Introdução a Ciências de Dados

Aula 5: Modelos de regressão

Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br

Aula 5: Modelos de regressão

- Regressão Linear,
- Simplificando Modelos via Regularização,
- Avaliando e Interpretando modelos.

Modelos de regressão

Dado um conjunto de variáveis, como podemos prever o valor de outra variável?

Exemplos:

- A partir de dados relacionados à renda, idade, profissão, e estado civil, podemos prever o quanto o cliente gasta por mês em supermercados?
- Considerando dados de temperatura, umidade, nível pluviométrico, latitude, longitude e altitude, podemos prever o número de casos de dengue em uma cidade?
- Qual dessas variáveis mais influenciam o número de casos de dengue?

Modelos de regressão

- No processo de regressão, temos um conjunto de atributos que serão usados para prever uma variável de saída (resposta).
- O modelo de regressão é descrito por: ($y \in \mathbb{R}$)

$$y = f(X, \theta) + \epsilon$$

- Onde ϵ é o erro (ou ruído), que possui média 0 e variância σ^2 ￼, descrevendo o que não pode ser capturado pelo modelo.

Regressão linear

Modelos de regressão podem ser usados principalmente para duas tarefas:

- Prever dados desconhecidos a partir do modelo treinado.
- Determinar a importância de cada variável independente na previsão.

O modelo de regressão linear é um dos mais usados na literatura:

Variáveis independentes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d$$

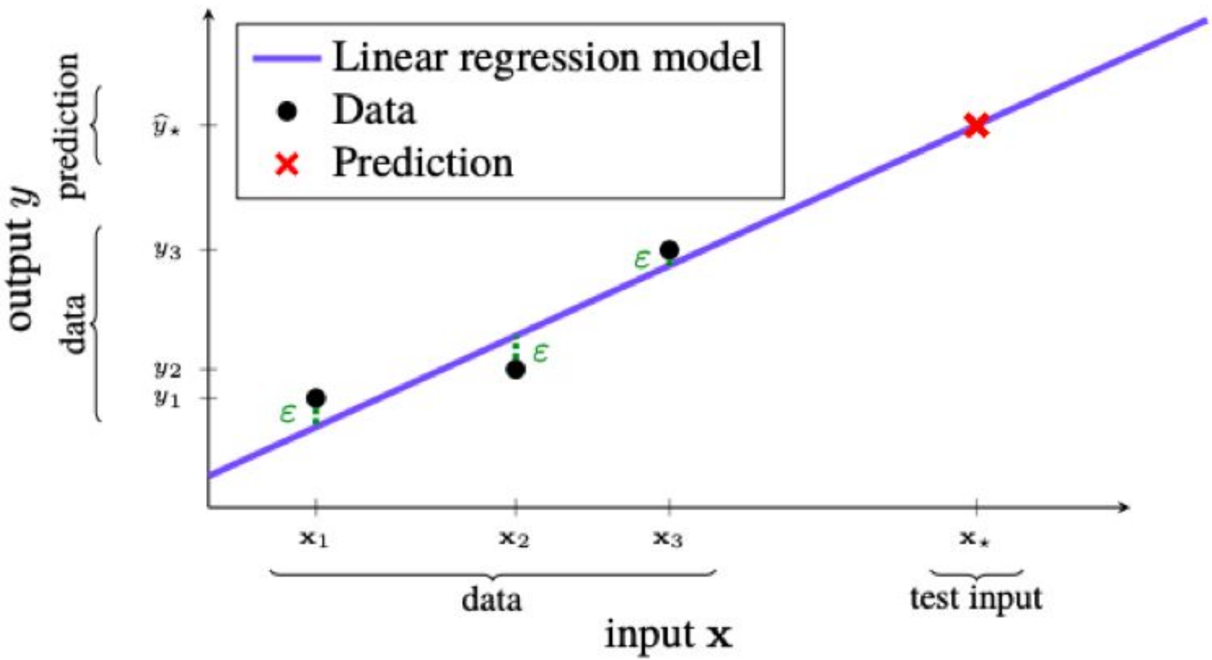
Variável resposta (saída)

Parâmetros

Regressão Linear Simples

Regressão linear simples

$$Y \approx \beta_0 + \beta_1 X$$



Regressão linear simples

Vamos considerar um exemplo.

- Uma empresa está interessada em determinar quanto o tempo de exposição na televisão ajuda no aumento das vendas.
- São coletados dados do volume de recursos gastos com exposição na TV e vendas.
- A partir desses dados, podemos analisar como a campanha publicitária influência nas vendas.

Regressão linear simples

Perguntas que podemos responder com o modelo de regressão:

- Há alguma relação entre os recursos e vendas?
- Se houver, quão forte é essa relação?
- Se houver dados de outras mídias, como essas mídias influenciam nas vendas?
- Qual delas mais influencia nas vendas?
- Podemos estimar o volume de vendas através dos recursos gastos em propaganda?
- Quão acurada pode ser a previsão de vendas no futuro?

Regressão linear simples

Vamos assumir que uma relação linear entre duas variáveis:

$$Y \approx \underbrace{\beta_0}_{\text{Intercepto}} + \underbrace{\beta_1}_{\text{Inclinação}} X$$

Parâmetros do modelo

Para o exemplo anterior:

$$Vendas \approx \beta_0 + \beta_1 TV$$

Regressão linear simples

Para estimar os parâmetros (coeficientes) do modelo, usamos o conjunto de treinamento.

O modelo ajustado é representado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Que permite determinar estimativas do valor de Y.

A acurácia do modelo é verificada no conjunto de teste.

Regressão linear simples

Na prática, os parâmetros do modelo não são conhecidos.

Para estimar os coeficientes, consideramos um conjunto de dados de treinamento:

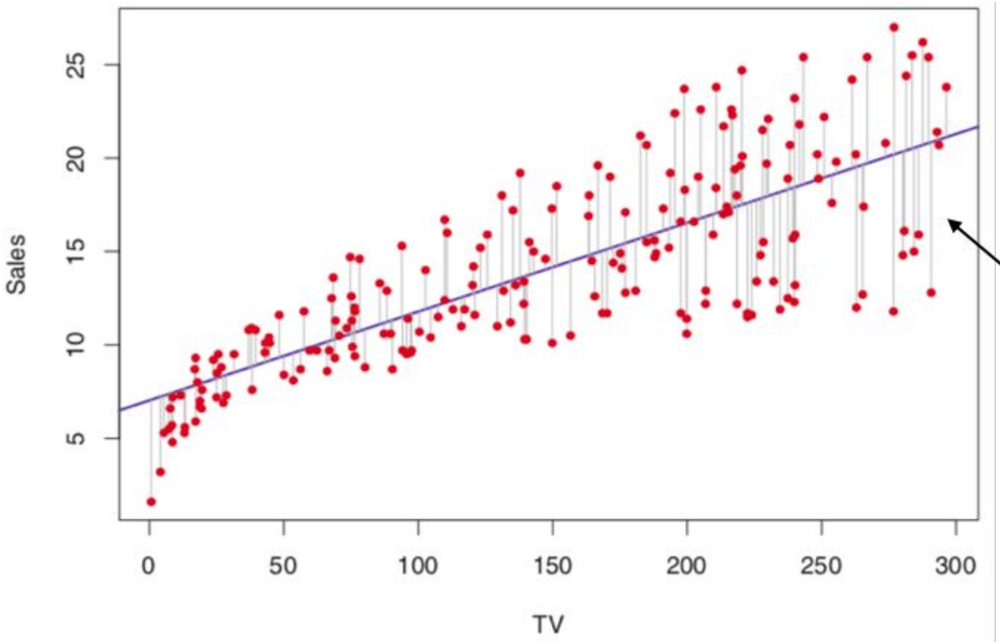
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

que representa **n pares de observações.**

Na regressão linear simples, queremos encontrar os valores dos parâmetros tal que a curva obtida se aproxime o máximo possível dos pontos.

Regressão linear simples

Exemplo: Para realizar a regressão, vamos minimizar o erro em relação à reta estimada.



$$e_i = y_i - \hat{y}_i$$

Erro de uma observação

Regressão linear simples

Seja a predição da variável Y: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

O resíduo é definido pela diferença entre o valor real (do conjunto de treinamento) e o valor predito:

$$e_i = y_i - \hat{y}_i$$

A soma dos quadrados dos resíduos é dada por:

$$\mathrm{RSS} = e_i^2 + e_2^2 + \ldots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_2 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_n x_n)^2$$

Regressão linear simples

Devemos encontrar os parâmetros que minimizem o RSS:

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_2 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_n x_n)^2$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = 0 \qquad \frac{\partial \mathrm{RSS}}{\partial \beta_1} = 0$$

Resolvendo, obtemos:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_{xx}}}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Regressão linear simples

Esse é o método dos mínimos quadrados para estimar os coeficientes do modelo de regressão.

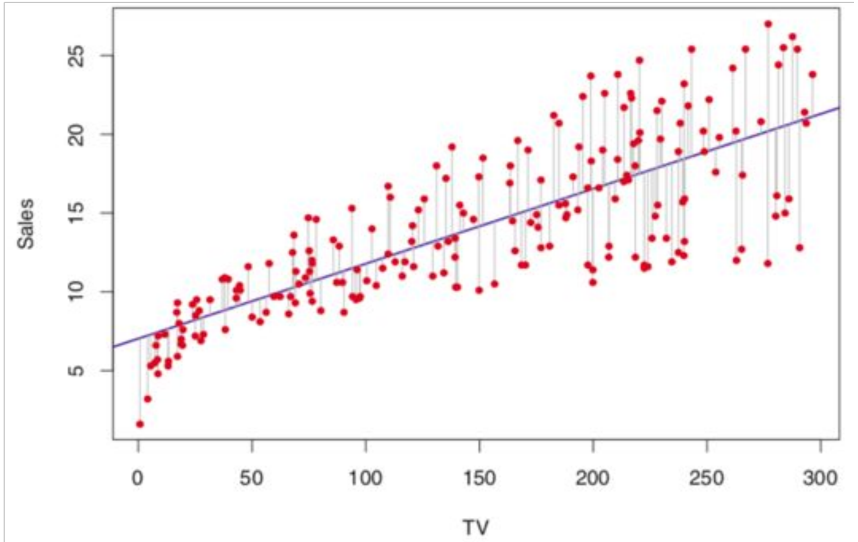
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = 7{,}03$$

$$\hat{\beta}_1 = 0{,}0475$$

A cada 1000 dólares gastos com o anúncio na TV, vende-se aproximadamente 47,5 unidades do produto.



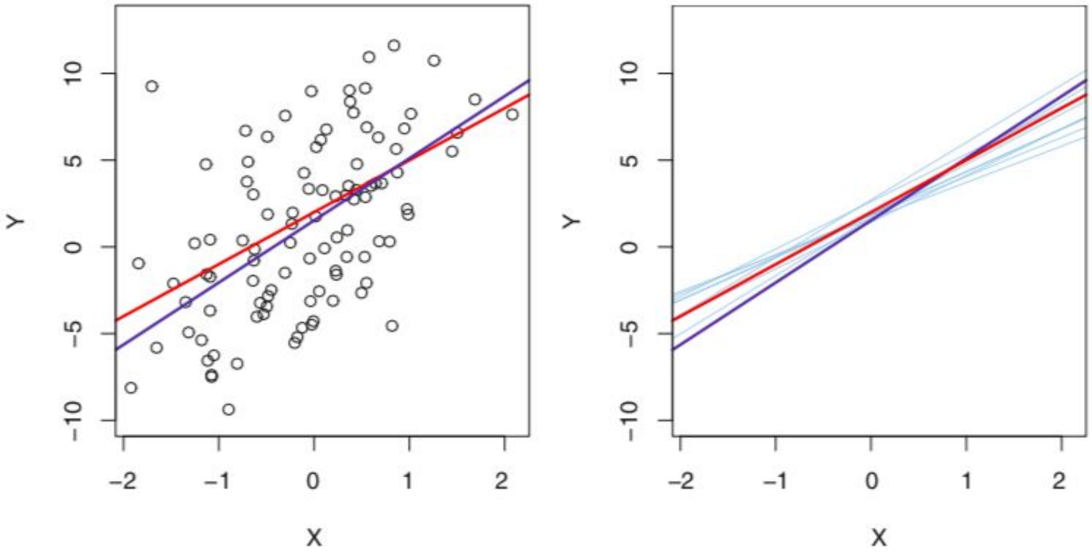
James et al., Introduction to statistical learning with applications in R, 2014.

Regressão linear simples

A reta em vermelha representa a relação original f(X) = 2 + 3X.

A reta em azul é a reta predita.

Na direita, retas em azul claro representam predições a partir de diferentes conjuntos de dados gerados por f(X) = 2 + 3X + $\mathcal{E}$.



James et al., Introduction to statistical learning with applications in R, 2014.

Regressão linear simples

Para quantificar a acurácia do modelo, usamos o erro padrão residual (residual standard error):

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Essa medida fornece o erro absoluto, medido em unidades de Y.

Uma alternativa é a medida R^2, que mede a proporção da variabilidade em Y que pode ser explicada a partir de X.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad 0 \leq R^2 \leq 1$$

Regressão linear simples

R^2 é uma medida de relação linear entre X e Y.

Valores de R^2 próximo de 1 indicam que uma grande proporção da variabilidade dos dados são explicadas pelo modelo de regressão.

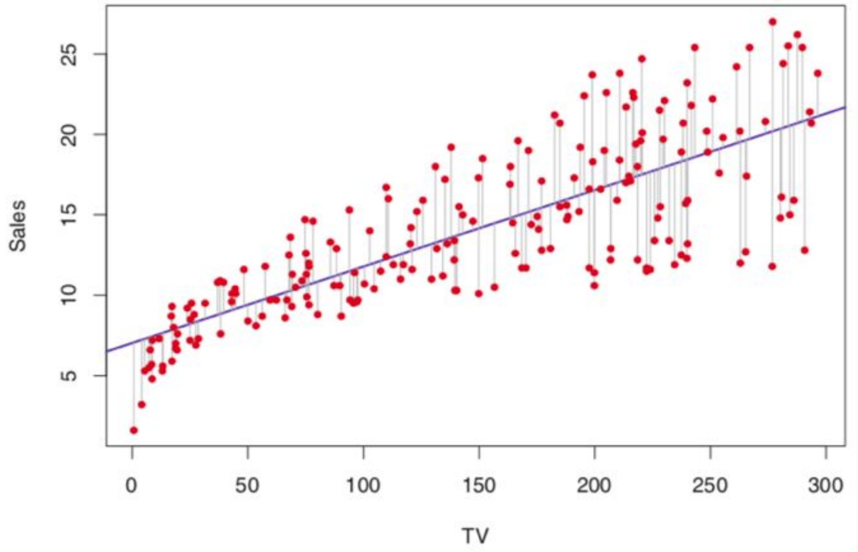
Valores de R^2 próximo de zero indicam que o modelo não explica muito da variabilidade, podendo ser que os dados não seguem uma relação linear ou erro é muito grande.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

Regressão linear simples

No exemplo, R^2 = 0,61, indicando que a variável TV explica apenas 2/3 da variabilidade nas vendas.



$$e_i = y_i - \hat{y}_i$$

Erro de uma observação

James et al., Introduction to statistical learning with applications in R, 2014.

Regressão linear simples

Teste de hipóteses:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

O erro associado na estimação dos parâmetros:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \qquad \boxed{\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},}$$

O intervalo de confiança de 95%:

$$\hat{\beta}_1 \pm 2\mathrm{SE}(\hat{\beta}_1) \qquad \hat{\beta}_0 \pm 2\mathrm{SE}(\hat{\beta}_0)$$

James et al., Introduction to statistical learning with applications in R, 2014.

Regressão linear simples

As hipóteses:

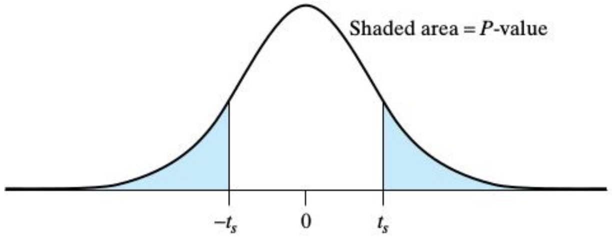
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- H_0: Não há relação entre X e Y.
- H_a: Há alguma relação entre X e Y.

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

Usamos a distribuição t:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$



Quanto mais próximo de zero for
o valor p, maior é confiança em rejeitar H0.

Regressão linear simples

No exemplo anterior, rejeitamos a hipótese nula, ou seja, há uma relação entre o investimento em anúncios na TV e número de itens vendidos.

$$Vendas \approx \beta_0 + \beta_1 TV$$

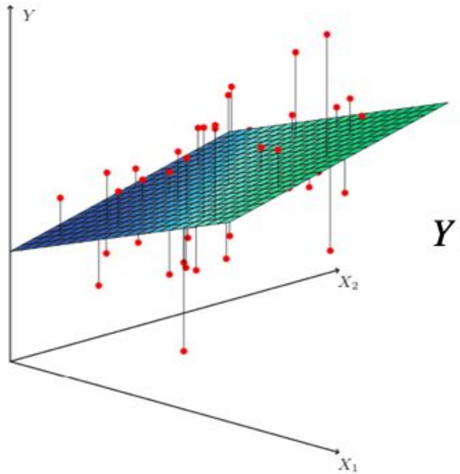
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

James et al., Introduction to statistical learning with applications in R, 2014.

Regressão Linear Múltipla

Regressão linear múltipla

Na maioria dos casos, estamos interessados na influência de várias variáveis em uma variável alvo.



$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Por exemplo, podemos analisar o efeito do investimento em TV, rádio e jornal na quantidade de itens vendidos, ao invés de considerar apenas TV.

Regressão linear múltipla

Suponha que temos **d** preditores distintos para a variável **Y**. Então, o modelo de regressão linear múltipla é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d$$

No exemplo:

$$\text{Vendas} = \beta_o + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Jornal}$$

Importante: o modelo é linear nos parâmetros e não nas variáveis, que podem ser não lineares (ex. X^2, sen(X), ln(X)).

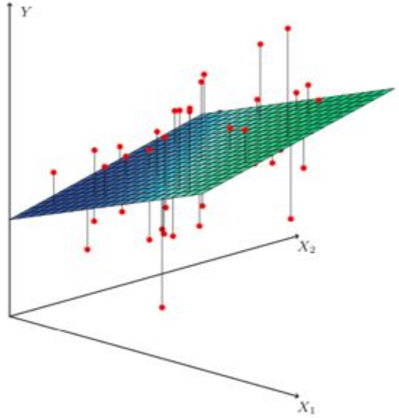
Regressão linear múltipla

Queremos inferir o modelo:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_d x_d$$

Usando o mesmo método que anteriormente, objetivamos minimizar o erro:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$



$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_d x_{id})^2$$

Regressão linear múltipla

Usando notação matricial:

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id}$$

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1d} \\ 1 & x_{21} & x_{22} & \ldots & x_{2d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nd} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_d \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \cdots \\ \epsilon_d \end{pmatrix}$$

- Queremos minimizar a soma dos quadrados dos resíduos:

$$\mathrm{RSS} = \epsilon^T \epsilon = \sum_{i=1}^{N} e_i^2$$

Regressão linear múltipla

Minimizando o erro:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$$\text{RSS} = \epsilon^T\epsilon = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\text{RSS} = \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$$

$$\boxed{\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

Regressão linear múltipla

Assim:

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- $\mathbf{X}^T\mathbf{X}$ deve ser positiva definida.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_d \end{pmatrix}$$

Inversa de Moore-Penrose $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Regressão linear múltipla

Exemplo:

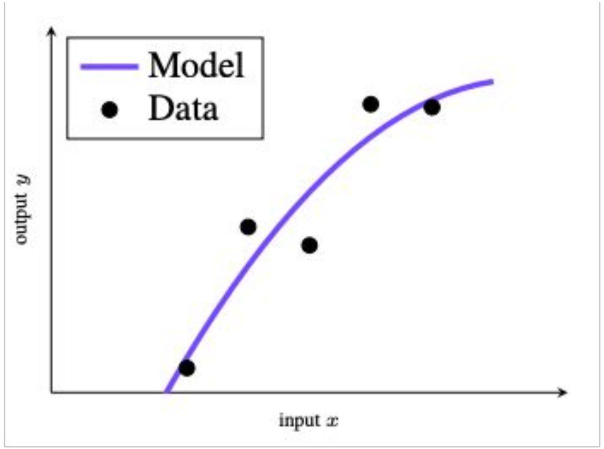
- A variável rádio é a que mais contribui para aumentar o volume de vendas.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

James et al., Introduction to statistical learning with applications in R, 2014.

Regressão linear múltipla

Importante: Notem que o modelo não precisa ter termos lineares em X, mas apenas nos parâmetros:



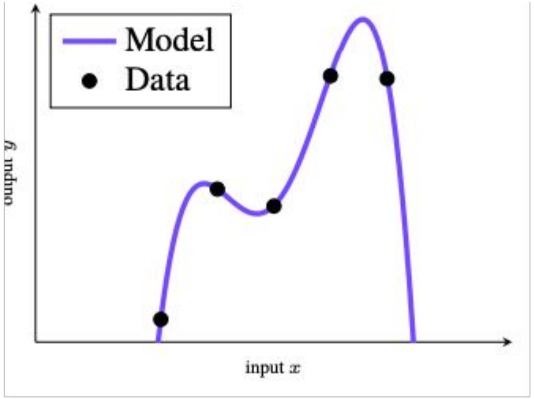
Ainda é um modelo linear!

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Simplificando Modelos via Regularização

Regularização

Nos modelos de regressão, pode ocorrer overfitting:



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$$

- Exemplo: Polinômio de grau 4 para 5 pontos!

Regularização

Uma das maneiras de evitar overfitting é usar regularização.

Um dos métodos mais populares é chamada **Ridge Regression** (ou *Tikhonov regularization*).

O critério de mínimos quadrados é modificado:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \;\; \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_2^2.$$

- Onde $\gamma \geq 0$ é o parâmetro de regularização.
- Para $\gamma = 0$, temos o caso sem regularização.
- Se $\gamma \to \infty$, os parâmetros se aproximam de zero.

$$||\beta||_p = (\sum_{i=1}^{n} |\beta_i|^p)^{1/p}$$

Regularização

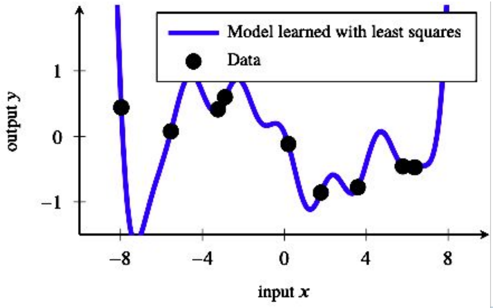
Resolvendo o modelo, como anteriormente, obtemos as equações normais:

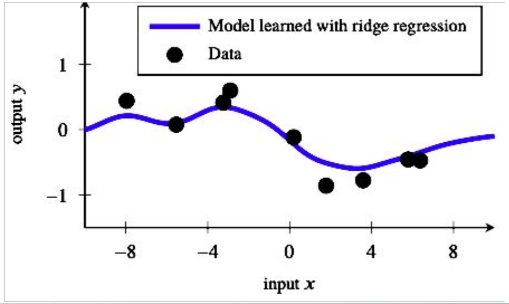
$$(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \gamma I_{d+1})\widehat{\beta} = \mathbf{X}^{\mathsf{T}}\mathbf{y},$$

E a solução:

$$\boxed{\widehat{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \gamma I_{d+1})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.}$$

Sem regularização





Com regularização

Regularização

Outro método bastante popular é chamado *Least Absolute Shrinkage and Selection Operator* (ou **LASSO**).

Nesse caso, o critério de mínimos quadrados será:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \ \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_1,$$

Nesse caso, não há uma solução analítica e métodos de otimização devem ser usados.

$$||\beta||_p = (\sum_{i=1}^{n} |\beta_i|^p)^{1/p}$$

Regularização

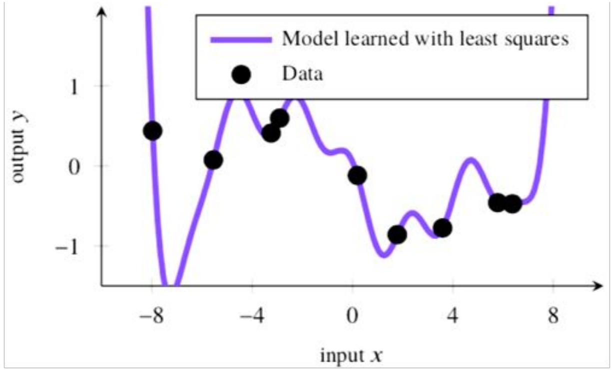
Notem que os resultados obtidos por **ridge regression** e **Lasso são diferentes**.

Enquanto ridge regression mantém os valores dos parâmetros pequenos, **LASSO** tende a selecionar alguns valores para serem diferentes de zero, enquanto que outros são exatamente iguais a zero.

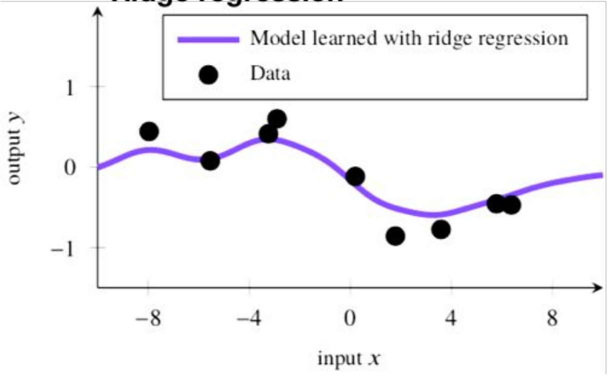
Assim, **LASSO pode ser usado para selecionar atributos.**

Regularização

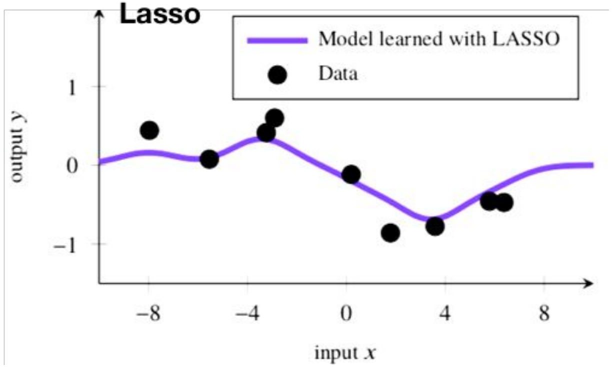
Regressão linear



Ridge regression



Lasso



Regularização

Outros métodos podem ser definidos para regularização.

De maneira geral, a função objetivo pode ser definidas por:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; \underbrace{V(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})}_{\text{data fit}} + \gamma \underbrace{R(\boldsymbol{\beta})}_{\substack{\text{model} \\ \text{flexibility} \\ \text{penalty}}} .$$

- Temos um termo que descreve o quão bem o modelo se ajusta aos dados.
- Um termo que penaliza a complexidade do modelo.
- A regularização é obtida através de um balanço entre esses dois termos.

Sumário

Modelos de regressão

- Regressão Linear,
- Simplificando Modelos via Regularização,
- Avaliando e Interpretando modelos.

Leitura adicional

- Lindholm et al., **Supervised Machine Learning**, 2019. http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf

- James et al., **Introduction to statistical learning with applications in R**, 2014.