Análise de Regressão: Teoria e Prática

Departamento de Estatística e Matemática Aplicada

Rafael Braz, Ronald Targino, Juvêncio Nobre e Manoel Santos-Neto

2025-08-18

Índice

Prefácio

Este livro....

1 Introdução

2 Preliminares

3 Análise de Regressão

3.1 Regressão Linear Simples

3.1.1 Motivação

Regressão Linear Simples: É um método estatístico que nos permite resumir e estudar as relações entre duas variáveis quantitativas:

- Uma variável, denotada por x, é considerada como preditora, explicativa ou variável independentes.

- A outra variável, denotada por y, é considerada como a resposta, resultado ou variável dependente.

  Usaremos os termos **"preditor"** e **"resposta"** para nos referirmos às variáveis utilizadas neste curso. Os outros termos são mencionados apenas para torná-lo ciente deles caso você os encontre em outros materiais. A regressão linear simples recebe o adjetivo *"simples"*, porque diz respeito ao estudo de apenas uma variável preditora. Em contraste, a regressão linear múltipla, que estudaremos mais adiante neste curso, recebe o adjetivo *"múltipla"*, porque diz respeito ao estudo de duas ou mais variáveis preditoras.

No slide anterior, foi possível observar que se você conhece a temperatura em graus Celsius, pode usar uma equação para determinar exatamente a temperatura em graus Fahrenheit.

Agora serão apresentadas outros exemplos de relações determinística.

1. Circuferência $= \pi \times$ diâmetro.

2. **Lei de Hooke:** $Y = \alpha + \beta X$, em que Y é a quantidade de alogamento em uma mola e X é o peso aplicado.

3. **Lei de Ohm:** $I = V/r$, em que V é a tensão aplicada, r é a resistência elétrica e I é a corrente elétrica.

4. **Lei de Boyle:** Para uma temperatura constate, $P = \alpha/V$, em que P é a pressão, α é uma constante para cada gás e V é o volume do gás.

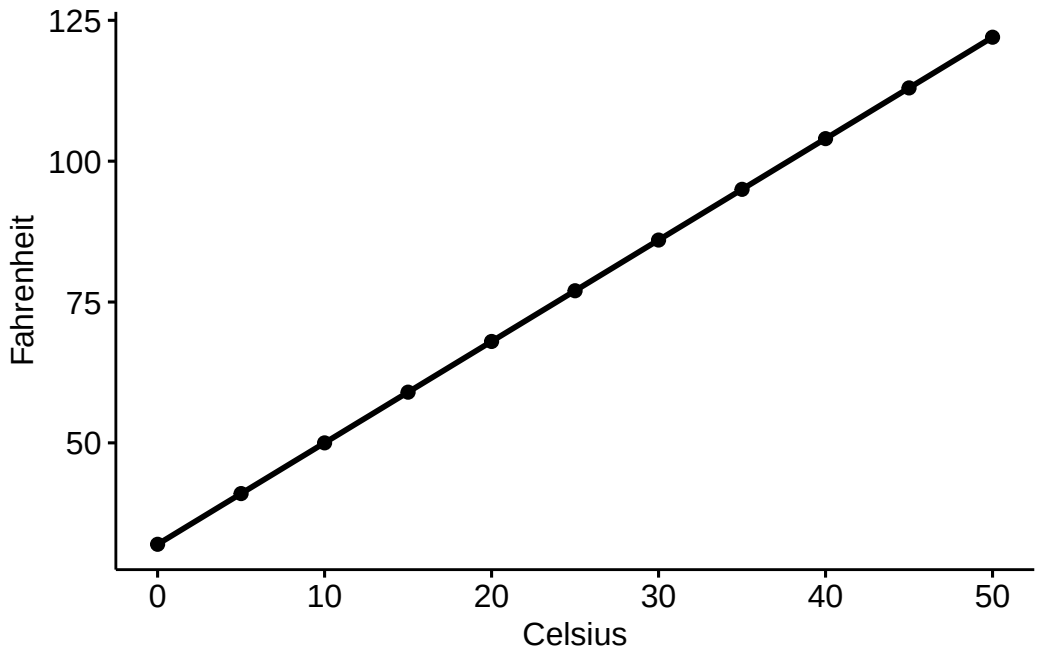
Para cada uma dessas relações determinísticas, a equação descreve exatamente a relação entre as duas variáveis. Esta disciplina não examina relacionamentos determinísticos. Em vez disso, estamos interessados em relações estatísticas, nas quais a relação entre as variáveis não é perfeita.

Primeiro devemos deixar claro quais tipos de relacionamentos não estudaremos neste curso, ou seja, relacionamentos determinísticos (ou funcionais). Abaixo está um exemplo de uma relação determinística.

```
library(ggpubr)
```

```
Carregando pacotes exigidos: ggplot2
```

```
cels <- seq(0, 50, by = 5)
fahr <- (9/5)*cels + 32
data <- data.frame(x = cels, y = fahr)
ggscatter(data,
          x = "x",
          y = "y",
          xlab = "Celsius",
          ylab = "Fahrenheit",
          add = "reg.line")
```



Observe que os pontos de dados observados caem diretamente em uma linha. Como você deve se lembrar, a relação entre graus Fahrenheit e graus Celsius é conhecida como:

$$\text{Fahrenheit} = (9/5) \times \text{Celsius} + 32.$$

Agora iremos apresentar um exeplo de relação estatística. A variável resposta Y é a mortalidade por cancêr de pele (por 10 milhões de pessoas) e a variável preditora X é a latitude no centro de cada um dos 48 estados americanos (dados de câncer de pele dos EUA). Os dados foram obtidos na década de 1950, então o Alasca e o Havaí ainda não eram estados. Além disso, Washington, DC está incluído no conjunto de dados, embora não seja tecnicamente um estado.

```
library(tidyverse)
library(DT)
skincancer <- read_table("skincancer.txt")
datatable(skincancer,
          options = list(pageLength = 5, scrollY = "200px"))
```

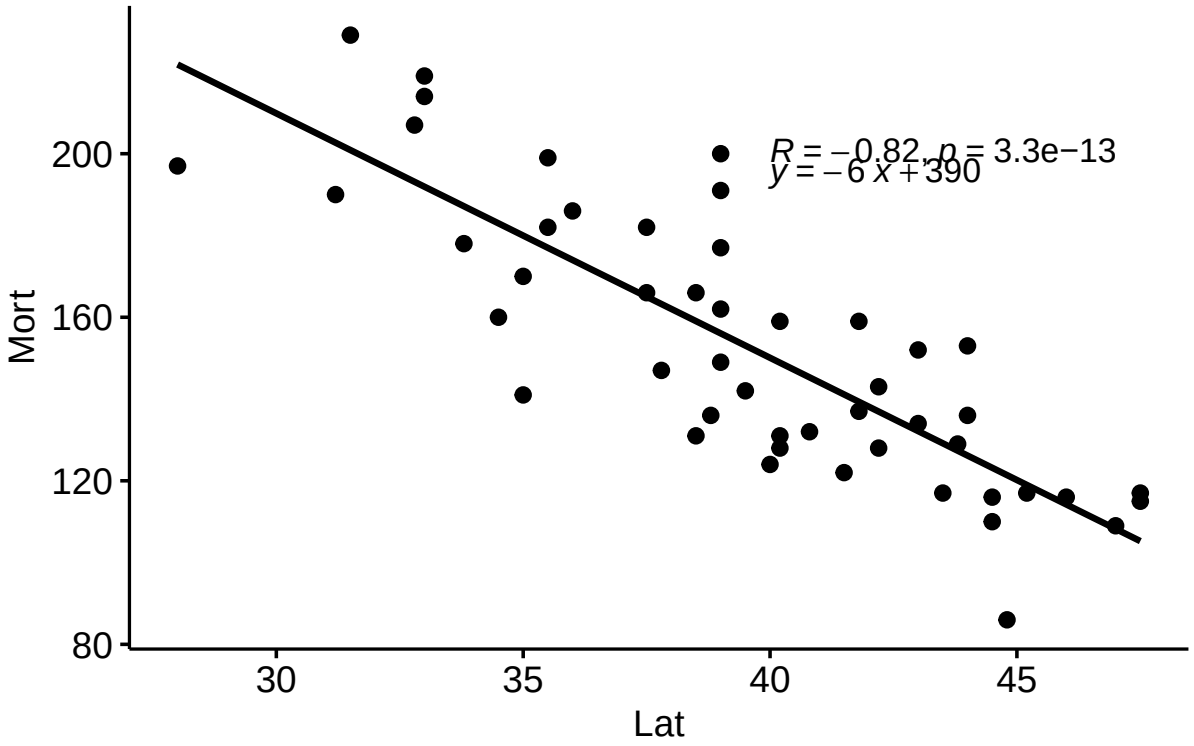
Show 5 entries Search:

	State				I
1	Alabama	33	219	1	87
2	Arizona	34.5	160	0	112
3	Arkansas	35	170	0	92.5
4	California	37.5	182	1	119.5
5	Colorado	39	149	0	105.5

Showing 1 to 5 of 49 entries

Previous 1 2 3 4 5 … 10 Next

Note que viver nas latitudes mais altas do norte dos Estados Unidos, diminuiria a exposição aos raios nocivos do sol e, portanto, menos risco teria de morrer devido ao câncer de pele. O gráfico de dispersão suporta tal hipótese. Parece haver uma relação linear negativa entre latitude e mortalidade por câncer de pele, mas a relação não é perfeita. De fato, o enredo exibe alguma "tendência", mas também exibe alguma "dispersão". Portanto, é uma relação estatística, não determinística.



Alguns outros exemplos de relações estatísticas podem incluir:

- Altura e peso – à medida que a altura aumenta, você esperaria que o peso aumentasse, mas não perfeitamente.
- Álcool consumido e teor alcoólico no sangue — à medida que o consumo de álcool aumenta, você esperaria que o teor alcoólico no sangue aumentasse, mas não perfeitamente.
- Capacidade pulmonar vital e maços-ano de tabagismo — à medida que a quantidade de fumo aumenta (conforme quantificado pelo número de maços-ano de tabagismo), você esperaria que a função pulmonar (conforme quantificada pela capacidade pulmonar vital) diminuísse, mas não perfeitamente.
- Velocidade de direção e consumo de combustível — à medida que a velocidade de direção aumenta, você esperaria que o consumo de combustível diminuísse, mas não perfeitamente.

Portanto, vamos estudar as relações estatísticas entre uma variável de resposta y e uma variável preditora x!

3.1.2 Pressupostos do modelo

Observação:

Importante destacar que o termo regressão linear significa .blue[regressão linear nos parâmetros], ou seja, da forma

$$y_i = \alpha + \beta x_i^2 + u_i$$

ou da forma

$$\log(y_i) = \alpha + \beta \log(x_i) + u_i,$$

também são considerados **regressões lineares**.

O parâmetro

$$E(Y|X = x) = \alpha + \beta x$$

que representa a média da variável aleatória Y, condicionada a $X = x$, será estimada por

$$\widehat{E(Y|X = x)} = a + bx,$$

em que a e b são estimativas para α e β. A quantidade

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i), \quad i = 1, \dots, n,$$

é chamada de resíduo.

Assim, o valor e_i pode ser interpretado como o erro cometido por prever $y_i (i = 1, \dots, n)$ a partir de $\hat{y}_i$.

Voltando ao Exemplo (Semana 1)

Quais as estimativas do modelo de regressão linear simples de interesse?

$$\hat{y} = 390 - 6x.$$

$$y_i = E(Y|X = x_i) + u_i = \alpha + \beta x_i + u_i,$$

em que α é o intercepto e β é o coeficiente angular da reta de regressão.

Na prática, nem sempre α (intercepto) apresenta interpretação.

Como as estimativas devem ser interpretadas?

Voltando ao Exemplo (Semana 1)

$$\widehat{y} = 390 - 6x.$$

- 390: valor médio de mortes por cancêr de pele em um estado com latitude central igual a zero. (Faz sentido essa interpretação?)
- -6: variação média no número de mortes quando aumenta-se a latitude em 1 unidade.

Exercício: Encontre a matriz hessiana e verifique sob quais condições a mesma é definida como positiva. Ainda, discuta se os estimadores encontrados geram o mínimo da função de interesse.

3.1.3 Estimação dos parâmetros pelo método dos mínimos quadrados

3.1.4 Propriedades dos estimadores

3.1.5 Decomposição da Soma de Quadrados Total

O modelo de regressão proposto está bem ajustado? Como medir a qualidade de ajuste do modelo?

> **Objetivo:** Construir uma medida que indique, mesmo que de modo imperfeito, a qualidade do ajuste do modelo de regressão.

- $y - \bar{y}$: erro ao se prever y pela média geral.
- $y - \widehat{y}$: erro ao se prever y pelo valor estimado para $E(Y|X)$.
- $\widehat{y} - \bar{y}$: "ganho" ao se prever y pelo valor estimado para $E(Y|X)$ em comparação ao se prever y pela média geral.
- Soma de Quadrados Total (SQT): $SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
- Soma de Quadrados devido aos Resíduos (SQE): $SQE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$.
- Soma de Quadrados devido ao modelo de regressão (SQReg): $SQReg = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$.

Resultado: $SQT = SQReg + SQE$

- Na SQT temos $n - 1$ graus de liberdade.
- Na SQE temos $n - 2$ graus de liberdade.

- Assim, chegamos aos chamados **quadrados médios**:

$$QMT = \frac{SQT}{n-1} = S^2, \quad QME = \frac{SQE}{n-2} \quad \text{e} \quad QMReg = \frac{SQReg}{1}.$$

- Desta forma, chegamos na estatística $F = \frac{QMreg}{QME} \sim F(1, n-2)$.

3.1.6 Tabela de ANOVA

Esses resultados nos levam a seguinte tabela de análise de variâncias:

	CV	GL	SQ	QM	F
Regressão	1	$SQReg$		$QMReg$	$QMReg/QME$
Resíduo	$n-2$	SQE		QME	
Total	$n-1$	SQT			

3.1.7 Coeficiente de Determinação

O coeficiente de determinação (explicação) é definido por

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQE}{SQT},$$

e mostra a proporção da variabilidade de y que é "explicada" pelos regressores do modelo adotado. Já o coeficiente de determinação ajustado é definido por:

$$\bar{R}^2 = 1 - \frac{SQE/(n-2)}{SQT/(n-1)}.$$

Voltando ao Exemplo (Semana 1)

Interpretação: 67% das variações das mortes por câncer de pele são explicadas pela latitude no centro de cada um dos 48 estados americanos.

Conclusão: Desta forma, parece que a posição geográfica do estado é relevante para a explicação da mortalidade por câncer de pele uma vez que tal regressor explica mais da metade das variações da variável resposta.

Exercício: Prove que, no caso do modelo de regressão linear simples com intercepto, o coeficiente de correlação linear de Pearson elevado ao quadrado é igual ao coeficiente de explicação (ou determinação) – R^2. Ou seja,

$$R^2 = \frac{SQReg}{SQT} = \frac{S_{XY}^2}{S_{XX}S_{YY}} = b\left(\frac{S_{XY}}{S_{YY}}\right).$$

3.1.8 Testes de Hipóteses sobre a inclinação e o intercepto

3.1.9 Intervalos de Confiança para a inclinação e para o intercepto

3.1.10 Intervalos de Confiança para a variância e para a média da variável resposta para um valor fixo da variável independente

3.1.11 Intervalos de Previsão

3.1.12 Teste para Falta de Ajustamento

Quando dispomos, para um ou mais valores de X, de mais de um valor observado Y, é possível obter uma outra estimativa da variância do erro. Essa outra estimativa de σ^2 é dada pelo quadrado médio do resíduo de uma análise de variância em que cada valor distinto de X é encarado como um diferente "tratamento" a que está sendo submetida a variável Y. Temos neste caso, portanto, dois resíduos, vamos nos referir ao primeiro, explicitamente, como "resíduo da regressão" e ao segundo, simplesmente como "resíduo".

Seja K o número de valores distintos de X, representamos por $T_k (k = 1, \ldots, K)$ os totais dos tratamentos, isto é, as soma dos valores de Y_i, para cada valor distinto de X_i.

X_k	Y_i	T_k
0	3	3
1	2 e 3	5
2	5	5
3	4 e 4	8
4	7	7
5	6 e 7	13
6	9	9

$$SQTrat = \sum_{k=1}^{K} \frac{T_k}{n_k} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}, \quad \text{e} \quad SQRes = \sum_{i=1}^{n} Y_i^2 - \sum_{k=1}^{K} \frac{T_k}{n_k}.$$

Com base nas esperanças dessas soma de quadrados, justifica-se a associação de $K-1$ e $n-K$ graus de liberdade à $SQTrat$ e $SQRes$, respectivamente.

As diferenças entre as médias de tratamentos $\bar{Y}_k$ e os respectivos valores de Y estimados pela regressão $\hat{Y}_k$ associamos a soma de quadrados de "falta de ajustamento", definida por

$$SQFA = \sum_{k=1}^{K} n_k(\bar{Y}_k - \hat{Y}_k)^2 \quad \text{com } K-2 \text{ graus de liberdade.}$$

E pode ser obtido também da seguinte maneira:

$$SQFA = SQResReg - SQRes.$$

C.V	G.L	SQ	QM	F	
Total		9	44		
Regressão		1	36	36	36
Res. de Regressão		8	8	1	
Falta de Ajustamento		5	7	1,4	4,2
Resíduo		3	1	1/3	

O resultado obtido mostra que a "falta de ajustamento" não significativa ao nível de significância de 5%.

Nos caso em que a "falta de ajustamento" é significativa, concluímos que o modelo linear utilizado não é apropriado, pois o quadrado médio do resíduo da regressão não estimaria corretamente a variância residual (σ^2), uma vez que estaria incluindo um erro sistemático devido ao uso de um modelo inapropriado.

```
X <- c(0, 1, 1, 2, 3, 3, 4, 5, 5, 6)
Y <- c(3, 2, 3, 5, 4, 4, 7, 6, 7, 9)
ajuste <- lm(Y ~X)
ajusteFA <- lm(Y ~factor(X))
anova(ajuste)
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1     36      36      36 0.0003234 ***
Residuals  8      8       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(ajuste, ajusteFA)
```

```
Analysis of Variance Table

Model 1: Y ~ X
Model 2: Y ~ factor(X)
  Res.Df RSS Df Sum of Sq   F Pr(>F)
1      8   8
2      3   1  5         7 4.2 0.1336
```

3.1.13 Análise de Resíduos

O -ésimo resíduo e dado por

$$\hat{u}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n.$$

Agora iremos estudar o comportamento individual e conjunto destes resíduos, comparando com as suposições feitas sobre os verdadeiros erros u_i. Existem várias técnicas formais para conduzir essa análise, mas aqui iremos ressaltar basicamente métodos gráficos.

Uma representação gráfica bastante útil é obtida plotando-se pares $(x_i, \hat{u}_i), i = 1, \ldots, n$. Outras vezes, é de maior utilidade fazer a representação gráfica dos chamados **resíduos padronizados**,

$$\hat{z}_i = \frac{\hat{u}_i}{s^2},$$

plotando-se os pares $(x_i, \hat{z}_i)$. Outro resíduo usado é o chamado **resíduo estudentizado**, definido por

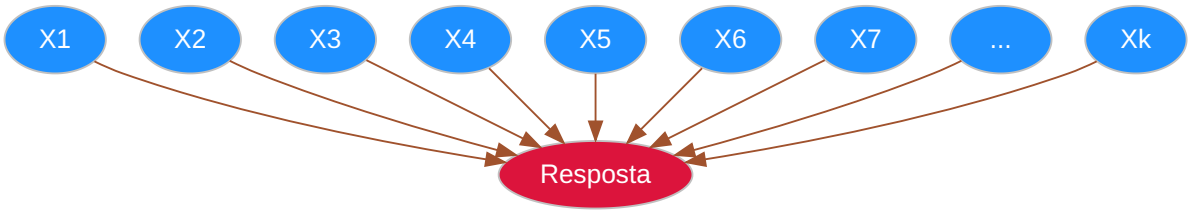
$$\hat{r}_i = \frac{\hat{u}_i}{s^2\sqrt{1 - v_{ii}}},$$

em que $v_{ii} = \frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^{n} x_i^2}$.

3.1.14 Analisar dados usando o R

3.2 Regressão Linear Múltipla

Até aqui, aprendemos a trabalhar com o modelo de regressão linear simples, com uma variável explicativa, e notamos que o mesmo pode ser utilizado em diversas situações. Porém, vários problemas envolvem duas ou mais variáveis explicativas influenciando o comportamento da variável resposta (Y). Qualquer modelo de regressão linear com dois ou mais regressores recebe o nome de .red[**modelo de regressão linear múltipla**



Agora iremos admitir que $X_1, X_2, \dots, X_k$ sejam variáveis independentes e Y a variável resposta. Dada uma amostra aleatória de n observações $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i), \quad i = 1, 2, \dots, n$, o modelo de regressão linear múltipla será dado por

$$E(Y_i|x_{1i}, x_{2i}, \dots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}, \quad i = 1, 2, \dots, n,$$

ou

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, \dots, n,$$

em que $n > k + 1$.

Iremos considerar uma estrutura similar a do modelo de regressão linear simples. Especificamente, estamos considerando o seguinte:

- Modelo de regressão linear, ou **linear nos parâmetros**.
- Valores fixos de X.
- O termo de erro ϵ_i tem valor médio zero.
- Homocedasticidade ou variância constante de ϵ_i.
- Ausência de autocorrelação, ou de correlação serial, entre os termos de erro.
- Não há colinearidade exata entre as variáveis X.
- Ausência de viés de especificação.

3.2.1 Interpretação da equação de regressão múltipla

Considerando que temos apenas duas variáveis regressoras, então

$$E(Y_i|x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

Desta forma, a equação acima fornece o **valor esperado ou média de Y condicional aos valores dados ou fixados de** X_1 e X_2.

Os coeficientes de regressão β_1 e β_2 são conhecidos como **coeficientes parciais de regressão** ou **coeficientes parciais angulares**. Seu significado é o seguinte: β_1 mede a *variação* no valor médio de Y, $E(Y)$, por unidade de variação em X_2, mantendo-se o valor de X_2 constante. Em outras palavras, ele nos dá o efeito "direto" ou "liquido" de uma unidade de variação em X_2 sobre o valor médio em Y, excluídos os efeitos que X_2 possa ter sobre a média de Y. De modo análogo, β_2 mede a variação do valor médio de Y por unidade de variação em X_2, mantendo-se constante o valor de X_1. Eles nos dá o efeito "direto" ou "liquido" de uma unidade de variação de X_2 sobre o valor médio de Y, excluídos quaiquer efeitos que X_1 possa ter sobre o valor médio de Y.

3.2.2 Abordagem Matricial

Por patricidade iremos utilizar a abordagem matricial, que no permitirar, entre outras coisas: i) encontrar o vetor de estimadores; ii) verificar as propriedade estatísticas dos estimadores; iii) obter a distribuição dos estimadores; qualquer que seja o número de variáveis independentes no modelo.

Sendo assim, podemos escrever o modelo de regressão linear múltipla como:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

que é conhecido como **modelo linear geral**.

Para determinarmos os estimadores de mínimos quadrados ordinários devemos minimizar

$$S(\beta) = \sum_{i=1}^{n} (\epsilon_i)^2 = \epsilon_1^2 + \cdots + \epsilon_n^2 = \epsilon^\top \epsilon,$$

ou

$$S(\beta) = \epsilon^\top \epsilon = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$