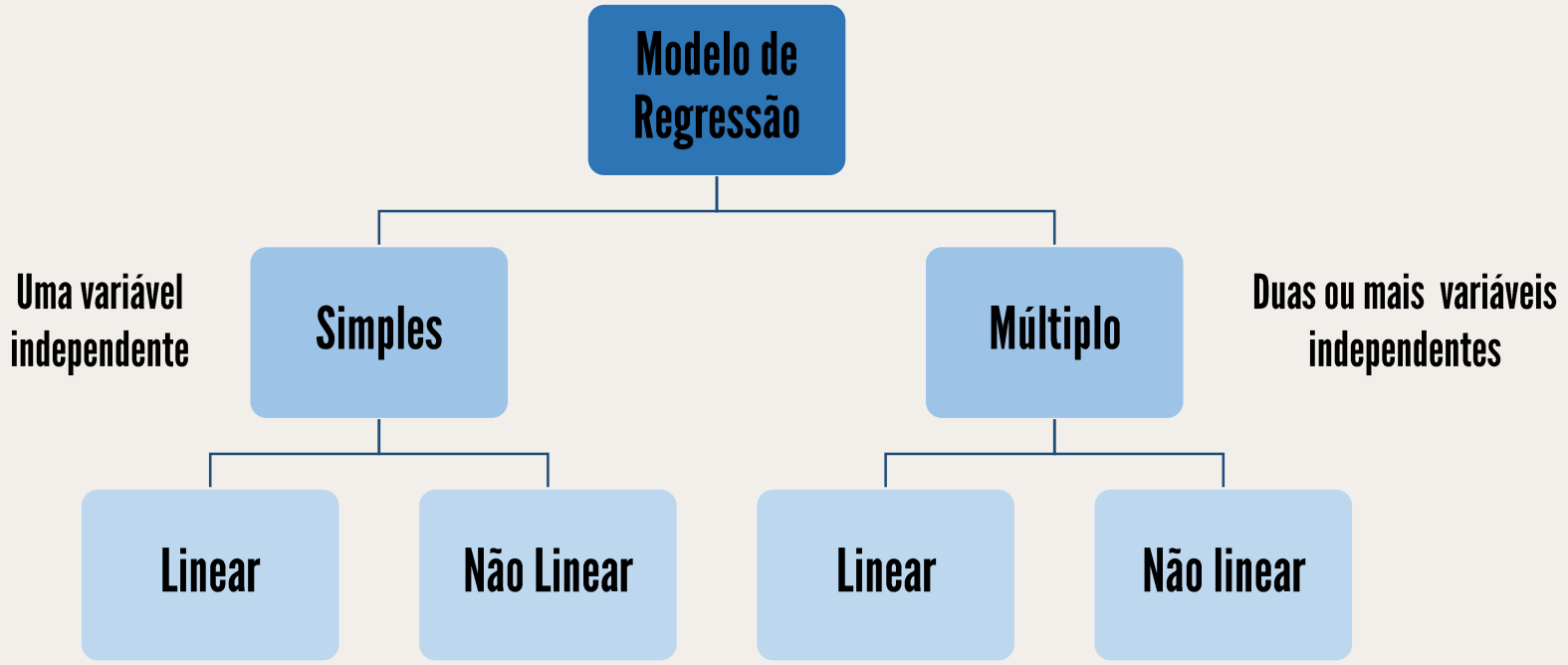
REGRESSÃO LINEAR MÚLTIPLA

PROF. LETÍCIA RAPOSO
profleticiaraposo@gmail.com

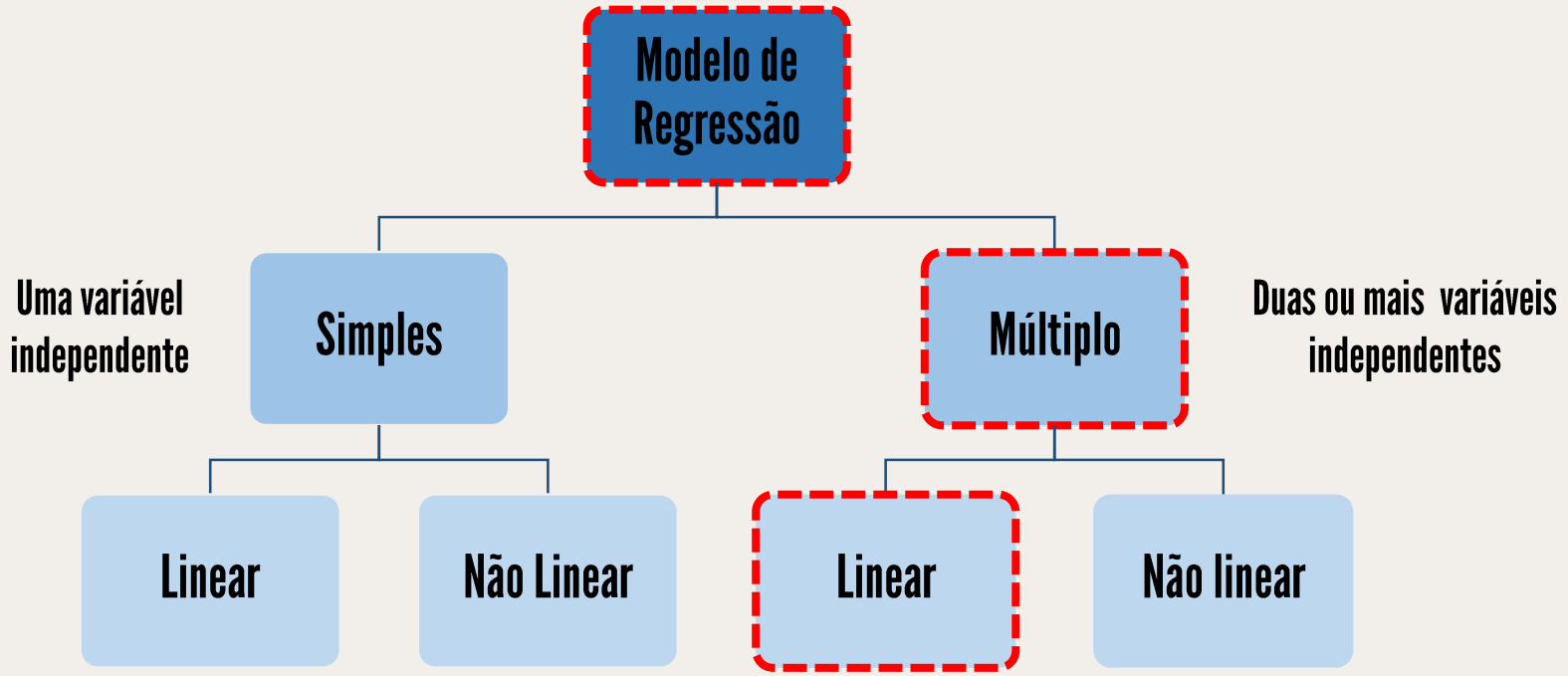
TÓPICOS DA AULA

- Modelo de regressão linear múltipla;
- Pressupostos do modelo;
- Significado dos parâmetros;
- Estimação dos parâmetros;
- Testes de significância;
- Avaliação da regressão.

TIPOS DE MODELOS DE REGRESSÃO



TIPOS DE MODELOS DE REGRESSÃO



REGRESSÃO LINEAR MÚLTIPLA

- Extensão da regressão simples.
- Mais de uma variável preditora (independente).
- Lidar com mais de uma variável é mais difícil, pois:
  - É mais *difícil escolher o melhor modelo*, uma vez que diversas variáveis candidatas podem existir;
  - É mais *difícil visualizar a aparência do modelo ajustado*, mais difícil a representação gráfica (>3D);
  - Às vezes, é *difícil interpretar o modelo ajustado*;
  - *Cálculos difíceis* de serem executados sem auxílio de computador.

REGRESSÃO LINEAR MÚLTIPLA

- Suponha que o peso de crianças dependa da altura e da idade. Um modelo de regressão múltipla que pode descrever esta relação é

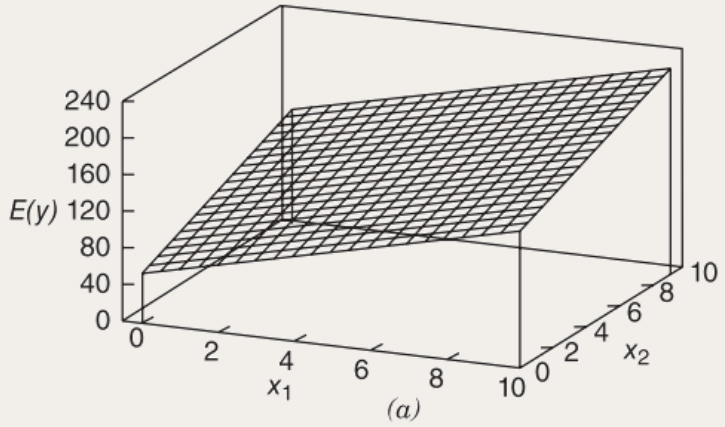
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (1)$$

em que y denota o peso, x_1 denota a altura, e x_2 denota a idade.

- Este é um modelo de regressão linear múltipla com duas variáveis preditoras.
- O termo linear é usado porque a equação (1) é uma função linear dos parâmetros desconhecidos $\beta_0, \beta_1, \beta_2$.

REGRESSÃO LINEAR MÚLTIPLA

O modelo de regressão (1) descreve um plano no espaço tridimensional de y, x_1 e x_2.



β_0 é o intercepto do plano de regressão. Se $x_1 = x_2 = 0$, β_0 é a média de y.

$$E(y) = \overbrace{50}^{\beta_0} + \underbrace{10x_1}_{\beta_1} + \underbrace{7x_2}_{\beta_2}$$ (Assumindo que $E(\varepsilon) = 0$)

β_1: indica a mudança esperada na resposta (y) para cada mudança de unidade em x_1 quando x_2 é mantido constante. β_2: indica a mudança esperada em y para cada mudança de unidade em x_2 quando x_1 é mantido constante.

REGRESSÃO LINEAR MÚLTIPLA

- Em geral, a resposta y pode estar relacionada a k variáveis preditoras.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon \qquad (2)$$

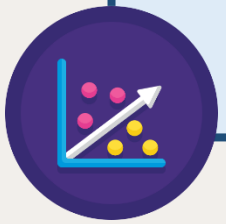
- O modelo é chamado de modelo de regressão linear múltipla com k preditores.
- Os parâmetros $\beta_j, j = 0,1, \ldots, k$ são chamados de coeficientes de regressão.
- Este modelo descreve um hiperplano no espaço (k+1)-dimensional das variáveis preditoras x_j.
- O parâmetro β_j representa a mudança esperada na resposta y por unidade de mudança em x_j quando todas as variáveis restantes do modelo x_i $(i \neq j)$ são mantidas constantes.

REGRESSÃO LINEAR MÚLTIPLA

- Na maioria dos problemas do mundo real, os valores dos parâmetros (os coeficientes de regressão) e a variância do erro σ^2 não são conhecidos e devem ser estimados a partir de dados amostrais.

Modelos de regressão linear múltipla são frequentemente usados para:

- Obter uma equação para predizer valores de y a partir dos valores de vários preditores.
- Estudar o efeito de uma variável x, ajustando ou levando em conta outras variáveis preditoras.
- Explorar as relações entre múltiplas variáveis para determinar quais influenciam y.



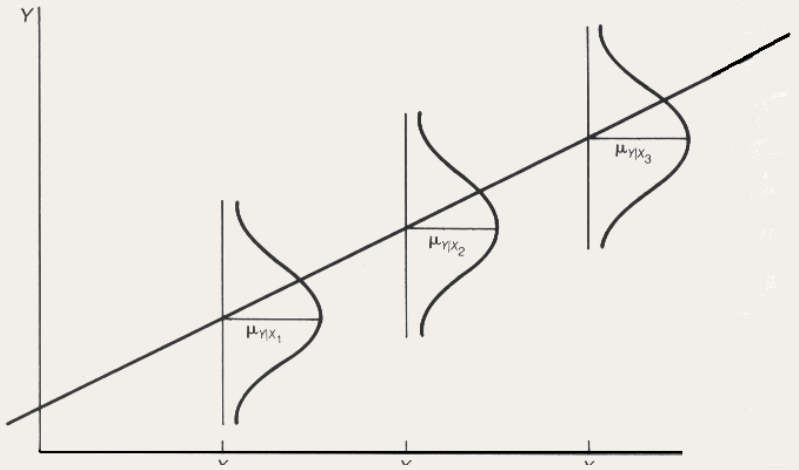
PRESSUPOSTOS

Os pressupostos da regressão linear simples podem ser estendidos para a regressão linear múltipla.

- *Existência:* para uma combinação específica das variáveis independentes $x_1, x_2, \ldots, x_k$, y é uma variável aleatória com uma certa distribuição de probabilidade, com média e variância finitas.
- *Independência*: as observações de y são estatisticamente independentes umas das outras. Este pressuposto é violado quando mais de uma observação é feita de um mesmo indivíduo.
- *Linearidade*: o valor médio de y para cada combinação específica de $x_1, x_2, \ldots, x_k$ é uma função linear de $x_1, x_2, \ldots, x_k$.

PRESSUPOSTOS

- *Homocedasticidade*: a variância de y é a mesma para qualquer combinação fixa de $x_1, x_2, \ldots, x_k$.
- *Amostra aleatória ou representativa* da população.
- *Normalidade*: para uma combinação fixa de $x_1, x_2, \ldots, x_k$, a variável y tem distribuição normal.



PRESSUPOSTOS

Normalidade de y:

- No caso de não normalidade, transformações matemáticas de y podem gerar conjunto de dados com distribuição aproximadamente normal ($\log y$, $\sqrt{y}$);
- No caso de variável y categórica nominal ou ordinal, métodos de regressão alternativos são necessários (logística - dados binários, Poisson - dados discretos).

ESTIMAÇÃO DOS PARÂMETROS DOS MODELOS

Método dos Mínimos Quadrados

Método da Máxima Verossimilhança

ALGUMAS PERGUNTAS IMPORTANTES

- Uma vez que tenhamos estimado os parâmetros no modelo, nos deparamos com algumas questões:

1. *Pelo menos um dos preditores é útil na predição da resposta?*
2. *Todos os preditores ajudam a explicar* $\boldsymbol{y}$ *ou apenas um subconjunto dos preditores é útil?*
3. *Quão bem o modelo se ajusta aos dados?*
4. *Dado um conjunto de valores de predição, que valor de resposta devemos predizer e quão precisa é a nossa predição?*

- Os testes formais exigem que nossos erros aleatórios sejam independentes e sigam uma distribuição normal com média $E(\varepsilon_i) = 0$ e variância $Var(\varepsilon_i) = \sigma^2$.

TESTE PARA SIGNIFICÂNCIA DA REGRESSÃO

- Teste para determinar se existe uma relação linear entre a resposta y e qualquer uma das variáveis preditoras $x_1, x_2, \ldots, x_k$.
- As hipóteses apropriadas são

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

$$H_1: \beta_j \neq 0 \text{ para, pelo menos, um j}$$

- Rejeitar H0 indica que pelo menos uma das variáveis preditoras contribui significativamente para o modelo.
- O procedimento deste teste é uma generalização da análise de variância usada na regressão linear simples.

TESTE PARA SIGNIFICÂNCIA DA REGRESSÃO

- A soma dos quadrados total SS_T é dividida na soma dos quadrados devida à regressão SS_R e na soma dos quadrados dos resíduos SS_{Res}.

$$SS_T = SS_R + SS_{Res}$$

- E a razão F é dada por

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

e segue uma distribuição $F_{k,n-k-1}$, em que k é o número de variáveis preditoras e n o número de observações.

- Se $F_0 > F_{\alpha,k,n-k-1}$, rejeita-se H0.

TESTE PARA SIGNIFICÂNCIA DA REGRESSÃO

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Regressão	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	n-k-1	MS_{Res}	
Total	SS_T	n-1		

TESTE PARA SIGNIFICÂNCIA DA REGRESSÃO

- Às vezes, queremos testar se um subconjunto específico de q dos coeficientes é zero.
- Isso corresponde a uma hipótese nula

$$H_0: \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$$

onde, por conveniência, colocamos as variáveis escolhidas para omissão no final da lista.

- Neste caso, ajustamos um segundo modelo que usa todas as variáveis, exceto as últimas q variáveis.
- A razão F apropriada é

$$F = \frac{SS_{Res}(RM) - SS_{Res}(FM)/q}{SS_{Res}(FM)/(n-k-1)}$$

RM: modelo reduzido
FM: modelo completo

TESTES PARA COEFICIENTES DA REGRESSÃO

- Se determinamos que pelo menos uma das variáveis preditoras é importante, precisamos descobrir qual (is) é (são).
- Adicionar uma variável a um modelo de regressão sempre faz com que a SS_R aumente e a SS_{Res} diminua. Devemos decidir se o aumento na SS_R é suficiente para garantir o uso da variável adicional no modelo.
- A adição de uma variável também aumenta a variância de $\hat{y}$, portanto devemos ter o cuidado de incluir apenas as variáveis que tenham valor na explicação da resposta.
- Além disso, adicionar uma variável sem importância pode aumentar o quadrado médio residual, o que pode diminuir a utilidade do modelo.

TESTES PARA COEFICIENTES DA REGRESSÃO INDIVIDUAIS

- As hipóteses para testar a significância de qualquer coeficiente de regressão individual, como β_j, são

$$H_0: \beta_j = 0, H_1: \beta_j \neq 0$$

- Se H0 não é rejeitada, então há uma indicação de que a variável x_j pode ser deletada do modelo.
- A estatística do teste é

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- H0 é rejeitada se $|t_0| > t_{\frac{\alpha}{2},n-k-1}$
- Intervalo de confiança 100(1-α) (IC):

$$\hat{\beta}_j - t_{\frac{\alpha}{2},n-k-1} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2},n-k-1}$$

SELEÇÃO DE VARIÁVEIS

- É possível que todos os preditores estejam associados à resposta, mas é mais comum que a resposta esteja relacionada apenas a um subconjunto dos preditores.
- Idealmente, gostaríamos de realizar a seleção de variáveis experimentando um monte de modelos diferentes, cada um contendo um subconjunto diferente dos preditores.
  - Total de 2^k modelos que contêm subconjuntos de k variáveis.
  - k = 2, $2^2 = 4$ modelos; k = 30, $2^{30} = 1.073.741.824$ modelos!
  - Precisamos de uma abordagem automatizada e eficiente para escolher um conjunto menor de modelos a serem considerados.

Como determinamos qual modelo é o melhor?

FORWARD SELECTION

1. Modelo nulo;
2. Ajustamos k regressões lineares simples e adicionamos ao modelo nulo a variável que resulta no SS_{Res} mais baixo;
3. Adicionamos a esse modelo a variável que resulta no SS_{Res} mais baixo para o novo modelo de duas variáveis;
4. Essa abordagem é continuada até que alguma regra de parada seja satisfeita.

BACKWARD SELECTION

1. Modelo com todas as variáveis;
2. Removemos a variável com o maior valor-p - ou seja, a variável que é menos significativa estatisticamente;
3. O novo modelo é ajustado e a variável com o maior valor-p é removida.
4. Esse procedimento continua até que uma regra de parada seja atingida. Por exemplo, podemos parar quando todas as variáveis restantes tiverem um valor p abaixo de um limite.

SELEÇÃO DE VARIÁVEIS

Mixed selection: combinação das duas outras seleções.

1. Começamos sem variáveis no modelo e adicionamos a variável que fornece o melhor ajuste.
2. Continuamos adicionando as variáveis uma a uma. Naturalmente, os valores-p para as variáveis podem se tornar maiores à medida que novos preditores são adicionados ao modelo. Portanto, se em algum ponto o valor-p de uma das variáveis no modelo ultrapassar um certo limite, então removemos essa variável do modelo.
3. Continuamos a executar essas etapas de "avanço" e "retrocesso" até que todas as variáveis no modelo tenham um valor-p suficientemente baixo, e todas as variáveis fora do modelo teriam um valor-p grande se fossem adicionadas ao modelo.

SELEÇÃO DE VARIÁVEIS

Várias estatísticas podem ser usadas para julgar a qualidade de um modelo. Estes incluem:

- Critério de informação de Akaike (AIC)

$$\text{AIC} = {}^{1}/_{n\hat{\sigma}^2}\,(SS_{Res} + 2k\hat{\sigma}^2), k = nº\ preditores$$

- Critério de informação Bayesiano (BIC)

$$\text{BIC} = {}^{1}/_{n}\,(SS_{Res} + \log(n)\,k\hat{\sigma}^2), k = nº\ preditores$$

- R^2 ajustado.

R^2 E R^2 AJUSTADO

- Duas outras formas de avaliar a adequação geral do modelo são o R^2 e o R^2 ajustado (R^2_{Adj}).
- Em geral, o R^2 nunca diminui quando uma variável preditora é adicionada ao modelo, independentemente do valor da contribuição dessa variável. Portanto, é difícil julgar se um aumento de R^2 realmente está nos dizendo algo importante.
- Por isso, muitas vezes é utilizado o R^2_{Adj}, que considera o número de variáveis preditoras.

Aumentará apenas se a adição da variável reduzir o quadrado médio residual

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-k)}{SS_T/(n-1)}$$

Quadrado médio residual

Constante independente do número de variáveis no modelo

O R^2_{Adj} penaliza a adição de termos que não são úteis, por isso é muito útil na avaliação e comparação de modelos de regressão.

REGRESSÃO MÚLTIPLA COM FATORES

- Os métodos de regressão vistos até aqui assumiram valores numéricos.
- Algumas variáveis preditoras podem ser qualitativas/categóricas, também chamadas de fatores. P. ex., sexo.
- As variáveis *dummy* são incorporadas nos modelos de regressão com o objetivo de representar o efeito desses fatores.
  - Se somente dois níveis são usados: $x_k = 0$ para o primeiro valor e $x_k = 1$ para o segundo;
  - Se temos mais de dois níveis: precisa-se de $k - 1$ variáveis preditoras para k níveis.

REGRESSÃO MÚLTIPLA COM FATORES

- Exemplo: variável etnia – 3 categorias: asiático, caucasiano e afro-americano.
- A primeira variável *dummy* poderia ser

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

- A segunda

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

Diferença no valor médio de y entre caucasiano e afro-americano

Diferença no valor médio de y entre as categorias asiática e afro-americana

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{se i}-\text{ésima pessoa é asiática} \\ \beta_0 + \beta_2 + \epsilon_i, \text{se i}-\text{ésima pessoa é caucasiana} \\ \beta_0 + \epsilon_i, \text{se i}-\text{ésima pessoa é afro}-\text{americana} \end{cases}$$

Valor médio de y para americanos

POSSÍVEIS PROBLEMAS

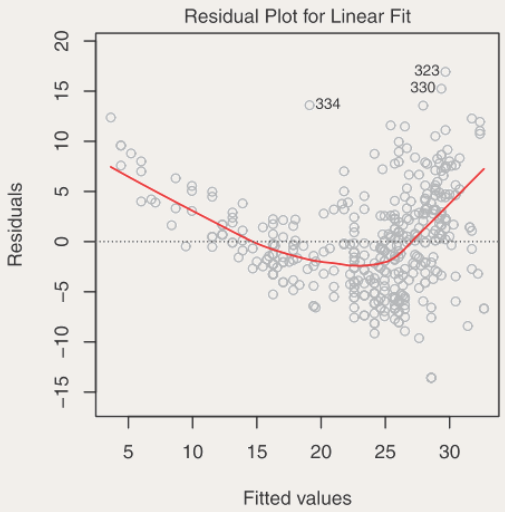
Quando ajustamos um modelo de regressão linear a um conjunto de dados específico, muitos problemas podem ocorrer. Os mais comuns são :

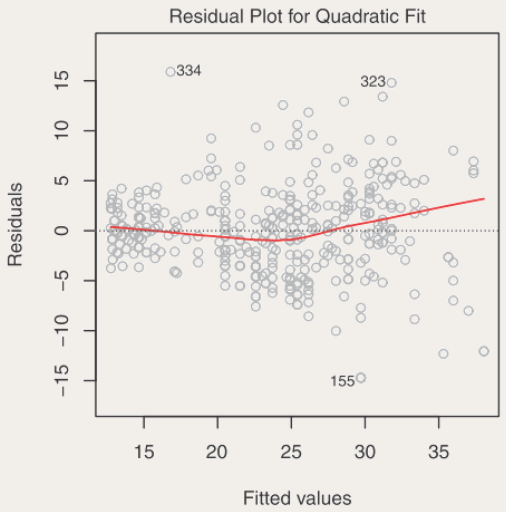
1. Não-linearidade das relações de previsão-resposta.
2. Correlação de termos de erro.
3. Variância não constante dos termos de erro.
4. *Outliers*.
5. Pontos de alta alavancagem.
6. Colinearidade.

NÃO LINEARIDADE DOS DADOS

- O modelo de regressão linear assume que existe uma *relação linear* entre os preditores e a resposta.
- Gráficos residuais são uma ferramenta gráfica útil para identificar a não linearidade.
- No caso de um modelo de regressão múltipla, uma vez que existem múltiplos preditores, nós plotamos os *resíduos versus os valores preditos (ou ajustados)*.
- Idealmente, o gráfico residual não mostrará nenhum padrão discernível. A presença de um padrão pode indicar um problema com algum aspecto do modelo linear.

NÃO LINEARIDADE DOS DADOS





Estratégia: usar transformações não-lineares nos preditores: $\log x\,, \sqrt{x}, x^2$

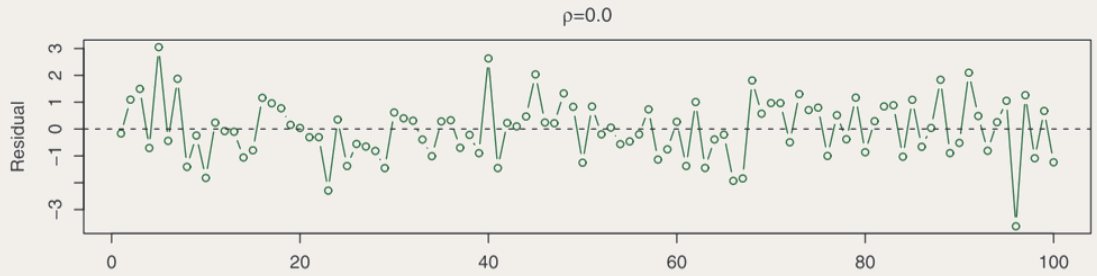
- A linha vermelha é um ajuste suave aos resíduos, exibido para facilitar a identificação de qualquer tendência.
- À esquerda, os resíduos exibem uma forma em U, que fornece uma forte indicação de não linearidade nos dados. Em contraste, à direita, a figura exibe o gráfico residual de um modelo com um termo quadrático.
- Parece haver pouco padrão nos resíduos, sugerindo que o termo quadrático melhora o ajuste aos dados.

CORRELAÇÃO DOS TERMOS DE ERRO

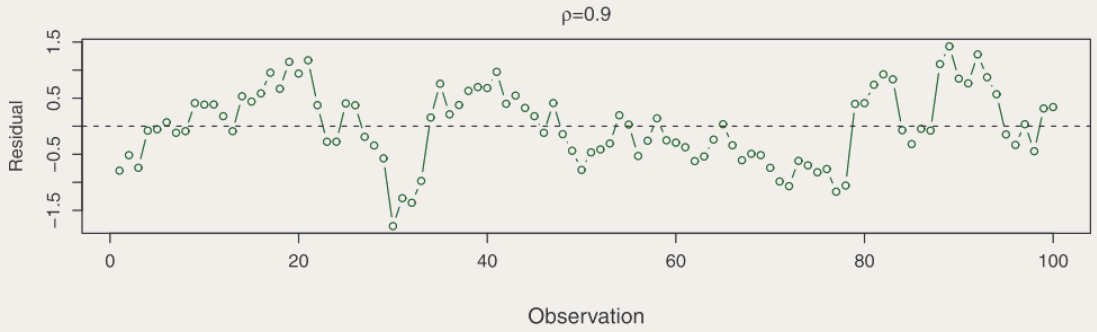
- Se houver correlação entre os termos de erro, os erros-padrão calculados para os coeficientes de regressão estimados tenderão a subestimar os erros-padrão verdadeiros, gerando
  - Intervalos de confiança e predição mais estreitos;
  - Menores valores-p associados ao modelo (pode nos levar a concluir erroneamente que um parâmetro é estatisticamente significativo).
- Se os termos de erro estiverem correlacionados, podemos ter um senso de confiança indevido em nosso modelo.

CORRELAÇÃO DOS TERMOS DE ERRO

- Comum em dados de séries temporais, que consistem em observações para as quais as medidas são obtidas em pontos discretos no tempo.
- Em muitos casos, observações obtidas em pontos de tempo adjacentes terão erros correlacionados positivamente.
- Para determinar se os erros estão correlacionados, podemos traçar os *resíduos do nosso modelo em função do tempo*.
  - Erros não são correlacionados → nenhum padrão discernível.
  - Erro correlacionados positivamente → os resíduos adjacentes podem ter valores semelhantes.



Resíduos de uma regressão linear ajustada a dados gerados com erros não correlacionados. Não há evidência de uma tendência relacionada ao tempo nos resíduos.

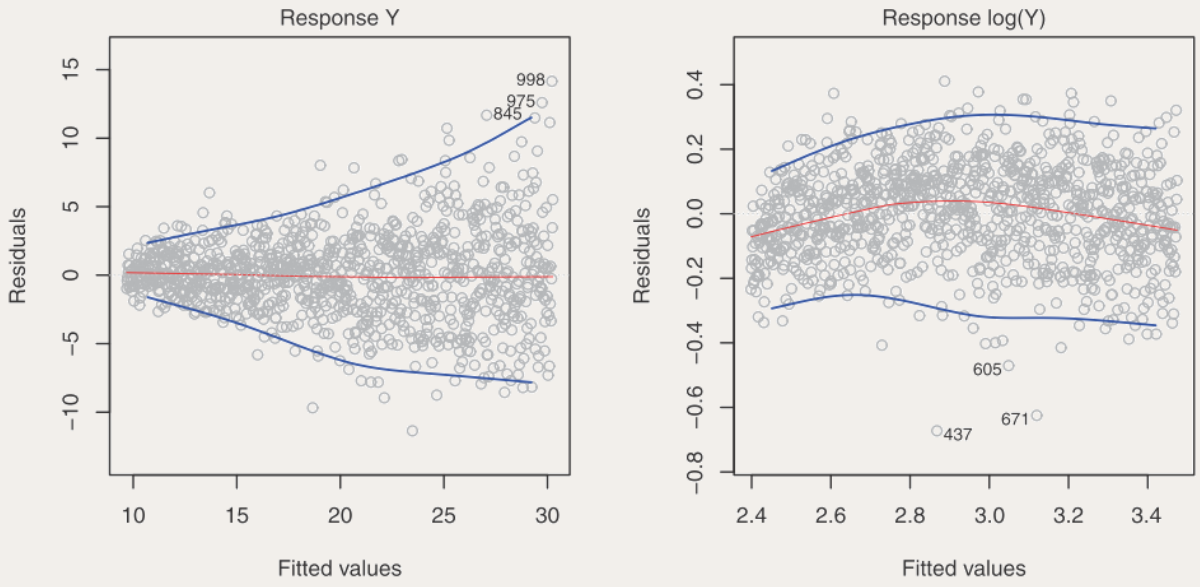


Resíduos de um conjunto de dados no qual erros adjacentes tiveram uma correlação de 0,9. Agora há um padrão claro nos resíduos - os resíduos adjacentes tendem a assumir valores semelhantes.

VARIÂNCIA NÃO CONSTANTE DOS ERROS

- É frequente que as variâncias dos termos de erro não sejam constantes.
  - Por exemplo, as variações dos termos de erro podem aumentar com o valor da resposta.
- Pode-se identificar o problema da heterocedasticidade a partir da presença de uma *forma de funil* no gráfico de *resíduos versus valores ajustados*.
- Possível solução: transformação da resposta y usando uma função côncava como $\log y, \sqrt{y}$ → maior encolhimento das respostas maiores, levando a uma redução na heterocedasticidade.

VARIÂNCIA NÃO CONSTANTE DOS ERROS

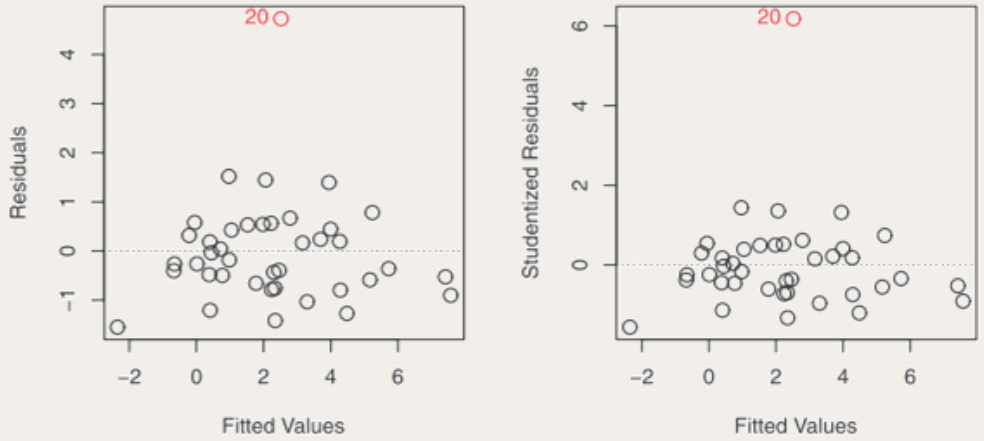


Os resíduos agora parecem ter variância constante, embora haja alguma evidência de um leve relacionamento não linear nos dados.

OUTLIERS

- *Outlier*: ponto para o qual o valor observado está longe do valor previsto pelo modelo.
  - Os *outliers* podem surgir por vários motivos, como o registro incorreto de uma observação durante a coleta de dados.
- Podem ter pouco efeito no ajuste dos mínimos quadrados, mas podem aumentar o erro padrão dos resíduos e reduzir o R^2.

OUTLIERS



- Em vez de plotar os resíduos, podemos plotar os resíduos *studentizados*, calculados dividindo cada resíduo por seu erro padrão estimado. Observações cujos resíduos estudados são maiores que 3 em valor absoluto são possíveis *outliers*.
- Se acreditarmos que um *outlier* ocorreu devido a um erro na coleta ou registro de dados, uma solução é simplesmente remover a observação.
- No entanto, deve-se tomar cuidado, pois um *outlier* pode indicar uma deficiência do modelo, como um preditor ausente.

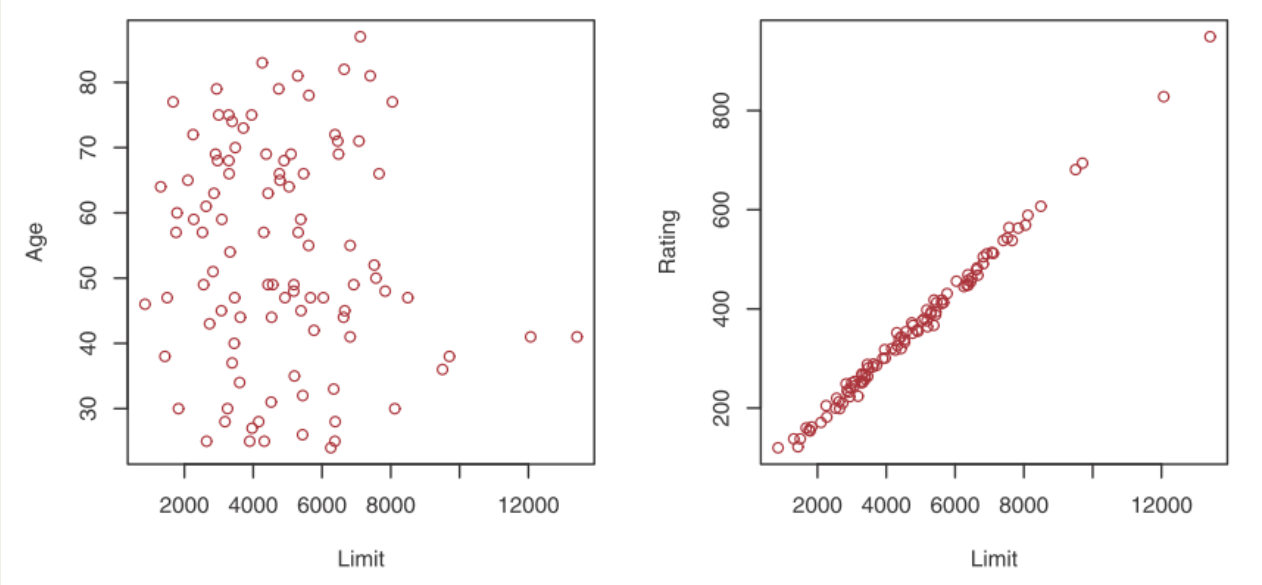
PONTOS DE ALTA ALAVANCAGEM

- Observações com alta alavancagem têm um valor incomum para x_i.
- As observações de alta alavancagem tendem a ter um impacto considerável na linha de regressão estimada.
- Assim como na regressão linear simples, podemos utilizar a distância de Cook para avaliar a presença de pontos de alta alavancagem.

COLINEARIDADE

- Colinearidade refere-se à situação em que duas ou mais variáveis preditoras estão intimamente relacionadas entre si.
- A presença de colinearidade pode representar problemas no contexto de regressão, uma vez que pode ser difícil separar os efeitos individuais de variáveis colineares na resposta.
- A colinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

COLINEARIDADE



À esquerda, os dois preditores parecem não ter relação óbvia. Em contraste, à direita, os preditores estão altamente correlacionados entre si, e dizemos que eles são colineares.

COLINEARIDADE

Algumas indicações da presença de colinearidade são:

1. Coeficientes de correlação linear entre pares de variáveis explicativas ficam muito próximos de -1 ou 1;
2. Grandes alterações nas estimativas dos coeficientes de regressão quando um preditor é adicionado ou retirado do modelo, ou quando uma observação é alterada ou eliminada;
3. Coeficientes de regressão apresentam sinais algébricos opostos ao esperado a partir de conhecimento teórico;
4. Rejeição da $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, mas nenhuma rejeição das hipóteses $H_0: \beta_j = 0, j = 1,2, \ldots, k$, sobre os coeficientes individuais de regressão;
5. Variáveis explicativas, que teoricamente são consideradas importantes, apresentam coeficientes de regressão com estatística t muito baixa.
6. Os erros-padrão dos coeficientes de regressão são muito altos.

FORMAS DE DETECÇÃO DA COLINEARIDADE

Matriz de correlação dos preditores:

- Um elemento dessa matriz que é grande em valor absoluto indica um par de variáveis altamente correlacionadas → colinearidade nos dados.
- Nem todos os problemas de colinearidade podem ser detectados pela inspeção da matriz de correlação: é possível que exista colinearidade entre três ou mais variáveis, mesmo que nenhum par de variáveis tenha uma correlação particularmente alta → Multicolinearidade.

FORMAS DE DETECÇÃO DA COLINEARIDADE

Fator de inflação da variância (VIF):

O VIF é o fator de variação da razão de variância de $\widehat{\beta}_j$ quando ajusta o modelo completo dividido pela variância de $\widehat{\beta}_j$ se ajustado sozinho. O menor valor possível para o VIF é 1, o que indica a completa ausência de colinearidade.

$$VIF(\widehat{\beta}_j) = \frac{1}{1 - R_j^2}$$

R_j^2 coeficiente de determinação da regressão de x_j sobre as outras variáveis preditoras.

- Ela mede o grau em que cada variável preditora é explicada pelas demais variáveis independentes.
- Maior VIF, mais severa a multicolinearidade.
- Se VIF > 10, a multicolinearidade causará efeitos nos coeficientes de regressão (outros autores – não mais que 4 ou 5).

MULTICOLINEARIDADE

Duas soluções simples para resolver o problema de multicolinearidade:

- *Eliminar uma das variáveis problemáticas da regressão*. Isso geralmente pode ser feito sem muito comprometimento com a regressão, já que a presença de colinearidade implica que as informações que essa variável fornece sobre a resposta são redundantes na presença das outras variáveis.
- *Combinar as variáveis colineares em um único preditor*, criando uma nova variável.

REFERÊNCIAS

- JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.
- MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to linear regression analysis**. John Wiley & Sons, 2012.
- MORETTIN, Pedro; BUSSAD, Wilton. **Estatística Básica**, 6. ed.
- MCDONALD, John H. **Handbook of biological statistics**. Baltimore, MD: sparky house publishing, 2009.
- RAWLINGS, John O.; PANTULA, Sastry G.; DICKEY, David A. **Applied regression analysis: a research tool**. Springer Science & Business Media, 2001.
- https://onlinecourses.science.psu.edu/stat501, acessado em 05 de abril de 2020.