UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE ESTASTÍSTICA

Monografia

Erros com Distribuição Log-Beta-Weibull em Regressões Lineares

Helson Barcelos Resende

Orientador:
Antônio Eduardo Gomes

Brasília, 9 de dezembro de 2014

Capítulo 1

Introdução

Na aplicação de modelos de regressão linear, comumente assumimos que os resíduos seguem uma distribuição normal, o que nem sempre é conveniente. Em situações onde a distribuição dos resíduos difere da normal, a suposição de normalidade pode ocasionar decisões incorretas quanto ao efeito de covariáveis sobre a variável resposta. Nesse trabalho de conclusão de curso será estudada a aplicação de uma família de distribuições pouco conhecida e relativamente nova: A família Beta-Weibull. Em especial, uma alternativa contínua nos reais, a Log-beta-Weibull (LBW) que pode vir a ser útil na modelagem dos resíduos em casos onde a normalidade não se mostra adequada.

Capítulo 2

Objetivo

Estudar casos onde se pode aplicar regressões lineares e testar a regressão Log-Beta-Weibull (LBW) comparando-a com a regressão linear simples (resíduos com distribuição normal). Observar quais aspectos levam a distribuição LBW a ser indicada na modelagem dos erros.

Capítulo 3

Metodologia

3.1 Regressão Linear Simples

O modelo de regressão linear simples consiste em uma função de média e de variância

$$E(Y|X = x) = \beta_0 + \beta_1 x \tag{3.1}$$

$$Var(Y|X = x) = \sigma^2 \tag{3.2}$$

Os parâmetros da função de média são o intercepto β_0 , o qual é o valor de $E(Y|X = x)$ quando x é igual a zero, e a inclinação β_1 , que é a taxa de variação em $E(Y|X = x)$ para uma unidade de variação em X. Variando os parâmetros, pode se obter todas as possíveis retas. Na maioria das aplicações, os parâmetros são desconhecidos e devem ser calculados usando os dados. A função de variância é assumida como sendo constante, com um valor positivo σ^2, que é geralmente desconhecido. Como a variância $\sigma^2 > 0$, o valor observado da i-ésima resposta y_i normalmente não é igual ao seu valor esperado $E(Y|X = x_i)$. Para explicar esta diferença entre os dados observados e o valor esperado, os estatísticos têm inventado uma quantidade chamada de erro estatístico, ou e_i, para o caso i definida implicitamente pela equação $y_i = E(Y|X = x_i) + e_i$ ou explicitamente por $e_i = y_i - E(Y|X = x_i)$. Se a função média considerada estiver incorreta, então a diferença entre os dados observados

e os valores da função terá uma componente não aleatória. Podemos fazer duas suposições importantes sobre os erros. Primeiro, assumimos que $E(e_i|x_i) = 0$, por isso, se nós construirmos um gráfico de dispersão dos e_i contra os x_i, teríamos um gráfico de dispersão nulo, sem padrões. O segundo pressuposto é que os erros são todos independentes, o que significa que o valor do erro para um dos casos não fornece qualquer informação sobre o valor do erro para o outro caso.

3.2 A Distribuição Log-Beta-Weibull (LBW)

Se X é uma variável aleatória, tendo a função de densidade

$$f(x) = \frac{c\lambda^c}{B(a,b)} x^{c-1} e^{-(\lambda x)^c} \{1 - e^{-(\lambda x)^c}\}^{a-1} e^{-(b-1)(\lambda x)^c}, x > 0, \tag{3.3}$$

então $Y = \log(X)$ tem distribuição Log-Beta-Weibull . A função densidade de Y, parametrizada em termos de $\sigma = c^{-1}$ e $\mu = -\log(\lambda)$, pode ser expressa na forma

$$\begin{aligned} f(y;\mu,\sigma,a,b) \\ &= \frac{1}{\sigma B(a,b)} \exp\left[\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right] \left(1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right)^{a-1} \\ &\quad \times \left(\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right)^{(b-1)}, \end{aligned} \tag{3.4}$$

sendo $-\infty < y < \infty$ e $-\infty < \mu < \infty$. Nós definimos a variável aleatória padronizada $Z = \frac{Y-\mu}{\sigma}$ com função densidade

$$\begin{aligned} \pi(z;a,b) &= \frac{1}{B(a,b)} \exp(z - \exp z)[1 - \exp(-\exp z)]^{a-1} \\ &\quad \times (\exp[-\exp z])^{(b-1)}, -\infty < z < \infty \end{aligned} \tag{3.5}$$

3.3 O modelo de regressão Linear Log Beta Weibull

Tome $v_i = (v_{i1}, ..., v_{ip})^T$ o vetor variável explicativa associado à variável resposta y_i com para $i = 1, ..., n$. Construímos um modelo de regressão linear para a variável

resposta y_i baseado na distribuição LBW dado por

$$y_i = \nu_i^T \beta + \sigma z_i, i = 1, ..., n \tag{3.6}$$

sendo que o resíduo zi tem função densidade (3.5), $\beta = (\beta_1, ..., \beta_p)^T$, $\sigma > 0$, $a > 0$ e $b > 0$ são parâmetros escalares desconhecidos e v_i é o vetor de variáveis explicativas modelando o parâmetro de locação $\mu_i = \nu_i^T \beta$. Assim, o vetor de parâmetros de locação $\mu = (\mu_1, ..., \mu_n)^T$ do modelo LBW tem uma estrutura linear $\mu = \nu\beta$, em que $v = (v_1, ..., v_n)^T$ é uma matriz conhecida do modelo. A função de log-verossimilhança total para os parâmetros do modelo $\theta = (a, b, \sigma, \beta^T)^T$ pode ser obtida a partir das Equações (3.5) e (3.6) como:

$$\begin{aligned} L(\theta) &= n\{-\log[\sigma B(a,b)]\} + \sum_{i=1}^{n}\left[\left(\frac{y_i - \nu_i^T\beta}{\sigma}\right) - \exp\left(\frac{y_i - \nu_i^T\beta}{\sigma}\right)\right] \\ &\quad +(a-1)\sum_{i=1}^{n}\log\left\{1 - \exp\left[\exp\left(\frac{y_i - \nu_i^T\beta}{\sigma}\right)\right]\right\} \\ &\quad +(b-1)\sum_{i=1}^{n}\exp\left[\exp\left(\frac{y_i - \nu_i^T\beta}{\sigma}\right)\right] \end{aligned} \tag{3.7}$$

As estimativas de máxima verossimilhança $\hat{\theta}$ de θ podem ser obtidas através da maximização da função log-verossimilhança (3.7). Sob condições gerais de regularidade, a distribuição assintótica de $\sqrt{n}(\hat{\theta} - \theta)$ é uma normal multivariada $N_{p+3}(0, K(\theta)^{-1}))$, onde $K(\theta)$ é a matriz de informação esperada. A matriz de covariância assintótica $K(\theta)^{-1}$ de $\hat{\theta}$ pode ser aproximada pela inversa da $(p+3) \times (p+3)$ matriz de informação observada $J(\theta)$ e, em seguida, a inferência sobre o parâmetro do vetor θ pode ser baseada na aproximação normal $N_{p+3}(0, J(\theta)^{-1})$ para $\hat{\theta}$. Essa distribuição normal multivariada pode ser usada para construir intervalos de confiança aproximados para alguns parâmetros em θ. De fato, um intervalo de $100(1-\alpha)\%$ confiança assintótico (ICA) para cada parâmetro θ_r é dado por

$$ICA_r = (\hat{\theta}_r - z_{\frac{\alpha}{2}}\sqrt[2]{J^{r,r}}; \hat{\theta}_r + z_{\frac{\alpha}{2}}\sqrt[2]{J^{r,r}})$$

onde $J^{r,r}$ representa a r_{th} diagonal da inversa da matriz de informação observada $J(\hat{\theta}^{-1})$ e $z_{\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição normal padrão.

3.4 As funções do R

Para encontrar os coeficientes da nossa reta de regressão, será necessário maximizar a função de verossimilhança (3.7) , tarefa essa que será realizada utilizando o software R. Para encontrar esse máximo utilizamos a função "optim", que é baseada em Nelder-Mead, quasi-Newton e algoritmos de gradiente conjugado. Essa função geralmente é utilizada para encontrar mínimos, bastando programar o negativo da verossimilhança e minimizar, achando assim o ponto de máximo da verossimilhança. Resta-nos, ainda, comparar nossa regressão com a usual (com erros normais) uma vez que não faria sentido utilizar uma ferramenta tão mais rebuscada para resolver um problema que tem solução simples se esta se mostrar satisfatória. Para comparar os modelos, utilizaremos a função AIC, que corresponde ao critério de Akaike, um critério de informação para um ou vários objetos de um modelo para o qual um valor de log-verossimilhança pode ser obtido, de acordo com a fórmula [$-2\times$ log-verossimilhança $+k\times$ NPAR] , onde NPAR representa o número de parâmetros no modelo ajustado, e k = 2 para a habitual AIC, ou $k = log(n)$ (sendo "n"o número de observações) para o chamado BIC ou SBC (critério Bayesiano de Schwarz). Ao comparar modelos ajustados por máxima verossimilhança para os mesmos dados, quanto menor AIC ou BIC, melhor o ajuste. A teoria do critério de Akaike exige que a log-verossimilhança seja maximizada: o AIC pode ser computado para modelos não ajustados por máxima verossimilhança, mas seus valores AIC não devem ser comparados.

Capítulo 4

Resultados

4.1 Sobre os casos

Caso 1 (Insulation) : Referem-se a um tipo de material para isolamento e o tempo até sua ruptura. Cada amostra foi mantida a uma temperatura elevada por um tempo especificado em semanas. Então a voltagem de quebra era medida - um teste destrutivo. Os dados estão tipicamente completos, ou seja , sem censura. Foram retirados de Nelson [p.535] e consistem na tensão de ruptura (em kV) de quatro tipos (igual alocação) para cada combinação de quatro temperaturas de ensaio (180, 225, 250 e 275 °C) e oito tempos (1, 2, 4, 8, 16, 32, 48 e 64 semanas). Temos assim então 4 x 4 x 8 = 128 observações.

Caso 2 (NO2): Sub amostra de 500 observações de um conjunto de dados que se originam em um estudo em que a poluição do ar em uma estrada está relacionada com o volume de tráfego e variáveis meteorológicas, coletadas pelo Norwegian Public Roads Administration. Os dados foram retirados do sítio [http://lib.stat.cmu.edu/datasets/].

4.2 Sobre as regressões

Nos dois casos a função de verossimilhança (3.7) aplicada nos respectivos dados foi maximizada, através da função optim do R. Os resíduos foram encontrados, no caso das regressões LBW, sem nenhuma função especifica do R. Simplesmente atribuimos à uma variável o valor respectivo aos modelos e seus coeficientes. Depois plotamos o gráfico dos valores estimados pelo modelo menos o valores observados. O valor AIC foi calculado de maneira análoga : uma variavel recebeu o valor calculado, para isso utilizamos o valor de máxima da função de verossimilhança disponivel nos resultados da função optim. No caso da regressão simples utilizamos a função "residuals"já própria da função "lm"e plotamos os valores encontrados. O valor AIC pode ser obtido diretamente da função AIC no R.

Caso 1: Tomaremos por variavel explicativa a temperatura (em ^{o} C) e como resposta a tensão de ruptura (em kv). Os valores estimados dos coeficientes ,com seus devidos erros, além do valor AIC estão na tabela (4.1). Os gráficos de dispersão dos resíduos na regressão LBW e são dados nas figuras (4.1) e (4.2)

Caso 2: A variável resposta consiste em valores horários do logaritmo da concentração de NO2 (partículas), medido pelo Alnabru em Oslo, Noruega, entre outubro de 2001 e agosto de 2003. As variáveis de previsão são o logaritmo do número de carros por hora e a velocidade do vento (m/s). Os valores estimados dos coeficientes ,com seus devidos erros, além do valor AIC estão na tabela (4.2). Os gráficos de dispersão dos resíduos são dados nas figuras (4.3) e (4.4).

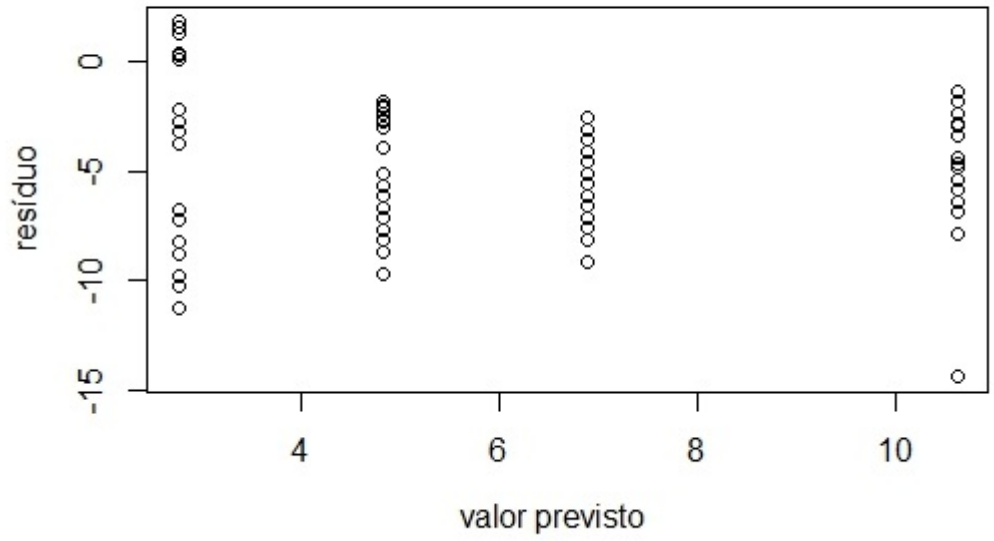

Figura 4.1: Gráfico de dispersão dos resíduos: Regresão LBW Caso1 : Insulation

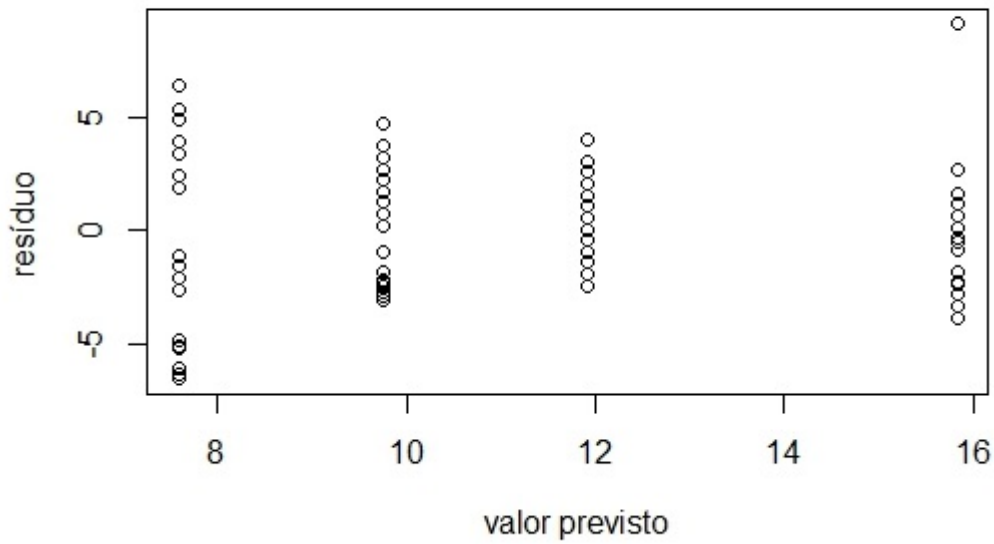

Figura 4.2: Gráfico de dispersão dos resíduos: Regresão Simples Caso1 : insulation

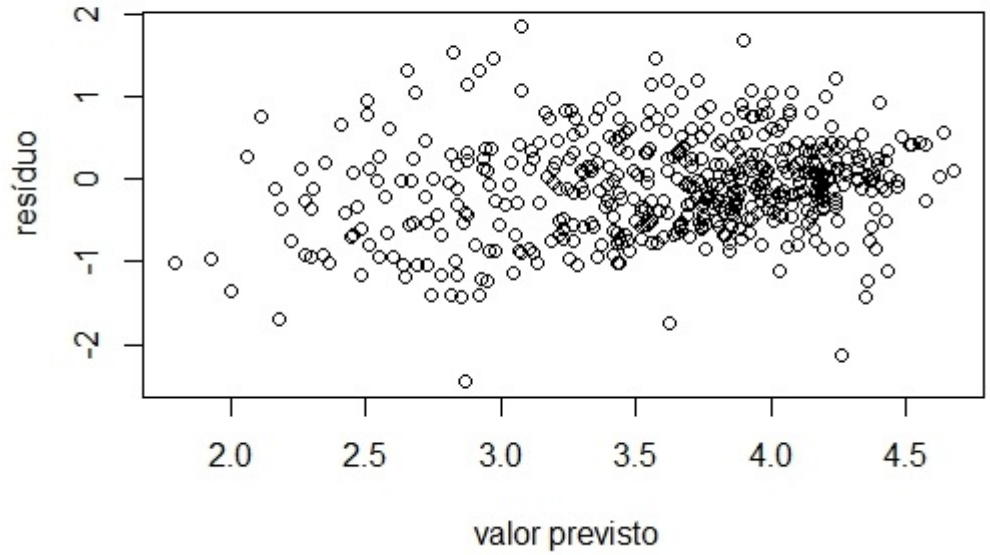

Figura 4.3: Gráfico de dispersão dos resíduos: Regresão LBW Caso 2 : No2

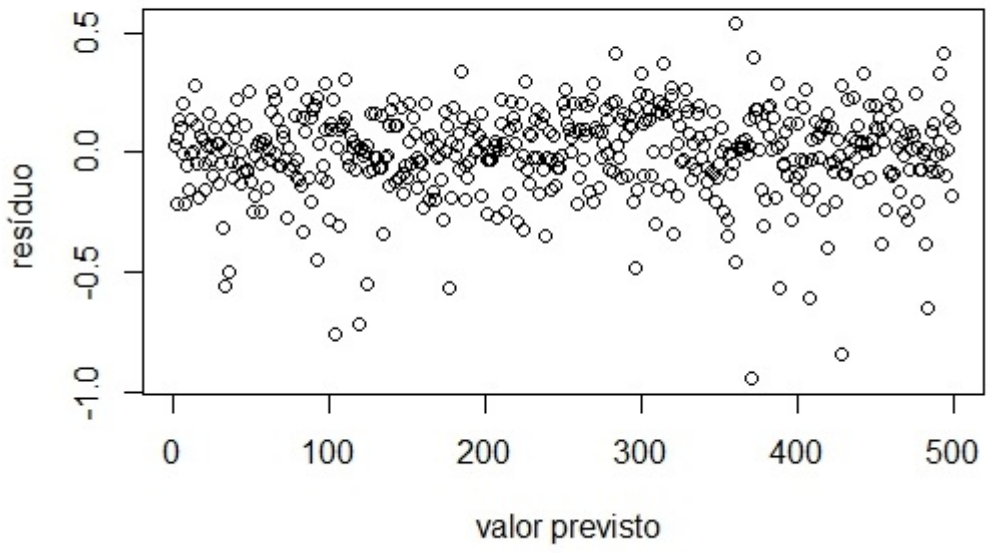

Figura 4.4: Gráfico de dispersão dos resíduos: Regresão Simples Caso 2 : No2

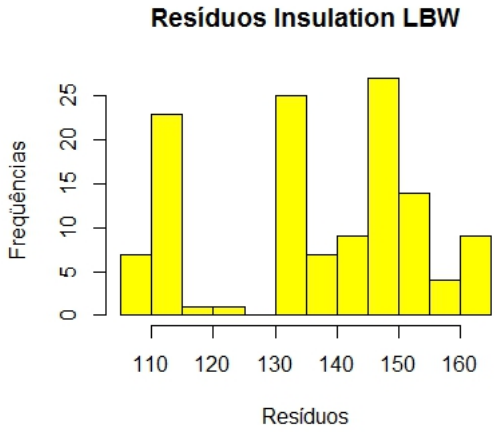

Figura 4.5: Histograma dos resíduos: Regresão LBW Caso1 : insulation

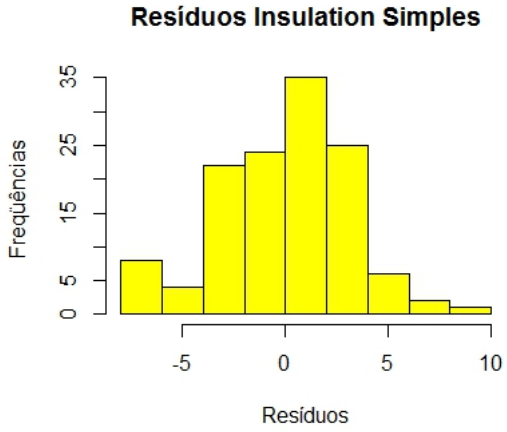

Figura 4.6: Histograma dos resíduos: Regresão Simples Caso1 : insulation

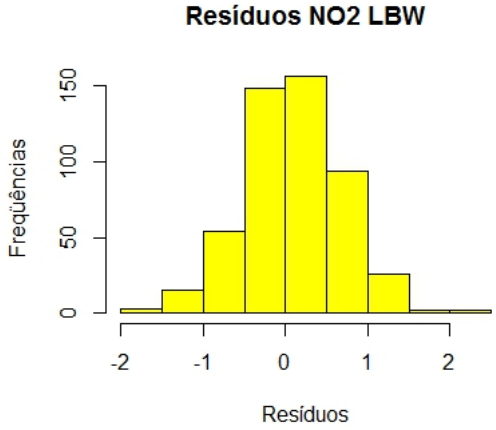

Figura 4.7: Histograma dos resíduos: Regresão LBW Caso2 : NO2

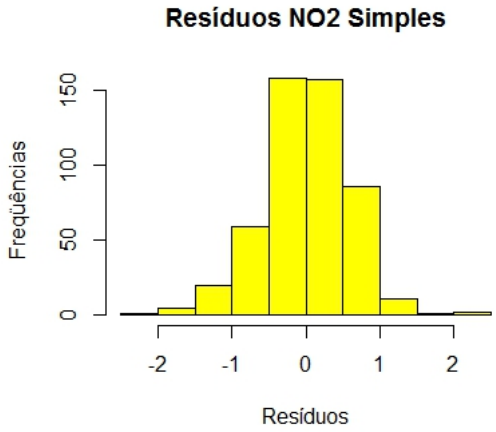

Figura 4.8: Histograma dos resíduos: Regresão Simples Caso2 : NO2

Tabela 4.1: Parâmetros estimados e seus erros Caso 1: Insulation

Modelo	a	b	σ	β_0	β_1	AIC
LBW	10.2454	2.2472	8.5276	25.5222	-0.0827	- 632
	(1.8154)	(1.7458)	(1.3195)	(3.0319)	(0.0081)	
Simples	.	.	3.068	31.4496	-0.0867	-488
				(1.8180)	(0.0077)	

Tabela 4.2: Parâmetros estimados e seus erros Caso2 : No2

Modelo	a	b	σ	β_0	β_1	β_2	AIC
LBW	3.1455	1.7283	1.0541	0.6266	0.4972	-0.1587	1.66
	(0.1170)	(0.0603)	(0.0312)	(0.0276)	(0.0039)	(0.0077)	
Simples	.	.	0.1828	0.6147	0.1167	-0.0475	-275
				(0.0538)	(0.0075)	(0.0046)	

Capítulo 5

Conclusão

Caso 1: Nesse primeiro caso os valores dos coeficientes angulares da reta de regressão foram praticamente o mesmo (algo em torno de - 0,08). O fato desse valor ser negativo indicaria uma correlação negativa entre a variável explicativa e a resposta. Assim, quanto mais tempo se passa menor a tensão necessária para o isolamento falhar. A diferença entre os modelos ficou então a cargo do intercepto; como não faz sentido testar o isolamento com tensão zero, o intercepto serve apenas para deslocar nossa reta. Os gráficos dos resíduos parecem indicar a necessidade de incluir alguma covariável no modelo Os valores AIC encontrados sugerem que a regressão LBW seria mais indicada, decisão amparada pelos valores dos erros que são menores que os da simples. Esse gráfico de dispersão torna a regressão não confiável uma vez que não poderíamos supor que a esperança dos erros seja zero. Os valores dos resíduos na LBW têm, ainda, amplitude maior que os da regressão simples e são majoritariamente negativos. Seria, nesse caso, mais confiável utilizar a regressão simples. Caso 2: Nesse modelo a quantidade de poluição é explicada pelo número de carros e pela velocidade do vento. Os valores dos coeficientes são coerentes: quanto mais carros tivermos, mais poluição e quanto mais vento menos poluição (era mesmo de se esperar uma vez que o vento vindo de um lugar menos poluído diminua a concentração de poluente). Diferente do primeiro caso, aqui os valores do

intercepto foram muito próximos enquanto os dos outros dois coeficientes diferiram bastante. Os valores dos erros apontam para o uso da regressão LBW enquanto o valor AIC sugere a regressão simples. Os gráficos de dispersão não descartam nenhum dos modelos. No caso da regressão LBW observa-se que o modelo tende a acertar mais valores altos de poluição e tem amplitude um pouco maior que a do caso simples. Novamente seria mais indicada a regressão simples.

Capítulo 6

Referências

.

[1] M. Cordeiro, Gomes, da-Silva, M.M. Ortega (2011): The beta exponentiated Weibull distribution, Journal of Statistical Computation and Simulation, DOI:10.1080/00949655.2011.615838

[2] G. Casela e R.L. Berger, Inferência Estatística. Tradução da 2º Edição Norte-Americana.

[3] Applied Linear Statistical Models by Neter, Kutner, et. al.

[4] W.B. Nelson, Accelerated Testing Statistical Models, Test, Plans and Data Analysis,Wiley, Hoboken, NJ, 2004.