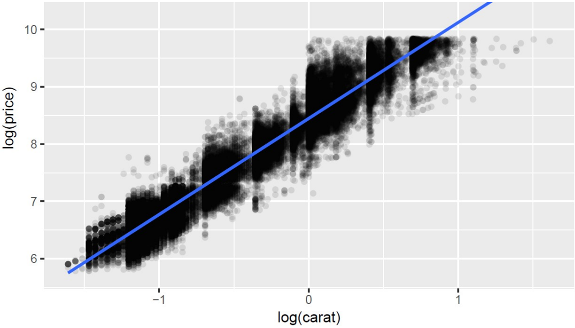
Introduction to AI
Linear Regression



Rodrigo Cabral, Lionel Fillatre and Michel Riveill

EUR DS4H-LIFE-SPECTRUM

cabral@unice.fr

Outline

1. Simple linear regression
2. Model adequacy for simple linear regression
3. Beyond lines: multiple linear regression
4. Model adequacy for multiple linear regression
5. Conclusions

Source: some of these slides are based on lectures 2 and 3 from Machine Learning for Big Data (Polytech Nice-Lionel Fillatre)

1. Simple linear regression

2. Model adequacy for simple linear regression

3. Beyond lines: multiple linear regression

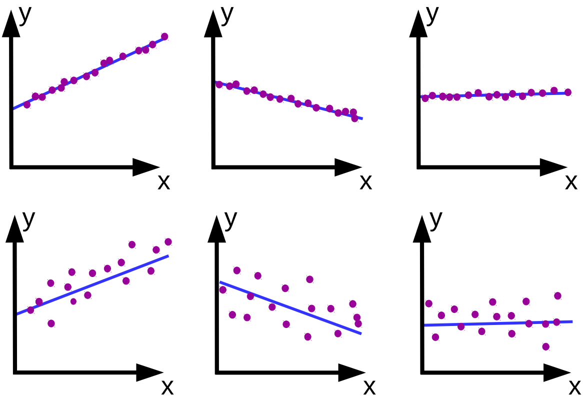
4. Model adequacy for multiple linear regression

5. Conclusions

Linear data

When do we use simple linear regression?

- One input feature $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, one output feature $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ with continuous amplitude.
- Scatter plot look like one of these



Simple linear prediction

What is the prediction model?

- In simple linear regression the prediction function $\hat{f}(x_i)$ is specified as

$$\hat{y}_i = \beta_1 x_i + \beta_0$$

where β_0 and β_1 parametrize the prediction model $\hat{f}_\beta(x_i)$.

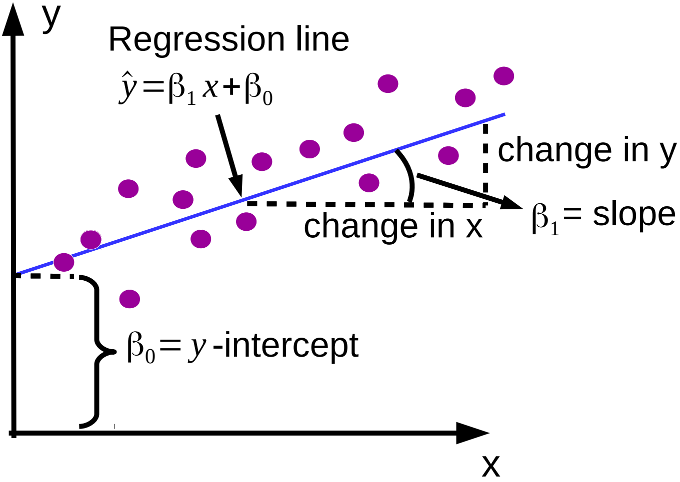
- Or in matrix notation (using matrix-vector product) for K predictions

$$\hat{\mathbf{y}} = [\mathbf{1} \quad \mathbf{x}]\boldsymbol{\beta} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_K \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_K \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Note that the vector $\mathbf{1}$ is an implicit feature vector in this model.

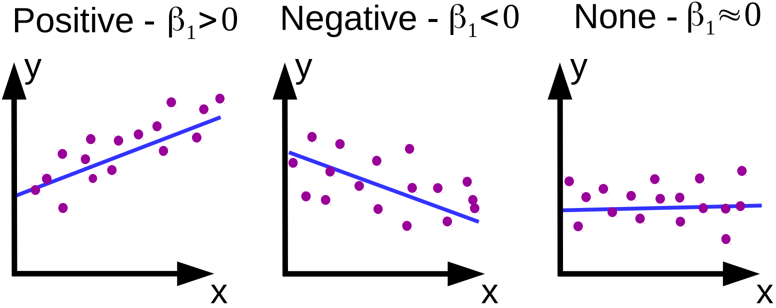
Simple linear prediction

What is the meaning of the parameters β_0 and β_1?



Simple linear prediction

Relationship between *y* and *x* depending on β_1

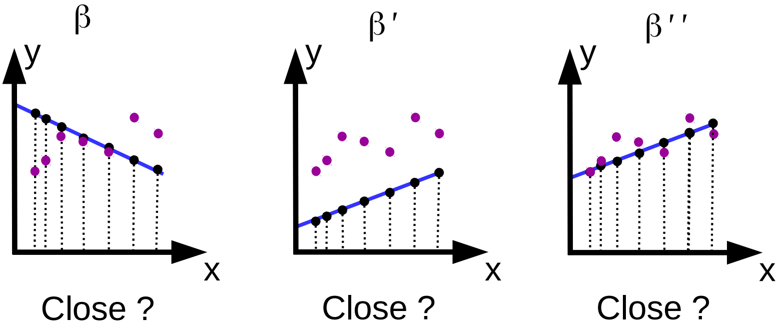


Simple linear prediction

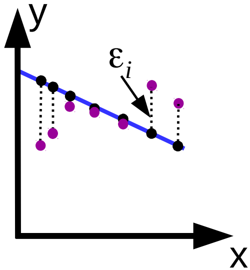
How do we learn from data in this case?

- Answer: find values of β such that $\hat{f}_{\beta}$ gives $\hat{\mathbf{y}}$ close to available $\mathbf{y}$ (training data).

How do we measure closeness between $\mathbf{y}$ and $\hat{\mathbf{y}}$?



Simple linear prediction: least squares criterion



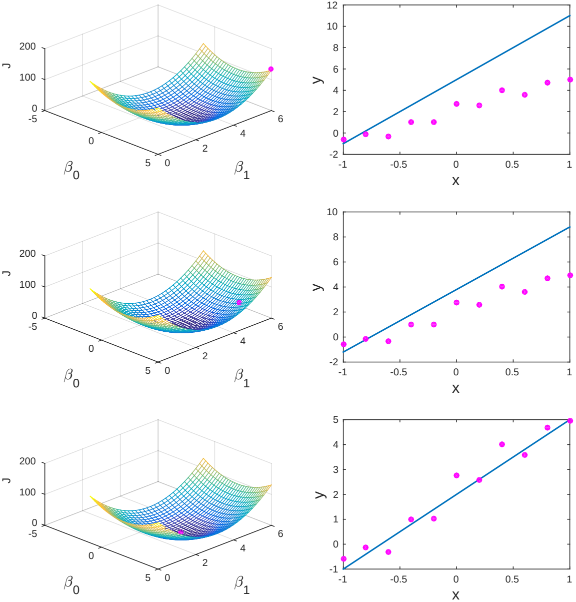
Minimize the sum of the squared residuals $\varepsilon_i = y_i - \hat{y}_i$

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

- J is a function of $\boldsymbol{\beta}$, since y_i and x_i are fixed.
- J should be minimized as a function of $\boldsymbol{\beta}$ - J is a cost function.
- This is called the **least squares** approach.
  - It can be applied to any regression problem.
  - The difficulty lies in solving the minimization problem.

Simple linear prediction: least squares criterion

What is the shape of the cost function?



Simple linear prediction: least squares criterion

What is the shape of the cost function?

- The cost function is convex: bowl-shaped function.
- There is only one global minimum, the point $\boldsymbol{\beta}$ at the bottom, which fits quite well the data.
- Points far from this minimum give quite bad solutions.

How do we find the optimal parameters $\hat{\boldsymbol{\beta}}$?

- Do we test many values for $\boldsymbol{\beta}$?

Simple linear prediction: least squares solution

How do we find the optimal parameters $\hat{\boldsymbol{\beta}}$?

- Hopefully, we have a closed form solution for the optimal parameters $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

where $\quad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad$ and $\quad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$

This can be shown with simple *calculus* (see Appendix 1).

1. Simple linear regression

2. Model adequacy for simple linear regression

3. Beyond lines: multiple linear regression

4. Model adequacy for multiple linear regression

5. Conclusions

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Absolute criteria:
  - **MSE**: evaluate mean squared error (MSE) on the dataset:

$$\text{MSE} = \frac{1}{N-2} J(\hat{\beta}) = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3}$$

    where the prediction is given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

    - The higher it is, the worse the model is.
    - Note that MSE is measured in square units of y.
    - Difficult interpretation from its value.

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Absolute criteria:
  - **RMSE**: evaluate root mean squared error (RMSE) on the dataset:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4}$$

    - The higher it is, the worse the model is.
    - Easier interpretation since in the same units of y.

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Analysis of variance:
  It can be shown that

$$
\begin{array}{ccccc}
\text{Total sum of squares} & = & \text{Sum of squares for regression} & + & \text{Sum of squared errors} \\
\text{SST} & = & \text{SSR} & + & \text{SSE} \\
\sum_{i=1}^{N}(y_i - \bar{y})^2 & = & \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 & + & \sum_{i=1}^{N}(y_i - \hat{y}_i)^2
\end{array}
\tag{5}
$$

- Good model: most variation on y should be explained by variation on $\hat{y}$.

$\Longrightarrow$ SSR should be close to SST.

$\Longrightarrow \dfrac{\text{SSR}}{\text{SST}}$ should be close to 1.

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Relative criterion:
  - **$\mathbf{R^2}$**: evaluate the **coefficient of determination** (R^2) on the dataset:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6}$$

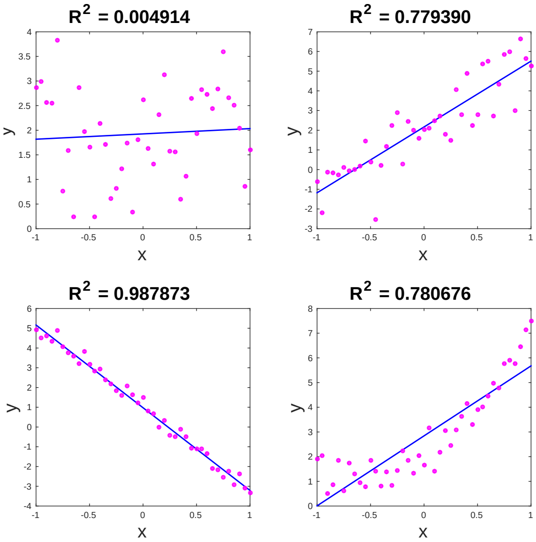
- $0 \leq R^2 \leq 1$
- $R^2 \approx 1 \Longrightarrow$ the model is adequate.
- $R^2 \approx 0 \Longrightarrow$ the model is inadequate.
- What is the limit value on R^2 to decide on adequacy?

$\Longrightarrow$ It depends strongly on the application.

Simple linear prediction: model adequacy

How do we test model adequacy ?

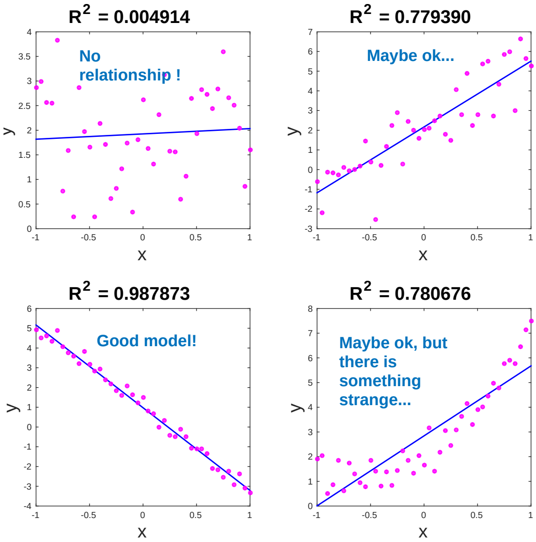
- Different examples and their R^2 :



Simple linear prediction: model adequacy

How do we test model adequacy ?

- Different examples and their R^2 :



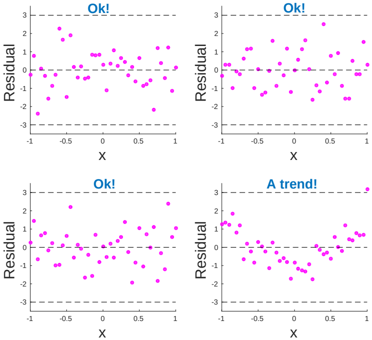
Simple linear prediction: model adequacy

How do we test model adequacy ?

- Residuals analysis :
- Evaluate Studentized (normalized) residuals:

$$\varepsilon_i = \frac{y_i - \hat{y}_i}{\text{RMSE}\sqrt{1 - h_i}} \tag{7}$$

where $h_i = \dfrac{1}{N} + \dfrac{(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$ are the leverage scores.



Simple linear prediction: model adequacy

How do we test model adequacy ?

- Residuals analysis :

Studentized residuals should

- present no trends and no bias

$\Longrightarrow$ linear model explains all deterministic behavior.

$\Longrightarrow$ If not the case, then consider other model

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Residuals analysis :

Studentized residuals should

- present no trends and no bias
  $\Longrightarrow$ linear model explains all deterministic behavior.
  $\Longrightarrow$ If not the case, then consider other model
- be inside or not far from the interval $[-3, 3]$
  $\Longrightarrow$ the approach we use is optimal in statistical sense when the residuals are assumed to be Gaussian distributed with constant variance.
  $\Longrightarrow$ If a small number of points lie far from this interval, these are **outliers** and they should be analyzed carefully (errors in data collection).
  $\Longrightarrow$ If lots of points lie quite far from this interval, then you should look for a different approach than least squares.

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Outliers :

- They can be often detected in scatter plots.

- They can be even more easily detected in residual analysis.

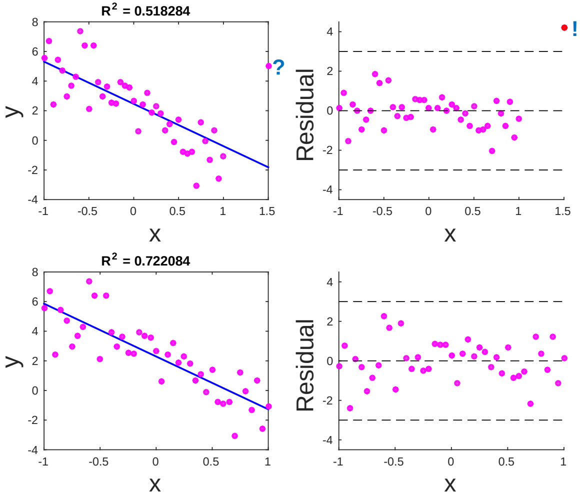
- Analysis of these data points is required.
  $\Longrightarrow$ Errors in the dataset?

- If they are errors, you should remove them.
  $\Longrightarrow$ They can have great effect on model adequacy indicators (MSE, RMSE and R^2), thus leading to wrong conclusions.

Simple linear prediction: model adequacy

How do we test model adequacy ?

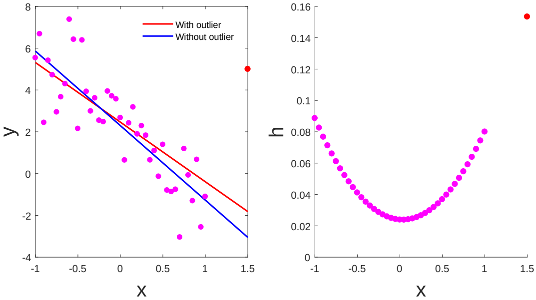
- Outliers and residual analysis:



Simple linear prediction: model adequacy

How do we test model adequacy ?

- Outliers and leverage:



- The leverage scores h_i indicate sensitivity of the prediction line to the i-th observation.
- High leverage outliers have great influence on the prediction line. $\Longrightarrow$ They should be analyzed very carefully, to see if they can be removed from the dataset.

1. Simple linear regression

2. Model adequacy for simple linear regression

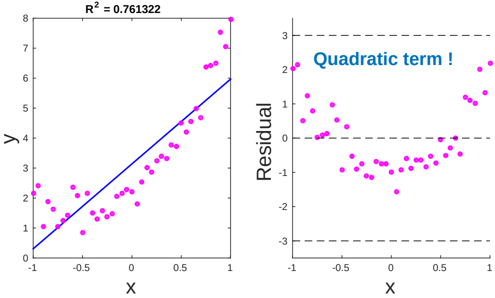
3. Beyond lines: multiple linear regression

4. Model adequacy for multiple linear regression

5. Conclusions

Multiple linear regression: polynomials

Polynomial trend in data



- Quadratic prediction model:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \tag{8}$$

Multiple linear regression: polynomials

Polynomial trend in data

- Quadratic prediction model:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

- You can form a new feature vector with the squares

$$\mathbf{x}^{(2)} = \begin{bmatrix} x_1^2 \\ \vdots \\ x_N^2 \end{bmatrix}$$

then modify the feature matrix to include this new term:

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} \end{bmatrix}$$

where we have redefined $\mathbf{x}^{(1)} = \mathbf{x}$.

Multiple linear regression: polynomials

Polynomial trend in data

- Quadratic prediction model:

$$\hat{y}_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)}$$

- Or in matrix notation (using matrix-product) for K predictions

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_K \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_K & x_K^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

  This is an example of **multiple linear regression**.
- Multiple linear regression: the prediction is a linear combination of one **or more features**.

Multiple linear regression

How do we learn from data in this case?

- Answer: we use the same least squares approach that we have used in simple linear regression.

  $\Longrightarrow$Minimization of $J(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left[y_i - \left(\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} \right) \right]^2$$

- Using vector notation gives a more compact expression

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{9}$$

  - $\cdot^{\mathsf{T}}$ is matrix/vector transpose
  - $\| \cdot \|_2$ is the L_2 norm of a vector.

Multiple linear regression

How do we find the optimal parameters $\hat{\boldsymbol{\beta}}$?

- Closed-form expressions for β_0, β_1 and β_2 minimizing $J(\boldsymbol{\beta})$ in the quadratic case are quite cumbersome.
- It can be shown that a closed-form minimizer for the general problem in vector form, with any **X** and any size of $\boldsymbol{\beta}$, is a solution $\hat{\boldsymbol{\beta}}$ of

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})\,\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{10}$$

  This can be shown with simple *calculus* (see Appendix 2).
- The equations forming this linear system are called **normal equations**.

Multiple linear regression

How do we find the optimal parameters $\hat{\beta}$?

- If none of the columns of $\mathbf{X}$ can be written as a linear combination of the others, then (10) has only one solution.
- Furthermore, the matrix inverse $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ exist and we have

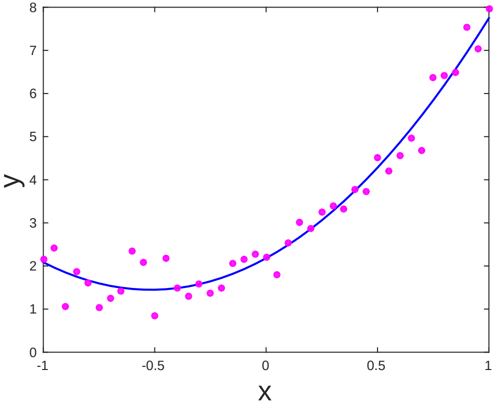
$$\hat{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{11}$$

- Matrix inverse can be calculated with numerical algorithms.
  $\Longrightarrow$ In practice, it is preferred to numerically solve the normal equations then to use the matrix inverse.
- Challenge for the curious: obtain the optimal $\hat{\beta}$ previously given for simple linear regression from this expression.

Multiple linear regression: polynomials

Quadratic model example:

- Prediction with optimal parameters given by (12):



Multiple linear regression: polynomials

Generalization to *p*-th degree polynomials:

- Set a $p+1$ parameter vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

  then create a new feature matrix $\mathbf{X}$ in a similar way:

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \cdots & \mathbf{x}^{(p)} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & & & \\ 1 & x_N & \cdots & x_N^p \end{bmatrix}$$

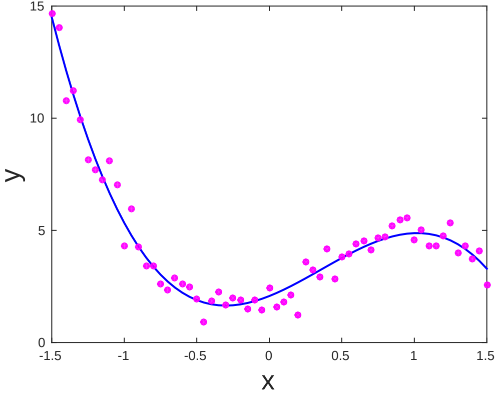
- $\hat{\boldsymbol{\beta}}$ is given again by (12) (or (10)):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \tag{12}$$

Multiple linear regression: polynomials

Cubic model example:

- Prediction with optimal parameters given by (12):



Multiple linear regression: hyperplanes

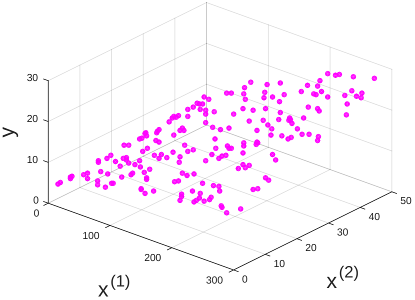
Multiple features:

- What if I really have another measured explanatory variable $\mathbf{x}^{(2)}$?

e.g.

$\mathbf{y}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$
Sales in a market	Spending in TV advertisement	Spending in radio advertisement

- and the data seems to lie approximately in a plane surface



Then mult. linear regression

$$\hat{y}_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_1 x_i^{(2)}$$

is a good prediction model.
$\Longrightarrow$ It is the equation of a plane!

Source: advertisement.csv dataset from `http://faculty.marshall.usc.edu/gareth-james/ISL/data.html`

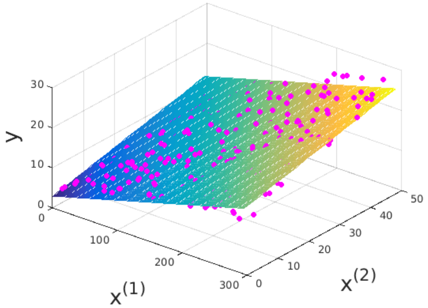
Multiple linear regression: hyperplanes

Multiple features:

- What if I really have another measured explanatory variable $\mathbf{x}^{(2)}$?
- In this case, set $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ and simply concatenate **1** vector

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} \end{bmatrix}$$

- Prediction results:



Multiple linear regression: hyperplanes

Multiple features:

- Generalization to p explanatory variables
- In this case, set $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ and concatenate $\mathbf{1}$ with other feature vectors:

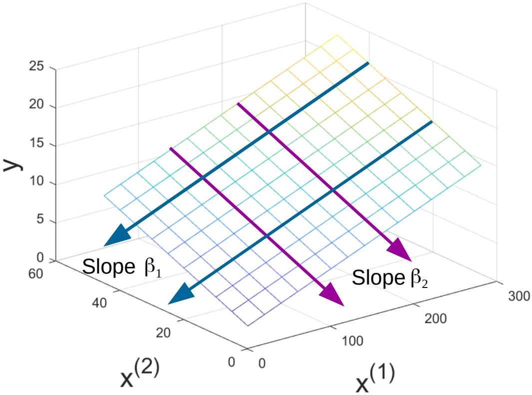
$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \cdots & \mathbf{x}^{(p)} \end{bmatrix}$$

- Prediction model is geometrically an hyper plan.
- β_0 is the y-intercept when all features are zero.
- β_i is the change in $\hat{y}$ for a unitary change in $x^{(i)}$ when all other variables are kept fixed.

Multiple linear regression: hyperplanes

Multiple features:

- β_i is the change in $\hat{y}$ for a unitary change in $x^{(i)}$ when all other variables are kept fixed.



Multiple linear regression: hyperplanes

Interaction models:

- Can we model interactions between variables?
- Prediction model with interaction:

$$\hat{y}_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_{1,2} x_i^{(1)} x_i^{(2)}$$

- You can form a new feature vector with the products

$$\mathbf{x}^{(1,2)} = \begin{bmatrix} x_1^{(1)} x_1^{(2)} \\ \vdots \\ x_N^{(1)} x_N^{(2)} \end{bmatrix}$$

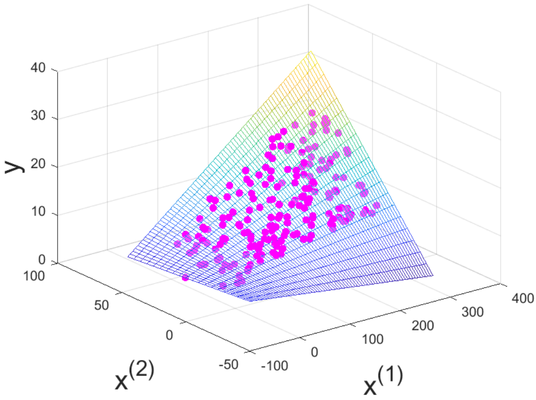
then modify the feature matrix and the parameter vector accordingly

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(1,2)} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{1,2} \end{bmatrix}$$

Multiple linear regression: hyperplanes

Interaction models:

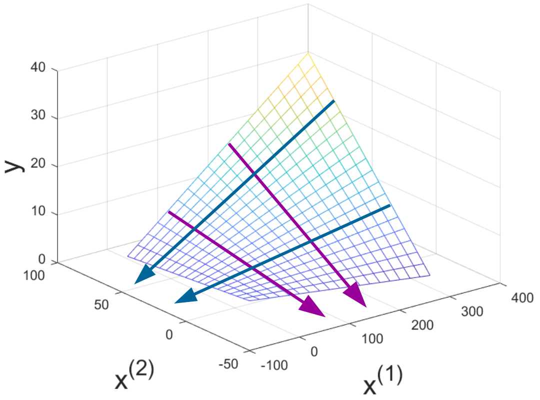
- Twisted planes



Multiple linear regression: hyperplanes

Interaction models:

- The slope for one variable changes as we change the value of the other:



1. Simple linear regression

2. Model adequacy for simple linear regression

3. Beyond lines: multiple linear regression

4. Model adequacy for multiple linear regression

5. Conclusions

Simple linear prediction: model adequacy

How do we test model adequacy ?

- Absolute criteria for p features:
  - **MSE**:

$$\text{MSE} = \frac{1}{N-(p+1)} J(\hat{\beta}) = \frac{1}{N-(p+1)} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{13}$$

  - **RMSE**:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{N-(p+1)} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{14}$$

    - We do not count the **1** vector in the features.
    - There is no change in interpretation with respect to what is presented in simple linear regression.

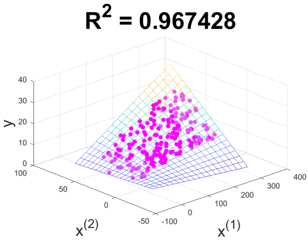
Simple linear prediction: model adequacy

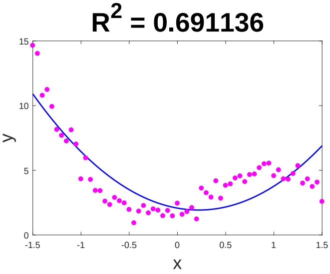
How do we test model adequacy ?

- Relative criterion:
  - **$\mathbf{R^2}$**: it is the same as for linear regression

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

  - Same interpretation on R^2 values as for simple linear regression.





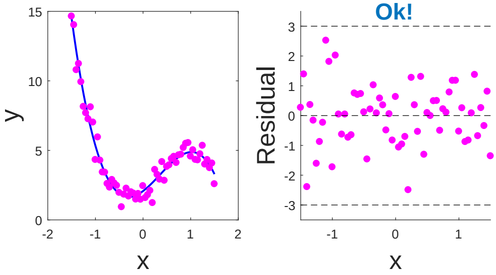
Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Residuals analysis :
- Studentized (normalized) residuals are evaluated in a slightly different way than before:

$$\varepsilon_i = \frac{y_i - \hat{y}_i}{\mathrm{RMSE}\sqrt{1 - h_i}} \tag{15}$$

where $h_i = \left\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right\}_{i,i}$ are the leverage scores for multiple linear regression (i-th element of the diagonal).



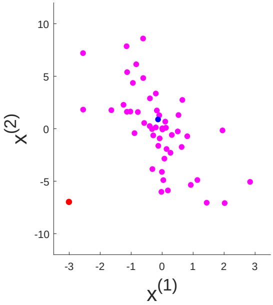
Multiple linear prediction: model adequacy

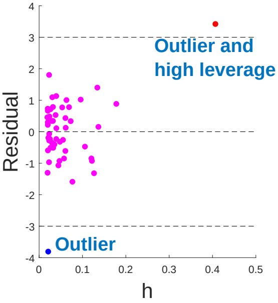
How do we test model adequacy ?

- Residuals/leverage analysis :
- In higher dimensions $p > 3$, residuals analysis is done with observation index on x axis.

  $\Longrightarrow$ More difficult to detect a trend.
- Visualization of outliers and outliers with high leverage cannot be done directly.

  $\Longrightarrow$ Use residuals *vs.* leverage scatter plots:





Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Collinearity :
- If some features can be written as linear combination of the others, we say that the data suffers from **collinearity**.

Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Collinearity :
- If some features can be written as linear combination of the others, we say that the data suffers from **collinearity**.
- Effects of collinearity:
  1. Estimated parameters $\hat{\boldsymbol{\beta}}$ have large variations with addition or removal of a few observations.
     $\Longrightarrow$ Interpretation of $\hat{\beta}_i$ as specific rates of change have no meaning.
  2. Predictions can be quite bad for points outside the collinear pattern.
     $\Longrightarrow$ The model generalization power can be quite poor.

Multiple linear prediction: model adequacy

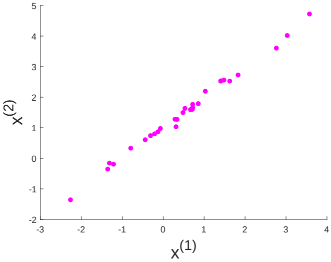
How do we test model adequacy ?

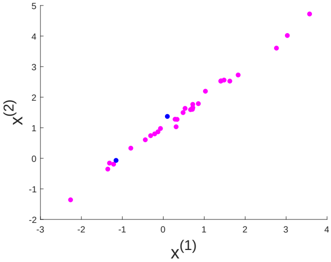
- Collinearity :
- If some features can be written as linear combination of the others, we say that the data suffers from **collinearity**.
- Effects of collinearity:
  1. Estimated parameters $\hat{\beta}$ have large variations with addition or removal of a few observations.
     $\Longrightarrow$ Interpretation of $\hat{\beta}_i$ as specific rates of change have no meaning.
  2. Predictions can be quite bad for points outside the collinear pattern.
     $\Longrightarrow$ The model generalization power can be quite poor.
- How do we deal with collinearity?
  $\Longrightarrow$ Remove features such that the data becomes not collinear.

Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Collinearity :
- An example of fluctuation on $\hat{\beta}$: addition of 2 observations to a collinear dataset with 30 observations.





2 features

$$\hat{\beta} = \begin{bmatrix} 4.356 \\ 5.505 \\ -6.409 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} 3.879 \\ 5.000 \\ -5.906 \end{bmatrix}$$

1 feature ($\mathbf{x}^{(1)}$)

$$\hat{\beta} = \begin{bmatrix} -1.989 \\ -0.886 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} -2.046 \\ -0.863 \end{bmatrix}$$

Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Collinearity:
- How do we measure the level of collinearity of a feature with respect to the other features?

  1. Do linear regression on the dataset $(\mathbf{X}_{-i}, \mathbf{x}_i)$
     i.e. remove $\mathbf{x_i}$ from $\mathbf{X}$ and consider it as the output for regression.
  2. Evaluate the corresponding coefficient of determination R_i^2.
  3. Evaluate the **variance inflation factor** VIF_i:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{16}$$

Multiple linear prediction: model adequacy

How do we test model adequacy ?

- Collinearity:
- How do we measure the level of collinearity of a feature with respect to the other features?
  1. Do linear regression on the dataset $(\mathbf{X}_{-i}, \mathbf{x}_i)$
     i.e. remove $\mathbf{x_i}$ from $\mathbf{X}$ and consider it as the output for regression.
  2. Evaluate the corresponding coefficient of determination R_i^2.
  3. Evaluate the **variance inflation factor** VIF_i:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{16}$$

- If $\mathrm{VIF}_i > 10$, feature $\mathbf{x}_i$ is substantially collinear to the others.
- If for some features $\mathrm{VIF}_i > 10$, remove one of them and re-evaluate collinearity, if you still have some $\mathrm{VIF}_i > 10$, repeat the procedure, otherwise stop removing features.

1. Simple linear regression

2. Model adequacy for simple linear regression

3. Beyond lines: multiple linear regression

4. Model adequacy for multiple linear regression

5. Conclusions

Conclusions

- Linear regression is one of the simplest parametric regression models in machine learning. It assumes that the response is linear with respect to the features.

  $\Longrightarrow$ Data is assumed to live in an hyperplan. Lines and 2D planes are special cases.

Conclusions

- Linear regression is one of the simplest parametric regression models in machine learning. It assumes that the response is linear with respect to the features.

  $\Longrightarrow$ Data is assumed to live in an hyperplan. Lines and 2D planes are special cases.

- Proper transformation of variables can be used to model some non linear trends in data: polynomials, twisted surfaces, *etc*.

  $\Longrightarrow$ Model may have nonlinear dependency on independent variables but has to be linear on the unknown parameters ($\boldsymbol{\beta}$).

Conclusions

- Linear regression is one of the simplest parametric regression models in machine learning. It assumes that the response is linear with respect to the features.

  $\Longrightarrow$ Data is assumed to live in an hyperplan. Lines and 2D planes are special cases.
- Proper transformation of variables can be used to model some non linear trends in data: polynomials, twisted surfaces, *etc*.

  $\Longrightarrow$ Model may have nonlinear dependency on independent variables but has to be linear on the unknown parameters ($\boldsymbol{\beta}$).
- Different tools exist to assess model adequacy to data: MSE, RMSE, R^2 and residual/leverage analysis.

  $\Longrightarrow$ Interpretation of results is application-dependent and should be done with care.

Appendix 1

Appendix 2

Optimal parameters for simple linear regression

- The solution of simple linear regression in the least squares approach are the β_0 and β_1 minimizing (1)

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2$$

- The minima are the critical points, *i.e.* β_0 and β_1 such that $\frac{\partial J}{\partial \beta_0} = 0$ and $\frac{\partial J}{\partial \beta_1} = 0$, with least $J(\boldsymbol{\beta})$.
- For β_0 we have

$$\frac{\partial J}{\partial \beta_0} = \frac{\partial \left\{ \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \right\}}{\partial \beta_0} = -2N\left(\bar{y} - \beta_0 - \beta_1 \bar{x}\right) = 0$$

Therefore,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Optimal parameters for simple linear regression

- For β_1 we have

$$\frac{\partial J}{\partial \beta_1} = \frac{\partial \left\{ \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2 \right\}}{\partial \beta_1} = -2 \sum_{i=1}^{N} x_i \left(y_i - \beta_0 - \beta_1 x_i \right) = 0$$

- Using $\hat{\beta}_0$ previously obtained,

$$\frac{\partial J}{\partial \beta_1} = -2 \sum_{i=1}^{N} x_i \left(y_i - \bar{y} - \beta_1 \bar{x} - \beta_1 x_i \right) = 0$$

- If all x_i are equal, then any β_1 is a solution to the equation above. Otherwise, we have the following,

$$\beta_1 \sum_{i=1}^{N} x_i (x_i - \bar{x}) = \sum_{i=1}^{N} x_i (y_i - \bar{y})$$

Optimal parameters for simple linear regression

- Since $\beta_1 \bar{x} \sum_{i=1}^{N} (x_i - \bar{x}) = 0$ and $\bar{x} \sum_{i=1}^{N} (x_i - \bar{x}) = 0$ we can subtract them from the left and right side respectively, leading to

$$\beta_1 \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

and finally $$\hat{\beta}_1 = \left[\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \right] / \left[\sum_{i=1}^{N} (x_i - \bar{x})^2 \right].$$

- It can be shown that $J(\boldsymbol{\beta})$ is a convex function, as a consequence $\hat{\beta}_0$ and $\hat{\beta}_1$ correspond to a minimum and they are the optimal solution to the least squares approach.

Appendix 1

Appendix 2

Optimal parameters for multiple linear regression

- The solution of multiple linear regression in the least squares approach is the vector $\boldsymbol{\beta}$ minimizing (9)

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- The minima are the critical points which give the least value of $J(\boldsymbol{\beta})$. The critical points are the vectors $\boldsymbol{\beta}$ which gives a zero gradient vector $\nabla_{\boldsymbol{\beta}} J = \mathbf{0}$.
- Developing $J(\boldsymbol{\beta})$

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}$$

- Applying the gradient and using the linearity property we have

$$\begin{array}{ccccc} \nabla_{\boldsymbol{\beta}} J = \nabla_{\boldsymbol{\beta}} \left(\mathbf{y}^{\mathsf{T}}\mathbf{y}\right) & - \nabla_{\boldsymbol{\beta}} \left(\mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right) - \nabla_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}\right) & + \nabla_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0} \\ = \mathbf{0} & - \mathbf{X}^{\mathsf{T}}\mathbf{y} - \mathbf{X}^{\mathsf{T}}\mathbf{y} & + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \end{array}$$

- The optimal solution is then given by (10)

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})\,\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Optimal parameters for multiple linear regression

- Convexity of $J(\boldsymbol{\beta})$ can be easily shown, it is required only to show that the Hessian operator (matrix with second partial derivatives with respect to β_i) is positive semi-definite, which is the case for any $\mathbf{X}$ and $\mathbf{y}$. Therefore the solutions of the normal equations are the minimizers of $J(\boldsymbol{\beta})$, thus representing the optimal solutions of the least squares approach.