Modelos Lineares Generalizados - Introdução

Erica Castilho Rodrigues

18 de Março de 2014

Introdução

Revisão - Regressão Mútipla

Por que queremos modelar os dados?

- A forma do modelo revela padrões de interação e associação nos dados.
- Através de procedimentos de inferência podemos verificar:
  - quais variáveis explicativas estão relacionadas com a variável resposta;
  - enquanto controlamos outras variáveis relevantes.
- A estimativa dos parâmetros fornece a importância de cada variável no modelo.

- Suponha que temos uma variável resposta contínua Y.
- Temos ainda um conjunto de variáveis explicativas:

$$X_1,\ X_2, \ldots,\ X_k\ .$$

- Queremos modelar Y como função das explicativas.
- Podemos usar um modelo de regressão

- Suponha que temos uma variável resposta contínua Y.
- Temos ainda um conjunto de variáveis explicativas:

$$X_1,\ X_2, \ldots,\ X_k\ .$$

- Queremos modelar Y como função das explicativas.
- Podemos usar um modelo de regressão

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

onde

- Y_i é a variável resposta observada no i-ésimo indivíduo;
- $\beta_0, \ldots, \beta_k$ são os coeficientes que descrevem o modelo;
- x_{ki} é a k-ésima variável explitiva observada no i-ésimo indivíduo;
- ϵ_i são erros aleatórios iid tais que

$$\epsilon_i \sim N(0, \sigma^2)\ .$$

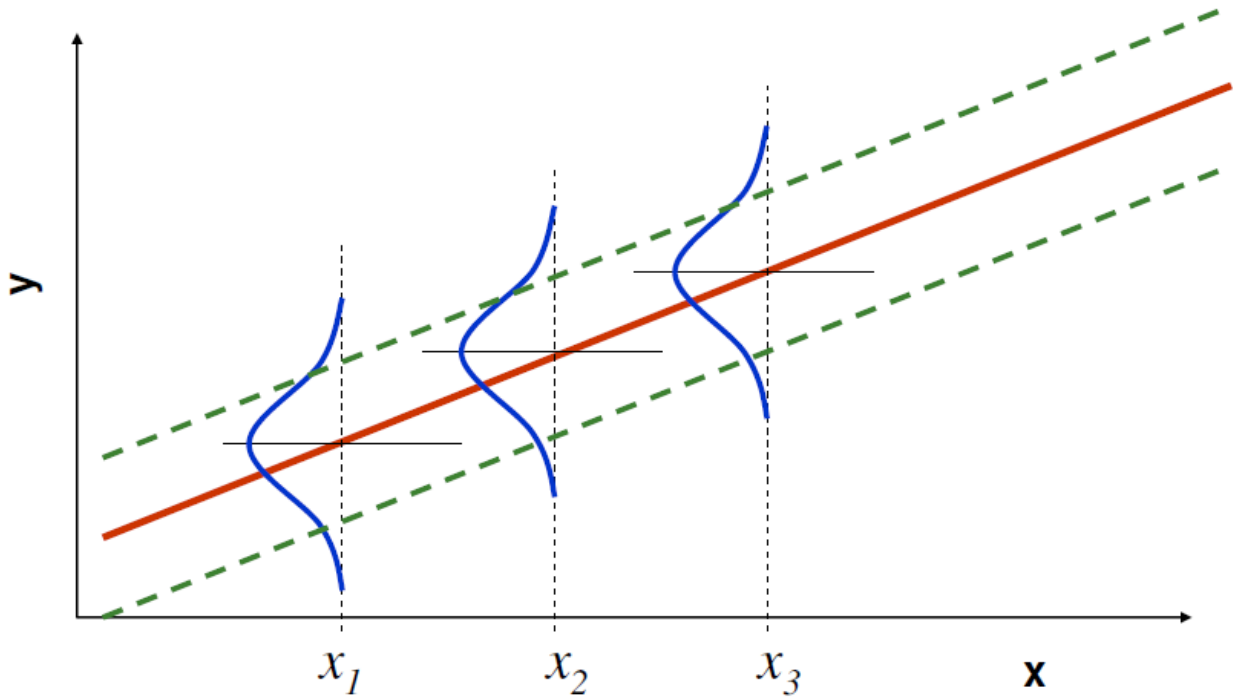
- Em geral podemos escrever

$$\mathbf{Y} = E(\mathbf{Y}) + \boldsymbol{\epsilon}$$

onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- O modelo de regressão linear simples pode ser representado da seguinte maneira.



- Algumas técnicas que podem ser aplicadas para esse tipo de modelo:
  - Regressão Linear Simples;
  - Regressão Linear Mútipla;
  - ANCOVA;
  - ANOVA.

- O estimador de mínimos quadrados do vetor β é dado por

- O estimador de mínimos quadrados do vetor $\boldsymbol{\beta}$ é dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- O valor predito de $\mathbf{Y}$ é dado por

- O estimador de mínimos quadrados do vetor β é dado por

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- O valor predito de $\mathbf{Y}$ é dado por

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

- A matriz $\mathbf{H}$ pode ser usada para fazer vários diagnósticos de ajuste.
- Nos Modelos Lineares Generalizados veremos muitas semelhanças.

- Todas as conclusões do modelo estão pautadas em um suposição forte:
  - o vetor **Y** tem distribuição normal.
- Muitas vezes essa suposição não será satisfeita.
- Se a resposta vor uma variável categórica, isso não será verdade.
- Se **Y** for quantitativa discreta a suposição também será violada.
- Precisamos, então, usar os Modelos Lineares Generalizados.

- Considere as seguintes situações:

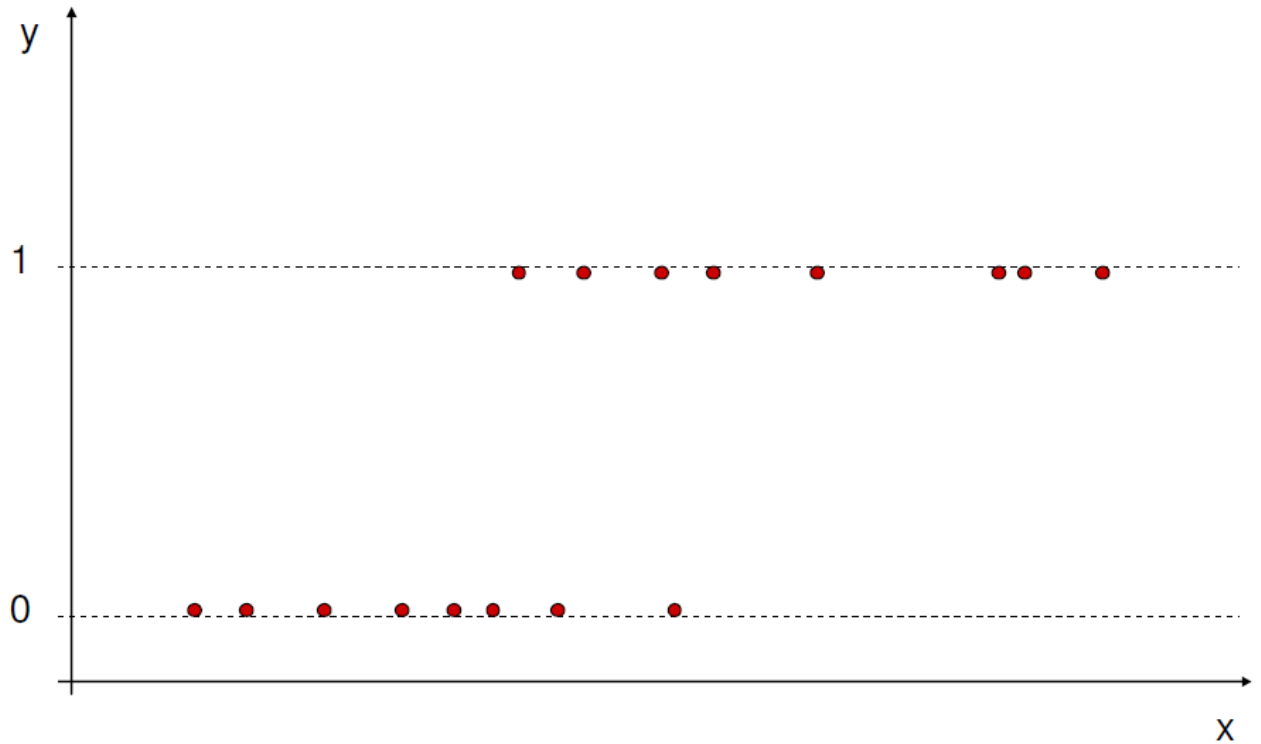
- Um paciente é submetido a um cirurgia.
- Deseja-se prever a chance do paciente sobreviver.
- Estima-se essa chance com base em dados clínicos pré-operatórios.

- Estamos analisando casos de dengue em municípios.
- Queremos tentar prever o número de casos.
- Podemos usar informações sócio-econômicas do município.

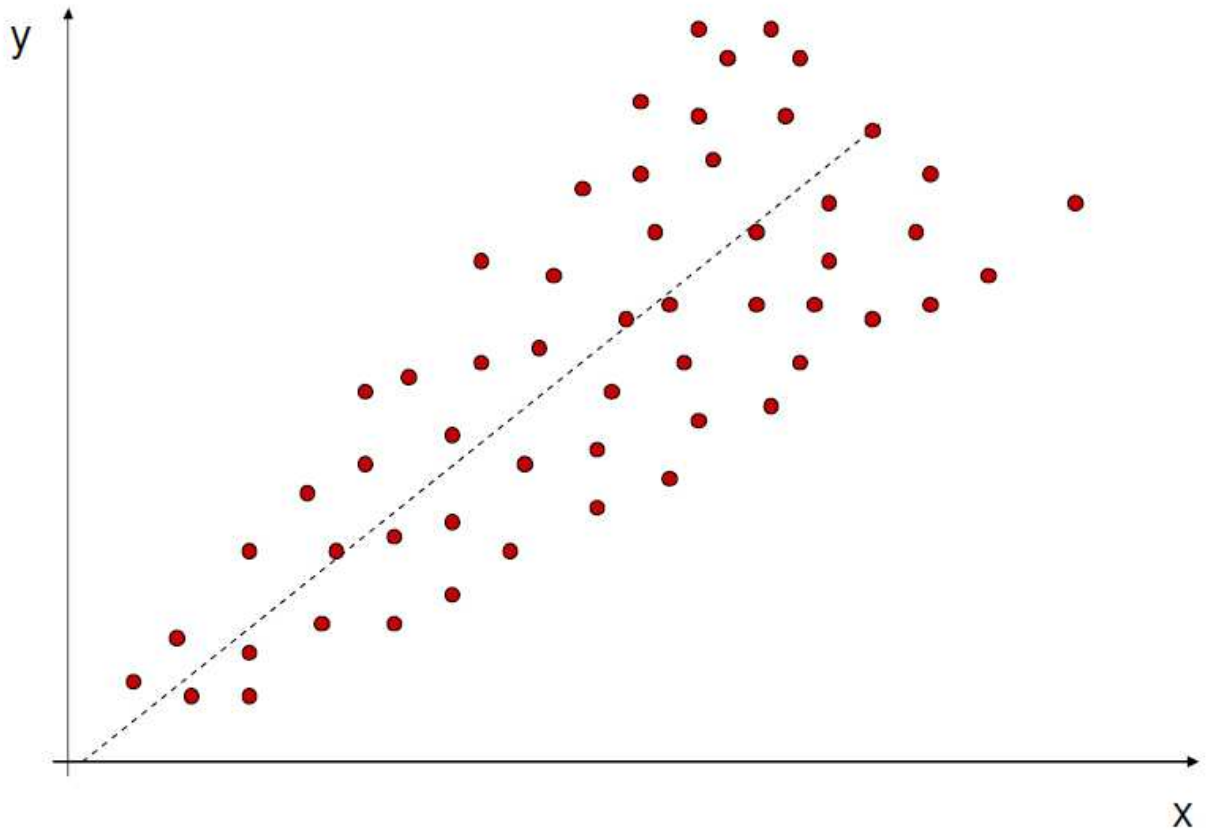
Objetivo da Disciplina

- Fornecer ferramentas para a análise de dados que não apresentam uma distribuição Normal:
  - Bernoulli/Binomial;
  - Poisson;
  - Binomial Negativa;
  - Gama.

- Veremos modelos para tratar com dados binários.



- Modelos apra lidar com dados heterocedásticos.



Revisão - Regressão Mútipla

Interpretação dos Coeficientes

Interpretação do β_j pra $j \geq 1$

- Considere o modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \alpha_1 Z_i + \epsilon_i \, .$$

Interpretação dos Coeficientes

Interpretação do β_j pra $j \geq 1$

- Considere o modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \alpha_1 Z_i + \epsilon_i \, .$$

- Mantemos constantes todos X_i com exceção do X_j.
- β_j representa a variação no Y quando X_j aumenta em uma unidade.
- Se $\beta_j > 0$, Y aumenta.
- Se $\beta_j < 0$, Y diminui.

Suposições do Modelo

- Os erros

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$$

tem média igual a zero e variância igual a σ^2.

- Isso implica que

$$E(Y_i|\mathbf{X}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

$$Var(Y_i|\mathbf{X}) = \sigma^2 \ .$$

- Os erros

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$$

são não correlacionados.

- Isso implica que os Y_i são não correlacionados.
- Os erros

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$$

têm distribuição normal.

- Isso implica que os Y_i têm distribuição normal.

Estimador de Mínimos Quadrados

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \, .$$

- Propriedades:

$$E(\hat{\beta}) =$$

Estimador de Mínimos Quadrados

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \, .$$

- Propriedades:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \, .$$

- A Tabela ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Regressão	p	$\hat{\beta}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2$	QMR=SQR/p
Resíduo	$n-p-1$	$\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$	QME=SQE/(n-p-1)
Total	$n-1$	$\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2$	

Tabela : Tabela ANOVA

Teste F da significância da Regressão

- Queremos testar as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0, \text{ ou seja, nenhum dos } \beta_j \text{ é significante}$$

$$H_1 : \beta_j \neq 0 \text{ para pelo menos um } j \geq 1, \text{ ou seja, } \beta_j \text{ é significante}$$

- A estatística de teste é dada por

$$F = \frac{QMR}{QME} \sim F_{p,n-p-1} \text{ sob } H_0$$

- Região crítica da forma

Teste F da significância da Regressão

- Queremos testar as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0, \text{ ou seja, nenhum dos } \beta_j \text{ é significante}$$

$$H_1 : \beta_j \neq 0 \text{ para pelo menos um } j \geq 1, \text{ ou seja, } \beta_j \text{ é significante}$$

- A estatística de teste é dada por

$$F = \frac{QMR}{QME} \sim F_{p,n-p-1} \text{ sob } H_0$$

- Região crítica da forma $F_{obs} > F$.
Observação: não estamos testando o intercepto, queremos verificar quais variáveis são significativas.

Testes t individuais

- Para $j = 1, 2, \ldots, q$, podemos testar a signicância do coeficiente β_j na presença dos demais coeficientes no modelo:

$$H_0 : \beta_j = 0 \text{ na presença dos demais coeficientes ;}$$

$$H_1 : \beta_j \neq 0 \text{ na presença dos demais coeficientes.}$$

- A estatística de teste é dada por

$$t_j = \frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \sim t_{n-p-1} \text{ sob } H_0 \, .$$

- Teste bilateral.

Variáveis Indicadoras

- A variáveis categóricas podem também ser vistas como subgrupos dos dados.
- Nesse caso as variáveis indicadoras recebem 1 se o indivíduo pertence ao grupo e 0 caso contrário.
- Elas permitem representar vários grupos em uma única equação.
- Não precisamos escrever uma equação para cada grupo.
- As variáveis indicadoras podem ser tratadas como qualquer outra no modelo de regressão.

- Suponha que queremos comparar dois grupos:
  - grupo controle e tratamento.
- Vamos definir a variável Z_i tal que

$$Z_i = \begin{cases} 1 & \text{se o i-ésimo indivíduo percente ao grupo tratamento} \\ 0 & \text{se o i-ésimo indivíduo percente ao grupo controle.} \end{cases}$$

- O modelo fica

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \alpha_1 Z_i + \epsilon_i$$

onde

- Y_i é a variável resposta,
- X_i's são variáveis explicativas quantitativas.

- Qual parâmetro estima a difrenreça entre os grupos?

- Qual parâmetro estima a difrenreça entre os grupos? α_1.
- Interpretação de α_1?
- Aumento (ou redução) esperado em Y quando passamos do grupo controle para o grupo tratamento.

Análise de Resíduos

- O erro ϵ_i é estimado pelo resíduo e_i

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

- Representa a quantidade da varilibilidade que Y que o modelo ajustado não consegue explicar.
- Os resíduos contem informação sobre o motivo do modelo não ter se ajustado bem aos dados.
- Conseguem indicar se uma ou mais suposições do modelo foram violadas.

- Principais problemas detectados através da análise dos resíduos:
  - Não-linearidade da relação entre X e Y ;
  - Não normalidade dos erros;
  - Variância não-constante dos erros (heterocedasticidade);
  - Correlação entre os erros;
  - Presença de *outliers* ou observações atípicas;
  - O modelo foi mal especificado.

Gráficos para análise de resíduos

Gráficos para análise de resíduos

- Gráfico de Probabilidade Normal dos resíduos;
- Gráfico dos resíduos versus valores de Y;
- Gráfico dos resíduos versus valores de X (incluída no modelo);
- Gráfico dos resíduos versus outras Xs (não incluídas no modelo);
- Gráfico dos resíduos versus tempo ou ordem de coleta dos dados.

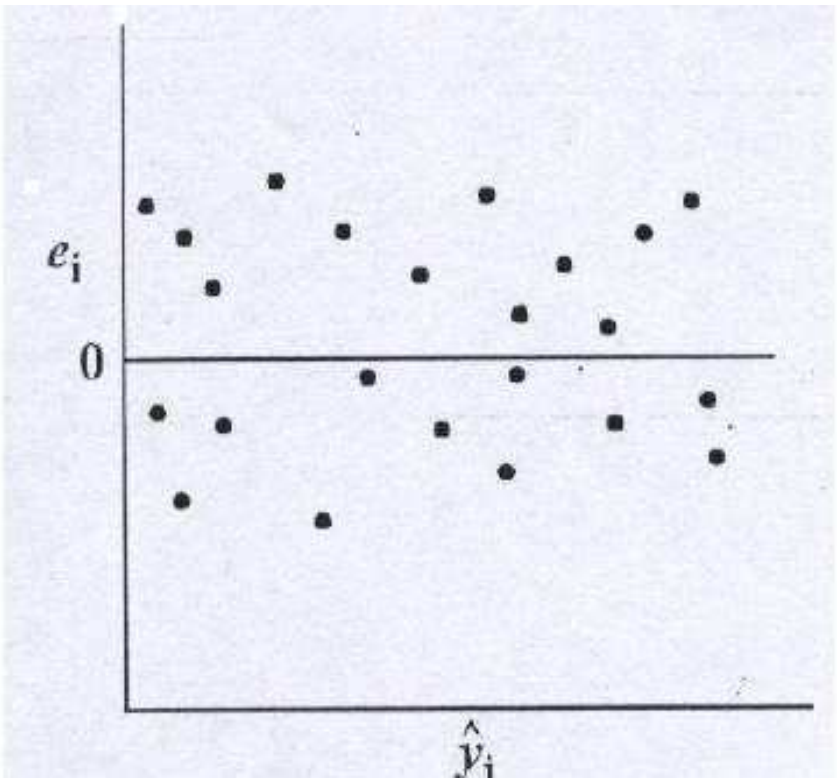
Verificação de Não-Normalidade dos Erros

- Assumimos que os erros $\epsilon_i \sim N(0, \sigma^2)$ para $i = 1, \ldots, n$.
- Desvios da normalidade afetam:
  - os intervalos de confiança;
  - testes t e F.
- Usamos os resíduos como estimativa do erro para verificar a suposição.

- Para testar normalidade podemo usar:
  - Histograma: deve ser simétrico em torno de zero;
  - Gráfico de Probabilidade Normal: verifica visualmente se os dados seguem uma normal;
  - Testes de normalidade (Shapiro-Wilk, Anderson Darling).
    - A hipótese nula é de que os dados são normais e deverá ser rejeitada se o p-valor é pequeno.

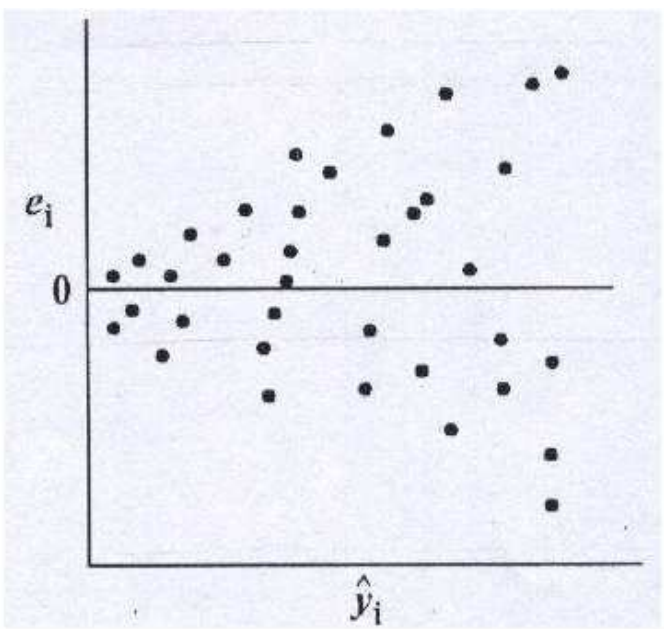
Gráfico dos Resíduos e_i contra Valores Ajustados $\hat{Y}_i$

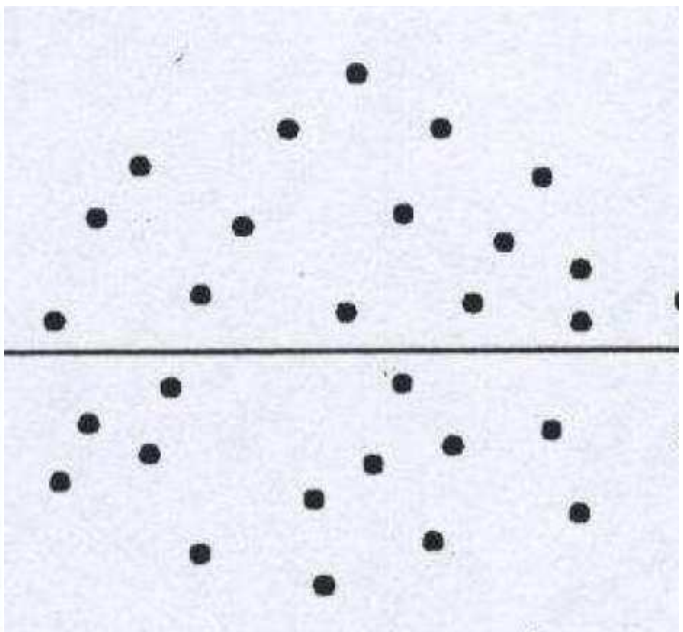
- Aparência desejada:
  - nuvem de pontos aleatória e homogênea em torno do eixo horizontal $Y = 0$.



Útil para detectar as seguintes inadequções do modelo:

- A variância do erro não é constante.
  - Solução: fazer transformação em Y ou usar Mínimos Quadrados Ponderados.





A homocedasticidade é provavelmente violada se...

- Se os resíduos aumentam ou diminuem com os valores ajustados.
- Se os pontos formam uma curva ao redor de zero e não estão dispostos aleatoriamente.
- Poucos pontos no gráfico ficam muito distantes dos demais.

- A equação de regressão não é linear.
  - Solução: transformações em Y e/ou X; inclusão do termo quadrático de X.

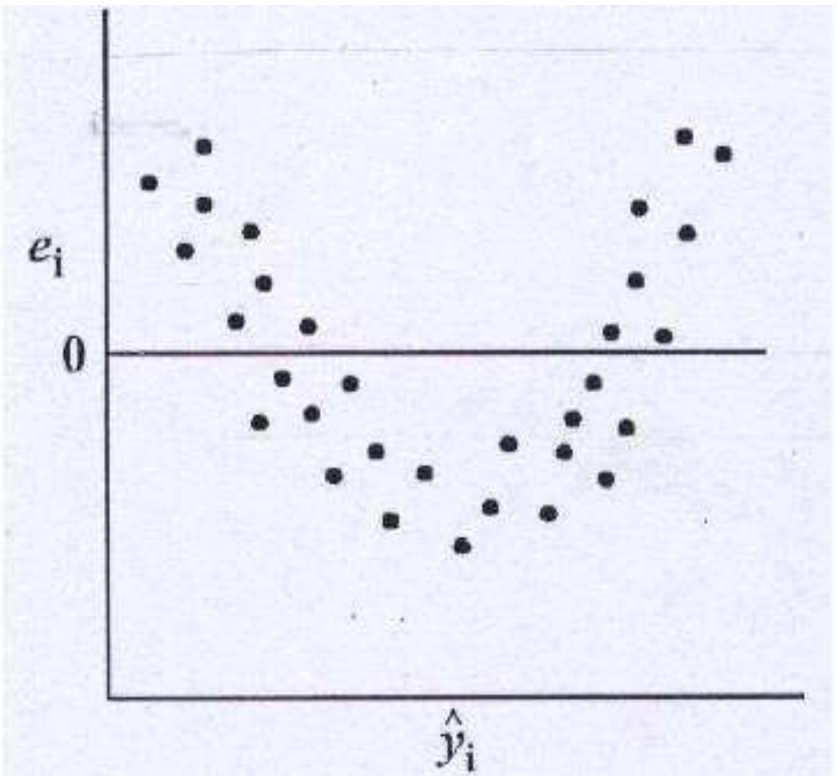


Gráfico dos Resíduos contra a Variável Explicativa

- Na Regressão Linear Simples, tem o mesmo papel do gráfico e_i vs $\hat{Y}_i$.
- Em Regressão Múltipla, pode ser usado para verificar a necessidade de se incluir variáveis.
- Nesse último caso, é feito o gráfico dos resíduos vs variáveis não incluídas no modelo.
- Se houver algum padrão, significa que a variável deve ser incluída.

Gráfico dos Resíduos contra o Tempo ou Ordem de Coleta

- Os erros devem ser independentes entre si.
- Esse gráfico verifica apenas se eles estão correlacionados no tempo.
- Só pode ser usado caso os dados sejam coletados sequencialmente.
- Os erros são plotados na ordem em que foi feita a coleta.
- A presença de algum padrão indica correlação entre eles.
- A existência de correlação temporal pode ser pode ser consequência da:
  - não inclusão que uma variável explicativa relacionada ao tempo.

Resumindo...

- Os erros são assumidos não correlacionados e com variância constante.
- Usamos os resíduos (estimativas do erro) para verificar essas suposições.

Gráfico dos Resíduos	Suposições Avaliadas
e_i vs $\hat{Y}_i$	Variância Constante
e_i vs X_i	Linearidade
e_i vs Variáveis não incluídas	Suficiência das variáveis incluídas.
Probabilidade Normal	Normalidade
e_i vs tempo de coleta	Ausência se autocorrelação temporal.

Tramsformações de Variáveis

- Podemos fazer uma transformação na variável resposta Y e/ou na preditora X para:
  - solucionar problemas de variância não constante;
  - não normalidade dos erros;
  - não linearidade do modelo.
- Violações que geralmente ocorrem ao mesmo tempo:
  - variância constante e distribuição normal.
- Suposições básicas

$$Y_i \text{ tem distribuição normal} \quad Var(Y_i) = \sigma^2.$$

Transformação para estabilizar a variância

- Frequentemente a variância das observações varia com sua média.
- Situações comuns:

$$Y \text{ são contagens} \Rightarrow Var(Y) \propto E(Y)$$

$$Y \text{ são proporções} \Rightarrow Var(Y) \propto E(Y)(1 - E(Y))$$

- Uma solução: fazer transformações em Y.
- Uma outra possibilidade é usar Mínimos Quadrados Generalizados.

- A tabela a seguir mostra:
  - tipos de relação mais comuns entre variância e esperança
  - tipo de transformação geralmente adequada.

Relação entre $Var(Y)$ e $E(Y)$	Transformação
$Var(Y) \propto E(Y)$	$Y^* = \sqrt{Y}$
$Var(Y) \propto E(Y)(1 - E(Y))$	$Y^* = \arcsin Y$
$Var(Y) \propto E(Y)^2$	$Y^* = \ln Y$
$Var(Y) \propto E(Y)^3$	$Y^* = 1/\sqrt{Y}$
$Var(Y) \propto E(Y)^4$	$Y^* = 1/Y$

Métodos analíticos para seleção de transformação

- Em muitos casos é possível escolher a transformação empiricamente.
- O gráficos e o tipo de dado, nos mostram indícios de qual é a transformação mais adequada.
- Pode ser interessante ter um técnica mais objetiva e automática de escolha da transformação.
- Uma técnica usa para isso é:

Métodos analíticos para seleção de transformação

- Em muitos casos é possível escolher a transformação empiricamente.
- O gráficos e o tipo de dado, nos mostram indícios de qual é a transformação mais adequada.
- Pode ser interessante ter um técnica mais objetiva e automática de escolha da transformação.
- Uma técnica usa para isso é:
  - **procedimento de Box-Cox.**

Procedimento de Box-Cox.

- É um método para escolher transformações de maneira automática.
- Essas transformações são escolhidas na família de transformações potência.
- Pode ser aplicada apenas se a variável resposta assumi só valores positivos.
- A transformação da variável Y é dada por

$$g_\lambda(y) = \begin{cases} \frac{(y^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \log(\lambda) & \lambda = 0\,. \end{cases}$$

Transformação para corrigir não-linearidade

- Vimos até agora transformações para estabilizar a variância.
- Porém essas transformações também afetam as relações entre as variáveis.
- Resultando geralemente em relações curvas entre as variáveis.
- Maneira de voltar a termos uma relação linear é:
  - fazer transformações na variável explicativa.
- Variáveis explicativas são aleatórias?

Transformação para corrigir não-linearidade

- Vimos até agora transformações para estabilizar a variância.
- Porém essas transformações também afetam as relações entre as variáveis.
- Resultando geralemente em relações curvas entre as variáveis.
- Maneira de voltar a termos uma relação linear é:
  - fazer transformações na variável explicativa.
- Variáveis explicativas são aleatórias? Não.
- Ao fazermos transformações nelas vamos afetar a variância?

Transformação para corrigir não-linearidade

- Vimos até agora transformações para estabilizar a variância.
- Porém essas transformações também afetam as relações entre as variáveis.
- Resultando geralemente em relações curvas entre as variáveis.
- Maneira de voltar a termos uma relação linear é:
  - fazer transformações na variável explicativa.
- Variáveis explicativas são aleatórias? Não.
- Ao fazermos transformações nelas vamos afetar a variância? Não.

- Transformar a variável resposta afeta a variância e a linearidade.
- Transformar a variável explicativa só afeta linearidade.
- Então devemos primeiro trabalhar com a variância:
  - transformar a variável Y.
- Em seguida corrigir a linearidade:
  - transformar a variável explicativa.

Transformação log-log

- Já vimos que podemos transformar a variável resposta e a explicativa.
- A transformação mais comum é a log-log.
- Tiramos o log da variável resposta e da explicativa.
- Essa transformação é últil quando a verdadeira relação entre X e Y é dada por

$$Y_i = \beta_0 x_i^{\beta_1} \times \epsilon_i \ .$$

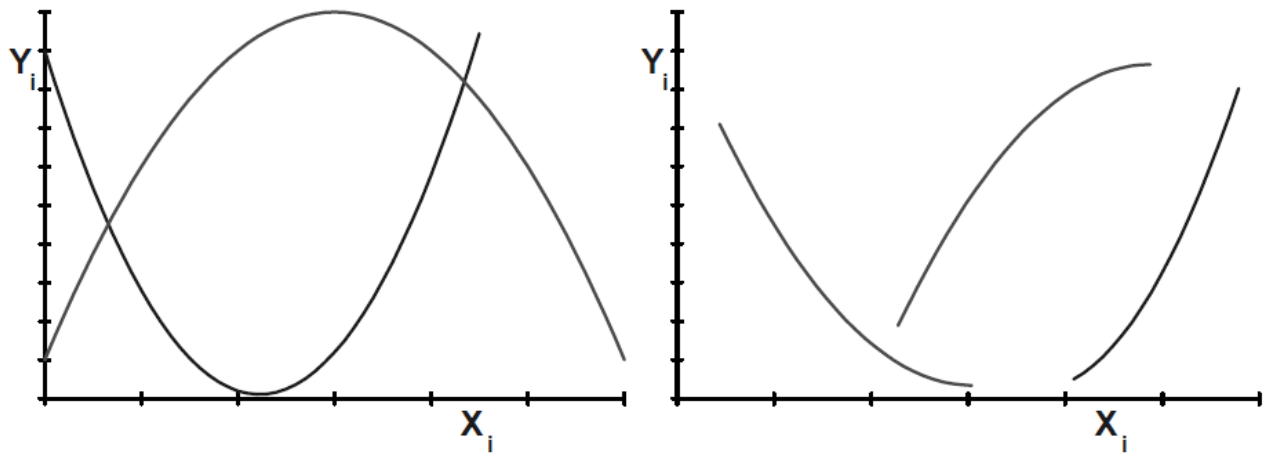
- O erro aleatório é multiplicativo.

Regressão Polinomial

- A relação entre a variável resposta e a explicativa é modelada por um polinômio de ordem p.
- Ajustamos um polinômio de ordem 2, 3 ou 4 e então verificamos se podemos eliminar alguns termos do modelo.
- Com polinômios podemos:
  - determinar se existe uma relação curvilínea entre Y e X;
  - determinar se essa curva é quadrática, cúbica, etc;
  - obter a equação polinomial de Y em função de X (podemos ter várias preditoras).

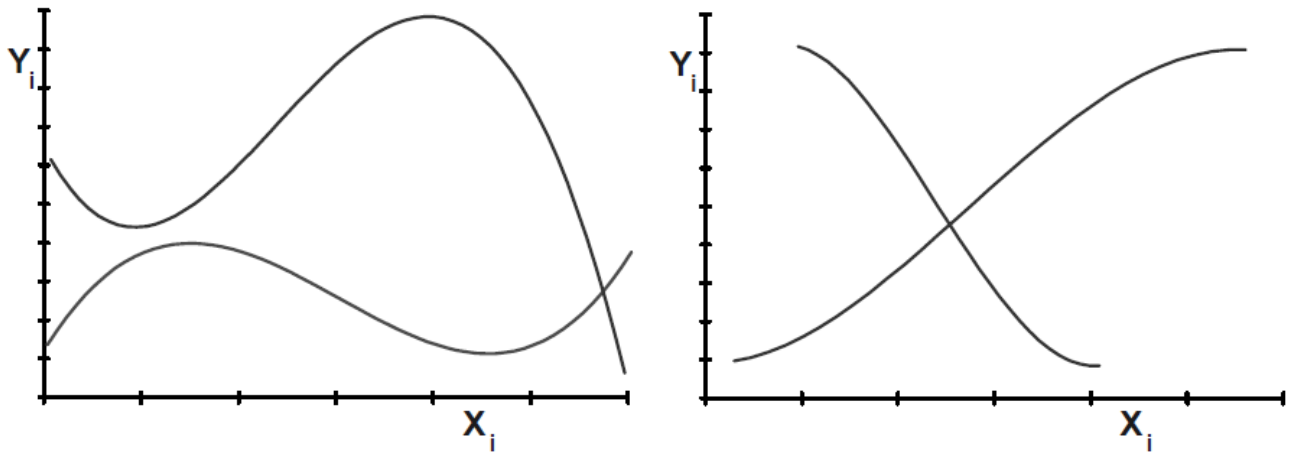
Polinômio mais simples - de segunda ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$



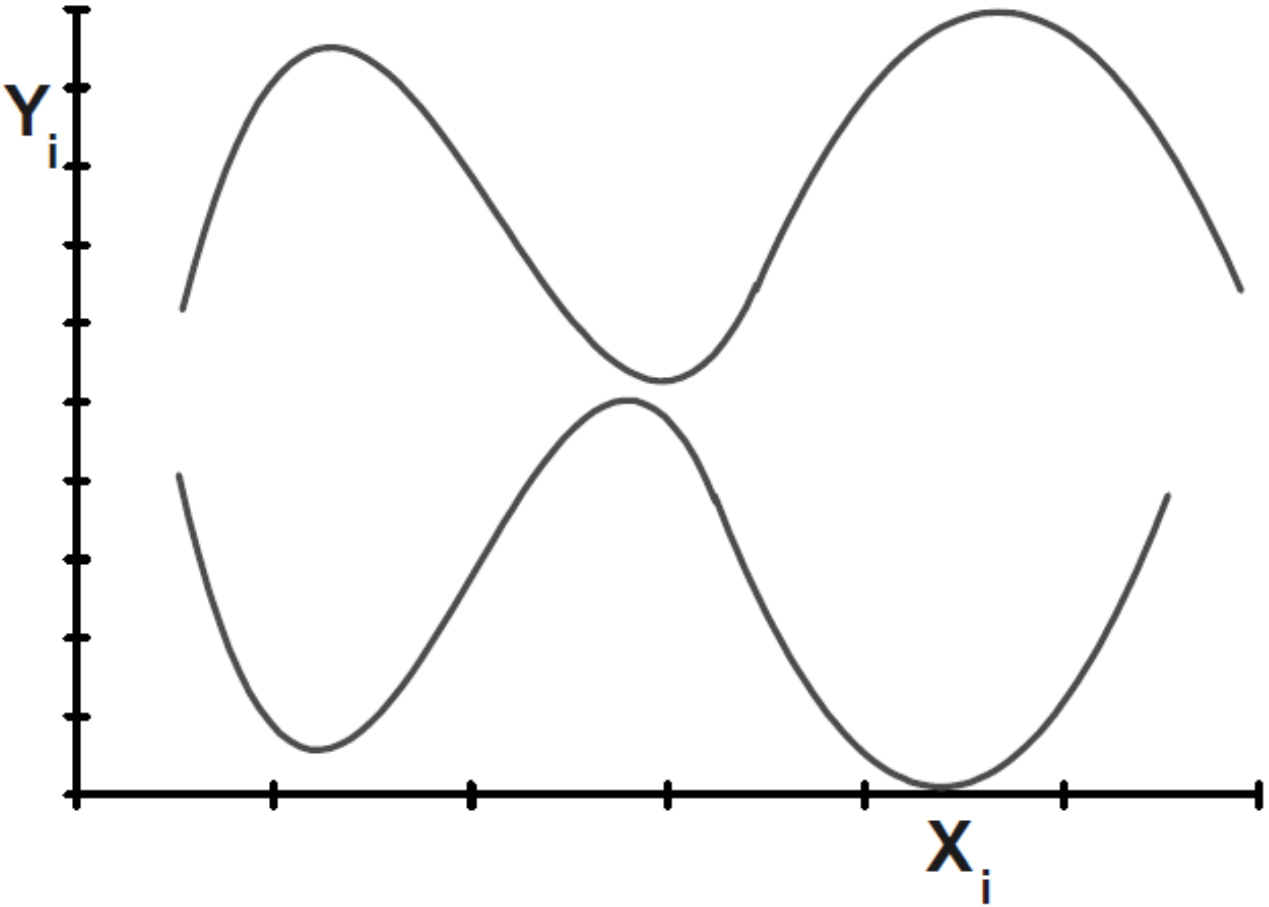
Um polinômio um pouco mais complexo - de terceira ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$$



Um polinômio um pouco mais complexo - de quarta ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \epsilon_i$$



Multicolinearidade

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

a solução de mínimos quadrados é dada por

$$\hat{\boldsymbol{\beta}} =$$

Multicolinearidade

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

a solução de mínimos quadrados é dada por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- Porém, se $(\mathbf{X}'\mathbf{X})$ é singular:
  - o estimador não pode ser obtido dessa forma;
  - as equações normais não têm solução única.
- Isto acontece porque?

Multicolinearidade

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

a solução de mínimos quadrados é dada por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- Porém, se $(\mathbf{X}'\mathbf{X})$ é singular:
  - o estimador não pode ser obtido dessa forma;
  - as equações normais não têm solução única.
- Isto acontece porque?
- A colunas da matriz $\mathbf{X}$ não são linearmente independentes.
- Uma delas é combinação linear das demais.

- Dizemos que existe multicolinearidade entre as colunas de X:
  - as variáveis explicativas são fortemente correlacionadas entre si.
- Se as variáveis são altamente correlacionadas, mas a correlação não é um:
  - uma não será exatamente combinação linear das demais.
- A matriz $(\mathbf{X}'\mathbf{X})$ tem inversa, ou seja $\det((\mathbf{X}'\mathbf{X})) \neq 0$.
- Porém $\det((\mathbf{X}'\mathbf{X})) \approx 0$.
- Dizemos que a matriz é mal condicionada.
- A inversa

$$(\mathbf{X}'\mathbf{X})^{-1}$$

é muito instável.
- Alterações pequenas na matriz modificam muito sua inversa.

- Esse tipo de comportamento não é desejável em um modelo de regressão.
- Veremos como identificar o problema e algumas possíveis soluções.
- Consequências para o ajuste do modelo:
  - estimadores dos coeficientes não são confiáveis;
  - estimadores com alta variância e covariância.

Indicações da presença de multicolinearidade:

Indicações da presença de multicolinearidade:

- Coeficientes de correlação linear entre pares de variáveis explicativas ficam muito próximos de -1 ou 1.
- Gráficos de dispersão entre pares de varráveis explicativas apresentam configurações especiais:
  - indicando algum tipo de relação entre elas.
- Coeficientes de regressão apresentam sinais algébricos opostos ao esperado a partir de conhecimento teórico.
- Coeficientes de regressão sofrem grandes alterações quando:
  - uma coluna ou linha da matrix **X** é extraída;
  - ou seja, quando uma variável explicativa é retirada do modelo ou uma observação é retirada da amostra.

Indicações da presença de multicolinearidade:

Indicações da presença de multicolinearidade:

- O teste F rejeita

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

mas nenhuma das

$$H_0 : \beta_j = 0$$

é rejeitada pelos testes *t* individuais.

- Variáveis explicativas que teoricamente são consideradas importantes:
  - apresentam coeficientes de regressão com estatística *t* muito baixa.
- Os erros padrão dos coeficientes de regressão são muito altos.

- Estes métodos de diagnósticos são informais.
- Possuem limitações importantes:
  - não fornecem uma medida do impacto da multicolinearidade;
  - não identificam a natureza da multicolinearidade.
- Diagramas de dispersão e coeficientes de correlação revelam relações entre pares de variáveis.
- Não mostram a relação entre grupos de variáveis.
- Como por exemplo, a relação entre X_1 e uma combinação linear de X_2, X_3 e X_4.

- Método formal de detectar e medir multicolinearidade:
  - análise dos fatores de inflação da variância (VIF).
- Para o modelo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

ajustado por mínimos quadrados. A matriz de Covariâncias dos estimadores dos coeficientes é dada por

$$Var(\hat{\beta}) =$$

- Método formal de detectar e medir multicolinearidade:
  - análise dos fatores de inflação da variância (VIF).
- Para o modelo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

ajustado por mínimos quadrados. A matriz de Covariâncias dos estimadores dos coeficientes é dada por

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

os elementos da diagonal são iguais a

$$Var(\hat{\beta}_k) = VIF_k \left(\frac{\sigma^2}{S_k}\right) \quad \text{com} \quad S_k = \sum_{i=1}^{n}(X_{ki} - \overline{X}_k)^2$$

- Se X_k não está relacionada às outras variáveis X, tem-se $VIF_k = 1$.

- Podemos tomar algumas medidas para corrigir o problema de multicolineridade.
- Coletar mais dados planejados para “quebrar” a relação entre as variáveis nos dados existentes.
- Considere que X_1 e X_2 são positivamente correlacionadas.
- Devemos coletar novos pares de observações (X_1, X_2) tais que:
  - sejam coletados valores baixos de X_1 e altos de X_2, e vice-versa.
- Isso pode ser uma característica intrínseca das variáveis.
- Por exemplo, as variáveis sócio-econômicas renda da família e tamanho da casa.

- Podemos ainda redefinir as variáveis preditoras.
- Por exemplo, se X_1, X_2 e X_3 são aproximadamente dependentes.
- É possível encontrar alguma função delas,

$$X = \frac{(X_1 + X_2)}{X_3} \quad X = X_1 X_2 X_3$$

que preserve a relação das preditoras, mas elimine a multicolineridade.

- Uma outra solução é eliminar variáveis explicativas.
- Essa não é uma solução satisfatória,
  - pois as preditoras podem ter grande poder de explicação da resposta;
  - podemos estar descartando informação importante.

Observações Atípicas

- **Outliers:** não se ajustam bem ao modelo (resíduo grande);
- **Alavancas:** têm valor não usual da variável explicativa;
- **Influentes:** quando presentes, mudam o ajuste do modelo substancialmente.

- Nem todo outlier é influente.
- Assim como nem toda alavanca é influente.
- Mas um ponto influente é um outlier e/ou uma alavanca.
- Os resíduos exercem importante papel na análise de observações não usuais.

A matriz H

- Lembre que no modelo de regressão

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

o estimador de mínimos quadrados é dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- Os valores estimados para $\mathbf{Y}$ são dados por

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{H}\,\mathbf{Y}\,.$$

- A matriz

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

é chamada de matriz **hat** (chapéu), pois liga $\mathbf{Y}$ a $\hat{\mathbf{Y}}$

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}\,.$$

Outlier

Outlier

- Observação com valor discrepante de Y.
- É um ponto que não se ajusta bem ao modelo.
- O valor ajustado $\hat{Y}_i$ é muito distante de Y_i.
- O resíduo e_i é grande.
- Podem afetar o ajuste do modelo.

- Nem todo ponto que possui um valor alto de e_i é um *outlier*.
- Como identificar pontos que são de fato *outliers*?

- Nem todo ponto que possui um valor alto de e_i é um *outlier*.
- Como identificar pontos que são de fato *outliers*?
- Valores extremos de t_i na distribuição t_{n-p} indicam que a i-ésima observação é um *outlier*.
- Um *outlier* só deve ser excluído se um erro é identificado.
- Caso isso ocorra deve-se destacar essa exclusão nas conclusões finais do trabalho.

There should be strong non-statistical evidence before we discard any of the outliers. Otherwise, it can be dangerous as "deleting a desirable response" may give the user a false sense of precision in estimation or prediction.

Observações alavanca

Observações alavanca

- Observações extremas em X.
- Observações tais que o valor de X está muito afastado dos demais.
- Elas não necessariamente afetam a reta de regressão.
- Como detectamo esses pontos?

Observações alavanca

- Observações extremas em X.
- Observações tais que o valor de X está muito afastado dos demais.
- Elas não necessariamente afetam a reta de regressão.
- Como detectamo esses pontos?

- A matriz **H** depende apenas dos valores **X**.
- Então h_{ii} é alto por causa de um valor atípico de X e não de Y.
- Já vimos que

$$\sum_{i=1}^{n} h_{ii} = p$$

onde p é a dimensão do vetor de coeficientes β.

- Se nenhum ponto "força" o ajuste, devemos ter
  - todos h_{ii} aproximadamente iguais, nenhum "puxa" mais;

$$h_{ii} \approx \frac{p}{n} .$$

- Uma regra empírica é olhar com mais atenção pontos tais que

$$h_{ii} > \frac{2p}{n} .$$

Ponto Influente

Ponto Influente

- É aquele cuja remoção do banco de dados causa grande mudança no ajuste.
- Pode ou não ser um *outlier*.
- Pode ou não ser uma *alavanca*.
- Tende a ter pelo menos um dessas duas propriedades.
- A medida mais usada para detectar pontos influentes é

Ponto Influente

- É aquele cuja remoção do banco de dados causa grande mudança no ajuste.
- Pode ou não ser um *outlier*.
- Pode ou não ser uma *alavanca*.
- Tende a ter pelo menos um dessas duas propriedades.
- A medida mais usada para detectar pontos influentes é a **Distância de Cook**.

- Vejamos como é definida a **Distância de Cook**.
- Sejam $\hat{\beta}$ e $\hat{\mathbf{Y}}$ os estimadores de β e $\mathbf{Y}$.
- Esses estimadores são calculados usando as n observações.
- Retiramos a i-ésima observação do banco de dados.
- Recalculamos os estimadores.
- Vamos denotá-los por $\hat{\beta}_{i}$ e $\hat{\mathbf{Y}}_{i}$.

- A Ditância de Cook mensura:
  - o quanto muda as estimativas dos coeficientes de regressão com a retirada da observação *i*

$$\hat{\beta}_{(i)} - \beta_{(i)} \, .$$

- A distância Cook da i-ésima observação é dada por:

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\hat{\sigma}^2} = \frac{r_i h_{ii}}{p(1 - h_{ii})}$$

onde

  - $\hat{\sigma}^2$ é a estimativa de σ^2 (baseada no modelo com as n observações);
  - p é a dimensão dao vetor β;
  - r_i é o resíduo padronizado;
  - h_{ii} é a alavanca.

- Um valor alto de D_i (em relação às demais)
  - indica que a observação *i* é influente, os valores estimados com e sem a observação são bem próximos.
- A influência pode ser consequência de:
  - um valor extremo do resíduo padronizado (o ponto é *outlier*);
  - um valor alto (próximo de 1) de h_{ii} (o ponto é uma alavanca).

Métodos de Seleção de Variáveis

- Devemos incluir no modelo todas as variáveis explicativas disponíveis?
- Ou devemos escolher apenas um subconjunto delas?
- Podemos usar procedimentos automáticos para auxiliar na escolha desse subconjunto.
- Alguns métodos que podem ser usados:

Métodos de Seleção de Variáveis

- Devemos incluir no modelo todas as variáveis explicativas disponíveis?
- Ou devemos escolher apenas um subconjunto delas?
- Podemos usar procedimentos automáticos para auxiliar na escolha desse subconjunto.
- Alguns métodos que podem ser usados:
  - Todas as Regressões Possíveis (*All Regressions*)
  - Inclusão Passo a Frente (*Forward*)
  - Eliminação Passo Atrás (*Backward*)
  - Seleção Passo-a-Passo (*Stepwise*)

Critérios de Seleção da melhor Regressão

- Modelo “cheio”: modelo com todas variáveis explicativas disponíveis.
- Precisamos escolher uma variável a ser descartada.
- Podemos escolher com base em alguma das medidas:
  - menor valor de S^2 (estimativa de σ^2);
  - maior valor de $R^2_{ajustado}$

$$R^2_{ajustado} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right);$$

  - Estatística C_p de Mallows

$$C_p = \frac{SQE}{S^2} - (n - 2p)$$

quanto mais próxima essa estatística estiver de p, melhor é o modelo.

Todas as Regressões Possíveis (*All Regressions*)

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

  $2^k - 1$ pois não contamos o modelo se nenhuma variável.

- Por exemplo, se $k = 10$, temos $2^{10} - 1 = 1023$ possibilidades.
- Escolhemos algum dos modelos com base nos critérios.
- Se o número de variáveis é grande, o método é computacionalmente inviável.

Inclusão Passo a Frente (Forward)

Inclusão Passo a Frente (Forward)

- Começamos com o modelo só com intercepto.
- Vamos incluindo as variáveis mais significativas.
- O algoritmo para quando uma variável não significativa é encontrada.
- Uma vez que a variável entra no modelo, ela não sai mais.
- Mesmo que sua contribuição deixe de ser significante quando uma outra variável entrar.

Eliminação Passo Atrás (*Backward*)

Eliminação Passo Atrás (*Backward*)

- Começamos com o modelo com todas variáveis.
- Vamos excluindo as variáveis menos significativas.
- Quando encontramos uma variável significativa o algoritmo para.
- Uma vez que a variável sai no modelo, ela não entra mais.
- Mesmo que sua contribuição passe a ser significante quando uma outra variável entrar.

Seleção Passo a Passo (*Stepwise*)

Seleção Passo a Passo (*Stepwise*)

- É uma mistura de *Forward* e *Backward*.
- A variável que entra em um passo pode sair nos próximos.
- A variável que sai em um passo pode entrar nos próximos.

Etapas para Análise de Regressão

- Inicie o procedimento realizando uma análise exploratória dos dados;
- Ajuste o modelo e realize um teste sobre a validade do mesmo;
- Caso necessário, faça uma transformação na variável resposta (y) para estabilizar a variância;
- Faça uma análise dos resíduos para justificar a aleatoriedade e normalidade dos mesmos e para detectar possíveis outliers;
- Caso seja identificado algum outlier, procure por evidências que justifiquem a ocorrência do mesmo antes que seja retirado das observações.