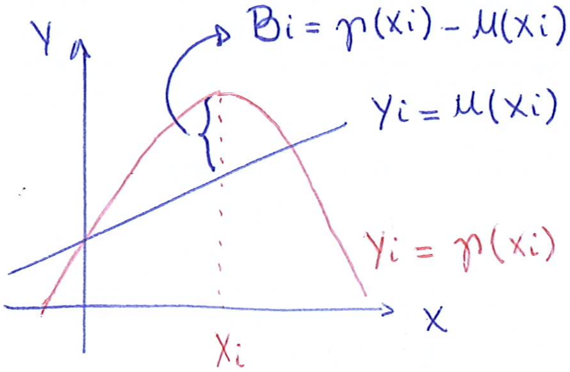
# TESTE PARA FALTA DE AJUSTE

Vimos que $QMRes = \dfrac{SQRes}{n-2} = \dfrac{\sum \hat{\varepsilon}_i^2}{n-2} = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}$ é uma estimativa não viesada da variância residual $\sigma^2$, sob a suposição de que: "o modelo ajustado é correto"

MODELO PROPOSTO: $y_i = \mu(x_i) + \varepsilon_i$

MODELO CORRETO: $y_i = \gamma(x_i) + \varepsilon_i^*$ $\begin{cases} E(\varepsilon_i^*) = 0 \\ Var(\varepsilon_i^*) = \sigma^2 \end{cases}$

$B_i = \gamma(x_i) - \mu(x_i) \Rightarrow$ VIÉS




↳ Esse viés estará incluído em $\varepsilon_i$

Desta forma,

$E(\varepsilon_i) = \underline{B_i}$ $\quad E(\varepsilon_i^* + B_i) = B_i$

$Var(\varepsilon_i) = E(\varepsilon_i^2) = E[(\varepsilon_i^* + B_i)^2] =$

$= E[\varepsilon_i^{*2} + 2\varepsilon_i^* B_i + B_i^2] = E(\varepsilon_i^{*2}) + 2B_i E($

$+ E(B_i^2)$

$= \underline{\sigma^2 + B_i^2}$

Proposta: Obter uma estimativa da variância residual ($\sigma$ que independa do modelo que está sendo ajustado $\Rightarrow$ planejando a coleta de observações repetidas de Y para cada nível de X distinto.

Considerar K níveis para variável X e $n_i$ repetições de Y em cada nível de X

| | X | Y | | | Totais | Médias |
|---|---|---|---|---|---|---|
| K níveis | $X_1$ | $y_{11}$ | $y_{12}$ | $\cdots$ $y_{1n_1}$ | $y_{1\cdot}$ | $\bar{y}_1$ |
| | $X_2$ | $y_{21}$ | $y_{22}$ | $\cdots$ $y_{2n_2}$ | $y_{2\cdot}$ | $\bar{y}_2$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $X_K$ | $y_{K1}$ | $y_{K2}$ | $\cdots$ $y_{Kn_K}$ | $y_{K\cdot}$ | $\bar{y}_K$ |

Essa outra estimativa para $\sigma^2$, que independe do modelo, é dada pelo QMErro de uma análise de variância em que os níveis da variável x são vistos como k tratamentos em um delineamento experimental (DIC, DBC, DQL)

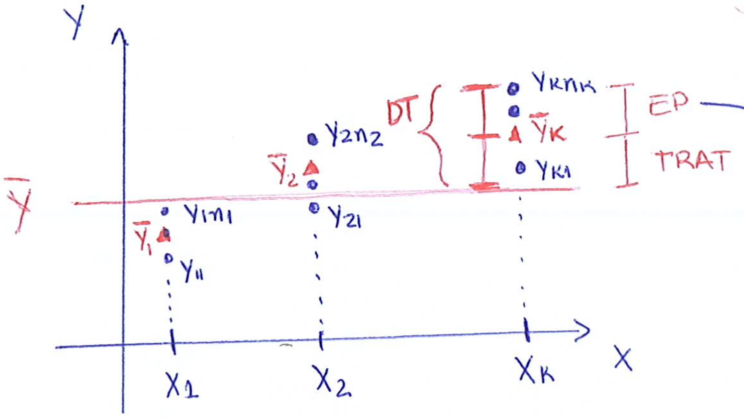
Considerando um DIC

$$Y_{ij} = \underset{\downarrow \; \bar{Y}_i}{\mu_i} + \varepsilon_{ij} \qquad \begin{array}{l} i = 1, \dots, k \\ j = 1, \dots, n_i \end{array}$$

ANAVA

| FV | GL |
|---|---|
| Tratamentos | $k-1$ |
| Erro | $N-k$ |
| Total | $N-1$ |

$N = \sum n_i = n_1 + \dots + n_k$




▲ $Y_{ij} = \bar{Y}_i + \varepsilon_{ij}$

$\varepsilon_{ij} = Y_{ij} - \bar{Y}_i$

$$\underset{\substack{\text{DESVIO}\\ \text{ERRO}\\ \text{PURO}}}{Y_{ij} - \bar{Y}_i} = \underset{\substack{\text{DESVIO}\\ \text{TOTAL}}}{(Y_{ij} - \bar{Y})} - \underset{\substack{\text{DESVIO}\\ \text{TRAT}}}{(\bar{Y}_i - \bar{Y})}$$

Considerando $i = 1, \dots, k$ e $j = 1, \dots, n_i$, temos:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij} - \bar{Y}_i)^2 + 2(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) + (\bar{Y}_i - \bar{Y})^2\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + 2\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})\overset{0}{\sum_{j=1}^{n_i}(\underbrace{Y_{ij}}_{Y_{i\cdot}} - \underbrace{\bar{Y}_i}_{n_i \frac{Y_i}{n_i}})} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2$$

$$\Downarrow$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2$$

$$\Downarrow$$

SQTotal = SQErroPuro + SQTratamento

$$SQTotal = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}^2 - 2y_{ij}\bar{y} + \bar{y}^2)$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}^2 - 2\bar{y}\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij} + N\bar{y}^2$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}^2 - 2\frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right)}{N}\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right) + \cancel{N}\frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N^{\cancel{2}}}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N}$$

TOTAL GERAL $y_{oo}$

$$SQTrat = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{y}_i^2 - 2\bar{y}_i\bar{y} + \bar{y}^2)$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\bar{y}_i^2 - 2\bar{y}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\bar{y}_i + N\bar{y}^2$$

$$= \sum_{i=1}^{K}\cancel{n_i}\frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i^{\cancel{2}}} - 2\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}}{N}\sum_{i=1}^{K}\cancel{n_i}\frac{\sum_{j=1}^{n_i} y_{ij}}{\cancel{n_i}} + \cancel{N}\frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N^{\cancel{2}}}$$

$$= \sum_{i=1}^{K}\frac{y_{i\circ}^2}{n_i} - \frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N}$$

SQErroPuro = SQTotal − SQTrat

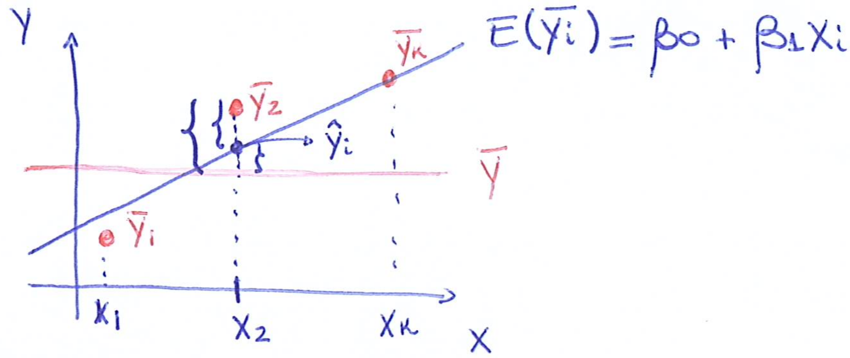
↳ $\hat{\sigma}^2$

Após obter uma estimativa para $\sigma^2$, o interesse está em verificar se existe uma relação linear entre as médias dos tratamentos ($\bar{y}_i$) e os níveis da variável X, ou seja, vamos desdobrar os K−1 graus de liberdade de TRAT em 1 grau de liberdade para RL e o restante para o que chamamos de DESVIO DE REGRESSÃO

O modelo estatístico para média de tratamentos é

$$\bar{Y}_i = \beta_0 + \beta_1 X_i + \varepsilon_i \longrightarrow E(\bar{Y}_i) = \beta_0 + \beta_1 X_i$$
$$\longrightarrow \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Voltando aos desvios




Para um dado $X_i$

$$\underbrace{\bar{Y}_i - \bar{Y}}_{\substack{\text{Desvio}\\ \text{entre os}\\ \text{níveis de X}}} = \underbrace{(\bar{Y}_i - \hat{Y}_i)}_{\substack{\text{falta de}\\ \text{ajuste}}} + \underbrace{(\hat{Y}_i - \bar{Y})}_{\substack{\text{efeito do}\\ \text{modelo}\\ \text{linear}}}$$

Considerando $i = 1, \dots, K$ e $j = 1, \dots, n_i$, temos:

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\bar{Y}_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right]^2$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\bar{Y}_i - \hat{Y}_i)^2 + 2(\bar{Y}_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2\right]$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{K}\sum_{j=1}^{n_i}\underbrace{(\bar{Y}_i - \hat{Y}_i)}_{\substack{\text{DESVIOS}\\ =0}}(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\hat{Y}_i - \bar{Y})^2$$

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{Y}_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\hat{Y}_i - \bar{Y})^2$$

$$\Downarrow \qquad\qquad \Downarrow$$

SQTrat = SQFaltaAj + SQRegressão

$$SQReg = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{K} n_i(\hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{K} n_i(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$

$$= \sum_{i=1}^{K} n_i(\cancel{\bar{y}} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}})^2$$

$$= \sum_{i=1}^{K} n_i[\hat{\beta}_1(x_i - \bar{x})]^2$$

$$= \sum_{i=1}^{K} n_i \hat{\beta}_1^2 x_i^2$$

$$= \hat{\beta}_1^2 \sum_{i=1}^{K} n_i x_i^2$$

$$= \frac{(\sum_{i=1}^{K} n_i x_i \bar{Y}_i)^2}{(\sum_{i=1}^{K} n_i x_i^2)^{\cancel{2}}} \sum_{i=1}^{K} \cancel{n_i x_i^2}$$

$$= \frac{(\sum_{i=1}^{K} n_i x_i \bar{Y}_i)^2}{\sum_{i=1}^{K} n_i x_i^2}$$

VERIFIQUEM

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{K} n_i x_i \bar{Y}_i}{\sum_{i=1}^{K} n_i x_i^2}$$

$$SQTrat = SQFalta\,Aj + SQReg$$

$$SQFalta\,Aj = SQTrat - SQReg$$

Com esse desdobramento, a Anava fica

| FV | GL | SQ | QM | Fc |
|---|---|---|---|---|
| Tratamentos | $K-1$ | SQTrat | QMTrat | QMTrat / QMEP |
| Reg linear | 1 | SQReg | QMReg | QMReg / QMEP |
| Falta Ajuste | $K-2$ | SQF.A. | QMF.A | QMF.A / QMEP |
| Erro Puro | $N-K$ | SQEP | QMEP | |
| Total | $N-1$ | SQTotal | | |

Para Tratamentos: $H_0: \mu_1 = \mu_2 = \dots = \mu_K = 0$ vs $H_a$: pelo menos $\mu_i \neq 0$

$$F_c = \frac{QMTrat}{QMEP} \sim F_{K-1;\,N-K}$$

Rejeitamos Ho, a um nível p de significância, se $F_c > F_{tab}$. Isso significa que pelo menos uma média de tratamento difere das demais.

Para Regressão: $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$$F_c = \frac{QMReg}{QMEP} \sim F_{1,\,N-K}$$

Rejeitamos Ho, a um nível p de significância, se $F_c > F_{tab}$. Isso significa que há evidência da tendência linear entre as variáveis X e Y.

Para Falta de Ajuste: $H_0: \mu(x) = \beta_0 + \beta_1 X$ vs $H_a: \mu(x) \neq \beta_0 + \beta_1 X$

$$F_c = \frac{QMF.A}{QMEP} \sim F_{K-2;\,N-K}$$

Rejeitamos Ho, a um nível p de significância, se $F_c > F_{tab}$. Isso significa que existem evidências de que o modelo linear não é adequado, havendo necessidade de se procurar outro modelo.

4 Casos:

Caso 1: { Falta Ajuste → não sig; Regressão → não sig } → Modelo: $\hat{Y}_{ij} = \hat{\beta}_0 = \bar{Y}$

Caso 2: { Falta Ajuste → não sig; Regressão → signif. } → Modelo: $\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Caso 3: { Falta Ajuste → signif.; Regressão → não sig. } → Modelo: $\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$ ou maior

Caso 4: { Falta Ajuste → signif.; Regressão → signif. } → Modelo: $\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$ ou maior