8076 - Modelos Mistos

Vanderly Janeiro

Dep. de Estatística - UEM

Conteúdo


```
## [1] FALSE
```

Tradução cap. 01:

Galwey, Nicholas W. (2014). Introduction to mixed modelling : beyond regression and analysis of variance. 2a Ed.

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

1 A necessidade de mais de um termo de efeito aleatório ao ajustar uma linha de regressão

1.1 Um conjunto de dados com várias observações da variável Y em cada valor da variável X

Um dos usos mais comuns e mais simples da análise estatística é o ajuste de uma linha reta, conhecida por razões históricas como linha de regressão, para descrever a relação entre uma variável explicativa, X e uma variável de resposta, Y. O afastamento dos valores de Y em relação a essa linha são chamados de variação residual e são considerados aleatórios. É natural perguntar se a parte da variação em Y que é explicada pela relação com X é mais do que se poderia razoavelmente esperar por acaso: ou mais formalmente, se é significativa em relação à variação residual. Esta é uma análise de regressão simples e, para muitos conjuntos de dados, é tudo o que é necessário. No entanto, em alguns casos, várias observações de Y são tomadas em cada valor de X. Os dados então formam grupos naturais, e pode não ser mais apropriado analisá-los como se cada observação fosse independente: observações de Y no mesmo valor de X pode estar a uma distância semelhante da linha. Podemos então ser capazes de reconhecer duas fontes de variação aleatória, a saber,

- variação entre grupos
- variação entre observações dentro de cada grupo.

Esta é uma das situações mais simples em que é necessário considerar a possibilidade de haver mais de um único estrato de variação aleatória – ou, na linguagem da modelagem mista, que um modelo com mais de um termo de efeito aleatório possa ser necessário. Neste capítulo, examinaremos um conjunto de dados desse tipo e exploraremos como a análise de regressão usual é modificada pelo fato de os dados formarem grupos naturais.

Exploraremos essa questão em um conjunto de dados que relaciona os preços das casas na Inglaterra à sua latitude. Não há dúvida de que as casas custam mais no sul da Inglaterra do que no norte: esses dados não levarão a novas conclusões, mas ilustrarão essa tendência e os métodos usados para explorá-la. Os dados são exibidos em uma planilha na saída do R a seguir. As variáveis 'latitude' e 'price_pounds' são variações – listas de observações que podem assumir qualquer valor numérico. No entanto, as observações da variável 'cidade' podem ter apenas alguns valores permitidos - neste caso, os nomes das 11 cidades em consideração. A variável 'cidade' é um fator. As cidades são grupos de observações: dentro de cada cidade, todas as casas estão quase na mesma latitude, e a latitude da cidade é representada por um único valor neste conjunto de dados. Em contraste, o preço de cada casa é potencialmente único.

Antes de iniciar uma análise formal desse conjunto de dados, devemos observar suas limitações. Uma investigação completa da relação entre latitude e preço da casa levaria em conta muitos fatores além daqueles registrados aqui:

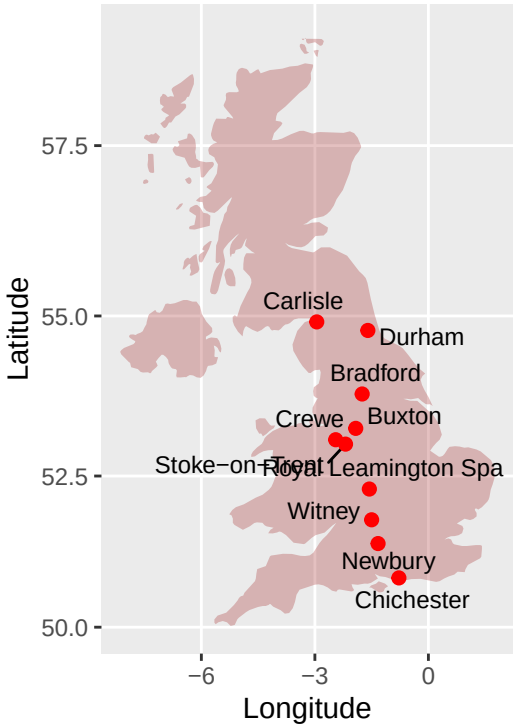


Figura 1: Cidades da Inglaterra pesquisadas

- o número de quartos em cada casa;
- seu estado de conservação e melhoria;
- outros indicadores observáveis de sua localização; e assim por diante.

Até certo ponto, tais fontes de variação foram eliminadas da presente amostra pela escolha de casas que são amplamente semelhantes: são todas casas "comuns" (não apartamentos, duplex, etc.) e todas têm três, quatro ou cinco quartos. As demais fontes de variação de preço contribuirão para a variação residual entre as casas em cada cidade, e serão tratadas de acordo. Devemos também considerar em que sentido podemos pensar na latitude das casas como "explicando" a variação de seus preços. A variável facilmente mensurável 'latitude' está associada a muitas outras variáveis, como clima e distância de Londres, e provavelmente são algumas delas, e não a latitude em si, que têm uma conexão causal com o preço. No entanto, uma variável explicativa não precisa estar causalmente ligada à resposta para servir como um preditor útil. Finalmente, devemos considerar o valor de estudar a relação entre latitude e preço em uma amostra tão pequena. Os dados sobre este tema são extensos e amplamente interpretados. No entanto, este pequeno conjunto de dados, ilustrando um exemplo simples e familiar, é altamente adequado para um estudo dos métodos pelos quais julgamos a significância de uma tendência, estimamos sua magnitude e colocamos limites de confiança nas estimativas. Acima de tudo, este exemplo mostrará que, para fazer essas coisas de maneira confiável, devemos reconhecer que nossas observações – as casas – não são mutuamente independentes, mas formam grupos naturais – as cidades.

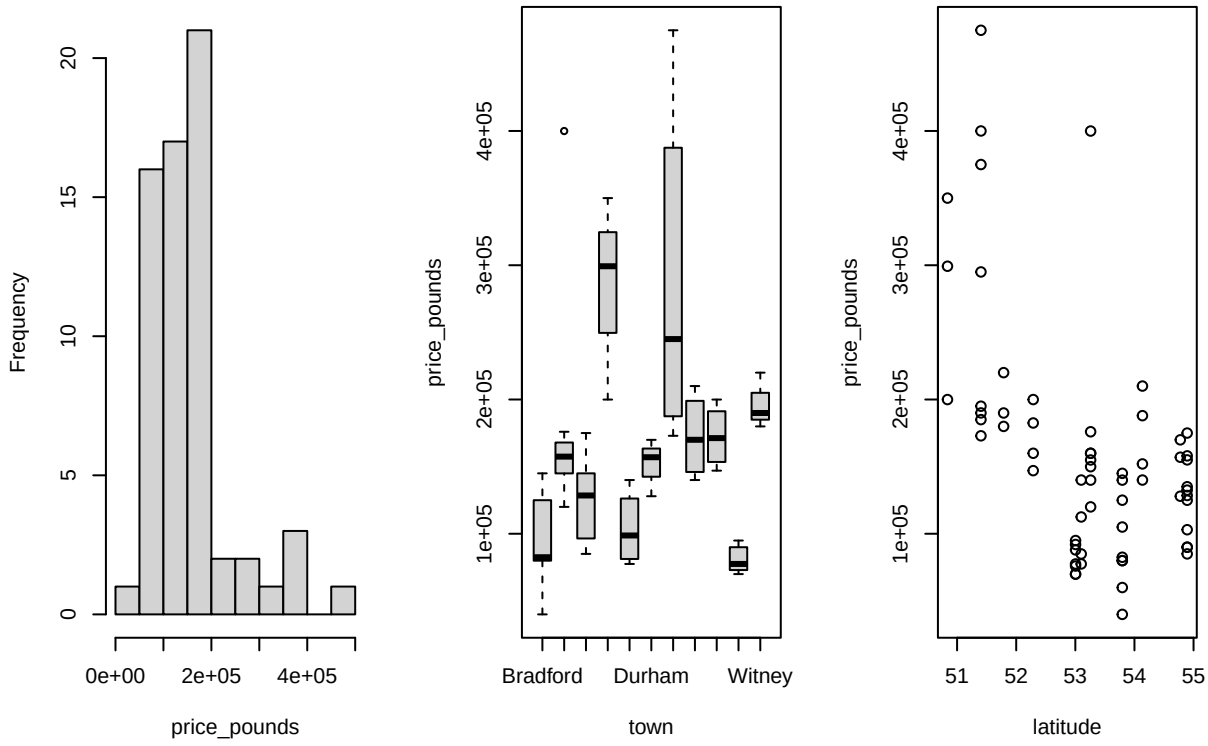
Leitura dos dados:

```
load("dadosGalwey/Galwey-tab1-1.RData")
psych::headTail(dat, top = 5, bottom = 5)
```

```
##      id           town latitude price_pounds    y
## 1     2       Bradford    53.79        39950  4.6
## 2     3       Bradford    53.79        59950 4.78
## 3     4       Bradford    53.79        79950  4.9
## 4     5       Bradford    53.79        79995  4.9
## 5     6       Bradford    53.79        82500 4.92
## ... ...           <NA>      ...          ...  ...
## 60   61 Stoke-On-Trent       53        92000 4.96
## 61   62 Stoke-On-Trent       53        94950 4.98
## 62   63         Witney    51.79       179950 5.26
## 63   64         Witney    51.79       189950 5.28
## 64   65         Witney    51.79       220000 5.34
```

```
attach(dat)

par(mfrow=c(1,3))
hist(price_pounds, main="")
boxplot(price_pounds~town)
plot( latitude,price_pounds )
```




```
layout(1)
```

1.2 Análise de regressão simples: Uso do software

Começaremos fazendo uma análise de regressão simples sobre esses dados, antes de considerar como isso deve ser modificado para levar em conta o agrupamento em cidades. O modelo de regressão linear padrão (Modelo 1) é

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \tag{1}$$

sendo:

- x_i : valor de X (latitude) para a i-ésima cidade,
- y_{ij}: valor observado de Y, preço da casa em libras, para a j-ésima casa na i-ésima cidade,
- β_0, β_1: constantes a serem estimadas, definindo a relação entre X e Y e
- ε_{ij}: o efeito residual, ou seja, o desvio de y_{ij} do valor previsto com base em x_i, β_0 e β_1.

Ajuste do modelo:

```r
fm1<- lm(price_pounds ~ latitude , data=dat)
summary(fm1)
```

```
##
## Call:
## lm(formula = price_pounds ~ latitude, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98579 -51488  -7715  32974 245443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2189739     399243   5.485 8.10e-07 ***
## latitude      -38133       7499  -5.085 3.64e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74160 on 62 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.283
## F-statistic: 25.86 on 1 and 62 DF,  p-value: 3.639e-06
```

Como a análise dos resíduos apresenta não conformidade com os pressupostos do modelo, tentaremos uma transformação por Box-Cox:

O gráfico do perfíl do logaritmo da função de verossimilhança e o valor de máximo obtido ($\hat{\lambda} = 0.0202$) sugerem uma transformação logarítmica para a variável resposta:

```r
dat$y<- log10(price_pounds);  attach(dat)
```

Assim a variável resposta passa a ser y o logaritmo na base 10 do preço da casa em libras. Ajustando o modelo novamente temos:

```r
fm2<- lm(y ~ latitude , data=dat)
```

```r
anova(fm2)
```

```
## Analysis of Variance Table
##
```

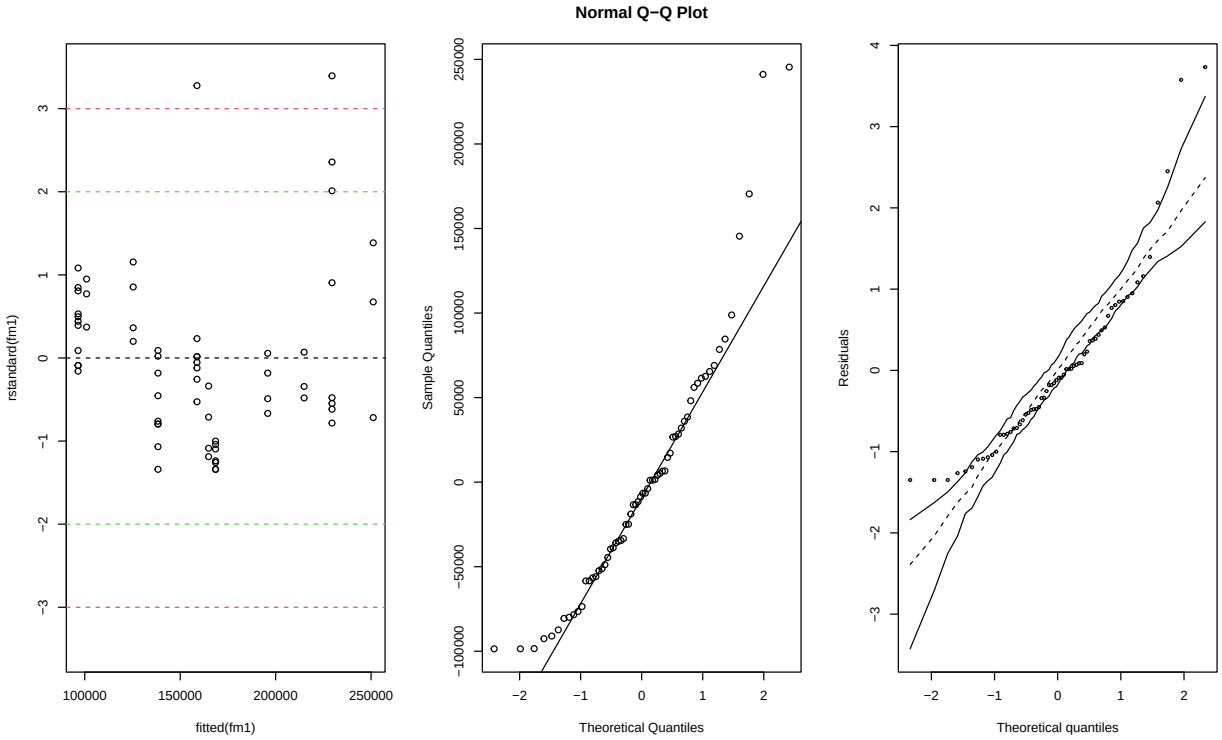


Figura 2: Resíduos do modelo (1) ajustado aos dados originais - `fm1`

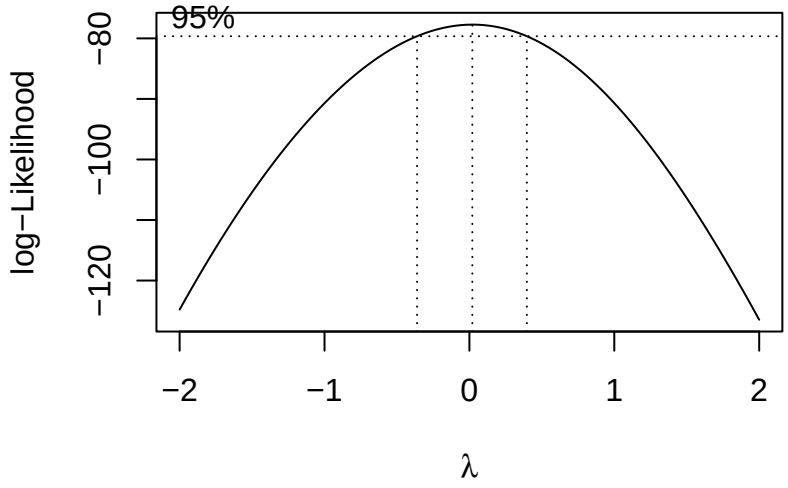


Figura 3: Perfíl do logaritmo da função de verossimilhança da transformação Box-Cox considerando o modelo (1)

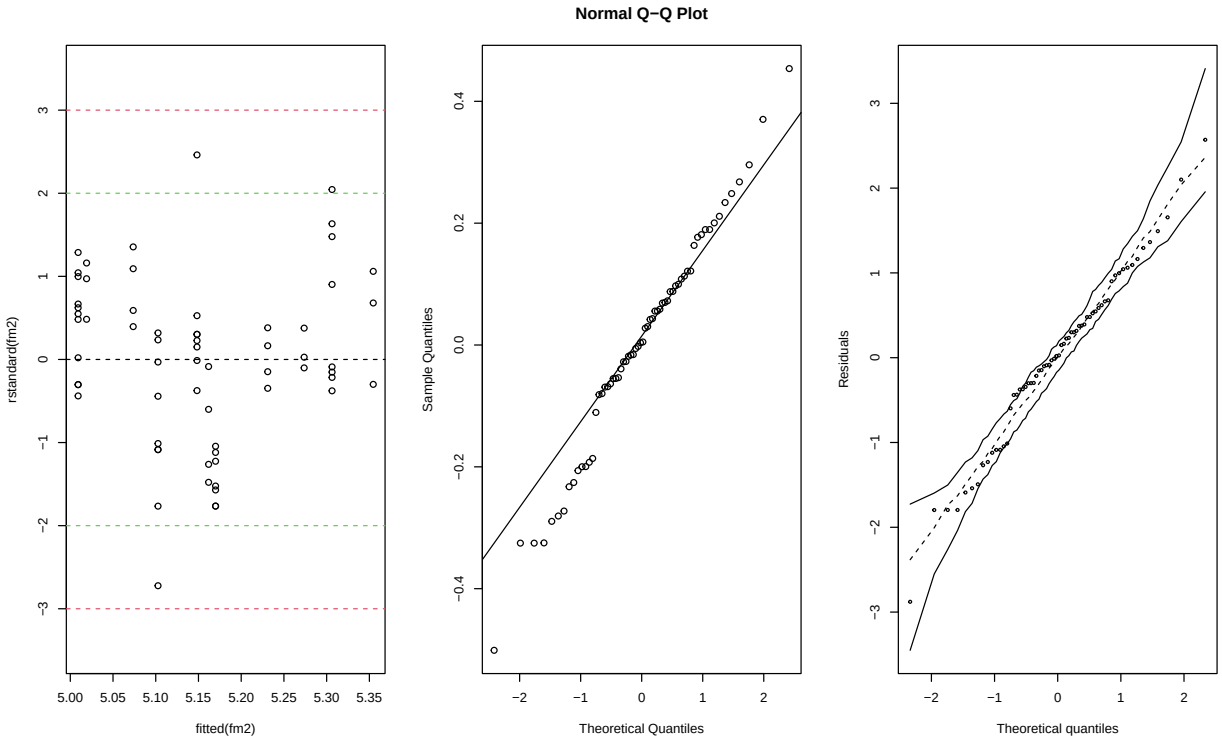


Figura 4: Resíduos do modelo (1) ajustado aos dados transformados - `fm2`

```
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## latitude   1 0.70955 0.70955  20.552 2.71e-05 ***
## Residuals 62 2.14056 0.03453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fm2)
```

```
##
## Call:
## lm(formula = y ~ latitude, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50115 -0.08019  0.00432  0.10942  0.45372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  9.68473    1.00036   9.681 5.14e-14 ***
## latitude    -0.08518    0.01879  -4.533 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1858 on 62 degrees of freedom
## Multiple R-squared:  0.249,  Adjusted R-squared:  0.2368
## F-statistic: 20.55 on 1 and 62 DF,  p-value: 2.71e-05
```

Observe que existem 64 casas na amostra, cada uma das quais dá um valor de ε_{ij}, então pode-se pensar que o termo residual incluiria 64 informações. No entanto, duas informações foram 'usadas' pela estimativa do intercepto e o efeito da latitude.

Na Figura 5, as observações da mesma cidade geralmente ficam do mesmo lado da linha – ou seja, os valores residuais dentro de cada cidade não são mutuamente independentes. Por exemplo, conforme observado as observações de Ripon, Durham e Carlisle geralmente ficam acima da linha, enquanto os de Stoke-On-Trent e Crewe situam-se abaixo da linha.

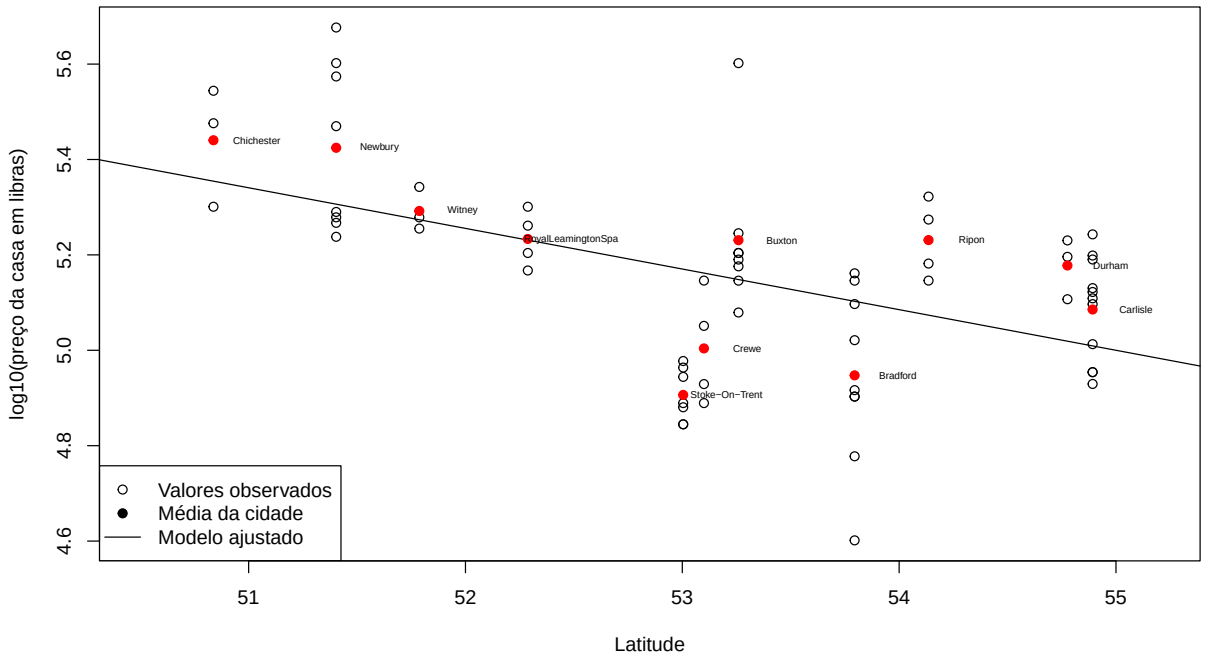


Figura 5: Disperção dos valores observados transformados (log10) e ajustado considerando o modelo (1)

1.3 Análise de regressão nas médias dos grupos

Como os valores residuais não são mutuamente independentes, a análise apresentada acima não pode ser confiável, mesmo que a linha de regressão pareça razoável. Em particular, os graus de liberdade residuais (DFResidual =62) são uma superestimativa: nos enganamos se acreditarmos que os dados compreendem 64 observações independentes. A maneira mais simples de superar esse problema é ajustar o modelo de regressão usando o valor médio de log(preço da casa) para cada cidade. Para isso realizaremos o seguinte código R:

```
library(dplyr)
grp <-group_by(dat, town)
dat.m <- summarise(grp, m.y=mean(y),latitude=mean(latitude), n=n())
dat.m
```

```
## # A tibble: 11 x 4
##    town                 m.y latitude     n
##    <chr>              <dbl>    <dbl> <int>
##  1 Bradford            4.95     53.8     9
##  2 Buxton              5.23     53.3     8
##  3 Carlisle            5.09     54.9    11
##  4 Chichester          5.44     50.8     3
##  5 Crewe               5.00     53.1     4
##  6 Durham              5.18     54.8     3
##  7 Newbury             5.42     51.4     8
##  8 Ripon               5.23     54.1     4
##  9 RoyalLeamingtonSpa  5.23     52.3     4
## 10 Stoke-On-Trent      4.91     53.0     7
## 11 Witney              5.29     51.8     3
```

Ajustando o modelo (1), as médias dos valores observados por cidade:

```
fm3<- lm(m.y ~ latitude , data=dat.m)
anova(fm3)
```

```
## Analysis of Variance Table
##
## Response: m.y
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## latitude   1 0.11601 0.116013   5.194 0.04864 *
## Residuals  9 0.20102 0.022336
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fm3)
```

```
##
## Call:
## lm(formula = m.y ~ latitude, data = dat.m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27468 -0.08741  0.05631  0.09985  0.14098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.44475    1.87211   5.045 0.000695 ***
## latitude    -0.08044    0.03530  -2.279 0.048639 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1495 on 9 degrees of freedom
## Multiple R-squared:  0.3659, Adjusted R-squared:  0.2955
## F-statistic: 5.194 on 1 and 9 DF,  p-value: 0.04864
```

Os DF residuais são muito menores (DFResidual = 9), refletindo a suposição mais razoável de que cada cidade pode ser considerada como uma observação independente. Consequentemente, a relação entre latitude e preço da casa é muito menos significativa: $p = 0,049$, comparado com $p < 0,001$ na análise anterior. No entanto, as estimativas do intercepto e inclinação da linha de regressão não são muito alteradas.

1.4 Um modelo de regressão com um termo para os grupos

A análise das médias das cidades, apresentada na seção anterior, não dá conta da variação dentro das cidades. Tão pouco leva em conta a variação no número de casas amostradas e, portanto, na precisão da média, de uma cidade para outra. Uma abordagem alternativa, que supera essas deficiências, é adicionar um termo ao modelo de regressão para levar em conta a variação entre as cidades que não é contabilizada pela latitude, de modo que o modelo se torna:

$$y_{ij} = \beta_0 + \beta_1 x_i + \tau_i + \varepsilon_{ij} \tag{2}$$

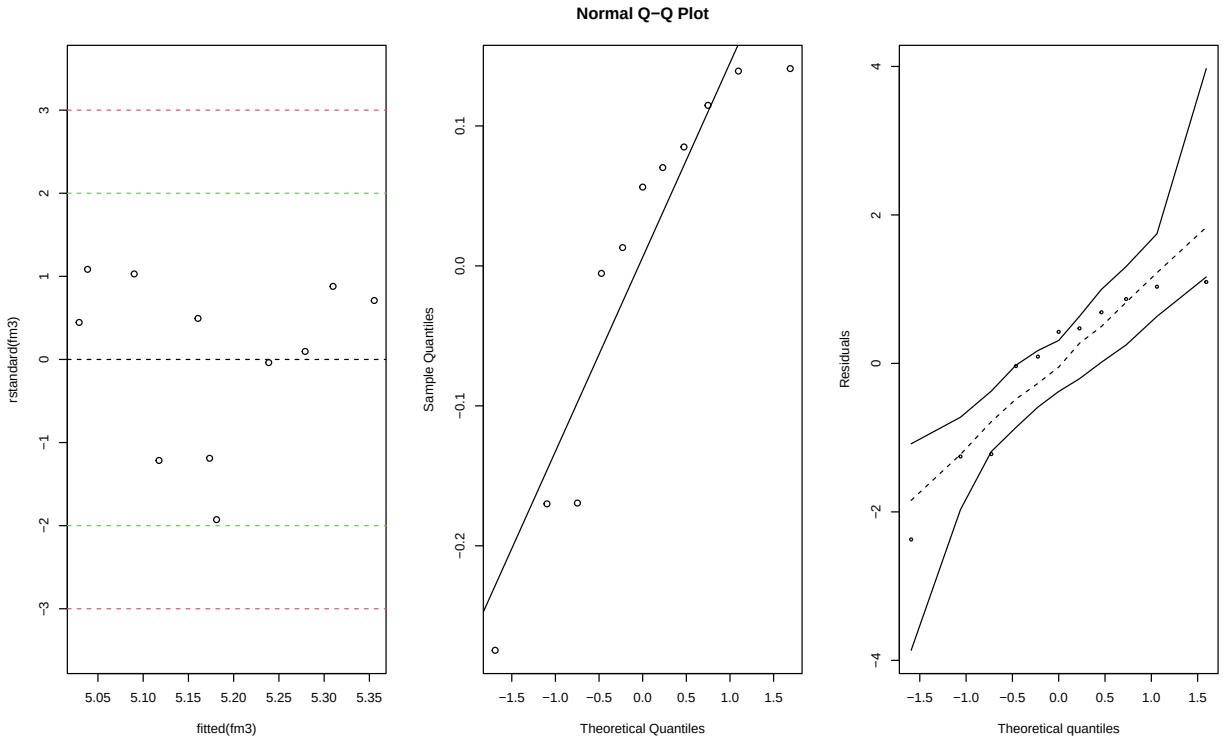


Figura 6: Resíduos do modelo (1) ajustado a médias dos dados transformados por cidade - `fm3`

sendo:

- τ_i : desvio médio da linha de regressão das observações da i-ésima cidade.

Ajustando o modelo (2), as médias dos valores observados por cidade:

```
fm4<- lm(y ~ latitude + town , data=dat)
anova(fm4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## latitude   1 0.70955 0.70955  41.354 3.710e-08 ***
## town       9 1.23119 0.13680   7.973 2.439e-07 ***
## Residuals 53 0.90937 0.01716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fm4)
```

```
##
## Call:
## lm(formula = y ~ latitude + town, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34594 -0.07121 -0.01527  0.05992  0.37117
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             14.18167    2.31841   6.117 1.18e-07 ***
## latitude                -0.17166    0.04350  -3.946 0.000235 ***
## townBuxton               0.19144    0.05981   3.201 0.002315 **
## townCarlisle             0.32648    0.08849   3.689 0.000531 ***
## townChichester          -0.01470    0.13621  -0.108 0.914461
## townCrewe               -0.06280    0.07609  -0.825 0.412873
## townDurham               0.39878    0.10636   3.749 0.000440 ***
## townNewbury              0.06668    0.10162   0.656 0.514585
## townRipon                0.34212    0.08404   4.071 0.000157 ***
## townRoyalLeamingtonSpa   0.02721    0.08736   0.312 0.756617
## townStoke-On-Trent      -0.17675    0.06355  -2.781 0.007483 **
## townWitney                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.131 on 53 degrees of freedom
## Multiple R-squared:  0.6809, Adjusted R-squared:  0.6207
## F-statistic: 11.31 on 10 and 53 DF,  p-value: 4.823e-10
```

Primeiro vem uma mensagem observando que, como a variação entre as médias das cidades é parcialmente explicada pela latitude, o termo 'cidade' não pode ser totalmente incluído no modelo de regressão. Diz-se que está parcialmente associado à latitude. (A consequência técnica deste aliasing parcial é que o efeito de uma das cidades – Witney, escolhida arbitrariamente porque vem por último quando as cidades estão organizadas em ordem alfabética – é uma função dos efeitos das outras cidades e da latitude, mas os detalhes numéricos dessa relação, dados na mensagem, não precisam nos preocupar.)

As estimativas do parâmetros levam corretamente à média amostral para cada cidade quando

substituídas na fórmula

$$mean(log(houseprice)) = 14.18 - 0.1717 \times \text{latitude} + \text{effect of town}$$

Por exemplo, para Durham,

$$mean(log(houseprice)) = 14.18 - 0.1717 \times 54.7762 + 0.399 = 5.1739.$$

No entanto, as próprias estimativas dos parâmetros são arbitrárias e não informativas, conforme ilustrado na Figura 1.5. A linha ajustada é especificada arbitrariamente para passar pelos valores médios de Bradford e Witney (a primeira e a última cidades na sequência alfabética), e os efeitos das outras cidades – suas distâncias verticais da linha ajustada – são determinados de acordo.

Apesar da natureza arbitrária das estimativas dos parâmetros, a anova é informativa. O termo 'latitude' representa a parte da variação do preço da habitação que se deve ao efeito da latitude, e o termo 'cidade' representa a parte devido à variação entre as cidades após ter em conta a latitude, ou seja, os desvios da cidade significa da linha original de melhor ajuste (não a linha ajustada arbitrária apresentada acima). Observe que o MS para latitude é o mesmo que o valor correspondente do Modelo (1):

$$SQ_{\text{latitude}} = SQ_{\text{Regressão, Modelos 1.1}} = 0.70955$$

Ou seja, a quantidade de variação explicada pela latitude é consistente entre os Modelos 1.1 e 1.3. Observe também que

$$SQ_{\text{Total Modelo 1.1}} = SQ_{\text{Total, Modelos 1.3}} = 0.70955 + (1.23119 + 0.90937) = 2.85011$$

De forma que, parte da variação antes inexplicável agora é atribuída aos efeitos das cidades.

1.5 Teste F apropriado para variável explicativa quando grupos estão presentes

A anova acumulada pode ser adaptada para fornecer uma avaliação realista do significado do efeito da latitude.

```
# Regressão baseada nas médias das cidade
anova(fm3)
```

```
## Analysis of Variance Table
##
## Response: m.y
```

```
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## latitude   1 0.11601 0.116013   5.194 0.04864 *
## Residuals  9 0.20102 0.022336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Regressão com o termo cidade
anova(fm4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## latitude   1 0.70955 0.70955  41.354 3.710e-08 ***
## town       9 1.23119 0.13680   7.973 2.439e-07 ***
## Residuals 53 0.90937 0.01716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabela 1: Comparação entre as estatísticas F das análises de regressão com base nas casas individuais e nas médias das cidades.

	Regressão com base nas médias da cidade	Regressão com termo para cidades(Modelo **??**)
Fórmula para $F_{latitude}$	$QM_{Reg}/QM_{Resíduo}$	$QM_{latitude}/QM_{town}$
valores	$0.11601/0.02234 = 5.19293$	$0.70955/0.13680 = 5.18677$
G.L.$_{Numerador}$	1	1
G.L.$_{Denomidor}$	9	9
valor-p	0.0487	0.0488

F_{town} é altamente significativo ($p < 0,001$), confirmando que há variação real entre os municípios além daquela contabilizada pela latitude. Consequentemente, $F_{latitude}$ é enganosa: é muito maior que o valor de 5,19, obtido quando a regressão é ajustada usando as médias da cidade. Para obter o valor apropriado da presente análise, precisamos calcular

$$F_{latitude} = \frac{QM_{latitude}}{QM_{town}} = \frac{0.70955}{0.13680} = 5.186769$$

Isso é quase exatamente equivalente à estatística F para latitude na análise baseada nas médias das cidades, conforme mostrado na Tabela 1. A razão pela qual os dois conjuntos de valores não concordam exatamente será explicada na Seção 1.9.

1.6 A decisão de especificar um termo do modelo como aleatório: Um modelo misto

Os valores de F apresentados na Tabela 1 são baseados na comparação da variação explicada pela linha de regressão com a variação das médias da cidade sobre a linha, enquanto o valor muito maior na anova acumulada ($F_{latitude} = 41,35$) é baseado na comparação com a variação dos valores individuais de log(preço da casa) sobre suas respectivas médias da cidade. Quando usamos a variação das médias das cidades como base de comparação, estamos considerando essas médias como valores de uma variável aleatória. Uma maneira de justificar isso é considerar as cidades deste estudo como uma *amostra representativa* de uma grande população de cidades. Formalmente falando, assumimos então que os valores τ_i no Modelo (2) são valores independentes de uma variável T, tal que

$$T \sim N(0, \sigma_T^2) \tag{3}$$

ou seja, eles são muito parecidos com os valores ε_{ij}, exceto que sua variância, σ_T^2, é diferente. Essa variação aleatória influencia nossa estimativa do efeito da latitude: se retirarmos uma cidade do estudo e substituí-la por uma cidade recém-escolhida, isso teria um efeito na inclinação da linha de regressão – um efeito maior do que seria produzido simplesmente tomando uma nova amostra de casas de dentro da mesma cidade. Isso ocorre porque há mais variação entre as médias das cidades, mesmo levando em conta o efeito da latitude, do que entre as casas individuais da mesma cidade: isso fica claro pelo valor altamente significativo de F_{town}, 7,97.

A suposição de que τ_i são valores de uma variável aleatória pode parecer radical, mas é necessária se quisermos fazer qualquer afirmação geral sobre a relação entre latitude e preços de casas: vimos anteriormente (Seção 1.4) que, se os efeitos da cidade forem especificados como fixo, a estimativa do efeito da latitude é arbitrária. Mas se nosso conjunto de cidades for uma amostra aleatória de uma população, a estimativa fornece uma previsão do que seria observado em outra cidade da mesma população em uma determinada latitude. **Na linguagem da modelagem mista, para obter uma estimativa significativa do efeito da latitude nos preços das casas, devemos especificar o efeito de cada cidade como um efeito aleatório e especificar a cidade como um termo de efeito aleatório no modelo de regressão**. O efeito da latitude em si é um efeito não aleatório ou fixo. Como nosso modelo contém efeitos de ambos os tipos, é um modelo misto. (A suposição de que a variável aleatória T tem uma distribuição normal não é um requisito, mas outras distribuições requerem métodos de modelagem mais avançados – veja o Capítulo 10.)

Ao decidir se um fator deve ser especificado como aleatório, é útil perguntar se os níveis estudados podem ser considerados uma amostra representativa de uma grande população de níveis. No presente caso, as cidades amostradas podem ser consideradas representativas da população das cidades inglesas? Se assim for, seremos capazes de prever os preços das casas em uma cidade não incluída na amostra a partir da latitude da cidade e da linha de regressão – reconhecendo que a precisão de

nossa previsão será limitada não apenas pela variação residual, mas também pela variação aleatória variação entre as cidades. Estaríamos dispostos a fazer tal previsão apenas se a cidade estivesse na Inglaterra e fosse uma cidade substancial e distinta – não, por exemplo, uma vila ou um subúrbio de uma aglomeração maior – isto é, se pertencesse à população da qual nossa amostra é representativa.

No entanto, mesmo que tal população não possa ser especificada, ainda pode ser válido especificar um fator como aleatório e fazer inferências sobre seus níveis coletivamente. Para determinar se esse é o caso, é útil perguntar se, se perdêssemos os valores da variável de resposta de um nível de fator, consideraríamos que vale a pena predizê-los a partir do(s) valor(es) do(s) termo(s) de efeito fixo) e o modelo ajustado. No presente caso, se perdêssemos os preços das casas de, digamos, Durham, consideraríamos valer a pena predizê-los, embora com precisão reduzida, a partir da latitude de Durham e da linha de regressão? Em caso afirmativo, os níveis de fator compreendem um conjunto intercambiável e o fator pode ser especificado como aleatório. Um conjunto de níveis amostrados aleatoriamente de uma população infinita são permutáveis por definição: isto é, a amostragem aleatória é uma suposição mais forte do que a permutabilidade, mas a permutabilidade é suficiente para permitir que um termo seja especificado como aleatório em um modelo misto. Se decidirmos que os preços das casas em cidades diferentes são tão desconexos que tal previsão não faria sentido, devemos especificar "cidade" como um termo de efeito fixo e retirar "latitude" do modelo. É claro que a linha de regressão terá algum valor preditivo porque se baseia parcialmente nos dados de Durham, mas isso é trapaça: ainda teríamos confiança em seu valor preditivo se nosso conjunto de cidades fosse muito grande, de modo que a contribuição de Durham tornou-se insignificante?

Formalmente, o conceito de conjunto permutável significa que, embora tenhamos algumas informações sobre as propriedades coletivas dos membros, não temos informações sobre seus valores individuais – como um baralho de cartas do qual se vê apenas as costas. No presente caso, nosso conhecimento das propriedades coletivas dos efeitos da cidade pode ser razoavelmente bem representado pela Distribuição (3) mesmo que eles não sejam uma amostra aleatória de uma população infinita: isto é, sabemos que eles estão centrados na linha de regressão (média(T)=0) com uma variância característica, σ_T^2, que estimaremos durante o processo de ajuste do modelo. Nossa ignorância de seus valores individuais significa que não temos ideia prévia de quais cidades ficarão acima da linha de regressão e quais abaixo. O conceito de permutabilidade será mais explorado nas Seções 2.6 e 4.1 no contexto de projetos experimentais aleatórios, e na Seção 5.6, no contexto de 'estimativas reduzidas' de efeitos aleatórios. Um relato mais rigoroso desse conceito é dado por Spiegelhalter, Abrams e Myles (2004, Seção 3.4). A questão de como determinar quais termos de modelo devem ser especificados como aleatórios será discutida mais detalhadamente, no contexto de uma ampla gama de modelos, na Seção 6.3.

1.7 Comparação dos testes em modelo misto com teste de falta de ajuste

A partição da variação em torno da linha de regressão em dois componentes, um devido aos desvios das médias das cidades em relação à linha e outro aos desvios das casas individuais em relação às

médias das cidades, é semelhante ao teste de falta de ajuste descrito por Draper e Smith (1998, Seção 2.1, pp. 49-53). De fato, os cálculos realizados para obter os MSs são idênticos nas duas análises. No entanto, há uma diferença importante nas ideias que os fundamentam e, consequentemente, nos testes F especificados, conforme ilustrado na Tabela 1.4. As linhas angulares nesta tabela indicam os pares de MSs que são comparados pelos dois testes F. O objetivo do teste de falta de ajuste é determinar se a variação entre grupos de observações no mesmo valor de X (cidades no presente caso) é significativa, ou se pode ser absorvida no termo residual. Na análise do modelo misto, por outro lado, a realidade da variação entre os municípios não está em dúvida. A questão é se o efeito da latitude é significativo ou se pode ser absorvido pelo termo "cidade".

Tabela 2: Comparação entre o teste F para falta de ajuste e o teste F de modelo misto na análise do efeito da latitude sobre os preços das casas na Inglaterra.

Fonte de variação	G.L.	QM	Teste de falta de ajuste		Teste de modelo misto	
			F	valor-p	F	valor-p
latitude	1	0.70955			5.19	0.0488
cidade	9	0.13680	7.97	<0.001		
Resíduo	53	0.01716				

Tabela 3: Comparação entre os testes F realizados na análise de regressão múltipla ordinária e na análise de modelo misto dos efeitos da latitude e da cidade sobre os preços das casas na Inglaterra.

Fonte de variação	G.L.	QM	Teste de falta de ajuste		Teste de modelo misto	
			F	valor-p	F	valor-p
latitude	1	0.70955	41.35	<0.001	5.19	0.0488
cidade	9	0.13680	7.97	<0.001	7.97	<0.001
Resíduo	53	0.01716				

Mesmo que o teste de falta de ajuste leve à conclusão de que a variação entre os municípios é significativa, as duas análises não são equivalentes. Como o teste de falta de ajuste trata apenas o termo residual como aleatório, ele leva ao uso desse termo como denominador em todos os testes F, enquanto o modelo misto leva ao uso do termo de efeito aleatório town como denominador contra para testar a significância do efeito da latitude. O conjunto completo de testes especificados pelas duas análises é, portanto, mostrado na Tabela 1.5. As linhas angulares pontilhadas nesta tabela indicam os pares de MSs que são comparados pelos testes F adicionais. No caso considerado por Draper e Smith, o termo adicional necessário (equivalente a 'cidade' no presente exemplo) é quadrático, para permitir a curvatura na resposta à variável explicativa, e nesta situação o teste de falta de ajuste está correto.

1.8 O uso de REsidual Maximum Likelihood (REML) para ajustar o modelo misto

Até agora, adotamos uma abordagem improvisada para ajustar o modelo misto.

- Ajustamos dois modelos de regressão, um sem termo para os efeitos de cidades e outro incluindo tal termo;
- tirou a estimativa da inclinação do primeiro modelo e os MSs do segundo;
- obteve a estatística F para testar a significância do efeito da latitude dos MSs manualmente.

No entanto, a análise de modelos mistos pode ser realizada de forma mais unificada utilizando o critério Máxima verossimilhança REsidual (REsidual Maximum Likelihood - REML, também conhecido como REstricted Maximum Likelihood). O significado formal deste critério será explicado no Capítulo 11: aqui, vamos simplesmente aplicá-lo aos dados presentes. A seguir

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
fm1.me <- lme(y ~ latitude ,
                   random = ~ 1|town,
                   method= "REML", data = dat)

summary(fm1.me)
```

```
## Linear mixed-effects model fit by REML
##   Data: dat
##         AIC       BIC   logLik
##   -42.12122 -33.61268 25.06061
##
## Random effects:
##  Formula: ~1 | town
##         (Intercept)  Residual
```

```
## StdDev:   0.1401131 0.1307764
##
## Fixed effects:  y ~ latitude
##                 Value Std.Error DF   t-value p-value
## (Intercept)  9.496793 1.9247991 53  4.933914  0.0000
## latitude    -0.081470 0.0362724  9 -2.246063  0.0513
##  Correlation:
##          (Intr)
## latitude -1
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.7577350 -0.5397447 -0.1141964  0.4664333  2.8930977
##
## Number of Observations: 64
## Number of Groups: 11
```

Note que a estimativa da variância residual do modelo REML ($\sigma^2 = 0.1307764^2 = 0.0171$) é aproximadamente a mesma da anova (`fm4`) acumulada incluindo o termo da cidade ($\sigma^2 = 0.01716$): os dois modelos são quase equivalentes em sua capacidade para explicar a variação dos preços das casas. A significância do efeito da latitude é testada por uma estatística F que tem quase o mesmo valor (5,04) que o valor de $F_{latitude}$ (5,19) obtido anteriormente (Seção 1.5). Os valores de p correspondentes a essas estatísticas de teste também são muito semelhantes ($p = 0,050$ e $0,049$, respectivamente). A pequena diferença entre eles decorre não só da diferença dos valores de F, mas também da diferença entre o valor DFDdenominador =9 obtido anteriormente e o valor obtido na presente análise, $DFDdenominador = 9,4$. Observe que enquanto em uma análise de regressão comum DF são sempre números inteiros, essa restrição é removida em uma análise de modelo misto. O valor $DFNumerator = 1$ é o mesmo em ambas as análises. Outra estatística de teste, a estatística Wald, também é apresentada. Geralmente está relacionado à estatística F pelas fórmulas

$$(EstatísticadeWald) = F \times DF_{NumeradordeF} \tag{4}$$

e

$$DF_{EstatísticadeWald} = F \times DF_{NumeradordeF} \tag{5}$$

Para uma pequena exceção, consulte a Seção 9.6; isso será considerado com mais detalhes posteriormente (Seção 1.9). O GenStat fornece testes separados dos efeitos de adicionar 'latitude' e retirá-lo do modelo, mas no caso deste modelo simples com apenas um termo de efeito fixo (além da constante), não há diferença entre esses testes. Finalmente, a saída fornece as estimativas dos

parâmetros $\hat{\beta}_1$ e $\hat{\beta}_1$ e seus SEs.

As estimativas e SEs produzidas pelos diferentes métodos de análise são apresentadas na Tabela 1.6. As três estimativas do efeito da latitude concordam intimamente, assim como as três estimativas da constante. No entanto, o SE do efeito estimado da latitude do Modelo 1.1 é muito menor do que os da análise de regressão nas médias das cidades e do Modelo 1.3. Isso ocorre porque o SE para o Modelo 1.1 é baseado na suposição de que toda casa é uma observação independente. Ele superestima a precisão da linha de melhor ajuste, assim como a estatística F desse modelo superestima a significância da relação entre o preço da casa e a latitude.

```
library(rsq)
rsq(fm1.me)
```

```
## $model
## [1] 0.6314344
##
## $fixed
## [1] 0.2465151
##
## $random
## [1] 0.3849192
```

Tabela 4: Comparação das estimativas dos parâmetros e seus SEs, obtidos a partir de diferentes métodos de análise do efeito da latitude sobre os preços da habitação na Inglaterra.

	Método de análise					
	Análise de regressão ignorando cidade		Análise de regressão com médias de cidades		Análise modelo misto	
	Estimativa	$EP_{estimativa}$	Estimativa	$EP_{estimativa}$	Estimativa	$EP_{estimativa}$
intercepto	9.68	1.00	9.44	1.87	9.497	1.9248
latitude	-0.0852	0.0188	-0.0804	0.0353	-0.08147	0.036272

As três linhas ajustadas podem ser transformadas de volta para as unidades originais (libras), usando a fórmula

$$\text{preço das casas estimados} = 10^{\log 10(\text{preço das casas})estimados}$$

$$
\begin{aligned}
\hat{y}_t &= \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \text{em que } \hat{y}_t = \log_{10} \hat{y} \\
\log_{10} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
10^{\log_{10} \hat{y}} &= 10^{\hat{\beta}_0 + \hat{\beta}_1 x_i} \\
\hat{y} &= 10^{\hat{\beta}_0 + \hat{\beta}_1 x_i} \\
\hat{y} &= 10^{\hat{y}_t}
\end{aligned}
$$

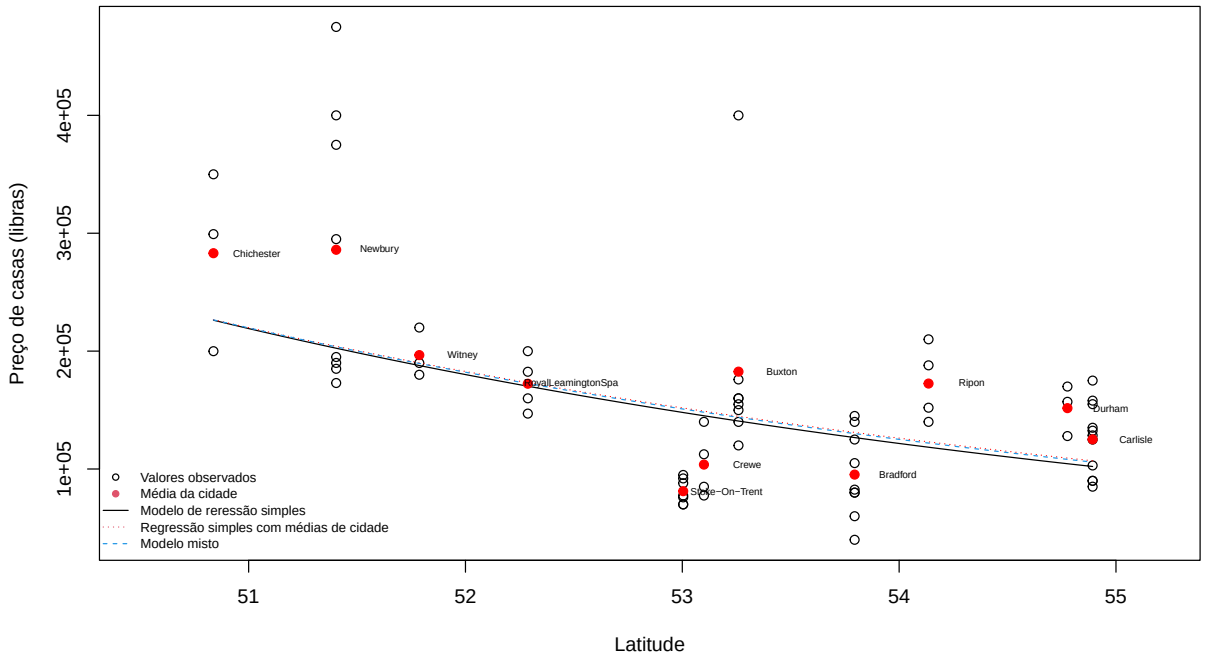


Figura 7: Relação entre latitude e preços de casas em uma amostra de cidades inglesas, comparando as linhas de melhor ajuste de análise de regressão simples, análise de médias de cidade e análise de modelo misto.

1.9 Testando os pressupostos das análises: Inspeção dos valores residuais

Para que os testes de significância e EPs nas análises anteriores sejam estritamente válidos, é necessário que o ε_{ij} cumpra a suposição impostas ao modelo.

```
(ef<- ranef(fm1.me))
```

```
##                     (Intercept)
```

```
## Bradford           -0.151967968
## Buxton              0.065890019
## Carlisle            0.056347314
## Chichester          0.066106136
## Crewe              -0.136976302
## Durham              0.111272875
## Newbury             0.104277612
## Ripon               0.118821731
## RoyalLeamingtonSpa -0.002909835
## Stoke-On-Trent     -0.242005847
## Witney              0.011144266
```

```
options(digits = 5)
# valores preditos
pred<- predict(fm1.me, level = 0:1)  ;pred[1:10, ]
```

```
##        town predict.fixed predict.town
## 1  Bradford        5.1141       4.9622
## 2  Bradford        5.1141       4.9622
## 3  Bradford        5.1141       4.9622
## 4  Bradford        5.1141       4.9622
## 5  Bradford        5.1141       4.9622
## 6  Bradford        5.1141       4.9622
## 7  Bradford        5.1141       4.9622
## 8  Bradford        5.1141       4.9622
## 9  Bradford        5.1141       4.9622
## 10   Buxton        5.1578       5.2237
```

Relembrando o modelo (2)

$$y_{ij} = \beta_0 + \beta_1 x_i + \tau_i + \varepsilon_{ij}$$
$$y_{ij} = (\beta_0 + \tau_i) + \beta_1 x_i + \varepsilon_{ij}$$

Estimativas dos parâmetros fixos:

```
(tb<- summary(fm1.me)$tTable)
```

```
##                Value Std.Error DF t-value    p-value
```

```
## (Intercept)  9.49679  1.924799 53  4.9339 8.3763e-06
## latitude    -0.08147  0.036272  9 -2.2461 5.1332e-02
```

Digamos que há o interesse no valor predito para uma casa da cidade `Buxton` ($x_i = 53.259$), temos duas situações:

* Casa da amostra:

$$\hat{y_{ij}} = (\hat{\beta}_0 + \tilde{\tau}_i) + \hat{\beta}_1 x_i$$
$$\hat{y_{ij}} = 9.49679 + 0.06589 - 0.08147 x_i$$
$$\hat{y_{ij}} = 9.56268 - 0.08147(53.259)$$
$$\hat{y_{ij}} = 9.56268 - 4.33902$$
$$\hat{y_{ij}} = 5.22366$$

Resíduo: 5.07918-5.22366=-0.14448

* Uma nova casa:

$$\hat{y_{ij}} = (\hat{\beta}_0 + \tilde{\tau}_i) + \hat{\beta}_1 x_i$$
$$\hat{y_{ij}} = 9.49679 + 0 - 0.08147(53.259)$$
$$\hat{y_{ij}} = 9.49679 - 4.33902$$
$$\hat{y_{ij}} = 5.15777$$

Resíduo: 5.07918-5.15777= -0.07859

1.9.1 Resíduos

```
RE.resp<- residuals(fm1.me, type = "response", level = 0:1); RE.resp[1:10,]
```

```
##        fixed      town
## 1  -0.512615 -0.360647
## 2  -0.336342 -0.184374
## 3  -0.211313 -0.059345
## 4  -0.211069 -0.059101
## 5  -0.197677 -0.045709
## 6  -0.092942  0.059026
## 7  -0.017221  0.134747
## 8   0.031842  0.183809
## 9   0.047237  0.199205
## 10 -0.078586 -0.144476
```

```
RE.per<- residuals(fm1.me, type = "pearson", level = 0:1)
RE.nor<- residuals(fm1.me, type = "normalized", level = 0:1)
```

```
apply(RE.resp, 2, range)
```

```
##          fixed     town
## [1,] -0.51261 -0.36065
## [2,]  0.44424  0.37835
```

```
apply(RE.per, 2, range)
```

```
##        fixed    town
## [1,] -3.9198 -2.7577
## [2,]  3.3969  2.8931
```

```
apply(RE.nor, 2, range)
```

```
##        fixed    town
## [1,] -3.9198 -2.7577
## [2,]  3.3969  2.8931
```

```
par(mar=c(7,6,1,1), mfrow=c(1,2))
boxplot(RE.resp[,2] ~ town, data = dat,las = 2, ylab = "Resíduos", xlab = "")
abline(h = 0, col="red", lty="dashed")
boxplot(RE.per[,2]  ~ town, data = dat,las = 2, ylab = "Resíduos", xlab = "")
abline(h=c(-3,-2,0,2,3), col=c(3,4,2,4,3),lty=2)
```

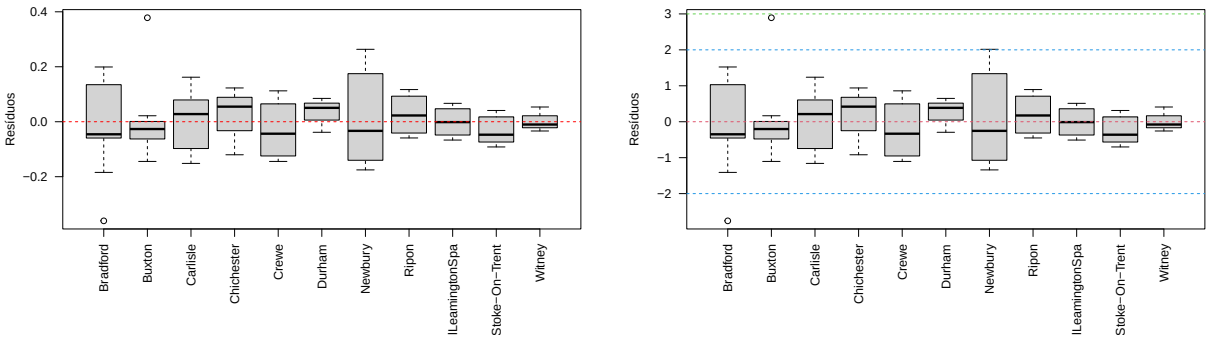


Figura 8: Resíduos brutos por cidade para o ajuste REML de um modelo de efeitos aleatórios aos dados de preços de residências.

```
par(mfrow=c(1,3))
plot(pred[,3], RE.resp[,2], xlab = expression(hat(y) ~ " - modelo de efeitos
```

```
                                aleatórios"), ylab = "Resíduo Bruto"  )
abline(h=0, lty=2)
plot(pred[,3], RE.per[,2], ylim = c(-3.5,3.5), xlab = expression(hat(y) ~ " - modelo de efeitos
abline(h=c(-3,-2,0,2,3), col=c(3,2,1,2,3),lty=2)
plot(pred[,3], RE.nor[,2], ylim = c(-3.5,3.5), xlab = expression(hat(y) ~ " - modelo de efeitos
abline(h=c(-3,-2,0,2,3), col=c(3,2,1,2,3),lty=2)
```

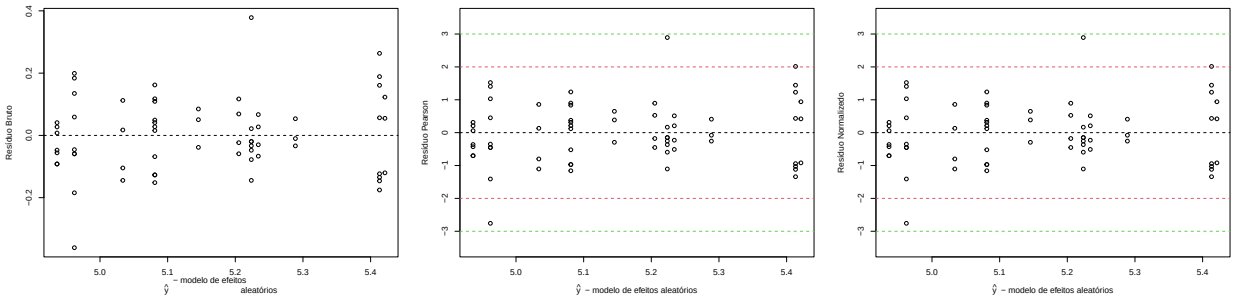


Figura 9: kkk Resíduos padronizados versus os valores ajustados para o ajuste REML de um modelo de efeitos aleatórios aos dados de preços de residências.

```
layout(1)
```

```
par(mfrow=c(1,3))
hist(RE.per[,2], breaks = 10, probability = TRUE)
curve(dnorm(x,0,1), from = -4, to = 4, add = TRUE, col=2)

qqnorm(RE.per[,2]);qqline(RE.per[,2])

library(hnp)
hnp(RE.per[,2], halfnormal = FALSE)
```

```
## Half-normal plot with simulated envelope generated assuming the residuals are
##         normally distributed under the null hypothesis.
```

Para fazer o gráfico de resíduos *studentizado*, pode ser usado a função `plot.lme`:

```
plot(fm1.me)
```

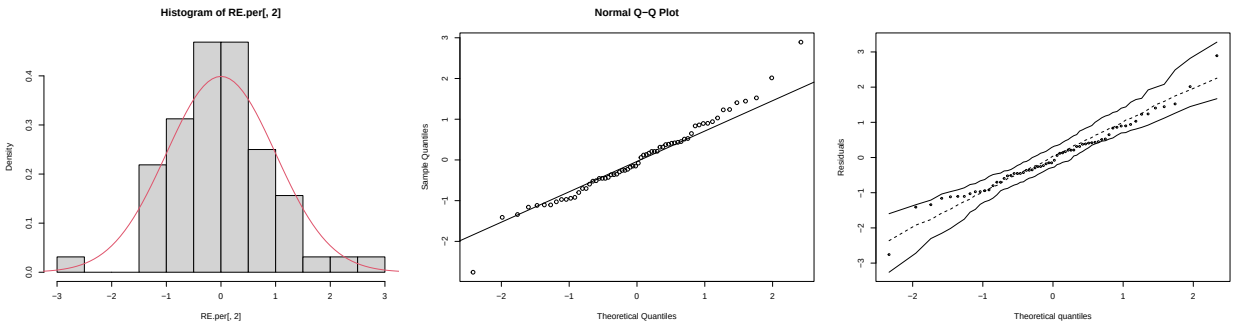
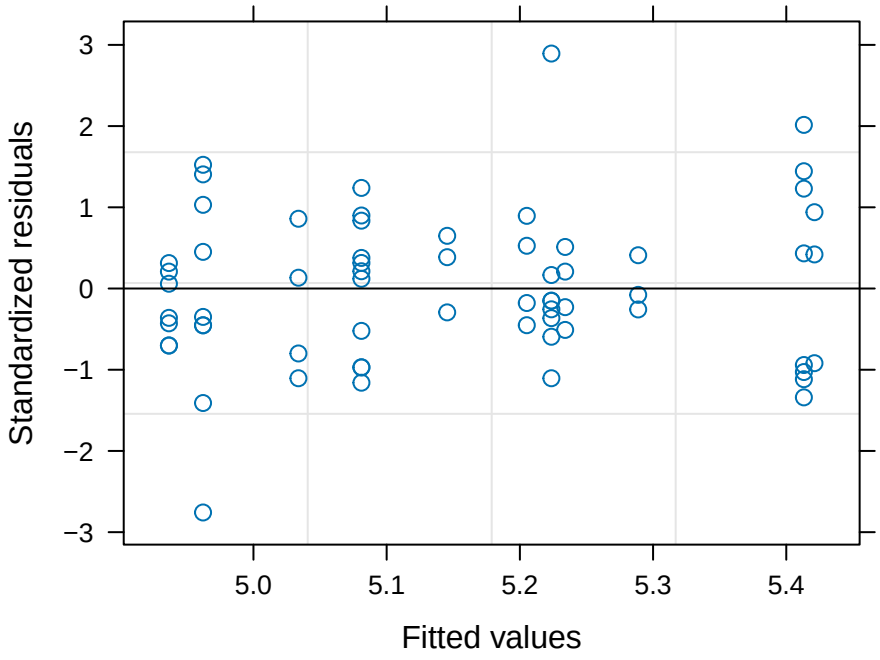


Figura 10: Resíduos do modelo de efeitos aleatórios ajustado por REML aos dados de preços de residências.



Matrizes importantes, considerando a cidade de Chichester:

```
(V <- getVarCov(fm1.me, type = "marginal", individual = "Chichester"))
```

```
## town Chichester
## Marginal variance covariance matrix
##          1        2        3
## 1 0.036734 0.019632 0.019632
## 2 0.019632 0.036734 0.019632
## 3 0.019632 0.019632 0.036734
##   Standard Deviations: 0.19166 0.19166 0.19166
```

```
(D <- getVarCov(fm1.me, type = "random.effects", individual = "Chichester"))
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)    0.019632
##   Standard Deviations: 0.14011
```

```
(R_1 <- getVarCov(fm1.me, type="conditional", individual = "Chichester"))
```

```
## town Chichester
## Conditional variance covariance matrix
##          1        2        3
## 1 0.017102 0.000000 0.000000
## 2 0.000000 0.017102 0.000000
## 3 0.000000 0.000000 0.017102
##   Standard Deviations: 0.13078 0.13078 0.13078
```

Entendendo as matrizes que compões Var(Y):

```
options(digits = 3)

Z <- model.matrix(~town-1,data=dat)

sig2b<- 0.14011^2
sig2 <- 0.13078^2

D<- diag(11)*sig2b
ZDZ<- Z%*% D %*% t(Z)
R<- diag(64)*sig2
V<- ZDZ + R
dim(V)
```

```
## [1] 64 64
```

```
# cidades:  Chichester, Crewe  e Durham
table(dat$town)
```

```
##
##           Bradford             Buxton           Carlisle         Chichester
##                  9                  8                 11                  3
##              Crewe             Durham            Newbury              Ripon
##                  4                  3                  8                  4
## RoyalLeamingtonSpa     Stoke-On-Trent             Witney
##                  4                  7                  3
```

```
V[29:38, 29:38]
```

```
##        29     30     31     32     33     34     35     36     37     38
## 29 0.0367 0.0196 0.0196 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## 30 0.0196 0.0367 0.0196 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## 31 0.0196 0.0196 0.0367 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## 32 0.0000 0.0000 0.0000 0.0367 0.0196 0.0196 0.0196 0.0000 0.0000 0.0000
## 33 0.0000 0.0000 0.0000 0.0196 0.0367 0.0196 0.0196 0.0000 0.0000 0.0000
## 34 0.0000 0.0000 0.0000 0.0196 0.0196 0.0367 0.0196 0.0000 0.0000 0.0000
## 35 0.0000 0.0000 0.0000 0.0196 0.0196 0.0196 0.0367 0.0000 0.0000 0.0000
## 36 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0367 0.0196 0.0196
## 37 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0196 0.0367 0.0196
## 38 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0196 0.0196 0.0367
```