Linear Regression

Simple and multiple

Introduction (Context, Motivation)

- Supervised method (we have a dataset (X,Y))

- Objective: to estimate a relationship f: X—>Y from the data in order to:

1) Predict new values of Y (Prediction)

(Example: predicting sales based on the TV, radio, and newspaper budget)

2)Understand the effect of each explanatory variable on Y (Inference)

(Example: measure the average impact of an additional €1000 on TV ad sales and test if this effect is significant)

Hypothesis: there is a linear relationship between X and Y.

$$Y = \beta_0 + \beta_1 X \text{ (ie : Sales = } \beta_0 + \beta_1 * TV)$$

Intercept

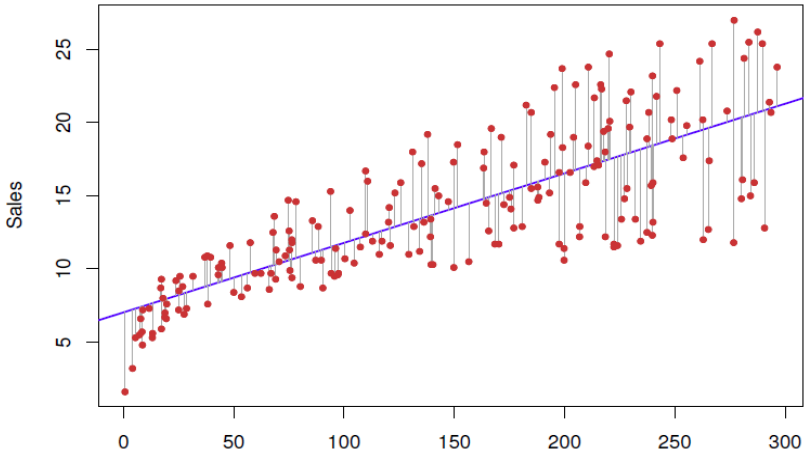
Slope : average incrase in Y associated with one unit increase in X

- β_0, β_1 : model parameters, we use the training data to
- ϵ random error, part of Y that cannot be explained by X
- Once estimated, we can predict Y:
  - $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} * x + \epsilon_i$

Predictor, estimated value of y

- To predict Y we need to determine $\widehat{\beta_0}$ et $\widehat{\beta_1}$ using the **OLS** (least squares) method

$$\min_{\beta} \sum (y_i - (\widehat{\beta_0} + \widehat{\beta_1}(x_i))^2 = y_i - \hat{y}_i = \underbrace{\epsilon_1{}^2 + \epsilon_2{}^2 + \ldots + \epsilon_n{}^2}_{\text{Residual Sum Of Squares}}$$



Minimization problem

- In order to minimize the function also called the cost function SSE:

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ which is a convex function (global minimum) we

1) Calculate the derivatives in relation to β_0, β_1

$$\frac{dSSE}{d\beta_1} = -2\sum_{i=1}^{n} x_i(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

2) We are looking for the zeros of these functions $\Leftrightarrow$

$$\frac{dSSE}{d\beta_0} = -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

We have closed Formulas :

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^{n}(x_i - \hat{x})^2} \quad , \quad \widehat{\beta_0} = \hat{y} - \widehat{\beta_1}\,\hat{x}$$

Model assumptions

- $E[\epsilon_i] = 0$ (no bias)
- $Var(\epsilon_i) = \sigma^2$ (homoscedasticity)
- ϵ_i independant of each other
- $\epsilon_i \sim N(0, \sigma^2)$
- $E[\varepsilon|X] = 0$, ϵ_i independance of x_i from the error (essential for unbiased estimators) **

** Often criticized: if $E[\varepsilon|X]$ is different from 0, then part of the information contained in X is hidden in the error, the estimators are biased (for example, we forgot an explanatory variable).

Statistical reminders

- $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Xi$ estimator (the simplest) of μ the mean
- If Xi i.i.d ~ N(μ, σ^2)
- Then E[$\hat{\mu}$] = μ et Var($\hat{\mu}$) =$\frac{\sigma^2}{n}$ d'où $\hat{\mu}$ ~ N(μ,$\frac{\sigma^2}{n}$)
- And a confidence interval CI at 95% for μ is given by$\hat{\mu}$ ±1.96·SE($\hat{\mu}$) (* next slide illustration)

As with the average of a sample, the regression estimators$\hat{\beta}_0, \hat{\beta}_1$are linear combinations of random variables. They therefore inherit a normal distribution (under the assumption of normally distributed errors), with a finite variance that can be calculated. This is why their confidence intervals have exactly the same form as those of the mean.

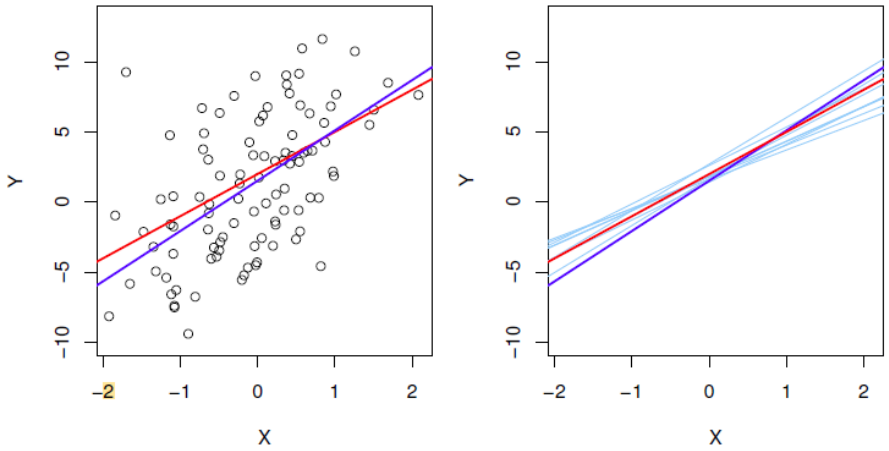
Illustration to understand the CI

- Let μ (true mean) = 170 cm (height of students)
- We take several samples of 25 students and calculate the mean ($\hat{\mu}$) of each sample.
- Around each $\hat{\mu}$, we construct the 95% confidence interval.

Sample	Estimated mean $\hat{\mu}$	CI at 95 %
#1	168.5	[166.0 ; 171.0]
#2	172.2	[169.5 ; 174.9]
#3	169.0	[166.5 ; 171.5]
#4	174.0	[171.0 ; 177.0]
#5	165.5	[163.0 ; 168.0] ❌

Uncertainty about the estimators

- Each sample(x1,y1),…,(xn,yn), or dataset gives a different line
- In red real line (if we had an infinite number of observations)
- In dark blue least square line
- In light blues: least square line across several subpopulations



- OLS estimators are unbiased : $E[\widehat{\beta_j}] = \beta_j$

$$SE(\widehat{\beta_0})^2 = \sigma^2[\frac{1}{n} + \frac{\overline{x^2}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}] \qquad SE(\widehat{\beta_1})^2 = [\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$$

(These formulas show that the more the x values are spread out, the better we estimate the slope)

- With assumptions and stats :

- RSE = $\sqrt{\frac{RSS}{n-2}}$ estimator of σ^2

- CI : $[\widehat{\beta_{0/1}}\text{-}2\,SE(\widehat{\beta_{0/1}}), \widehat{\beta_{0/1}}\text{-}2\,SE(\widehat{\beta_{0/1}})]$

Hypothesis testing on the coefficients

- Null hypothesis H_0 : There is no relationship btw X and Y ⇔ $\beta_1 = 0$
- Alternative hypothesis H_a : There is no relationship btw X and Y ⇔ $\beta_1 \neq 0$
- Question : When is it far enough from 0?
- To quantify it we compute t-statistic $t=\frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$
- P-value weak we reject H_0

Accuracy of the model

- Once we have ensured that H_0 is rejected, we calculate the model's accuracy.

- Two quantities are used for that **RSE** et $\boldsymbol{R^2}$

- RSE = $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ $\quad$ $R^2 = \frac{TSS-RSS}{RSS}$

- RSE : mean standard deviation of the residuals: average size of the prediction error

- TSS = $\sum_{i=1}^{n}(y_i - \hat{y})^2$: total variability of Y, measured in relation to its mean (error if we always predict "$\hat{y}$")
- RSS= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: variability of Y remaining after regression (error if predicted with the estimated regression)

- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$: proportion of variability explained by the model

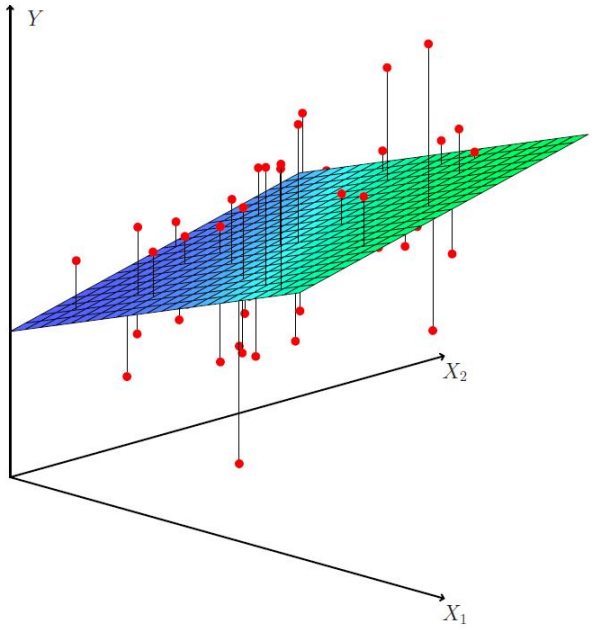
- If RSS = 0 ⇔ the model explains perfectly R^2 =1. If RSS = 1 ⇔ the model explains nothing RSS = TSS R^2 =0

- Limits: In simple regression, you risk overestimating the importance of one variable because it 'takes away' part of the effect of another correlated variable.

Multiple regression

- Idea: expand the analysis to several explanatory variables.
- Give each predictor a different slope coefficient.
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

- No more visualizable from 3 predictors
- Maths part with matrix



Matrix model

- $Y = X\beta + \varepsilon$

- Let n be the number of observations, and p the number of explanatory variables.
- Y vector of observations : $Y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$ vector of size n
- X matrix of predictors $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ \dots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$ matrix (n,p+1)
- β: vecteur des coefficients $\beta = \begin{pmatrix} \beta_0 \\ \dots \\ \beta_p \end{pmatrix}$ vector of size p+1 What we are trying to estimate
- ε: errors vector $\varepsilon = \begin{pmatrix} \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$ vector of size n ($Var(\varepsilon) = \sigma^2 I$)

- We want to find the matrix $\bar{\beta}$ that minimizes residual sum of squares

- $\min_{\beta} S(\beta)=||y-X\beta||^2 \Leftrightarrow (y-X\beta)^T(y-X\beta)$ cause $||v||^2 \Leftrightarrow v^T v$

- This is equivalent to find β such that $\frac{dS(\beta)}{d\beta} = 0$ (**computations on the next slide)

- $\frac{dS(\beta)}{d\beta} = 0 \quad \Leftrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y$

Calculations

$S(\beta) = (Y - X\beta)^T (Y - X\beta) = Y^T Y - Y X\beta - X^T \beta^T Y + X^T \beta^T X \beta$

**(If AX is a scalar (dim = 1*1) $(AX)^T = X^T A^T = AX$)*

D'où $Y^T X\beta = (y^T X\beta)^T = X^T \beta^T Y$ et (1) $= Y^T Y - 2 X^T \beta^T Y + X^T \beta^T X \beta$

Et $\frac{dS(\beta)}{d\beta} = -2X^T Y + 2\beta X^T X$

$$\frac{dS(\beta)}{d\beta} = 0$$

$$\Leftrightarrow -2X^T Y = -2\beta X^T X$$

$$\Leftrightarrow \beta = X^T Y (X^T X)^{-1}$$

Hypotheses and global tests

- Null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 \ldots = 0$
- Alternative hypothesis H_a : Au moins un β_j != 0
- Question : When is it far enough from 0 ?
- To quantify it, we calculate the F-stat $F=\frac{(TSS-RSS)/p}{RSS/(n-p-1)}$
- Interpretation : if F is large -> at least one prédictor explains Y

Partial effect and individuals p-values

- Each $\widehat{\beta}_j$ as a t-test like in simple regression
- Difference
    1. In simple regression : average effect of X_j on Y
    2. In multiple regression: effect of X_j keeping the other variables constant

- A variable may appear significant in simple regression but become useless in multiple regression when it is correlated with another variable.

Fit quality

- We take again R^2 and RSE but dépend on the number of variables
- RSE = $\sqrt{\frac{RSS}{n-p-1}}$ et R^2
- Warning R^2 always increases when new variables are added -> ${R^2}_{ajd}$

$${R^2}_{ajd} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

If we add a variable that brings nothing, it decreases.

- AIC/BIC: criteria that penalize the complexity of the model (choose the model with the smallest AIC/BIC)
- AIC : balance quality adjustment and number of parameters
- BIC: If n is large, BIC becomes more severe against complex models (more conservative)

Selection of variables and overfitting

- Too many variables : risk of overfitting

- Classic methods

  Forward selection : we add variables one by one

  (We add the variable that improves the model the most (reduction of the RSS, improvement of the AIC/BIC or $R^2{}_{ajd}$)

  Backward selection : we remove the least significant ones

  (We start with all the variables in the model. We progressively remove the least significant variable (the one with the highest p-value, or the one that improves the least the AIC/BIC). We stop when all remaining variables are significant) (p<n)

  Stepwise : mix of both

It is not enough to maximize R^2 , find a good compromise between complexity and performance)

Collinearity problem

- Collinearity: if the explanatory variables are highly correlated → unstable coefficients, difficult interpretation.

Let's imagine 2 predictors: radio and newspapers, very correlated ($\rho \approx 0.9$) if radio increases, newspapers increase almost the same.
So, it's impossible to know if the effect on sales really comes from radio or from newspapers.

In multiple regression, we try to separate their effects $Y = \beta_0 + \beta_1 Radio + \beta_2 newspaper + \cdots + \beta_p X_p + \varepsilon$

Radio and Newspaper move together -> many pairs(β_1, β_2) that explain sales almost as well the coefficients become unstable
  - sometimes β_1 is large and β_2 small, sometimes the inverse, sometimes one is negative while the other is positive.

- This is one of the reason why we introduce penalization regression

Extension (penalization regression) – Ridge

- When the variables are highly correlated, as we have seen, the coefficients can take unstable values (small variation in the data leads to large changes in the coefficients).
- Ridge stabilizes this by adding an L2 penalty to the function to be minimized

$$\min_{\beta} \sum(y_i - (\widehat{\beta_0} + \widehat{\beta_1}(x_i))^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

  - ⇔ minimize RSS under the constraint $\sum_{j=1}^{p} {\beta_j}^2 \leq c$ (sphere/circle constraint)
- Result : coefficients remains low in comparison with linear regression but never equal to 0 :

$$\frac{d}{d\beta_j}(RSS + \lambda \sum_{j=1}^{p} {\beta_j}^2) = \frac{dRSS}{d\beta_j} + 2\lambda\beta_j = 0$$

$$\beta_j = \frac{-1 dRSS}{2\lambda d\beta_j} \text{ smooth solution } \beta_j = 0 \Leftrightarrow RSS = 0$$

- Here we can see that $\beta_j = 0$ is an optimal solution for $\frac{dRSS}{d\beta_j} = 0$; RARE

- Visual example with 2 predictors :

  We want to minimize $\mathrm{RSS} + \Lambda({\beta_1}^2 + {\beta_2}^2)$

  $\beta_1 = 10$, $\beta_2 = -9$ $\sum_{j=1}^{2} {\beta_j}^2 = 181$

  $\beta_1 = 0.5$, $\beta_2 = 0.5$ $\sum_{j=1}^{2} {\beta_j}^2 = 0.5$ -> more stable , more likely to be choosed

About Λ

- Λ > 0 forces the coefficients $\widehat{\beta_j}$ to stay small (shrinkage)
- The more Λ is high the smaller the betas are ; low variance ; high bias ; risk of underfitting
- The more Λ is small the bigger the betas are ; low bias ; high variance ; risk of overfitting (normal linear regression)
- How to choose the best Λ ? : Cross validation
    - Split the datasets in k subset we train/ test several time in different subset / we compute the error / choose Λ that minimize this error

Extension (penalization regression) – Lasso

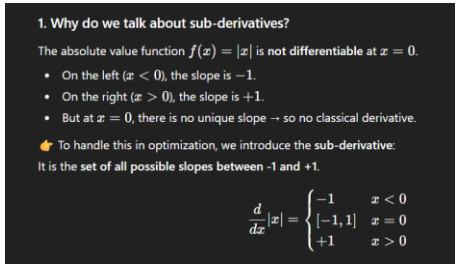
- When the variables are highly correlated, as we have seen, the coefficients can take unstable values (small variation in the data leads to large changes in the coefficients).
- Ridge stabilizes this by adding an L1 penalty to the function to be minimized

$$\min_{\beta} \sum(y_i - (\widehat{\beta_0} + \widehat{\beta_1}(x_i))^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - ⇔ minimize RSS under the constraint $\sum_{j=1}^{p} {\beta_j}^2 \leq c$ (sphere/circle constraint)
- Result : coefficients remains low in comparison with linear regression and CAN BE equal to 0 :

$$\frac{d}{d\beta_j}(RSS + \lambda \sum_{j=1}^{p} |\beta_j|) = \frac{dRSS}{d\beta_j} + \lambda \operatorname{sign}(\beta_j) = 0$$

$$\Leftrightarrow \begin{cases} \beta_j > 0, \frac{dRSS}{d\beta_j} + \lambda \\ \beta_j < 0, \frac{dRSS}{d\beta_j} - \lambda \\ \beta_j = 0, \frac{dRSS}{d\beta_j} + \lambda u \text{ , u belongs to } [-1,1] \; ** \end{cases}$$



- Here we can see that $\beta_j = 0$ is an optimal solution for - $\lambda < \frac{dRSS}{d\beta_j} < \lambda$; it happens more often (thos explain why lasso allows nul coeff)