student possuem mesma tendência do modelo Normal e as mesmas observações 7,8 e 9 aparecem como pontos aberrantes na Figura 3.6b, neste caso encontram-se fora do intervalo [-2.67, 2.67], referente ao quantil 95% da distribuição t-Student com $\nu = 4,45$ graus de liberdade.

O gráfico de envelope dado na Figura (3.5)b acomoda melhor as observações aberrantes do que sob erros normais. Quando eliminamos os três pontos aberrantes não

Tabela 3.6: EMV dos parâmetros do modelo (3.16) ajustado aos dados de luminosidade sob erros t-Student, com e sem pontos aberrantes (obs. 7,8 e 9).

Efeito	Com todos os pontos			Sem pontos aberrantes		
	Estimativa	t-valor	p-valor	Estimativa	t-valor	p-valor
Constante	78,90	195,75	0,00	79,08	215,57	0,00
Grupo B	-0,39	-1,04	0,30	-0,56	-1,61	0.11
Grupo C	-0,79	-2,18	0,03	-0,95	-2,81	0.00
Grupo D	-1,52	-3,95	0,00	-1,67	-4,61	0,00
Grupo E	-1,03	-2,91	0,00	-1,18	-3,57	0,00
γ_1	-1,43	-19,57	0,00	-1,42	-20,59	0,00
γ_2	0,05	14,44	0,00	0,05	15,05	0,00
σ^2	1,36	5,18	0,00	1,19	5,96	0,00
ν	4,45	2,86	0,00	4,45	3,48	0,00

Figura 3.6: Gráfico (a) referente a suposição de homocedasticidade dos erros; Gráfico (b) referente a presença de observações aberrantes.

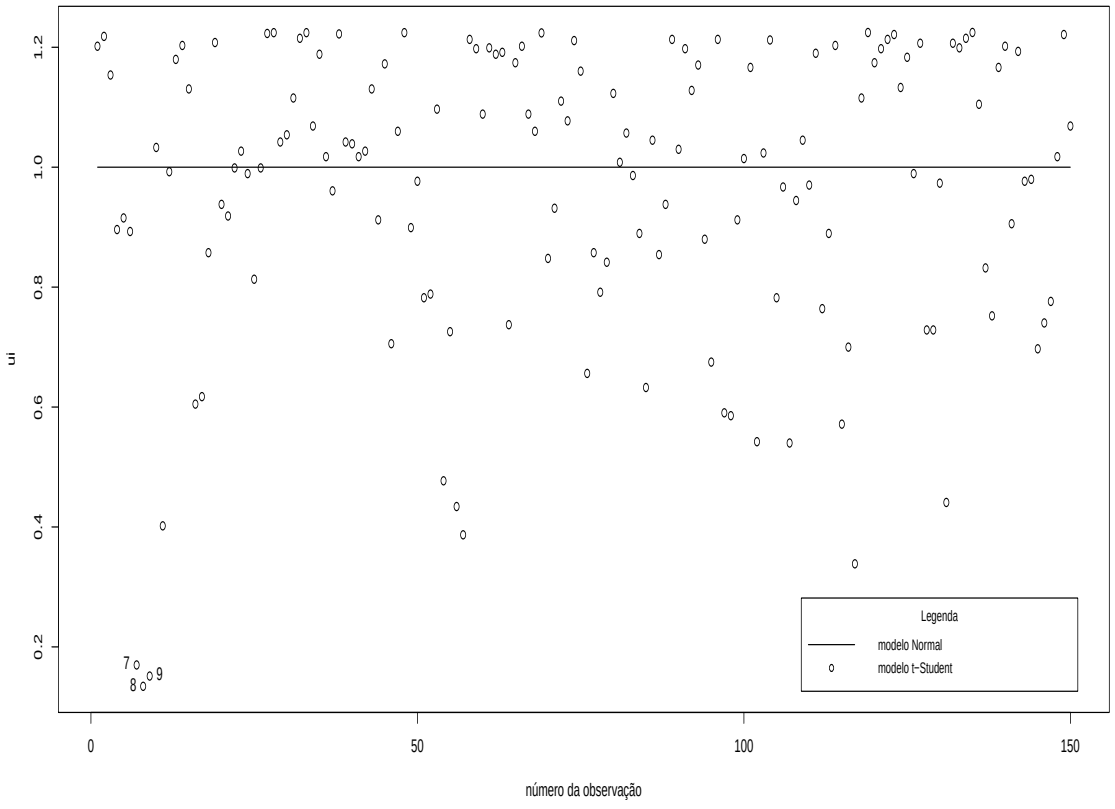


notamos mudanças inferenciais ao nível de significância de 5% (Tabela 3.6)e as estimativas variam bem menos do que sob erros normais (Tabela 3.5), neste caso não encontramos, ao nível de 5%, diferenças significativas entre os velores médios dos grupos A e B, havendo fortes indícios de que os demais grupos têm índice médio de luminosidade inferiores ao

grupo A. Esses resultados não mudam quando as observações aberrantes são eliminadas, confirmando a robustez das estimativas de máxima verossimilhança sob erros t-Student contra pontos extremos.

Para finalizar, temos a Figura 3.7 que nos mostra os valores de $\widehat{u}_i$ para os dois modelos ajustados, como dito, no modelo sob erros normais $\widehat{u}_i = 1$ para todo i e no modelo sob erros t-Student $\widehat{u}_i = \frac{(\nu+1)}{(\nu+d_i^2)}$, ver equação (3.8). Observamos, pela Figura 3.7 que $\widehat{u}_i$ atua como um "peso"para cada observação, ou seja, no modelo normal todas as observações contribuem da mesma forma para as estimativas, enquanto no modelo t-Student as observações aberrantes são as que recebem menor peso, influenciando em menor grau nas estimativas para tal modelo.

Figura 3.7: Gráfico de $\widehat{u}_i\times$ número da observação, para os modelos assumindo erros normais e erros t-Student.



Após todas as análises acima concluímos que o modelo t-Student com $\nu = 4,45$ graus de liberdade parece se ajustar melhor aos dados do que o modelo normal no sentido de robustez das estimativas obtidas contra os pontos aberrantes.

4 Conclusão

Apresentamos neste trabalho situações onde as usuais suposições de normalidade em modelos estatísticos não são verificadas, tornando necessário a aplicação de novos métodos. Citamos a classe de distribuições de modelos robustos simétricos com enfoque na distribuição t-Student. Verificamos porém, que a obtenção de estimadores de máxima verossimilhança (EMV) para tal modelo não é tão simples analiticamente quanto para o modelo normal, apresentamos então no Capítulo 2 o algoritmo EM, um método iterativo para obter tais estimadores. Introduzimos também o conceito de modelos hierárquicos através de alguns exemplos, para que a estruturação do algoritmo EM se tornasse de mais fácil compreensão.

No Capítulo 3 referente ao modelo de regressão com erros aleatórios t-Student demonstramos de forma empírica que o algoritmo EM proposto realmente convergiu para o EMV e tentamos explicar o motivo da estimativa de ν não ter se aproximado de seu verdadeiro valor, bem como sugerimos para um próximo trabalho, mudar o critério de parada utilizado, usamos neste trabalho $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| < \epsilon$ que poderia ser substituído por $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})|| < \epsilon$, pois este critério possivelmente evitaria a ocorrência do platô. Apresentamos como é feita a análise de resíduos para a classe de modelos simétricos, sob o enfoque de Cysneiros, Paula e Galea (2005), bem como o chamado gráfico de envelope (Atkison, 1981, 1985). E através de uma aplicação prática confirmamos que o modelo t-Student possui estimativas mais robustas contra pontos aberrantes do que o modelo normal, ou seja, o processo de estimação assumindo para os erros aleatórios uma distribuição t-Student é mais robusto.

5 Apêndice - Derivadas Matriciais - Modelo t-Student

Neste capítulo apresentaremos o desenvolvimento das derivadas listadas nos capítulos anteriores. Antes, porém para melhor comprensão de tais derivadas vamos a duas definições e seus resultados.

Definição 5.0.1. Sejam $\mathbf{x}$ um vetor de argumentos e $\mathbf{a}$ um vetor numérico, ambos de dimensão n. Então, $\mathbf{a}^\top\mathbf{x}$ é uma combinação linear dos elementos de $\mathbf{x}$, ou uma forma linear em $\mathbf{x}$ e seu vetor de derivadas parciais, $\frac{\partial \mathbf{a}^\top\mathbf{x}}{\partial \mathbf{x}}$, é um vetor cujo i-ésimo elemento é dado por $\frac{\partial \mathbf{a}^\top\mathbf{x}}{\partial x_i}$.

Resultado 5.0.1.

$$\frac{\partial \mathbf{a}^\top\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

Definição 5.0.2. Sejam $\mathbf{x}$ um vetor de argumentos de dimensão n e $\mathbf{A}$ uma matriz quadrada numérica, $n \times n$. Então, $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ é uma forma quadrática em $\mathbf{x}$ e seu vetor de derivadas parciais, $\frac{\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial \mathbf{x}}$, é um vetor cujo i-ésimo elemento é dado por $\frac{\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial x_i}$.

Resultado 5.0.2.

$$\frac{\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{A}^\top\mathbf{x}$$

e se $\mathbf{A}$ é simétrica,

$$\frac{\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{2}\mathbf{A}\mathbf{x}$$

- Desenvolvimento equações (3.10) e (3.11):

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = -\frac{n}{2}\log\sigma^{2(k)} - \frac{1}{2\sigma^{2(k)}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top\mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})$$

Pelos resultados 5.0.1 e 5.0.2.

$$\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial \boldsymbol{\beta}^{(k)}} = \frac{1}{2\sigma^2}[2\mathbf{X}^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})]$$

De $\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial \boldsymbol{\beta}^{(k)}} = 0$, temos que

$$\begin{aligned} \mathbf{X}^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{y} &= \mathbf{X}^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{X}\boldsymbol{\beta}^{(k)}, \text{ portanto} \\ \widehat{\boldsymbol{\beta}}^{(k)} &= (\mathbf{X}^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})\mathbf{y} \end{aligned}$$

e,

$$\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial \sigma^{2(k)}} = - \frac{n}{2\sigma^{2(k)}} + \frac{1}{2\sigma^{4(k)}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})$$

De $\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial \sigma^{2(k)}} = 0$, temos que

$$\begin{aligned} \frac{n}{2\sigma^{2(k)}} &= \frac{1}{2\sigma^{4(k)}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)}), \text{ portanto} \\ \widehat{\sigma^2}^{(k)} &= \frac{1}{n}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})^\top \mathbf{D}(\widehat{\mathbf{u}}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)})] \end{aligned}$$

- Desenvolvimento função escore (3.12):

  A partir de

$$\ell(\boldsymbol{\theta}) = n \log \mathrm{c}(\nu) - \frac{n}{2}\log \sigma^2 - \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{n} \log\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right],$$

calculamos

$$\begin{aligned} U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= -\frac{(\nu+1)}{2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(-2)\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i}{\nu\sigma^2} \\ &= \frac{(\nu+1)}{\nu\sigma^2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i \\ &= \frac{(\nu+1)}{\nu\sigma^2}\mathbf{X}^\top \mathbf{D}(\mathbf{d})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} U(\sigma^2) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{(\nu+1)}{2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(-1)\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\nu}{\nu^2\sigma^4} \\ &= -\frac{n}{2\sigma^2} + \frac{(\nu+1)}{2\nu\sigma^4}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \\ &= -\frac{n}{2\sigma^2} + \frac{(\nu+1)}{2\nu\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{D}(\mathbf{d})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$
\begin{aligned}
U(\nu) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \nu} &= \frac{n}{c(\nu)} - \frac{1}{2}\sum_{i=1}^{n} \log\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right] \\
&\quad - \frac{(\nu+1)}{2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(-1)\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\sigma^2}{\nu^2\sigma^2} \\
&= n\frac{\partial}{\partial \nu}[\log c(\nu)] - \frac{1}{2}\sum_{i=1}^{n}\log d_i^{-1} + \frac{(\nu+1)}{2\nu^2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{D}(\mathbf{d})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

- Desenvolvimento dos elementos da matriz de informação de Fisher observada (3.13):

Como a derivada da função $\ell(\boldsymbol{\theta})$ em relação a $\boldsymbol{\beta}$ nos retorna uma matriz coluna, a derivada da segunda tem que ser obtida através do transposto $\boldsymbol{\beta}^\top$. Sabemos que

$$
I_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\left(\underbrace{\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top}}\right)
$$

e

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top} &= \frac{(\nu+1)}{\nu\sigma^2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i^\top \\
&= \frac{(\nu+1)}{\nu\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{D}(\mathbf{d})\mathbf{X}
\end{aligned}
$$

portanto

$$
\begin{aligned}
I_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} = \frac{(\nu+1)}{\nu\sigma^2}(-1)\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-2)\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i\nu\sigma^2}{\nu^2\sigma^2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i^\top \\
&\quad - \frac{(\nu+1)}{\nu\sigma^2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}\mathbf{x}_i\mathbf{x}_i^\top \\
&= \frac{2(\nu+1)}{\nu^2\sigma^4}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i^\top \\
&\quad - \frac{(\nu+1)}{\nu\sigma^2}\sum_{i=1}^{n}\left[1 + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}\mathbf{x}_i\mathbf{x}_i^\top \\
&= \frac{(\nu+1)}{\nu\sigma^2}\left[\frac{2}{\nu\sigma^2}\mathbf{X}^\top \mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2\mathbf{X} - \mathbf{X}^\top \mathbf{D}(\mathbf{d})\mathbf{X}\right] \\
&= \frac{(\nu+1)}{\nu\sigma^2}\mathbf{X}^\top\left[\frac{2}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right]\mathbf{X}
\end{aligned}
$$

$$
\begin{aligned}
I_{\boldsymbol{\beta}\sigma^2} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \sigma^2} = -\frac{(\nu+1)}{\nu\sigma^4}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&\quad +\frac{(\nu+1)}{\nu\sigma^2}(-1)\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\nu}{\nu^2\sigma^4}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&= \frac{(\nu+1)}{\nu^3\sigma^8}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&\quad -\frac{(\nu+1)}{\nu\sigma^4}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&= \frac{(\nu+1)}{\nu\sigma^4}\left[\frac{1}{\nu\sigma^2}\mathbf{X}^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\mathbf{X}^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\
&= \frac{(\nu+1)}{\nu\sigma^4}\mathbf{X}^\top\left[\frac{1}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2-\mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
I_{\boldsymbol{\beta}\nu} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial \nu\partial \boldsymbol{\beta}} = -\frac{1}{\nu^2\sigma^2}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&\quad +\frac{(\nu+1)}{\nu\sigma^2}(-1)\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\sigma^2}{\nu^2\sigma^4}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&= \frac{(\nu+1)}{\nu^3\sigma^4}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&\quad -\frac{1}{\nu^2\sigma^2}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\
&= \frac{1}{\nu^2\sigma^2}\left[\frac{(\nu+1)}{\nu\sigma^2}\mathbf{X}^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\mathbf{X}^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\
&= \frac{1}{\nu^2\sigma^2}\mathbf{X}^\top\left[\frac{(\nu+1)}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2-\mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
I_{\sigma^2\sigma^2} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2\sigma^4} - \frac{(\nu+1)}{\nu\sigma^6}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad + \frac{(\nu+1)}{2\nu\sigma^4}(-1)\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\nu}{\nu^2\sigma^4}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&= \frac{n}{2\sigma^4} - \frac{(\nu+1)}{\nu\sigma^6}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad + \frac{(\nu+1)}{2\nu^2\sigma^8}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&= \frac{(\nu+1)}{\nu\sigma^6}\left[\frac{1}{2\nu\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\
&\quad + \frac{n}{2\sigma^4} \\
&= \frac{(\nu+1)}{\nu\sigma^6}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{1}{2\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad + \frac{n}{2\sigma^4}
\end{aligned}
$$

$$
\begin{aligned}
I_{\sigma^2\nu} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial\sigma^2\partial\nu} = -\frac{1}{2\nu^2\sigma^4}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad + \frac{(\nu+1)}{2\nu\sigma^4}(-1)\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\sigma^2}{\nu^2\sigma^4}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&= \frac{(\nu+1)}{2\nu^3\sigma^6}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad - \frac{1}{2\nu^2\sigma^4}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&= \frac{1}{2\nu^2\sigma^4}\left[\frac{(\nu+1)}{\nu\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\
&= \frac{1}{2\nu^2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{(\nu+1)}{\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
I_{\nu\nu} &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial\nu\partial\nu} = -\frac{1}{2}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\sigma^2}{\nu^2\sigma^4} \\
&\quad -\frac{(\nu+2)}{2\nu^3\sigma^2}\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-1}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad +\frac{(\nu+1)}{2\nu^2\sigma 2}(-1)\sum_{i=1}^{n}\left[1+\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-2}(-1)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2\sigma^2}{\nu^2\sigma^4}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&\quad +n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right] \\
&= \frac{1}{2\nu^2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{(\nu+2)}{2\nu^3\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad +\frac{(\nu+1)}{2\nu^4\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad +n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right] \\
&= \frac{(\nu+1)}{2\nu^4\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{\nu^3\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad +n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right] \\
&= \frac{1}{\nu^3\sigma^2}\left[\frac{(\nu+1)}{2\nu\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\mathbf{d})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \\
&\quad +n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right] \\
&= \frac{1}{\nu^3\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\left[\frac{(\nu+1)}{2\nu\sigma^2}\mathbf{D}(\mathbf{d})^2\mathbf{D}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2 - \mathbf{D}(\mathbf{d})\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad +n\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right]
\end{aligned}
$$

- Desenvolvimento das derivadas $\frac{\partial}{\partial\nu}[\log c(\nu)]$ e $\frac{\partial}{\partial\nu}\left[\frac{\partial}{\partial\nu}(\log c(\nu))\right]$ presentes em $U(\nu)$ e $I_{\nu\nu}$, respectivamente:

Assumindo $c(\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}$, $\Psi(\nu) = \frac{\partial}{\partial\nu}\log\Gamma(\nu)$ e $\Psi'(\nu) = \frac{\partial^2}{\partial\nu\partial\nu}\log\Gamma(\nu)$.

$$
\begin{aligned}
\log c(\nu) &= -\frac{1}{2}\log(\pi\nu) + \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) \\
&= -\frac{1}{2}\log(\pi) - \frac{1}{2}\log(\nu) + \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right)
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial \log c(\nu)}{\partial \nu} &= -\frac{1}{2\nu} + \frac{\partial}{\partial \nu} \log \Gamma \left(\frac{\nu+1}{2} \right) - \frac{\partial}{\partial \nu} \log \Gamma \left(\frac{\nu}{2} \right) \\
&= -\frac{1}{2\nu} + \frac{1}{2} \Psi \left(\frac{\nu+1}{2} \right) - \frac{1}{2} \Psi \left(\frac{\nu}{2} \right) \\
&= \frac{1}{2} \left[\Psi \left(\frac{\nu+1}{2} \right) - \Psi \left(\frac{\nu}{2} \right) - \frac{1}{\nu} \right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log c(\nu)}{\partial \nu \partial \nu} &= \frac{\partial}{\partial \nu} \left[\frac{\partial \log c(\nu)}{\partial \nu} \right] \\
&= \frac{1}{2} \left[\frac{1}{2} \Psi' \left(\frac{\nu+1}{2} \right) - \frac{1}{2} \Psi' \left(\frac{\nu}{2} \right) - \frac{1}{\nu^2} \right] \\
&= \frac{1}{4} \left[\Psi' \left(\frac{\nu+1}{2} \right) - \Psi' \left(\frac{\nu}{2} \right) - \frac{2}{\nu^2} \right]
\end{aligned}$$

Referências Bibliográficas

[1] Atkinson, A. C. (1981). Two graphical display for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

[2] Atkinson, A. C. (1985). *Plots, Transformation and Regression*. Oxford: Claredon Press.

[3] Boyles, R. A., On the Convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser.* B **45**, 47-50, 1983.

[4] Casella, G., Berger, R. L. *Inferência estatística - tradução da 2ª edição norte-americana*. Centage Learning, 2010. Página 147-151, 291-294, 329.

[5] Cysneiros, F. J. A., Paula, G.A., Galea, M. In: Escola de Modelos de Regressão, 9, 2005, Águas de São Pedro, São Paulo. *Modelos Simétricos Aplicados*.

[6] Charnet, R., et al. *Análise de Modelos de Regressão Liner com aplicações*. – Campinas, SP: Editora da Unicamp, 1999.

[7] Dempster, A. P., Laird, N.M. e Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B, **39**, 1-22.

[8] Ferreira, C. S. (2008) *Inferência e diagnóstico em modelos assimétricos*. Tese de Doutorado, Departamento de Estatística, IME-USP. São Paulo.

[9] Lange, K. L., Little, J.A. e Taylor, M.G.J. (1989). Robust modeling using the t distribution . *Journal of the American Statistical Association*, **84**, 881-896.

[10] MATLAB for Windows User's Guide, The Math Works Inc., 1991.

[11] Meng, X.; Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **81**, 633-648.

[12] Paula, G. A. *Modelos de Regressão com apoio computacional*. IME-USP, 2004. Página 28-55.

[13] R Development Core Team. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, 2011. Disponível em <http://www.R-project.org>.

[14] Wu, C. F. J. On the Convergence of the EM algorithm. *Ann. Statist.* **11**, 95-113, 1983.

[15] Zeller, C.B. (2009). *Distribuições Mistura de Escala Skew Normal: Estimação e Diagnóstico em Modelos Lineares*. Tese de Doutorado, Departamento de Estatística, IMECC-UNICAMP.

[16] Zhu, H. e Lee, S. (2001). Local influence for incomplete-data models.*Journal of the Royal Statistical Society*, B, **63**, 111-126.