Slides Semana 11

Correlação e Regressão Linear Simples

Problema

Focaremos nas observações referentes a 116 alunos que obtiveram, no máximo, 6.25 pontos nas atividades de uma disciplina.

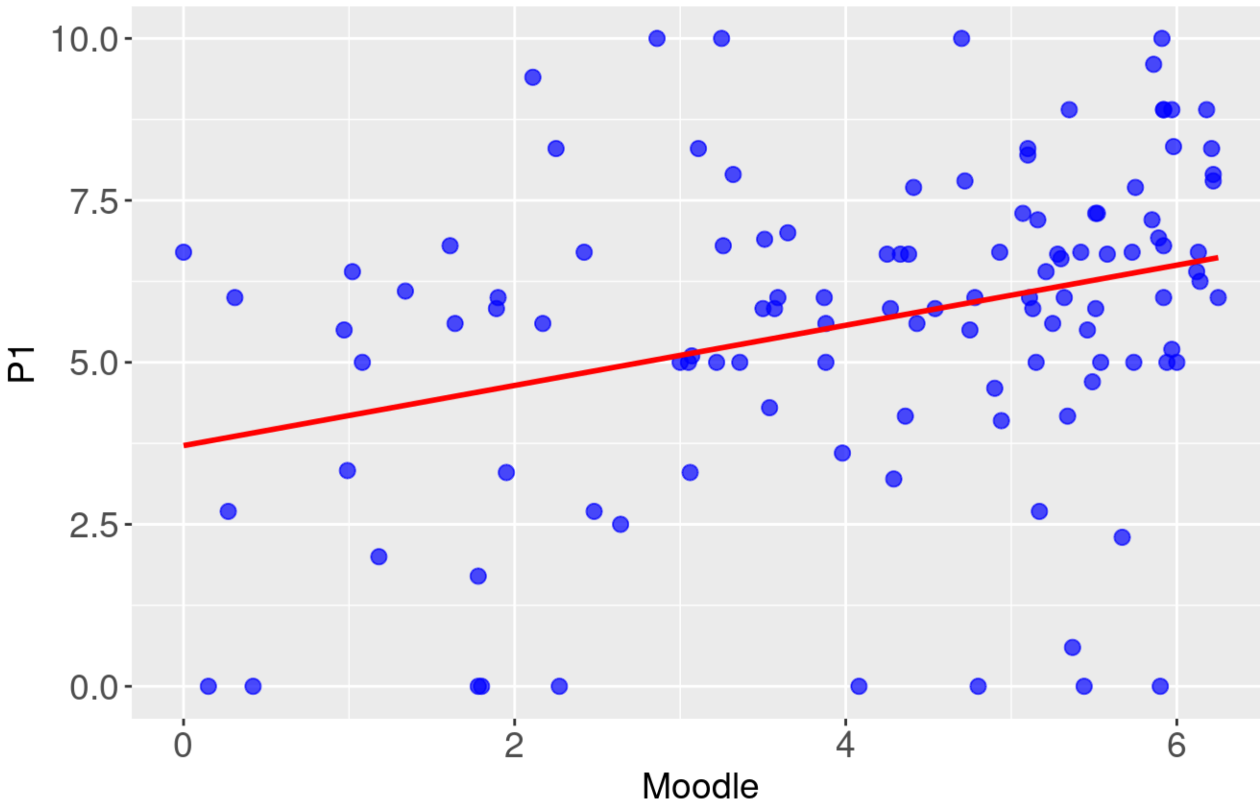
Nosso objetivo é inferir a respeito da associação das notas (absolutas) das atividades disponibilizadas com aquelas de uma Prova.

Atividade (≤ 6.25) e Notas da Prova

Moodle	P1
5.98	8.33
3.00	5.00
2.42	6.70
2.11	9.40
3.88	5.00
2.86	10.00

Como explicar essa associação?

```
## `geom_smooth()` using formula 'y ~ x'
```



Explicando Associação Linear

- **Coeficiente de Correlação**
  - Quantidade no intervalo $(-1, 1)$
  - Mede a força da associação linear em função da dispersão dos dados

- **Modelo de regressão linear simples**
  - Estima a forma $Y = \hat{\alpha} + \hat{\beta}X$;
  - O modelo é linear **nos parâmetros**

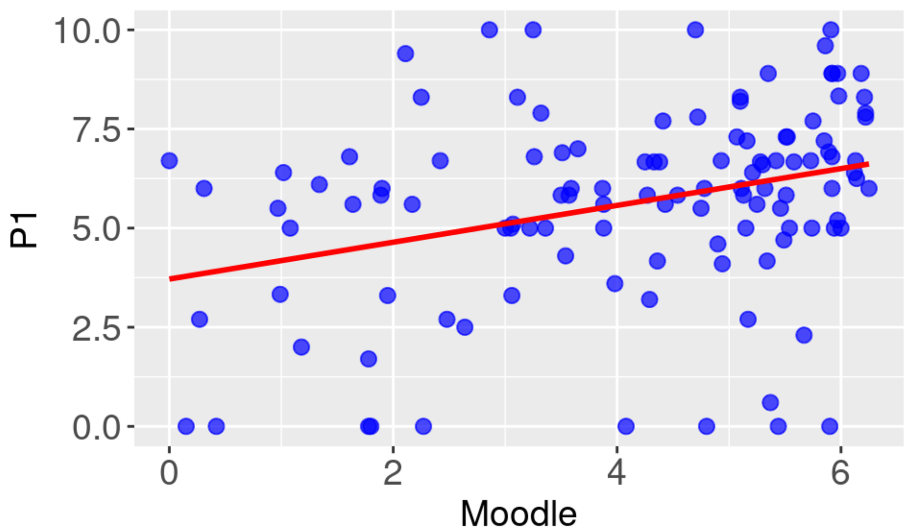
Coeficiente de Correlação

Introdução ao Coeficiente de Correlação

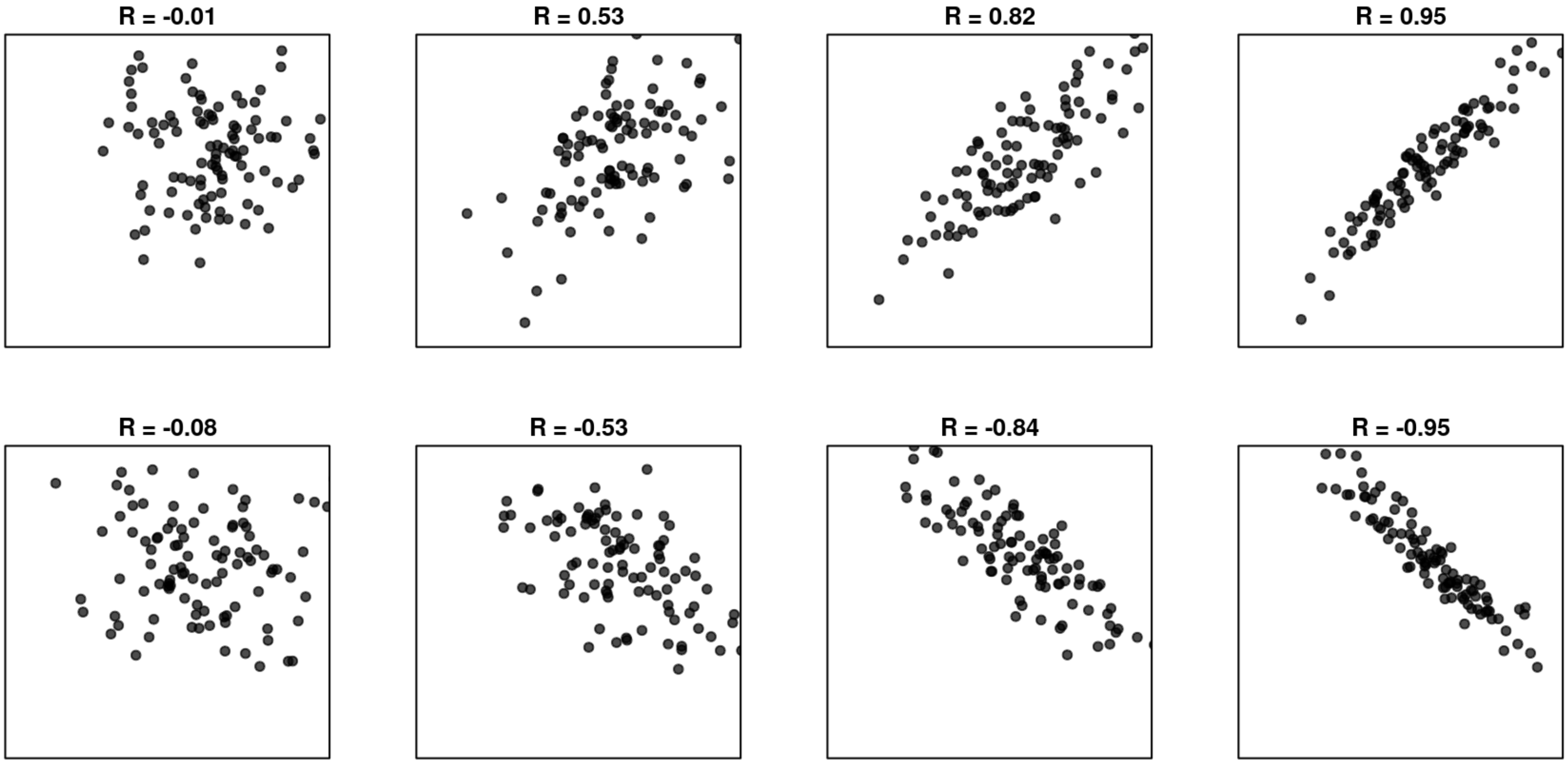
Denotamos a correlação por R.

- $R = -1$: associação linear negativa entre X e Y;
- $R = 0$: ausência de associação linear entre X e Y;
- $R = +1$: associação linear positiva entre X e Y;

```
## `geom_smooth()` using formula 'y ~ x'
```



Diferentes níveis de correlação



Determinação do Coeficiente de Correlação

Hipóteses:

- Duas variáveis contínuas: X e Y;
- n pares de observações: (X_i, Y_i);

Fórmula 1

$$
\begin{aligned}
R &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \\
&= \frac{S_{XY}}{\sqrt{S_{XX}^2 S_{YY}^2}} = \frac{S_{XY}}{S_{XX}S_{YY}}
\end{aligned}
$$

Notem que S_{XX}^2 e S_{YY}^2 são as somas de quadrados de X e Y corrigida por suas respectivas médias.

No exemplo das notas da P1 e Moodle:

$$S_{XY} = 157.99, \qquad S_{XX} = 18.45, \qquad S_{XY} = 26.41$$

Portanto, $R = 0.3243.$

Fórmula 2

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Notem que s_X e s_Y representam os desvios padrão amostrais de X e Y, respectivamente.

$$\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) = 37.29 \qquad \text{e} \qquad n - 1 = 115$$

Portanto, $R = 0.3243$.

Fórmula 3

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{1}{n-1}\,\frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{s_X s_Y}$$

Observem que $\bar{X}$, $\bar{Y}$, s_X e s_Y representam, respectivamente, as médias amostrais e desvios padrão amostrais de cada uma das variáveis.

$$\bar{X} = 4.14 \qquad \bar{Y} = 5.64 \qquad n - 1 = 115 \qquad \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{s_X s_Y} = 37.29$$

Portanto, $R = 0.3243$.

Regressão Linear Simples

Terminologia em Regressão Linear Simples

Um modelo de regressão possui, pelo menos, duas variáveis:

- X: variável independente, variável exploratória, variável preditora, covariável.
- Y: variável dependente, variável resposta.

Para alunos com notas de atividades de no máximo 6.25, como as notas das atividades se associam com a nota da prova P1?

- Variável dependente (resposta) Y: nota da prova P1
- Variável independente X: nota das atividades do Moodle

Forma do Modelo

O modelo de regressão usual descreve associação linear entre Y e X da seguinte forma:

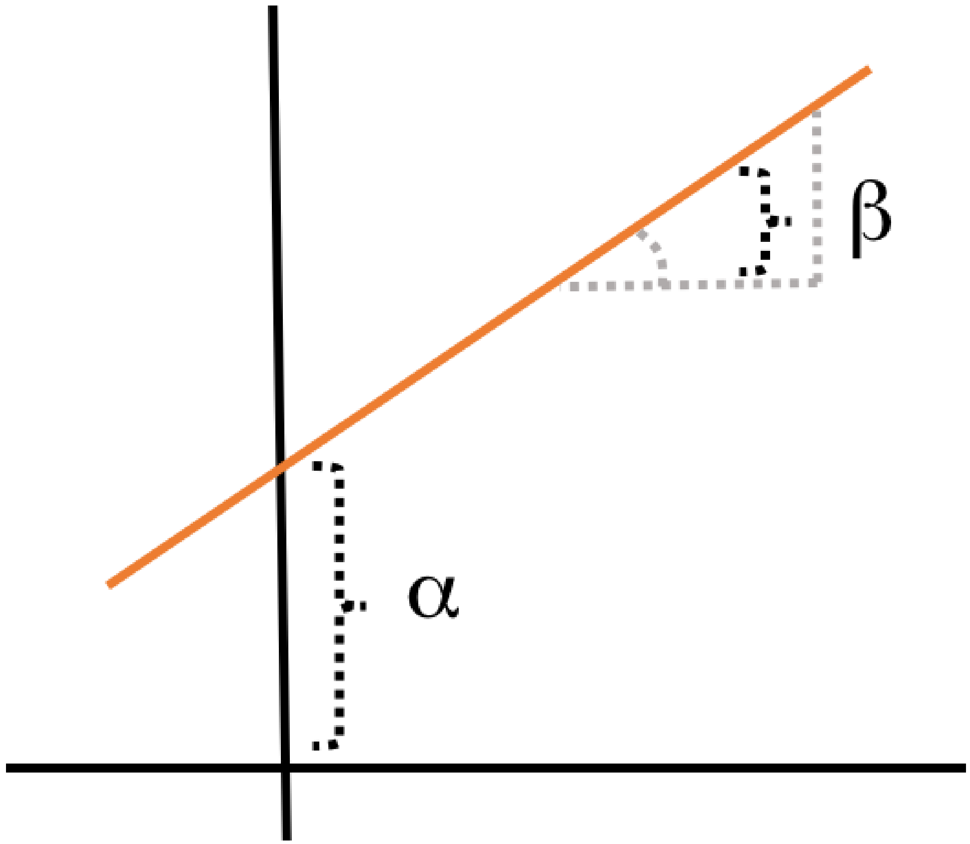
$$Y = \alpha + \beta X + \varepsilon.$$

Neste modelo, os termos adicionais são:

- α: intercepto
- β: coeficiente angular
- ε: erro observacional

Considerar o erro é necessário, pois associações perfeitas são improváveis.

Forma do Modelo



Hipóteses do Modelo de Regressão Linear

Modelo de regressão linear assume:

- Linearidade entre variáveis;
- Erros aleatórios independentes nas observações;
- Erro tem média zero;
- Variância constante do erro σ^2;

Desta forma, a variável aleatória Y, escrita como

$$Y = \alpha + \beta X + \varepsilon,$$

possui as seguintes características:

- $\mathbb{E}(Y) = \mathbb{E}(\alpha + \beta X + \varepsilon) = \alpha + \beta X + \mathbb{E}(\varepsilon) = \alpha + \beta X$
- $\mathrm{Var}(Y) = \mathrm{Var}(\alpha + \beta X + \varepsilon) = \mathrm{Var}(\varepsilon) = \sigma^2$

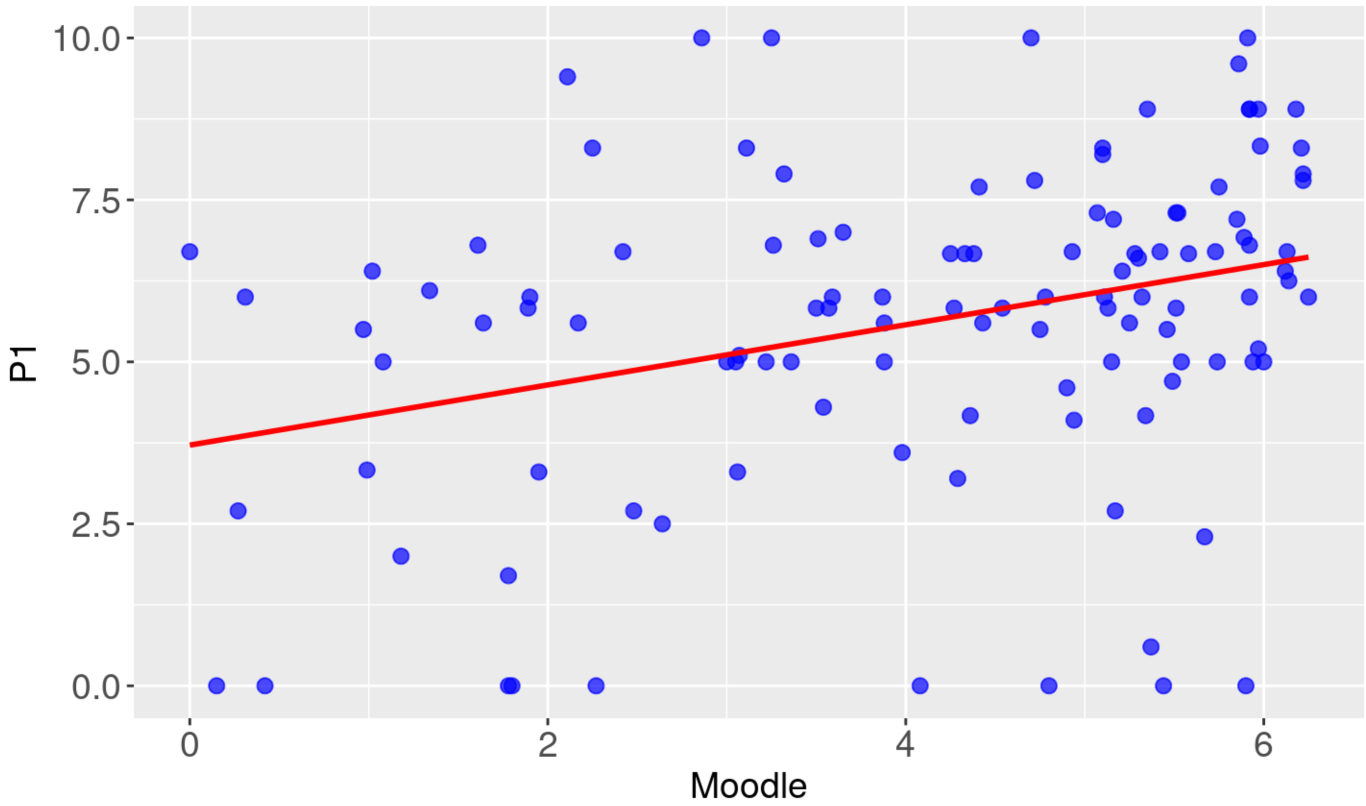
Como saber se a regressão linear é adequada?

- Utilizar diagramas de dispersão;
- Analisar visualmente a forma de associação entre X e Y;
- Buscar associação linear;
- Aferir a homogeneidade da variância;
- Buscar informação sobre independência das observações.

Exemplo: Notas

Voltando no exemplo das notas dpara 116 alunos.

```
## `geom_smooth()` using formula 'y ~ x'
```



Escolha da Melhor Reta

Um modo de determinar a melhor reta é escolhendo os parâmetros de forma que a distância entre os pontos e a reta seja mínimo, ou seja, pelo método conhecido como mínimos quadrados:

- Determinar a função a ser minimizada;
- Determinar a primeira derivada com respeito aos parâmetros de interesse;
- Igualar estas derivadas a zero;
- Verificar segundas derivadas.

Escolha da Melhor Reta

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

A função a ser minimizada é a soma de quadrados dos erros:

$$f(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

Tomando as derivadas em relação a α e β e igualando-as a zero temos:

$$\frac{\partial f(\alpha, \beta)}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) \qquad \frac{\partial f(\alpha, \beta)}{\partial \beta} = -2 \sum_{i=1}^{n} X_i (Y_i - \alpha - \beta X_i)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \qquad \text{e} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Exemplo: Notas

Para esses dados, calculou-se:

$$\bar{X} = 4.14, \qquad \bar{Y} = 5.64, \qquad S_{XY} = 157.99 \qquad \text{e} \qquad S_{XX} = 340.33.$$

Então, as estimativas dos coeficientes são:

$$\begin{aligned}
\hat{\beta} &= \frac{S_{XY}}{S_{XX}} = \frac{157.99}{340.33} = 0.46 \\
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
&= 5.64 - 0.46 \times 4.14 = 3.72
\end{aligned}$$

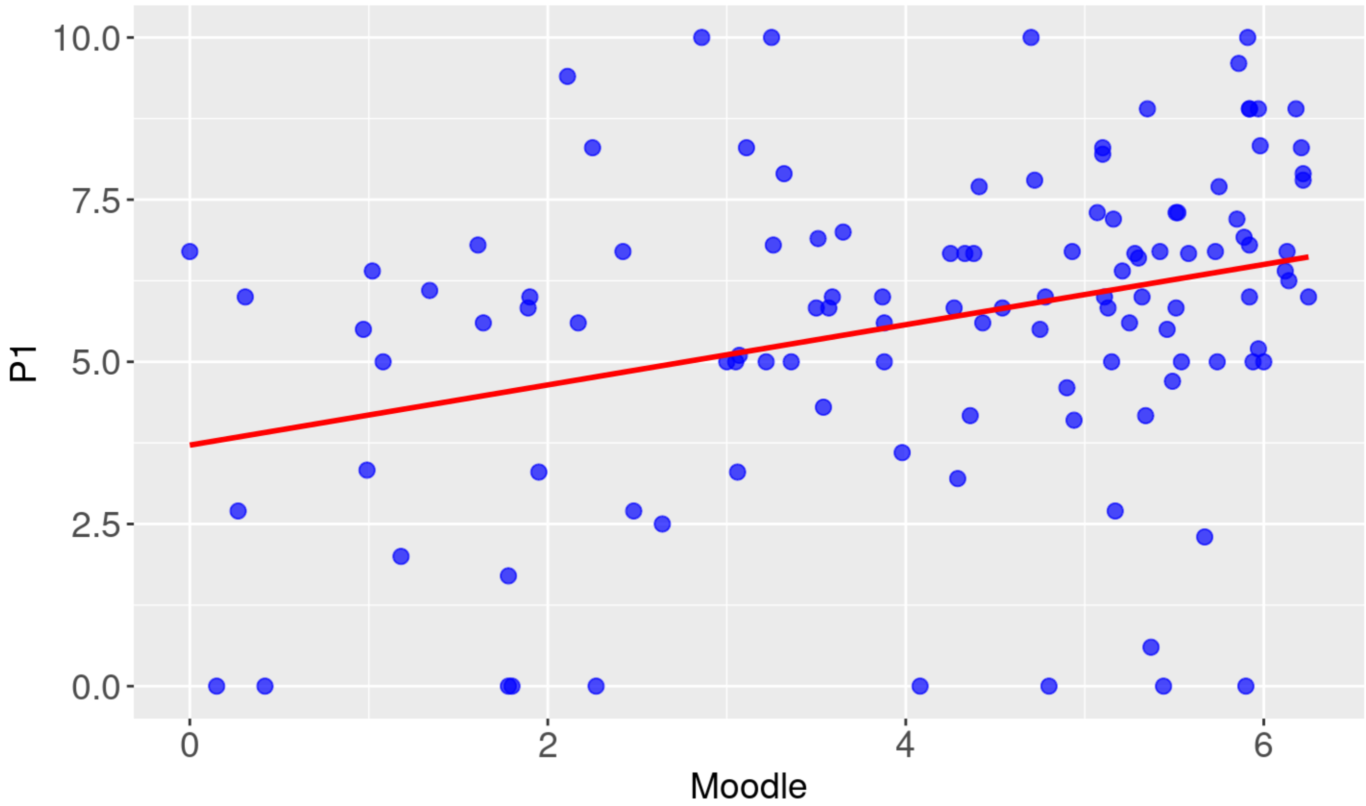
E a equação da reta estimada é dada por:

$$\text{P1} = 3.72 + 0.46 \times \text{Moodle}.$$

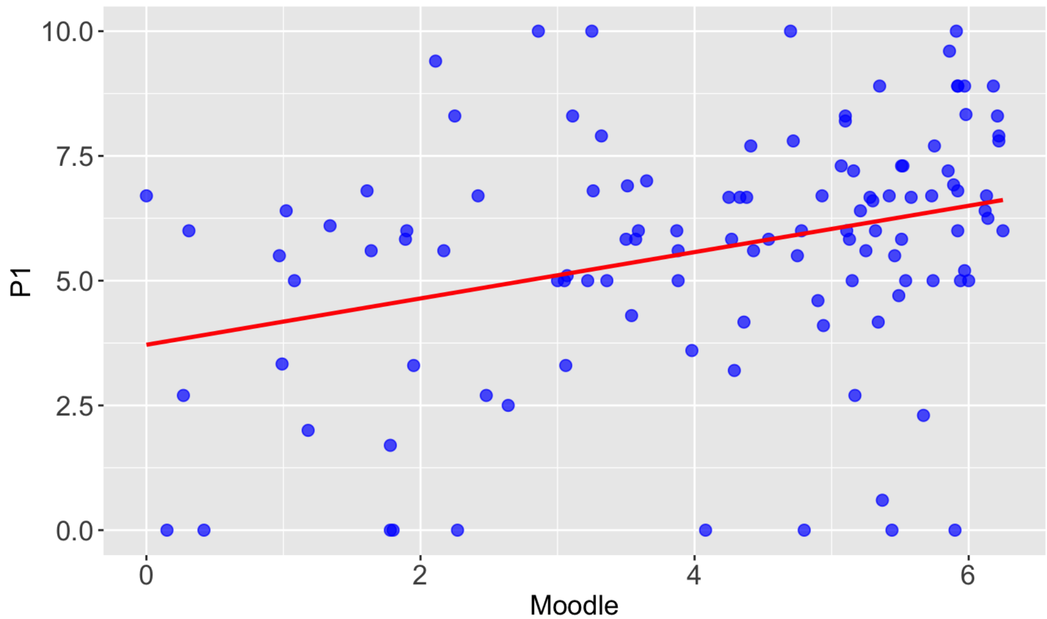
A Escolha da Melhor Reta

$$\text{P1} = 3.72 + 0.46 \times \text{Moodle}$$

```
## `geom_smooth()` using formula 'y ~ x'
```



Interpretação dos Parâmetros



$$\text{P1} = 3.72 + 0.46 \times \text{Moodle}$$

$\hat{\alpha} = 3.72$ é a nota média na prova para alunos com nota 0 na atividade **(intercepto)**.

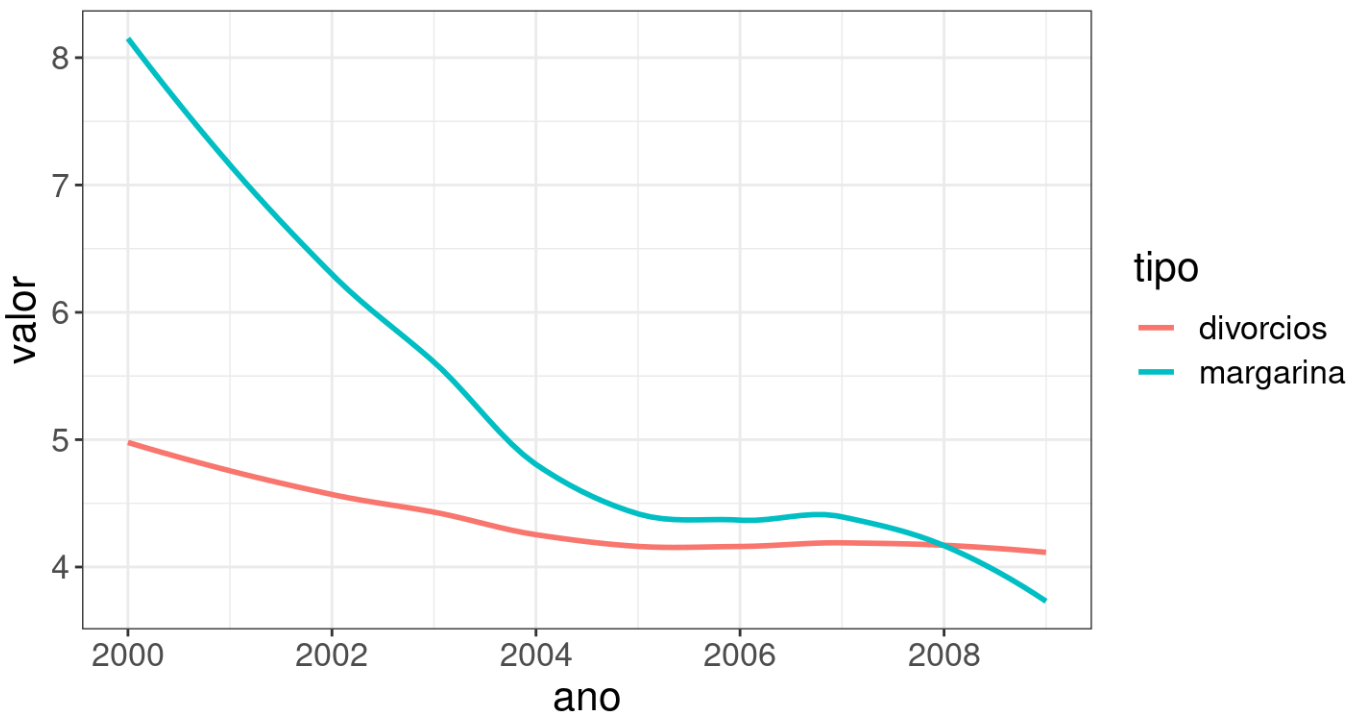
$\hat{\beta} = 0.46$ é o aumento médio na nota da Prova para cada ponto extra na atividade **(coeficiente angular)**.

Erros na Interpretação de Correlação e Regressão

- Correlação e regressão apresentam **associação**!
- Associação não indica causalidade!!!
- Extrapolações não devem ser feitas.

Associações

O gráfico abaixo apresenta o número de divórcios (por 1000 casamentos) no Maine/EUA e o consumo *per capita* de margarina (em libras) ao longo dos anos.



Associações

A correlação entre estas duas variáveis (número de divórcios e consumo de margarina) é 0.9926.

Considere o número de divórcios como variável resposta e o consumo de margarina como variável independente.

Temos o seguinte modelo de regressão linear:

	Estimativa	Erro Padrão	valor t	valor-de-p
(Intercept)	3.308626	0.0480316	68.88431	0
margarina	0.201386	0.0087350	23.05495	0

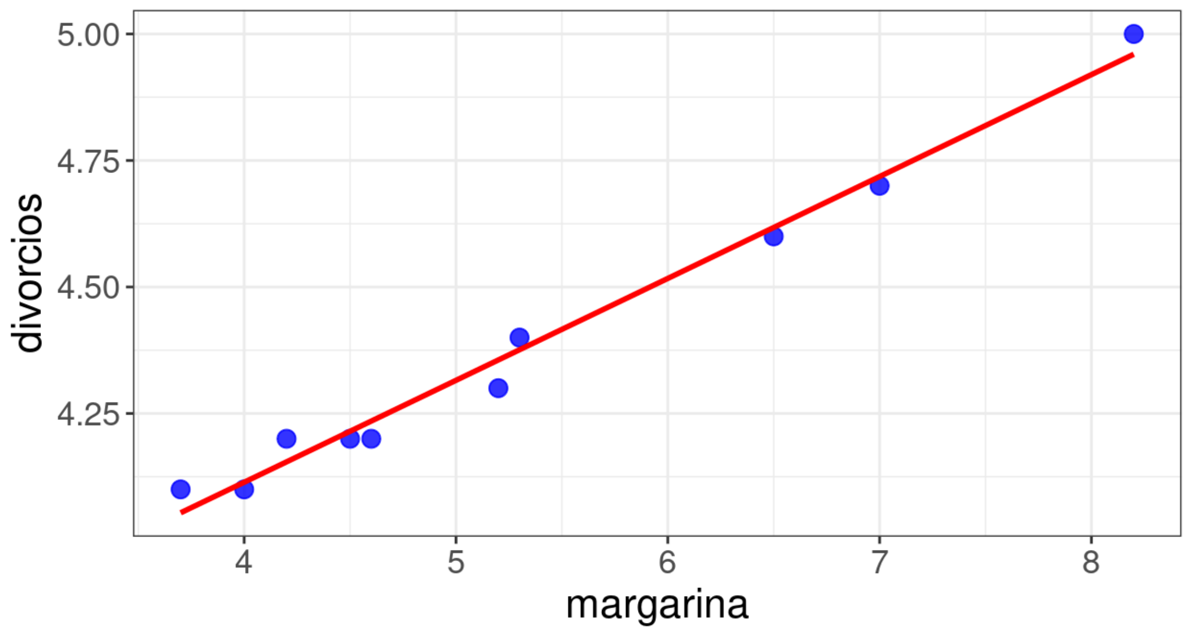
Ou seja,

$$\text{divórcios} = 3.30 + 0.20 \times \text{margarina}$$

Associações

$$\text{divórcios} = 3.30 + 0.20 \times \text{margarina}$$

```
## `geom_smooth()` using formula 'y ~ x'
```

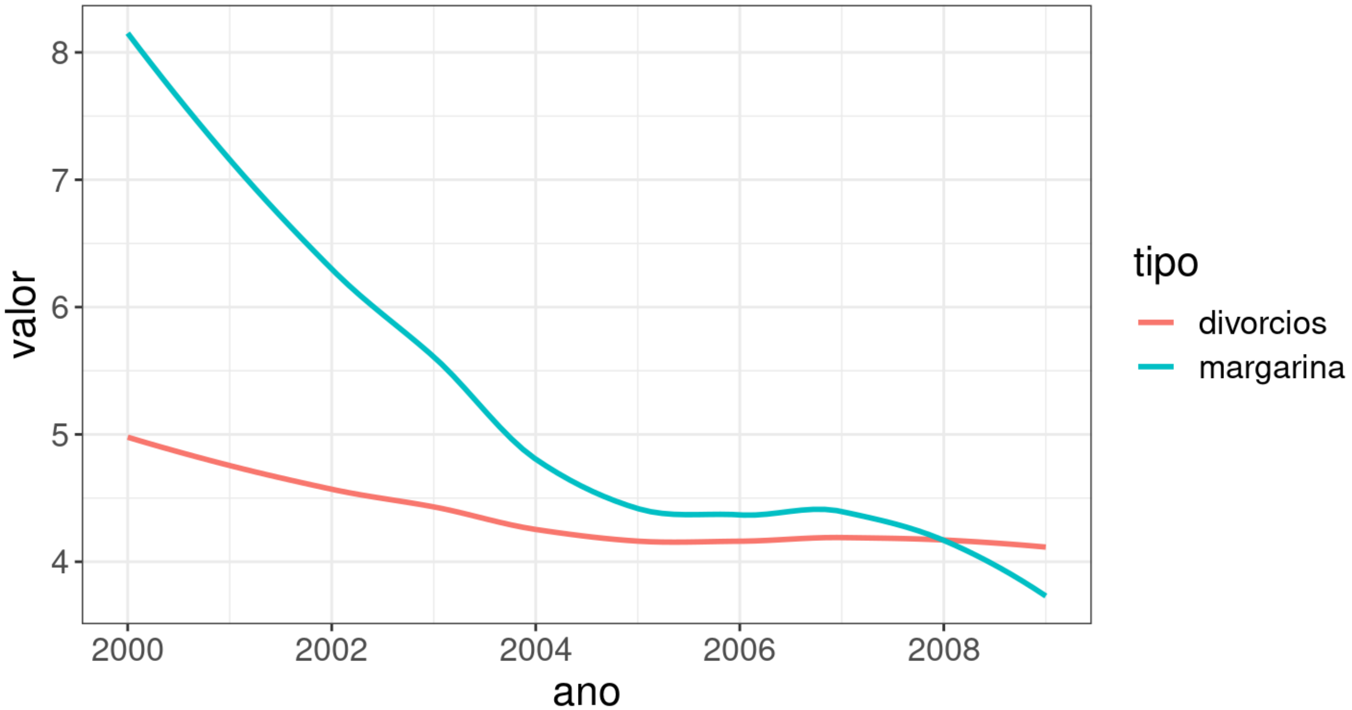


Importante: modelos de regressão descrevem **associação**, não causalidade.

Extrapolações

Qual o consumo esperado de margarina em 2016?

```
## `geom_smooth()` using formula 'y ~ x'
```



Extrapolações não devem ser feitas!!!