Linear Model Selection Regularization

Machine Learning

Prof. Neylson Crepalde

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Na unidade anterior, discutimos algumas possibilidades de estimação que vão além dos modelos de regressão linear. Vamos discutir nesta unidade algumas maneiras pelas quais os modelos lineares podem ser melhorados substituindo o método dos mínimos quadrados ordinários por algum outro procedimento de estimação.

Por quê, entretanto, nós gostaríamos de tentar modelos com métodos diferentes dos mínimos quadrados?

- *Acurácia nas predições* - métodos de *shrinkage* permitem reduzir a variância a um custo desprezível de aumento de viés;
- *Interpretabilidade dos modelos* - métodos de *feature selection* permitem selecionar as variáveis mais importantes reduzindo a complexidade e aumentando a interpretabilidade dos modelos.

1. **Subset selection**
2. **Shrinkage**
3. **Dimension reduction**

Best subset selection

No *best subset selection*, estimamos um modelo linear para todas as combinações possíveis de cada um dos *p* preditores (2^p)

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

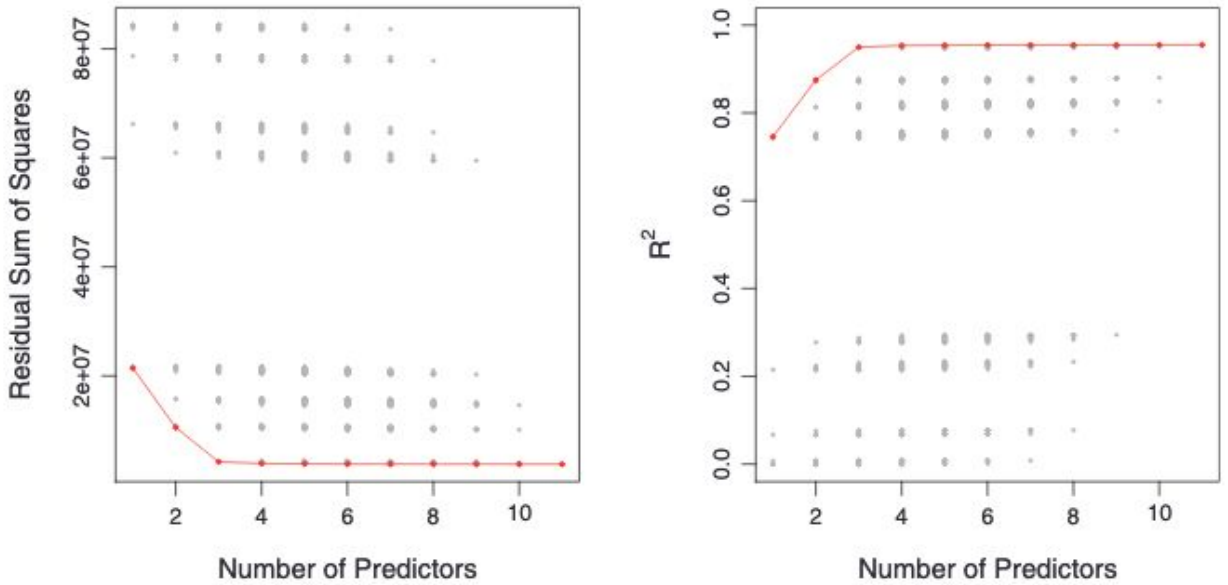

FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the x-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Obviamente, este método é computacionalmente custoso. A partir de 20 preditores serão mais de 1 milhão de modelos para treinar. Será um procedimento lento realizar o *best subset selection* com cerca de 40 variáveis até mesmo numa máquina potente.

Há algumas alternativas computacionalmente menos custosas do que o *Best Subset Selection.*

Forward Stepwise selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

O custo computacional do *Forward stepwise selection* é de

$$1+p(p+1)/2:$$

Desse modo, enquanto um *Best subset* com 20 preditores precisaria de 1.048.576 modelos estimados, o *Forward stepwise* precisa de apenas 211 modelos (!!!).

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

Backwards Stepwise selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Como escolher o melhor modelo?

É necessário verificar o modelo que possui o melhor ajuste para os dados de *teste!!!* Há duas abordagens possíveis

1. Podemos *ajustar* o erro de treino para contabilizar o viés de *overfitting* (usamos Cp, AIC, BIC, e R2 ajustado)

2. Podemos diretamente estimar o erro de teste usando a abordagem de um *validation set* ou a técnica do *cross-validation.*

Para d preditores,

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

$$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$ (Não tão recomendável)

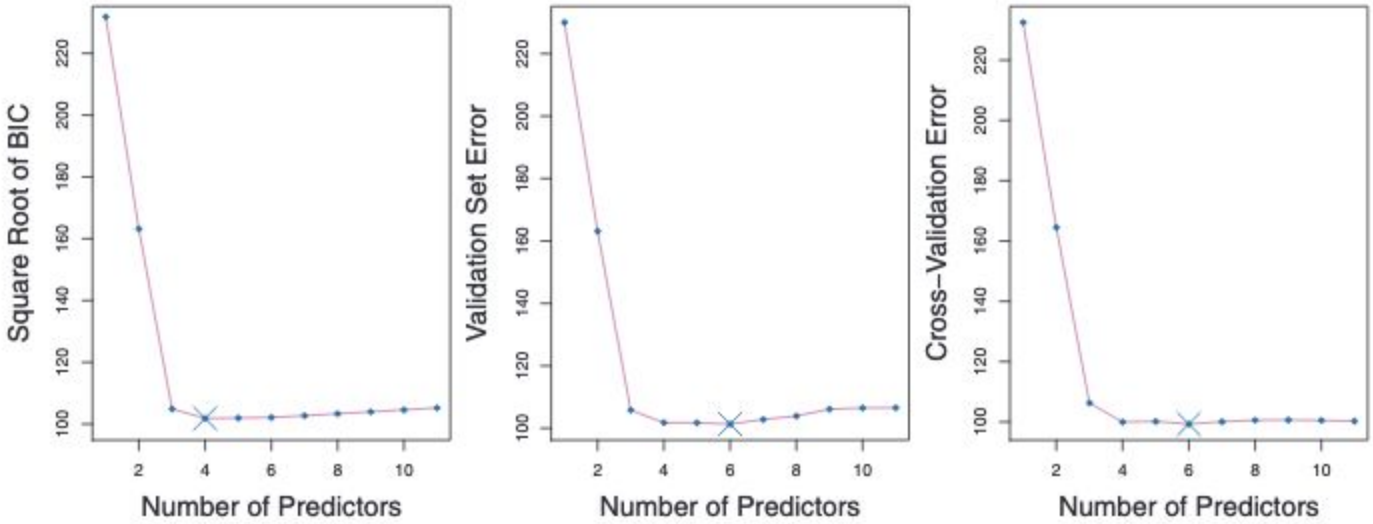

FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Exercício!

Shrinkage

Os métodos de *shrinkage* usam um parâmetro penalizador para *regularizar* os coeficientes da regressão ou diminuí-los tendendo a zero. Esse métodos podem melhorar a estimação reduzindo a variância dos coeficientes.

As duas técnicas mais conhecidas são a *Ridge Regression* e o *LASSO.*

Ridge Regression

A Ridge Regression adiciona um parâmetro penalizador à soma dos quadrados dos resíduos da seguinte forma:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

onde *lambda* é o **parâmetro de ajuste (*tuning parameter*)** e a parte seguinte é chamada de ***shrinkage penalty*.** O objetivo do modelo é reduzir RSS. Desse modo, quanto mais próximos de 0 forem os coeficientes, melhor o ajuste do modelo. O parâmetro de ajuste controla o impacto dos termos na regressão. Se *lambda* = 0, o penalizador não terá nenhum efeito e os coeficientes serão os mesmos da regressão. Quando *lambda* tende ao infinito, os coeficientes tendem a zero.

A escolha do valor desse parâmetro não é algo trivial; pode ser feita por cross-validation.

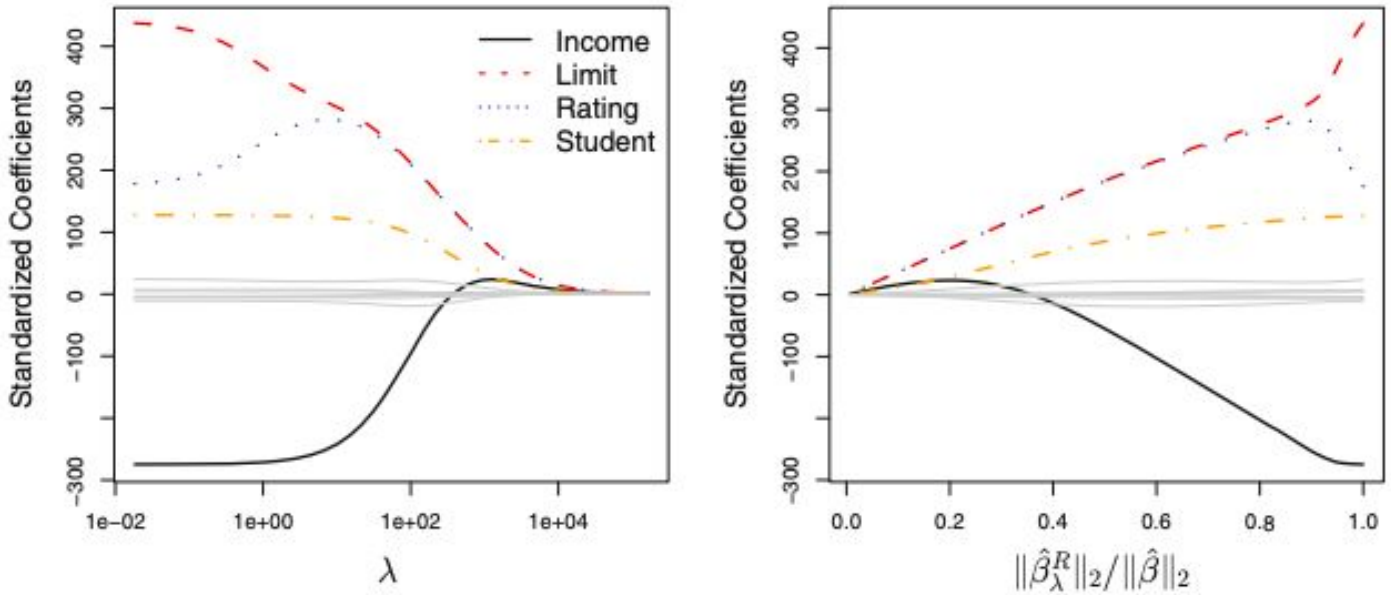

FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

ATENÇÃO!!!

Os resultados da *Ridge Regression* são sensíveis à escala das variáveis preditoras. Desse modo, é importante **padronizar** os dados antes de aplicar a técnica.

LASSO

A Ridge Regression possui uma desvantagem em relação aos métodos de seleção de *features* estudados anteriormente: muito embora os coeficientes sejam penalizados, o modelo final terá sempre todos os preditores porque os coeficientes penalizados tendem a zero mas nunca chegam a zero.

A Ridge Regression utiliza um parâmetro penalizador conhecido como "L2". Já LASSO utiliza o mesmo princípio mas com o parâmetro penalizador "L1".

O *LASSO* penaliza os coeficientes da seguinte forma:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Ciências de Dados

Nesse caso os coeficientes chegarão de fato a zero quando *lambda* tende ao infinito. LASSO opera, portanto, *shrinkage* e *feature selection.*

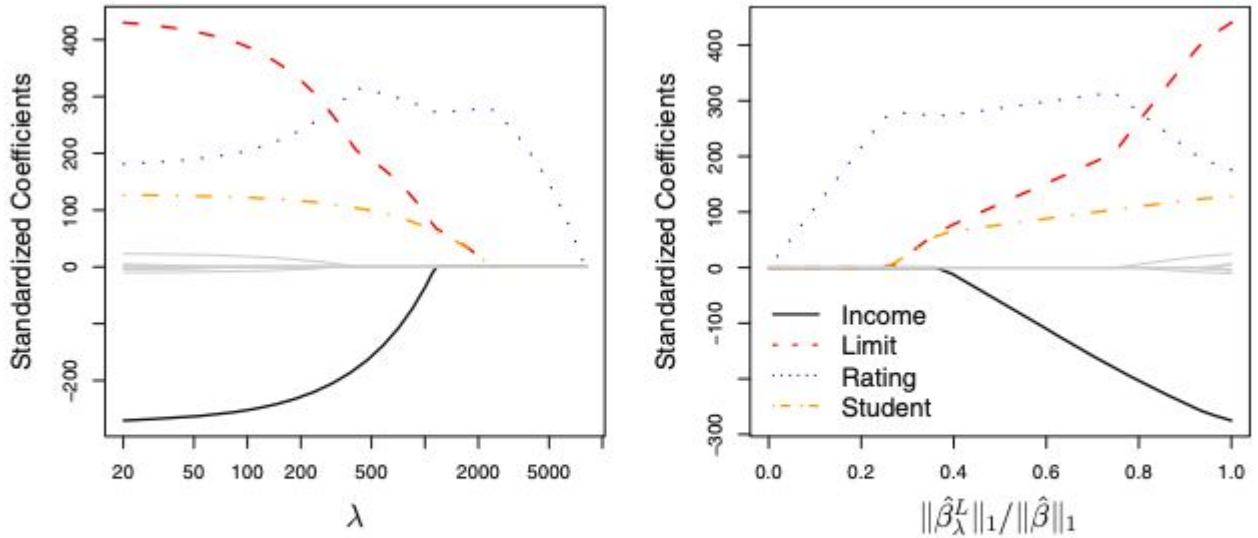

FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

Dimension Reduction

Uma outra forma de controlar a variância dos modelos de ML está relacionada às técnicas de **redução de dimensionalidade**. Ao invés de selecionar um subset de *features* ou de regularizar os coeficientes, podemos construir variáveis latentes que agreguem a informação de várias outras.

Uma das técnicas mais utilizadas para esse fim é a **Principal Components Analysis**.

Principal Components Regression

Condição:

- Variáveis **numéricas** ou **categóricas ordinais**

O PCA cria variáveis latentes, construtos, a partir de combinações lineares das variáveis originais como na equação abaixo:

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}})$$

0.839 e 0.544 são os *loadings* do componente principal Z1 gerado.

É possível construir p-1 componentes principais. Utilizamos a quantidade de componentes que dá um melhor *trade-off* entre quantidade de preditores e variância explicada. Utilizamos as novas variáveis criadas no modelo de regressão.

Partial Least Squares

A *Partial Least Squares* possui abordagem semelhante ao PCA. Ela produz um set de features que são combinações lineares dos *p* preditores mas também leva em conta a variável dependente *y*. Os componentes são portanto gerados de uma maneira supervisionada.

De uma maneira geral, a abordagem do PLS tenta identificar direções que ajudam a explicar tanto a resposta *y* quanto os preditores.

A PLS é computada da seguinte maneira:

Após padronizar os *p* preditores, PLS computa a primeira direção Z1 definindo cada φj1 como o coeficiente de uma regressão linear simples de Y sobre Xj. Pode-se mostrar que esse coeficiente é proporcional à correlação entre Y e Xj. Portanto, ao computar $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, PLS coloca o máximo peso nas variáveis que são mais fortemente correlacionadas com a resposta.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

O parâmetro *m* correspondente ao número de direções usadas na PLS é um *tuning parameter* tipicamente escolhido por cross-validation.

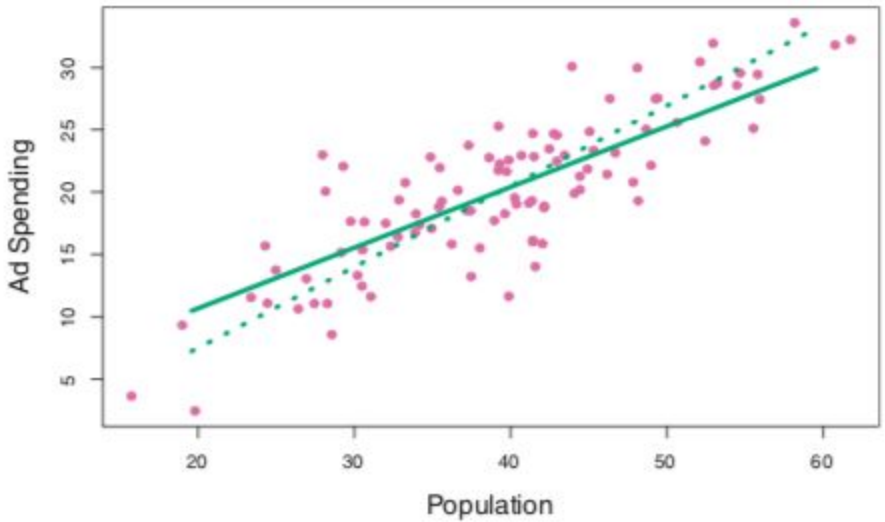

FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

A maldição da dimensionalidade

As técnicas estatísticas mais tradicionais foram desenvolvidas para situações de *baixa dimensionalidade* onde *n,* o número de observações, é bem maior do que *p,* o número de preditores. Com os avanços tecnológicos, há situações onde *p* é maior do que *n.* Por exemplo:

1. Ao invés de predizer a pressão com base em apenas idade, gênero e afins, pode-se também coletar medidas para meio milhão de *single nucleotide polymorphisms* (SNPs; mutações individuais de DNA relativamente comuns na população) para inclusão em modelos preditivos. Nesses casos n ≈ 200 and p ≈ 500,000.
2. Um analista de marketing interessado em entender os padrões de compra online das pessoas pode tratar como *features* todos os termos de pesquisa utilizados pelos usuários num mecanismo de busca. Isso é conhecido como o modelo "bag-of-words". O mesmo pesquisador pode ter acesso ao histórico de consultas de apenas algumas centenas ou milhares de usuários que consentiram em compartilhar suas informações. Para um dado usuário, cada uma das *p* buscas está codificada como ausente (0) ou presente (1) criando um vetor binário bastante grande. Portanto *n* ≈ 1,000 e *p* é muito maior.

O que pode dar errado em alta dimensionalidade?

Quando o número *p* de *features* é tão grande quanto ou maior que o número *n* de observações, OLS não pode (ou não deveria) ser aplicado. A razão é simples: independente se a relação entre os preditores e a resposta é ou não verdadeira, OLS sempre vai dar uma estimativa de coeficientes que resulta em um ajuste perfeito dos dados (*overfitting*) tal que os resíduos sejam iguais a zero.

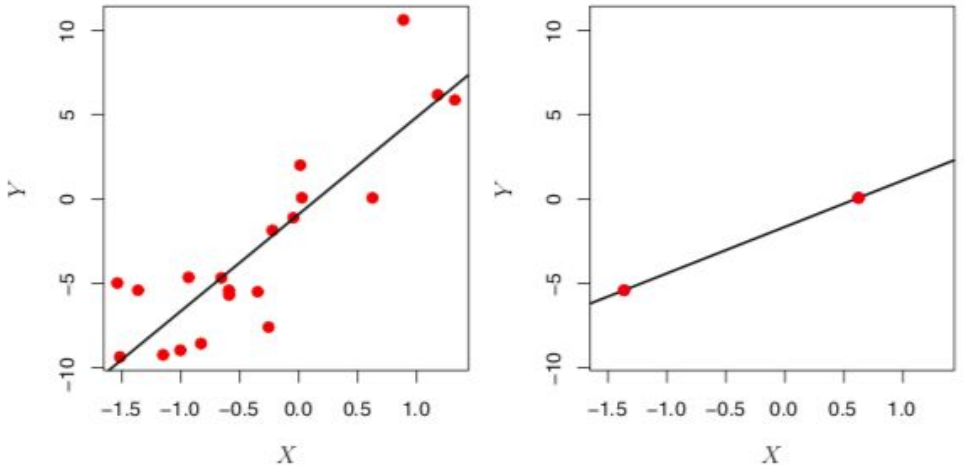

FIGURE 6.22. Left: *Least squares regression in the low-dimensional setting.* Right: *Least squares regression with* $n = 2$ *observations and two parameters to be estimated (an intercept and a coefficient).*

Possíveis soluções

- Best subset selection;
- Stepwise;
- Ridge Regression;
- LASSO;
- PCR;
- PLS;
- etc...

Atividade avaliativa!!!

Ciências de Dados