ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula 11

Análise de Resíduos

Alexandre Ribeiro Leichsenring
alexandre.leichsenring@usp.br

Organização

1. Regressão com variáveis qualitativas

2. Análise de Resíduos
   - Homocedasticidade
   - Normalidade
   - Independência

Regressão com variáveis qualitativas

- Até agora, as variáveis dependente e independentes nos nossos modelos tinham significado quantitativo (salário, anos de educação, taxa de aprovação, etc)
- Em trabalhos empíricos precisamos incorporar fatores qualitativos aos modelos de regressão:
  - Sexo ou raça de um indivíduo, região geográfica, tipo de escola (pública, privada), etc
- Variáveis qualitativas podem ser facilmente incorporadas aos modelos de regressão
- Discutiremos a introdução de variáveis qualitativas através de um tipo específico: variáveis binárias.
- A generalização é direta e os *softwares* de análise estatística possibilitam seu uso.

- Fatores qualitativas frequentemente aparecem na forma de informação binária:
  - Masculino x Feminino
  - Tem x não tem computador
  - Firma tem plano de saúde x não tem
  - Etc
- Em todos esses exemplos a informação relevante pode ser capturada através da definição de uma variável binária, ou uma variável zero-um.
- No contexto da análise de regressão, as variáveis binárias são chamadas de variáveis *dummy*

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- Ao definir uma variável *dummy*, é preciso decidir a qual categoria será atribuído o valor 1 e qual será atribuído o valor 0
- Por exemplo:

$$\begin{cases} 1, & \text{mulher} \\ 0, & \text{homem} \end{cases}$$

ou

$$\begin{cases} 1, & \text{homem} \\ 0, & \text{mulher} \end{cases}$$

?

- A maneira como definimos não é importante... a mesma informação será capturada em qualquer das duas formulações
- Mas é importante que o nome dado à variável reflita a formulação (no primeiro caso, devemos chamar a variável de algo como *sexo feminino*, enquanto no segundo caso, *sexo masculino*)
- Por que usamos os valores zero e um para descrever informação qualitativa?
- Num certo sentido, esses valores são arbitrários: qualquer par de valores diferentes serviria
- A formulação acima resultam em modelos cujos parâmetros relacionados têm interpretações naturais

Uma única variável *dummy*

- Como incorporamos informações binárias nos modelos de regressão?
- No caso mais simples, com uma única variável *dummy*, a adicionamos diretamente como variável explicativa na equação.
- Ex: considere a equação de determinação de salário:

$$salario = \beta_0 + \delta_0 feminino + \beta_1 educ + u$$

- Usamos δ_0 para ressaltar o seu papel de variável binária (depois podemos usar a representação padrão)
- No modelo acima, apenas dois fatores afetam o salário: sexo e escolaridade
- Como:

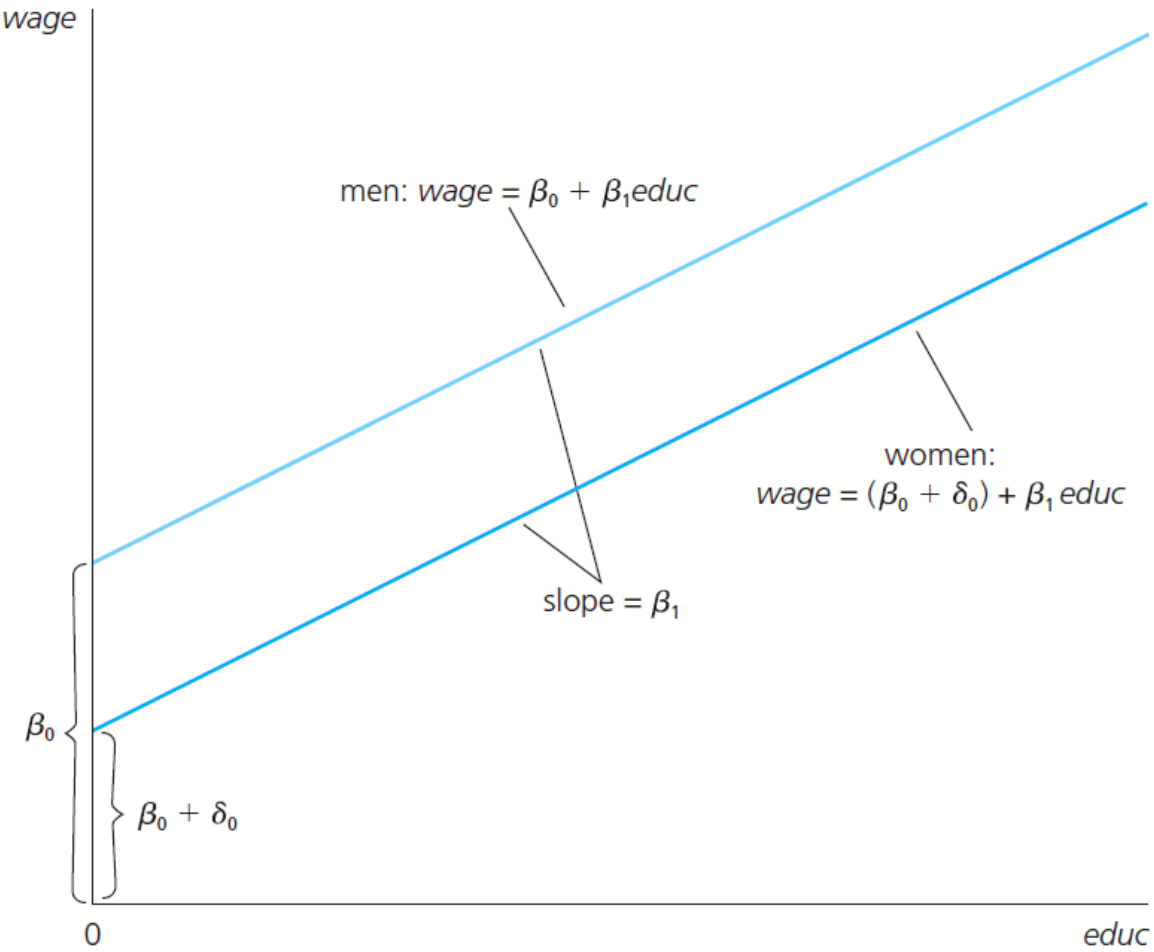
$$feminino = \begin{cases} 1, & \text{se mulher} \\ 0, & \text{se homem} \end{cases}$$

  o parâmetro δ_0 terá a seguinte interpretação:
  - δ_0 é a diferença de salário entre mulheres e homens, dada a mesma escolaridade (e o mesmo termo de erro u)
- O coeficiente δ_0 determina uma possível existência de discriminação contra mulheres:
  - Se $\delta < 0$, então, para outros fatores no mesmo nível, a mulher ganha menos do que o homem em média

- Assumindo a hipótese de média condicional zero:

$$\begin{aligned}\delta_0 &= \mathbf{E}(salario|feminino = 1, educ) - \mathbf{E}(salario|feminino = 0, educ) \\ &= \mathbf{E}(salario|mulher, educ) - \mathbf{E}(salario|homem, educ)\end{aligned}$$

- A escolaridade é a mesma $\Rightarrow \delta_0$ é devido apenas ao sexo



- No modelo acima, escolhemos $sexo = masculino$ como categoria de referência (a comparação é feita contra esse grupo)
- δ_0 é a diferença dos interceptos de mulheres e homens
- Poderíamos escolher as mulheres como categoria de referência:

$$salario = \alpha_0 + \gamma_0 masculino + \beta_1 educ + u$$

e nesse caso

  - α_0 será o intercepto para mulheres e
  - $\alpha_0 + \gamma_0$ o intercepto para homens.

- Não importa como escolhemos a categoria de referência, apenas devemos ter isso em mente qual é ela
- Nada muda muito quando incluímos outras variáveis explicativas. Por exemplo:

$$salario = \beta_0 + \delta_0 feminino + \beta_1 educ + \beta_2 exper + \beta_3 permanencia + u$$

- Se $educ$, $exper$ e $permanencia$ são características relevantes de produtividade, a hipótese nula de não-diferença entre homens e mulheres é:

H_0: $\delta_0 = 0$

e a hipótese alternativa:

H_1: $\delta_0 < 0$

Exemplo

Usando os dados de wage1.sav, estimamos o modelo acima proposto. Por hora, usamos $salario$ ao invés de $\log(salario)$ como variável dependente:

$$\hat{salário} = -1,57 - 1,81\ feminino + 0,572\ educ + 0,25\ exper + 0,141\ permanencia$$

- O intercepto negativo nesse caso não tem muito significado (pois ninguém tem valor zero para todos as variáveis $educ$, $exper$ e $permanencia$ na amostra
- O coeficiente de $feminino$ é interessante, pois mede a diferença média dos salários horários entre homens e mulheres que têm os mesmos níveis de $educ$, $exper$ e $permanencia$.
- Se pegamos uma mulher e um homem com os mesmos níveis de escolaridade, experiência e permanência, a mulher ganha em média $\$1,81$ a menos do que o homem

Análise de Resíduos

Outras questões estão em jogo na avaliação da qualidade do ajuste além de R^2:

- Como saber se o modelo é adequados aos dados?
- Função de regressão é linear?
- Resíduos têm distribuição normal?
- As hipóteses RLM.1 a RLM.5 assumidas estão satisfeitas?

Uma ou mais características do modelo podem não ser apropriadas para um conjunto de dados.

Vamos apresentar alguns métodos para a análise das hipóteses que envolvem os termos de erro.

As hipóteses

Hipótese RLM.1 (Linear nos parâmetros)

No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro (ou perturbação) u como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Hipótese RLM.2 (Amostragem Aleatória)

Temos uma amostra aleatória de n observações

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : 1 = 1, 2, ..., n\}$$

proveniente do modelo populacional descrito na Hipótese RLM.1

Hipótese RLM.3 (Média condicional zero)

O erro u tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes:

$$\mathbf{E}(u|x_1, x_2, \ldots .x_k) = 0$$

As hipóteses

Hipótese RLM.4 (Colinearidade não perfeita)

Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante, e não há relações lineares *exatas* entre as variáveis independentes.

Hipótese RLM.5 (Homocedasticidade)

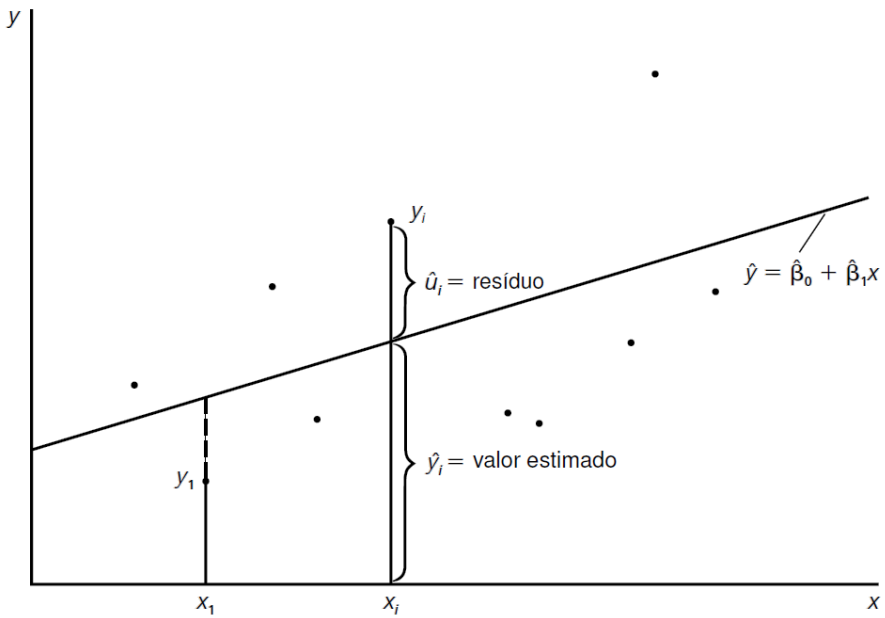
$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

Ideia

Fazer uma análise dos resíduos gerados pelo ajuste do modelo de regressão, isto é, uma análise dos $\hat{u}_i$.

Vamos recordar as suposições sobre os termos de erro u_i:

- $u_i \sim N(0, \sigma^2)$, independentes, de maneira que Cov$(u_i, u_j) = 0, i \neq j$
- Cov$(u_i, X_i) = 0$



Algumas **Propriedades dos Estimadores de Mínimos Quadrados**.

- A soma, e portanto a média amostral dos resíduos resultantes do método MQ, é zero.

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

- A covariância entre os regressores e os resíduos do método MQ é zero, isto é $\text{Cov}(X_i, \hat{u}_i) = 0$, ou

$$\sum_{i=1}^{n} X_i \hat{u}_i = 0$$

resta verificar as suposições referentes à **heterocedasticidade** (variâncias constantes), à **independência** e à **normalidade** dos resíduos.

Lembramos que os resíduos $\hat{u}_i$ observados na amostra são dados por

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

Após estimarmos o modelo de regressão, temos então n observações de $\hat{u}_i$. Os testes sobre a validade dessas suposições são baseados nessa amostra.

A maioria dos testes são baseados em medidas descritivas e/ou representação gráfica dos resíduos.

Homocedasticidade

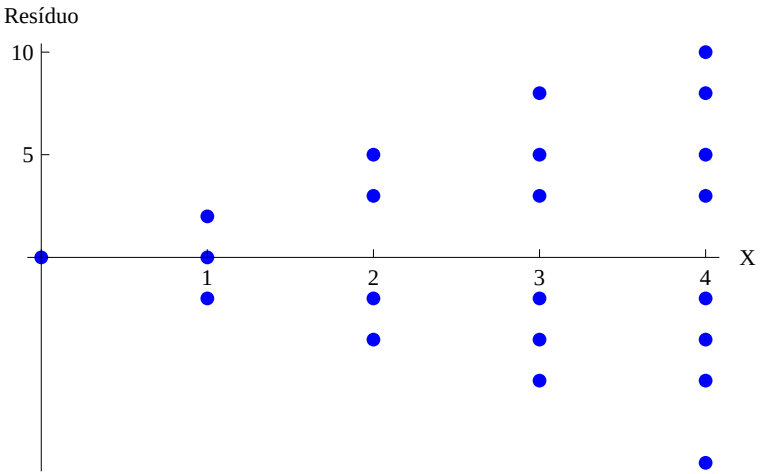
Quando assumimos que $u_i \sim N(0, \sigma^2)$, estamos implicitamente assumindo que a variância de $Y|X_i$ não depende de um particular valor X_i. Em outros termos

$$Var(u \mid X_i) = \sigma^2 \ \ \forall X_i$$

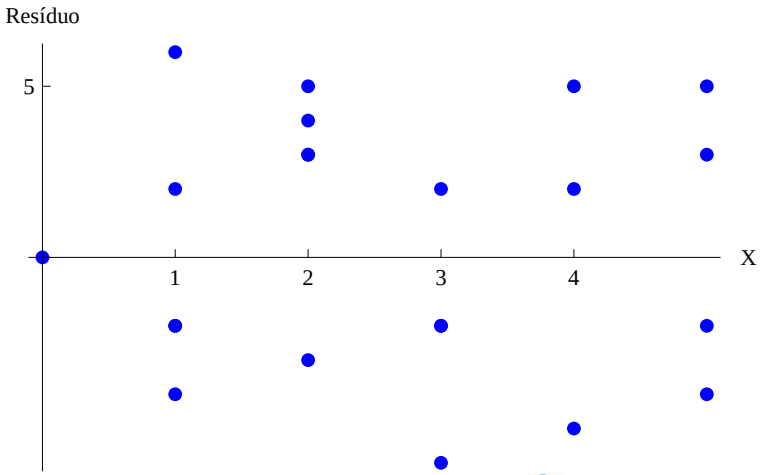
A homocedasticidade diz respeito à constância da variância dos resíduos.

Teste: Gráfico de dispersão dos resíduos.

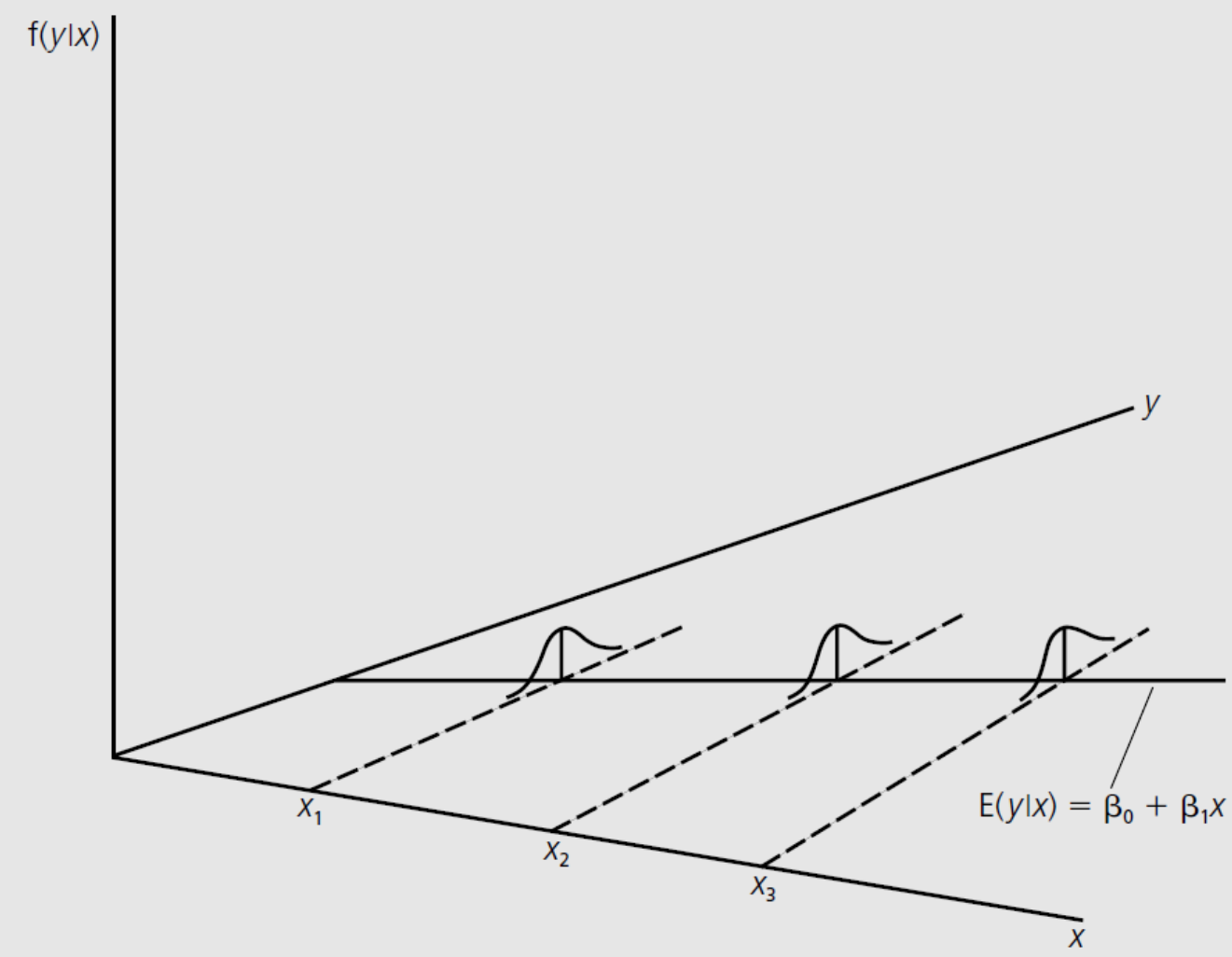
Tendência nos resíduos.



Dispersão nos resíduos.

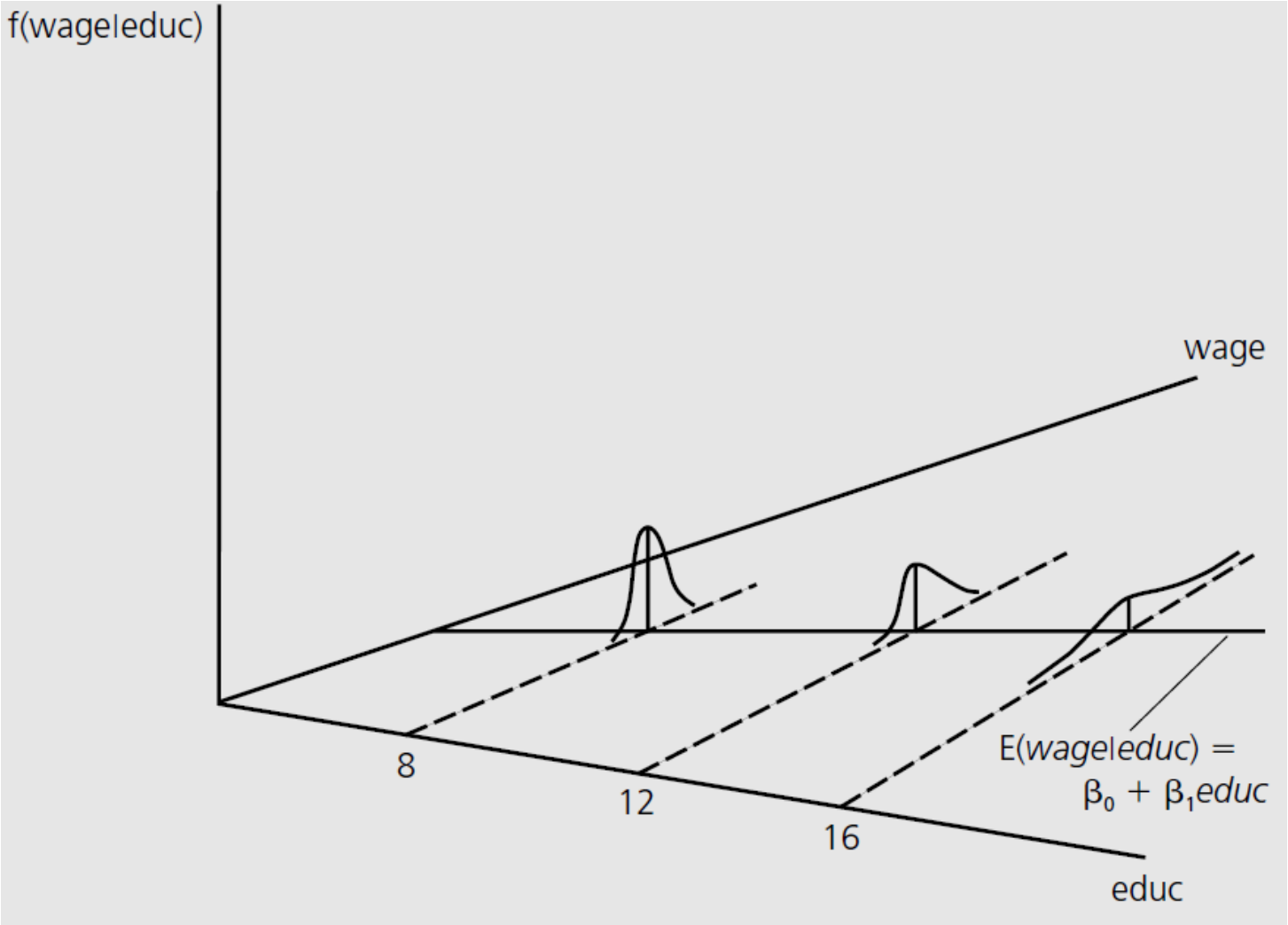


O modelo de regressão simples sob homocedasticidade



Normalidade

Variância do salário crescente com escolaridade



Análise da distribuição de frequências dos resíduos.

Histograma

Informações sobre a forma da função densidade dos resíduos

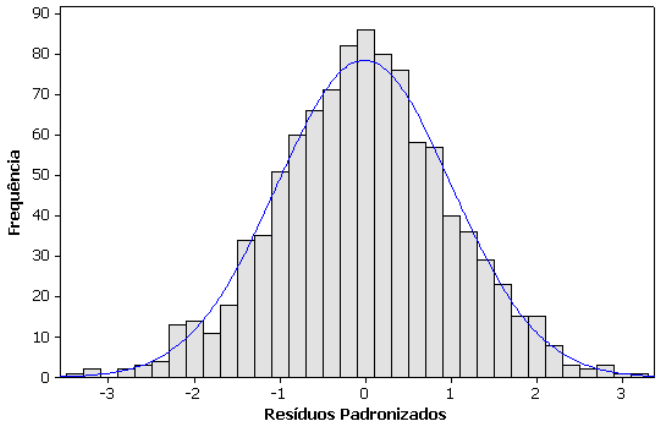
Gráfico de probabilidade normal

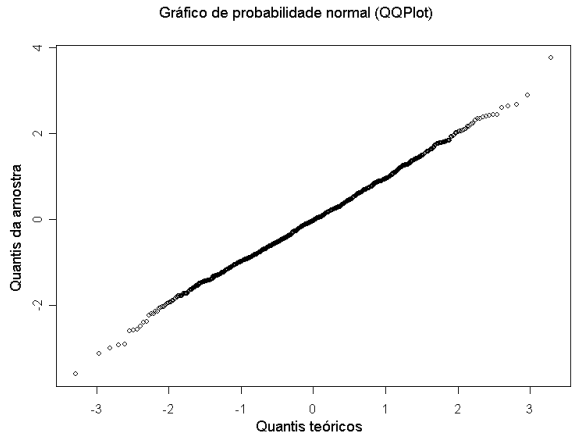
Ferramenta gráfica que compara os quantis observados com os quantis esperados de uma distribuição normal. No caso de resíduos normalmente distribuídos esperamos que os pares (quantis observados , quantis esperados) estejam alinhados em uma linha reta. Assim, comparamos os resíduos com os quantis de uma normal com média 0 e variância dada pelo valor estimado para a variância u, isto é:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}.$$

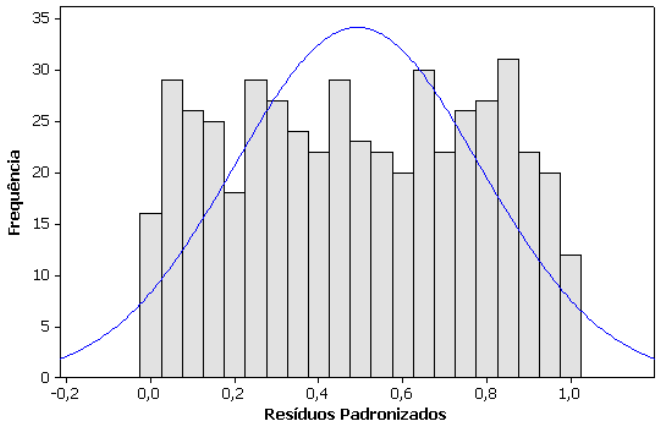
Alternativamente, pode-se comparar os resíduos padronizados (i.e., $\tilde{e}_i = \frac{\hat{u}_i}{\hat{\sigma}}$) com os quantis correspondentes da normal padrão.

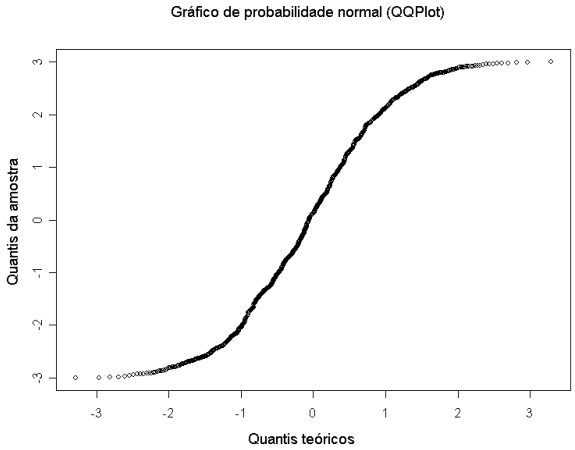
Evidências de Normalidade





Afastamento da Normalidade





Exemplo

Vamos fazer a análise dos resíduos gerados pelo modelo ajustado para os dados de desempenho em matemática (mate10) do arquivo meap93.sav.

Queremos:

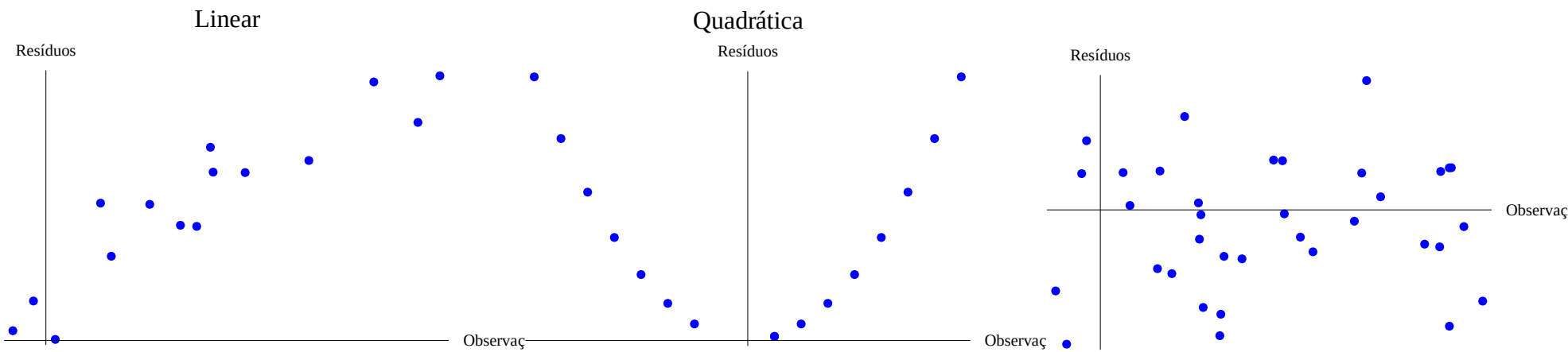
- *Gráfico de dispersão dos resíduos por X_i (ou valores ajustados para Y_i)*
- *Gráfico de dispersão dos resíduos ordenados pela ordem de observação (particularmente importantes em observações coletadas através do tempo)*
- *Histograma dos resíduos*
- *Gráfico de probabilidade normal*

Independência

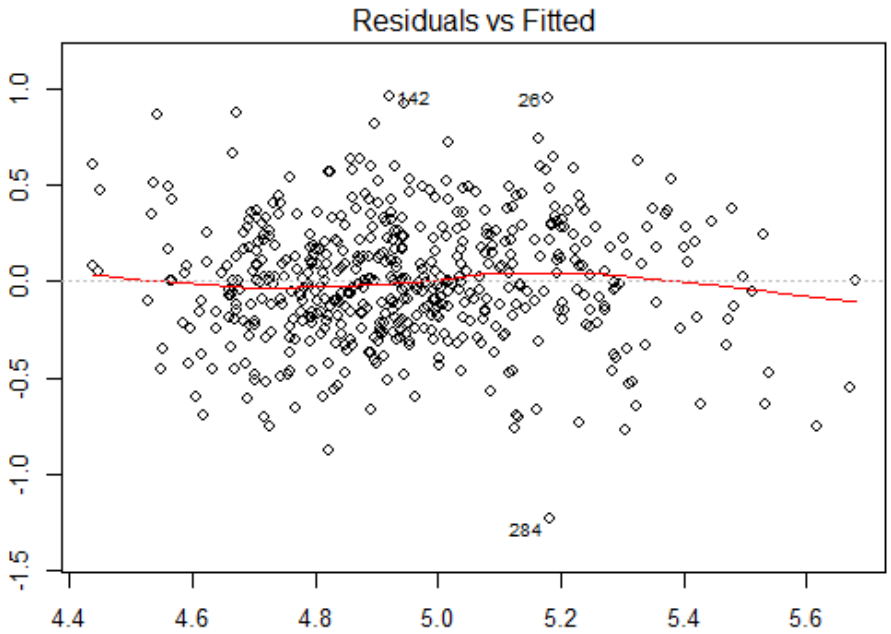
Os dados coletados ao longo de períodos de tempo algumas vezes exibem um efeito de *autocorrelação* entre observações sucessivas. Nestes casos, existe uma relação entre resíduos consecutivos.

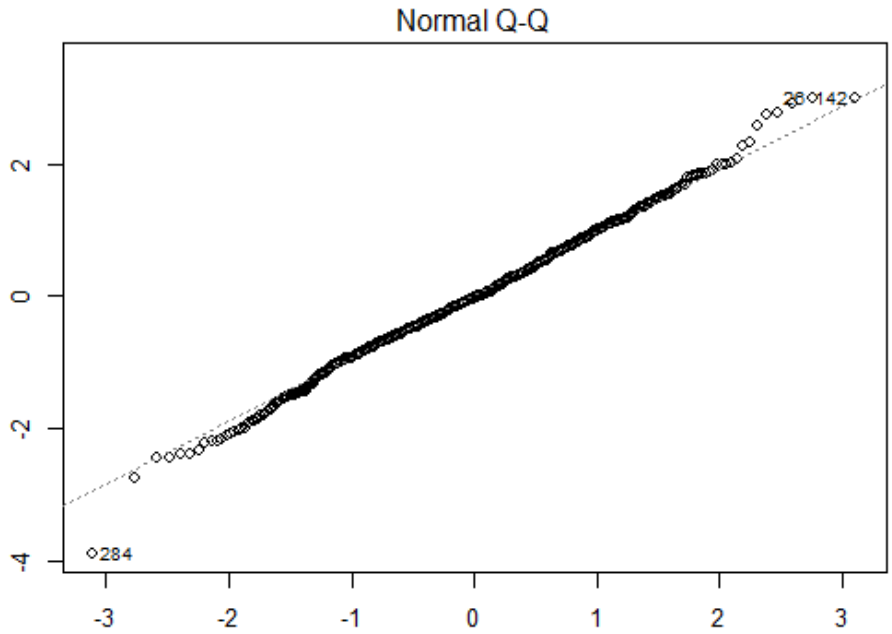
Teste:
- Gráfico de dispersão dos resíduos ordenados pelo período de tempo em que foram coletados.
- Gráfico de resíduos contra resíduos defasados: $\hat{u}_i \times \hat{u}_{i-1}$.

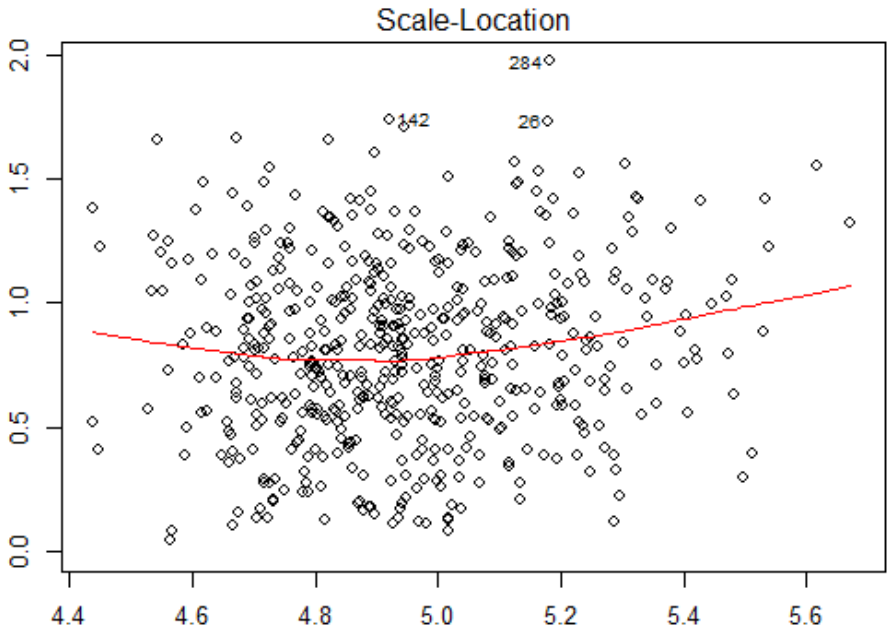
Resíduos ordenados: Dependência linear, quadrática e independência

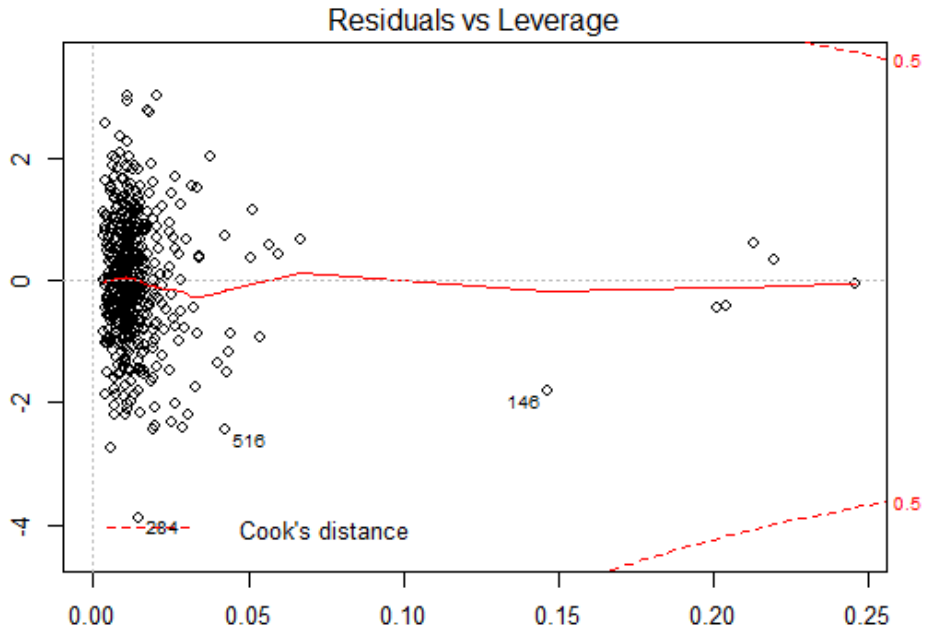


Análise de Resíduos no R - função plot()





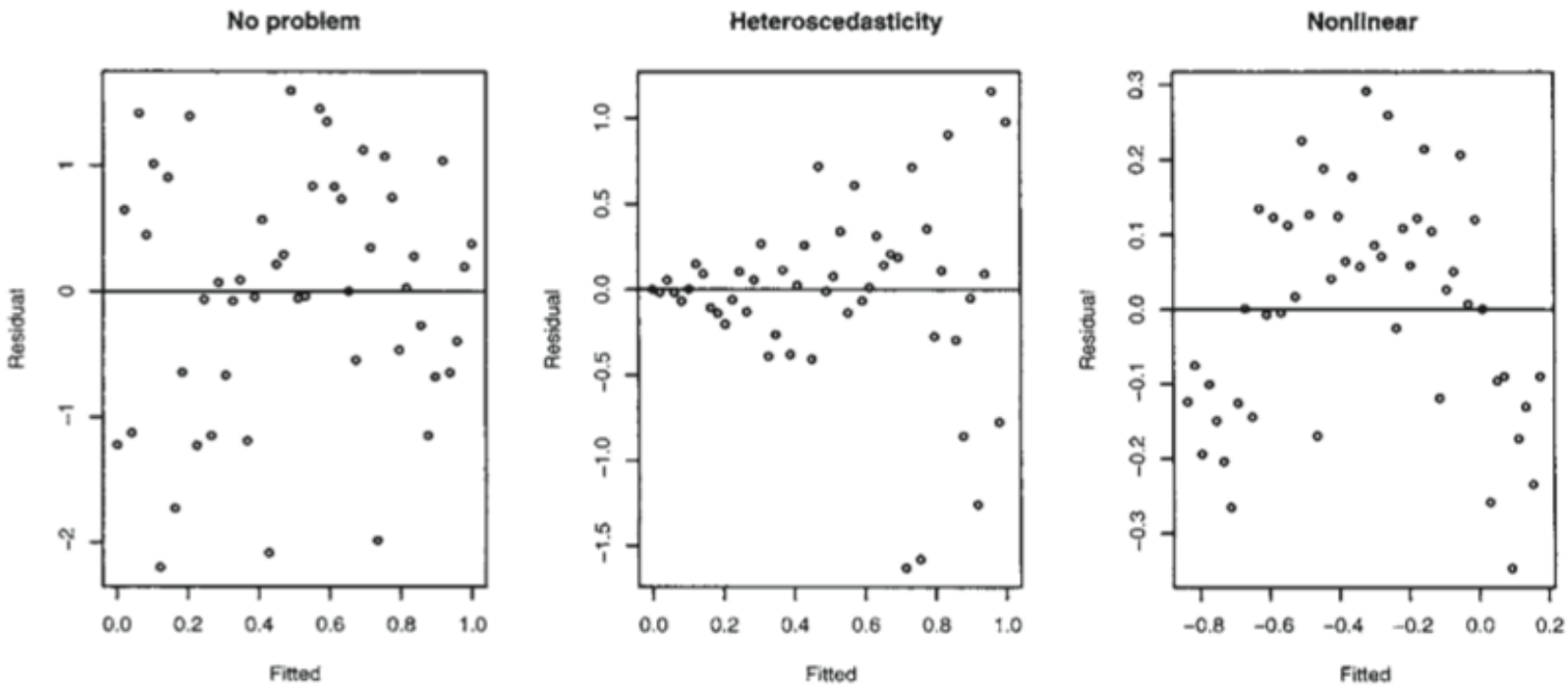




1. **Residuals vs Fitted**
   - Usado para checar a suposição de relações lineares e homoscedasticidade
   - Uma linha aproximadamente horizontal, sem padrão definido indica linearidade
2. **Normal Q–Q**
   - Usado para examinar se os resíduos têm distribuição Normal
   - Quando normalmente distribuídos, pontos acompanham a linha
3. **Scale-Location**
   - Usado para checar homoscedasticidade (variância constante)
   - Linha vermelha deve se apresentar aproximadamente horizontal
4. **Residuals vs Leverage**
   - Usado para identificar casos influentes, ou seja, valores extremos que podem influenciar a regressão conforme forem incluídos ou excluídos da análise

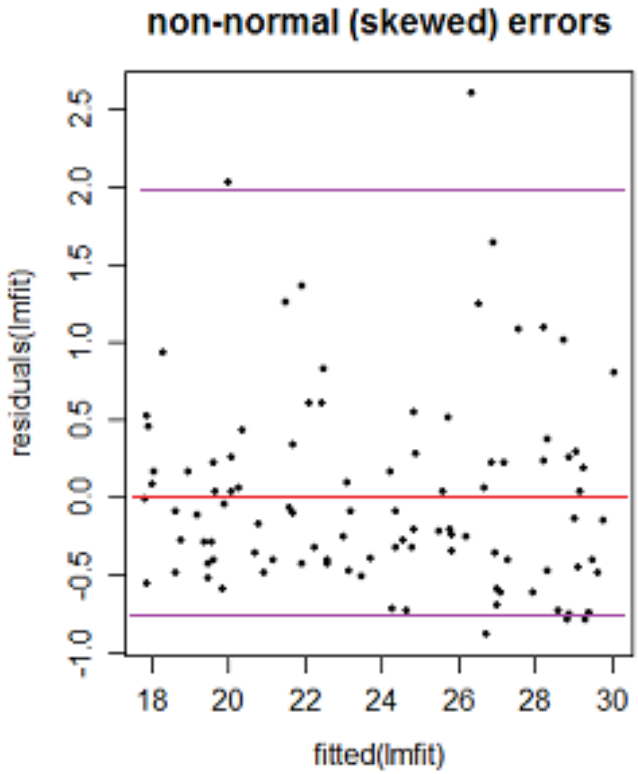
1 Residuals vs Fitted

- Usado para checar a suposição de relações lineares e homoscedasticidade
- Uma linha aproximadamente horizontal, sem padrão definido indica linearidade



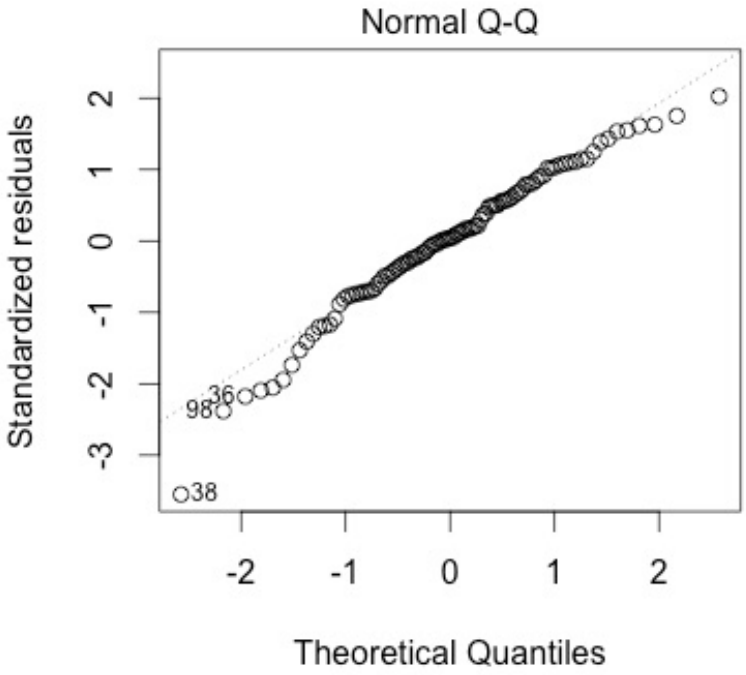
1 Residuals vs Fitted

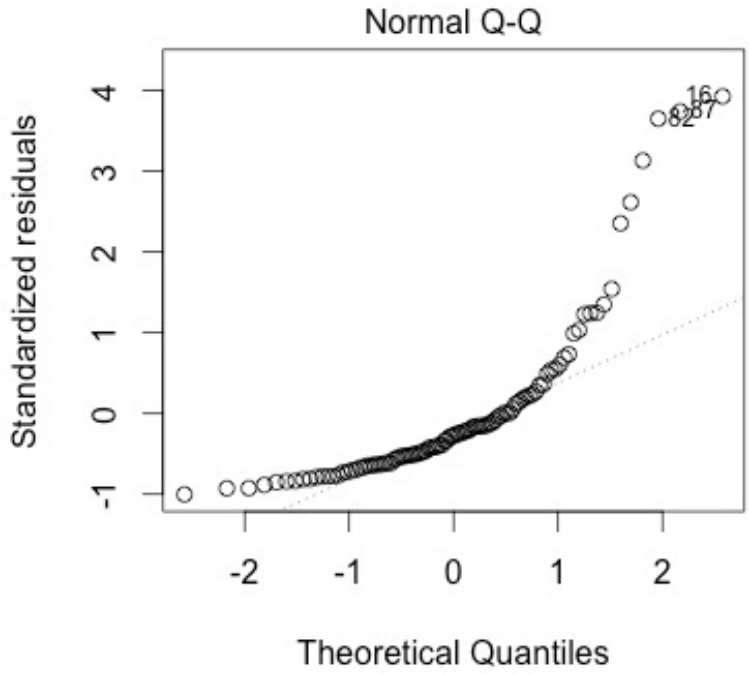
- Usado para checar a suposição de relações lineares e homocedasticidade
- Uma linha aproximadamente horizontal, sem padrão definido indica linearidade



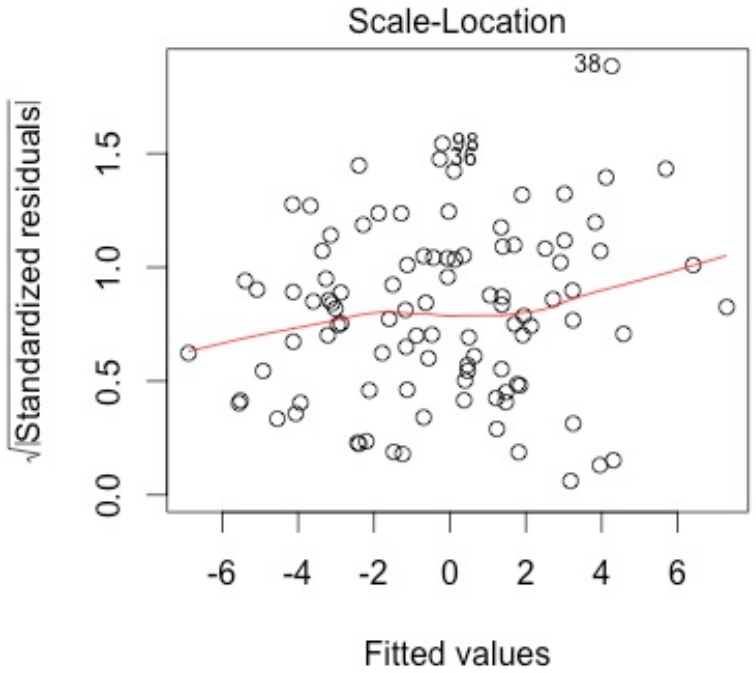
Normal Q–Q

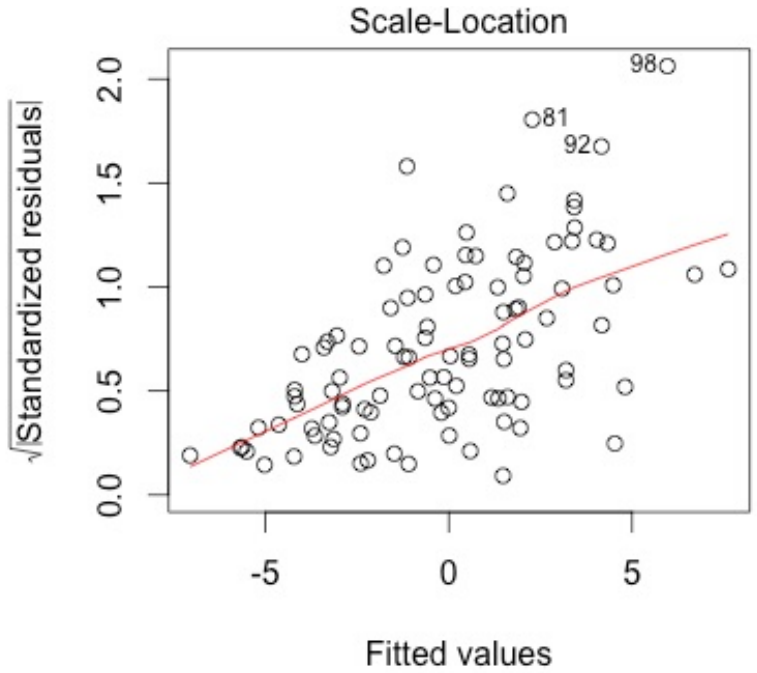
- Usado para examinar se os resíduos têm distribuição Normal
- Quando normalmente distribuídos, pontos acompanham a linha



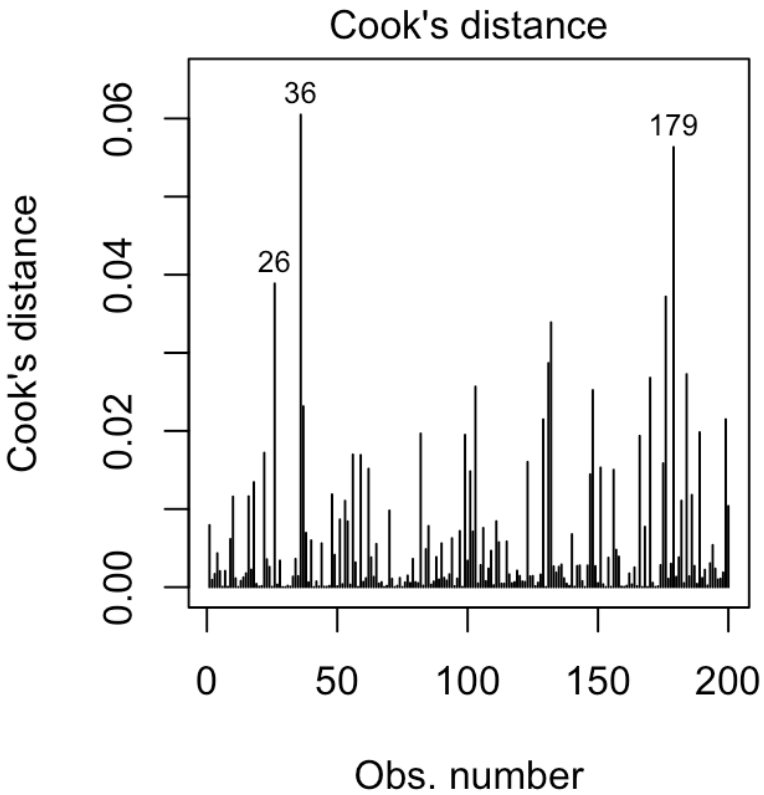


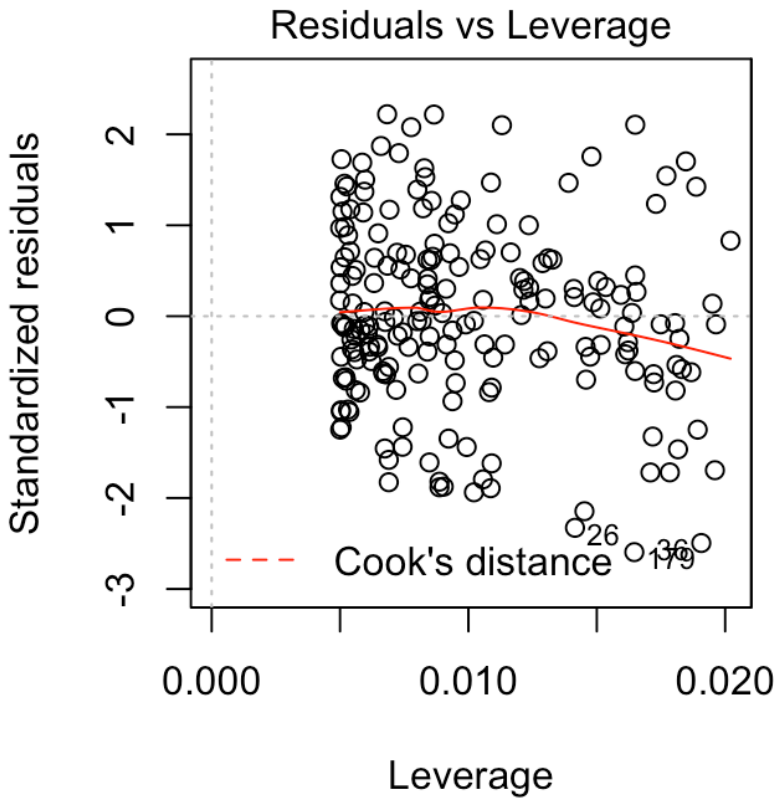
- **Scale-Location**
  - Usado para checar homoscedasticidade (variância constante)
  - Linha vermelha deve se apresentar aproximadamente horizontal



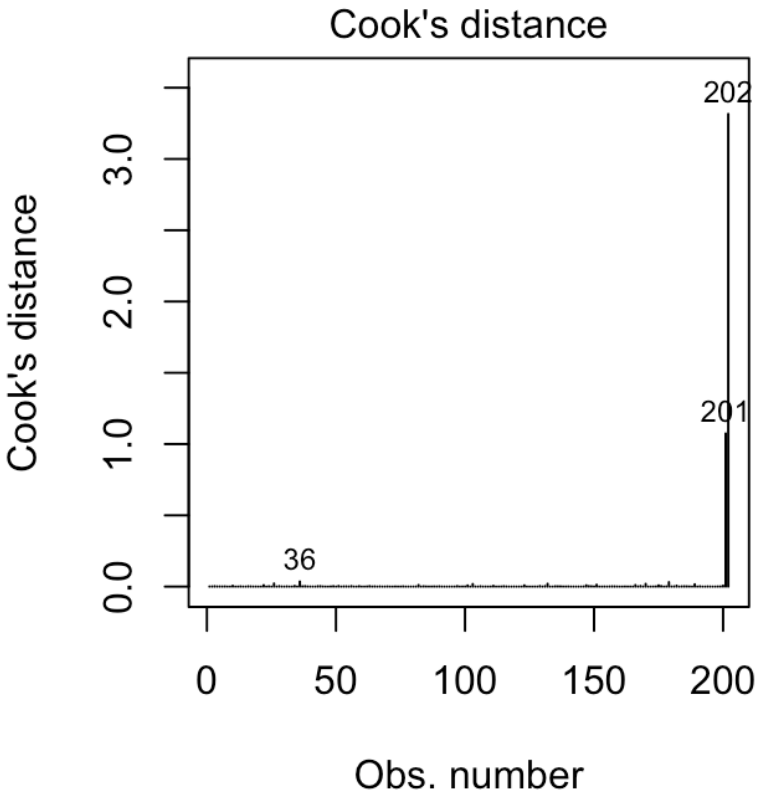


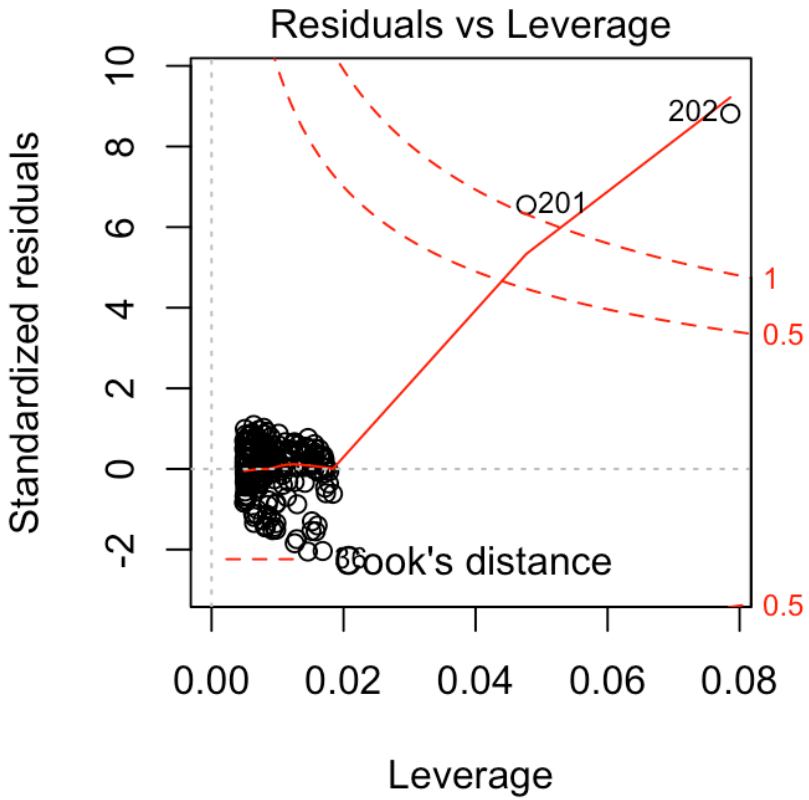
- **Residuals vs Leverage**
  - Usado para identificar casos influentes, ou seja, valores extremos que podem influenciar a regressão conforme forem incluídos ou excluídos da análise





- **Residuals vs Leverage**
  - Usado para identificar casos influentes, ou seja, valores extremos que podem influenciar a regressão conforme forem incluídos ou excluídos da análise





Alertas na construção do modelo de regressão

- A ampla disponibilidade de *softwares* de estatística e de planilhas removeu a difícil barreira dos cálculos
- Acesso às técnicas avançadas nem sempre foi acompanhado por um entendimento claro sobre como utilizar a análise de regressão de forma apropriada.

Dificuldades envolvidas na análise de regressão:

1. Falta de atenção às suposições do modelo de regressão (homocedasticidade, normalidade e independência).
2. Saber quais alternativas utilizar na regressão linear quando as suposições são violadas.
3. Utilizar um modelo de regressão sem conhecimento do assunto.

Passos recomendados:

1. Sempre iniciar com gráfico de dispersão para observar a possível relação entre as variáveis explicativas x_i e a resposta y.
2. Verificar se as suposições do modelo de regressão (homoscedasticidade, normalidade e independência) estão satisfeitas, antes de prosseguir e utilizar seus resultados.
3. Se as suposições em (2) forem violadas deve-se utilizar métodos alternativos para remediar as violações, como por exemplo transformações nas variáveis explicativas ou respostas.
4. Se as suposições forem satisfeitas, deve-se realizar testes em relação aos coeficientes de regressão e construir intervalos de confiança e previsão.