O método dos mínimos quadrados: análise de regressão

- O métodos dos mínimos quadrados é um método famoso para lidar com dados ruidosos. Sua justificativa segue diretamente do método da máxima verossimilhança. Para Y_i dados medidos en valores das variável independente X_i, se a distribuição dos resíduos é gaussiana, então o log a verossimilhança é a soma daso quadrados na seguinte forma:

$$\ln \mathcal{L} = cte. - \sum_{i=1}^{N} \xi_i (Y_i - \mu(X_i, \alpha_1, \alpha_2, ...))^2,$$

onde ξ_i são os pesos, usualmente inversamente proporcionais à variância.

- De fato, o método dos mínimos quadrados é apenas isso: Buscamos os parâmetros do modelo que minimiza

$$\ln \mathcal{L} = cte. - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \mu(X_i, \alpha_1, \alpha_2, ...))^2$$

- Trata-se apenas de um MLE e tudo que dissemos antes se aplica. Caso os erros não forem conhecidos, ainda é possível obter-se os parâmetros, mas não se pode estimar a qualidade do modelo.

O método dos mínimos quadrados: análise de regressão

- A matriz das segundas derivadas que define a matriz de covariância das estimativas, a matriz Hessiana, é de particular importância no caso dos mínimos quadrados, pois é usado pelos métodos numéricos para a determinação dos parâmetros.

- O valor do Hessiano no mínimo sai como um sub-produto da minimização. Esta pode ser usada diretamente para a determinação da matriz de covariância, sempre que os modelos forem lineares em relação aos parâmetros (i.e. como $\alpha z^2 + \beta exp(-z)$ e não como $\alpha\, exp(-\beta z)$).

- Claro que um modelo pode ser aproximadamente linear próximo do mínimo, mas quanto? Em geral os problemas aparecem primeiro nas 'asas', das distribuições, de modo que altos graus de significância são difíceis de se confiar, a menos que tenham sido calculados exatamente ou testados via métodos de monte Carlo.

- Exemplo W&J P. 135

O método dos mínimos quadrados: análise de regressão

- Frequentemente não se pode fiar que estamos lidando com resíduos gaussianos e a razão mais frequente para isso são os contaminantes - resíduos que são extremamente improváveis numa hipótese gaussiana.

- Uma distribuição que tem uma "cauda gorda", e contrasta com a gaussiana é a exponencial

$$prob(x) = \frac{1}{2a} exp\left[-\frac{|x-\mu|}{a}\right]$$

- Se os resíduos são distribuídos exponencialmente é fácil ver que a máxima verossimilhança resulta numa minimização da soma dos valores absolutos dos resíduos.

- O problema são os rotinas numéricas que são um pouco mais pesadas. O método dos mínimos quadrados, em comparação é muito mais desenvolvido.

O método dos mínimos quadrados: análise de regressão

- Voltemos agora ao ajuste do modelo $y = ax + b$ via método dos mínimos quadrados. Isso leva às expressões bem conhecidas para a inclinação (a) e o intercepto (b):

$$a = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum Y_i - (\sum X_i)^2}$$

e

$$b = \frac{\sum Y_i - a \sum Y_i}{N}$$

- Na ausência de informações do "como e por que" da relação entre x_i e y_i qualquer modelo de dois parâmetros pode se ajustados aos dados por simples transformações de coordenadas, por exemplo:
  - Um exponencial, $y = b, exp(ax)$, requer a mudança de y_i para $\ln y_i$ na expressões acima.
  - Para uma lei de potência, $y = b\, x^a$, mude y_i para $\ln y_i$ e x_i para $\ln x_i$
  - Para uma parábola, $y = b + ax^2$, mude x_i para $\sqrt{x_i}$

- Note que os resíduos podem não ser gaussianos para todas essas transformações (e de fato não são). Claro que sempre é possível minimizar os quadrados dos resíduos, mas a justificativa formal para isso pode não ser mais válida. Os testes do capítulo 5 são úteis para verificar se o ajuste é aceitável, em particular o teste de *runs*.

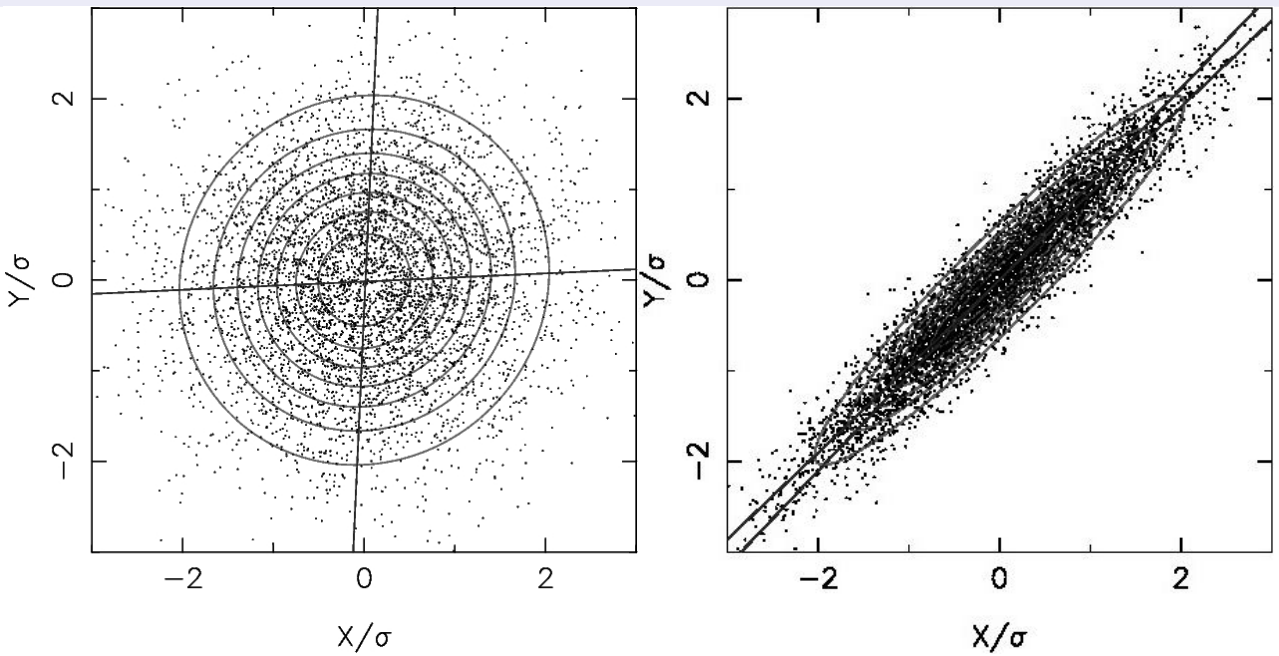
O método dos mínimos quadrados: análise de regressão

- Essa formulação simples do ajuste por mínimos quadrados é apenas a ponta do iceberg. Existe uma enorme variedade de procedimentos para regressão linear via mínimos quadrados. Dentre estes estão as seguintes questões envolvidas na escolha do procedimento:
  - Os dados devem ser ponderados ou não?
  - (Relacionado à anterior) Os dados podem ser tratados igualmente? O mesmo σ_y se aplica a todos ou σ_Y depende de y? O primeiro caso é chamado de *homocedastico* e o segundo de *heterocedastico*.
  - Devemos fazer um ajuste dos mínimos quadrados ordinário de y em função de x (OLS(Y/X)) ou de x em y? Ou ainda algo diferente?
  - Em caso de heterocedasticidade , com incertezas diferentes em cada Y_i e possivelmente em cada X_i, como podemos usar essa informação para estimar as incertezas do ajuste?
  - Os dados foram truncados ou censurados de alguma forma? gostaríamos de colocar limites no nosso ajuste?

O método dos mínimos quadrados: análise de regressão

- Há uma série de artigos de Feigelson e colaboradores (Isobe et al. 90; Babu & Feigelson, 92; Feigelson & Babu 92) que leva todas essas questões em consideração, descreve suas complexidades e indica como encontrar erros a partir das técnicas de reamostragem como *bootstrap* e *jacknife*.
- Recomenda-se a leitura dessas referências ao "ajustador" sério (bem como Sereno 2015), mas nesse ínterim vamos a um exemplo
-

Exemplo W&J P. 137



O método dos mínimos quadrados: análise de regressão

- O ponto do exemplo é que sabemos a resposta de antemão: uma linha com inclinação de 45°. Para uma correlação pequena, ainda que significativa, as linhas OLS erram dramaticamente. Claramente a linha bissetriz chagaria na resposta correta, bem como a linha obtida por regressão ortogonal, que minimiza as distâncias perpendiculares. Para o primeiro caso uns poucos contaminantes rapidamente levariam ao erro. O segundo caso é o PCA.

- Como já foi dito anteriormente, quando a dependência entre as variáveis não for bem conhecida, o PCA é a melhor alternativa. Ele dá o resultado correto nesse exemplo pois diz qual é a relação entre x e y, sem a necessidade de assumir que uma delas é "independente".

- Frequentemente é necessário ajustar uma reta com erros em ambas as variáveis. Isso é surpreendentemente difícil de se fazer, ainda que existam algumas soluções práticas. Veja por exemplo em Gull (1989; *Bayesian data-analysis: straight-line fitting*) ou Zellner (1987), para um tratamento completo.

O Método do mínimo qui-quadrado

- O processo de modelagem clássico na literatura de nossa área é o mínimo qui-quadrado, uma simples extensão do teste de qui-quadrado da qualidade do ajuste descrito anteriormente.
- Veremos ainda que ele está intimamente relacionado com o método dos mínimos quadrados e, de fato, a estatística no mínimo qui-quadrado está relacionada com o método da máxima verossimilhança, para erros distribuídos de modo Normal.
- Considere dados observacionais os quais podem ser (ou já são) "intervalados", e um modelo/hipótese que prediz a população em cada intervalo. Já vimos que a estatística qui-quadrado é:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

  onde O_i corresponde à observação e E_i ao valor esperado, em cada intervalo.
- O método dos mínimos quadrados para ajuste de modelos consiste em encontrar os parâmetros que minimizam a estatística χ^2
- A premissa dessa técnica é de que o modelo está correto e que as diferenças entre dados e modelo são apenas devidas a flutuações estatísticas.

O Método do mínimo qui-quadrado

- Na prática a busca pelos melhores parâmetros é fácil, sempre quando o número de parâmetros não seja maior do que 4. em caso contrário, simples "varreduras" pelos valores dos parâmetros são extremamente ineficientes e métodos mais sofisticados são necessários.

- O número de graus de liberdade associado com o χ^2 para k intervalos e N parâmetros é: $\nu = k - 1 - N$.

- Uma questão essencial é, tendo achado os parâmetros apropriados, é estimar os seus limites de confiança. A resposta (ver *Numerical Recipes*) é que a região de confiança é definida pela zona do espaço de parâmetros na qual

$$\chi^2 = \chi^2_{min} + \Delta(\chi^2),$$

onde o $\Delta(\chi^2)$ vem da tabela abaixo.

Confiança	Número de parâmetros		
c	1	2	3
0.68	1.00	2.30	3.50
0.90	2.71	4.61	6.25
0.99	6.63	9.21	11.30

O Método do mínimo qui-quadrado

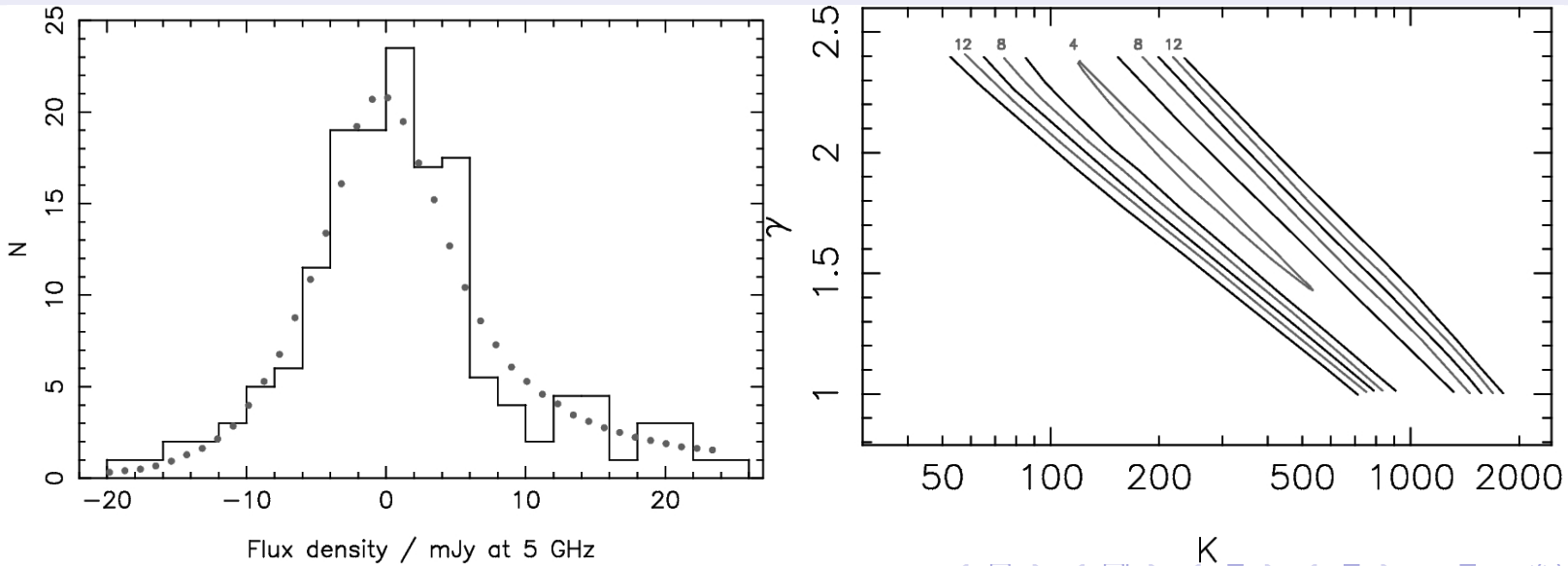
- O significado dessa tabela é que existe uma probabilidade c de que essa dada região vai conter os valores verdadeiros dos parâmetros.
- Isso é calculado via a equação

$$c(\nu, \Delta\chi^2) = P(\nu/2, |\Delta\chi^2/2),$$

onde P é a função gama incompleta.

- É interessante notar que a probabilidade $\Delta(\nu, \chi^2)$ depende apenas do número de parâmetros envolvidos e não na qualidade do ajuste que se conseguiu (via χ_{min})
-

Exemplo W&J P. 140



O Método do mínimo qui-quadrado

- Há três características boas sobre o método do mínimo qui-quadrado e duas más. As boas:
  1. Como o χ^2 é aditivo os resultados advindos de diferentes conjuntos de dados ou que se aplicam a diferentes aspectos do mesmo modelo podem ser usados ao mesmo tempo.
  2. A contribuição do χ^2 em cada intervalo pode ser examinada e regiões de ajuste excepcionalmente bom ou ruim podem ser delineadas.
  3. Você obtém uma medida de qualidade do ajuste "de graça". A tabela B.6 indica probabilidades do χ^2 para graus de liberdade dados. É de se esperar que o modelo tenha ao final uma probabilidade da ordem de 0.5. O pico da distribuição de χ^2 é em $\sim \nu$, quando $\nu > 4$. No exemplo acima, há 7 intervalos e dois parâmetros ($\nu = 7 - 2 - 1 = 4$). O valor de χ_{min} de 4 é o que se esperava. A probabilidade desse valor ser excedido, caso o modelo esteja correto, é de 0.4, de forma que o modelo ótimo é um ajuste satisfatório.
- As ruins:
  1. Baixas populações nas somas de χ^2 causarão grande instabilidade. Como uma "regra do dedão", 80% dos intervalos devem ter $E_i > 5$
  2. Finalmente é importante repetir o mantra: divisão dos dados em intervalos é ruim. Causa perda de resolução e informação. Ainda pior é a possibilidade de resultados com viés isso pode causar. Suponha uma distribuição muito distorcida com poucos dados. A necessidade da divisão por intervalos pode apagar essa distorção inteiramente.
- Fim da aula 10