

PES310 MODELAGEM E INFERÊNCIA ESTATÍSTICA

Revisão final

Resumo visual da disciplina

Modelagem e Inferência Estatística



SEMANA 1 | Regressão linear simples Atividade Avaliativa

Nesta semana, eu já:

- Conheci os principais assuntos que serão abordados nesta matéria;
- Lembrei os conceitos de estatística e probabilidade;
- Entendi o que é a regressão linear e a quais tipos de dados ela se aplica;
- Resolvi problemas com dados onde pode ser aplicada a regressão linear.



SEMANA 2 | Análise de ajustes Atividade Avaliativa

Nesta semana, eu já:

- Entendi os procedimentos para fazer inferências;
- Obtive medidas qualitativas mediante o coeficiente de correlação;
- Estimei os coeficientes de regressão usando o método dos mínimos quadrados.



SEMANA 3 | Teste de hipótese e Previsão da resposta Atividade Avaliativa

Nesta semana, eu já:

- Realizei o teste de hipótese para verificar se o modelo de regressão de primeira ordem é um ajuste apropriado aos dados;
- Estimei a resposta esperada, predizi valores de observações futuras, e encontrei seus intervalos de confiança usando os coeficientes de confiança;
- Fiz inferências sobre coeficiente de correlação entre a variável resposta e as variáveis preditoras.



SEMANA 4 | Adequações do modelo e modelos não lineares Atividade Avaliativa

Nesta semana, eu já:

- Entendi os processos de padronização de variáveis.
- Usei o método de mínimos quadrados para estimar os coeficientes de regressão em um modelo de regressão não linear.
- Realizei teste de hipótese para a determinação de quais coeficientes de regressão são significantes.



SEMANA 5 | Métodos de regressão gerais, Modelos Logístico e polinomial

Atividade Avaliativa

Nesta semana, eu já:

- Estimei os coeficientes de regressão em um modelo de regressão múltipla e realizei teste de hipótese para a determinação de quais coeficientes de regressão são significantes.
- Ajustei modelos de regressão linear múltipla a um conjunto de dados ao usar duas ou mais variáveis preditoras, e realizar análise de resíduos.
- Ajustei modelos de regressão linear múltipla a um conjunto de dados que envolve variáveis preditoras qualitativas ou categóricas.
- Determinei a presença de multicolinearidade e sua possível eliminação.



SEMANA 6 | Regressão múltipla parte 1 Atividade Avaliativa

Nesta semana, eu já:

- Lembrei modelos existentes na teoria de confiabilidade;
- Entendi os modelos de regressão linear múltipla usando variáveis preditoras e variáveis qualitativas;
- Analisei dados provenientes de experimentos com efeitos fixos, aleatórios ou mistos;
- Realizei análise de resíduos para verificar a adequação dos modelos em consideração;
- Usei técnicas não paramétricas em certos tipos de planejamentos quando as condições de normalidade não são válidas;
- Resumi e interpretar os resultados desses experimentos;
- Estimei observações ausentes em certos tipos de planejamentos e, subsequentemente, analisar os dados como dados balanceados;
- Fiz inferências sobre coeficiente de correlação entre a variável resposta e as variáveis preditoras.



SEMANA 7 | Regressão múltipla parte 2 Atividade Avaliativa

Nesta semana, eu já:

- Conheci os modelos lineares generalizados e as aplicações de cada um deles;
- Lembrei os conceitos de contagens, categorias e dados assimétricos;
- Apliquei os modelos a diferentes dados de acordo com critérios estabelecidos.



SEMANA 8 | Revisão

Nesta semana, eu já:

- Revisei os principais conteúdos vistos;
- Identifiquei os conteúdos mais sensíveis e que merecem mais atenção no momento da revisão.

Algumas perguntas iniciais...

- Iremos aprender todas as bases estatísticas para fazer a análise de dados? A ciência de dados se preocupa em validar se os dados são confiáveis?
- Essa disciplina contempla cálculo numérico?

O modelo linear simples

Regressão linear simples

A regressão linear simples corresponde ao **modelo mais simples** que se pode propor para descrever a relação que eventualmente exista entre uma **variável explicatória x** e uma **variável resposta Y**

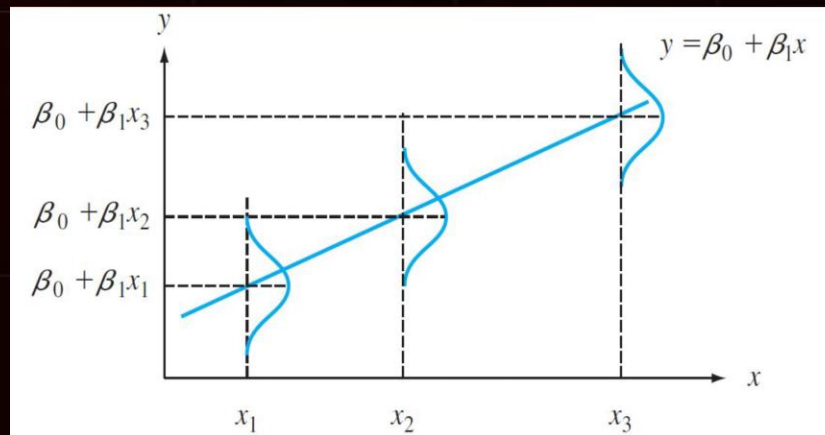
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Como não temos acesso às populações de Y , usamos **dados para estimar os parâmetros β_0 e β_1**

O modelo linear simples

Hipóteses do modelo linear simples

- Para cada valor da variável explicativa x , a distribuição populacional da resposta Y é normal, com média $\mu = \beta_0 + \beta_1 x$ e desvio padrão σ .



- A média geralmente muda para diferentes valores de x , enquanto o desvio padrão é o mesmo para todo x

O modelo linear simples

Hipóteses do modelo linear simples

- Se (x_i, y_i) denota um par de valores observados e \hat{y}_i o valor previsto da resposta em $x = x_i$, então os resíduos

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$$

são independentes

- Uma vez que tenhamos obtido a melhor estimativa dos parâmetros β_0 e β_1 e validado o modelo, podemos usá-lo para fazer previsões para a resposta y em função de x e vice-versa

O modelo linear simples

Formulário

- A partir de um conjunto de observações $(x_1, y_1), \dots, (x_n, y_n)$ —isto é, dados—**estimamos os parâmetros** da regressão linear pelas fórmulas

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

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

Validação do modelo linear

O modelo é útil?

- Geralmente propomos um modelo de regressão linear para fazer previsões. Portanto, devemos tanto quanto possível tentar validar a utilidade da reta de regressão estimada resultante $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Avaliação do coeficiente de determinação

$$r^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{S_{yy}}$$

Teste de significância

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$$

Validação do modelo linear

O coeficiente de determinação r^2

- Pode ser interpretado como a proporção da variação observada de y que pode ser atribuída a uma relação linear aproximada entre y e x (i.e., pelo modelo linear)
- O modelo é tanto mais adequado quanto r^2 for mais próximo de 1

A hipótese $H_0: \beta_1 = 0$

- Caso a hipótese H_0 seja plausível, o modelo não é adequado, pois $\beta_1 = 0$ significa que não existe relação entre y e x (poder de previsão nulo)

Validação do modelo linear

Teste da hipótese $H_0: \beta_1 = 0$

- Para realizar este teste de significância, precisamos conhecer a **distribuição amostral de $\hat{\beta}_1$**
- Sob as suposições para regressão linear simples $\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1})$, onde a variância e o dp de $\hat{\beta}_1$ valem

$$V(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}} \Rightarrow \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

Na prática desconhecemos $\sigma_{\hat{\beta}_1}$ e utilizamos a estimativa empírica $s_{\hat{\beta}_1}$ em seu lugar

Validação do modelo linear

Teste da hipótese $H_0: \beta_1 = 0$

- A fim de construir um intervalo de confiança (IC) para o estimador $\hat{\beta}_1$ de β_1 normalizamos essa variável

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

- Essa estatística t corresponde à razão de uma v. a. normal por outra v. a. do tipo χ^2 e possui portanto distribuição t de Student—no caso, de $n - 2$ graus de liberdade (gl ou df)

Validação do modelo linear

Teste da hipótese $H_0: \beta_1 = 0$

- O intervalo de confiança de nível $1 - \alpha$ para a estatística t é dado por

$$P(-t_{\alpha/2, n-2} < t < t_{\alpha/2, n-2}) \geq 1 - \alpha$$

- Encontrando o valor crítico $t_{\alpha/2, n-2}$ em uma tabela da distribuição t de Student estabelecemos o IC de nível $100\%(1 - \alpha)$ para a inclinação β_1 da linha de regressão da população na forma

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot s_{\hat{\beta}_1}$$

Validação do modelo linear

Teste da hipótese $H_0: \beta_1 = 0$

Normalmente $\alpha = 0,05$, i.e., confiança $c = 1 - \alpha = 0,95$

Hipótese nula: $H_0: \beta_1 = \beta_{10}$

Valor da estatística do teste: $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Hipótese alternativa

Determinação do valor- p

$H_a: \beta_1 > \beta_{10}$

Área sob a curva t_{n-2} à direita de t

$H_a: \beta_1 < \beta_{10}$

Área sob a curva t_{n-2} à esquerda de t

$H_a: \beta_1 \neq \beta_{10}$

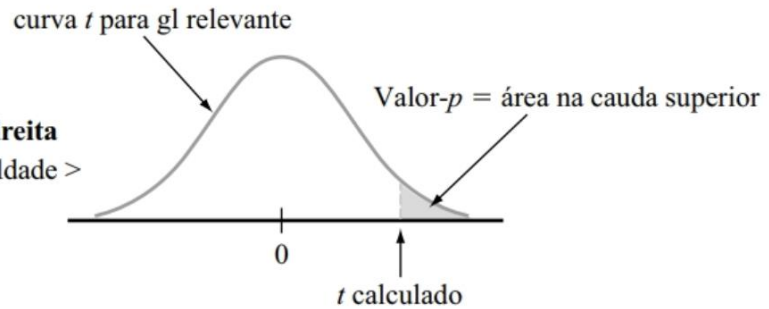
$2 \cdot (\text{Área sob a curva } t_{n-2} \text{ à direita de } |t|)$

O teste de utilidade do modelo é o teste de $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, sendo o valor da estatística de teste a razão $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

Validação do modelo linear

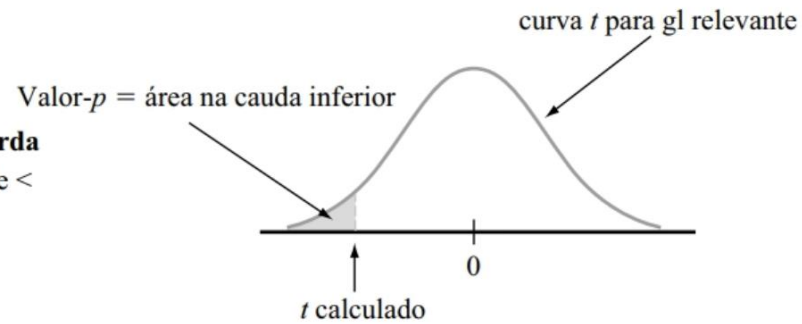
1. Teste unilateral à direita

H_a contém a desigualdade $>$



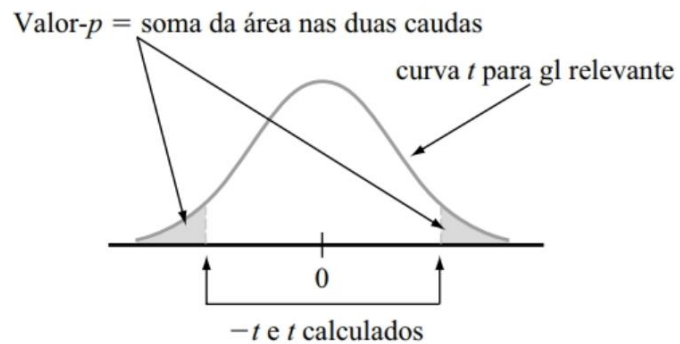
2. Teste unilateral à esquerda

H_a contém a desigualdade $<$



3. Teste bilateral

H_a contém a desigualdade \neq



Predição de valores futuros

Um intervalo de confiança para \hat{Y}

A partir das estimativas de $\hat{\beta}_0$ e $\hat{\beta}_1$ podemos estimar o valor de $\mu_{Y|x^*}$ para um valor dado de x^* a partir do estimador $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Essa **estimativa pontual** não fornece informações sobre a precisão da estimativa de \hat{Y} ...

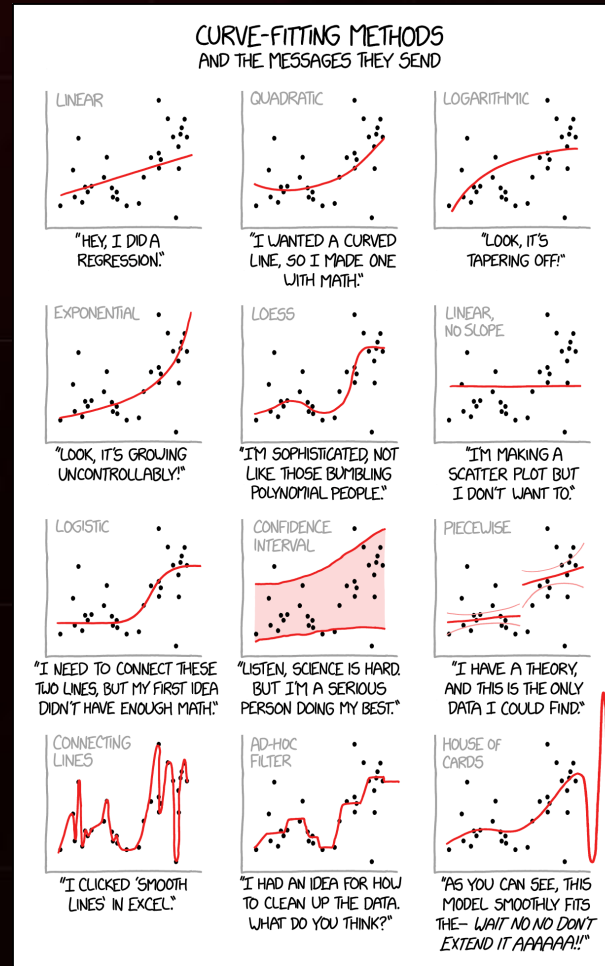
Solução: construir um **intervalo de confiança** para \hat{Y} usando o desvio padrão de $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

$$s_{\hat{Y}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Modelos não lineares

Às vezes a regressão linear
simples não funciona...

(Fonte xkcd #2048)



Modelos não lineares

Inadequação de modelos lineares

- Caso se constate que um modelo de regressão linear simples não descreve adequadamente os dados, pode-se tentar modelos não lineares ou de regressão múltipla
- Análise de adequação: **resíduos padronizados**

$$e_i^* = \frac{y_i - \hat{y}_i}{s \sqrt{1 - \frac{1}{n} - \frac{(x_i^* - \bar{x})^2}{S_{xx}}}}$$

quase todos os e_i^* recaem no intervalo $(-2, 2)$

Modelos não lineares

Nem todo modelo não linear é difícil

Os principais modelos não lineares usados em uma primeira abordagem são obtidos de um modelo linear por transformações de variáveis

Tabela 13.1 Funções intrinsecamente lineares úteis*

Função	Transformação(ões) para linearizar	Forma linear
a. Exponencial: $y = \alpha e^{\beta x}$	$y' = \ln(y)$	$y' = \ln(\alpha) + \beta x$
b. Potência: $y = \alpha x^{\beta}$	$y' = \log(y), x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
c. $y = \alpha + \beta \times \log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$
d. Recíproca: $y = \alpha + \beta \cdot \frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$

* Quando $\log(\cdot)$ aparece, tanto a base 10 quanto a base e podem ser usadas.

Modelos não lineares

Nem todo modelo não linear é difícil

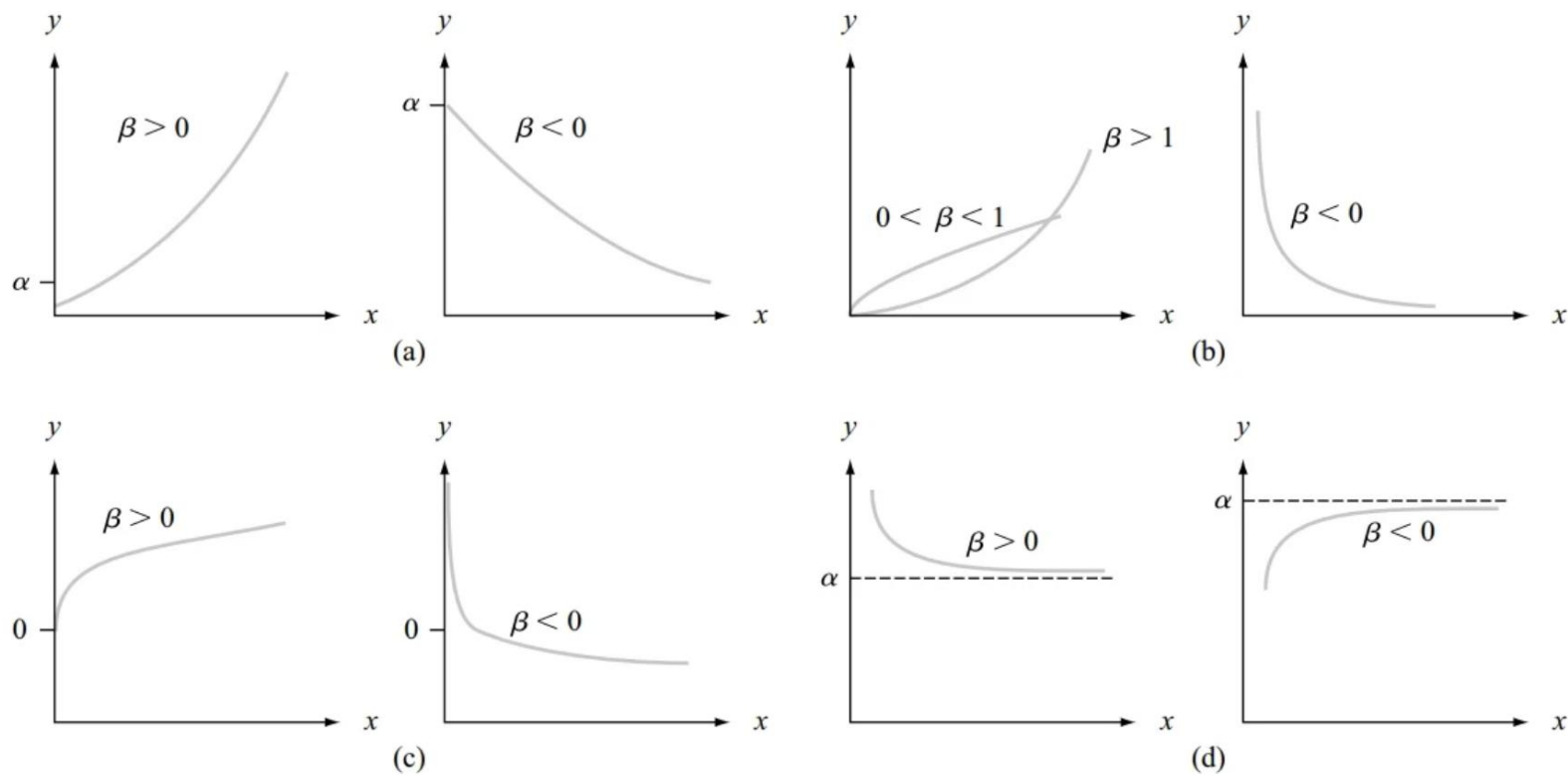


Figura 13.3 Gráficos das funções intrinsecamente lineares dadas na Tabela 13.1.

Modelos não lineares

Modelos intrinsecamente lineares

- Um modelo relacionando Y com x é **intrinsecamente linear** se por meio de uma transformação em Y ou x ou ambos puder ser **reduzido** a um modelo probabilístico linear $Y' = \beta_0 + \beta_1 x' + \varepsilon'$
- A vantagem de um modelo intrinsecamente linear é que os parâmetros β_0 e β_1 do modelo transformado podem ser imediatamente estimados usando o princípio dos mínimos quadrados simplesmente substituindo x' e y' nas fórmulas de estimativa

Métodos de regressão gerais

Modelos logístico e polinomial

- Muitas vezes a variável de resposta Y não tem suporte em toda a reta—por exemplo, é uma variável binária (0 ou 1), de contagem, estritamente positiva...
- Nestes casos, podemos empregar **modelos lineares generalizados** para ajustar os dados
- Os dois modelos lineares generalizados são o **modelo logístico** e o **modelo polinomial**

Métodos de regressão gerais

Modelos logístico

- Y assume dois valores (0 ou 1) dependendo do valor de alguma variável quantitativa $P(Y_i = 1 \mid x_i) = p(x_i)$
- Um modelo para $p(x)$ bastante útil é aquele que relaciona a **razão das chances** $p(x)/(1 - p(x))$ com

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

- Tomando o logaritmo natural em ambos os lados obtemos o **modelo logístico (linear no preditor x)**

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Métodos de regressão gerais

Modelos polinomiais

- Y não se comporta de maneira monotônica, às vezes crescendo em um intervalo e decrescendo em outro
- Nestes casos, podemos empregar **modelos polinomiais** para ajustar os dados

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_k x^k + \varepsilon$$

- As estimativas para $\beta_0, \beta_1, \dots, \beta_k$ são obtidas pela resolução de um **sistema linear** $(k + 1) \times (k + 1)$

Métodos de regressão gerais

Modelos polinomiais

- No modelo de regressão polinomial, a estimativa da **variância do erro** é dada por

$$\hat{\sigma}^2 = s^2 = \frac{SQE}{n - (k + 1)}$$

o denominador $n - (k + 1)$ indica que $k + 1$ graus de liberdade são perdidos na estimativa de $\beta_0, \beta_1, \dots, \beta_k$

- O **coeficiente de determinação múltipla ajustado** vale

$$R^2(\text{ajustado}) = 1 - \frac{n - 1}{n - (k + 1)} \cdot \frac{SQE}{SQT}$$

Modelos de regressão múltipla

Modelos de regressão múltipla

- Modelo probabilístico que relacione uma variável dependente Y a mais de uma variável preditora

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

- O coeficiente de regressão β_i fornece a variação esperada em Y devido a uma variação em x_i enquanto as outras variáveis preditoras $x_j, j \neq i$, se mantêm constantes

Modelos de regressão múltipla

Modelos com interação e preditores quadráticos

- Modelos de regressão múltipla podem ser usados para formular **modelos com interação entre as variáveis preditoras**
- O modelo com interação mais geral envolvendo no máximo **produtos quadráticos** possui a forma geral (usando somente duas variáveis preditoras, x_1 e x_2 , no exemplo, mas poderiam ser em maior número)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

Modelos de regressão múltipla

Teste de utilidade de modelos com interação e preditores quadráticos

- Testamos a hipótese H_0 de que todos os $\beta_i = 0$ usando a estatística de teste

$$f = \frac{R^2 / k}{(1 - R^2) / [n - (k + 1)]}$$

onde $R^2 = 1 - SQE/SQT = SQR/SQT$

- A estatística f possui distribuição $F_{k, n-(k+1)}$ de Fisher-Snedecor com k graus de liberdade “em cima” e $n - (k + 1)$ graus de liberdade “embaixo”

Dicas finais

- As videoaulas de “**Revisando conhecimentos**” são valiosas, em particular nas **3 primeiras semanas**
- As videoaulas são integralmente baseadas nos exemplos resolvidos do texto-base de Jay L. Devore, *Probabilidade e Estatística para Engenharia e Ciências*, (trad. 9ª ed.), Capítulos 12 e 13
- Estudem com base nos **exemplos resolvidos do texto-base**—eles estão muito mais claros e detalhados do que nas videoaulas (devido à curta duração das videoaulas)

Bom estudo e bom final de ano a tod@s!

