Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

(rachid.muleia@uem.mz)

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Regressão Linear- inferência sobre os parâmetros

Ano lectivo: 2023

Inferência sobre β_1

Considere o modelo de regressão populacional

$$y_i = \beta_0 + \beta_1 x_i + \epsilon, \quad \epsilon_i \sim N(0, \sigma^2)$$

- No modelo de regressão linear, a estimação dos parâmetros pode ser feita usando o MQO, o qual não exige o pressuposto da normalidade
- A partir do MQO, pode-se obter os estimadores pontuais, $\hat{\beta}_0$ e $\hat{\beta}_1$, para os parâmetros populacionais β_0 e β_1
- A inferência estatística sobre os parâmetros do modelo de regressão, baseia-se no pressuposto da normalidade
  - ➜ Os termos de erro ϵ_i seguem uma distribuição normal e distribuídos de forma idêntica e independente.
  - ➜ Considerando o teorema do limite central, assume-se que o procedimento inferencial é correcto

Cont.

- A inferência sobre os parâmetros de regressão pode ser feito por via dos testes de hipóteses e intervalos de confiança
- No modelo de regressão, o interesse primário está sobre o parâmetro β_1
  - ➜ $\beta_1 = 0 \longleftrightarrow$ não existe relação linear entre X e Y
  - ➜ mas, $\beta_1 = 0 \neq$ não existe relação entre X e Y
- A inferência sobre o parâmetro β_1, exige que se conheça a variabilidade a volta $\hat{\beta}_1$. Isto significa que, se tomarmos uma outra amostra e calcular o valor de $\hat{\beta}_1$, quão próximo este valor estará do parâmetro β_1
- A distribuição de $\hat{\beta}_1$, resultante de repetidas amostra, é designada de distribuição amostral de $\hat{\beta}_1$
  - ➜ $E(\hat{\beta}_1) = \beta_1 \rightarrow \hat{\beta}_1$ é um estimador não enviesado
  - ➜ $V(\hat{\beta}_1) = \dfrac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \dfrac{\sigma}{S_x\sqrt{n-1}}$, onde $S_x = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

Cont.

- Como reduzi a variabilidade de $\hat{\beta}_1$, $V(\hat{\beta}_1)$ (de tal forma que $\hat{\beta}_1$ esteja próximo de β_1)?
  - ➜ Aumentar o tamanho da amostra n
  - ➜ Aumentar a variabilidade dos valores de X
- Para calacular a variabilidade de $\hat{\beta}_1$ precisamos de de conhecer σ, que é o desvio padrão do termo de erro, ϵ_i
  - ➜ Uma forma intuitiva de estimar o σ é através de $\hat{\sigma} = \sqrt{\frac{\sum_{i=s1}^{n}(\epsilon_i - \bar{\epsilon})^2}{n-1}}$, onde $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Contudo, isto não é possível, visto que β_0 e β_1 são desconhecidos
  - ➜ β_0 e β_1 podem ser estimados por $\hat{\beta}_0$ e $\hat{\beta}_1$, logo o termo de erro ϵ_i pode ser aproximado pelo resíduos $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Então podemos usar o desvio padrão dos resíduos, $S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$, para estimar σ

Distribuição amostral de $\hat{\beta}_1$

A distribuição amostral de $\hat{\beta}_1$ é normal

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right) \quad \rightarrow \quad Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum(x_i - \bar{x})^2}}$$

Esta aproximação é válida:

- se os erros, ϵ_i distribuídos de forma idêntica e independente e $\epsilon_i \sim N(0, \sigma)$
- ou se os erros forem independentes e o tamanho da amostra n for suficientemente grande

Visto que σ é desconhecido, ao substituir por S_e, a estatística do teste T abaixo terá distribuição t-student com $n - 2$ graus de liberdade

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_e / \sqrt{\sum(x_i - \bar{x})^2}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

Intervalo de confiança para β_1

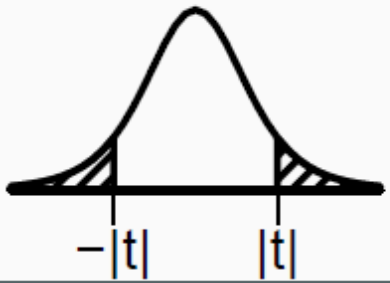
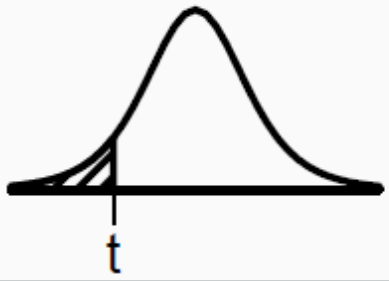
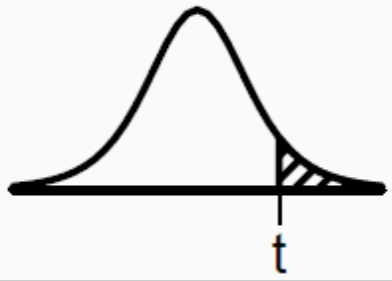
O intervalo de confiança a $(1-\alpha)$ para β_1 é dado por

$$\left[\hat{\beta}_1 - t_{1-\alpha/2;n-2} \times SE(\hat{\beta}_1); \hat{\beta}_1 + t_{1-\alpha/2;n-2} \times SE(\hat{\beta}_1)\right]$$

teste de hipótese para β_1

Para testar a hipótese $H_0 : \beta_1 = a$, usamos a estatística de teste T,

$$T = \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)}$$

H_a	$\beta_1 \neq a$	$\beta_1 < a$	$\beta_1 > a$
P-value			

ue testar H_0 : $beta_1 = 0$ é equivalente a testar se x é útil na previsão de y linearmente

- é possível r (coeficiente de correlação de Pearson) ser bastante pequeno, mas β_1 ser significativamente diferente de zero

Inferência para o intercepto β_0

- Embora o β_0 raramente seja de interesse, todos os resultados para β_1 têm a contraparte para β_0:

- A distribuição amostral de $\hat{\beta}_0$ também é normal

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}}\right)$$

  - ➜ de forma similar, podemos definir o IC para β_0 : $\hat{\beta}_0 \pm t_{1-\alpha/2,n-1} \times SE(\hat{\beta}_0)$
  - ➜ A estatística de teste para $H_0 : \beta = a$ é $T = \dfrac{\hat{\beta}_0 - a}{SE(\hat{\beta}_0)} \sim t_{n-2}$
  - ➜ o valor de p pode ser calculado usando o procedimento similar ao de teste de hipótese para β_1

Exemplo

- Considere o a base de dados do meuse sobre a poluição no rio meuse
- Vamos analisar a relação entre a concentração de *Pb* e *Zn*
- O coeficiente de correlação entre Pb e Zn é de 0.95, o que significa que a Pb e Zn podem estar relacionados linearmente

Exemplo

```
> reg <- lm(lead ~ zinc, data = meuse)
> summary(reg)

Call:
lm(formula = lead ~ zinc, data = meuse)

Residuals:
    Min      1Q  Median      3Q     Max
-79.853 -12.945  -1.646  15.339 104.200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.367688   4.344268   3.998 9.92e-05 ***
zinc         0.289523   0.007296  39.681  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 33.24 on 153 degrees of freedom
Multiple R-squared:  0.9114,Adjusted R-squared:  0.9109
```

Exemplo: Relação entre concentração de Pb e Zn

- A coluna **`Estimates`** nos outputs, representa as estimativas de MQO para $\hat{\beta}_0 = 17.367688$ e $\hat{\beta}_1 = 0.289523$
- A coluna **`Std. Error`** contem as estimativas do erro padrão para o intercepto e a inclinação , $SE(\hat{\beta}_0) = 4.344268$ e $SE(\hat{\beta}_1) = 0.007296$
- O intervalo de confiança a 95% para o β_1 é

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2} \times SE(\hat{\beta}_1) = 0.289523 \pm 1.975 \times 0.007296 \approx (0.2751; 0.3039)$$

- Interpretação: a 95% de confiança, espera-se que a cada aumento unitário na concentração de zinco, espera-se que a concentração de chumbo aumente entre $(0.2751; 0.3039)$

Exemplo: Relação entre concentração de Pb e Zn

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.367688   4.344268   3.998 9.92e-05 ***
zinc         0.289523   0.007296  39.681  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- Observe que a coluna `t value` é a razão entre os valores na coluna `Estimate` e `Std. Error`

$$\frac{17.367688}{4.344268}, \qquad T = \frac{0.289523}{0.007296} \approx 39.68243$$

- Testar $H_0 : \beta_1 = 0$ é equivalente a verificar se zinco está linearmente relacionado com a concentração de chumbo. O $p - value < 2 \times 10^{-16} < 0.05$ atesta que existe uma relação estatisticamente significativa entre o chumbo e zinco

Como Interpretar os demais valores do output

```
Residual standard error: 33.24 on 153 degrees of freedom
Multiple R-squared:  0.9114,Adjusted R-squared:  0.9109
F-statistic:  1575 on 1 and 153 DF,  p-value: < 2.2e-16
```

- `Residual standard error: 33.24 on 153 degrees of freedom`
  Aqui temos a estimativa para a desvio-padrão do termo de erro σ, $S_e = 33.24$ e os graus de liberdade $df = 155 - 2 = 153$

- `Multiple R-squared: 0.9114, Adjusted R-squared: 0.9109`
  Temos aqui o coeficiente de determinação $r^2 = 0.9114$. Isto significa que a concentração de Zinco explica 91% da variação na concentração do chumbo
  `Adjusted R-squared: 0.9109`- Podemos ignorar

- `F-statistic: 1575 on 1 and 153 DF, p-value: < 2.2e-16`
  Valor da estatística F para a anova. Existe uma relação entre o teste t e a estatística $F = t^2$. Podemos, por enquanto, ignorar este valor.

Predição condicional

Existem dois tipos de predição condicional da variável resposta Y dado $X = x_0$, considerando o model de regressão linear simples, $Y = \beta_0 + \beta_1 X + \epsilon$

- Estimação da resposta média, dado que $X = x_0$, $E\left[Y|X = x_0\right] = \beta_0 + \beta_1 x_0$, que pode ser estimado por $\hat{\beta}_0 + \hat{\beta}_1 x_0$
- Predição da variável resposta para um valor específico dado que $X = x_0$, $Y = \beta_0 + \beta_1 x_0 + \epsilon$, pode ser estimado por $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 + 0$

  ➜ Note que a melhor predição para o termo de erro ϵ é a média, que é zero

A diferença nas duas predições está na Incerteza

- Pode-se mostrar que a variância envolvida na estimação da resposta média, $E\left[Y|X = x_0\right] = \beta_0 + \beta_1 x_0$ é dada por:

$$V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

- Para predição de um valor específico, temos uma variabilidade adicional resultante da variabilidade no termo de erro ϵ

$$V(\hat{Y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] + \sigma^2$$

- A variabilidade a volta do novo valor previsto é maior que a variabilidade a volta da estimativa da resposta média para um dado valor

Intervalo de confiança e Intervalo de predição

- O intervalo de confiança a $100 \times (1-\alpha)\%$ para a estimativa da resposta média, dado que $X = x_0$, é dado por

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{1-\alpha/2,n-2} \times \sigma \sqrt{\frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- O intervalo de predição para $Y = \beta_0 + \beta_1 x_0 + \epsilon$ é dado por

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{1-\alpha/2,n-2} \times \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Observe que σ é estimado com base no desvio padrão dos resíduos S_e

Intervalo de confiança e Intervalo de predição

```
> predict(reg, data.frame(zinc=500), interval="confidence",level=0.95)
       fit     lwr      upr
1 162.1292 156.837 167.4213

> predict(reg, data.frame(zinc=500), interval="prediction",level=0.95)
       fit      lwr      upr
1 162.1292 96.25415 228.0042
```

- Quando a concentração de zinco é 500 ppm , espera-se que a concentração de chumbo, em média, seja 162.1 ppm, com um intervalo de confiança a 95% de 156.837 – 167.421
- Quando a concentração de zinco é 500 ppm, a predição para um único valor é de 162.1 ppm, com um intervalo de confiança de 96.25415 – 228.0042
- Observe que o intervalo de predição para uma única observação é mais largo

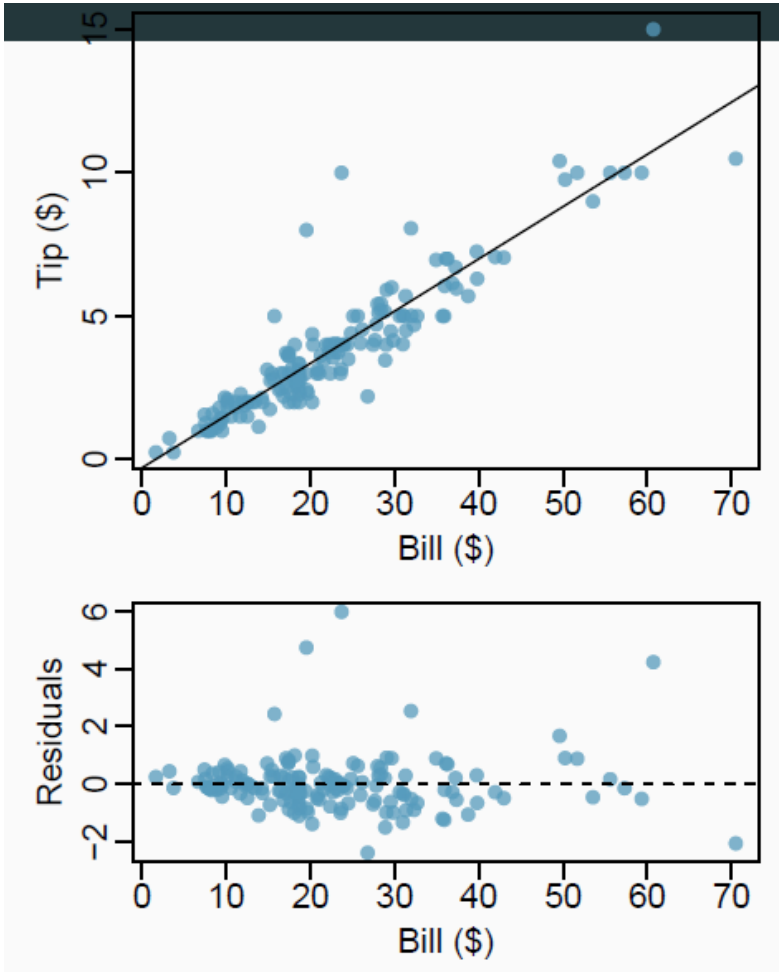
Pressupostos do modelo de regressão linear

- Linearidade
- Constância da variância dos resíduos (homocedasticidade)
- Normalidade
- Independência

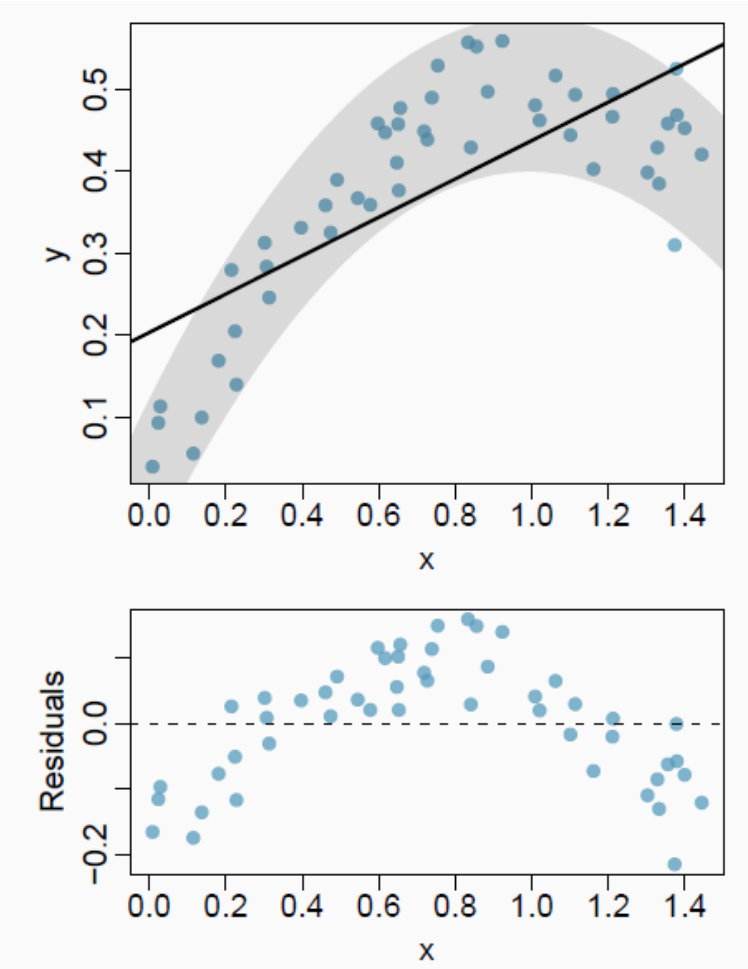
Como verificar os pressupostos?

- Através do gráfico dos resíduos

Se as condições forem satisfeitas, os pontos deverão se espalhar uniformemente em torno da linha zero no gráfico residual



Linearidade

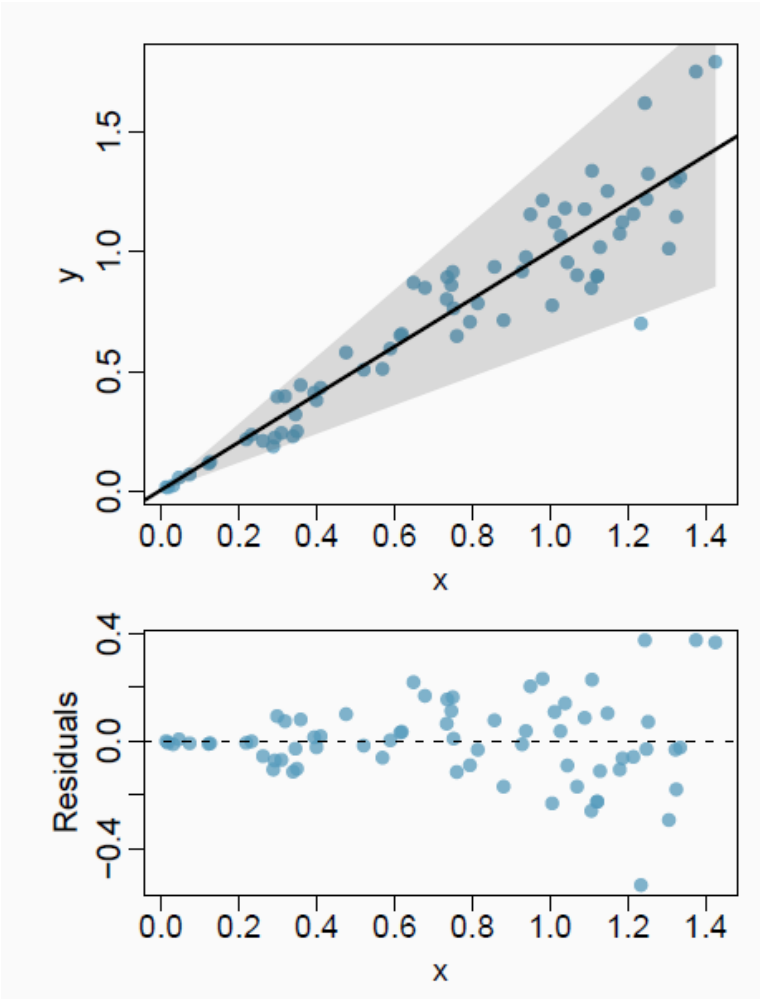


Que pressuposto é violado neste modelo de regressão linear?

- Constância da variância
- Linearidade
- Normalidade dos resíduos

Observe que a correlação entre os resíduos e a variável x é igual a zero, mas correlação zero $\neq$ nenhuma associação. Pode ser uma associação não linear

Homocedasticidade

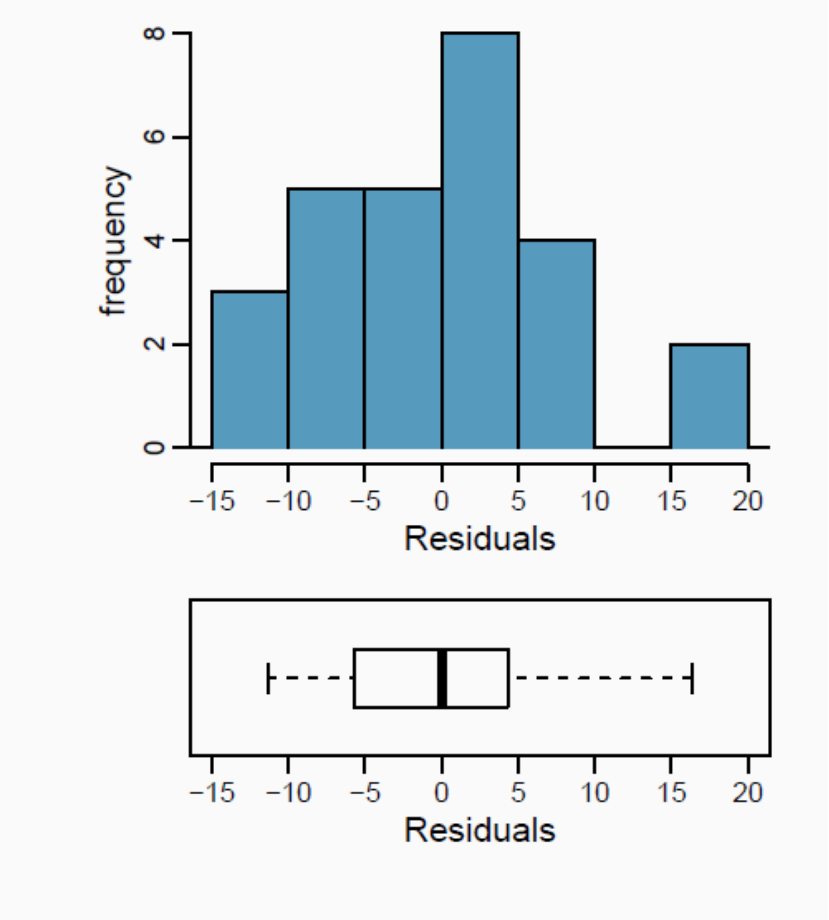


A variabilidade dos pontos em torno da linha de mínimos quadrados deve ser aproximadamente constante, implicando que a variabilidade dos resíduos em torno da linha zero também deve ser aproximadamente constante

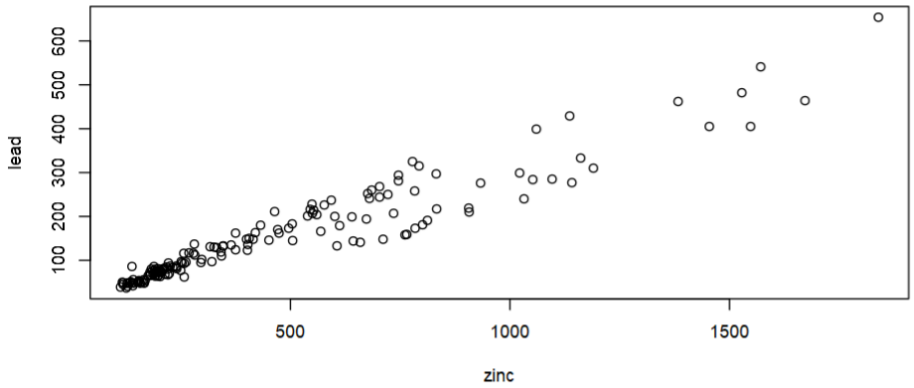
Caso contrário, no caso de heterocedasticidade, as previsões feitas em áreas de maior variabilidade serão piores. Alternativamente, pode-se tentar o método dos mínimos quadrados ponderados

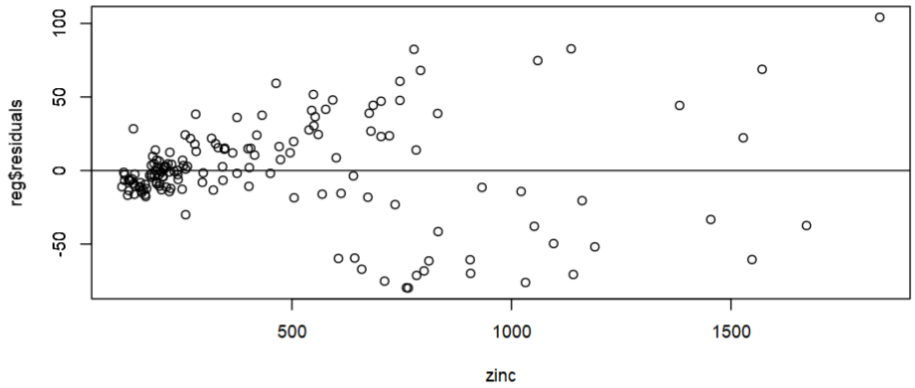
Normalidade dos resíduos

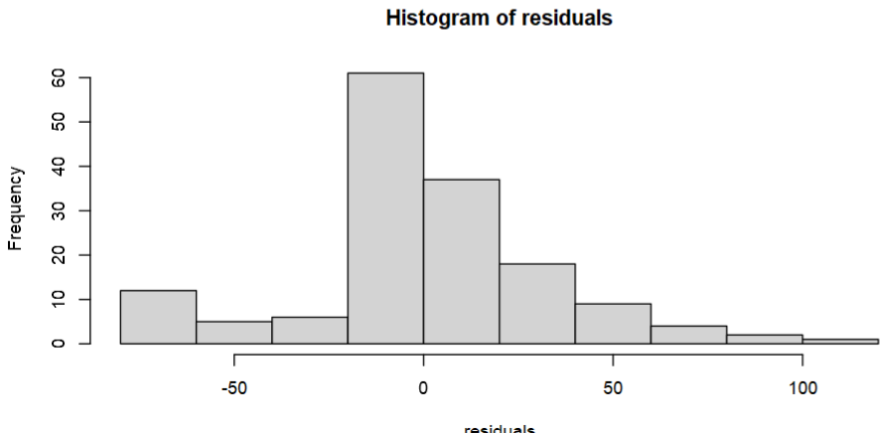
- O pressuposto de normalidade não bastante imprescindível, desde que a amostra seja suficientemente grande
- Pode-se verificar este pressuposto através do histograma,
- Caso haja violação do pressuposto da linearidade e da constância da variância, não há necessidade de verificar a normalidade

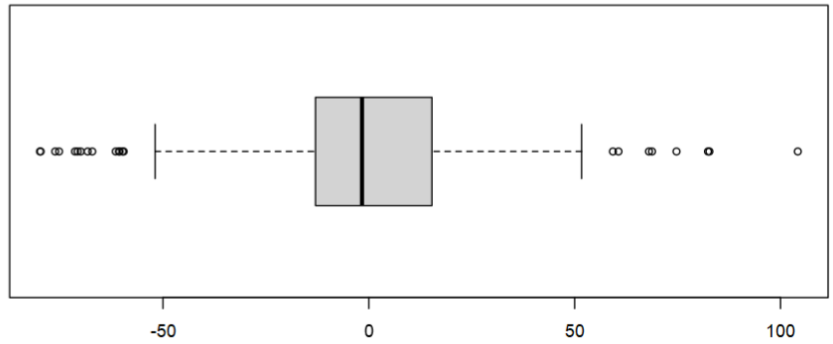


Exemplo: relação de chumbo vs zinco









- Pode-se notar que o pressuposta da homocedasticidade é violado