REGRESSÃO LINEAR MÚLTIPLA

- **MODELO:** $\mathbf{f(x) = \beta_0 + \sum x_j \beta_j}$ **assume que a f.r.** $\mathbf{E(Y|x)}$ **é linear ou aproximadamente linear.**

- **As variáveis** $\mathbf{x_j}$ **podem ser:**

  - ⋆ **quantitativas**
  - ⋆ **transformações tais como** $\mathbf{\log(x_1)}$, $\mathbf{x_3^2}$, $\mathbf{\sqrt{x_2}}$, $\mathbf{x_1^3}$.
  - ⋆ **Codificação numérica ou "dummy" de variáveis categóricas. Uma v.c. com** $\mathbf{L}$ **valores origina** $\mathbf{L}$ **v. binárias.**
  - ⋆ **interações entre varáveis, tais como** $\mathbf{x_1.x_2}$.

- **MMQ: $\hat{\beta}$ minimiza a soma dos quadrados dos erros,**
  $\mathbf{SSE}(\beta) = \sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{y_i} - \mathbf{f(x_i)})^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{y_i} - \beta_{\mathbf{0}} - \sum \mathbf{x_{ij}} \beta_{\mathbf{j}})^{\mathbf{2}} = (\mathbf{Y} - \mathbf{X}\beta)^{\mathbf{T}} (\mathbf{Y} - \mathbf{X}\beta)$

- $\frac{\partial \mathbf{SSE}}{\partial \beta} = -\mathbf{2X^T}(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$ **sse** $\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$

- $\frac{\partial^{\mathbf{2}} \mathbf{SSE}}{\partial \beta \partial \beta^{\mathbf{T}}} = \mathbf{2X^T X}$**. Se** $\mathbf{X^T X}$ **for não singular, é definida positiva.**

- **Os valores previstos para o conjunto de treino são dados por**
  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X^T X})^{-1} \mathbf{X^T Y}$

INTERPRETAÇÃO GEOMÉTRICA

- **Em** $\mathbb{R}^{p+1}$

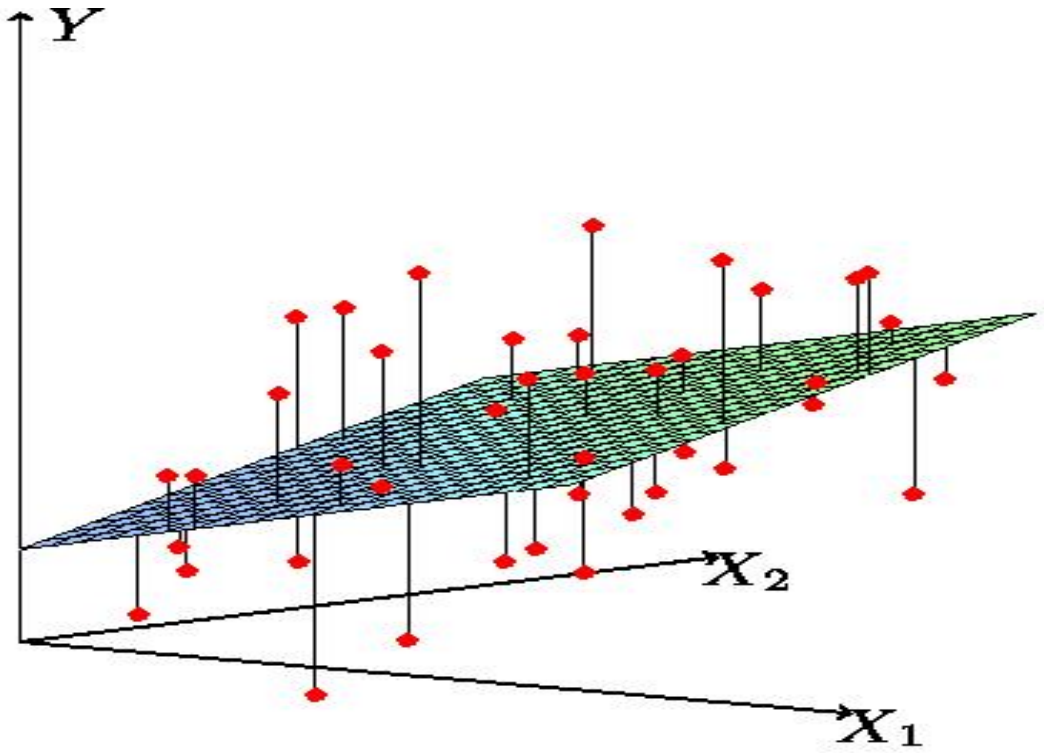

Figure 1: *Linear least squares fitting with* $X \in \mathbb{R}^2$. *We seek the linear function of* X *that minimizes the sum of squared residuals from* Y.

- **Em $\mathbb{R}^N$ $\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_p}$ designam agora as colunas da matriz da dados $\mathbf{X}$ ($\mathbf{x_0} \equiv 1$). Pretende-se minimizar $\| \mathbf{y} - \mathbf{X}\beta \|^2$**

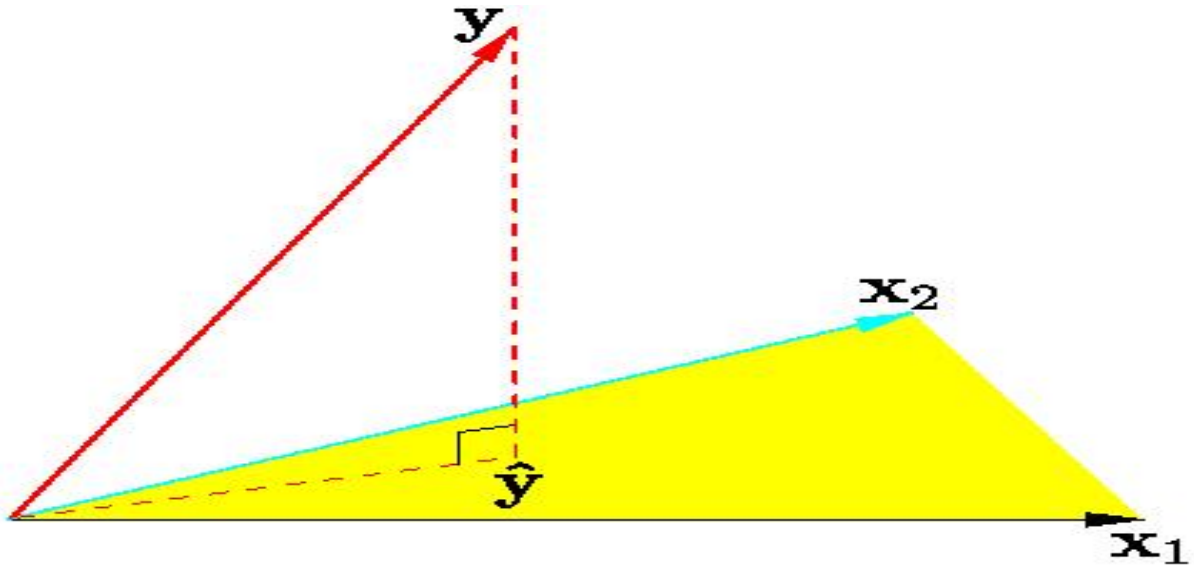

Figure 2: *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x_1}$ *and* $\mathbf{x_2}$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

- **Se as colunas de $\mathbf{X}$ não forem lineares, então $\mathbf{X}^T\mathbf{X}$ é singular. Neste caso convém começar por remover as colunas redundantes.**

INFERÊNCIA

- **Voltemos a $\mathbb{R}^{p+1}$ onde portanto x_i é uma linha da matriz X.**
- **Suponhamos que os Y_i são não correlacionados, $Var(Y_i) = \sigma^2$; para além disso, os x_i são fixos.**
- $Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$
- **Tipicamente usa-se $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$**
- **Suponhamos ainda que de facto a f.r. é linear e que**
- $Y = E(Y|x_1, \ldots, x_p) + \epsilon = \beta_0 + \sum x_j\beta_j + \epsilon$, **em que $\epsilon \sim N(0, \sigma^2)$.**
- **Então:**

- $\hat{\beta} \sim \mathbf{N}(\beta, (\mathbf{X^T X})^{-1}\sigma^2)$
- $\frac{(\mathbf{N-p-1})\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\mathbf{N-p-1}}$
- $\hat{\beta}$ **e** $\hat{\sigma}^2$ **são independentes**

- **Seja** $\mathbf{H_0} : \beta_\mathbf{j} = \mathbf{0}$ **vs** $\mathbf{H_1} : \beta_\mathbf{j} \neq \mathbf{0}$.
- **Sob** $\mathbf{H_0}$, $\hat{\beta}_\mathbf{j} \sim \mathbf{N}(0, \nu_\mathbf{j}\sigma^2)$ **em que** $\nu_\mathbf{j}$ **é o j-ésimo elemento da diagonal de** $(\mathbf{X^T X})^{-1}$.
- **Logo,** $\frac{\hat{\beta}_\mathbf{j}}{\sigma\sqrt{\nu_\mathbf{j}}} \sim \mathbf{N}(0,1)$ **e** $\frac{\hat{\beta}_\mathbf{j}}{\hat{\sigma}\sqrt{\nu_\mathbf{j}}} \sim \mathbf{t_{N-p-1}}$
- **Podemos agora testar** $\mathbf{H_0}$ **ou construir intervalos de confiança:** $\hat{\beta}_j \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\nu_j}$
- **Podemos também querer testar se um grupo de variáveis é insignificante, como por exemplo no caso das variáveis binárias correspondentes a uma var. categórica:**

- $\mathbf{H_0 : \beta_1 = \beta_2 =, \ldots, = \beta_k = 0.}$

- **Seja** $\mathbf{F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(N - p_1 - p_0)}}$

- **Sob as condições anteriores e a condição adicional de que o modelo "menor" é o correcto,** $\mathbf{F \sim F_{p_1 - p_0, N - p_1 - 1}}$

- **Para** $\mathbf{N}$ **grande,** $\mathbf{F_{p_1 - p_0, N - p_1 - 1} \longrightarrow \chi^2_{p_1 - p_0}}$

- **Podemos também obter uma região de confiança para o vector** β**:**

- $\mathbf{C_\beta = \left\{ \beta : (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi^2_{p+1}(1 - \alpha) \right\}}$

- **Isto origina um I.C. para a função** $\mathbf{f(x) = x^T \beta}$**, que é**

- $\mathbf{\left\{ x^T \beta : \beta \in C_\beta \right\}}$

- **EXEMPLO: Prever o nivel de antigénio da próstata a partir das variáveis predictivas:**

  - ⋆ **lcavol, lweight (logaritmo do volume de cancro e do peso da próstata)**
  - ⋆ **age, lbph,svi,lcp,gleason,pgg45.**

- **Existem muitas correlações fortes entre as v. predictivas.**

- **Após estandardização das v.p., e divisão dos dados em 67 para treino e 30 para teste, aplicou-se o modelo linear.**

- **Obteve-se o valor de Z Score** $= \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\nu_j}}$ **para cada v.pred. Um valor maior do que 2 em valor absoluto é aprox. sig. a 95%.**

Var.	Intercept	lcavol	lweight	age
Z Score	27.66	5.37	2.75	-1.40

lbph	svi	lcp	gleason	pgg45
2.06	2.47	-1.87	-0.15	1.74

- **Considerou-se de seguida retirar do modelo as var. não sig. age, lcp, gleason e pgg45. Obteve-se**

- $\mathbf{F = \frac{(32.81-29.43)/(9-5)}{29.43/(67-9)} = 1.67}$

- $\mathbf{P(F_{4,58} > 1.67) = 0.17}$ **e portanto não é significativo.**

- **O erro médio no teste é de 0.545. Se usarmos o valor médio de** $\mathbf{Y}$ **para prever, o erro é de 1.050. Portanto o modelo linear reduz o erro base em cerca de 50%.**

TEOREMA DE GAUSS-MARKOV: De todos os estimadores lineares centrados de β, o EMMQ é o que tem menor variância.

- **Seja $\theta = \mathbf{a^T}\beta$ e $\hat{\theta} = \mathbf{a^T}\hat{\beta} = \mathbf{a^T}(\mathbf{X^T X})^{-1}\mathbf{X^T Y}$ o EMMQ de θ.**
- **$\hat{\theta}$ é uma função linear de Y. $\mathbf{Var}(\mathbf{a^T}\hat{\beta}) \leq \mathbf{Var}(\mathbf{c^T y}), \forall \mathbf{c} : \mathbf{E}\{\mathbf{c^T y}\} = \theta$.**
- **Seja $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$. O EMMQ de β é o que tem menor variância entre os estimadores lineares centrados de β:**
- **Se $\hat{\mathbf{V}}$ é a matriz de variâncias-covariâncias de $\hat{\beta}$ e $\tilde{\mathbf{V}}$ a de outro estimador linear centrado, então:**
- **$\hat{\mathbf{V}} \preceq \tilde{\mathbf{V}}$, isto é, $\tilde{\mathbf{V}} - \hat{\mathbf{V}}$ é semi-definida positiva.**
- **$\mathbf{MSE}(\tilde{\theta}) = \mathbf{E}(\tilde{\theta} - \theta)^2 = \mathbf{Var}(\tilde{\theta}) + [\mathbf{E}(\tilde{\theta}) - \theta]^2 = \mathbf{Var}(\tilde{\theta}) + \mathbf{Vies}^2(\tilde{\theta})$.**
- **De todos os estimadores sem viés e lineares, os obtidos pelo MMQ minimizam MSE.**

- **No entanto, pode haver estimadores viesados com menor MSE, como é o caso de "Ridge Regression" e outros métodos que veremos adiante.**

- **NOTA:** $\mathbf{EPE} = \sigma^2 + \mathbf{MSE}$

- **Decomposição** $\mathbf{QR}$ **da matriz de dados** $\mathbf{X}$**:**

$$\mathbf{X} = \mathbf{QR}$$

- $\mathbf{Q_{Nx(p+1)}}$ **é ortogonal;** $\mathbf{Q^T Q = I}$**;** $\mathbf{R_{(p+1)x(p+1)}}$ **é triangular superior.**

- **A solução do MMQ é** $\hat{\beta} = \mathbf{R^{-1} Q^T y}$ **e** $\hat{\mathbf{y}} = \mathbf{Q Q^T y}$**.**

SELECÇÃO DE UM SUBCONJUNTO DE VARIÁVEIS

- **Quando temos muitas var. o viés diminui, mas a variância aumenta. Pode haver um menor sub. com EPE menor.**
- **Um modelo com menos variáveis é mais interpretável**
- **MELHOR SUBCONJUNTO DE TODOS:**
  - ⋆ $\forall \mathbf{k} \in \{0, 1, \ldots, \mathbf{p}\}$ **escolher o subconjunto de k var. que minimiza SSE.**
  - ⋆ **Em 1974, Furnival e Wilson forneceram um algoritmo que torna este processo eficiente para p até 30 ou 40.**

- **Para o exemplo dos dados da próstata:**

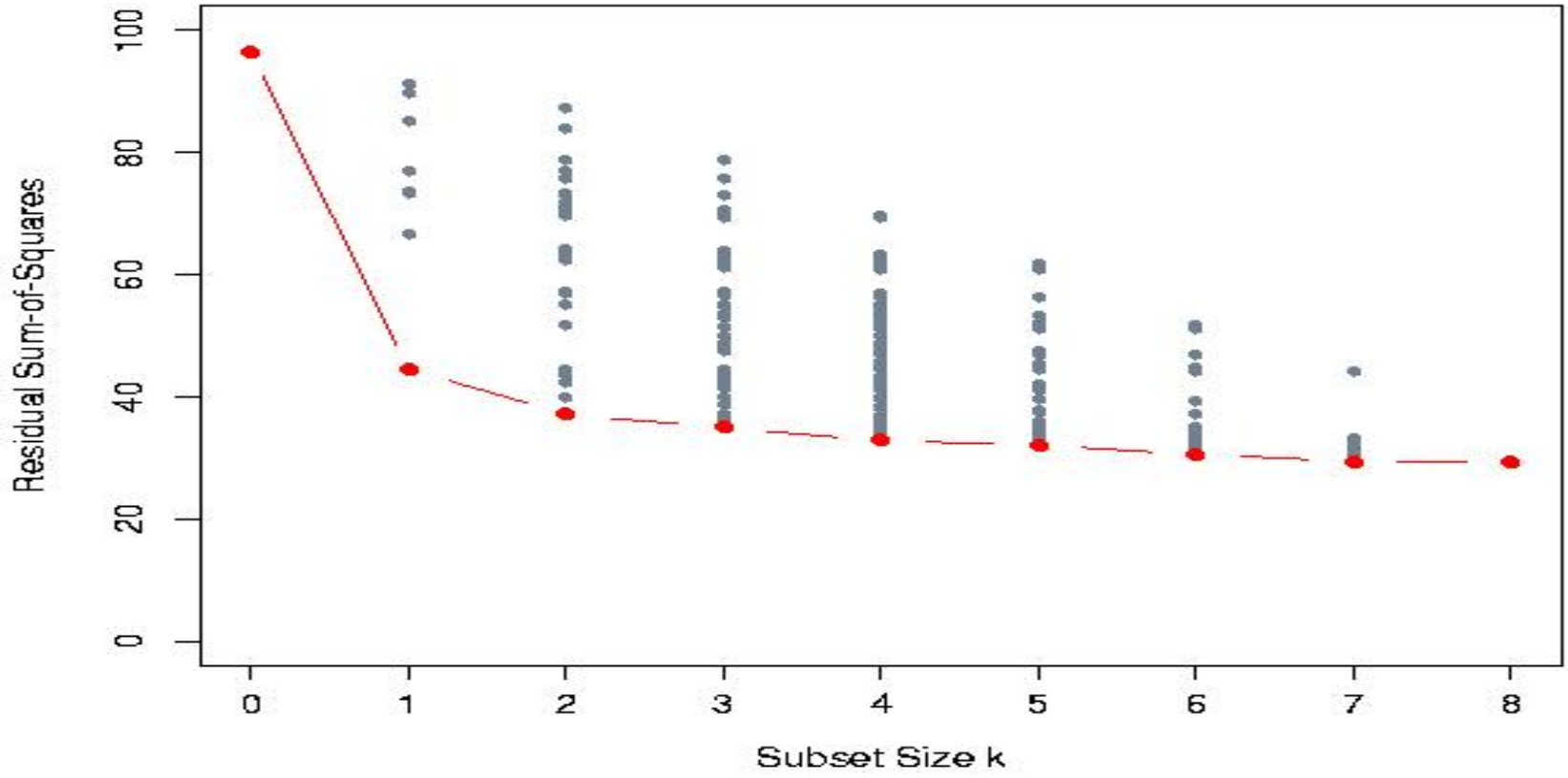

Figure 5: *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

- **A curva a vermelho decresce necessáriamente. Para o melhor modelo geralmente escolhe-se o que minimiza uma estimativa de EPE.**

- **Para $\mathbf{p > 40}$ é comum usar-se :**

  - **SELECÇÃO PASSO A PASSO PARA A FRENTE (FORWARD STEPWISE SELECTION):**

    - **Começa-se com β_0 e adiciona-se seq. a var. que mais contribui para a melhoria do modelo, isto é:**

    - **Seja $\hat{\beta}$ o vector de est. com k comp. antes e $\tilde{\beta}$ o vector depois.**

    - **Seja $\mathbf{F} = \frac{\mathbf{SSE}(\hat{\beta}) - \mathbf{SSE}(\tilde{\beta})}{\mathbf{SSE}(\tilde{\beta})/(\mathbf{N-k-2})}$**

    - **Adiciona-se a var. que maximiza $\mathbf{F}$, parando se esse valor é menor que o percentil 90 ou 95 de $\mathbf{F_{1,N-k-2}}$.**

- **SELECÇÃO PASSO A PASSO PARA TRÁS (BACKWARD STEPWISE SELECTION):**
  - **Aqui começa-se com o modelo completo e retira-se seq. variáveis.**
  - **A cada passo retira-se a var. que minimiza o valor de F e pára-se quando esse valor é maior do que o percentil 90 ou 95 de $\mathbf{F_{1,N-k-2}}$.**
- **Existem também procedimentos mistos (BACKWARD AND FORWARD). Aqui, em cada passo escolhe-se o melhor movimento (frente ou tràs). Isto requer um par. adicional para decidir quando é melhor adicionar ou retirar.**

- **No final temos uma seq. de modelos e podemos escolher aquele que minimiza por exemplo uma est. de EPE.**

MÉTODOS DE CONTRACÇÃO (SHRINKAGE METHODS)

- **Nos métodos anteriores, uma var. ou entrava no modelo ou tinha coef. 0. Vamos agora ver métodos mais "contínuos".**

- **Começemos por analisar o estimador dos mínimos quadrados (MMQ) que temos usado, $\hat{\beta}^{MMQ}$:**

- **O estimador $\hat{\beta}^{MMQ}$ tem boas propriedades (Gauss-Markov, EMV). MAS, podemos arranjar um melhor?**

- **Para ver se um novo estimador $\hat{\beta}$ é um bom candidato, temos de responder às duas questões seguintes:**

  1. **$\hat{\beta}$ está próximo do verdadeiro β ?**
  2. **$\hat{Y} = x^t\hat{\beta}$ é bom a prever o valor de Y de futuras observações ?**

- **O problema com o $\hat{\beta}^{MMQ} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^{MMQ}$ é que alguns dos $\hat{\beta}_j$ podem "explodir": se duas variáveis são muito correlacionadas, os coeficientes que lhes correspondem podem ser muito grandes, e com sinal contrário, anulando-se. Isso pode influenciar bastante o erro de previsão.**

- **RIDGE REGRESSION**

  - **Impõe uma penalização ao tamanho dos coef.**

  - $\hat{\beta}^{\mathbf{ridge}} = \mathbf{argmin}_{\beta}$
    $\left\{\sum_{i=1}^{N}(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right\}$

  - **Ou, de forma equivalente:**

  - $\hat{\beta}^{\mathbf{ridge}} = \mathbf{argmin}_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{Y_i} - \beta_0 - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{x_{ij}}\beta_j)^2$,
    sujeito a $\sum_{j=1}^{p} \beta_j^2 \leq s$

  - **Para var. muito correlacionadas dá bons resultados.**
  - **Deve-se começar por estandardizar as var. predictivas.**

- $\mathbf{SSE}(\lambda) = (\mathbf{Y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^{\mathbf{T}}\beta$

- **A solução é** $\hat{\beta}^{\mathbf{ridge}} = (\mathbf{X^T X} + \lambda\mathbf{I})^{-1}\mathbf{X^T Y}$

- **Foi desta forma que o método foi introduzido em 1970 por Hoerl & Kennard.**

- **A decomposição SVD da matriz** $\mathbf{X_{Nxp}}$ **é**

$$\mathbf{X = UDV^T}$$

- $\mathbf{U_{Nxp}}$ **e** $\mathbf{V_{pxp}}$ **são ortogonais.**
  $\mathbf{D_{pxp}}$ **é diagonal com elementos** $\mathbf{d_1 \geq d_2 \geq \ldots \geq d_p}$**,**
  chamados os valores próprios de $\mathbf{X}$**.**

- **Se usarmos a dec. SVD na reg. lin. vem**

$\hat{\beta}^{\mathbf{LS}} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$ **e que** $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}^{\mathbf{LS}} = \mathbf{UU^T Y}$**.**

- **A solução é semelhante se usarmos a dec. QR:**

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}^{\mathbf{LS}} = \mathbf{Q}\mathbf{Q}^{\mathbf{T}}\mathbf{Y}.$$

- **$\mathbf{Q}$ e $\mathbf{U}$ são geralmente bases diferentes.**

- **Em RIDGE REGRESSION**

$$\mathbf{X}\hat{\beta}^{\mathbf{ridge}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} = \mathbf{U}\mathbf{D}(\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^{\mathbf{T}}\mathbf{Y}$$
$$= \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{p}} \mathbf{u_j} \frac{\mathbf{d_j^2}}{\mathbf{d_j^2}+\lambda} \mathbf{u_j^T}\mathbf{Y}$$

- **Portanto este método, contrariamente à reg. linear, "encolhe" as cloordenadas na nova base.**

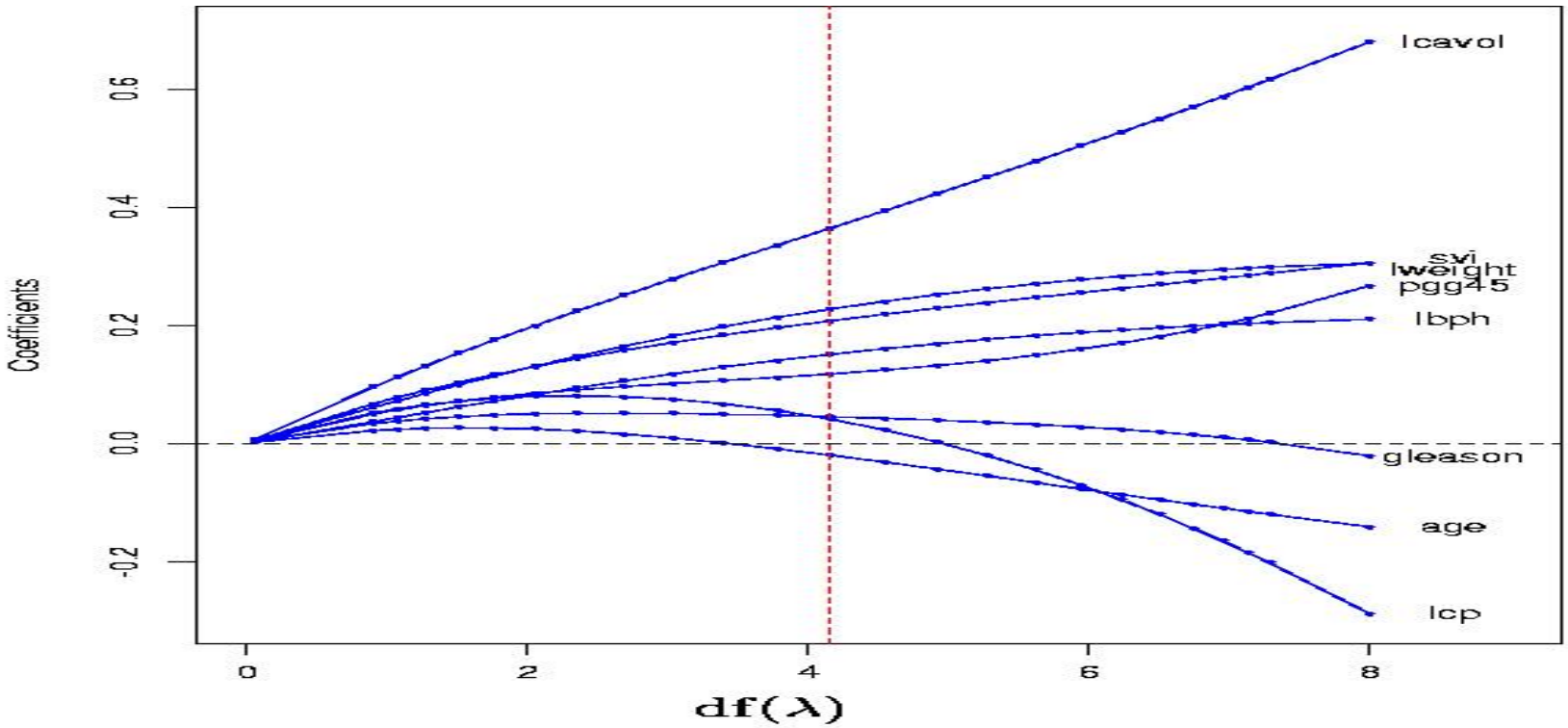

Figure 7: *Profiles of ridge coefficients for the prostate cancer example, as tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 4.16$, the value chosen by cross-validation.*

- $\mathbf{X} = \mathbf{UDV^T}$, **logo,** $\mathbf{X^TX} = \mathbf{VD^2V^T}$
- $\mathbf{v_j}$ **são as direcções principais de** $\mathbf{X}$.

- **A 1a com. principal, $\mathbf{Z_1 = Xv_1}$, é a que tem variância máxima.**

- $\mathbf{Var(Z_i) = Var(Xv_i) = \frac{d_i^2}{N}}$
- **As últimas c.p. tem menor variância. Ridge Regression encolhe as suas coordenadas mais do que as outras.**

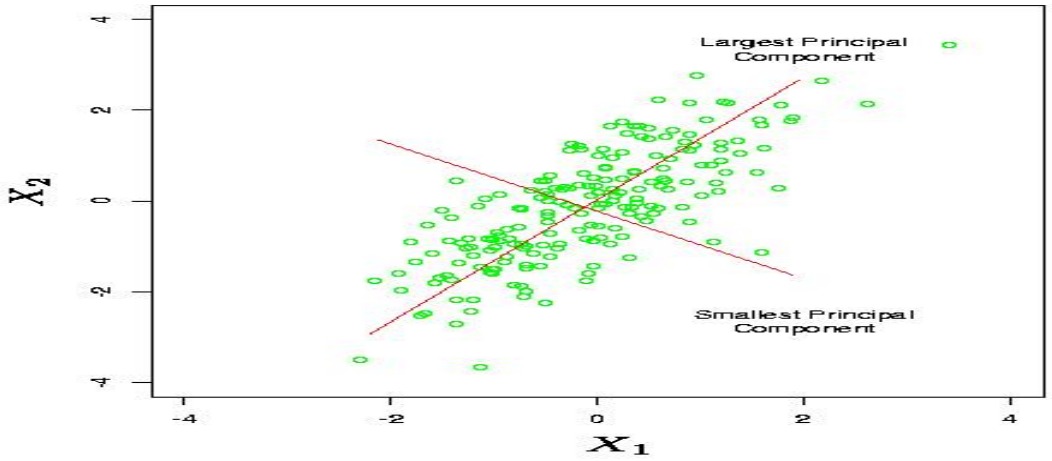

Figure 8: *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

- **Método LASSO (Tibshirani 1996)**

  - $\hat{\beta}^{\mathbf{ridge}} = \mathbf{argmin}_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{Y_i} - \beta_{\mathbf{0}} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{x_{ij}} \beta_{\mathbf{j}})^{\mathbf{2}}$,

    sujeito a $\sum_{j=1}^{p} |\beta_j| \leq s$

- **A solução é não linear nos $\mathbf{Y_i}$.**

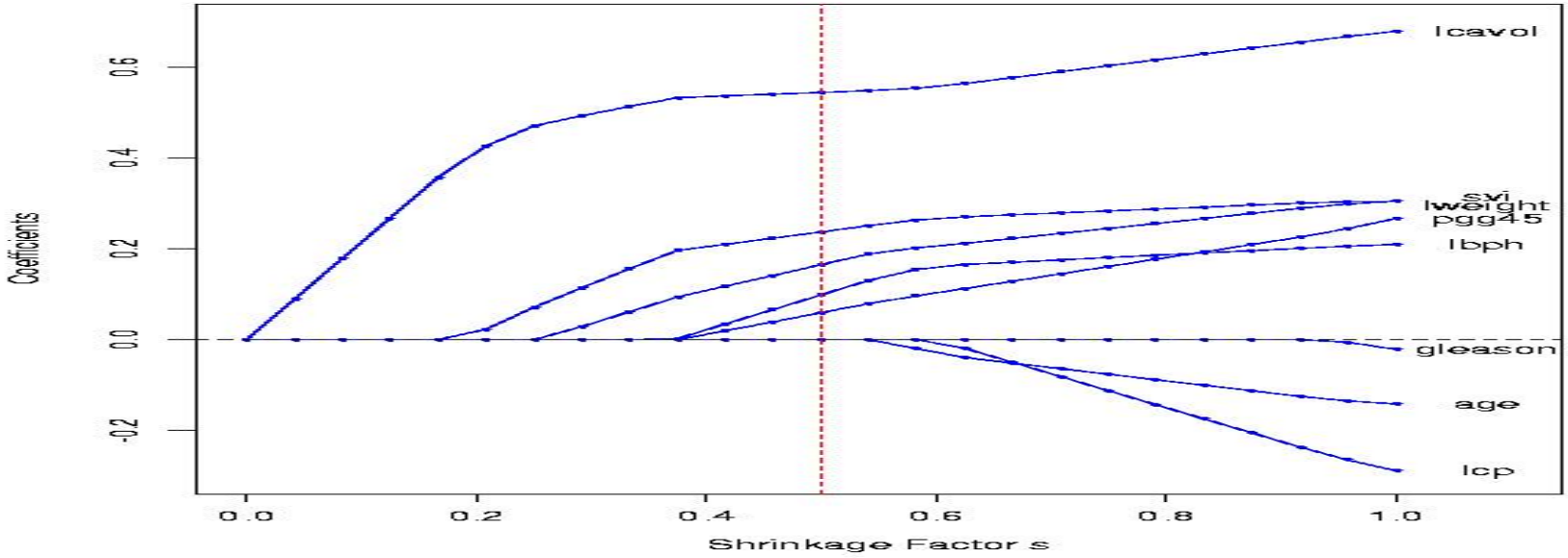

Figure 9: *Profiles of lasso coefficients, as tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.5$, the value chosen by cross-validation. Compare Figure 7 on page 7; the lasso profiles hit zero, while those for ridge do not.*

REGRESSÃO COM COMPONENTES PRINCIPAIS

- **Em muitas situações temos um grande número de v. pred. e freq. bastante correlacionadas.**
- **Como as comp. p. $\mathbf{Z_m}$ são ortogonais a reg. é uma soma de reg. univariadas:**
- $\hat{\mathbf{Y}}^{\mathbf{pcr}} = \bar{\mathbf{Y}} + \sum_{\mathbf{m=1}}^{\mathbf{M}} \hat{\theta}_{\mathbf{m}} \mathbf{Z_m}$ **em que** $\hat{\theta}_{\mathbf{m}} = \frac{<\mathbf{Z_m}, \mathbf{Y}>}{<\mathbf{Z_m}, \mathbf{Z_m}>}$
- **Como em Ridge Reg. é comum estandardizar-se os inputs primeiro.**
- **PCR é muito semelhante a Ridge Regression, pois ambas utilizam as comp. principais. Ridge Reg. encolhe os coef. nessas componentes enquanto que PCR deita fora as $\mathrm{p} - \mathrm{M}$ componentes menos importantes.**

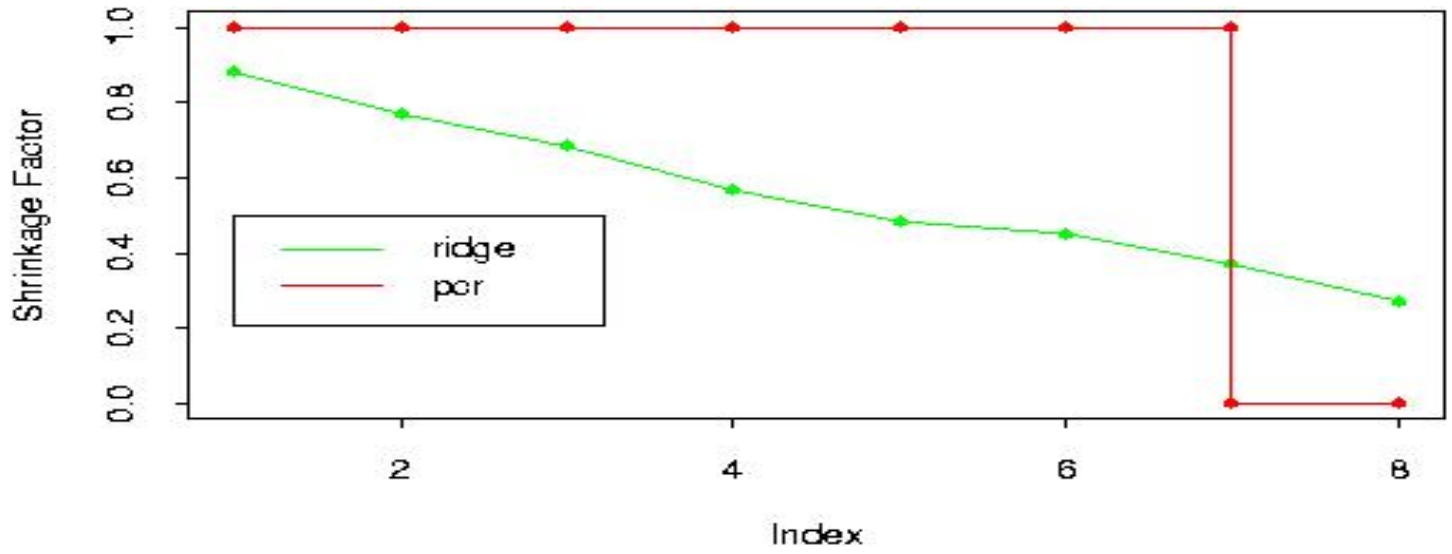

Figure 10: *Ridge regression shrinks the regression coefficients of the principal components, using shrinkage factors $d_j^2/(d_j^2+\lambda)$ as in (3.47). Principal component regression truncates them. Shown are the shrinkage and truncation patterns corresponding to Figure 6, as a function of the principal component index.*