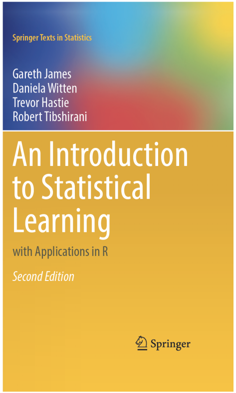
Aula 2: Regressão Linear

Machine Learning

Paulo Orenstein

Verão, 2025
IMPA



Capítulo 3: Métodos lineares para regressão

Créditos de figuras e slides: James, Witten, Hastie e Tibshirani

Introdução: componentes de aprendizado

- Dados de treino: $\{(x_i, y_i)\}_{i=1}^n$, onde $x_i \in \mathcal{X}$ e $y_i \in \mathcal{Y}$
  - Vamos assumir que $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}^p$ (com $n > p$)
- Classe de funções preditivas: $\hat{f}_{\text{tr}} : \mathcal{X} \to \mathcal{Y}$
  - Vamos assumir que $\hat{f}_{\text{tr}}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- Função-perda: $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Vamos assumir que $L(y, \hat{f}(x)) = (y_i - \hat{f}(x))^2$
- Otimizador: encontrar $\hat{f}$ (ou, equivalentemente, $\hat{\beta}_0, \ldots, \hat{\beta}_p$) que minimiza o erro médio:

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \underset{\tilde{\beta}_0, \ldots, \tilde{\beta}_p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_p x_p))^2.$$

Nesse caso, é possível encontrar solução em forma fechada (regressão linear)

Capítulo 3

- Regressão linear simples: previsão e inferência
- Regressão linear múltipla: previsão e inferência
- Extensões: previsores qualitativos, interações, não-linearidades
- Problemas: correlação nos erros, heterocedasticidade, pontos de alavanca e colinearidade

Regressão linear simples: previsão

- Modelo: $Y = \beta_0 + \beta_1 X + \varepsilon$, onde $\beta_0, \beta_1 \in \mathbb{R}$ são coeficientes a serem estimados
- Previsão: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimativas
- Vamos escolher $\hat{\beta}_0$ e $\hat{\beta}_1$ para minimizar a soma residual ao quadrado (RSS):

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\tilde{\beta}_0, \tilde{\beta}_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

- Solução da minimização: como o problema é convexo, derive e iguale a zero, obtendo

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

onde $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ e $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ são as médias amostrais

Regressão linear simples: previsão

- Vamos derivar a solução do último slide para

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname*{argmin}_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

- Primeiro igualamos a derivada em β_0 a zero:

$$0 = (-2) \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{x}) \Longrightarrow \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{y} - \hat{\beta}_1 \overline{x}$$

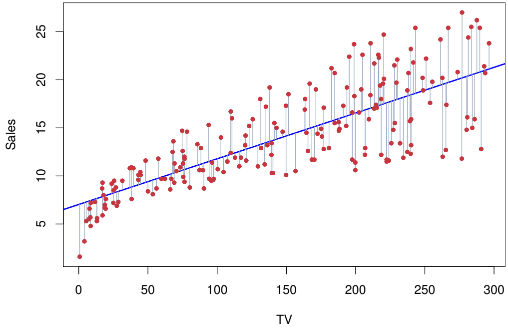
- Daí, igualando a derivada em β_1 a zero:

$$0 = (-2) \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} x_i (y_i - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i),$$

  então podemos reescrever

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sum_{i=1}^{n} x_i (x_i - \overline{x})} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} \quad \left(= \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

Regressão linear simples: exemplo



- Regredindo **`sales`** em **`TV`**, temos $\hat{\beta}_0 = 7.03$ e $\hat{\beta}_1 = 0.0475$ (*i.e.*, $\widehat{\texttt{sales}} = 7.03 + 0.0475 \times \texttt{TV}$)
- Para cada 100 dólares em **`TV`**, há um adicional de 4.75 em vendas
- O fit é razoável, apesar de deficiente nos extremos

Regressão linear simples: inferência

- A inferência sobre $\hat{\beta}_0, \hat{\beta}_1$ depende do único objeto aleatório: ε_i, com $\mathbb{E}[\varepsilon_i] = 0$ e $\mathbb{V}[\varepsilon_i] = \sigma^2$
- Como $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[y_i] = f(x_i) = \beta_0 + \beta_1 x_i$. Os estimadores são não-viesados:

$$\mathbb{E}[\hat{\beta}_1] = \frac{\sum_{i=1}^n (x_i - \overline{x})(\mathbb{E}[y_i - \overline{y}])}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \beta_1$$

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\overline{y}] - \mathbb{E}[\hat{\beta}_1]\overline{x} = \beta_0 + \beta_1 \overline{x} - \mathbb{E}[\hat{\beta}_1]\overline{x} = \beta_0$$

- O erro (ou desvio) padrão de um estimador reflete a sua variabilidade:

$$\mathrm{SE}^2(\hat{\beta}_1) = \mathbb{V}[\hat{\beta}_1] = \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \mathbb{V}[y_i]}{\left(\sum_{i=1}^n (x_i - \overline{x}^2)\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\mathrm{SE}^2(\hat{\beta}_0) = \mathbb{V}[\hat{\beta}_0] = \mathbb{V}[\overline{y}] - 2\,\mathrm{Cov}(\overline{y}, \hat{\beta}_1 \overline{x}) + \mathbb{V}[\hat{\beta}_1]\overline{x}^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 \cdot \overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

pois $\mathrm{Cov}(\overline{y}, \hat{\beta}_1 \overline{x}) = \overline{x} \cdot \mathrm{Cov}(\overline{y}, \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}) = \overline{x} \cdot \frac{\sum_{i=1}^n (x_i - \overline{x})\,\mathrm{Cov}(\overline{y}, y_i)}{\sum_{i=1}^n (x_i - \overline{x})^2} = \overline{x} \frac{\sum_{i=1}^n (\sigma^2/n)(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0$

Regressão linear simples: intervalo de confiança

- Sabemos média e variância de $\hat{\beta}_0, \hat{\beta}_1$, mas e a sua distribuição?
- A seguir, vamos adicionar uma hipótese importante:

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Usando propriedades das distribuições Normais, $\hat{\beta}_0 \sim N(\beta_0, \mathbb{V}[\hat{\beta}_0])$ e $\hat{\beta}_1 \sim N(\beta_1, \mathbb{V}[\hat{\beta}_1])$
- Ou seja, $Z = (\hat{\beta}_1 - \beta_1)/\text{SE}(\hat{\beta}_1) \sim N(0, 1)$ e como $\mathbb{P}[|Z| \leq 2] \approx 0.95$,

$$\mathbb{P}\left[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)\right] \approx 0.95$$

- Isto é, há 95% de chance do intervalo $\left[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)\right]$ conter o verdadeiro valor de β_1, sob repetidas amostragens — importante: β_1 é um valor fixo, mas desconhecido
- Problema: não sabemos σ^2, então não podemos calcular $\text{SE}(\hat{\beta}_1) = \sqrt{\sigma^2 / \sum_{i=1}^n (x_i - \overline{x})^2}$

Regressão linear simples: intervalo de confiança

- Solução: estimar $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Mas qual é a distribuição de $\hat{\sigma}^2$?
- Distribuição χ^2_n: se $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0,1)$, então $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$
- Fato: Vale que $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ (vamos provar esse fato mais adiante)
- Com isso, ao invés de $\text{SE}(\hat{\beta}_1)$, podemos usar o estimador

$$\widehat{\text{SE}}^2(\hat{\beta}_1) = \hat{\sigma}^2 / \sum_{i=1}^{n}(x_i - \overline{x}^2)$$

- Nesse caso, ao invés de $Z = (\hat{\beta}_1 - \beta_1)/\text{SE}(\hat{\beta}_1) \sim N(0,1)$, que não sabíamos calcular por conta de $\text{SE}(\hat{\beta}_1)$ depender de σ^2, usamos

$$t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{SE}}(\hat{\beta}_1)}$$

- Usando $\text{SE}(\hat{\beta}_1)$, a distribuição era Normal. Mas e agora usando $\widehat{\text{SE}}(\hat{\beta}_1)$?

Regressão linear simples: intervalo de confiança

- Distribuição t_n: se $Z \sim N(0,1)$ e $K \sim \chi^2_n$ são independentes, $Z/\sqrt{K/n} \sim t_n$

- Como sabemos que $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$,

$$t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2/\sigma^2}{n-2}}} \sim t_{n-2},$$

  a independência segue do fato de que $(y_i - \hat{y}_i)$ é independente de $\hat{\beta}_1$ (vamos provar depois)

- Para n grande, a distribuição t_{n-2} é muito próxima da Normal: ainda vale $\mathbb{P}[|t| \leq 2] \approx 0.95$ e

$$\mathbb{P}\left[\hat{\beta}_1 - 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_1)\right] \approx 0.95.$$

- Regredindo `sales` em `TV` vimos que $\hat{\beta}_1 = 0.0475$; o intervalo de confiança de 95% é $[0.042, 0.053]$

Regressão linear simples: teste de hipótese

- Se o intervalo de confiança a 95% não contém zero, então provavelmente $\beta_1 \neq 0$

- Ou seja, podemos reformular o intervalo de confiança como um teste de hipótese:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0, \end{aligned}$$

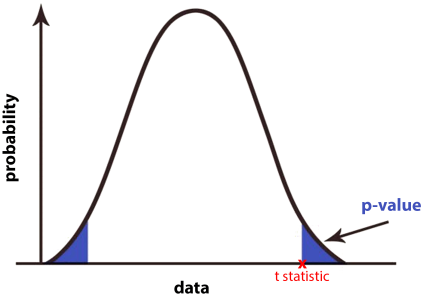
  o que equivale a testar se X tem algum impacto em Y

- Quão provável é H_0? Sob H_0, $\beta_1 = 0$, então, de antemão, $t \stackrel{H_0}{=} (\hat{\beta}_1 - 0)/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ deveria ser uma distribuição t_{n-2}. Depois de observado os dados, temos t_{obs}

- Usando computadores, é fácil achar a probabilidade de uma variável de distribuição t_{n-2} ser igual ou maior, em módulo, a t_{obs}. Essa probabilidade é o chamado p-valor

Regressão linear simples: p-valor

- Intuição: assumindo H_0, quão improvável seria observar o t que observamos?
- Matematicamente, o p-valor é dado por: $\mathbb{P}[|T| > t_{\text{obs}}]$, onde T tem distribuição conhecida
- Se o p-valor é muito pequeno (*e.g.*, menor que 5%), rejeitamos a hipótese H_0



Regressão linear simples: exemplo

	Coefficient	Std Error	t-statistic	p-value
`Intercept` ($\hat{\beta}_0$)	7.0325	0.4578	15.36	<0.0001
`TV` ($\hat{\beta}_1$)	0.0475	0.0027	17.67	< 0.0001

- **`Intercept`** e **`TV`** têm alto valor de estatística t ($=$ coeficiente/erro padrão)
- Com isso, os p-valores são muito baixos (abaixo de 0.0001)
- É estatisticamente improvável que os coeficientes β_0 e β_1 sejam zero
- Concluímos que provavelmente existe uma relação entre X e Y nesse caso

Medindo a acurácia do modelo linear: R^2

- Um diagnóstico importante é o erro padrão residual:

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

- O coeficiente de R^2 é a fração de variância explicada pelo modelo (versus $\overline{y}$):

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

- Quanto maior R^2, melhor o modelo

- Essa medida é cotada por cima e por baixo: $0 \leq R^2 \leq 1$ (com casos extremos $\hat{y}_i = y_i$ e $\hat{y}_i = \overline{y}$)

Capítulo 3

- Regressão linear simples: previsão e inferência
- Regressão linear múltipla: previsão e inferência
- Extensões: previsores qualitativos, interações, não-linearidades
- Problemas: correlação nos erros, heterocedasticidade, pontos de alavanca e colinearidade

Regressão linear múltipla: estimação

- Agora, queremos usar mais de uma variável explicativa. O modelo é:

$$Y = f(X_1, \dots, X_p) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Vale a pena usar notação matricial:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & & \vdots & \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

- Com isso, conseguimos um enorme poder de síntese:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \forall i \quad \Longleftrightarrow \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Vamos escolher $\boldsymbol{\beta}$ para minimizar os resíduos

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Regressão linear múltipla: estimação

- Tomando derivadas em $\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, obtemos:

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad \frac{\partial^2 \text{RSS}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2\mathbf{X}^T\mathbf{X}$$

- Supondo que $\mathbf{X}$ tem posto cheio e portanto que $\mathbf{X}^T\mathbf{X}$ é positiva definida, a condição de primeira ordem é $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$, ou seja,

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

- As previsões são dadas por

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y},$$

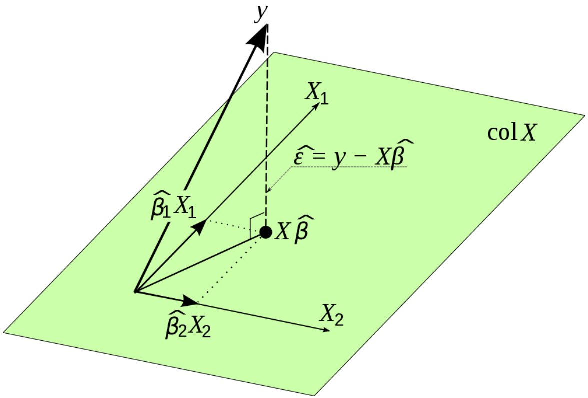
  onde $\mathbf{H}$ é a "hat matrix", que leva $\mathbf{y}$ à sua previsão $\hat{\mathbf{y}}$.

- $\mathbf{H}$ é uma matriz de projeção no espaço coluna de $\mathbf{X}$, daí simétrica e idempotente (*i.e.*, $\mathbf{H}\mathbf{H} = \mathbf{H}$)

- Aqui, ajuda visualizar o que está acontecendo

Regressão linear múltipla: estimação

- Pela condição de primeira ordem: $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T(\mathbf{y} - \mathbf{H}\mathbf{y}) = 0$



Credit: Wikipedia

- Ou seja, $\hat{\boldsymbol{\varepsilon}} \perp \text{col}(\mathbf{X})$, e, em particular, $\hat{\boldsymbol{\varepsilon}} \perp \mathbf{X}\hat{\boldsymbol{\beta}}$

Regressão linear múltipla: BLUE

Teorema de Gauss-Markov

Dentre todos os estimadores lineares não-viesados, o estimador $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ é o que possui a menor variância. Ou seja, para qualquer vetor $\mathbf{c} \in \mathbb{R}^{p\times 1}$ e $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$ com $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$,

$$\mathbb{V}[c^T\hat{\boldsymbol{\beta}}] \leq \mathbb{V}[c^T\tilde{\boldsymbol{\beta}}].$$

Isso significa que $\hat{\boldsymbol{\beta}}$ é o melhor estimador linear não-viesado (BLUE) de $\boldsymbol{\beta}$.

Demonstração. Escrevendo $\tilde{\boldsymbol{\beta}} = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D})\mathbf{y}$, vale que $\mathbf{D}\mathbf{X} = \mathbf{0}$, pois para qualquer $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = \mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbb{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D})\mathbf{y}] = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D})(\mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]) = \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta}.$$

Aí, como $\mathbb{V}[\mathbf{y}] = \sigma^2\mathbf{I}$,

$$\begin{aligned}
\mathbb{V}[\mathbf{c}^T\tilde{\boldsymbol{\beta}}] &= \mathbf{c}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D})\ \mathbb{V}[\mathbf{y}]\ ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D})^T\mathbf{c} \\
&= \mathbb{V}[\mathbf{c}^T\hat{\boldsymbol{\beta}}] + 2\sigma^2\mathbf{c}^T\mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c} + \sigma^2\mathbf{c}^T\mathbf{D}\mathbf{D}^T\mathbf{c} = \mathbb{V}[\mathbf{c}^T\hat{\boldsymbol{\beta}}] + \sigma^2(\mathbf{D}^T\mathbf{c})^T(\mathbf{D}^T\mathbf{c}) \\
&= \mathbb{V}[\mathbf{c}^T\hat{\boldsymbol{\beta}}] + \tilde{\mathbf{c}}^T\tilde{\mathbf{c}} \geq \mathbb{V}[\mathbf{c}^T\hat{\boldsymbol{\beta}}]
\end{aligned}$$

□

Regressão linear múltipla: inferência

- Como antes, $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ e $\mathbb{V}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
- Assumindo que $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$,
$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$
- A variância σ^2 pode ser estimada via
$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-(p+1)}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$$
- O que sabemos sobre um $\hat{\beta}_j$? Como $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, segue que $\hat{\beta}_j = \mathbf{e}_j^T\hat{\boldsymbol{\beta}}$ é Normal
  - Média: $\mathbb{E}[\hat{\beta}_j] = \mathbf{e}_j^T\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbf{e}_j^T\boldsymbol{\beta} = \boldsymbol{\beta}_j$
  - Variância: $\mathbb{V}[\hat{\beta}_j] = \mathbf{e}_j^T\mathbb{V}[\hat{\boldsymbol{\beta}}]\mathbf{e}_j = \mathbf{e}_j^T(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})\mathbf{e}_j = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$.
- Se valer $\frac{1}{n-(p+1)}\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$, conseguimos recuperar o teste t para $H_0 : \beta_j = 0$

Regressão linear múltipla: inferência

- Note que $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$, pois $\mathbf{H}\mathbf{X} = \mathbf{X}$, logo

$$(n - (p+1))\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{Q}^T\boldsymbol{\varepsilon})^T\boldsymbol{\Lambda}(\mathbf{Q}^T\boldsymbol{\varepsilon}) \stackrel{d}{=} \sum_{i=1}^n \lambda_i \varepsilon_i^2$$

$$\sim \sigma^2 \cdot \chi^2_{\sum_{i=1}^n \lambda_i} = \sigma^2 \cdot \chi^2_{\mathrm{tr}(\mathbf{I}-\mathbf{H})} = \sigma^2 \cdot \chi^2_{n-(p+1)},$$

onde usamos a decomposição espectral $\mathbf{I} - \mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, e, por idempotência, $\boldsymbol{\Lambda}$ tem 0s e 1s

- Além disso, note que $\hat{\boldsymbol{\beta}}$ é independente de $\hat{\boldsymbol{\varepsilon}}$ (e, portanto $\hat{\sigma}^2$), já que são Normais e:

$$\mathbb{V}\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\varepsilon}}\end{bmatrix} = \mathbb{V}\left[\begin{bmatrix}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ \mathbf{I} - \mathbf{H}\end{bmatrix}\mathbf{y}\right] = \begin{bmatrix}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ \mathbf{I} - \mathbf{H}\end{bmatrix}\sigma^2\mathbf{I}\begin{bmatrix}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} & \mathbf{I} - \mathbf{H}\end{bmatrix} = \sigma^2\begin{bmatrix}(\mathbf{X}^T\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{H}\end{bmatrix},$$

já que $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Regressão linear múltipla: teste de hipóteses

- Para testar $H_0 : \beta_j = 0$, basta notar que

$$t^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}} = \frac{\frac{\hat{\beta}_j - 0}{\sqrt{\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}}}{\sqrt{\frac{(n-p+1)\hat{\sigma}^2}{(n-p+1)\sigma^2}}} \overset{H_0}{\sim} t_{n-(p+1)}$$

- Para testar $H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$ conjuntamente, é possível usar

$$F^{(k)} = \frac{(\sum_{i=1}^n (y_i - \hat{y}_i^{(k)})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2)/(p-k)}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-(p+1))} \overset{H_0}{\sim} F_{p-k,n-(p+1)},$$

onde $\hat{y}_i^{(k)}$ denota a i-ésima previsão quando $\beta_{k+1} = \cdots = \beta_p = 0$, e $F_{a,b}$ é a distribuição F com graus de liberdade a e b

Regressão linear múltipla: prova do teste F

- A distribuição $F_{a,b}$ é definida como $\frac{\chi_a^2/a}{\chi_b^2/b} \sim F_{a,b}$ para χ_a^2 e χ_b^2 são independentes
- Particione $\mathbf{X} = \begin{bmatrix} \mathbf{X}_k & \mathbf{X}_{-k} \end{bmatrix}$ e $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{\beta}_{-k} \end{bmatrix}$
- Objetivo: testar $H_0 : \boldsymbol{\beta}_{-k} = \mathbf{0}$ versus $H_a : \boldsymbol{\beta}_{-k} \neq 0$.
- Chame as matrizes de projeção de $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ e $\mathbf{H}_{-k} = \mathbf{X}_{-k}(\mathbf{X}_{-k}^T\mathbf{X}_k)^{-1}\mathbf{X}_{-k}$
- Pelos cálculos anteriores,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} \sim \chi^2_{n-(p+1)},$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i^{(k)})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}^T(\mathbf{H} - \mathbf{H}_{-k})\mathbf{y} \sim \chi^2_{p-k}$$

Regressão linear múltipla: prova do teste F

- Agora, resta mostrar que $\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}$ e $\mathbf{y}^T(\mathbf{H}-\mathbf{H}_{-k})\mathbf{y}$ são independentes
- É suficiente mostrar a independência entre $(\mathbf{I}-\mathbf{H})\mathbf{y}$ e $(\mathbf{H}-\mathbf{H}_{-k})\mathbf{y}$
- Como $\mathbb{V}[\mathbf{y}]=\sigma^2\mathbf{I}$,

$$\mathbb{V}\left[\begin{bmatrix}\mathbf{I}-\mathbf{H}\\ \mathbf{H}-\mathbf{H}_{-k}\end{bmatrix}\mathbf{y}\right]=\begin{bmatrix}\mathbf{I}-\mathbf{H}\\ \mathbf{H}-\mathbf{H}_{-k}\end{bmatrix}\sigma^2\mathbf{I}\begin{bmatrix}\mathbf{I}-\mathbf{H}\\ \mathbf{H}-\mathbf{H}_{-k}\end{bmatrix}^T=\sigma^2\begin{bmatrix}\mathbf{I}-\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}-\mathbf{H}_{-k}\end{bmatrix},$$

pois $\mathbf{I}-\mathbf{H}$ e $\mathbf{H}-\mathbf{H}_{-k}$ são idempotentes e $(\mathbf{I}-\mathbf{H})(\mathbf{H}-\mathbf{H}_{-k})=\mathbf{H}-\mathbf{H}-\mathbf{H}_{-k}+\mathbf{H}_{-k}=\mathbf{0}$

- Logo, por definição,

$$\frac{\mathbf{y}^T(\mathbf{H}-\mathbf{H}_{-k})\mathbf{y}/(p-k)}{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(n-(p+1))}\sim\frac{\chi^2_{p-k}/(p-k)}{\chi^2_{n-(p+1)}/(n-(p+1))}\sim F_{p-k,n-(p+1)}$$

Regressão linear múltipla: exemplo

- A interpretação dos resultados nem sempre é direta:

	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

- Faz sentido que **`newspaper`** não seja relevante?
- Como $\mathrm{Corr}(\texttt{newspaper}, \texttt{radio}) = 0.35$, algum efeito de **`newspaper`** pode estar em **`radio`**
- Ou seja, agora a interpretação dos coeficientes é menos clara: β_j indica o efeito médio em Y de um aumento de uma unidade de X_j, mantendo fixo todos os outros previsores
- É sabido que $\texttt{shark attacks} \sim \beta_0 + \beta_1 \cdot \texttt{ice cream sales}$ tem $\hat{\beta}_1$ significativamente positivo. Não é possível fazer associação causal.

Questões fundamentais

1. Vale que ao menos um previsor $X_1, \ldots, X_p$ é estatisticamente significante para prever Y?

2. Existe algum subconjunto de previsores que ajudam a prever Y?

3. Quão bom é o fit do modelo?

4. Estimados os coeficientes, que valor devemos prever para Y e quão acurada é essa previsão?

Questões fundamentais: quais variáveis são significativas?

- Necessário: hipótese sobre a distribuição de ε_i: $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$
- Para testar $H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$ conjuntamente, usamos o teste F:

$$F_k = \frac{(\sum_{i=1}^n (y_i - \hat{y}_i^{(k)})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2)/(p-k)}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-(p+1))} \overset{H_0}{\sim} F_{p-k, n-(p+1)},$$

- Para testar se certo β_j é significativo, basta colocá-lo por último e testar com $k = p - 1$ (isso é equivalente ao teste t)
- Para testar se ao menos uma variável é estatisticamente significativa, tome $k = 0$

Questões fundamentais: quais variáveis são preditivas?

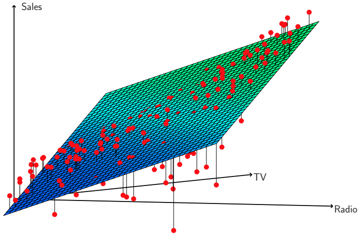
- Há 2^p escolhas de subconjuntos preditivos. Precisamos de estratégias computacionalmente aceitáveis.
  - Forward selection: começando pelo modelo nulo, adicione a variável que minimiza o RSS, uma de cada vez
  - Backward selection: começando pelo modelo cheio, elimine variáveis, uma de cada vez, escolhendo aquela com maior p-valor em cada etapa
  - Mixed selection: começando pelo modelo nulo, adicione a variável que minimiza o RSS, uma de cada vez; se o p-valor for maior do que um certo valor, descarte-a
- Mais adiante discutiremos estratégias mais avançadas

Questões fundamentais: fit

- Para medir o fit, focamos nos resíduos $\hat{\varepsilon}_i = y_i - \hat{y}_i$
- Note que $\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ sempre diminui com mais previsores
- O erro padrão residual (RSE) inclui uma correção para esse problema:

$$\text{RSE} = \sqrt{\frac{1}{n - (p+1)} \text{RSS}}$$

- Também é útil visualizar os resíduos



Questões fundamentais: previsão

- Dado um novo x_{n+1}, nossa previsão seria $\hat{y}_{n+1} = x_{n+1}^T \hat{\boldsymbol{\beta}}$
- É possível encontrar um intervalo preditivo para y_{n+1}:

$$\mathbb{E}[y_{n+1} - \hat{y}_{n+1}] = 0$$

$$\mathbb{V}[y_{n+1} - \hat{y}_{n+1}] = \sigma^2 + x_{n+1}^T \mathbb{V}[\hat{\beta}] x_{n+1} = \sigma^2 (1 + x_{n+1}^T (\mathbf{X}^T \mathbf{X})^{-1} x_{n+1})$$

$$(y_{n+1} - \hat{y}_{n+1}) / \sqrt{\hat{\sigma}^2 (1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})} \sim t_{n-(p+1)}$$

- Daí, com probabilidade aproximadamente 95%,

$$y_{n+1} \in \left[\hat{y}_{n+1} - 2 \cdot \sqrt{\hat{\sigma}^2 (1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})}, \hat{y}_{n+1} + 2 \cdot \sqrt{\hat{\sigma}^2 (1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})}\right]$$

Capítulo 3

- Regressão linear simples: previsão e inferência
- Regressão linear múltipla: previsão e inferência
- Extensões: previsores qualitativos, interações, não-linearidades
- Problemas: correlação nos erros, heterocedasticidade, pontos de alavanca e colinearidade

Extensões: previsores qualitativos

- O que fazer quando alguns previsores são qualitativos ou categóricos?
- Nesse caso, os valores que o previsor toma são discretos
- Suponha que queiramos prever o saldo do cartão de crédito (**`balance`**) a partir de 6 variáveis quantitativas e 4 qualitativas
  - **`own`**: se o indivíduo tem casa própria
  - **`student`**: se é estudante
  - **`status`**: estado civil
  - **`region`**: zona norte, oeste ou sul

Extensões: previsores qualitativos

- Para investigar se casa própria tem um efeito, ignorando outras variáveis, defina

$$x_i = \begin{cases} 1, & \text{se } i\text{-ésima pessoa tem casa própria} \\ 0, & \text{caso contrário} \end{cases}$$

- Isso equivale ao modelo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{se } i\text{-ésima pessoa tem casa própria} \\ \beta_0 + \varepsilon_i, & \text{caso contrário} \end{cases}$$

- Ou seja, β_1 captura o impacto médio de ter casa própria no saldo

Extensões: previsores qualitativos

- Resultados para o modelo:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`own[Yes]`	19.73	46.05	0.429	0.6690

- Saldo médio para quem não tem casa: \$509.80

- Saldo médio para quem tem casa: $\$509.80 + 19.73 = 529.53$

- Mas p-valor não parece ser muito pequeno

Extensões: previsores qualitativos

- Para mais de dois níveis, usamos mais variáveis
- No caso de **`region`**, criamos variáveis

$$x_{i1} = \begin{cases} 1, & \text{se } i\text{-ésima pessoa é da zona sul} \\ 0, & \text{se } i\text{-ésima pessoa não é da zona sul} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{se } i\text{-ésima pessoa é da zona oeste} \\ 0, & \text{se } i\text{-ésima pessoa não é da zona oeste} \end{cases}$$

- Isso equivale ao modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da zona sul} \\ \beta_0 + \beta_2 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da zona oeste} \\ \beta_0 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da zona norte} \end{cases}$$

Extensões: previsores qualitativos

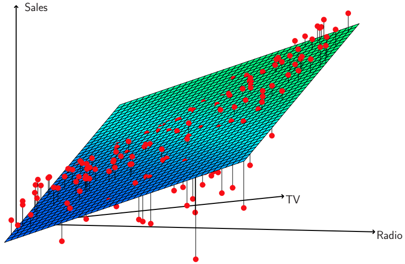
- O nível sem variável (zona norte) é entendido como um valor de base
- Resultados para o modelo:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`region[South]`	-18.69	65.02	-0.287	0.7740
`region[West]`	-12.50	56.68	-0.221	0.8260

- Morar na região oeste significa um decréscimo médio de -12.50 no saldo em relação à morar na zona norte
- Mas será que essas diferenças de zona são significativas? Teste F com $H_0 : \beta_1 = \beta_2 = 0$ tem p-valor de 0.96 — não é possível rejeitar H_0

Extensões: interações

- Regressão linear tem uma hipótese aditiva: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon$
- Ou seja, um aumento de \$100 dólares em anúncios de TV causa um aumento fixo nas vendas, independentemente do gasto com radio. Isso não parece ser o caso:



- Quando gastos com **TV** ou **radio** são baixos, modelo superestima vendas; quando os gastos são divididos, modelo subestima vendas: interação entre **TV** e **radio**

Extensões: interações

- Solução: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{TV} \times \texttt{radio}) + \varepsilon$

- A interação entre **TV** e **radio** parece importante:

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
radio	0.0289	0.009	3.24	0.0014
TV × **radio**	0.0011	0.000	20.73	<0.0001

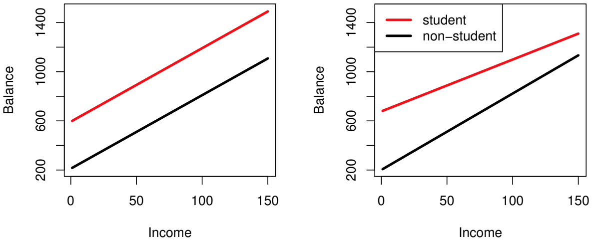
- O R^2 sobe de 89.7% para 96.8% com a inclusão da interação

- Um aumento de \$1000 em **TV** aumenta as vendas em $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$; um aumento de \$1000 em **radio** aumenta vendas em $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$

Extensões: interações

- Princípio da hierarquia: ao incluir uma interação entre previsores, os previsores também devem ser adicionados independentemente. (Senão a interpretação muda: *e.g.*, $y = \beta_0 + \beta_1(\texttt{student} \times \texttt{own}) + \varepsilon$ tem β_1 medindo o impacto de ser estudante com casa própria versus todo o resto)
- É possível misturar interações entre previsores quantitativos e qualitativos:

$$\texttt{balance} \approx \beta_0 + \beta_1 \texttt{income} + \beta_2 \texttt{student}$$
$$\texttt{balance} \approx \beta_0 + \beta_1 \texttt{income} + \beta_2 \texttt{student} + \beta_3 (\texttt{student} \times \texttt{income})$$

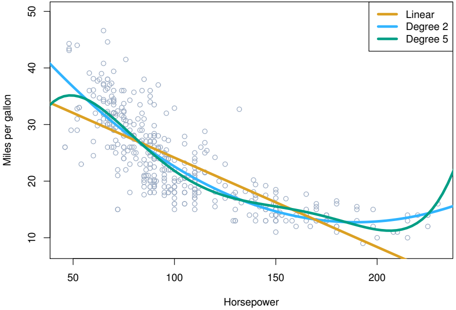


Extensões: não-linearidades

- Uma maneira de lidar com não-linearidades é usando potências de previsores:

$$y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 = \mathbf{X}\boldsymbol{\beta}$$

- Exemplo:



Extensões: não-linearidades

- A figura sugere a regressão

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \varepsilon$$

- Temos evidências para concluir que a adição do termo quadrático é útil:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	59.9001	1.8004	3.16	<0.0001
`horsepower`	-0.4662	0.0311	-15.0	<0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	<0.0001

- A ideia de acrescentar potências de previsores é também chamada de regressão polinomial

Perguntas para revisão

- Como encontrar o coeficiente $\hat{\beta}$ de regressão linear?
- Que hipóteses adicionais são necessárias para inferência?
- Como testar a hipótese $H_0 : \beta_j = 0$? E a hipótese $H_0 : \beta_j = \beta_{j+1} = \cdots = \beta_{p+1} = 0$?
- O que é p-valor? O que significa um p-valor ser alto ou baixo para a hipótese H_0 sendo testada?
- Como é possível interpretar cada coeficiente?
- O que é R^2? E RSE? Como eles medem a acurácia do modelo linear?
- Como estender regressão linear para usar previsores qualitativos?
- Como incluir interações entre features numa regressão?
- Como incluir features não-lineares numa regressão?