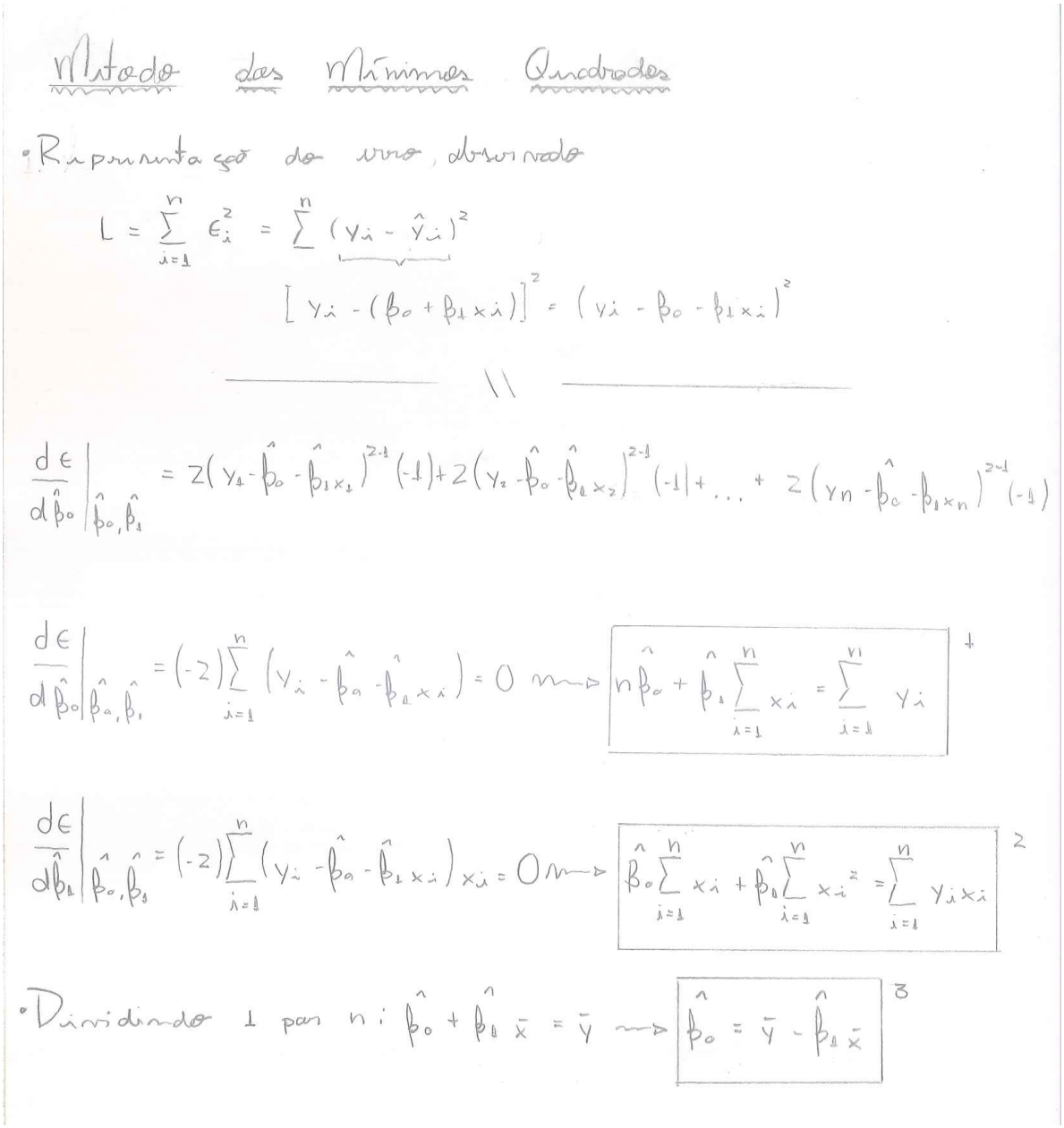
Ciência dos Dados Projeto 3 Parte 2
–
Desenvolvimento Teórico da Técnica de Regressão

A seguir segue uma demonstração da dedução do método dos mínimos quadrados.



Método das Mínimas Quadradas

• Representação do erro absoluto

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum^{n} (y_i - \hat{y}_i)^2$$

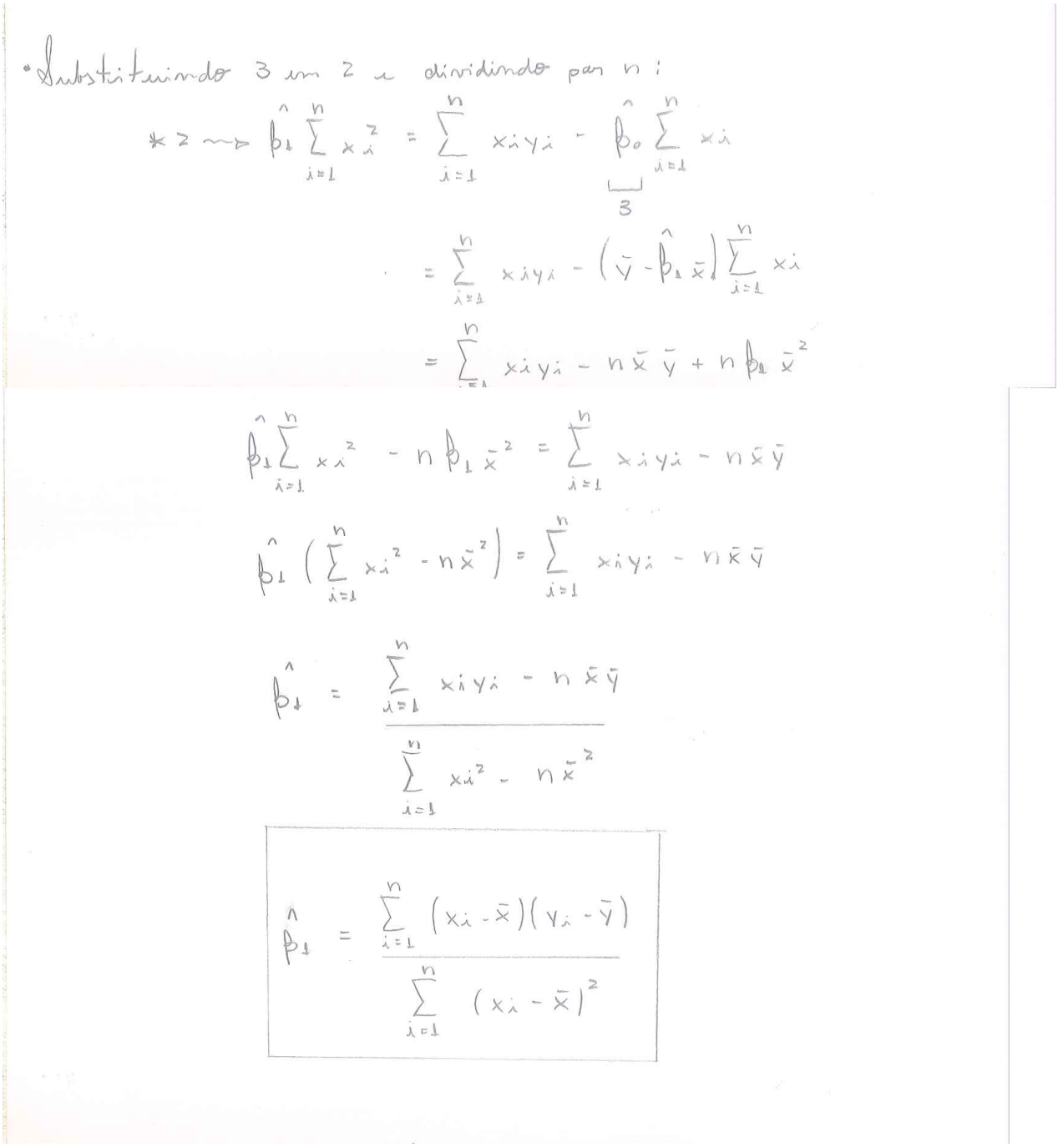
$$[y_i - (\beta_0 + \beta_1 x_i)]^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{d\epsilon}{d\hat{\beta}_0}\bigg|_{\hat{\beta}_0,\hat{\beta}_1} = 2(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^{2-1}(-1) + 2(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^{2-1}(-1) + \dots + 2(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^{2-1}(-1)$$

$$\frac{d\epsilon}{d\hat{\beta}_0}\bigg|_{\hat{\beta}_0,\hat{\beta}_1} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightsquigarrow \boxed{n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i} \quad 1$$

$$\frac{d\epsilon}{d\hat{\beta}_1}\bigg|_{\hat{\beta}_0,\hat{\beta}_1} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \rightsquigarrow \boxed{\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i} \quad 2$$

• Dividindo 1 por n: $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \rightsquigarrow \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$ 3



Sobre a distribuição dos erros podemos supor que este respeita uma distribuição normal de média zero e variância constante ou seja os dados regredidos encontram-se dispersos de forma homogênea em torno da reta de regressão, apresentando homocedasticidade. Na prática essas suposições podem ser checadas a partir da análise das funções de distribuição acumulada (fda) e da própria distribuição normal dos erros estabelecendo hipóteses e valores limites para que esta seja aceita.

Para uma regressão simples o teste de hipótese seria realizado para a hipótese nula $\beta 1 = 0$, dessa forma se a hipótese nula após o teste fosse

rejeitada poderíamos concluir que não existe relação linear entre as duas variáveis analisadas.

Além de regressões simples que estabelecem a relação entre uma variável explicativa e uma variável resposta, as regressões podem demonstrar a relação entre mais de duas variáveis sendo assim classificada como regressão múltipla que resultaria em uma equação mais complexa que envolvesse as outras variáveis de modo que a nova equação da reta seria: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon_i$. E para verificar a relação entre as variáveis deveriam ser feitos testes de hipóteses para cada uma das variáveis explicativas (n testes).