Machine Learning for Big Data: Multiple Linear Regression

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Multivariate regression model
- Estimating model parameters
- Testing significance
- Conclusion

1 Introduction

What is MLR?

- Multiple Linear Regression (MLR) is a statistical method for estimating the relationship between a dependent variable and two or more independent (or predictor, or regressor) variables $\{x_1, x_2, \ldots x_k\}$.

- Find the subset of all possible predictor variables that explains a significant and appreciable proportion of the variance of Y, trading off adequacy of prediction against the cost of measuring more predictor variables.

- Purposes:
  - Prediction
  - Explanation
  - Theory building

Expanding Simple Linear Regression

- Quadratic model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

- General polynomial model.

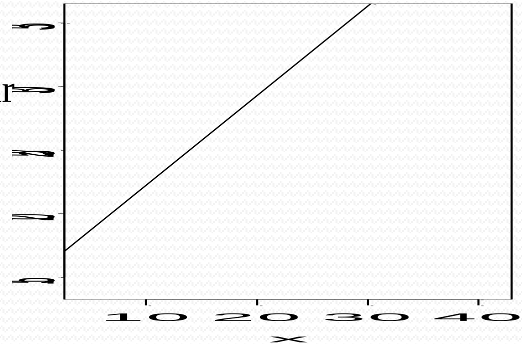
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \ldots + \beta_k x_1^k + \varepsilon$$

Adding one or more polynomial terms to the model.

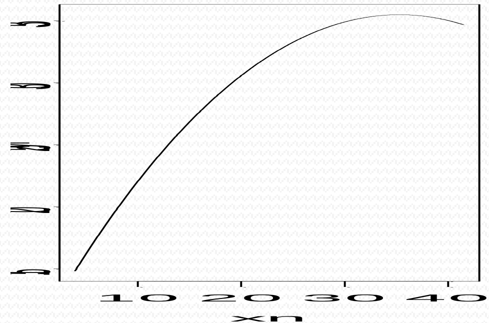
Any independent variable, x_i, which appears in the polynomial regression model as x_i^k is called a **k^{th}-degree term**.

Polynomial model shapes

Linear



Quadratic



Adding one more terms to the model significantly improves the model fit.

2 Multivariate regression model

Incorporating Additional Predictors

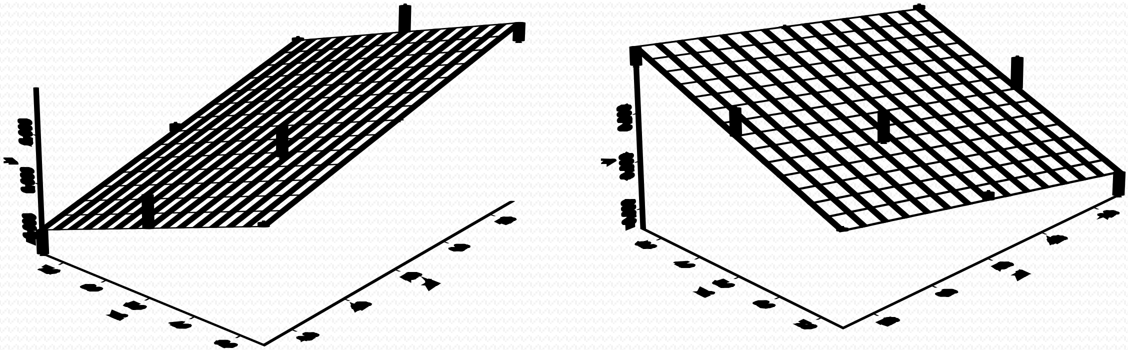
- **Simple additive multiple regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

- **Additive (effect) Assumption**:
  - The expected change in y per unit increment in x_j is constant
  - It does not depend on the value of any other predictor
  - This change in y is equal to β_j.

Additive regression models

For two independent variables, the response is modeled as a surface.



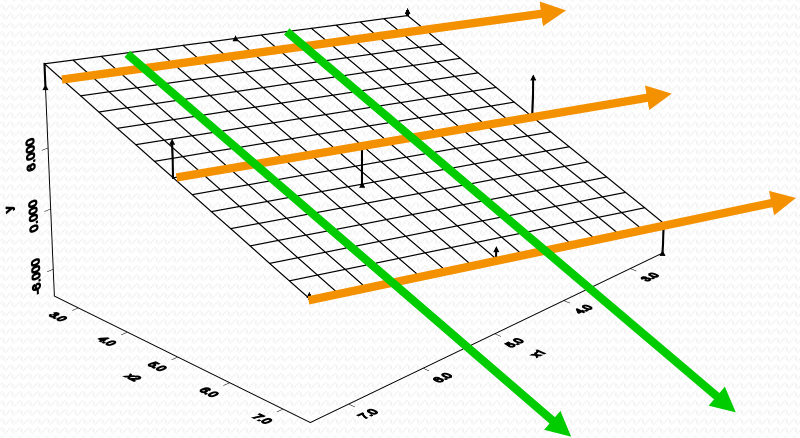
Interpreting Parameter Values

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

- “Intercept” β_0: value of y when all predictors are 0.
- “Partial slopes” $\beta_1, \beta_2, \ldots, \beta_k$
- β_j describes the expected change in y per unit increment in x_j when all other predictors in the model are held at a constant value.

Additive regression models β_j

β_1 : slope in direction of x_1



β_2 : slope in direction of x_2

Regression with Interaction Terms

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +$$

$$\beta_3 x_3 + \ldots + \beta_k x_k +$$

$$\beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 +$$

$$\ldots + \beta_{1k} x_1 x_k + \ldots$$

$$+ \beta_{k-1,k} x_{k-1} x_k + \varepsilon$$

cross-product terms quantify the interaction among predictors.

Interactive (effect) assumption:
The effect of one predictor, x_i, on the response, y, will depend on the value of one or more of the other predictors.

Interpreting Interaction

Interaction Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \underline{\beta_{12} x_{i1} x_{i2}} + \varepsilon_i$$

or define: $$x_{i3} = x_{i1} x_{i2}$$

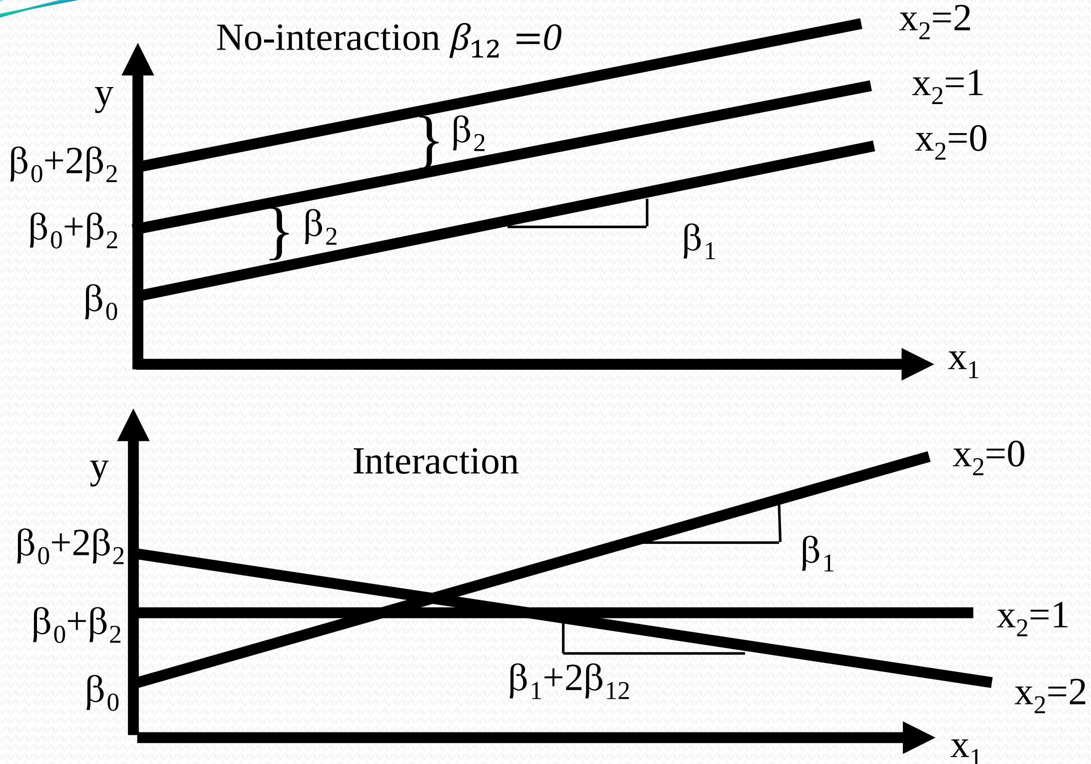
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i3} + \varepsilon_i$$

No difference

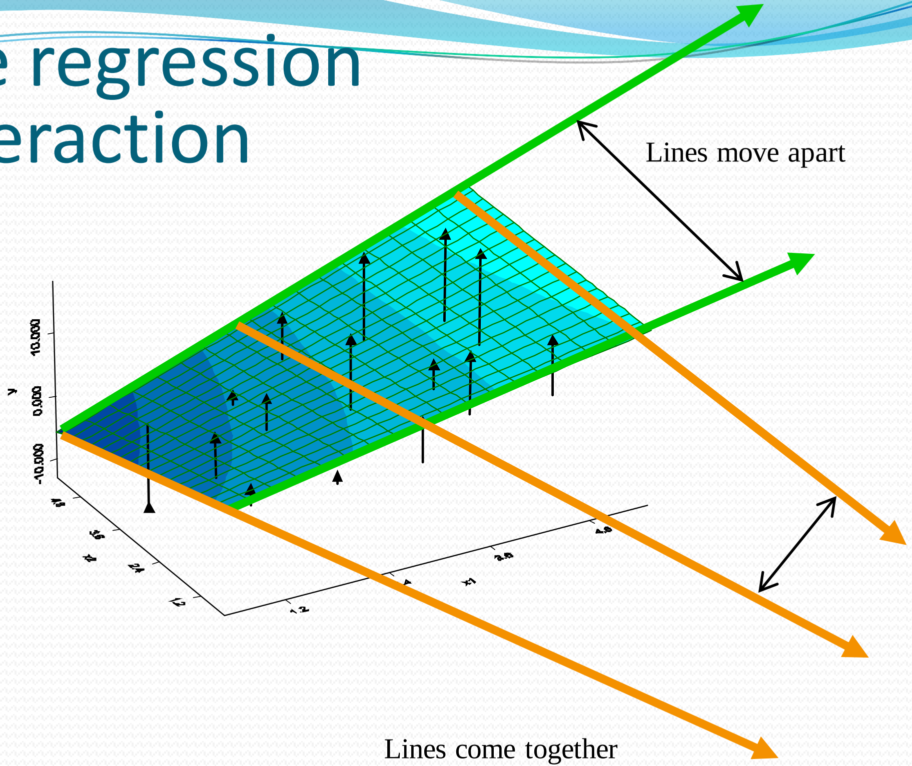
$\boldsymbol{\beta_1}$: No longer the expected change in y per unit increment in x_1!

$\boldsymbol{\beta_{12}}$: No easy interpretation! The effect on y of a unit increment in x_1, now depends on x_2.

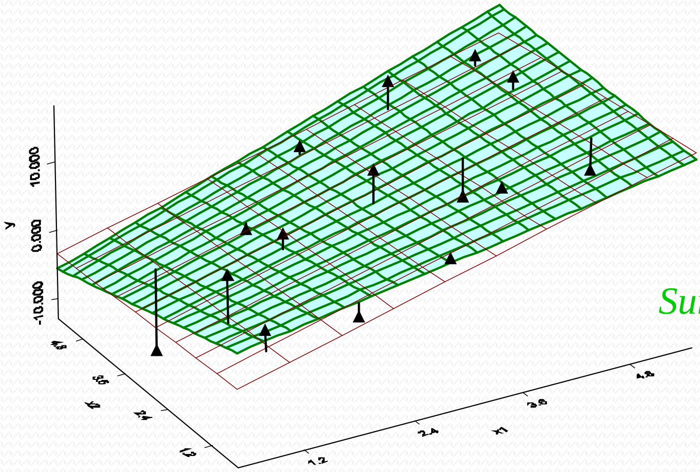
Visual interpretation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$



Multiple regression with interaction



Effect of the interaction term



Surface is twisted.

A Protocol for Multiple Regression

Identify all possible predictors.

Establish a method for estimating model parameters and their standard errors.

Develop tests to determine if a parameter is equal to zero (i.e. no evidence of association).

Reduce number of predictors appropriately.

Develop predictions and associated standard error.

3 Estimating Model Parameters

Least Squares Estimation

- Assuming n random samples

$$(y_i, x_{i1}, x_{i2}, \dots, x_{ik}), i = 1,2, \dots, n.$$

- The estimates of the parameters for the best predicting equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

is found by choosing the values: $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$

which minimize the expression:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2$$

Matrix form and optimal solution

- $SSE = SSE(\beta) = \|Y - A\beta\|^2$ with

$$\mathrm{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ & \vdots & & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

- The optimal solution $\hat{\beta}$ satisfies (**normal equations**)

$$\nabla SSE = 2A^T A\hat{\beta} - A^T Y = 0.$$

- The optimal solution is then $\hat{\beta} = (A^T A)^{-1} A^T Y$.

- Of course, $A^T A$ must be invertible!

4 Testing significance

An Overall Measure of How Well the Full Model Performs

- **Coefficient of Multiple Determination:**
  - Denoted as R^2.
  - Defined as the proportion of the variability in the dependent variable y that is accounted for by the independent variables, x_1, x_2, …, x_k, through the regression model.
  - With only one independent variable ($k=1$), $R^2 = r^2$, the square of the simple correlation coefficient.

The coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}, \qquad 0 \le R^2 \le 1$$

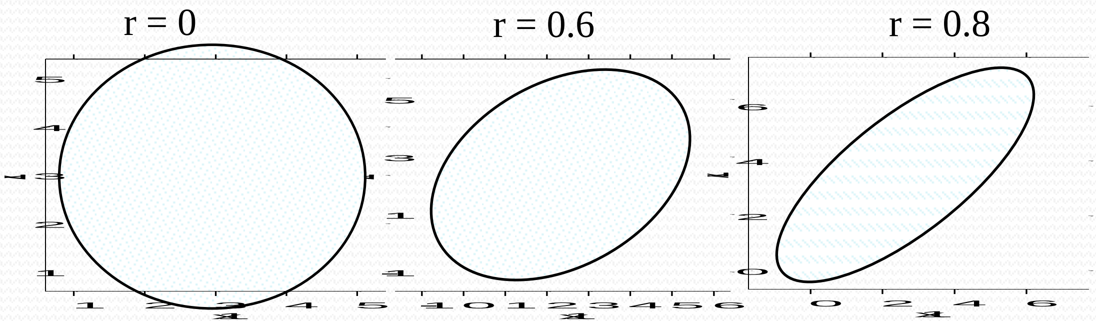
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

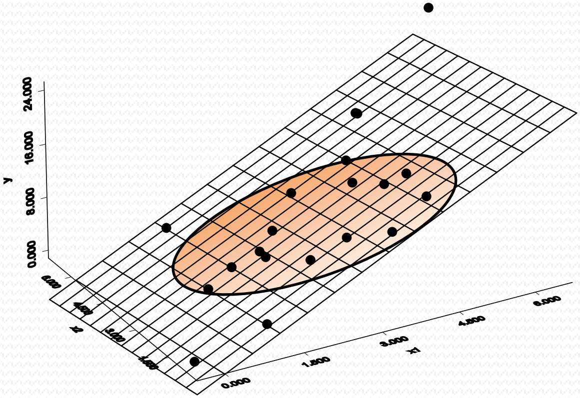
Multicollinearity

- A further **assumption** in multiple regression (absent in SLR), is that the predictors (x_1, x_2, … x_k) are statistically uncorrelated. That is, the predictors do not co-vary.

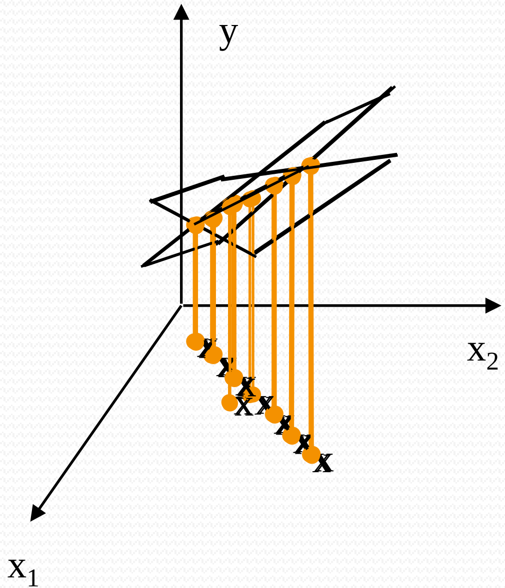
- When the predictors are significantly correlated (correlation greater than about 0.6) then the multiple regression model is said to suffer from problems of multicollinearity.



Effect of Multicollinearity on the Fitted Surface



Extreme collinearity



Effect of Multicollinearity

- **Multicollinearity leads to**
  - Numerical instability in the estimates of the regression parameters: wild fluctuations in these estimates if a few observations are added or removed.
  - No longer have simple interpretations for the regression coefficients in the additive model.
- **Ways to detect multicollinearity:**
  - Scatterplots of the predictor variables.
  - Correlation matrix for the predictor variables: the higher these correlations the worse the problem.
  - Variance Inflation Factors (VIFs): values larger than 10 usually signal a substantial amount of collinearity.
- **What can be done about multicollinearity:**
  - Regression estimates are still computable, but the resulting confidence/prediction intervals are very wide.
  - Choose explanatory variables wisely! (e.g. consider omitting one of two highly correlated variables.).

Variance Inflation Factor

- It measures how much the variances of the estimated regression coefficients are inflated as compared to when the independent variables are not linearly related.

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \cdots k$$

- R_j^2 is the coefficient of determination from the regression of the *j*-th variable on the remaining k-1 variables:

$$x_j = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_{j-1} x_{j-1} + c_{j+1} x_{j+1} + \ldots + c_k x_k + e_j$$

Standard Error for Partial Slope Estimate

- The estimated standard error for $\hat{\beta}_j$

$$s_{\hat{\beta}_j} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{S_{x_j x_j}(1-R_j^2)}}$$

where

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{SSE}{n-(k+1)}}$$

$$S_{x_j x_j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

and R_j^2 is the coefficient of determination from the regression of the *j*-th variable

What happens if all the predictors are truly independent of each other?

$$R_j^2 \to 0 \quad s_{\hat{\beta}_j} \to \frac{\hat{\sigma}_\varepsilon}{\sqrt{S_{x_j x_j}}}$$

If there is high dependency?

$$R_j^2 \to 1 \quad s_{\hat{\beta}_j} \to \infty$$

Global Testing in Multiple Regression

- Testing individual parameters in the model.
- Computing predicted values and associated standard errors.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- Overall F-test
  - H_o: None of the explanatory variables is a significant predictor of Y

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{not all } \beta_i \, (i = 1, \cdots k) \text{ equal zero}$$

$$F = \frac{SSR / k}{SSE / (n-k-1)} = \frac{MSR}{MSE}$$

Reject if: $F > F_{k, n-k-1, \alpha}$

Testing a partial slope

$$H_0: \quad \beta_j = 0$$

Alternatives: Rejection Region:

$$H_1: \quad \beta_j > 0 \longrightarrow t > t_{n-(k+1),\alpha}$$

$$\beta_j < 0 \longrightarrow t < -t_{n-(k+1),\alpha}$$

$$\beta_j \neq 0 \longrightarrow |t| > t_{n-(k+1),\alpha/2}$$

Test Statistic: $t = \dfrac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$

Confidence interval

- 100(1-α)% Confidence Interval for $\hat{\beta}_j$

$$\boxed{\hat{\beta}_j \pm t_{n-(k+1),\alpha/2}\, s_{\hat{\beta}_j}}$$

Reflects the number of data points minus the number of parameters that have to be estimated.

Degree of freedom for SSE

5 Conclusion

Conclusion

- Multiple linear regression: a very useful parametric method to explain a variable with respect to some other variables!
- It is a parametric model since it depends on a known model characterized by a finite number of parameters.
- Many tools for interpreting the results,
- Interpretation should be done carefully,