Lista 5

MI406-Regressão

Caio Gomes Alves

1 Questão 1

1.1 Pergunta

Considere o conjunto de dados abaixo:

x	y
1	2.67
1	3.48
1	2.46
4	3.40
4	2.13
4	0.98
7	6.19
7	6.44
7	6.28
10	14.69
10	16.51
10	15.39

- (a) Calcule a Soma de Quadrados dos Resíduos considerando o modelo $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- (b) Calcule a Soma de Quadrados de Erro Puro.
- (c) Calcule a Soma de Quadrados da Falta de Ajuste ("Lack of Fit").
- (d) Faça um teste para determinar se o modelo linear é apropriado.

1.2 Resposta

a)

Considerando o modelo $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, teremos que:

```r
# Primeiramente, criemos o modelo de regressão especificado:
mod1 <- lm(y ~ x, data = df1)

# A soma dos quadrados dos resíduos pode ser
# facilmente calculada com os resíduos do
# objeto onde o modelo está armazenado:
(SQR1 <- sum(residuals(mod1)^2))
```

```
## [1] 79.14264
```

b)

Utilizando as funções do pacote `tidyverse`, podemos agregar as observações na base de dados e calcular a Soma de Quadrados de Erro Puro:

```r
# Soma de Quadrados de Erro Puro:
(SQEp1 <- df1 %>%
     group_by(x) %>%
     mutate(y_bar = mean(y)) %>%
     ungroup() %>%
     mutate(EP = y - y_bar) %>%
     summarise(SQEp = sum(EP^2)) %>%
     unlist())
```

```
##     SQEp
## 5.228467
```

c)

Semelhantemente ao efetuado no item **(b)**, temos que a Soma de Quadrados da Falta de Ajuste será dada por:

```r
# Soma de Quadrados da Falta de Ajuste
(SQLoF1 <- df1 %>%
     cbind(data.frame(y_pred = predict(mod1, df1))) %>%
     group_by(x) %>%
     summarise(y_bar = mean(y),
               y_pred = mean(y_pred)) %>%
     mutate(LOF = y_pred - y_bar) %>%
     summarise(SQLoF = round(3 * sum(LOF^2), 4)) %>%
     unlist())
```

```
##   SQLoF
## 73.9142
```

d)

Para verificar se o modelo linear é apropriado para esse conjunto de dados, usaremos o teste de falta de ajuste, dado por:

$$\frac{\text{SQLoF}/(m-2)}{\text{SQEp}/(n-m)} \sim F_{(m-2,n-m)}$$

Em que $n = 12$ é a quantidade de observações, $m = 4$ é a quantidade de observações distintas de x e $F_{(m-2,n-m)}$ é a distribuição F de Snedecor, com $(4-2, 12-4)$ graus de liberdade. Assim, se considerarmos um nível de significância de 95% para o teste, teremos:

```r
# Estatística de Teste:
(estat_teste1 <- unname((SQLoF1/2)/(SQEp1/8)))
```

```
## [1] 56.54752
```

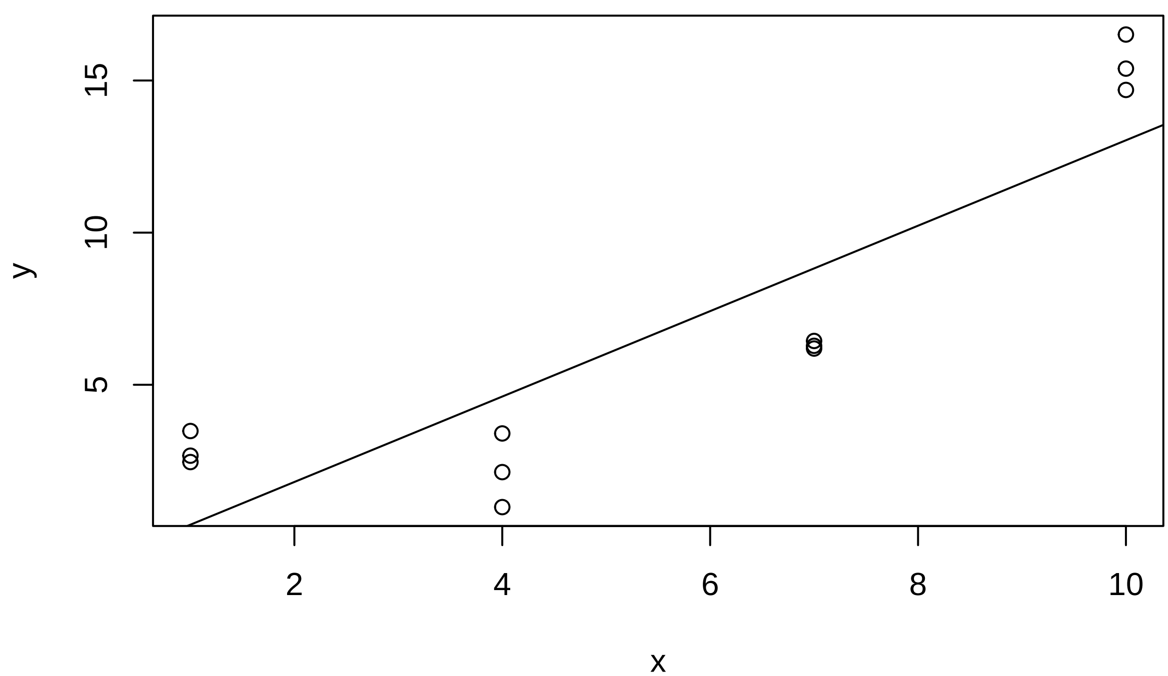
```r
# Valor do quantil 0.95 da distribuição F:
(f_1 <- qf(0.95, 2, 8))
```

```
## [1] 4.45897
```

```r
# Estatística de Teste > Valor de F?
estat_teste1 > f_1
```

```
## [1] TRUE
```

Como o valor da estatística de teste foi maior do que o valor do quantil 0.95 da distribuição F correspondente, rejeitamos a hipótese nula do teste, de modo que alguma das esperanças condicionais nos valores únicos de x não é expresso pela relação do modelo ajustado. Podemos verificar isso com o gráfico do modelo ajustado:



2 Questão 2

2.1 Pergunta

Considere o conjunto de dados abaixo:

- (a) Calcule a Soma de Quadrados dos Resíduos considerando o modelo $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

x	y
1	2.37
1	3.18
1	2.16
4	7.60
4	6.33
4	5.18
7	9.49
7	9.74
7	9.58
10	11.69
10	13.51
10	12.39

- (b) Calcule a Soma de Quadrados de Erro Puro.
- (c) Calcule a Soma de Quadrados da Falta de Ajuste ("Lack of Fit").
- (d) Faça um teste para determinar se o modelo linear é apropriado.

2.2 Resposta

a)

Considerando o modelo $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, teremos de maneira semelhante ao exercício anterior:

```
# Modelo de regressão especificado:
mod2 <- lm(y ~ x, data = df2)

# Soma dos Quadrados dos Resíduos:
(SQR2 <- sum(residuals(mod2)^2))
```

```
## [1] 5.81064
```

b)

Utilizando as funções do pacote `tidyverse`, podemos agregar as observações na base de dados e calcular a Soma de Quadrados de Erro Puro:

```
(SQEp2 <- df2 %>%
     group_by(x) %>%
     mutate(y_bar = mean(y)) %>%
     ungroup() %>%
     mutate(EP = y - y_bar) %>%
     summarise(SQEp = sum(EP^2)) %>%
     unlist())
```

```
##     SQEp
## 5.228467
```

c)

Semelhantemente ao efetuado no item **(b)**, temos que a Soma de Quadrados da Falta de Ajuste será dada por:

```
(SQLoF2 <- df2 %>%
     cbind(data.frame(y_pred = predict(mod2, df2))) %>%
     group_by(x) %>%
     summarise(y_bar = mean(y),
               y_pred = mean(y_pred)) %>%
     mutate(LOF = y_pred - y_bar) %>%
     summarise(SQLoF = round(3 * sum(LOF^2), 4)) %>%
     unlist())
```

```
##  SQLoF
## 0.5822
```

d)

Para verificar se o modelo linear é apropriado para esse conjunto de dados, usaremos o teste de falta de ajuste, dado por:

$$\frac{\text{SQLoF}/(m-2)}{\text{SQEp}/(n-m)} \sim F_{(m-2,n-m)}$$

Em que $n = 12$ é a quantidade de observações, $m = 4$ é a quantidade de observações distintas de x e $F_{(m-2,n-m)}$ é a distribuição F de Snedecor, com $(4-2, 12-4)$ graus de liberdade. Assim, se considerarmos um nível de significância de 95% para o teste, teremos:

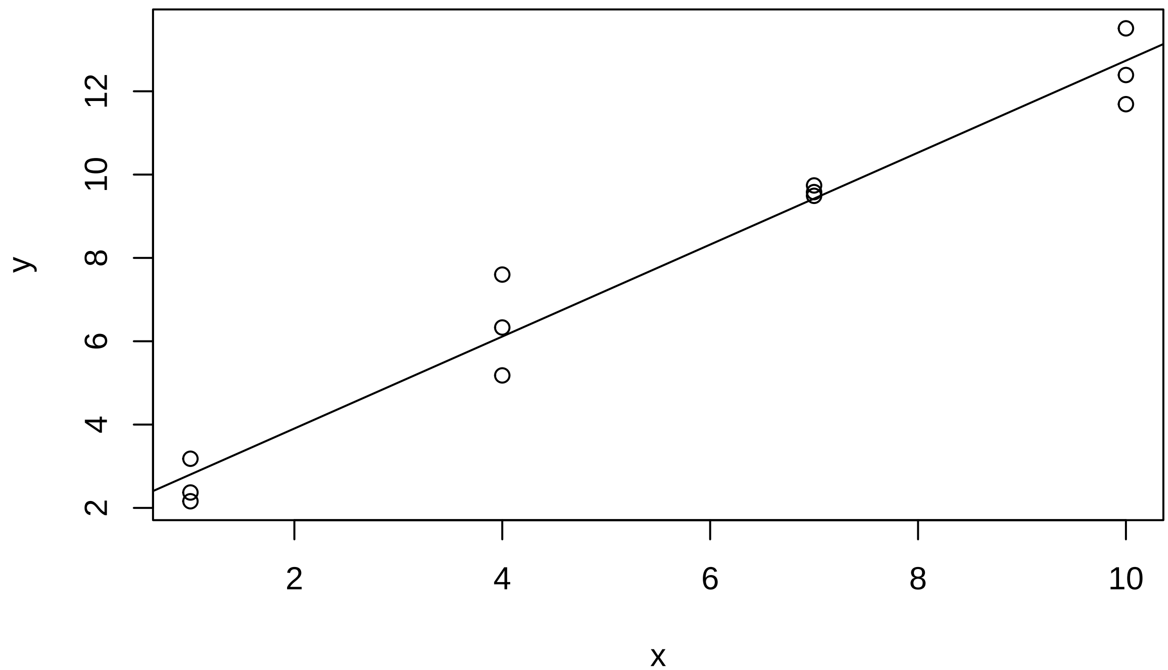
```
# Estatística de Teste:
estat_teste2 <- unname((SQLoF2/2)/(SQEp2/8))

# Valor do quantil 0.95 da distribuição F:
(f_2 <- qf(0.95, 2, 8))
```

```
## [1] 4.45897
```

```
# Estatística de Teste > Valor de F?
estat_teste2 > f_2
```

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

Desta vez, temos que o teste não rejeita a hipótese nula, de modo que todas as esperanças condicionais nos valores únicos de x são modeladas pela relação explicitada. Podemos ver isso com o gráfico do modelo:



3 Questão 3

3.1 Pergunta

Considerando o mesmo conjunto de dados da questão 2:

- (a) Calcule a Soma de Quadrados dos Resíduos considerando o modelo $Y_i = \beta_1 x_i + \epsilon_i$.
- (b) Calcule a Soma de Quadrados de Erro Puro.
- (c) Calcule a Soma de Quadrados da Falta de Ajuste ("Lack of Fit").
- (d) Faça um teste para determinar se o modelo linear sem intercepto é apropriado.

3.2 Resposta

a)

Podemos atualizar o modelo especificado anteriormente com o seguinte código:

```
mod3 <- lm(y ~ x - 1, data = df2)

(SQR3 <- sum(residuals(mod3)^2))
```

```
## [1] 15.18483
```

b)

Como a Soma de Quadrados do Erro Puro não depende do modelo, ela será igual à do modelo 2:

```
(SQEp3 <- SQEp2)
```

```
##     SQEp
## 5.228467
```

c)

Diferentemente ao efetuado no item **(b)**, temos que a Soma de Quadrados da Falta de Ajuste depende do modelo, e será dada por:

```
(SQLoF3 <- df2 %>%
     cbind(data.frame(y_pred = predict(mod3, df2))) %>%
     group_by(x) %>%
     summarise(y_bar = mean(y),
               y_pred = mean(y_pred)) %>%
     mutate(LOF = y_pred - y_bar) %>%
     summarise(SQLoF = round(3 * sum(LOF^2), 4)) %>%
     unlist())
```

```
##  SQLoF
## 9.9564
```

d)

Para verificar se o modelo linear é apropriado para esse conjunto de dados, usaremos o teste de falta de ajuste, dado por:

$$\frac{\text{SQLoF}/(m-2)}{\text{SQEp}/(n-m)} \sim F_{(m-2,n-m)}$$

Em que $n = 12$ é a quantidade de observações, $m = 4$ é a quantidade de observações distintas de x e $F_{(m-2,n-m)}$ é a distribuição F de Snedecor, com $(4-2, 12-4)$ graus de liberdade. Assim, se considerarmos um nível de significância de 95% para o teste, teremos:

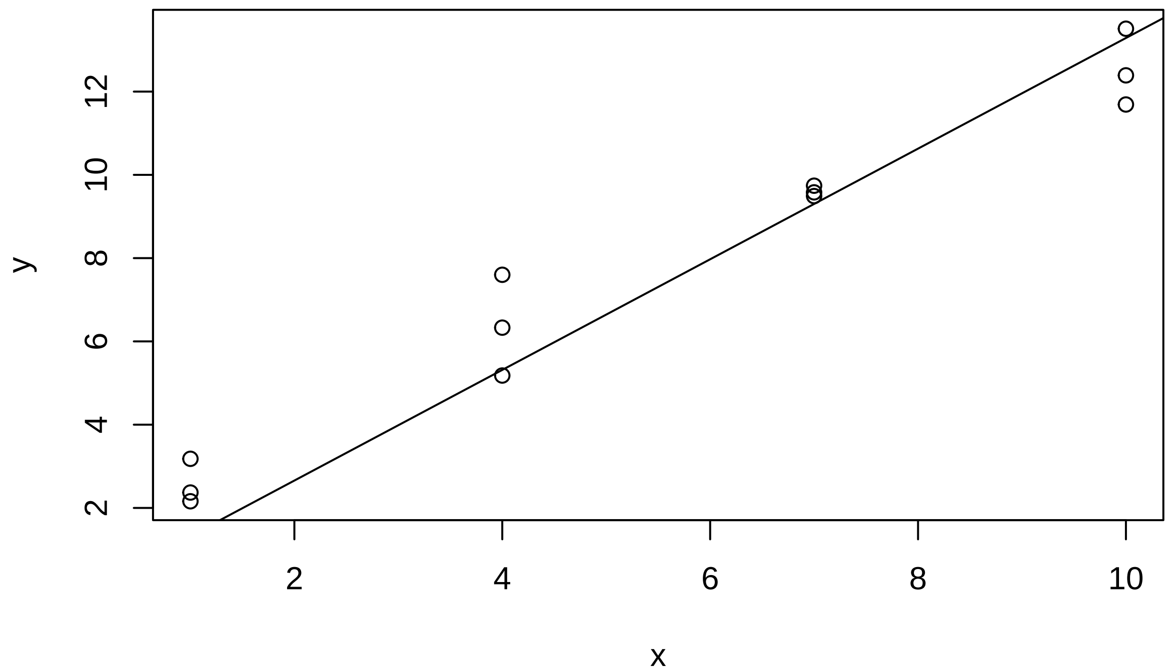
```
# Estatística de Teste:
estat_teste3 <- unname((SQLoF3/2)/(SQEp3/8))

# Valor do quantil 0.95 da distribuição F:
(f_3 <- qf(0.95, 2, 8))
```

```
## [1] 4.45897
```

```
# Estatística de Teste > Valor de F?
estat_teste3 > f_3
```

```
## [1] TRUE
```

Desta vez, temos que o teste rejeita a hipótese nula, de modo que alguma das esperanças condicionais nos valores únicos de x não é modelada pela relação explicitada. Podemos ver isso com o gráfico do modelo:



4 Questão 4

4.1 Pergunta

Um banco de dados contém informações de área e valor sobre 3 tipos de imóveis: Apartamentos, Casas e Terrenos. Defina dois modelos de regressão para determinação do valor dos imóveis de acordo com o tipo e a área. Em um deles o incremento do valor com respeito à área deve ser o mesmo para os 3 tipos, enquanto no outro, cada tipo de imóvel pode ter um incremento de valor em função da área distinto.

Interprete todos os parâmetros desses dois modelos.

4.2 Resposta

Sejam Y_i o valor do imóvel i, x_i a sua área e z_i o tipo de imóvel, podendo tomar valores $z \in \{0, 1, 2\}$, referente aos tipos Apartamentos, Casas e Terrenos, respectivamente. Para o primeiro tipo de modelo, em que o incremento de em valor em função da área é igual para os três tipos, podemos definir como:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbb{I}(z_i = 1) + \beta_3 \mathbb{I}(z_i = 2) + \epsilon_i$$

Aqui, β_1 representa o comportamento do aumento no valor do imóvel com relação à área, que é comum a todos os tipos de imóvel. β_0 representa o valor do intercepto quando o tipo de imóvel for Apartamentos, pois

$\mathbb{I}(z_i = 1) = 0$ e $\mathbb{I}(z_i = 2) = 0$. Por outro lado, β_2 e β_3 representam a diferença no intercepto para os tipos de imóveis Casas e Terrenos com relação ao tipo Apartamentos. Esse modelo representa 3 retas paralelas, com diferentes valores de intercepto.

Por outro lado, se considerarmos um modelo em que cada tipo de imóvel possui um incremento diferente, teremos:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 x_i + \beta_2 \mathbb{I}(z_i = 1) + \beta_3 \mathbb{I}(z_i = 2) + \\ & \beta_4 x_i \mathbb{I}(z_i = 1) + \beta_5 x_i \mathbb{I}(z_i = 2) + \epsilon_i \end{aligned}$$

Neste caso, teremos que β_0, β_2 e β_3 terão as mesmas interpretações que no caso anterior. Porém, β_1 agora representa o crescimento do valor em função da área para imóveis do tipo Apartamentos, enquanto que β_4 representa o crescimento do valor em função da área para imóveis do tipo Casas e β_5 representa o crescimento do valor em função da área para imóveis do tipo Terrenos. Assim, este modelo representa três retas com diferentes valores de intercepto e diferentes inclinações, dependendo do tipo do imóvel.

5 Questão 5

5.1 Pergunta

Sejam $x_i \in \mathbb{R}$ covariáveis com valores contínuos e z_i covariáveis com valores $z_i \in \{0, 1\}$. Considere os seguintes modelos de regressão:

$$Y_i = \beta_{a,0} + \beta_{a,1} x_i + \epsilon_i$$

$$Y_i = \beta_{b,0} + \beta_{b,1} x_i + \beta_{b,2} z_i + \epsilon_i$$

- (a) Explique, se existir, em que cenário os coeficientes dos estimadores $\hat{\beta}_{a,1}$ e $\hat{\beta}_{b,1}$ podem ter sinais diferentes.
- (b) O que podemos dizer sobre a igualdade $\hat{\beta}_{a,2} = \hat{\beta}_{b,2}$? Em que cenários, se existirem, essa igualdade é verdadeira? Interprete.
- (c) (Extra, opcional) Simule um banco de dados com base no modelo $Y_i = \beta_{b,0} + \beta_{b,1} x_i + \beta_{b,2} z_i + \epsilon_i$, de forma que o teste de falta de ajuste não rejeite a hipótese de que o modelo $Y_i = \beta_{a,0} + \beta_{a,1} x_i + \epsilon_i$ é apropriado. Interprete esse resultado.

5.2 Resposta

a)

Veja que, como $z \in \{0, 1\}$, podemos especificar o segundo modelo da seguinte forma:

$$Y_i = \begin{cases} \beta_{b,0} + \beta_{b,1} x_i + \epsilon_i & \text{, se } z_i = 0 \\ (\beta_{b,0} + \beta_{b,2}) + \beta_{b,1} x_i + \epsilon_i & \text{, se } z_i = 1 \end{cases}$$

Desse modo, podemos ver que o modelo representa duas retas paralelas, em que uma é deslocada $\beta_{b,2}$ caso $z = 1$. Isso pode representar o caso em que uma covariável seja binária, o que representaria um modelo que acomoda o comportamento das duas categorias (0 e 1) conjuntamente. A fim de ilustrar o caso, seja a seguinte base de dados de exemplo:

x	z	y
1.00	0	5.69
1.36	0	6.18
1.71	0	6.37
2.07	0	7.95
2.43	0	8.55
2.79	0	7.02
3.14	0	8.10
3.50	0	9.10
4.00	1	4.46
4.36	1	2.93
4.71	1	3.69
5.07	1	5.02
5.43	1	5.14
5.79	1	6.99
6.14	1	6.45
6.50	1	6.82

Podemos com ele ajustar os dois modelos, considerando os modelos

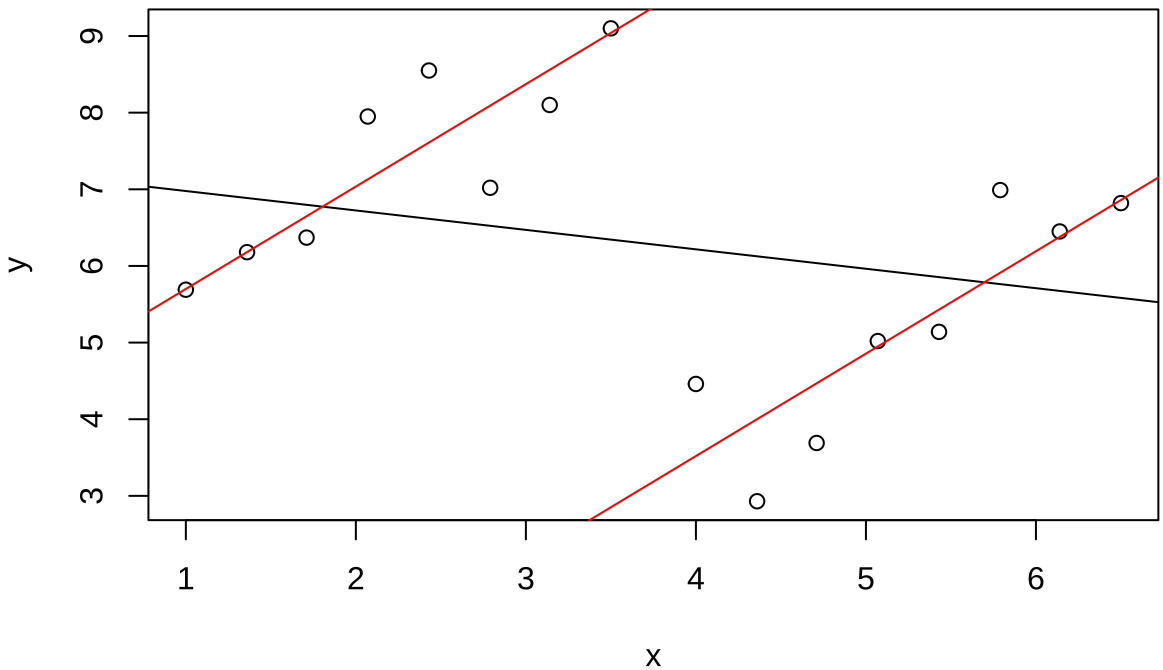
```
# Primeiro modelo:
(mod_exemplo_1 <- lm(y ~ x, data = df_exemplo))
```

```
##
## Call:
## lm(formula = y ~ x, data = df_exemplo)
##
## Coefficients:
## (Intercept)            x
##      7.2319      -0.2542
```

```
# Segundo modelo:
(mod_exemplo_2 <- lm(y ~ x + z, data = df_exemplo))
```

```
##
## Call:
## lm(formula = y ~ x + z, data = df_exemplo)
##
## Coefficients:
## (Intercept)            x            z
##       4.364        1.336       -6.191
```

Podemos ver que, não levando em consideração a variável z, as estimativas do β_1 tem os sinais invertidos. Vejamos graficamente como se comportam os dados e os modelos ajustados (o primreiro estará em preto, e o segundo estará em vermelho):



Essa inclusão de variáveis que separam os dados em conjuntos com comportamentos internos que são o contrário do comportamento "global" cria o chamado Paradoxo de Simpson, e é um problema recorrente de regressão.

b)

A igualdade ocorre quando a inclusão dessa variável z que "separa" os grupos nos dados não altera de maneira significativa a estimação do efeito "global" de $\beta_{\cdot,2}$. Assim, espera-se que os grupos possuam comportamento semelhante e que estejam distribuídos sem distinção ao longo do espaço da variável.

Veja que no exemplo de demonstração, ambos os grupos possuem o mesmo crescimento, mas eles possuem uma clara separação ao longo de x, de modo que ocore a inversão no sinal da estimação de β_1.