Modelos heterocedásticos com erros nas variáveis: *modelando a emissão de raios-X contra a massa de buracos negros*

Alexandre Galvão Patriota

IME - USP

Roteiro

Emissão de raios-X versus proporção de Eddington

Os **raios-X** são radiações altamente energéticas que são emitidas pelo disco de acresção de buracos negros supermassivos. Não podem ser observadas, contudo telecópios espaciais podem capta-las (e.g., Hubble e Chandra).

A **proporção de Eddington de quasares** é uma função da massa do buraco negro. Que por sua vez é estimada através da atividade do núcleo galáctico, também conhecido por quasar (quase estrela).

As observações são estimativas obtidas através de tratamento de dados. Cada observação possui uma medida de variabilidade. O interesse é relacionar a verdadeira **emissão de raios-X** com a verdadeira **proporção de Eddington** que não são observados.

OBS: As variáves são trasformadas aplicando o logaritmo na base 10.

Emissão de raios-X versus proporção de Eddington

Os **raios-X** são radiações altamente energéticas que são emitidas pelo disco de acresção de buracos negros supermassivos. Não podem ser observadas, contudo telecópios espaciais podem capta-las (e.g., Hubble e Chandra).

A **proporção de Eddington de quasares** é uma função da massa do buraco negro. Que por sua vez é estimada através da atividade do núcleo galáctico, também conhecido por quasar (quase estrela).

As observações são estimativas obtidas através de tratamento de dados. Cada observação possui uma medida de variabilidade. O interesse é relacionar a verdadeira **emissão de raios-X** com a verdadeira **proporção de Eddington** que não são observados.

OBS: As variáves são trasformadas aplicando o logaritmo na base 10.

Emissão de raios-X versus proporção de Eddington

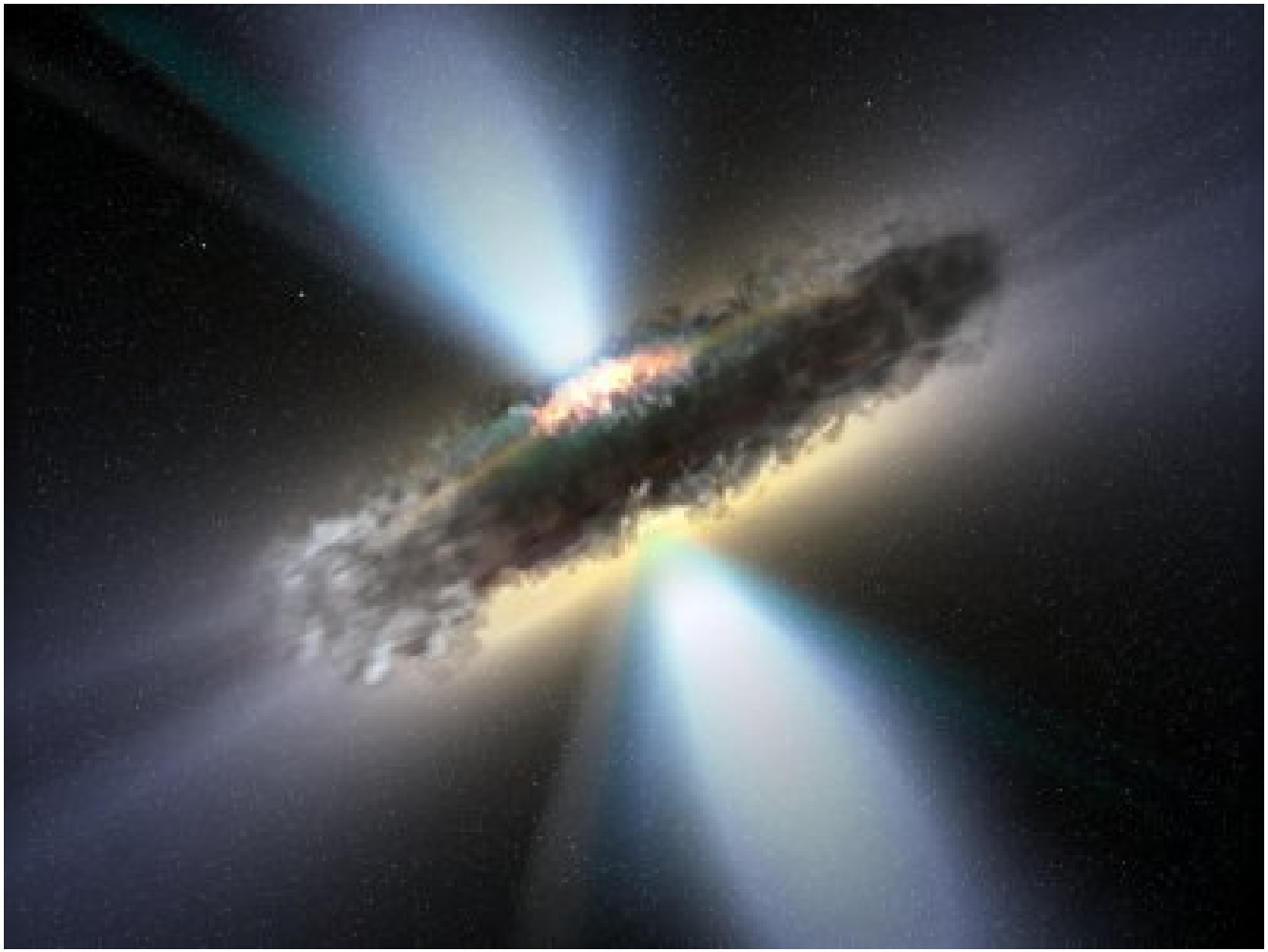
Os **raios-X** são radiações altamente energéticas que são emitidas pelo disco de acresção de buracos negros supermassivos. Não podem ser observadas, contudo telecópios espaciais podem capta-las (e.g., Hubble e Chandra).

A **proporção de Eddington de quasares** é uma função da massa do buraco negro. Que por sua vez é estimada através da atividade do núcleo galáctico, também conhecido por quasar (quase estrela).

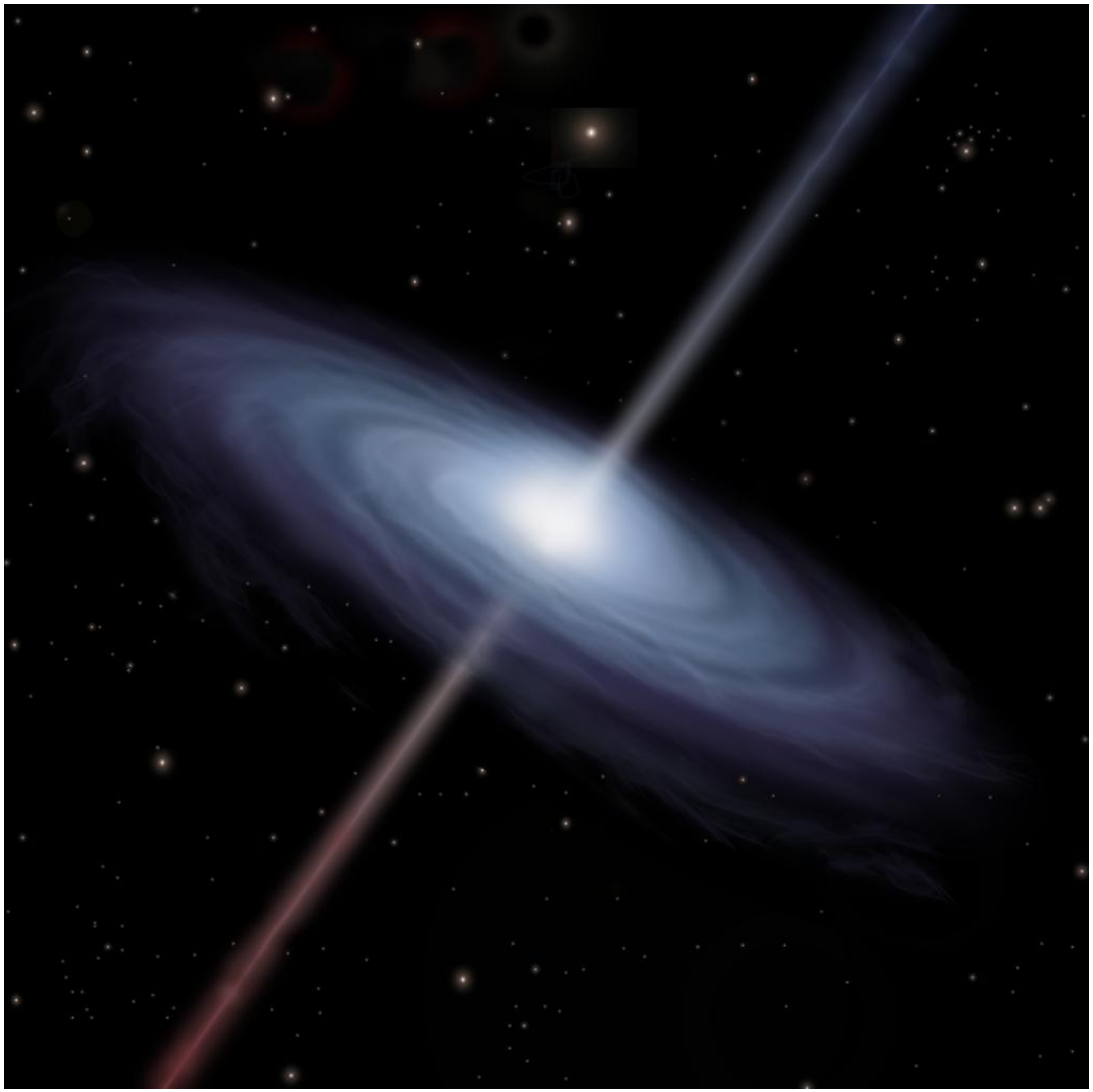
As observações são estimativas obtidas através de tratamento de dados. Cada observação possui uma medida de variabilidade. O interesse é relacionar a verdadeira **emissão de raios-X** com a verdadeira **proporção de Eddington** que não são observados.

OBS: As variáves são trasformadas aplicando o logaritmo na base 10.

Buraco Negro emitindo quasares

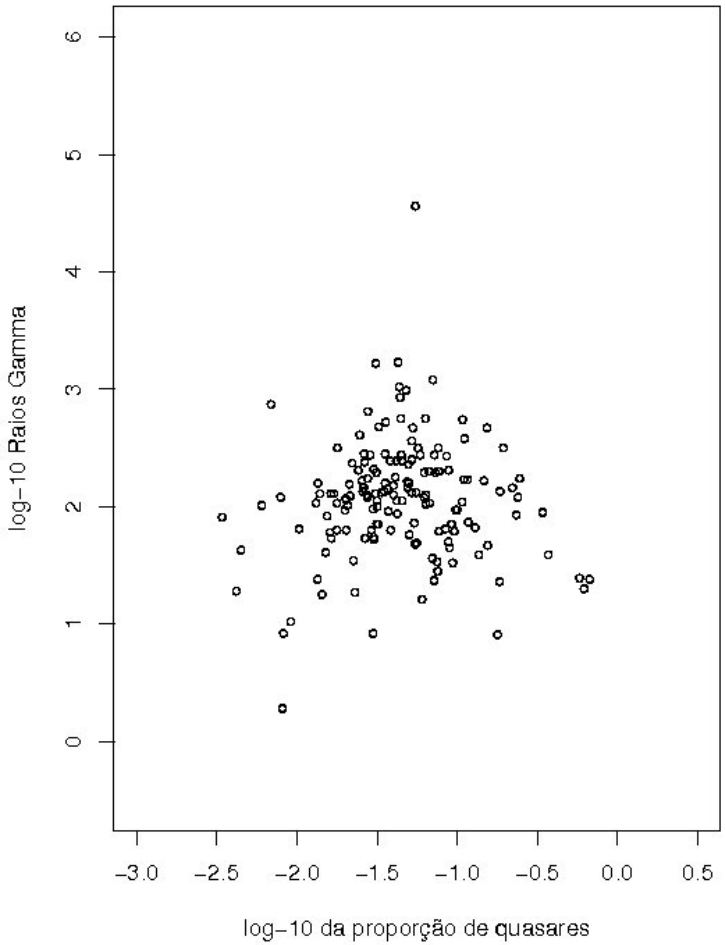


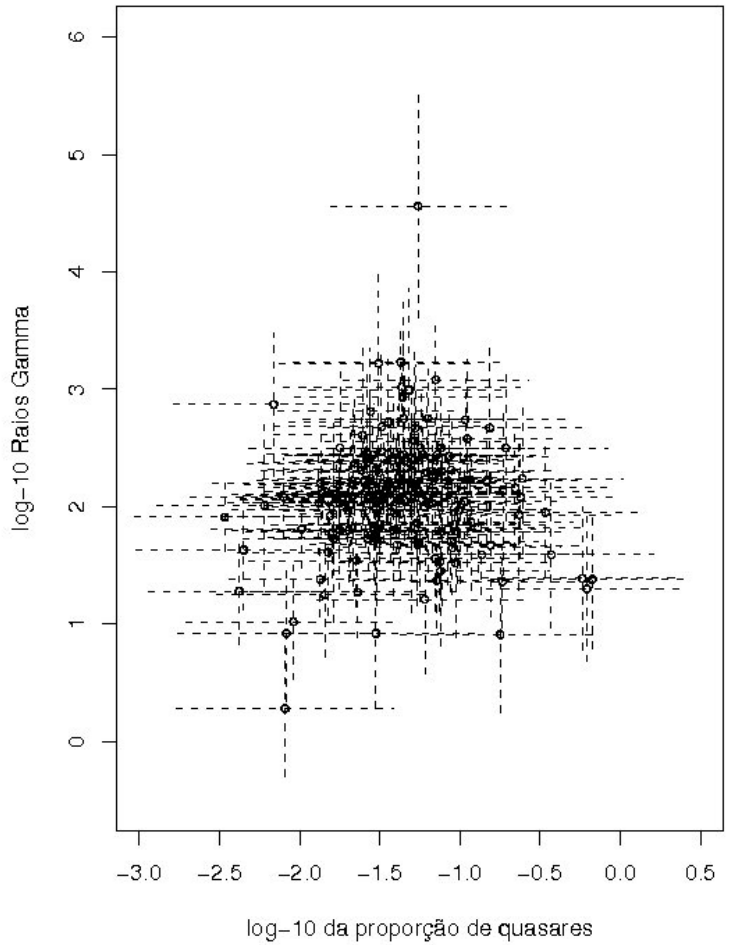
Buraco Negro emitindo quasares



Dados do Observatório de Chandra

O Observatório *Chandra* é o carro-chefe da NASA na missão para detectar raios-X. O tamanho da amostra é de 153 objetos.





Modelo com erros nas variáveis

$$\begin{aligned} Y_i &= y_i + e_i \\ X_i &= x_i + u_i \end{aligned} \tag{1}$$

onde Y_i e X_i são as variáveis observadas, y_i e x_i são as variáveis não observadas, e por fim e_i e u_i são os erros do modelo com dist. Normal de média zero e **variâncias conhecidas** λ_i e κ_i, respectivamente.

As variáveis não observadas y_i e x_i podem ser relacionadas da seguinte forma [Patriota and Bolfarine(2008)].

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + q_i \quad \text{para} \quad i = 1, 2, \ldots, n$$

onde n é o tamanho amostral, q_i é o erro da equação com distribuição normal de média zero e variância σ^2 (desconhecida). O modelo linear é siductio em [Patriota et al(2009)]

Modelo com erros nas variáveis

$$\begin{aligned} Y_i &= y_i + e_i \\ X_i &= x_i + u_i \end{aligned} \tag{1}$$

onde Y_i e X_i são as variáveis observadas, y_i e x_i são as variáveis não observadas, e por fim e_i e u_i são os erros do modelo com dist. Normal de média zero e **variâncias conhecidas** λ_i e κ_i, respectivamente.

As variáveis não observadas y_i e x_i podem ser relacionadas da seguinte forma [Patriota and Bolfarine(2008)].

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + q_i \quad \text{para} \quad i = 1, 2, \ldots, n$$

onde n é o tamanho amostral, q_i é o erro da equação com distribuição normal de média zero e variância σ^2 (desconhecida). O modelo linear é siductio em [Patriota et al(2009)]

Objetivos principais

1. Estimar os parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ e σ^2 consistentemente.

2. Testar se $\mathbf{G}\boldsymbol{\beta} = \mathbf{d}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, $\mathbf{G}$ é uma matriz de dimensões $q \times (p+1)$ e $\mathbf{d}$ um vetor de dimensão q.

3. Regiões de confiança para $\mathbf{G}\boldsymbol{\beta}$

Por exemplo, para testar se $(\beta_0, \beta_3) = (0, 0)$, temos que

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix} \quad \text{e} \quad \mathbf{d} = (0,0)^\top$$

Objetivos principais

1. Estimar os parâmetros β_0, β_1, ..., β_p e σ^2 consistentemente.

2. Testar se $\mathbf{G}\boldsymbol{\beta} = \mathbf{d}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, $\mathbf{G}$ é uma matriz de dimensões $q \times (p+1)$ e $\mathbf{d}$ um vetor de dimensão q.

3. Regiões de confiança para $\mathbf{G}\boldsymbol{\beta}$

Por exemplo, para testar se $(\beta_0, \beta_3) = (0, 0)$, temos que

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix} \quad \text{e} \quad \mathbf{d} = (0,0)^\top$$

Objetivos principais

1. Estimar os parâmetros β_0, β_1, ..., β_p e σ^2 consistentemente.

2. Testar se $\mathbf{G}\boldsymbol{\beta} = \mathbf{d}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, $\mathbf{G}$ é uma matriz de dimensões $q \times (p+1)$ e $\mathbf{d}$ um vetor de dimensão q.

3. Regiões de confiança para $\mathbf{G}\boldsymbol{\beta}$

Por exemplo, para testar se $(\beta_0, \beta_3) = (0, 0)$, temos que

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix} \quad \text{e} \quad \mathbf{d} = (0,0)^\top$$

Objetivos principais

1. Estimar os parâmetros β_0, β_1, ..., β_p e σ^2 consistentemente.

2. Testar se $\mathbf{G}\boldsymbol{\beta} = \mathbf{d}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, $\mathbf{G}$ é uma matriz de dimensões $q \times (p+1)$ e $\mathbf{d}$ um vetor de dimensão q.

3. Regiões de confiança para $\mathbf{G}\boldsymbol{\beta}$

Por exemplo, para testar se $(\beta_0, \beta_3) = (0, 0)$, temos que

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix} \quad \text{e} \quad \mathbf{d} = (0,0)^\top$$

Estimação via escore corrigido

A abordagem do escore corrigido foi proposta por Nakamura em 1990. A idéia principal é a seguinte:

(1) Encontre a log-verossimilhança, ℓ, considerando que x_i é observado diretamente sem erro de medição.

$$(Y_i = \underbrace{\beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + q_i}_{y_i} + e_i)$$

(2) Encontre uma pseudo log-verossimilhança, ℓ^*, que seja função apenas das variáveis observadas Y_i e X_i de tal forma que

$$E(\ell^*(\mathbf{Y}, \mathbf{X})|\mathbf{Y}, \mathbf{x}) = \ell(\mathbf{Y}, \mathbf{x})$$

Nakamura mostrou que, sob certas condições de regularidade, os estimadores obtidos da função ℓ^* são consistentes e têm distribuição normal assintótica com uma matriz sanduíche de covariâncias assintótica.

Simulações de Monte Carlo

- $X_i|x_i \sim \mathcal{N}(x_i, \kappa_i)$ onde $\sqrt{\kappa_i} \sim U(0.1, 0.2)$,
- $Y_i|x_i \sim \mathcal{N}(y_i, \lambda_i)$ onde $\sqrt{\lambda_i} \sim U(0.1, 0.4)$.
- $x_i \sim \mathcal{N}(-2, 0.13)$. A variância do erro na equação é $\sigma^2 = 0.1$

Dois tipos de relações foram estudadas, a saber: linear e cúbica. Para a relação linear, nós considerados (β_0, β_1) em uma vizinhança de $(0, 1)$. Para a relação cúbica, foi considerado $\beta_0 = \beta_2 = 0$ e (β_1, β_3) em uma vizinhança de $(0.1, 0)$.

Para cada simulação (em um total de 10 000) estimamos os parâmetros, a matriz de covariâncias e testamos se $(\beta_0, \beta_1) = (0, 1)$ para o caso linear e $(\beta_1, \beta_3) = (0.1, 0)$. Como resultado obtemos o tamanho e poder do tese empíricos.

Simulação para $n = 40$, $n = 80$ e $n = 160$

	Linear				*Cúbica*		
β_0		β_1		β_3		β_1	
$n = 40$	0.9	1	1.1		-0.1	0.1	0.3
-0.1	0.3732	0.3529	0.9923	-0.5	0.9713	0.9345	0.9437
0.0	0.8525	**0.1255**	0.8384	0.0	0.9517	**0.2705**	0.9702
0.1	0.9933	0.3541	0.3418	0.5	0.9291	0.9633	0.9761
$n = 80$	0.9	1	1.1		-0.1	0.1	0.3
-0.1	0.5447	0.5545	0.9999	-0.5	0.8828	0.7725	0.9867
0.0	0.9898	**0.0641**	0.9873	0.0	0.8268	**0.1168**	0.9955
0.1	0.9999	0.5432	0.5160	0.5	0.7801	0.8824	0.9988
$n = 160$	0.9	1	1.1		-0.1	0.1	0.3
-0.1	0.8692	0.8538	1.0000	-0.5	0.7042	0.6325	0.8718
0.0	1.0000	**0.0569**	1.0000	0.0	0.6590	**0.0598**	0.9156
0.1	1.0000	0.8483	0.8572	0.5	0.6340	0.7054	0.9556

Emissão de Raios-X versus proporção de Eddington

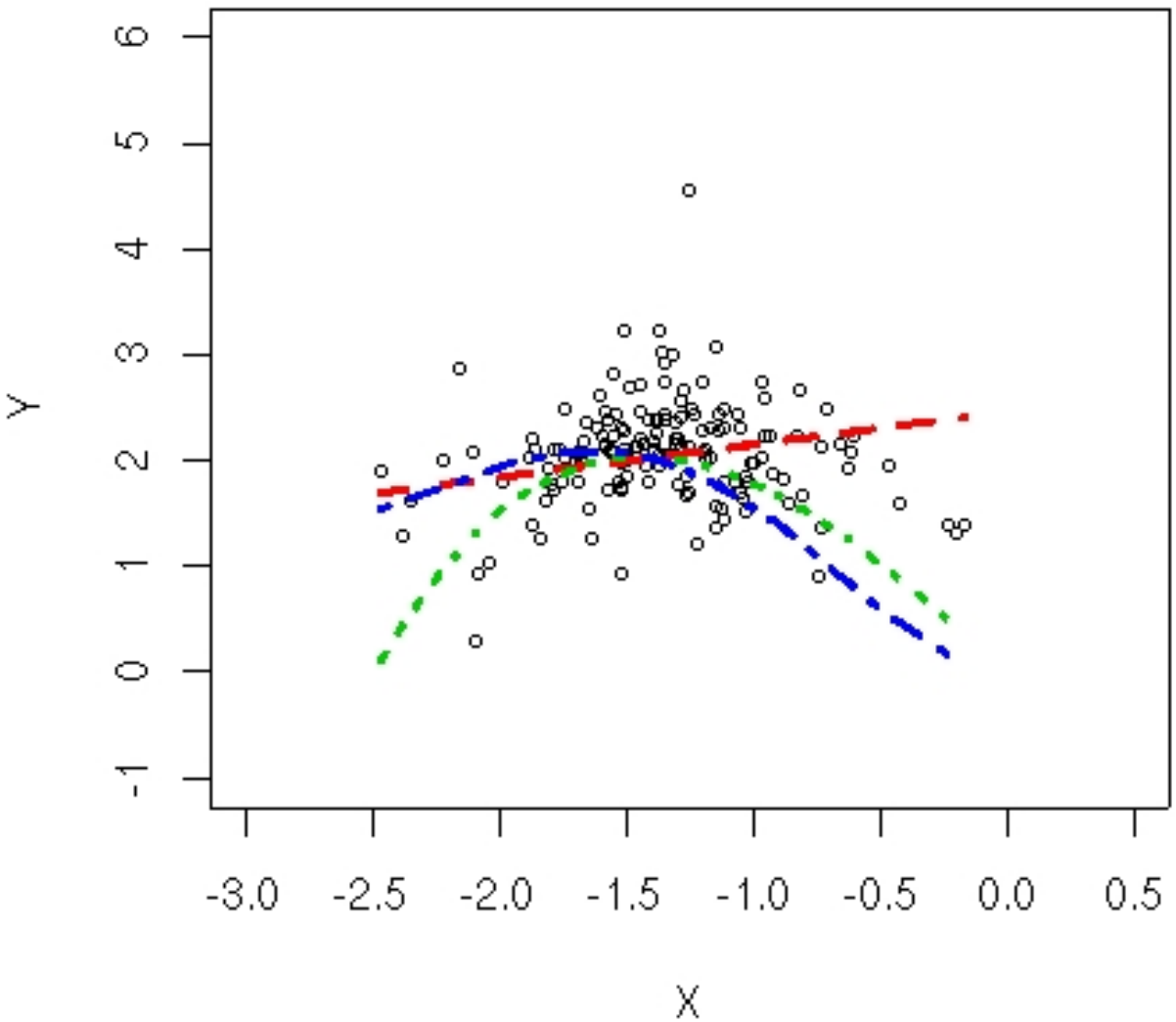
Consideramos um modelo cúbico e quártico:

$$y_i|x_i \sim \mathcal{N}(\beta_1 x_i + \beta_3 x_i^3, \sigma^2), \quad y_i|x_i \sim \mathcal{N}(\beta_2 x_i + \beta_3 x_i^3 + \beta_4 x_i^4, \sigma^2)$$

As estimativas são dadas por:

	Linear (MV)	Cúbica	Quártica
β_0	2.4618 (0.6765)	-	-
β_1	0.3102 (0.4987)	-2.1202 (0.3944)	-
β_2	-	-	3.3880 (0.5710)
β_3	-	0.3415 (0.2156)	2.2296 (0.7024)
β_4	-	-	0.3889 (0.2248)
σ^2	0.1279	0.1156 (0.0293)	0.1453 (0.0381)

Emissão de Raios-X versus proporção de Eddington



Considerações finais

Não foi possível estimar os modelos linear e quadrático neste conjunto de dados, pois não houve convergência para estes casos.

De acordo com os resultados das simulações, percebemos que a não convergência pode ocorrer quando:

- Um modelo polinomial de ordem p é verdadeiro e um modelo de ordem $q < p$ é estimado,
- Os erros de medida da variável X são muito grandes. Fato que ocorre nos dados da astrofísica.

O modelo proposto pode ser aplicado a problemas de regressão com erros em ambas as variáveis cuja relação não é linear.

- Akritas MG, Bershady MA. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*. **470**:706–714.
- Gimenez P and Bolfarine H. (1997). Corrected score functions in classical error-in-variables and incidental parameter models. *Australian Journal of Statistics* **39**(3): 325–344.
- Kelly BC, Bechtold J, Trump JR, Vertergaard M. (2008). Observational constraints on the dependence of ratio-quiet quasar X-ray emission on black hole mass and accretion disk. (To appear) *The Astrophysical Journal*.
- Kulathinal SB, Kuulasmaa K, Gasbarra D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement error vary between the observations. *Statistics in Medicine*. **21**:1089–1101.

- Nakamura T. (1990). Corrected score functions for errors-in-variables models: methodology and applications to generalized linear models. *Biometrika* **77**:127–137.
- Patriota, AG, Bolfarine, H. (2008). A heteroscedastic polynomial regression with measurement error in both axes. *Sankhya Series B*, **70**: 267-282.
- Patriota, AG, Bolfarine, H, de Castro, M. (2009). A heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology*, **6**: 408-423.
- Zavala AAZ, Bolfarine H and de Castro M. (2007). Consistent estimation and testing in heteroscedastic polynomial errors-in-variables models. *AISM* **59**:515–530.