Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
GUSTAVO L. GILARDONI

Regressão Linear Simples

25 de Abril de 2022

0 Introdução

- Nas unidades anteriores estudamos sobre a relação de uma variável *resposta* com uma variável qualitativa que podia tomar dois ou mais valores. Nesta unidade vamos estudar a relação entre duas variáveis quantitativas. O caso mais simples
- Durante muito tempo o ser humano procurava uma explicação *mecanicista* do mundo que o rodeia. A relação entre coisas ou variáveis era determinista. Se soltarmos um objeto de massa m de uma altura h ele vai chegar ao solo com a velocidade v em exatamente o tempo t.
- No século XX toma muita força uma visão do mundo onde acontecem coisas ao acaso, primeiro nas ciências ditas *duras* (física quântica, por exemplo) e posteriormente nas outras áreas.

Exemplo 1.

Exemplo 2.

1 O Modelo

- Dadas duas variáveis aleatórias Y e X, pensamos modelar o valor esperado condicional de Y dado que $X = x$. Dessa forma escrevemos $\mathrm{E}(Y \mid X = x) = g(x)$ para alguma função g.
- O *erro* $\epsilon = Y - g(x)$ é uma variável aleatória tal que $\mathrm{E}(\epsilon \mid X = x) = 0$ para todo x.
- O caso mais simples, que vamos estudar nesta unidade, ocorre quando (i) a função g é linear em x e (ii) a variância do erro não depende de x.

Suponha que observamos a variável resposta $y_1, \ldots, y_n$ para n individuos com valores $x_1, \ldots, x_n$ da *covariável* X. Assumimos que os valores de $x_1, \ldots, x_n$ são fixados pelo experimentador ou, alternativamente, que toda a inferência será feita *condicional* a esses valores. De acordo à discussão anterior o modelo é

$$\begin{cases} \mathrm{E}(y_i|x_i) = \alpha + \beta\, x_i \\ \\ \mathrm{Var}(y_i|x_i) = \sigma^2 \end{cases} \qquad i = 1, \ldots, n$$

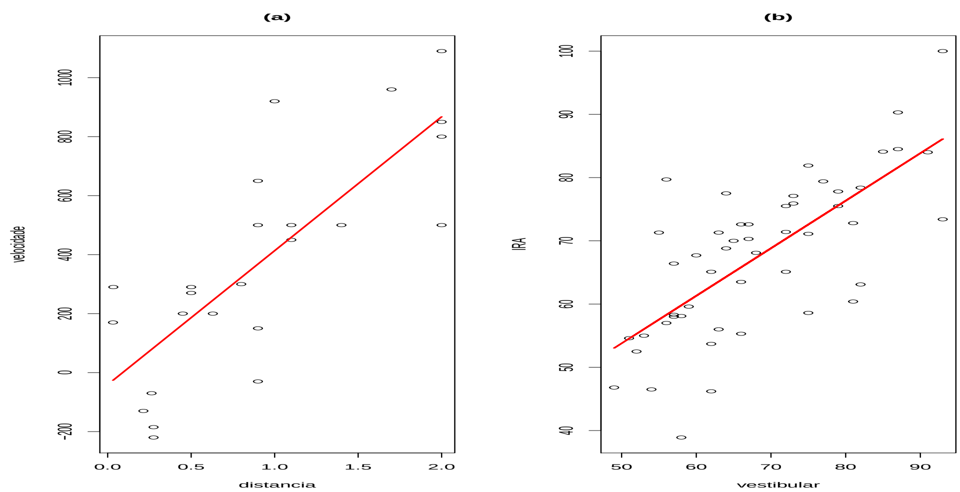

Figura 1: **(a)** Distância da Terra (em megaparsecs) e velocidade de recessão (em km/s) para 24 nebulosas de galáxias (dados coletados pelo astrônomo Edwin Hubble por volta de 1929); **(a)** notas de ingresso no vestibular e IRA durante o primeiro ano do curso para alunos de um curso de estatística.

ou, equivalentemente,

$$y_i = \alpha + \beta\, x_i + \epsilon_i \tag{1}$$

com os erros ϵ_i tais que

$$\begin{cases} \epsilon_1, \ldots, \epsilon_n \text{ são independentes;} \\ \mathrm{E}(\epsilon_i) = 0 \\ Var(\epsilon_i | x_i) = \sigma^2 \end{cases} \qquad i = 1, \ldots, n\,. \tag{2}$$

sendo independentes e tais que $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{Var}(\epsilon_i|x_i) = \sigma^2$. Veja que o modelo tem três parâmetros: α, β e σ^2.

2 Os estimadores de mínimos quadrados

Para ajustar uma reta de regressão como as da Figura 1 podemos proceder da seguinte forma. Considere o vetor $\boldsymbol{y} = (y_1, \ldots, y_n)$ em $\mathbb{R}^n$. Nos queremos aproximar esse vetor por um outro da forma $\tilde{\boldsymbol{y}} = (\hat{\alpha} + \hat{\beta}x_1, \ldots, \hat{\alpha} + \hat{\beta}x_n)$ com $\hat{\alpha}$ e $\hat{\beta}$ escolhidos convenientemente. Uma possibilidade conveniente pode ser escolher $\hat{\alpha}$ e $\hat{\beta}$ de forma a minimizar a distância euclidiana entre $\boldsymbol{y}$ e $\tilde{\boldsymbol{y}}$ ou, o que é a mesma coisa, o quadrado da distância

$$d^2(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = h(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2\,.$$

Derivando com respeito a $\hat{\alpha}$ e a $\hat{\beta}$ e igualando a zero, obtemos as assim chamadas *equações normais*

$$-\frac{1}{2}\frac{\partial h}{\partial \hat{\alpha}} = \sum_{i=1}^{n} y_i - n\,\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = 0 \tag{3}$$

$$-\frac{1}{2}\frac{\partial h}{\partial \hat{\beta}} = \sum_{i=1}^{n} x_i\,y_i - \hat{\alpha}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0\,. \tag{4}$$

Da equação (3) obtemos que

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \tag{5}$$

e substituindo em (4) obtemos

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,. \tag{6}$$

Estes estimadores são chamados *Estimadores de Mínimos Quadrados* e foram usados para ajustar as retas de regressão na Figura 1.

Falta achar um estimador para σ^2. Veja que se α e β fossem conhecidos, poderíamos usar o estimador $n^{-1}\sum_{i=1}^{n}\epsilon_i^2 = n^{-1}\sum_{i=1}^{n}(y_i-\alpha-\beta\,x_i)^2$. Porém, como os *erros* $\epsilon_i = y_i-\alpha-\beta\,x_i$ não são conhecidos, podemos substituir pelos *resíduos* $e_i = y_i-\hat{\alpha}-\hat{\beta}\,x_i$. Nesse caso, pode-se mostrar que o denominador correto para obter um estimador não-viesado de σ^2 é $(n-2)$, de forma que

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}\,x_i)^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 - \hat{\beta}\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})\right]. \tag{7}$$

Exemplo 1. (Continuação). Temos que $\sum_i y_i^2 = 6511425$, $\sum_i y_i = 8955$, $\sum_{i=1}^{2}(y_i-\bar{y})^2 = 651142-(8955)^2/24 = 3170091$; $\sum_i x_i^2 = 29.51779$, $\sum_i x_i = 21.873$, $\sum_i(x_i-\bar{x})^2 = 29.51779-(21.873)^2/24 = 9.583$; $\sum_i x_i\,y_i = 12513.69$, $\sum_i(x_i-\bar{x})\,(y_i-\bar{y}) = 12513.69-(8955)(21.873)/24 = 4352.332$. Assim, os estimadores de mínimos quadrados são $\hat{\beta} = 4352.332/9.58329 = 454.16$, $\hat{\alpha} = 373.125-(454.16)\,(0.911) = -40.78$ e $\hat{\sigma}^2 = (24-2)^{-1}\,[3170091-(454.16)\,(4352.332)] = (232.9)^2$. Alternativamente, se as observações estão guardadas sob os nomes `x` e `y` numa sessão do **R**, basta executar o código "`summary(lm(y~1+x))`" para obter esses resultados.

3 Propriedades dos estimadores e inferência

Assumindo somente (2) é possível mostrar que os estimadores (6) e (5) são:

- funções lineares das observações $y_1,\ldots,y_n$;

- não-viesados. Por exemplo, como $\mathrm{E}(y_i) = \alpha + \beta\, x_i + \mathrm{E}(\epsilon_i) = \alpha + \beta\, x_i$ e, daí, $\mathrm{E}(\bar{y}) = \alpha + \beta\, \bar{x} + \mathrm{E}(\bar{\epsilon}) = \alpha + \beta\, \bar{x}$,

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left\{\frac{\sum_{i=1}^n (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\} = \frac{\sum_{i=1}^n (x_i - \bar{x})\ \mathrm{E}(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})\ \mathrm{E}(\alpha + \beta\, x_i - \alpha - \beta\, \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta\, \frac{\sum_{i=1}^n (x_i - \bar{x})\,(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta\,.\end{aligned}$$

  De forma semelhante,

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}(\bar{y} - \hat{\beta}\, \bar{x}) = \mathrm{E}(\bar{y}) - \bar{x}\ \mathrm{E}(\hat{\beta}) = \alpha + \beta\, \bar{x} - \beta\, \bar{x} = \alpha\,.$$

- Dados quaisquer outros estimadores $\tilde{\alpha}$ e $\tilde{\beta}$ que sejam também lineares e não-viesados, segue necessariamente que que $\mathrm{Var}(\tilde{\alpha}) \geq \mathrm{Var}(\hat{\alpha})$ e $\mathrm{Var}(\tilde{\beta}) \geq \mathrm{Var}(\hat{\beta})$. Esse resultado é conhecido como *Teorema de Gauss-Markov*. Fala-se que os estimadores (6) e (5) são BLUE, pelo acrônimo em inglês de **B**est **L**inear **U**nbiased **E**stimators.

Para se fazer inferências tais como intervalos de confiança ou testes de hipóteses sobre α e/ou β, é necessário achar as distribuições amostrais de $\hat{\alpha}$ e $\hat{\beta}$, e para isso é necessário especificar a distribuição dos erros ϵ_i. Usualmente assume-se que eles são normalmente distribuídos e, nesse caso, é possível mostrar que:

- Como $\hat{\beta}$ e $\hat{\alpha}$ são combinações lineares de $Y_1, \ldots, Y_n$ que são normalmente distribuídas, então $\hat{\beta}$ e $\hat{\eta}$ também são normalmente distribuídas;

- Vimos acima que $\mathrm{E}(\hat{\beta}) = \beta$ e $\mathrm{E}(\hat{\alpha}) = \alpha$;

- Para caracterizar a distribuição dos estimadores, só falta calcular a variância deles. É possível mostrar que

$$\mathrm{Var}(\hat{\beta}) = \sigma^2\, \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{8}$$

  e que

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\, \frac{\sum_{i=1}^n x_i^2}{n\, \sum_{i=1}^n (x_i - \bar{x})^2}\,; \tag{9}$$

- O estimador $\hat{\sigma}^2$ é independente tanto de $\hat{\beta}$ quanto de $\hat{\alpha}$ e

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}\,.$$

- Portanto, usando o fato que a distribuição t_m de Student com m graus de liberdade pode ser obtida como a ração entre uma variável aleatória Normal(0,1) e a raíz quadrada de uma χ^2_m dividida pelos graus de liberdade m, segue que

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}}\, \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sim t_{n-2} \tag{10}$$

e que

$$\frac{\hat{\alpha}-\alpha}{\hat{\sigma}}\sqrt{\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}x_i^2}}\sim t_{n-2} \tag{11}$$

- Esses dois resultados podem ser usados para construir intervalos de confiança para α e β e para se fazer testes sobre eles. Por exemplo, um IC de nível $100(1-a)\%$ para β é

$$\hat{\beta}\pm t_{(n-2);a/2}\,\hat{\sigma}\,\frac{1}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

e o correspondente intervalo para α é

$$\hat{\alpha}\pm t_{(n-2);a/2}\,\hat{\sigma}\,\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}}\,.$$

- Para testar. por exemplo, a H_0 que $\beta=\beta_0$ contra a alternativa que $\beta\neq\beta_0$ ao nível de significância $100a\%$, rejeitamos a H_0 se

$$|t_{obs}|=\left|\frac{\hat{\beta}-\beta_0}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|>t_{(n-2);a/2}\,.$$

O p-valor desse teste será $2\,P(T_{n-2}>|t_{obs}|)$.

Exemplo 1. (Continuação). Suponha que queremos testar se $\alpha=0$ ao nível de significância de 10%. Nesse caso calculamos

$$t_{obs}=\frac{\hat{\alpha}}{\hat{\sigma}}\sqrt{\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}x_i^2}}\doteq\frac{-40.78}{232.9}\sqrt{\frac{(24)\,9.583}{29.518}}\doteq-0.489$$

O valor crítico é $t_{22;0.05}\doteq 1.717$ e portanto não rejeitamos a H_0 que $\alpha=0$. O p-valor desse teste é $2\,P(T_{22}>0.489)=0.629$. Um IC com confiança 90% para α é $-40.78\pm(1.717)\,(232.9)\sqrt{\frac{29.518}{(24)\,9.583}}\doteq(-184.05;102.49)$ De forma análoga, um IC com confiança 90% para β é $454.16\pm(1.717)\,(232.9)/\sqrt{9.583}\doteq(324.96;583.36)$.

4 A tabela ANOVA e o coeficiente R^2

Quando estudamos análise da variância a um fator, vimos uma decomposição da variabilidade da resposta y_{ij} que basicamente era da forma

$$SQ_{tot}=\sum_{ij}(y_{ij}-\bar{y}_{\cdot\cdot})^2=\sum_{ij}(y_{ij}-\hat{y}_{ij})^2+\sum_{ij}(\hat{y}_{ij}-\bar{y}_{\cdot\cdot}))^2=SQ_{res}+SQ_{reg}\,,$$

onde $\hat{y}_{ij}=\hat{\mu}_i=\bar{y}_{i\cdot}$. Comparando a SQ_{reg} com a SQ_{res} (ou, mais precisamente, os respectivos quadrados médios após dividir pelos respectivos graus de liberdade), podíamos testar a H_0 que todas as médias dos tratamentos eram iguais.

Fonte de variação	**gl**	**Soma de quadrados**	**Quadrados médios**	F_{obs}
Regressão	1	$SQ_{reg} = \hat{\beta}^2 \sum_i (x_i - \bar{x})^2$	$QM_{reg} = \frac{SQ_{reg}}{1}$	$\frac{QM_{reg}}{QM_{res}}$
Residuos	$n-2$	$SQ_{res} = \sum_i (y_i - \bar{y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2$	$\hat{\sigma}^2 = QM_{res} = \frac{SQ_{res}}{n-2}$	
Total	$n-1$	$SQ_{tot} = \sum_i (y_i - \bar{y})^2$		

Tabela 1: Tabela de Análise da Variância para testar a hipótese $\beta = 0$.

Existe uma decomposição semelhante no caso da regressão linear simples que pode ser usada para testar se $\beta = 0$. Defina neste caso $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ e $e_i = y_i - \hat{y}_i$ e note das equações (3) e (4) que $\hat{y}_i - \bar{y} = \hat{\beta}\,(x_i - \bar{x})$. Logo, a identidade fundamental é

$$
\begin{aligned}
SQ_{tot} = \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\
&= \sum_i e_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i e_i^2 + \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 = SQ_{res} + SQ_{reg}\,.
\end{aligned}
$$

Semelhante à discussão feita na unidade de Análise da Variância, sob a hipótese de normalidade dos erros é possível mostrar que

- SQ_{res}/σ^2 segue uma distribuição χ^2 com $(n-2)$ gl;
- SQ_{reg}/σ^2 segue, sob a H_0 que $\beta = 0$, uma distribuição χ^2 com 1 gl e
- Sob H_0, SQ_{res} e SQ_{reg} são independentes.

Dessa forma, sabemos da definição da distribuição F de Snedecor que, sob H_0,

$$
F = \frac{QM_{reg}}{QM_{res}} = \frac{\frac{1}{\sigma^2}\, SQ_{reg}/1}{\frac{1}{\sigma^2}\, SQ_{res}/(n-2)} = (n-2)\,\frac{SQ_{reg}}{SQ_{res}}
$$

é um quociente entre duas variáveis aleatórias independentes com distribuição χ^2 divididas pelos respectivos graus de liberdade e segue, portanto, uma distribuição $F_{1,(n-2)}$ e o nosso teste rejeita H_0 quando $F_{obs} > F_{1-\alpha;1,(n-2)}$.

Usualmente a informação anterior é apresentada na tabela ANOVA 1.

Note que o fato de rejeitar a H_0 que $\beta = 0$ não diz necessariamente qual fração da variabilidade de y é explicada pela introdução da covariável (ou preditor) x. Em outras palavras, poderia ser que $\beta \neq 0$ é relativamente pequeno, mas se a variância do erro σ^2 for também pequena, eventualmente a diferença entre β e 0 será capturada pelo

Fonte de variação	**gl**	**Soma de quadrados**	**Quadrados médios**	F_{obs}
Distância	1	1976602	1976602	36.4354
Residuos	22	1193489	54249.5	
Total	23	3170091		

Tabela 2: Tabela de Análise da Variância para o Exemplo 1.

procedimento e rejeitaremos que $\beta = 0$. Para medir a proporção da variabilidade total dos y_i ($SQ_{tot} = \sum_i (y_i - \bar{y})^2$) que é explicada pelo preditor x, usa-se o assim chamado *coeficiente de determinação*

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{res}}{SQ_{tot}} = \hat{\beta}^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2}$$

Exemplo 1. (Continuação). Neste caso temos que $SQ_{tot} = 3170091$, $\sum_i (x_i - \bar{x})^2 \doteq 9.583$ $\hat{\beta} \doteq 454.16$ e portanto $SQ_{reg} \doteq (454.16)^2\,(9.583) \doteq 1976602$. Logo, $SQ_{res} \doteq 3170091 - 1976602 \doteq 1193489$ e $R^2 = 0.62$, o que diz que 62% da variabilidade da `velocidade` é explicada introduzindo o preditor `distância`. Da tabela de análise da variância 2 segue que ao nível de significância de 1% rejeitamos a hipótese que $\beta = 0$ pois $F_{obs} \doteq 36.435 > F_{0.01;1,22} \doteq 7.945$. O p-valor do teste é $P(F_{1,22} > 36.435) \doteq 4 \times 10^{-6}$.

Exercício 1. Temos agora dois procedimentos para testar a $H_0 : \beta = 0$ contra a alternativa $H_a : \beta \neq 0$, um baseado no estatístico t_{obs} apresentado na Seção 3 e outro baseado no estatístico F_{obs} da tabela ANOVA. Mostre que esses dois testes são equivalentes, isto é, que um rejeita se e somente se o outro também rejeita (*dica*: verifique que $F_{obs} = t_{obs}^2$).

5 Previsão de valores "futuros"

Existem dois problemas parecidos quando queremos saber o que acontecerá com a variável resposta correspondente a valores não observados na amostra. Por um lado, podemos estar interessados em estimar a média μ_c dos valores da resposta y para todos os indivíduos da população que têm um valor $x = x_c$ dado do preditor x. Por outro, poderíamos estar interessados em "estimar" o valor do y_c para um **único** indivíduo da população que tem $x = x_c$ (as aspas em volta de "estimar" foram usadas aquí pois y_c não é um parâmetro fixo, mas uma variável aleatória, de forma que tecnicamente fala-se de *previsão* ao invés de *estimação*; independente disso, os dois problemas são semelhantes).

Abordamos primeiro o problema de estimar a média μ_c. De acordo ao nosso modelo (1), $\mu_c = \alpha + \beta\, x_c$. Como temos estimadores não-viesados para α e β dados nas equações

(3) e (4), segue que o estimador

$$\hat{\mu}_c = \hat{\alpha} + \hat{\beta}\, x_c$$

será também não-viesado para estimar μ_c.

Para construir um IC ou fazer testes de hipóteses sobre μ_c precisamos obter a distribuição do estimador $\hat{\mu}$, o que usualmente é feito sob o suposto de normalidade dos erros. Nesse caso, como $\hat{\mu}_c$ é uma combinação linear dos y_i, segue que $\hat{\mu}_c$ é normalmente distribuído com média μ_c e variância

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_c) &= \mathrm{Var}(\hat{\alpha} + \hat{\beta}\, x_c) = \mathrm{Var}(\bar{y} - \hat{\beta}\,\bar{x} + \hat{\beta}\, x_c) = \mathrm{Var}[\hat{\beta}\,(x_c - \bar{x}) + \bar{y}] \\ &= (x_c - \bar{x})^2\ \mathrm{Var}(\hat{\beta})^2 + \mathrm{Var}(\bar{y}) + 2\,(x_c - \bar{x})\ \mathrm{cov}(\hat{\beta}, \bar{y}) = \sigma^2\left\{\frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}\right\}, \quad (12)\end{aligned}$$

onde temos usado a equação (8) e os fatos que (i) $\mathrm{Var}(\bar{y}) = \sigma^2/n$ e (ii) $\mathrm{cov}(\hat{\beta}, \bar{y}) = 0$ (verifique!) Logo,

$$t = \frac{\frac{\hat{\mu}_c - \mu_c}{\sqrt{\mathrm{Var}(\hat{\mu}_c)}}}{\sqrt{\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{\hat{\mu}_c - \mu_c}{\hat{\sigma}\,\sqrt{\frac{(x_c-\bar{x})^2}{\sum_i (x_i-\bar{x})^2} + \frac{1}{n}}} \sim t_{n-2}\,. \quad (13)$$

Assim, por exemplo, um IC com confiança $100(1-\alpha)\%$ para μ_c será

$$\hat{\mu}_c \pm t_{\alpha/2;n-2}\,\hat{\sigma}\,\sqrt{\frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}}\,. \quad (14)$$

Um problema diferente é "prever" um único valor da resposta y_c correspondente a um indivíduo com $x = x_c$. A semelhança com o caso anterior é que, se for desejada uma previsão *pontual*, como $\mathrm{E}(y_c) = \alpha + \beta\, x_c = \mu_c$ e também $\mathrm{E}(\hat{\alpha} + \hat{\beta}\, x_c) = \hat{\mu}_c$, a nossa previsão "não-viesada" deveria ser $\hat{\mu}_c$, o mesmo estimador que usamos acima para μ_c. Em outras palavras, a previsão *pontual* de y_c é igual ao estimador *pontual* de μ_c. Porém, as coisas são diferentes quando queremos calcular por exemplo um intervalo de previsão para y_c. Veja que, intuitivamente, deveríamos ter consideravelmente mais incerteza ao prever um único valor de y_c que ao estimar a média de todos os indivíduos com $x = x_c$. Como veremos a seguir, esse é efetivamente o caso.

Para construir um intervalo de previsão para y_c, considere a diferença

$$e_c = y_c - \hat{\mu}_c = y_c - \hat{\alpha} - \hat{\beta}\, x_c\,.$$

É claro que $\mathrm{E}(e_c) = 0$ e, como o nosso modelo postula que y_c é independente de $y_1, \ldots, y_n$ e, portanto, também de $\hat{\alpha}$ e de $\hat{\beta}$, segue que

$$\mathrm{Var}(e_c) = \mathrm{Var}(y_c) + \mathrm{Var}(\hat{\mu}_c) = \sigma^2 + \sigma^2\left\{\frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}\right\} = \sigma^2\left\{1 + \frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}\right\} \quad (15)$$

(compare as equações (12) e (15): o termo adicional 1 em (15) fará que efetivamente a variância (15) seja maior do que a (12)). Logo, um argumento semelhante ao que usamos para chegar às equações (10) ou (11) ou ainda (13) mostra que

$$t = \frac{\frac{\hat{\mu}_c - y_c}{\sqrt{\mathrm{Var}(e_c)}}}{\sqrt{\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{\hat{\mu}_c - y_c}{\hat{\sigma}\,\sqrt{1 + \frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}}} \sim t_{n-2}\,. \tag{16}$$

Assim, um intervalo de previsão com confiança $100(1-\alpha)\%$ para y_c será

$$\hat{\mu}_c \pm t_{\alpha/2;n-2}\,\hat{\sigma}\,\sqrt{1 + \frac{(x_c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{1}{n}}\,. \tag{17}$$

[novamente, compare as equações (14) e (17)].

Exemplo 1. (Continuação). Suponha que estamos interessados em saber qual é a velocidade média μ_c de nebulosas que estão a 1.8mpc do sistema solar. A estimativa pontual para μ_c será $\hat{\mu}_c \doteq -40.78 + (454.16)\,(1.8) \doteq 776.71$km/s. De acordo à equação (14), um intervalo com confiança 90% para μ_c

$$776.71 \pm (1.717)\,(232.9)\,\sqrt{\frac{(1.8 - 0.911)^2}{9.583} + \frac{1}{24}} \doteq (276.16; 1277.26)km/s\,.$$

Já se o objetivo fosse prever a velocidade de uma única nebulosa situada a 1.8mpc do sistema solar, a previsão pontual sería a mesma ($\mu_c \doteq 776.71$km/s), mas o intervalo de previsão calculado de acordo à equação (17) sería

$$776.71 \pm (1.717)\,(232.9)\,\sqrt{1 + \frac{(1.8 - 0.911)^2}{9.583} + \frac{1}{24}} \doteq (136.01; 1417.40)km/s\,,$$

consideravelmente maior do que o IC para μ_c.

6 Diagnóstico do modelo

Existem basicamente dois supostos que foram feitos nas Seções anteriores. Primeiro, assumimos que todas as observações seguem uma distribuição Normal. Segundo, assumimos que as variâncias dos y_i (ou dos ϵ_i) são todas iguais. Nas aplicações, usualmente queremos checar se os dados suportam esses supostos.

O primeiro suposto ($\mathrm{Var}(y_i) = \sigma^2$) é chamado de homocedasticidade. Um diagnóstico visual pode ser feito olhando ao gráfico dos resíduos $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}\,x_i$ contra os valores ajustados $\hat{y}_i$. Usualmente, procuramos verificar visualmente se existe algum padrão que poderia implicar que a variância poderia estar variando com o valor esperado da resposta y_i. Por exemplo, é comum que quanto maior é o valor esperado de y_i ($= \alpha + \beta\,x_i$), maior é a variância de y_i. Nesse caso, o gráfico deveria mostrar uma dispersão maior dos e_is a medida que os $\hat{y}_i$s aumentam.

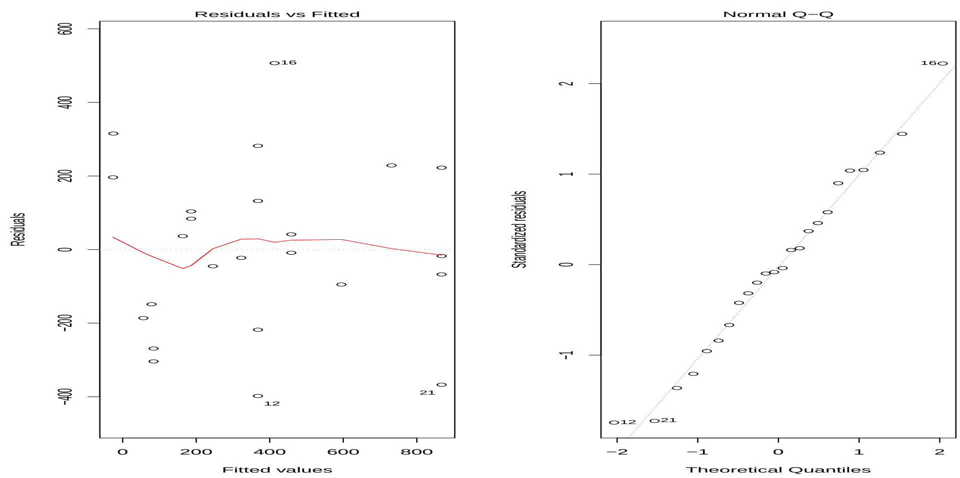

Figura 2: Gráficos para diagnóstico do modelo no Exemplo 1: (a) Resíduos versus valores ajustados e (b) gráfico q-q para os resíduos do modelo.

O suposto de normalidade pode ser checado visualmente de várias formas. A mais comum é usando o assim chamado gráfico Q-Q dos resíduos. Se os erros fossem normalmente distribuídos, esse gráfico deveria mostrar uma forte tendência linear.

Exemplo 1. (Continuação). A Figura 2 mostra esses dois gráficos construídos com a linguagem **R**. O primeiro gráfico (resíduos vs. valores ajustados) não parece mostrar nenhum padrão de variabilidade que fizesse duvidar o suposto de homocedasticidade. O segundo gráfico (q-q dos resíduos) mostra tendência linear, embora existem algumas observações com valores pequenos de $\hat{y}_i$ que poderiam fugir dessa tendência. Em geral, o diagnóstico visual sugere que os dois supostos podem ser razoáveis para esses dados.