Regressão em R: exemplos

Renato Martins Assunção

DCC - UFMG

2015

Fitting a regression model in R

- Slides de Masanao Yajima, UCLA
- To fit a linear regression model in R you use the `lm()` function

  ```
  <fit object> = lm( <outcome> ~ <predictor 1> + ... + <predictor p> )
  ```

- To look at the fitted linear regression model we will use the `summary()` function: `summary( <fit object > )`
- You can also directly obtain the fitted value, residual, and estimated coefficient(s) by

  ```
  fitted( <fit object > )
  resid( <fit object > )
  coef( <fit object > )
  ```

- Adding a regression line in a plot is simple also, you first plot the outcome vs predictor then call `abline( <fit object > )`.

Fitting a regression model in R

- You can also directly obtain the fitted value, residual, and estimated coefficient(s) by

  ```
  fitted( <fit object > )
  resid( <fit object > )
  coef( <fit object > )
  ```

- Adding a regression line in a plot is simple also, you first plot the outcome vs predictor then call

  ```
  abline( <fit object > )
  ```

Um exemplo

- Cognitive test scores of three- and four-year-old children and characteristics of their mothers.
- Survey of adult American women and their children (subsample of National Longitudinal Survey of Youth).
- Data and code are from Data Analysis Using Regression and Multilevel/Hierarchical Models by Gelman and Hill 2007

```
> kidiq <- read.table("kidiq.txt", header=T)
> attach(kidiq)
> kidiq
  kid.score mom.hs    mom.iq mom.work mom.age
1        65      1 121.11750        4      27
2        98      1  89.36188        4      25
3        85      1 115.44320        4      27
4        83      1  99.44964        3      25
5       115      1  92.74571        4      27
6        98      0 107.90180        1      18
> dim(kidiq)
[1] 434   5
```

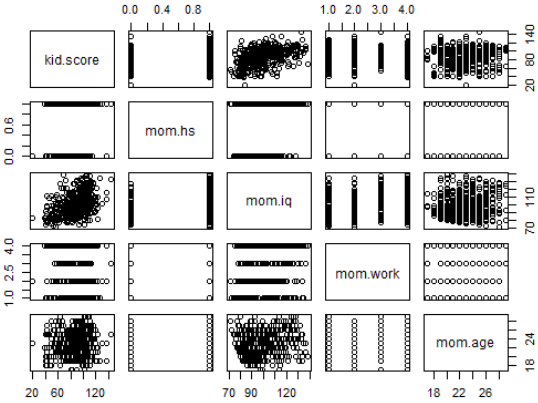
Um exemplo

- The variables:

```
> kidiq
  kid.score mom.hs    mom.iq mom.work mom.age
1        65      1 121.11750        4      27
2        98      1  89.36188        4      25
...
5       115      1  92.74571        4      27
6        98      0 107.90180        1      18
```

- `kid.score`: resultado do test de QI na criança de 3 ou 4 anos
- `mom.hs`: binária, mãe completou ou não o secundário (high school)
- `mom.work`: categórica: 1: mother did not work in first three years of child's life; 2: mother worked in second or third year of child's life; 3: mother worked part-time in first year of child's life; 4: mother worked full-time in first year of child's life.
- `mom.age`: mother's age at the time she gave birth

Visualizando...



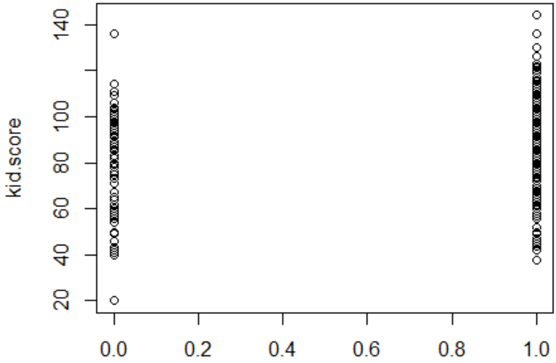
Linear Regression With One Binary Predictor

- Let's fit our first regression model.
- We will start with a simple model, then gradually build on it.
- As an illustrative purpose, we will start with a binary variable indicating whether mother graduated from high school or not
- Variable `mom.hs` is the predictor.

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \text{error}$$

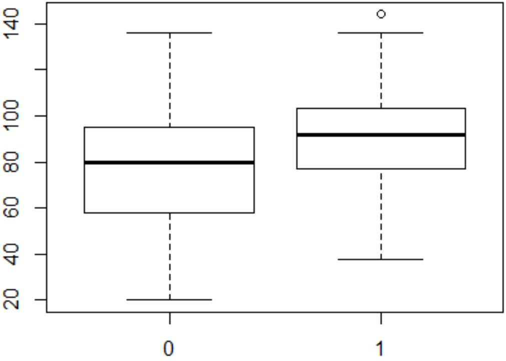
Visualizando...

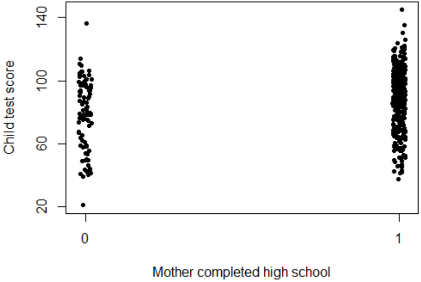
-
- `plot(mom.hs, kid.score)`



Visualizando: boxplot

- `plot(mom.hs, kid.score)`





```
plot(jitter(mom.hs, amount=0.02),
     jitter(kid.score, amount=1),
     xlab="Mother completed high school",
     ylab="Child test score",pch=20,
     xaxt="n", yaxt="n")
axis (1, seq(0,1))
axis (2, c(20,60,100,140))
## abline(fit.0)
```

Linear Regression With One Binary Predictor

- Since `mom.hs` is a binary predictor, the single model

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \text{error}$$

can be broken into two:

  - When mom.hs=0 we have

$$\text{kid.score} = \beta_0 + \beta_{hs} \ * \ 0 + \text{error} = \beta_0 + \text{error}$$

  - When mom.hs=1 we have

$$\text{kid.score} = \beta_0 + \beta_{hs} \ * \ 1 + \text{error} = \beta_0 + \beta_{hs} + \text{error}$$

- Since $\mathbb{E}(\text{error}) = 0$, we have
  - When mom.hs=0,

$$\mathbb{E}(\text{kid.score}) = \mathbb{E}(\beta_0 + \text{error}) = \beta_0 + \mathbb{E}(\text{error}) = \beta_0$$

  - When mom.hs=1,

$$\mathbb{E}(\text{kid.score}) = \beta_0 + \beta_{hs} + \mathbb{E}(\text{error}) = \beta_0 + \beta_{hs}$$

Linear Regression With One Binary Predictor

- The interpretation of the coefficients is clear now.
- $\beta_0 = \mathbb{E}(\text{kid.score}|\text{mom.hs} = 0)$, it is the expected value of the kid QI when the mother did not complete high school.
- $\beta_0 + \beta_{hs} = \mathbb{E}(\text{kid.score}|\text{mom.hs} = 1)$, the same for moms completing high school
- Therefore, $\beta_{hs} = \mathbb{E}(\text{kid.score}|\text{mom.hs} = 1) - \mathbb{E}(\text{kid.score}|\text{mom.hs} = 0)$
- That is, β_{hs} is the increment (positive or negative) that completing high school provides to the child QI.
- It is the expected effect on `kid.score` of changing from `mom.hs=0` to `mom.hs=1`.

Interpretação dos coeficientes de regressão

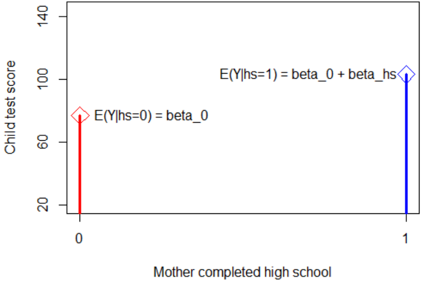


Figura: Modelo com um único preditor binário. Note que β_{hs} é a DIFERENÇA entre as alturas das duas barras representando $\mathbb{E}(\text{kid.score}|\text{mom.hs} = 0)$ e $\mathbb{E}(\text{kid.score}|\text{mom.hs} = 0)$

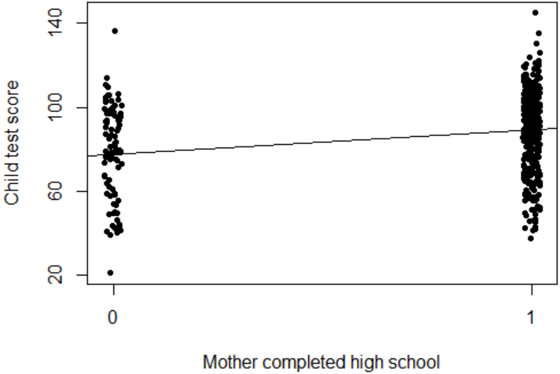
Linear Regression With One Binary Predictor

- Here is how to fit the model in R.

```
> fit.0 <- lm ( kid.score ˜ mom.hs )
> summary( fit.0 )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.548      2.059  37.670  < 2e-16 ***
mom.hs        11.771      2.322   5.069 5.96e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 19.85 on 432 degrees of freedom
Multiple R-squared:  0.05613, Adjusted R-squared:  0.05394
F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

- For now just look at the estimated coefficients $\hat{\beta}_0 = 77.55$ and $\hat{\beta}_{hs} = 11.77$


Child test score
20
60
100
140
0
1
Mother completed high school

- Now that we have a model let's try to interpret it.

$$\text{kid.}\hat{\text{s}}\text{core} = 77.55 + 11.77\ \text{mom.hs}$$

- $\hat{\beta_{hs}}$ is the estimated change on `kid.score` when `mom.hs` changes from 0 to 1.
- For a mother with high school education `mom.hs`
    - NO high school education (`mom.hs = 0`): expected IQ of a child is about 78.
    - WITH high school education (`mom.hs = 1`): expected IQ of a child is about 89.
- It looks like children whose mothers completed high school do better on this test.
- The regression coefficient β_{hs} for a SINGLE BINARY predictor is just the difference between the mean of the 2 groups.

```
# this is alpha
mean(kid.score[mom.hs==0])
[1] 77.54839
# this is beta
mean(kid.score[mom.hs==1]) - mean(kid.score[mom.hs==0])
[1] 11.77126
```

The matrix-version of the model

- Up to now, we have only one predictor, `mom.hs`, which is a binary predictor.
- In our matrix-version of linear regression this predictor is the second column in the design matrix $\mathbf{X}$:

$$\texttt{kid.score} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{hs} \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\bar{y}_1$ be the mean of the kid scores for the moms with mom.hs$_i = 1$ while $\bar{y}_0$ is the mean for those with mom.hs$_i = 0$
- We can show that the least squares solution is given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \bar{y}_0 \\ \bar{y}_1 - \bar{y}_0 \end{bmatrix}$$

Linear Regression With One Continuous Predictor

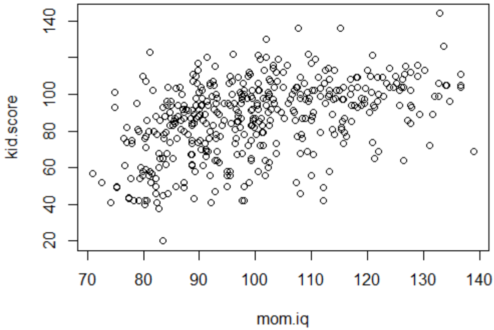


Figura: Kid IQ versus mom IQ. Correlation = 0.45. High IQ Moms tend to have high IQ kids but there is a LOT of variability and uncertainty.

The matrix representation

- Next we will model the child's IQ from mother's IQ score, a continuous predictor:

$$\text{kid.score} = \beta_0 + \beta_{iq}\ \text{mom.iq} + \text{error}$$

- In our matrix-version of linear regression the mother IQ is the second column in the design matrix **X**:

$$\texttt{kid.score} = \begin{pmatrix} \text{kid.score}_1 \\ \text{kid.score}_2 \\ \text{kid.score}_3 \\ \text{kid.score}_4 \\ \text{kid.score}_5 \\ \vdots \\ \text{kid.score}_n \end{pmatrix} = \begin{pmatrix} 1 & \text{mom.iq}_1 \\ 1 & \text{mom.iq}_2 \\ 1 & \text{mom.iq}_3 \\ 1 & \text{mom.iq}_4 \\ 1 & \text{mom.iq}_5 \\ \vdots & \vdots \\ 1 & \text{mom.iq}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{iq} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The matrix representation

- This example is a particular case of the general matrix representation of the linear regression model with a single predictor:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ 1 & x_{14} \\ 1 & x_{15} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- The second column in the design matrix **X** is any numerical attribute that is not constant in all rows.

Linear Regression With One Continuous Predictor

- We fit this single continuous predictor model the same way as we did for a single binary variable

```
> fit.1 <- lm (kid.score ˜ mom.iq)
> summary(fit.1)
Call:
lm(formula = kid.score ˜ mom.iq)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.79976    5.91741    4.36 1.63e-05 ***
mom.iq       0.60997    0.05852   10.42  < 2e-16 ***
...
```

Linear Regression With One Continuous Predictor

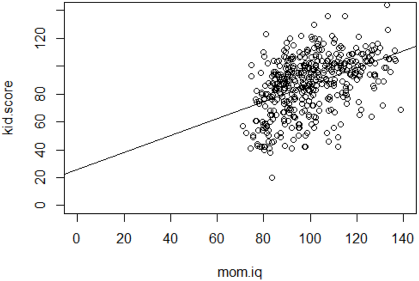
- Has the interpretation changed for the coefficients' model with a continuous variable?

$$\widehat{\text{kid.score}} = 25.80 + 0.61 \text{ mom.iq}$$

- Yes it has!
- Intercept β_0: expected IQ of a child for mother with IQ of 0 (that is, `mom.iq = 0`).
- Is this even possible?
- NO, very low IQ's tend to be between 60 and 70.
- The region `mom.iq = 0` is outside the region where the statistical data cloud lives. Just check the plot again (next slide).

Intercept: interpretation.

- The point $(0, \beta_0) = (0, 25.80)$ where the regression line crosses the vertical line is completely outside the data region.



- Therefore, the intercept β_0 does not have a clear empirical interpretation in this example as there is no mother with zero IQ.

Linear Regression With One Continuous Predictor

- The slope β_{iq} however is clearly interpretable.

$$\text{kid.}\hat{\text{s}}\text{core} = 25.80 + 0.61 \text{ mom.iq}$$

- Coefficient $\hat{\beta}_{iq}$: estimated expected increase in child's IQ with every unit increase in mother's IQ.
- Suppose a mom has her IQ increased from a certain value `mom.iq` to a new value `mom.iq* = mom.iq +1`.
- If we expect `kid.score` to increase to a new value `kid.score*`, what is this new value?

$$\begin{aligned}
\texttt{kid.score*} &= 25.80 + 0.61 \text{ mom.iq}^{*} \\
&= 25.80 + 0.61 \text{ mom.iq} + 1 \\
&= 25.80 + 0.61 \text{ mom.iq } + 0.61 \\
&= \texttt{old kid.score} + 0.61
\end{aligned}$$

- We can always expect an increase of 0.61 for each additional unit of mom's IQ, IT DOES NOT MATTER THE INITIAL VALUE OF mom.iq.

Linear Regression With One Continuous Predictor

- The impact of `mom.iq` in `kid.score` seems really small:
- Only an additional 0.61 in `kid.score` for an increase of 1 unit in `mom.iq` .
- But ... how different are two mothers with difference of just 1 point in their IQ's?
- The variation range of `mom.iq` goes from 70 up to 140.
- A variation of 1 point in this range is very small.
- If we take the whole range, a mom going from `mom.iq` =70 to `mom.iq` =140 increases her kid IQ, on average, in $0.61 \times 70 = 42.7$, a substantial amount.

Linear Regression With One Continuous Predictor

- Can we do better?
- Yes! if we center and scale the predictor variables.
    - centering: shifting the variable to a meaningful center so it is easier to interpret the coefficient(s)
    - scaling: re-scaling the variable to a meaningful unit
- Centering becomes more important when we add the interaction term (later).
- It also provides numerical stability when many predictors are used simultaneously (later)
- For the current case with mother's IQ
    - centering: we can center the mother's IQ by subtracting the mean IQ of the mothers
    - scaling: we can divide the mother's IQ by a unit that might be more meaningful (say, 10)

Linear Regression With One Continuous Predictor

- Given the attribute `mom.iq`, consider a new variable given by

$$\text{mom.iq.st} = \frac{\text{mom.iq} - \text{mean(mom.iq)}}{10}$$

- Suppose `mom.iq.st` is increased by 1 unit changing to `mom.iq.st*`.
- What this means in terms of the original variable `mom.iq`?

$$\begin{aligned}
\text{mom.iq.st}^* &= \text{mom.iq.st} + 1 \\
&= \frac{\text{mom.iq} - \text{mean(mom.iq)}}{10} + 1 \\
&= \frac{\text{mom.iq} - \text{mean(mom.iq)} + 10}{10} \\
&= \frac{\text{mom.iq}^* - \text{mean(mom.iq)}}{10}
\end{aligned}$$

- That is, changing `mom.iq.st` in 1 unit is equivalent to change the original variable `mom.iq` in 10 units.

Centering and scaling

- We will refit the model using the centered and scaled predictor variable.

```
> mom.iq.st = (mom.iq - mean(mom.iq))/10
> fit.2 <- lm (kid.score ~ mom.iq.st)
> summary(fit.2)
Call:
lm(formula = kid.score ~ mom.iq.st)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.7972     0.8768   98.99   <2e-16 ***
mom.iq.st     6.0997     0.5852   10.42   <2e-16 ***
...
```

- How has the interpretation changed?
  - Intercept: expected IQ of a child for mother with MEAN IQ (which is approx. 100)
  - Coefficient $\hat{\beta}_{iq}$: expected increase in child's IQ with every 10 points increase in mother's IQ.
- Much better, right? scaling and centering helps.

Increasing the number of predictors

- We will go a step further and combine both a continuous and a binary predictors

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \beta_{iq}\ \text{mom.iq} + \text{error}$$

- It turns out that this allows the regression line to have a different *intercept* depending on whether a child's mother completed high school.
- That is, we have two models:

$$\begin{aligned}
(\text{mom.hs} = 0) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs} \cdot 0 + \beta_{iq}\ \text{mom.iq} + \epsilon \\
&= \beta_0 + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 1) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs} \cdot 1 + \beta_{iq}\ \text{mom.iq} + \epsilon \\
&= (\beta_0 + \beta_{hs}) + \beta_{iq}\ \text{mom.iq} + \epsilon
\end{aligned}$$

Comments on the model

- Note that the difference between the model for `(mom.hs = 0)` and `(mom.hs = 1)` is only on the intercept.
- The expected effect of changing `mom.iq` is represented by β_{iq} and IT IS THE SAME VALUE in the two groups of `mom.hs`.
- The effect of changing `mom.hs` from 0 to 1 is to change the intercept from β_0 to $\beta_0 + \beta_{hs}$.
- Therefore, the coefficient $\beta_{hs}$is the increment (positive or negative) on the intercept of the initial model for `(mom.hs = 0)`.

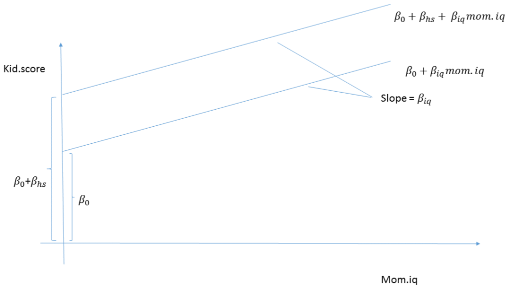

$\beta_0 + \beta_{hs} + \beta_{iq} mom.iq$
Kid.score
$\beta_0 + \beta_{iq} mom.iq$
Slope = β_{iq}
β_0+β_{hs}
β_0
Mom.iq

Figura: Modelo kid.score $= \beta_0 + \beta_{hs}$mom.hs $+ \beta_{iq}$ mom.iq $+$ error.

The matrix representation

- The matrix-version of linear regression has a design matrix $\mathbf{X}$ with three columns.
- The first is the constant 1, the second is the attribute `mom.hs` and the third is `mom.iq`.

$$\texttt{kid.score} = \begin{pmatrix} \text{kid.score}_1 \\ \text{kid.score}_2 \\ \text{kid.score}_3 \\ \text{kid.score}_4 \\ \text{kid.score}_5 \\ \vdots \\ \text{kid.score}_n \end{pmatrix} = \begin{pmatrix} 1 & \text{mom.hs}_1 & \text{mom.iq}_1 \\ 1 & \text{mom.hs}_2 & \text{mom.iq}_2 \\ 1 & \text{mom.hs}_3 & \text{mom.iq}_3 \\ 1 & \text{mom.hs}_4 & \text{mom.iq}_4 \\ 1 & \text{mom.hs}_5 & \text{mom.iq}_5 \\ \vdots & \vdots & \\ 1 & \text{mom.hs}_n & \text{mom.iq}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{hs} \\ \beta_{iq} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The matrix representation

- This example is a particular case of the general matrix representation of the linear regression model with two predictors:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Regression With a Continuous and a Binary Predictors

- We fit the model in the same way as before

```
> fit.3 <- lm(kid.score ~ mom.hs + mom.iq.st)
> summary(fit.3)
Call:
lm(formula = kid.score ~ mom.hs + mom.iq.st)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.1221     1.9437  42.250  < 2e-16 ***
mom.hs        5.9501     2.2118   2.690  0.00742 **
mom.iq.st     5.6391     0.6057   9.309  < 2e-16 ***
...
```

Regression With a Continuous and a Binary Predictors

- Here is the fitted model

$$\text{kid.}\hat{\text{score}} = 82.12 + 5.95\ \text{mom.hs} + 5.64\ \text{mom.iq.st}$$

- How has the interpretation changed?
  - Intercept β_0: expected IQ of a child for mother with mean IQ (that is, IQ $= 100$) that did NOT graduate high school
  - Coefficient β_{hs}: expected increase in child's IQ for a mother graduating high school (when `mom.hs` changes one unit from 0 to 1)
  - Coefficient β_{iq}: expected increase in child's IQ with every 10 points increase in mother's IQ (10 points due to the scaling)
- You may have noticed that although we fit only one regression model, because of the binary predictor we are actually fitting 2 PARALLEL regression lines, for each value of `mom.hs`.

As retas ajustadas

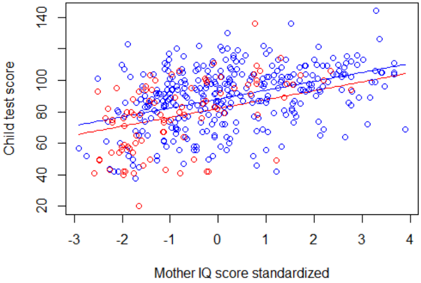


Figura: Regression line of child's IQ score for the mothers who graduated from high school (blue) and who did not graduate high school(red).

```
plot(mom.iq.st, kid.score,
     xlab="Mother IQ score standardized",
     ylab="Child test score", pch=20,
     type="n")
curve(coef(fit.3)[1] + coef(fit.3)[2]
       + coef(fit.3)[3]*x, add=TRUE,
       col="blue")
curve(coef(fit.3)[1] + coef(fit.3)[3]*x,
       add=TRUE,col="red")
points(mom.iq.st[mom.hs==1],
 kid.score[mom.hs==1], col="blue")
points(mom.iq.st[mom.hs==0],
 kid.score[mom.hs==0], col="red")
```

Continuous and a Binary Predictors + Interaction

- The previous model imposes the SAME slope for the two groups: (`mom.hs = 0`) and (`mom.hs = 1`).
- The next step is to fit a different line for each group allowing both, intercept and slope, to vary.
- We need to find a single representation of this two-lines model in terms of a single linear model.
- The most common representation uses the so-called interaction attribute, a variable derived from the product of the two variables, the binary `mom.hs` and the continuous `mom.iq`.
- That is, we create a new variable `mom.hs:mom.iq = mom.hs * mom.iq`

Continuous and a Binary Predictors + Interaction

- The new model is

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \beta_{iq}\ \text{mom.iq} + \beta_{hsiq}\ \text{mom.hs:mom.iq} + \text{error}$$

- Since `mom.hs:mom.iq = mom.hs * mom.iq`, we have the following two models:

$$\begin{aligned}
(\text{mom.hs} = 0) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs}\ \cdot\ 0 + \beta_{iq}\ \text{mom.iq} + \beta_{hsiq}\ \cdot\ 0 * \text{mom.iq} + \epsilon \\
&= \beta_0 + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 1) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs}\ \cdot\ 1 + \beta_{iq}\ \text{mom.iq} + \beta_{hsiq}\ \cdot\ 1 * \text{mom.iq} + \epsilon \\
&= (\beta_0 + \beta_{hs}) + (\beta_{iq} + \beta_{hsiq})\ \text{mom.iq} + \epsilon
\end{aligned}$$

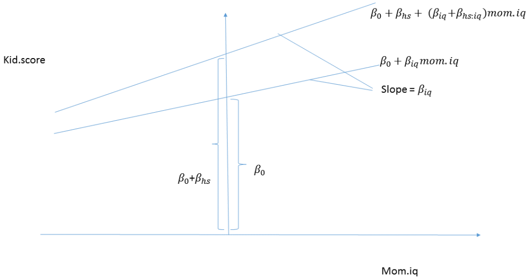

Kid.score
$\beta_0 + \beta_{hs} + (\beta_{iq}+\beta_{hs:iq})mom.iq$
$\beta_0 + \beta_{iq} mom.iq$
Slope = β_{iq}
β_0+β_{hs}
β_0
Mom.iq

Figura: Modelo
kid.score $= \beta_0 + \beta_{hs}$mom.hs $+ \beta_{iq}$ mom.iq $+ \beta_{hsiq}$ mom.hs:mom.iq $+$ error.

The matrix representation

- The matrix-version of linear regression has a design matrix **X** with four columns.
- The same three from the last model plus one additional column created by the simple element-wise product between the columns `mom.hs` and `mom.iq`.

$$\texttt{kid.score} = \begin{pmatrix} \text{kid.score}_1 \\ \text{kid.score}_2 \\ \text{kid.score}_3 \\ \text{kid.score}_4 \\ \text{kid.score}_5 \\ \vdots \\ \text{kid.score}_n \end{pmatrix} = \begin{pmatrix} 1 & \text{mom.hs}_1 & \text{mom.iq}_1 & \text{mom.hs}_1 * \text{mom.iq}_1 \\ 1 & \text{mom.hs}_2 & \text{mom.iq}_2 & \text{mom.hs}_2 * \text{mom.iq}_2 \\ 1 & \text{mom.hs}_3 & \text{mom.iq}_3 & \text{mom.hs}_3 * \text{mom.iq}_3 \\ 1 & \text{mom.hs}_4 & \text{mom.iq}_4 & \text{mom.hs}_4 * \text{mom.iq}_4 \\ 1 & \text{mom.hs}_5 & \text{mom.iq}_5 & \text{mom.hs}_5 * \text{mom.iq}_5 \\ \vdots & \vdots & & \\ 1 & \text{mom.hs}_n & \text{mom.iq}_n & \text{mom.hs}_n * \text{mom.iq}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{hs} \\ \beta_{iq} \\ \beta_{hsiq} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- In the more compact notation, we have

$$\texttt{kid.score} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The matrix representation

- This example is a particular case of the general matrix representation of the linear regression model with three predictors where one of them is the interaction (product) odthe other two:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11} * x_{12} \\ 1 & x_{21} & x_{22} & x_{21} * x_{22} \\ 1 & x_{31} & x_{32} & x_{31} * x_{32} \\ 1 & x_{41} & x_{42} & x_{41} * x_{42} \\ 1 & x_{51} & x_{52} & x_{51} * x_{52} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & x_{n1} * x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

..

- Rather than creating a columns with the product of the other two, the interaction is coded in R with ":"
- Later, with multi-category columns, it will be clear the advantages of using ":".

```
 > fit.4 = lm (kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st)
> summary(fit.4)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        85.407      2.218  38.502  < 2e-16 ***
mom.hs              2.841      2.427   1.171  0.24239
mom.iq.st           9.689      1.483   6.531 1.84e-10 ***
mom.hs:mom.iq.st   -4.843      1.622  -2.985  0.00299 **
...
```

Continuous and a Binary Predictors + Interaction

- Here is the fitted model

$$\text{kid.score} = 85.41 + 2.84\ \text{mom.hs} + 9.69\ \text{mom.iq.st} - 4.84\ \text{mom.hs} : \text{mom.iq}$$

- Intercept β_0: expected IQ of a child for mother with mean IQ that did NOT graduate high school
- Coefficient β_{hs}: expected increase in child's IQ for a mother graduating high school.
- Coefficient β_{iq}: expected increase in child's IQ with every 10 points increase in mother's IQ.
- Coefficient β_{hsiq}: difference of the β_{iq} for mothers who graduated high school and did not.
- Let's look closely at what that means.

Interpreting the results

- Since mom.hs only takes values 0 or 1, we have
  - For mothers who did not graduate high school

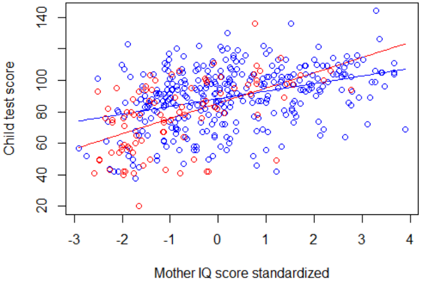
$$\hat{\text{kid.score}}_{hs=0} = 85.41 + 9.69 \text{ mom.iq}$$

  - For mothers who did graduate high school

$$\hat{\text{kid.score}}_{hs=1} = 85.41 + 2.84 + 9.69 \text{ mom.iq} - 4.84 \text{ mom.iq} = 88.25 + 4.85 \text{ mo}$$

- Based on the two intercepts: Mothers who did graduate from high school do have, on average, children with slightly higher IQ than mothers who did not graduate from high school, when evaluated at the mean IQ for the mothers.
- Based on the slope intercepts: However, the effect of mother's IQ is larger on the expected child IQ for mothers who did not graduate from high school.

As retas ajustadas



```
plot(mom.iq.st, kid.score,
 xlab="Mother IQ score standardized",
 ylab="Child test score", pch=20,
 type="n")
curve(coef(fit.4)[1] + coef(fit.4)[2]
 + (coef(fit.4)[3]+coef(fit.4)[4])*x,
 add=TRUE, col="blue")
curve(coef(fit.4)[1] + coef(fit.4)[3]*x,
 add=TRUE,col="red")
points(mom.iq.st[mom.hs==1],
 kid.score[mom.hs==1], col="blue")
points(mom.iq.st[mom.hs==0],
 kid.score[mom.hs==0], col="red")
```

Figura: Regression lines of child's IQ score for the mothers who graduated from high school (blue) and who did not graduate high school(red).

As retas ajustadas

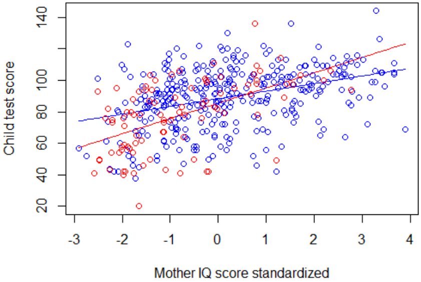


Figura: See anything strange here?

- The effect of mother's IQ is larger for `mom.hs=0`.
- But this effect seems to be so great that .... for moms with high IQ, the regression line for `mom.hs=0` IS ABOVE that for `mom.hs=0`!
- So...for high IQs moms, it is worth dropping from high school???
- *Extraordinary claims require extraordinary evidence.* (Carl Sagan)
- Let us look at this more closely.

As retas ajustadas

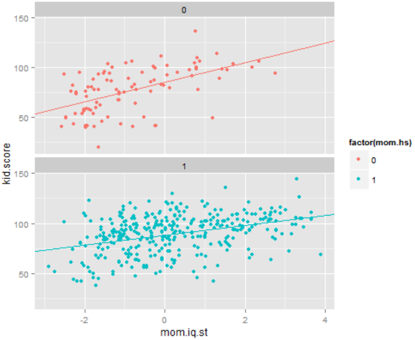


Figura: See anything strange here?

- Look at the `mom.hs=0` plot.
- There are very FEW moms with mom.iq.st $>$ 2 (only 3) and ZERO $>$ 3.
- This means that we are extrapolating beyond the data region when we are predicting what is the effect of mom.iq.st $>$ 2 for `mom.hs=0`.
- As we will see soon, the UNCERTAINTY of this prediction in this region will be large.
- There could be a saturation effect with the straight line curving down in these high IQ regions for mothers with no high school.
- We will return to this example in the next classes.

Regressão com variáveis categóricas

- Vamos expandir o modelo atual acrescentando a variável `mom.work`
- Esta é uma variável categórica com quatro valores:
    - `mom.work = 1`: mother did not work in first three years of child's life;
    - `mom.work = 2`: mother worked in second or third year of child's life;
    - `mom.work = 3`: mother worked part-time in first year of child's life;
    - `mom.work = 4`: mother worked full-time in first year of child's life.
- One may expect to find a negative correlation between `mom.work` and the outcome `kid.score`.
- Should we just add one more column to the design matrix $\mathbf{X}$? It is not a good idea.

How to deal with categorical attributes

- Suppose we consider a new model with `mom.work`:

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \beta_{iq}\ \text{mom.iq} + \beta_{hsiq}\ \text{mom.hs:mom.iq} + \beta_{wk}\text{mom.work} + \text{error}$$

- Fitting this new model we obtain $\beta_{wk} = 0.11$ (very small value).
- This means that when `mom.work` increase in 1 unit we expect a small increase of 0.11 in `kid.score`.
- When `mom.work` changes from 1 to 2, `kid.score` increases 0.11, on average.
- When `mom.work` changes from 3 to 4, `kid.score` increases on average THE SAME AMOUNT 0.11.
- Does this makes sense? It does not.

Categorical regressors

- `mom.work` has 4 values: 1, 2, 3, 4
- These are not just nominal values, there are clearly a semantic order.
- As `mom.work` increases, there is a general decreasing time commitment of the mother with the child.
- However, it is not clear that we should attach THE SAME MEANING to the 1 unit change when `mom.work` changes from 1 to 2 as when `mom.work` changes from 3 to 4.
- Increasing `mom.work` implies less time commitment but...
- ...it does not mean that changing from 1 to 2 is the same change of time commitment when changing from 3 to 4.

Categorical regressors

- To make this more clear, imagine another categorical variable.
- Suppose we have `mom.rel` coding the mother religion with four values: 1, 2, 3, 4.
- (1): catholic; (2) protestant; (3) other religion; (4) no religion.
- In this case, it is clear that the values are simply nominal labels.
- There is no sense in seeing a meaningful order on these values.
- Differences between them are meaningless.

Categorical regressors and dummy variables

- The best approach is to create dummy variables to represent the different levels of the categorical regressor.
- When we have a categorical regressor has k levels, we create $k - 1$ dummy (or binary) variables.
- We select one of the levels as a reference or base.
- Next, we create one binary variable to indicate the presence of each of the non-reference category.

One simple example

- Vamos voltar para um modelo mais simples apenas para ilustrar como uma variável categórica deve ser introduzida num modelo de regressão linear.
- Suponha que nosso objetivo seja predizer `kid.score` usando `mom.hs`, `mom.iq` e `mom.work`.
- Ignorando as possíveis interações vamos criar uma matriz de desenho que use cada uma dessas variáveis.
- Como `mom.work` possui quatro níveis distintos (1, 2, 3,4), nós vamos precisar criar $4 - 1 = 3$ variáveis binárias (ou dummies).
- Vamos fixar a categoria `mom.work = 1` como base ou referência e criar trés variáveis dummies Z_1, Z_2 e Z_3 para indicar a presenạ das demai categorias.
- Veja o esquema a seguir.

Uma matriz com as variáveis

- Vamos imaginar alguns dados da nossa amostra organizados como uma tabela e mostrar como as variáveis Z_1, Z_2 e Z_3 devem ser criadas EXCLUSIVAMENTE a partir do valor de `mom.work`:

i	kid.score$_i$	mom.hs$_i$	mom.iq$_i$	mom.work$_i$	Z_{i1}	Z_{i2}	Z_{i3}
1	65	1	121.12	1	0	0	0
2	98	1	89.36	2	1	0	0
3	85	0	115.44	3	0	1	0
4	83	1	99.45	4	0	0	1
5	115	1	92.75	1	0	0	0
6	98	0	107.90	1	0	0	0
7	69	0	138.89	3	0	1	0
8	106	0	125.15	2	1	0	0
9	102	1	81.62	4	0	0	1
10	95	1	95.07	2	1	0	0
11	91	0	88.58	3	0	1	0
⋮	⋮		⋮			⋮	
432	50	0	94.86	2	1	0	0
433	88	1	96.86	4	0	0	1
434	70	1	91.25	1	0	0	0

- Repetindo parcialmente:

i	kid.score$_i$	mom.hs$_i$	mom.iq$_i$	mom.work$_i$	Z_{i1}	Z_{i2}	Z_{i3}
1	65	1	121.12	1	0	0	0
2	98	1	89.36	2	1	0	0
3	85	0	115.44	3	0	1	0
4	83	1	99.45	4	0	0	1
5	115	1	92.75	1	0	0	0
6	98	0	107.90	1	0	0	0
7	69	0	138.89	3	0	1	0
⋮	⋮		⋮			⋮	

- Assim, a referência é a categoria `mom.work = 1`.
- Definimos:

$$Z_2 = \begin{cases} 1, & \text{se mom.work=2} \\ 0, & \text{caso contrário} \end{cases} \qquad Z_3 = \begin{cases} 1, & \text{se mom.work=3} \\ 0, & \text{c.c.} \end{cases} \qquad Z_4 = \begin{cases} 1, & \text{se mom.work=4} \\ 0, & \text{c.c.} \end{cases}$$

- Se uma mãe tem `mom.work = 1` então $Z_2 = Z_3 = Z_4 = 0$.
- Se `mom.work = 3` então $Z_2 = Z_4 = 0$ e $Z_3 = 1$.

The matrix representation

- The matrix-version of linear regression has a design matrix **X** with ?? columns.

$$
\texttt{kid.score} = \left[\begin{array}{c}
\text{kid.score}_i \\ \hline
65 \\ 98 \\ 85 \\ 83 \\ 115 \\ 98 \\ 69 \\ 106 \\ 102 \\ 95 \\ 91 \\ \vdots \\ 50 \\ 78 \\ 80
\end{array}\right]
= \left[\begin{array}{rrrrr}
\text{mom.hs}_i & \text{mom.iq}_i & Z_{i1} & Z_{i2} & Z_{i3} \\ \hline
1 & 121.12 & 0 & 0 & 0 \\
1 & 89.36 & 1 & 0 & 0 \\
0 & 115.44 & 0 & 1 & 0 \\
1 & 99.45 & 0 & 0 & 1 \\
1 & 92.75 & 0 & 0 & 0 \\
0 & 107.90 & 0 & 0 & 0 \\
0 & 138.89 & 0 & 1 & 0 \\
0 & 125.15 & 1 & 0 & 0 \\
1 & 81.62 & 0 & 0 & 1 \\
1 & 95.07 & 1 & 0 & 0 \\
0 & 88.58 & 0 & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 94.86 & 1 & 0 & 0 \\
1 & 96.86 & 0 & 0 & 1 \\
1 & 91.25 & 0 & 0 & 0
\end{array}\right]
\left[\begin{array}{c}
\beta_0 \\ \beta_{hs} \\ \beta_{iq} \\ \beta_{wk}^{(2)} \\ \beta_{wk}^{(3)} \\ \beta_{wk}^{(4)}
\end{array}\right]
+ \left[\begin{array}{c}
\epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_n
\end{array}\right]
$$

- In the more compact notation, we have

$$\texttt{kid.score} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The model with the dummy variables

- The model with the dummy variables (or columns) is

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error}$$

- This means that

$$\begin{aligned}
(\text{mom.hs} = 0) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs} \cdot 0 + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error} \\
&= \beta_0 + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error} \\
(\text{mom.hs} = 1) \Rightarrow \text{kid.score} &= \beta_0 + \beta_{hs} \cdot 1 + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error} \\
&= (\beta_0 + \beta_{hs}) + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error}
\end{aligned}$$

$$\text{kid.score} = \beta_0 + \beta_{hs}\text{mom.hs} + \beta_{iq}\ \text{mom.iq} + \beta_{wk}^{(2)}\ Z_2 + \beta_{wk}^{(3)}\ Z_3 + \beta_{wk}^{(4)}\ Z_4 + \text{error}$$

$$\begin{aligned}
(\text{mom.hs} = 0)\ \&\ (\text{mom.work} = 1) &\Rightarrow \beta_{hs}\text{mom.hs} = 0 \text{ and } Z_2 = Z_3 = Z_4 = 0 \\
\text{kid.score} &= \beta_0 + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 0)\ \&\ (\text{mom.work} = 2) &\Rightarrow \beta_{hs}\text{mom.hs} = 0 \text{ and } Z_3 = Z_4 = 0,\ Z_2 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{wk}^{(2)}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 0)\ \&\ (\text{mom.work} = 3) &\Rightarrow \beta_{hs}\text{mom.hs} = 0 \text{ and } Z_2 = Z_4 = 0,\ Z_3 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{wk}^{(3)}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 0)\ \&\ (\text{mom.work} = 4) &\Rightarrow \beta_{hs}\text{mom.hs} = 0 \text{ and } Z_2 = Z_3 = 0,\ Z_4 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{wk}^{(4)}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
\\
(\text{mom.hs} = 1)\ \&\ (\text{mom.work} = 1) &\Rightarrow \beta_{hs}\text{mom.hs} = \beta_{hs} \text{ and } Z_2 = Z_3 = Z_4 = 0 \\
\text{kid.score} &= (\beta_0 + \beta_{hs}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 1)\ \&\ (\text{mom.work} = 2) &\Rightarrow \beta_{hs}\text{mom.hs} = \beta_{hs} \text{ and } Z_3 = Z_4 = 0,\ Z_2 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{hs} + \beta_{wk}^{(2)}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 1)\ \&\ (\text{mom.work} = 3) &\Rightarrow \beta_{hs}\text{mom.hs} = \beta_{hs} \text{ and } Z_2 = Z_4 = 0,\ Z_3 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{hs} + \beta_{wk}^{(3)}) + \beta_{iq}\ \text{mom.iq} + \epsilon \\
(\text{mom.hs} = 1)\ \&\ (\text{mom.work} = 4) &\Rightarrow \beta_{hs}\text{mom.hs} = \beta_{hs} \text{ and } Z_2 = Z_3 = 0,\ Z_4 = 1 \\
\text{kid.score} &= (\beta_0 + \beta_{hs} + \beta_{wk}^{(4)}) + \beta_{iq}\ \text{mom.iq} + \epsilon
\end{aligned}$$

Interpretation of the effect of categorical attributes

- We have a regression line given by

  $$\text{kid.score} = \beta_0 + \beta_{iq}\ \text{mom.iq} + \epsilon$$

  for mothers with `mom.hs = 0` and `mom.work = 1`).
- The categorical attributes (both, `mom.hs` and `mom.work`) impact this base model by just shifting the intercept up or down.
- Each combination of `mom.hs` and `mom.work` has a regression line associated and they are all parallel to each other.
- The distance between the parallel lines are produced by the coefficients of the categorical attributes.

Interpretation of the effect of categorical attributes

- For example, the average impact on `kid.score` of having `mom.work = 3` is to add the amount $\beta_{wk}^{(3)}$ to the base model.
- The average impact on `kid.score` of having `mom.hs = 1` is to add the amount β_{hs} to the base model.
- The combined effect of having `mom.work = 3` and `mom.hs = 1` is to add $\beta_{wk}^{(3)} + \beta_{hs}$ to the base model.

Regression with `mom.work` as a factor

- To run a regression in R with a categorical variable, we need to transform it in an object of class `factor`
- If the categorical variable is already binary (as `mom.hs`), this is not necessary.

```
> mom.work.ft = factor(mom.work)
> mom.work.ft[1:10]
 [1] 4 4 4 3 4 1 4 3 1 1
Levels: 1 2 3 4
> is.character(levels(mom.work.ft))
[1] TRUE
```

- The reference category is the first one in the lexicographic order of the levels by default.
- It can be changed, if necessary.

Regression with a factor

- Fitting in R:

```
> fit.5 = lm(kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st + mom.work.ft)
> summary(fit.5)
...
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       83.9221     2.7605  30.401  < 2e-16 ***
mom.hs             2.7922     2.4783   1.127  0.26052
mom.iq.st          9.4869     1.4924   6.357  5.3e-10 ***
mom.work.ft2       1.8354     2.8061   0.654  0.51343
mom.work.ft3       5.1586     3.2204   1.602  0.10993
mom.work.ft4       0.9189     2.4985   0.368  0.71321
mom.hs:mom.iq.st  -4.7436     1.6359  -2.900  0.00393 **
...
```

- Changing from `mom.work=1` to `mom.work=3` increases the INTERCEPT of the base model in 5.15.
- What this means?

Interpretation of the dummies coefficients

- Let a and b be two arbitrary values for `mom.hs` and `mom.iq.st`.
- Define

$$\Delta_{13} = \mathbb{E}(Y|\text{mom.hs=a, mom.iq.st=b, mom.work=3}) - \mathbb{E}(Y|\text{mom.hs=a, mom.iq.st=b, mom.work=1})$$

- The coefficient $\beta_{wk}^{(3)}$ associated with `mom.work.ft3` in the R output is an estimate of Δ_{13}:

$$\begin{aligned} \mathbb{E}(Y|\text{mom.hs=a, mom.iq.st=b, mom.work=3}) &= \beta_0 + \beta_{hs}\ a + \beta_{iq}\ b + \beta_{wk}^{(3)} \\ \mathbb{E}(Y|\text{mom.hs=a, mom.iq.st=b, mom.work=1}) &= \beta_0 + \beta_{hs}\ a + \beta_{iq}\ b \end{aligned}$$

- Therefore, the difference Δ_{13} between these values is simply $\beta_{wk}^{(3)}$.

Interpretation of the dummies coefficients

- We saw that $\beta_{wk}^{(3)}$ is the expected additional impact on `kid.score` of having a mother with `mom.work = 3`.
- This value $\beta_{wk}^{(3)}$ does not depend on the FIXED values *a* and *b* of `mom.hs` and `mom.iq.st`.
- Hence, the impact of `mom.work` does not depend on the `mom.hs`.
- What if we want it to vary?
- We may want to allow the impact of working outside home to impact heavily the kids when `mom.hs=0` but to have negligible effect when `mom.hs=1`.
- This is not allowable by the present model.
- We want to allow the OTHER attributes to change their effect depending on `mom.work`.
- We can enlarge the model using the operator ":". See the example next.

Interaction between a continuous and categorical

- We want the effect of `mom.iq` to vary according to `mom.hs` AND `mom.work`

```
> fit.6 = lm(kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st +
+                mom.work.ft + mom.work.ft:mom.iq.st)
> summary(fit.6)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            83.48570    2.76781  30.163  < 2e-16 ***
mom.hs                  3.15913    2.48116   1.273    0.204
mom.iq.st               8.98161    1.72058   5.220 2.81e-07 ***
mom.work.ft2            2.79984    2.84633   0.984    0.326
mom.work.ft3            5.33531    3.30784   1.613    0.108
mom.work.ft4            1.05074    2.50028   0.420    0.675
mom.hs:mom.iq.st       -4.29069    1.70018  -2.524    0.012 *
mom.iq.st:mom.work.ft2  3.02729    1.97610   1.532    0.126
mom.iq.st:mom.work.ft3 -0.08742    2.05808  -0.042    0.966
mom.iq.st:mom.work.ft4 -0.65338    1.57537  -0.415    0.679
```

One more step?

- The impact on the expected value of `kid.score` when we move from `mom.work = 1` to `mom.work = 3` IS THE SAME FOR `mom.hs=0` and `mom.hs=1`.
- If we want to allow for differences on this effect, we can.
- We should add a TRIPLE interaction, adding a factor `mom.work.ft:mom.iq.st:mom.hs`

One more step?

- We can allow also an TRIPLE interaction, adding a factor `mom.work.ft:mom.iq.st:mom.hs`

```
 > fit.7 = lm(kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st +
+   mom.work.ft + mom.work.ft:mom.iq.st + mom.work.ft:mom.iq.st:mom.hs )
> summary(fit.7)
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     84.1811     2.8786  29.244  < 2e-16 ***
mom.hs                           3.0816     2.5091   1.228   0.2201
mom.iq.st                       10.0622     2.1720   4.633 4.82e-06 ***
mom.work.ft2                     2.1427     2.9939   0.716   0.4746
mom.work.ft3                     4.0619     3.5623   1.140   0.2548
mom.work.ft4                     0.3974     2.6994   0.147   0.8830
mom.hs:mom.iq.st                -6.1031     2.8295  -2.157   0.0316 *
mom.iq.st:mom.work.ft2           1.6803     3.5403   0.475   0.6353
mom.iq.st:mom.work.ft3          -3.6560     4.3597  -0.839   0.4022
mom.iq.st:mom.work.ft4          -1.9290     3.3063  -0.583   0.5599
mom.hs:mom.iq.st:mom.work.ft2    2.1771     4.3196   0.504   0.6145
mom.hs:mom.iq.st:mom.work.ft3    4.9373     5.1763   0.954   0.3407
mom.hs:mom.iq.st:mom.work.ft4    2.0367     3.9129   0.520   0.6030
```

The last model (for now)

- Adding mother's age

```
 > mom.age.ct = mom.age - mean(mom.age)
> fit.8 = lm(kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st + mom.age.ct +
+                mom.work.ft + mom.work.ft:mom.iq.st + mom.work.ft:mom.iq.st:m
> summary(fit.8)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    84.8598     2.9364  28.899  < 2e-16 ***
mom.hs                          2.4152     2.5731   0.939   0.3485
mom.iq.st                      10.2982     2.1806   4.723 3.18e-06 ***
mom.age.ct                      0.3897     0.3362   1.159   0.2471
mom.work.ft2                    2.2492     2.9941   0.751   0.4529
mom.work.ft3                    3.6809     3.5760   1.029   0.3039
mom.work.ft4                    0.2179     2.7027   0.081   0.9358
mom.hs:mom.iq.st               -6.4799     2.8470  -2.276   0.0233 *
mom.iq.st:mom.work.ft2          1.7177     3.5391   0.485   0.6277
mom.iq.st:mom.work.ft3         -4.2564     4.3886  -0.970   0.3327
mom.iq.st:mom.work.ft4         -1.8642     3.3054  -0.564   0.5731
mom.hs:mom.iq.st:mom.work.ft2   2.2554     4.3183   0.522   0.6018
mom.hs:mom.iq.st:mom.work.ft3   5.6929     5.2151   1.092   0.2756
mom.hs:mom.iq.st:mom.work.ft4   2.0715     3.9115   0.530   0.5967
```

One more step?

- Variáveis criadas como interações (produtos) entre atributos devem ser criadas com cuidado.
- Elas aumentam rapidamente o número de colunas da matriz de desenho **X**.
- Uma matriz **X** com muitas colunas é problemática.
- Matrizes assim devem ser tratadas com métodos de regressão com *regularização*, um assunto que veremos mais tarde.

So, is it good?

- OK, we fit the best possible linear model considering a set of attributes (COLUMNS of the design matrix).
- We found the coefficients the minimize the difference between the vector of the observed values **Y** and a PREDICTOR that is a linear combination os the columns of the design matrix **X**.
- We got the best possible.
- The question now is: the best is good enough?
- The best can be excellent OR it can be really poor.
- May be the attributes in the design matrix **X** areNOT able to predict well **Y**.
- How to check the quality of the fit?
- Is there a measure for goodness of fit? Yes, it R^2.

Total Sum of Squares: SSTO

- **Y** is a vector with variation.
- How much does **Y** varies *without considering* any regressors?
- The total variation of **Y** around its mean $\bar{Y} = \sum_i Y_i/n$ is given by

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- SSTO measures the total amount of variability we have in the vector **Y** ignoring completely the presence of other possible regressors that could explain why **Y** varies.
- We could take the average variation by dividing $SSTO$ by n but we will consider instead the total variation rather than the average variability.
- However, it is more convenient mathematically to work with SSTO.

Total Sum of Squares: SSTO

- We can see SSTO as the squared length of a n-dimensional vector.
- Indeed, let $\mathbf{1} = (1, 1, \ldots, 1)'$ be a column-vector of dimension n.
- Let $\bar{Y}$ be the aithmetic mean of the vector $\mathbf{Y}$.
- Then, $\mathbf{Y} - \bar{Y}\mathbf{1}$ is a n-dimensional vector and its squared length is given by
$$||\mathbf{Y} - \bar{Y}\mathbf{1}||^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = SSTO$$
- The vector $\bar{Y}\mathbf{1}$ is the orthogonal projection of $\mathbf{Y}$ into the vector sub-space spanned by the multiple of $\mathbf{1}$.
- Indeed, $\bar{Y}\mathbf{1}$ is a multiple of $\mathbf{1}$ and $\bar{Y}\mathbf{1} \perp (\mathbf{Y} - \bar{Y}\mathbf{1})$.

Total Sum of Squares: SSTO

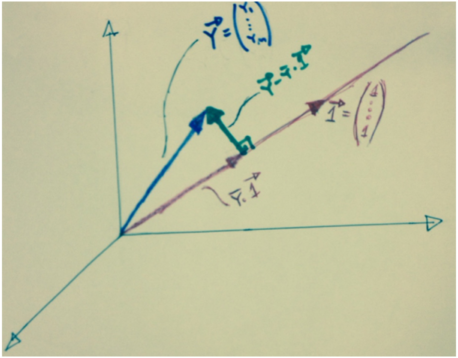

$\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$
$\vec{Y} - \bar{Y}.\vec{1}$
$\vec{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$
$\bar{Y}.\vec{1}$

Figura: Vector $\mathbf{Y}$ and its orthogonal projection $\bar{Y}\mathbf{1}$ into the vector $\mathbf{1}$. *SSTO* is the squared length of the vector $\mathbf{Y} - \bar{Y}\mathbf{1}$.

The Residual Vector

- The least squares coefficients that minimize the distance between the vector $\mathbf{Y}$ and a linear combination of the p columns of the design matrix $\mathbf{X}$ is given by
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$
- The linear regression predictor of the vector $\mathbf{Y}$ is given
$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$
- The $n \times n$ matrix $\mathbf{H}$ is the orthogonal projection matrix: given any vector $\mathbf{Y} \in \mathbb{R}^n$, the vector $\mathbf{H}\mathbf{Y}$ is the orthogonal projection of $\mathbf{Y}$ into the linear subspace of linear combination of columns of $\mathbf{X}$.
- The n-dimensional vector of the difference between the data $\mathbf{Y}$ and the best regression predictor $\hat{\mathbf{Y}}$ is called the residual vector
$$\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Pictorial representation

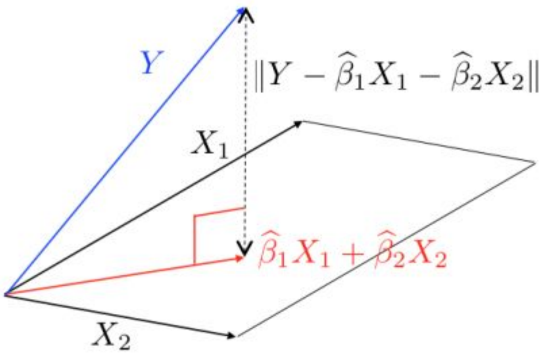

Y
$\|Y - \widehat{\beta}_1 X_1 - \widehat{\beta}_2 X_2\|$
X_1
$\widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2$
X_2

Figura: $\mathbf{Y} \in \mathbb{R}^n$ and its orthogonal projection $\hat{\mathbf{Y}}$ which is also in $\mathbb{R}^n$. However, $\hat{\mathbf{Y}}$ lives in the p-dimensional subspace composed by the linear combination of the columns of $\mathbf{X}$ (here, represented as a two-column $[\mathbf{X}_1|\mathbf{X}_2]$ matrix).The residual vector $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ is orthogonal to the projected vector $\hat{\mathbf{Y}}$.

When do we have a good fit?

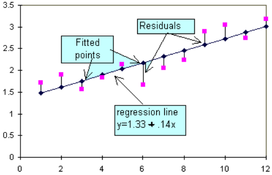
- If the regression model is able to predict well the observed data we should have a small residual vector.
- That is, $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}} \approx \mathbf{0}$.
- Rather than looking at each one of the n entries in the residual vector, we look globally at its length.
- We have a vector with small length if, and only if, each entry is small.
- So, a good fit implies that the (squared) length of residual vector is small

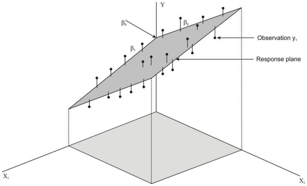
$$0 \approx ||\mathbf{r}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Now, how to evaluate if $||\mathbf{r}||^2$ is small? Small compared to what?

Looking at the residuals

- The residuals are easily visualized when we have one or two regressors besides the constant column $\mathbf{1} = (1, 1, \ldots, 1)'$.
- The *i*-th element of the *n*-dim vector $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the difference between each data point y_i and the predicted value $\hat{y}_i$.
- The predicted value $\hat{y}_i$ is the value in the regression line (in the 1-regressor case plus the column $\mathbf{1}$) or the value in the regression plane (in the 2-regressors case plus the columns $\mathbf{1}$).





The sum of squares

- How to evaluate if $||\mathbf{r}||^2$ is small? Small compared to what?
- Statisticians developed an intrinsic scale, that takes into account the scale in which the data has been measured.
- We compare the size of the residuals (that is, $||\mathbf{r}||^2$) with the size of the variation of the $\mathbf{Y}$ vector.
- The squared length of the residual vector measures what remains of *unpredicted* variability in the vector $\mathbf{Y}$ AFTER we consider the regressors as predictors:

$$||\mathbf{r}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The sum of squares

- When the residual vector

$$||\mathbf{r}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  is small?
- The idea is to compare this remaining variability with the original variability in $\mathbf{Y}$ BEFORE any regressors were considered.
- The variation of $\mathbf{Y}$ around $\bar{y}$, the mean of $\mathbf{Y}$, is equal to:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = ||\mathbf{Y} - \bar{y}\mathbf{1}||^2$$

Finally, the R^2

- That is, we consider the ratio

$$\frac{SSE}{SSTO} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{||\mathbf{Y} - \bar{y}\mathbf{1}||^2}$$

- If we have a good fit, we should have this ratio close to zero.
- We will prove that $SSE/SSTO$ is always smaller than 1.
- Hence, it is more common to use $1 - SSE/SSTO$, which is called R^2:

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{||\mathbf{Y} - \bar{y}\mathbf{1}||^2}$$

- A good fit should have $R^2 \approx 1$.

R^2 in R output

- The R^2 is such an important global measure of the quality of the regression model to predict Y that it is always part of any regression output.
- In R, it is at the end of the `summary()` command:

```
> summary(fit.8)
Call:
lm(formula = kid.score ~ mom.hs + mom.iq.st + mom.hs:mom.iq.st +
    mom.age.ct + mom.work.ft + mom.work.ft:mom.iq.st + mom.work.ft:mom.iq.st:mom.hs)

Residuals:
    Min      1Q  Median      3Q     Max
-57.783 -10.531   2.274  11.417  42.362

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    84.8598     2.9364  28.899  < 2e-16 ***
mom.hs                          2.4152     2.5731   0.939   0.3485
...

Residual standard error: 17.97 on 420 degrees of freedom
Multiple R-squared:  0.2478, Adjusted R-squared:  0.2246
F-statistic: 10.65 on 13 and 420 DF,  p-value: < 2.2e-16
```

Is $R^2 = 0.25$ a high value?

- The R^2 in this example disappointed us.
- It is very small.
- Hence, it is poor our ability to predict `kid.score` based on the mother's IQ and the other few variables we are looking at.
- The models has some prediction capacity, as we will learn from the other statistics in this utput.
- However, the low R^2 is saying that "some predicition capacity" is NOT a LOT of prediction ability.
- To predict very well, we need R^2 above 0.8, at least.
- In social and economic analysis, most of the time, the R^2 is small.
- In these studies, there are too many other variables that we do not take into account affecting the output.

Seeing $\mathbf{r}$ and $\hat{\mathbf{Y}}$ in R

- ...

Engineering-type studies

- In social and economic studies, there are too many other variables that we do not take into account affecting the output.
- In more controlled studies, such as those in engineering problems, there is a small number of important factors and the many others have much smaller effects.
- Hence, it is simpler to isolate the relevant factors and obtain a good predictive model in these types of studies.
- Medical studies lie in between the social and the engineering studies in their predictive capacity.

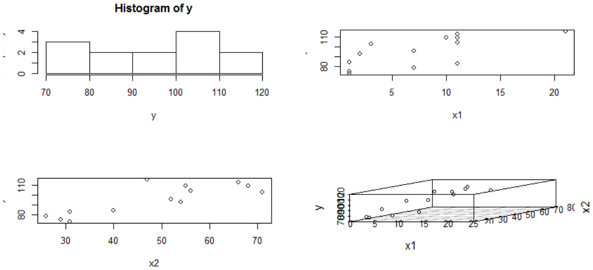
Um exemplo de engenharia

- Concreto é produzido pela mistura de água ao cimento produzindo uma massa pastosa.
- Ocorre então uma reação exotérmica, que libera calor e seca a massa pastosa deixando-a muito dura.
- Durante este processo, o concreto sofre variações de volume que podem gerar tensões perigosas para a estrutura final.
- Um estudo foi realizado para analisar a quantidade de calor produzido pelo cimento portland durante o endurecimento.
- Mediu-se o calor produzido (variável y) em resposta á quantidade de dois componentes do cimento medidos em termos da sua contibuição percentual para o peso do cimento.
- Os dois componentes foram *Gypsum* e *Tricalcium silicate* (regressores x_1 e x_2).
- O objetivo é verificar como o calor produzido responde às variações em x_1 e x_2.

Lendo e visualizando os dados

- R script:

```
cement <- read.table("CementHeat.txt", header=T)
head(cement); attach(cement)
par(mfrow=c(2,2)); hist(y); plot(x1, y); plot(x2, y);
library(scatterplot3d); scatterplot3d(cement)
```



Lendo e visualizando os dados

- Veja que o R^2 é muito alto: $R^2 = 0.97$.

```
> summary(lm(y ~ x1 + x2, data=cement))
Call:
lm(formula = y ~ x1 + x2, data = cement)
...
Residuals:
    Min      1Q  Median      3Q     Max
-5.4473 -1.5021 -0.4428  1.6925  4.4180
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.98189    2.90405  17.555 7.65e-09 ***
x1           1.40452    0.15408   9.115 3.69e-06 ***
x2           0.69728    0.05825  11.971 2.99e-07 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
...
Residual standard error: 3.057 on 10 degrees of freedom
Multiple R-squared:  0.9668, Adjusted R-squared:  0.9602
F-statistic: 145.8 on 2 and 10 DF,  p-value: 4.013e-08
```

..

- ...

Lingo

- There are several colorful expressions associated with the regression analysis.
- The residual variability

$$||\mathbf{r}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  is called *the unexplained variation* of $\mathbf{Y}$.
- It is also called the *residual sum of squares* and denoted by SSE.

Lingo

- The summary and global variation

$$||\mathbf{Y} - \bar{y}\mathbf{1}||^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  of $\mathbf{Y}$ around its mean $\bar{y}$ is called *the total variation* of $\mathbf{Y}$.
- It is also called the *total sum of squares* and denoted by *SSTO*.

Visualizando *SSTO* e *SSE*

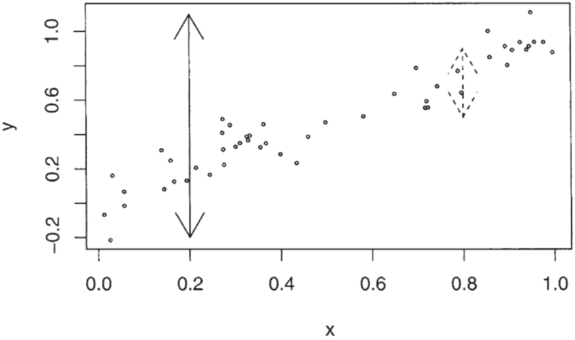


Figure 2.2 *Variation in the response y when x is known is denoted by dotted arrows while variation in y when x is unknown is shown with the solid arrows.*

Figura: SSTO é representado pela seta sólida e SSE pela seta tracejada.

More lingo

- Finally, the difference between *SSTO* and *SSE*.
- This is the *the variation of* $\mathbf{Y}$ *explained by the regressors* and it is called the *regression sum of squares*, and denoted by *SSR*.

$$SSR = SSTO - SSE$$

- A VERY IMPORTANT RESULT: It can be shown that

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = ||\hat{\mathbf{Y}} - \bar{y}\mathbf{1}||^2$$

Lingo

- This means that the total sum of squares $SSTO$ can be decomposed into TWO sums:

$$\begin{aligned} ||\mathbf{Y} - \bar{y}\mathbf{1}||^2 &= ||\hat{\mathbf{Y}} - \bar{y}\mathbf{1}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 \\ SSTO &= SSR + SSE \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

- The reason for this decomposition is the fact that we deal with orthogonal projections.
- ***Todos*** os resultados de regressão linear são derivados do Teorema de Pitágoras em espaços vetoriais.