

# Modelos Lineares

Adaptado dos slides da Profa. Suzi Camey

# Informações úteis

- Statistical Modeling for Biomedical Researchers
  - Bancos de dados:  
<http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/>
- Statistical Associates "Blue Book"  
<http://www.statisticalassociates.com/booklist.htm>
- UCLA  
<http://www.ats.ucla.edu/stat/>

# Informações úteis

- Blog sobre estatística:
  - <http://www.theanalysisfactor.com/>
  - <http://theanalysisinstitute.com/>
- Khan Academy:
  - <https://www.khanacademy.org/math/probability/regression>
- SPSS:
  - Conteúdo do HELP
    - *Show me*
    - *Algorithms*

# Modelos Estatísticos

- Ferramentas para abordar duas questões importantes
  - Predição (ou Previsão)
    - Ajuste de uma equação matemática de forma a prever resultados ainda não observados
    - Exemplo: prever o custo de um tratamento a partir de idade, sexo, tipo e tempo de tratamento
  - Explicação
    - Identificar e medir o impacto de uma ou mais variáveis em outra
    - Exemplo: quais fatores estão relacionados a pressão sistólica de um indivíduo (idade, status de fumo, IMC, ...)

# Modelos Estatísticos

- Alguns tipos
  - Modelos Lineares
  - Regressão Logística
  - Regressão de Poisson
  - Modelos de sobrevida
  - Modelos para dados longitudinais

# Modelos Lineares

- Modelos que buscam explicar um desfecho quantitativo através de um ou mais preditores
  - Desfecho = variável dependente
  - Preditores = variáveis independentes, explicativas, fatores
- Nos modelos lineares, o desfecho deve ter distribuição Normal.

# Modelos Lineares

- Casos especiais:
  - Regressão: preditor de interesse também é quantitativo
    - Pode ser simples ou múltipla
  - Análise de variância: preditor de interesse é qualitativo
  - Análise de covariância: preditor de interesse é qualitativo, mas é essencial “ajustar” por outro preditor quantitativo

# Regressão Linear Simples

Adaptado dos slides da Profa. Suzi Camey



# Regressão Linear Simples

- Para que serve?
  - Relacionar uma única variável independente quantitativa com o desfecho quantitativo
  - Ajusta uma reta
  - Explicação e/ou previsão

# Regressão Linear Simples

- Exemplo: Estudo entre a relação entre infecção por ancilóstomo (número de vermes) e perda de sangue. Estudo conduzido em 1970 na Tailândia.
  - 15 pacientes com anemia devida a este verme
  - Banco Suwit.sav

# Primeiro passo

- Descritiva
  - Primeiro contato com os dados
  - Encontrar observações estranhas
    - Erros de digitação
    - Observações raras
    - Informações duplicadas

# Descritiva

---

# Variáveis

---

- X: imc
- Y: fat\_Brozek

# Segundo passo

- “Verificar” linearidade
  - Gráfico de dispersão
  - Identificar pontos

# Gráfico

---

# Terceiro passo

---

- Estimar modelo



# Gráfico com reta

---

# Outro exemplo

- Banco: fat\_dat.sav
- Fitting Percentage of Body Fat to Simple Body Measurements
  - <http://www.amstat.org/publications/jse/datasets/fat.txt>

# Descritiva

---

# Correções de erros

- Ler as “Special Notes” no link de descrição do banco de dados.
  - The data are as received from Dr. Fisher. Note, however, that there are a few errors. The body densities for cases 48, 76, and 96, for instance, each seem to have one digit in error as can be seen from the two body fat percentage values. Also note the presence of a man (case 42) over 200 pounds in weight who is less than 3 feet tall (the height should presumably be 69.5 inches, not 29.5 inches)! The percent body fat estimates are truncated to zero when negative (case 182).

# Variáveis

- X:

- Y:

# Gráfico

---

# Gráfico com reta

---

# Modelo teórico

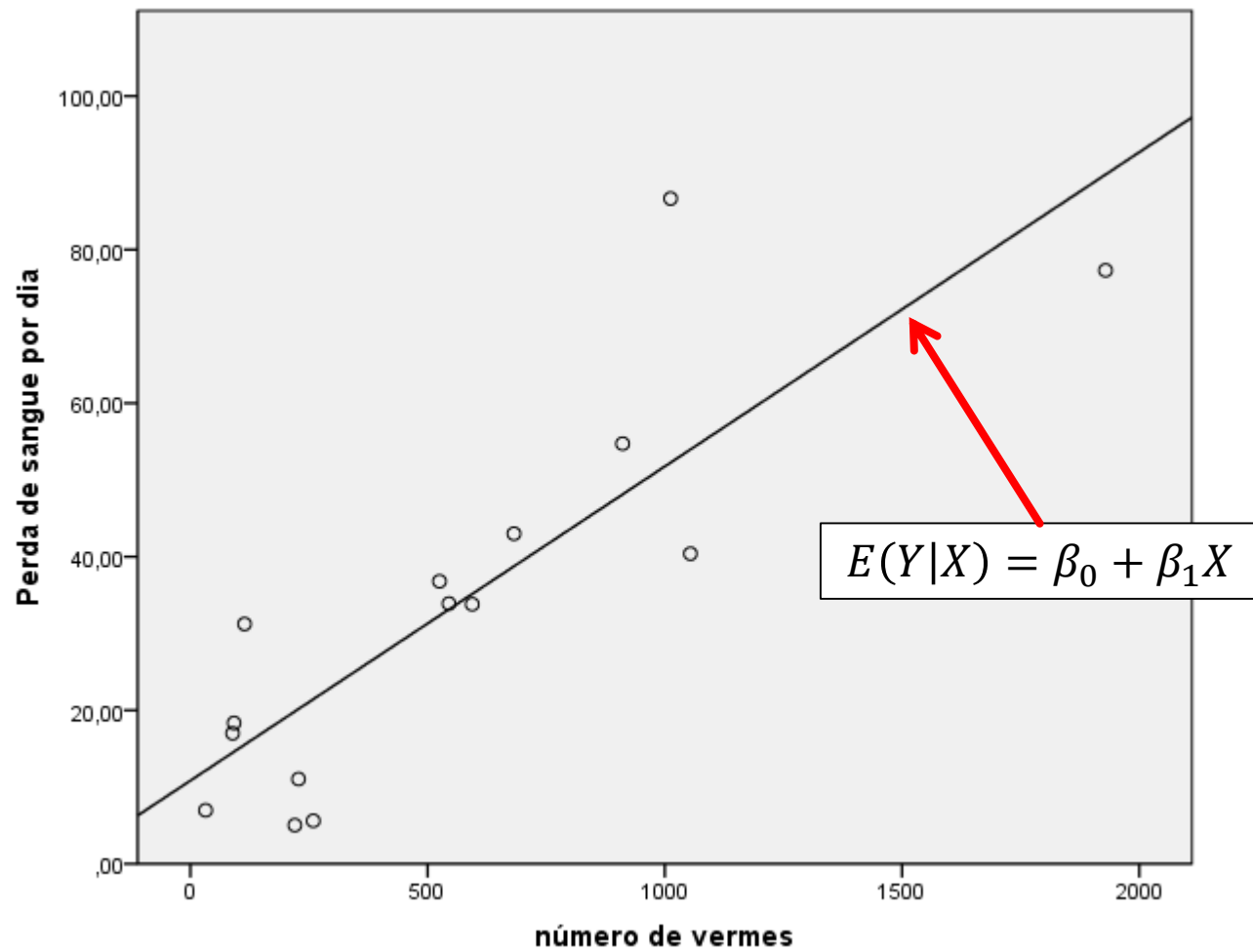
- Supondo uma relação linear entre X e Y:

$$E(Y|X) = \beta_0 + \beta_1 X$$

Intercepto

Inclinação





$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Parâmetros desconhecidos

Componente de erro aleatório

- $i=1, \dots, n$  onde  $n$  é o número de pacientes

# Estimação

- A partir dos dados temos uma estimativa, ou seja:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

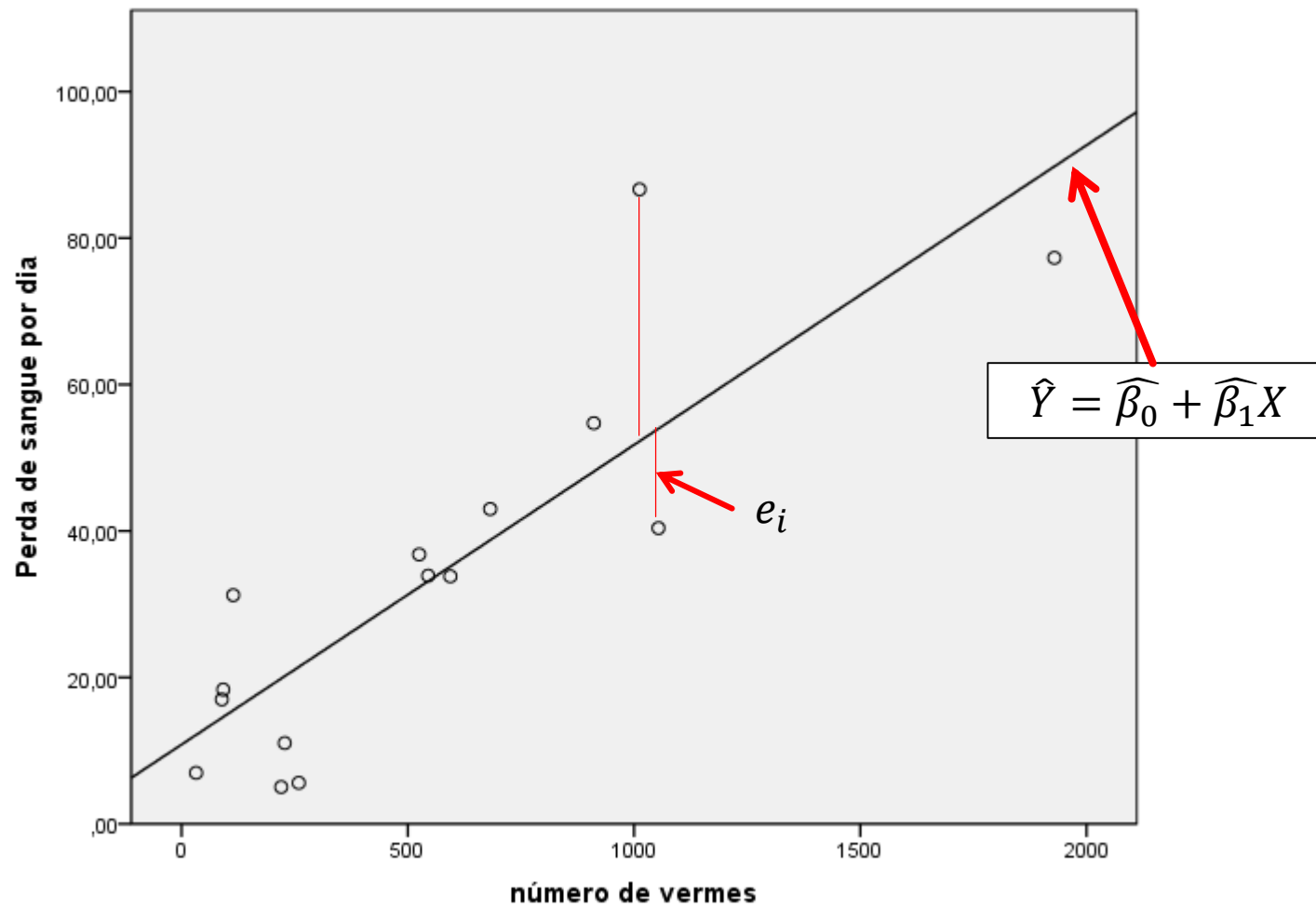
Diagram illustrating the components of the linear regression equation:

- $\hat{Y}_i$  is labeled as "Valor predito de Y".
- $\hat{\beta}_0$  and  $\hat{\beta}_1$  are labeled as "Estimativas dos parâmetros".

# Como estimar $\beta_0$ e $\beta_1$

- Resíduo=Observado – Predito

$$e_i = Y_i - \hat{Y}_i$$



# Qual é a melhor reta?

- Aquela com menores resíduos
  - Método dos mínimos quadrados

$$\begin{aligned}\sum_{i=1}^n (e_i^2) &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \\ &= \sum_{i=1}^n \left( Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X) \right)^2\end{aligned}$$

# Método dos mínimos quadrados

$$\widehat{\beta}_1 = \frac{SP_{XY}}{SQ_X} = \frac{Cov(X, Y)}{Var(X)}$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

Onde:  $SP_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$  e

$$SQ_X = \sum_{i=1}^n (X_i - \bar{X})^2$$

# Método dos mínimos quadrados

- Propriedades
  - Soma dos resíduos é igual a zero
  - A reta passa pelo ponto  $(\bar{X}, \bar{Y})$



# Suposições

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Para fazermos inferências sobre o modelo temos que fazer algumas suposições:

- Linearidade
- Distribuição dos erros

# Distribuição dos erros

- Tem média zero.
  - Tem distribuição Normal.
  - Tem variância constante ( $\sigma^2 = Var(\varepsilon)$ ) para cada valor de  $X$  (Homocedasticidade).
  - Independentes.
- 
- Resumindo: os erros são i.i.d. (independentes e identicamente distribuídos).

$$\varepsilon \sim N(0, \sigma^2)$$

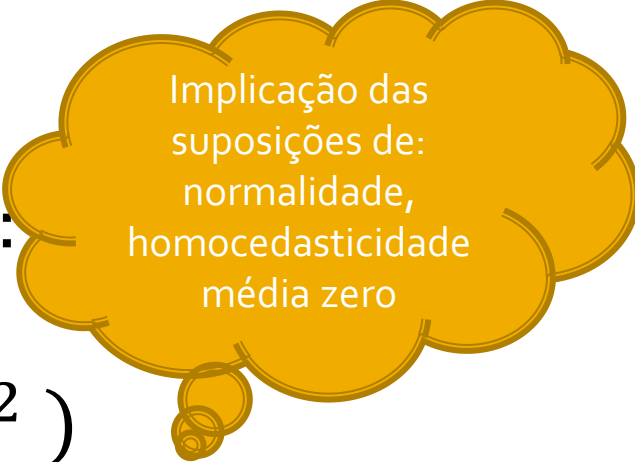
# Distribuição do erro

Como

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

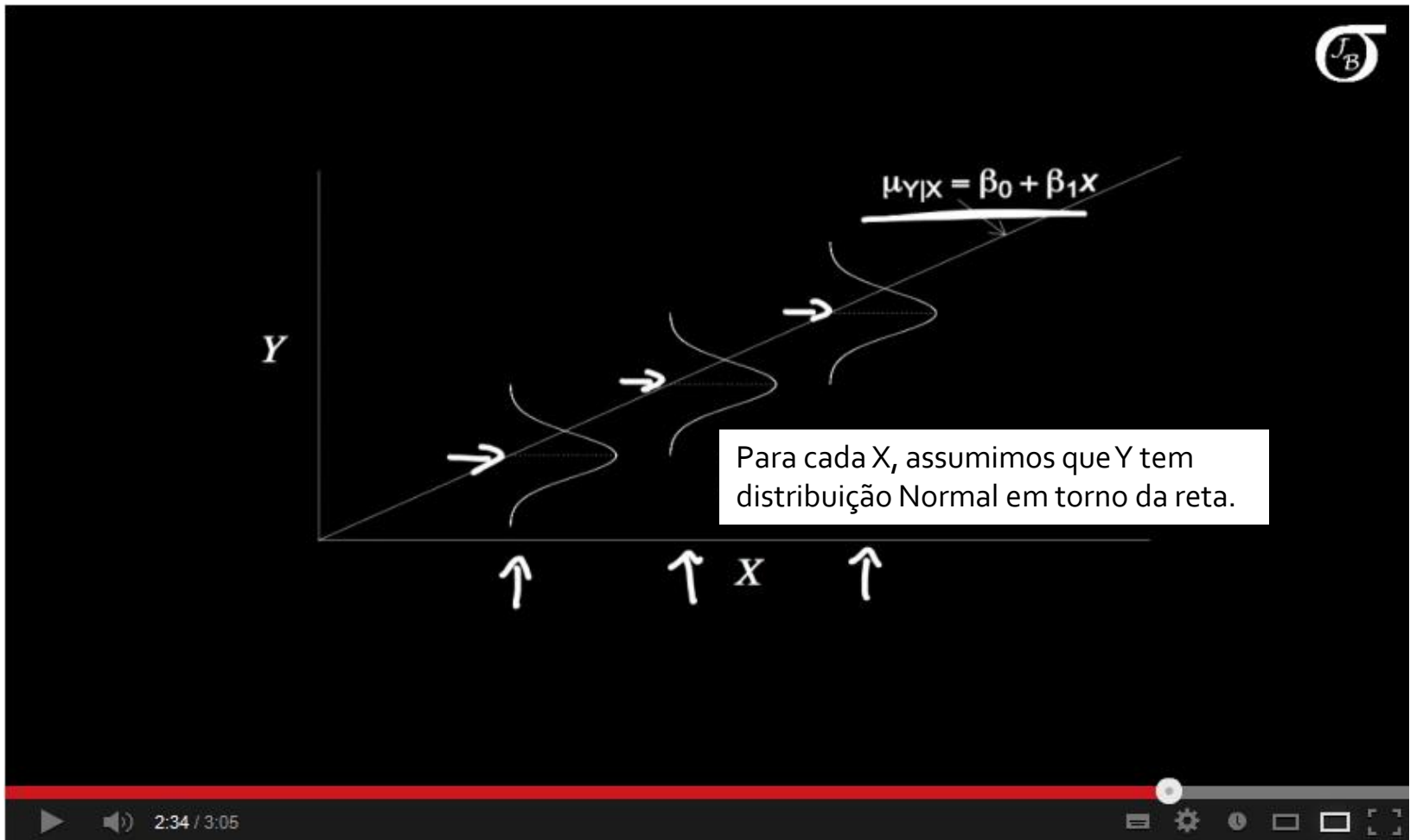
a distribuição do erro implica que:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i; \sigma^2)$$



Implicação das  
suposições de:  
normalidade,  
homocedasticidade  
média zero

# Gráfico



# Propriedades dos estimadores

- Não viciados.

$$E(\widehat{\beta}_0) = E(\bar{Y} + \widehat{\beta}_1 \bar{X}) = \beta_0$$

$$E(\widehat{\beta}_1) = E\left(\frac{SP_{XY}}{SQ_X}\right) = \beta_1$$

- Possuem a menor variância.

- Consistentes.

$$\widehat{\beta}_0 \xrightarrow{n \rightarrow \infty} \beta_0$$

$$\widehat{\beta}_1 \xrightarrow{n \rightarrow \infty} \beta_1$$

# Estimação de $\beta_1$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- Se  $\beta_1 = 0$ , então não há relação **linear** entre X e Y.
- $\widehat{\beta}_1$  é uma variável aleatória, logo tem uma distribuição:
  - Normal.
  - $E(\widehat{\beta}_1) = \beta_1$
  - $\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{SQ_X}$

# Estimação de $\sigma^2$

Implicações das  
suposições do modelo:  
 $Var(Y) = Var(\varepsilon) = \sigma^2$

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$$

E portanto o erro padrão de  $\hat{\beta}_1$  é

$$EP(\hat{\beta}_1) = \frac{s}{\sqrt{SQ_X}}$$

# Intervalo de confiança para $\beta_1$

Intervalo de  $(1 - \alpha) \times 100\%$  de confiança para  $\beta_1$  é:

$$\widehat{\beta}_1 \pm t_{(n-2; \alpha/2)} \times EP(\widehat{\beta}_1)$$



# Testar significância do preditor

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

Estatística do teste:

$$t = \frac{\widehat{\beta}_1 - \beta_1}{EP(\widehat{\beta}_1)} \sim t_{(n-2)}$$

# Regressão Banco Suwit

- ICs e teste de significância do preditor

# Sempre faça o gráfico de dispersão!

- Arquivo Anscombe.sav
  - Construído por Francis Anscombe (1973)
- Fazer as descritivas de  $X_1, X_2, X_3, X_4$

# Sempre faça o gráfico de dispersão!

- Fazer as regressões.

# Sempre faça o gráfico de dispersão!

- Fazer os 4 gráficos

1

2

3

4

# Diferença entre estimar $E(Y|X^*)$ e prever $Y|X^*$

- Preciso saber quanto em média um paciente com 100 vermes perde de sangue.
- Preciso saber quanto um paciente com 100 vermes perdeu de sangue.

# Intervalo de confiança para $E(Y|X^*)$ e Intervalo de predição para $Y|X^*$

Notação:  $\mu_{Y|X^*} = E(Y|X^*)$

- Intervalo de confiança para  $E(Y|X^*)$ :

$$\hat{\mu}_{Y|X^*} \pm t_{(n-2; \alpha/2)} \times EP(\hat{\mu}_{Y|X^*})$$

- Intervalo de predição para  $Y|X^*$  (Y para um valor  $X^*$ ):

$$\hat{Y} \pm t_{(n-2; \alpha/2)} \times EP(Y_{pred})$$

# Intervalo de confiança para $E(Y|X^*)$ e Intervalo de predição para $Y|X^*$

- $EP(\hat{\mu}_{Y|X^*})$ :

$$Var(\hat{\mu}_{Y|X^*}) = \sigma^2 \left( \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

$$EP(\hat{\mu}_{Y|X^*}) = s^2 \left( \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

- $EP(Y_{pred})$ :

$$Var(Y_{pred}) = \sigma^2 \left( \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) + \sigma^2$$

$$EP(Y_{pred}) = s^2 \left( 1 + \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$



# Banda de confiança para $E(Y|X^*)$

---

# Banda de predição para $Y|X^*$

---

# Tabela da ANOVA

---

# Medidas descritivas da associação linear entre X e Y

- Coeficiente de determinação ( $R^2$ ):

$$R^2 = \frac{SQR}{SQT}$$

- $0 < R^2 < 1$
- Limitações de  $R^2$ :
  - Alto  $R^2$  **não** implica em predições úteis. Ex.: Intervalo de predição do exemplo dos vermes.
  - Alto  $R^2$  **não** implica em bom ajuste. Ex.: Anscombe
  - $R^2$  próximo de zero **não** implica que X e Y são não relacionados. Ex.: Gerar X e  $X^2$
- Correlação ( $r$ ):

$$r = \pm\sqrt{R^2}$$

# Avisos sobre Regressão

- Não fazer previsões para observações fora do intervalo de variação das variáveis estudadas.
- $\beta_1 \neq 0$  não implica em relações de causa e efeito.
- Uma suposição que está implícita é:  
X é medida sem erros

# Avisos sobre Regressão

- Para todos os testes e intervalos serem válidos, as suposições têm que ser atendidas.
  - Independência
  - Linearidade
  - Homocedasticidade
  - Normalidade
    - Em quase todas as situações um tamanho de amostra grande contorna a falta de normalidade dos erros, a exceção é no caso do intervalo de predição para  $Y|X$

# EXERCÍCIO

- Para o banco `fat_dat`, rodar as regressões simples de IMC versus os dois índices de gordura.
  - Interpretar os coeficientes e intervalos de confiança.
  - Interpretar a ANOVA e  $R^2$ .
  - Obter os gráficos com predição individual e média.

# Diagnóstico em Regressão Linear Simples

Adaptado dos slides da Profa. Suzi Camey



# Análise de resíduos

- Não padronizado:  $e_i = Y_i - \hat{Y}_i$
- Padronizado:  $e_i^* = \frac{e_i}{s\sqrt{1-h_i}}$
- Studentizado:  $e_i^{**} = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}$ 
  - $h_i$  é uma medida de alavancagem;
  - $s_{(i)}$  é a estimativa do desvio padrão dos resíduos ao se deletar a i-ésima observação.
- O resíduo studentizado tem distribuição t de Student com n-3 graus de liberdade.

# Diagnósticos: gráficos de resíduos

- Preditor X Resíduos ( $X$  x  $e$ ) :
  - Linearidade
  - Heterocedasticidade
  - Outliers
- Valores preditos (ajustados) X Resíduos ( $\hat{Y}$  x  $e$ ):
  - Linearidade
  - Heterocedasticidade
  - Outliers
- Tempo x Resíduos ou Ordem x Resíduos
  - Não independência dos erros

# Diagnósticos: gráficos de resíduos

- Preditores não incluídos X Resíduos
  - Omissão de preditores
- Boxplot, histograma e qqplot dos resíduos
  - Outliers
  - Normalidade
- Usamos os resíduos padronizados para os gráficos.

# Diagnósticos: gráficos de residuos

- No SPSS:
  - <https://stats.idre.ucla.edu/spss/webbooks/reg/chapter2/spss-webbooksregressionwith-spsschapter-2-regression-diagnostics/>
  - <https://stats.idre.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson2/>

# Diagnósticos: gráficos de resíduos

Usar os gráficos apropriados nos exemplos indicados.

- Não linearidade. Exemplo Anscombe .
- Outliers. Exemplo Anscombe .
- Heterocedasticidade. Exemplo fat\_dat.
- Não independência dos erros. Exemplo fat\_dat.
- Não normalidade dos erros. Exemplo fat\_dat.
- Omissão de preditores. Exemplo fat\_dat.

# Exemplo Anscombe

---

# Exemplo fat\_dat

---

# Diagnósticos: gráficos de resíduos

Existem testes para verificar:

- Normalidade dos resíduos: Shapiro-Wilk
- Homocedasticidade: Breusch-Pagan
- Outliers: resíduo studentizado
- Independência: Durbin-Watson (d)
  - $0 < d < 4$
  - $d = 2$ ; ausência de autocorrelação



# Medidas para remediar problemas nos resíduos

- Abandonar a regressão linear e buscar modelos mais adequados.
- Aplicar transformações nos dados para adequar ao modelo de regressão.

# Como remediar violação das suposições

- Não linearidade: relação quadrática, exponencial
  - Transformações
- Heterocedasticidade:
  - usar mínimos quadrados ponderados
  - Transformações
- Outliers:
  - Estudar cuidadosamente se o valor não deve ser retirado, se não for retirado indicado usar regressão robusta

# Como remediar violação das suposições

- Não independência dos erros:
  - Usar modelos que incluam autocorrelação
- Não normalidade dos erros
  - Transformações ou regressão robusta
- Omissão de preditores
  - Incluir preditores

# Observações Influentes

- Leverage (valor de alavancagem)
  - Mede o afastamento do valor de X uma observação em relação às demais observações.
  - Varia de zero a um. Valores acima de 0,2 são considerados altos (Dupont et al)

# Observações Influentes

- Distância de Cook (D de Cook)
  - Aumenta quanto maior for o resíduo da observação e sua *leverage*.
  - Valores grandes do D de Cook (maiores que  $4/n$ ) indicam que a exclusão daquela observação mudaria a reta estimada de maneira importante.
- Gráfico Leverage x D de Cook, com identificação dos casos.
  - Identifica pontos extremos.

# Exemplo fat\_dat

---

# Lowess (locally weighted scatter plot smoothing)

- Ajusta uma forma no diagrama de dispersão sem fazer suposição sobre esta forma.
- Princípios:
  - Cada observação  $(X_i, Y_i)$  é ajustada à uma regressão linear separada baseada nos dados vizinhos.
  - A ponderação é feita de modo que pontos mais afastados de  $X_i$  tenham menos influência no ajuste.
  - Usamos o termo “janela” para designar a proporção de dados considerada para a estimação de  $\hat{Y}_i$ .

# Lowess (locally weighted scatter plot smoothing)

- Para bancos de dados grandes essa técnica é computacionalmente pesada.
- Para bancos de dados grandes janelas de 0,3 ou 0,4 geralmente funcionam melhor; uma janela de 0,99 é recomendada para bancos pequenos
- Referência: **Statistical Modeling for Biomedical Researchers : A Simple Introduction to the Analysis of Complex Data.**



# Lowess (locally weighted scatter plot smoothing)

---

- Exemplo no SPSS:
  - Banco Suwit

# EXERCÍCIO

- Para o banco Suwit, rodar a regressão e reproduzir as técnicas de diagnóstico vistas. Aplicar também a técnica LOWESS.