Supervised Learning
Regression

Claire Boyer & Demian Wassermann

November 2, 2020

Outline

Motivation

Unemployment

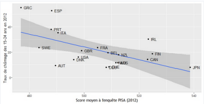
Le taux de chômage peut-il s'expliquer par la qualité de l'éducation ?

LE MONDE | 29.08.2016 à 14h07 • Mis à jour le 29.08.2016 à 14h20 |

Par Romain Damian

Plus d'écoles, moins de chômage ? C'est ce que semble affirmer une étude publiée récemment par la banque Natixis. A l'approche de la rentrée, l'économiste Patrick Artus, dans son « Flash économie » du 3 août 2016 (un papier non académique destiné aux clients de la banque), affirme que « *la qualité du système éducatif et du système de formation professionnelle joue […] un rôle majeur pour expliquer la performance économique et sociale* » d'un pays.

Corrélation ne veut pas dire causalité



On peut aller plus loin que la corrélation simple grâce à une modélisation. Cette dernière, plus complexe, exige de formuler des hypothèses plus simples, mais permettrait d'estimer au mieux l'impact réel de la qualité d'éducation sur la vitalité économique d'un pays en écartant au maximum les effets dus à d'autres facteurs (les politiques monétaires ou d'austérité, par exemple).

- **Data:** OECD study
- **Input:** PISA score
- **Output:** Unemployment rate

Source: OECD

Motivation

Chocolate and Nobel prizes

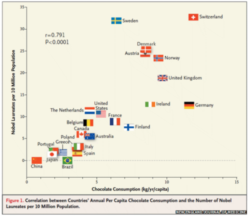


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Nombre de prix Nobel par dix millions d'habitants en fonction de la consommation nationale de chocolat en kilogrammes par personne et par an.

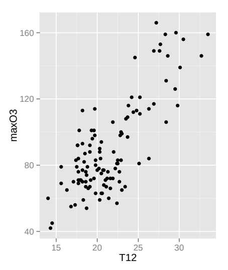
Image : Franz H. Messerli, *The New England Journal of Medicine* **367**(16) (2012), p. 1562-1564

- **Data:** 22 countries (The new England journal of medicine)
- **Input:** Per capita chocolate consumption
- **Output:** Number of Nobel prizes

Source: F. Messerli

Motivation

Ozone pollution

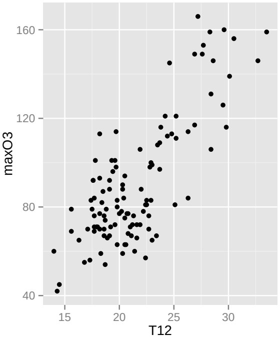


- **Data:** Air Breizh, Summer 2001
- **Input:** Temperature at 12h00
- **Output:** max Ozone concentration

Source: Air Breizh

Outline

Ozone Dataset



Ozone pollution

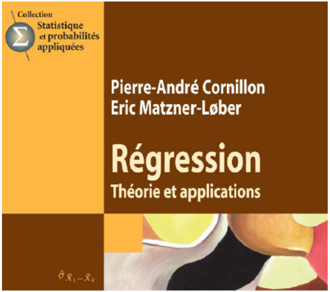
- Two quantities measured at the same location and day during n days:
  - $\underline{X}$: Temperature at 12h00
  - Y : Maximal ozone concentration

Regression Goal

From the practical point of view, the aim is two-fold:

- Adjust a model to *explain* Y from $\underline{X}$
- Adjust a model to *predict* the value of Y for new values of $\underline{X}$.

Bibliography : Pierre-André Cornillon, Eric Matzner-Lober

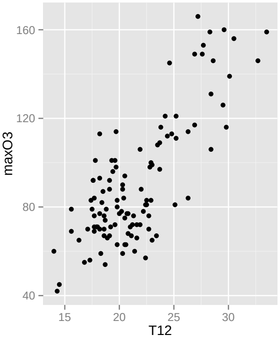


Ozone Dataset

We always start by looking and visualizing the data!

```
112 obs. of  13 variables:
max03 : int  87 82 92 114 94 80 79 79 101 106 ...
T9    : num  15.6 17 15.3 16.2 17.4 17.7 16.8 14.9 16.1 18.3 ...
T12   : num  18.5 18.4 17.6 19.7 20.5 19.8 15.6 17.5 19.6 21.9 ...
T15   : num  18.4 17.7 19.5 22.5 20.4 18.3 14.9 18.9 21.4 22.9 ...
Ne9   : int  4 5 2 1 8 6 7 5 2 5 ...
Ne12  : int  4 5 5 1 8 6 8 5 4 6 ...
Ne15  : int  8 7 4 0 7 7 8 4 4 8 ...
Vx9   : num  0.695 -4.33 2.954 0.985 -0.5 ...
Vx12  : num  -1.71 -4 1.879 0.347 -2.954 ...
Vx15  : num  -0.695 -3 0.521 -0.174 -4.33 ...
max03v: int  84 87 82 92 114 94 80 99 79 101 ...
vent  : Factor w/ 4 levels "Est","Nord","Ouest",..: 2 2 1 2 3 3 3 2 2 3 ...
pluie : Factor w/ 2 levels "Pluie","Sec": 2 2 2 2 2 1 2 2 2 2 ...
```

Ozone Dataset



Ozone pollution

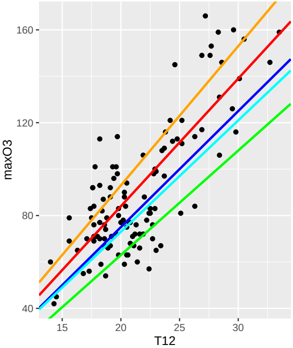
- Two quantities measured at the same location and day during n days:
  - $\underline{X}$: Temperature at 12h00
  - Y : Maximal ozone concentration

Least Squares Regression

Empirical Goal

- **Goal:** Finding a function f such that
$$y_i \approx f(\underline{x}_i).$$
- **$\approx$?:** Need to choose a criterion quantifying the quality of the fit of f to the data by a loss $\ell(y, f(\underline{x}))$.
- **Function?:** Need to specify a function class $\mathcal{S}$ in which to choose f.
- **Least Squares Regression:**
$$\hat{f} = \arg\min_{f \in \mathcal{S}} \sum_{i=1}^{n} (y_i - f(\underline{x}_i))^2$$
where we use the quadratic loss $\ell(y, f(\underline{x})) = (y - f(\underline{x}))^2$

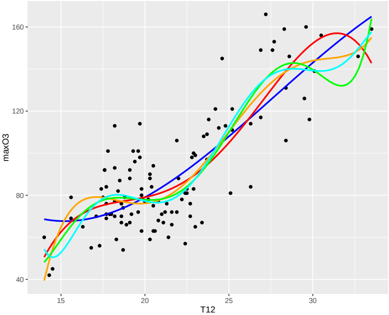
$\mathcal{S}$: Family of linear functions



Empirical Goal

- Find among all the possible lines the one that minimizes the sum of the squared distance between the line and the observations.

$\mathcal{S}$: Family of polynomial functions



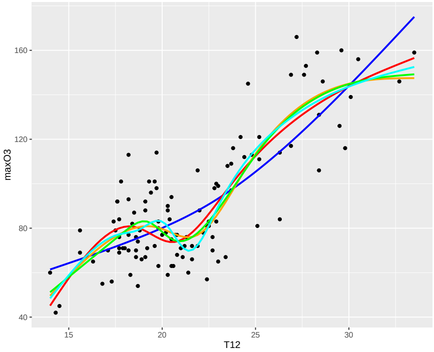
- Polynomials of degree 3, 4, 5, 6 and 7.

Empirical Goal

- Find among all the possible polynomials the one that minimizes the sum of the squared distance between the function and the observations

- **Issue:** How to select the *good* degree!

$\mathcal{S}$: More complex family

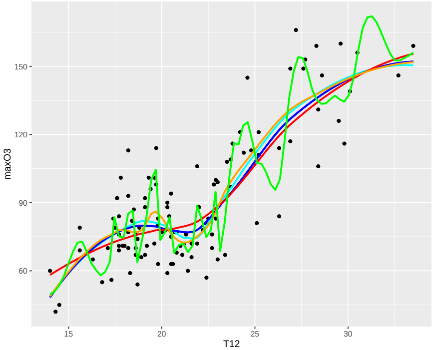


- Spline family...

Empirical Goal

- Find among all the possible splines the one that minimizes the sum of the squared distance between the function and the observations

$\mathcal{S}$: More complex family

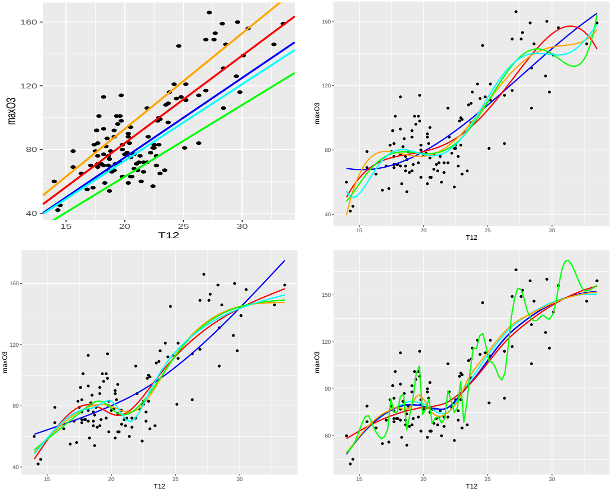


- Kernel estimate family...

Empirical Goal

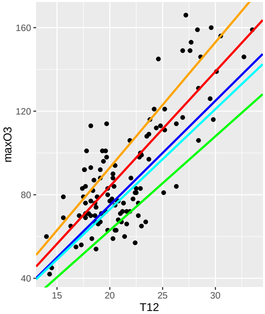
- Find among all the possible kernel estimates the one that minimizes the sum of the squared distance between the function and the observations

$\mathcal{S}$: Much more complex family



- Which model to choose? Linear, Polynomial, Spline, Kernel?

$\mathcal{S}$: Family of linear functions



$$\mathcal{S} = \{f : f_\beta(\mathtt{T12}) = \beta^{(0)} + \beta^{(1)}\mathtt{T12} \quad \beta^{(0)} \in \mathbb{R}, \beta^{(1)} \in \mathbb{R}\}$$

Empirical Goal

- Find among all the possible lines the one that minimizes the sum of the squared distance between the line and the observations.

Least Squares

Least Squares Approach

- Goodness criterion:

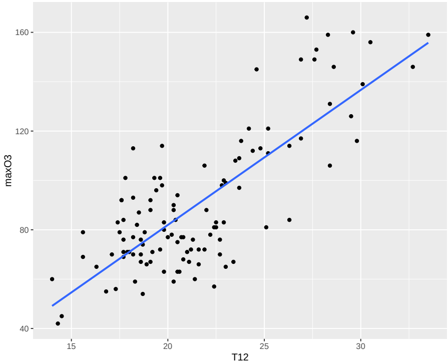
$$\sum_{i=1}^{n} |Y_i - f_\beta(\underline{X}_i)|^2 = \sum_{i=1}^{n} |\mathrm{maxO3}_i - f_\beta(\mathtt{T12}_i)|^2$$
$$= \sum_{i=1}^{n} |\mathrm{maxO3}_i - (\beta^{(0)} + \beta^{(1)}\mathtt{T12}_i)|^2$$

- Choice of β that minimizes this criterion!

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^2} \sum_{i=1}^{n} |\mathrm{maxO3}_i - (\beta^{(0)} + \beta^{(1)}\mathtt{T12}_i)|^2$$

- **Rk:** Easy minimization with an explicit solution!

Prediction



Linear prediction

- Linear prediction for the ozone maximum:

$$\widehat{\text{maxO3}} = f_{\widehat{\beta}}(\texttt{T12}) = \widehat{\beta}^{(0)} + \widehat{\beta}^{(1)}\texttt{T12}$$

Regression model

Statistical Modeling

- The collection of n observations $(\underline{x}_i, y_i)$, $i = 1, \ldots, n$ is assumed to Y_i satisfying
$$Y_i = f(\underline{x}_i) + \epsilon_i \text{ for all } i = 1, \ldots, n$$
where
  - $\underline{x}_i$ are some non (necessarily) random covariates
  - f is an unknown function.
  - ϵ_i are centered random variables.

- The random variable ϵ_i is often called an *error*.
- The *statisticians* may make assumption on the law of ϵ_i to obtain some results...

Outline

Data

Observations	Y	$\underline{X}^\top$				
1	y_1	$\underline{x}_1^{(1)}$	...	$\underline{x}_1^{(j)}$	...	$\underline{x}_1^{(d)}$
2	y_2	$\underline{x}_2(1)$	...	$\underline{x}_2^{(j)}$	...	$\underline{x}_2^{(d)}$
$\cdots$	...	...	...	...	...	...
i	y_i	$\underline{x}_i^{(1)}$	...	$\underline{x}_i^{(j)}$	...	$\underline{x}_i^{(d)}$
$\cdots$	...	...	...	...	...	...
n	y_n	$\underline{x}_n^{(1)}$	...	$\underline{x}_n^{(j)}$	...	$\underline{x}_n^{(d)}$

Regression Equation

Multivariate Regression

- Corresponds to $Y \in \mathbb{R}$ and $\underline{X} \in \mathbb{R}^d$
- Prediction model:

$$f_\beta(\underline{X}) = \beta^{(0)} + \sum_{j=1}^{d} \beta^{(j)} \underline{X}^{(j)}$$

with an unknown parameter $\beta \in \mathbb{R}^{d+1}$

- Example :
  - Ozone univariate regression:
    - $Y = \text{maxO3}$ and $\underline{X} = \left(\texttt{T12}\right)$
    - $f_\beta(\underline{X}) = \beta^{(0)} + \beta^{(1)} \times \texttt{T12} = \beta^{(0)} + \beta^{(1)}\texttt{T12}$
  - Ozone multivariate regression:
    - $Y = \text{maxO3}$ and $\underline{X} = \begin{pmatrix} \texttt{T12} \\ \texttt{Vx} \\ \texttt{Ne12} \end{pmatrix}$
    - $f_\beta(\underline{X}) = \beta^{(0)} + \beta^{(1)}\texttt{T12} + \beta^{(2)}\texttt{Vx} + \beta^{(3)}\texttt{Ne12}$

Least Squares

Least Squares

- Empirical quadratic loss:

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \left(\beta^{(0)} + \sum_{j=1}^{d}\beta^{(j)}\underline{X}_i^{(j)}\right)\right|^2$$

- Least Squares:

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \left(\beta^{(0)} + \sum_{j=1}^{d}\beta^{(j)}\underline{X}_i^{(j)}\right)\right|^2$$

- Simple minimization problem with an explicit solution.

Implicit Goal

- Minimization of the expected quadratic loss:

$$\ell(Y, f(\underline{X})) = \mathbb{E}\left[\left|Y - \left(\beta^{(0)} + \sum_{j=1}^{d}\beta^{(j)}\underline{X}^{(j)}\right)\right|^2\right]$$

Matrix Formulation

Linear Model and Scalar Products

- Scalar product notation:

$$\beta^{(0)} + \sum_{j=1}^{d} \beta^{(j)} \underline{X}^{(j)} = \left\langle \begin{pmatrix} 1 \\ \underline{X} \end{pmatrix}, \beta \right\rangle = \begin{pmatrix} 1 \\ \underline{X} \end{pmatrix}^{\top} \beta$$

- **For sake of simplicity, we will identify $\underline{X}$ and $\begin{pmatrix} 1 \\ \underline{X} \end{pmatrix}$**
- **Linear Predictor:** $f_\beta(\underline{X}) = \langle \underline{X}, \beta \rangle = \underline{X}^\top \beta$
- **Goal:** For all i, $f(\beta)(\underline{X}_i) = \underline{X}_i{}^\top \beta \simeq Y_i$.

Matrix Formulation

Matrix Rewriting

- **Goal:**

$$\begin{pmatrix} \underline{X}_1{}^\top \beta \\ \vdots \\ \underline{X}_n{}^\top \beta \end{pmatrix} = \begin{pmatrix} \underline{X}_1{}^\top \\ \vdots \\ \underline{X}_n{}^\top \end{pmatrix} \beta = \underbrace{\mathbb{X}}_{\text{design matrix}} \beta \simeq \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\underline{Y}}_{\text{output vector}}$$

Matrix Formulation

Squared Loss

- Squared loss:

$$\frac{1}{n}\sum_{i=1}^{n}|Y_i - f_\beta(\underline{X}_i)|^2$$

- Equivalent matrix formulation:

$$\|\underline{Y} - \mathbb{X}\beta\|^2$$

Least Squares Formula

- Least squares estimate:

$$\widehat{\beta} = \operatorname{argmin} \|\underline{Y} - \mathbb{X}\beta\|^2.$$

- First order optimality condition:

$$-2\mathbb{X}^\top(\underline{Y} - \mathbb{X}\beta) = 0 \Leftrightarrow \mathbb{X}^\top\mathbb{X}\beta = \mathbb{X}^\top\underline{Y}$$

- If $\mathbb{X}^\top\mathbb{X}$ is invertible, the unique solution is given by

$$\widehat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\underline{Y}$$

Gaussian regression and ML

- Gaussian model: ϵ i.i.d. $\mathcal{N}(0, \sigma^2)$ (very strong assumption)!
- Likelihood: $\mathbb{P}_{\beta,\sigma}\left(\underline{Y}|\mathbb{X}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\|\underline{Y}-\mathbb{X}\beta\|^2/(2\sigma^2)}$
- Opposite of the log-likelihood:
$$-\log \mathbb{P}_{\beta,\sigma}\left(\underline{Y}|\mathbb{X}\right) = \frac{n}{2}(\log(2\pi) + \log\sigma^2) + \frac{1}{2\sigma^2}\|\underline{Y} - \mathbb{X}\beta\|^2$$
- Maximum likelihood estimate = Least Squares for β!
- Small difference for σ^2...

Outline

Supervised Learning

Supervised Learning Framework

- Input measurement $\underline{X} \in \mathcal{X}$
- Output measurement $Y \in \mathcal{Y}$.
- $(\underline{X}, Y) \sim \mathbb{P}$ with $\mathbb{P}$ unknown.
- **Training data** : $\mathcal{D}_n = \{(\underline{X}_1, Y_1), \ldots, (\underline{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbb{P}$)

- Often
  - $\underline{X} \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$ (classification)
  - or $\underline{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (regression).
- A **predictor** is a function in $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y} \text{ meas.}\}$

Goal

- Construct a **good** predictor $\widehat{f}$ from the training data.

- Need to specify the meaning of good.
- Classification and regression are almost the **same** problem!

Loss and Probabilistic Framework

Loss function for a generic predictor

- **Loss function**: $\ell(Y, f(\underline{X}))$ measures the goodness of the prediction of Y by $f(\underline{X})$
- Examples:
  - Prediction loss: $\ell(Y, f(\underline{X})) = \mathbf{1}_{Y \neq f(\underline{X})}$
  - Quadratic loss: $\ell(Y, f(\underline{X})) = |Y - f(\underline{X})|^2$

Risk function

- Risk measured as the average loss for a new couple:

$$\boldsymbol{\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[\ell(Y, f(\underline{X}))\right]}$$

- Examples:
  - Prediction loss: $\mathbb{E}\left[\ell(Y, f(\underline{X}))\right] = \mathbb{P}\left(Y \neq f(\underline{X})\right)$
  - Quadratic loss: $\mathbb{E}\left[\ell(Y, f(\underline{X}))\right] = \mathbb{E}\left[|Y - f(\underline{X})|^2\right]$

- **Beware:** As $\widehat{f}$ depends on $\mathcal{D}_n$, $\mathcal{R}(\widehat{f})$ is a random variable!

Best Solution

- The best solution f^* (which is independent of $\mathcal{D}_n$) is

$$f^* = \arg\min_{f \in \mathcal{F}} R(f) = \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(\underline{X}))\right] = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\underline{X}}\left[\mathbb{E}_{Y|\underline{X}}\left[\ell(Y, f(\underline{X}))\right]\right]$$

Bayes Predictor (explicit solution)

- In regression with the quadratic loss

$$f^*(\underline{X}) = \mathbb{E}\left[Y|\underline{X}\right]$$

Issue: Solution requires to **know** $\mathbb{E}\left[Y|\underline{X}\right]$ for all values of $\underline{X}$!

Conditional Law and Expectation

Conditioning

- **Conditioning** is a powerful probabislitic tool!
- Behavior of a random variable when some values are **known**.
- Example: $Y|\underline{X}$
  - Behavior of Y for a fixed $\underline{X}$.
  - Depends on the value of $\underline{X}$.

- Subtle mathematical definition!

Conditional Law and Conditional Expectation

	Discrete Case	Continuous Case
Cond. Law	$\mathbb{P}\left(Y\|\underline{X}\right) = \frac{\mathbb{P}\left(Y \cap \underline{X}\right)}{\mathbb{P}\left(\underline{X}\right)}$	$d\mathbb{P}\left(Y\|\underline{X}\right) \simeq \frac{d\mathbb{P}\left(Y \cap \underline{X}\right)}{d\mathbb{P}\left(\underline{X}\right)}$
Cond. Exp.	$\mathbb{E}\left[Y\|\underline{X}\right] = \sum_y y\mathbb{P}\left(Y = y\|\underline{X}\right)$	$\mathbb{E}\left[Y\|\underline{X}\right] = \int_y y d\mathbb{P}\left(Y = y\|\underline{X}\right)$

Goal

Machine Learning

- Learn a rule to construct a **predictor** $\widehat{f} \in \mathcal{F}$ from the training data $\mathcal{D}_n$ s.t. **the risk** $\mathcal{R}(\widehat{f})$ is **small on average** or with high probability with respect to $\mathcal{D}_n$.

- In practice, the rule should be an algorithm!

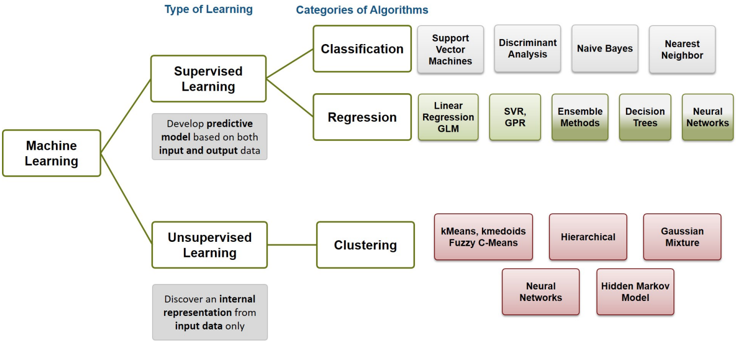
Canonical example: Empirical Risk Minimizer

- One restricts f to a subset of functions $\mathcal{S} = \{f_\theta, \theta \in \Theta\}$
- One replaces the minimization of the average loss by the minimization of the empirical loss

$$\widehat{f} = f_{\widehat{\theta}} = \operatorname*{argmin}_{f_\theta, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(\underline{X}_i))$$

- Example: univariate linear regression!

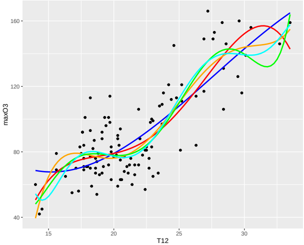
Machine Learning



Source: MathWorks

Outline

$\mathcal{S}$: Family of polynomial functions



- Polynomials of degree 3, 4, 5, 6 and 7.

Fixed degree model

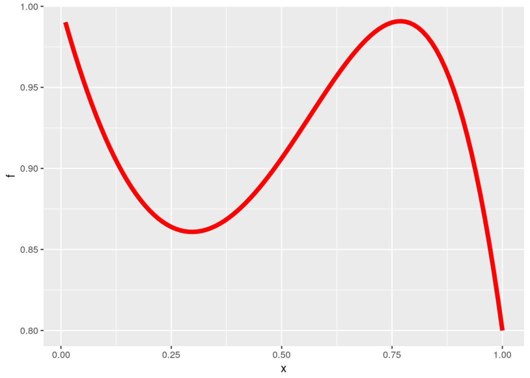
- Fixed degree polynomial model: $f_\beta(\underline{X}_i) = \sum_{l=0}^{d} \beta^{(l)} \underline{X}_i^l$
- Linear in β!
- Amounts to use $\underline{X}_i' = (1, \underline{X}_i, \ldots, \underline{X}_i^d)^\top$
- Easy least squares estimation!

- **Issue:** How to select the *good* degree!

Toy Example

- Illustration of the difficulty with an artifical dataset.

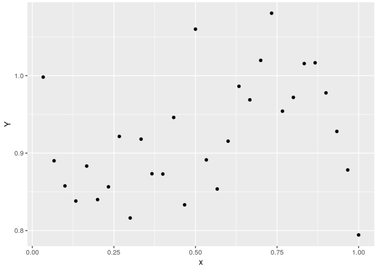
Known function to estimate



- Polynomial of degree 5:

$$f(\underline{x}) = 1 - \underline{x} + 2\underline{x}^2 - 0.8\underline{x}^3 + 0.6\underline{x}^4 - \underline{x}^5$$

Probabilistic Modeling



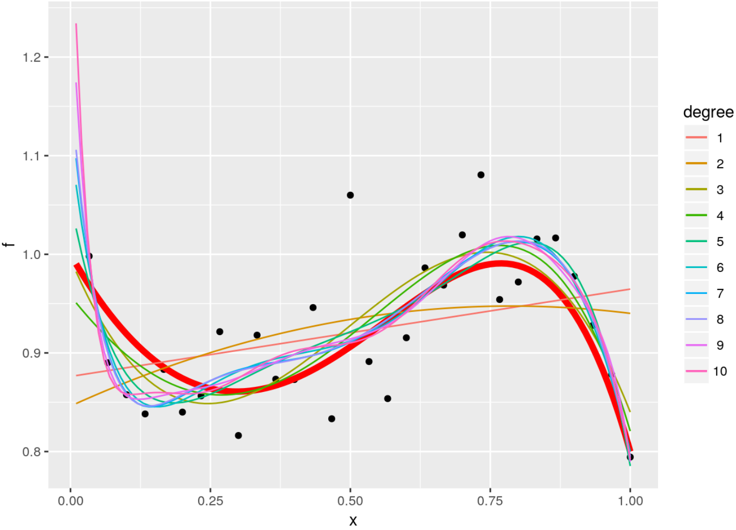
Fixed Design

- Observation on a uniform grid $\underline{x}_k = k/n$, with $1 \leq k \leq n$
- Observed values Y_k are the values of f at $\underline{x}_k$ corrupted by a noise:
$$Y_k = f(k/n) + \epsilon_k$$
- The noises ϵ_k are centered.

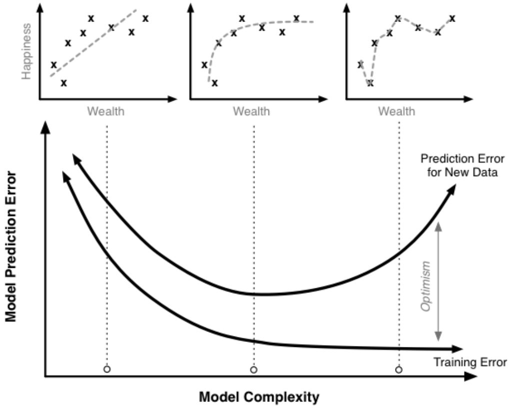
- Here, (ϵ_k) is an i.i.d. centered Gaussian seq. of variance σ^2.

Source: Unknown

Which Degree?

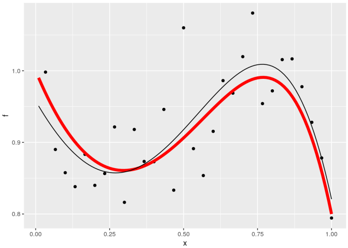


Overfitting Issue



Source: A. Ng

Best Polynomial

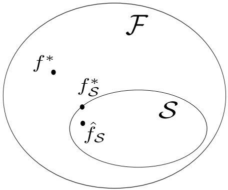


Estimation Result

- Best choice for $\hat{f}_{\hat{m}}$: regression with a polynomial of degree 4 (and not 5 as the true one)
- Degree depends on the amount of noise and the number of observations.

- **The best function for prediction may come from a different model than the true one!**

Bias-Variance Dilemma



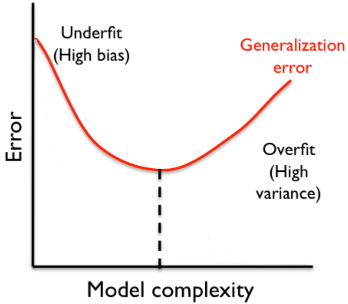
- General setting:
  - $\mathcal{F} = \{\text{measurable functions } \mathcal{X} \to \mathcal{Y}\}$
  - Best solution: $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
  - Class $\mathcal{S} \subset \mathcal{F}$ of functions
  - Ideal target in $\mathcal{S}$: $f^*_{\mathcal{S}} = \operatorname{argmin}_{f \in \mathcal{S}} \mathcal{R}(f)$
  - Estimate in $\mathcal{S}$: $\widehat{f}_{\mathcal{S}}$ obtained with some procedure

Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f^*_{\mathcal{S}}) - \mathcal{R}(f^*)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^*_{\mathcal{S}})}_{\text{Estimation error}}$$

- Approx. error can be large if the model $\mathcal{S}$ is not suitable.
- Estimation error can be large if the model is complex.

Under-fitting / Over-fitting Issue



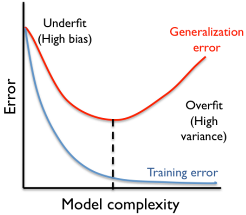
- Different behavior for different model complexity
- **Low complexity model** are easily learned but the approximation error (**bias**) may be large (**Under-fit**).
- **High complexity model** may contain a good ideal target but the estimation error (**variance**) can be large (**Over-fit**)

Bias-variance trade-off $\iff$ avoid **overfitting** and **underfitting**

Source: Unknown

Outline

Training Error Issue



Error behaviour

- Learning/training error (error made on the learning/training set) decays when the complexity of the model increases.
- Quite different behavior when the error is computed on new observations (generalization error).

- Overfit for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of polynomial fit...)
- Need to use a different criterion than the training error!

Source: JMP

Cross Validation and Error Correction

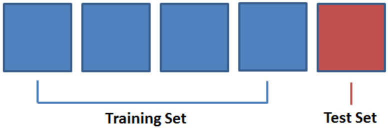
Two Approaches

- **Cross validation:** Very efficient (and almost always used in practice!) but slightly biased as it target uses only a fraction of the data.
- **Correction approach:** use empirical loss criterion but *correct* it with a term increasing with the complexity of $\mathcal{S}$
$$R_n(\widehat{f}_{\mathcal{S}}) \to R_n(\widehat{f}_{\mathcal{S}}) + \mathrm{cor}(\mathcal{S})$$
and choose the model with the smallest corrected risk.

Which loss to use?

- The loss used in the risk: most natural!
- The loss used to estimate $\widehat{\theta}$: penalized estimation!

Cross Validation



- **Very simple idea:** use a second learning/verification set to compute a verification error.
- Sufficient to remove the dependency issue!
- Implicit random design setting...

Cross Validation

- Use $(1-\epsilon) \times n$ observations to train and $\epsilon \times n$ to verify!
- Possible issues:
  - Validation for a learning set of size $(1-\epsilon) \times n$ instead of n ?
  - Unstable error estimate if ϵn is too small ?

- Most classical variations:
  - Hold Out,
  - Leave One Out,
  - V-fold cross validation.

Source: JMP

V-fold Cross Validation

Principle

- Split the dataset $\mathcal{D}$ in V sets $\mathcal{D}_v$ of almost equals size.
- For $v \in \{1, .., V\}$:
  - Learn $\widehat{f}^{-v}$ from the dataset $\mathcal{D}$ minus the set $\mathcal{D}_v$.
  - Compute the empirical error:

$$\mathcal{R}_n^{-v}(\widehat{f}^{-v}) = \frac{1}{n_v} \sum_{(\underline{X}_i, Y_i) \in \mathcal{D}_v} |Y_i - \widehat{f}^{-v}(\underline{X}_i)|^2$$

- Compute the average empirical error:

$$\mathcal{R}_n^{CV}(\widehat{f}) = \frac{1}{V} \sum_{v=1}^{V} \mathcal{R}_n^{-v}(\widehat{f}^{-v})$$

- Leave One Out : $\mathrm{V} = \mathrm{n}$.

Source: JMP

V-fold Cross Validation

Analysis (when n is a multiple of V)

- The $\mathcal{R}_n^{-v}(\widehat{f}^{-v})$ are identically distributed variable but are not independent!
- Consequence:

$$\mathbb{E}\left[\mathcal{R}_n^{CV}(\widehat{f})\right] = \mathbb{E}\left[\mathcal{R}_n^{-v}(\widehat{f}^{-v})\right]$$

$$\mathbb{V}\mathrm{ar}\left[\mathcal{R}_n^{CV}(\widehat{f})\right] = \frac{1}{V}\,\mathbb{V}\mathrm{ar}\left[\mathcal{R}_n^{-v}(\widehat{f}^{-v})\right] + (1-\frac{1}{V})\,\mathbb{C}\mathrm{ov}\left[\mathcal{R}_n^{-v}(\widehat{f}^{-v}), \mathcal{R}_n^{-v'}(\widehat{f}^{-v'})\right]$$

- Average risk for a sample of size $(1-\frac{1}{V})n$.
- Variance term much more complex to analyze!
- Fine analysis shows that the larger V the better...

- Accuracy/Speed tradeoff: $V=5$ or $V=10$!

Error Correction

- Empirical loss of an estimator computed on the dataset used to chose is is biased!
- Empirical loss is an optimistic estimate of the true loss.

Risk Correction Heuristic

- Estimate an upper bound of this optimism for a given family.
- Correct the empirical loss by adding this upper bound.

- **Rk:** Finding such an upper bound can be complicated!
- Correction often called a **penalty**.

Penalization

Penalized Loss

- Minimization of

$$\operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(\underline{X}_i)) + \text{pen}(\theta)$$

where $\text{pen}(\theta)$ is an error correction (penalty).

Penalties

- Upper bound of the optimism of the empirical loss
- Depends on the loss and the framework!

Instantiation

- Mallows Cp: Least Squares with $\text{pen}(\theta) = 2\frac{d}{n}\sigma^2$.
- AIC Heuristics: Maximum Likelihood with $\text{pen}(\theta) = \frac{d}{n}$.
- BIC Heuristics: Maximum Likelohood with $\text{pen}(\theta) = \log(n)\frac{d}{n}$.

Outline

Variable Selection

- **Which submodels is the most interesting one?**
- Keep only a subset of the variable.

Variable Selection Goal(s)

- **Description:** What are the most influent variables?
- **Prediction:** Due to the bias/variance tradeoff, the best prediction is not necessarily obtained with the most complex model.

- How to find a good **sparse** submodel?
- **Rk:** Performance criterion (Cp, AIC, BIC, CV...) often measures some prediction ability.

Exhaustive Search

- **Easy** least squares minimization for **a given** subset of variables.

Exact Minimization

- Compute the least squares for all the possibles subsets.
- Compute a performance criterion for all those subsets.
- Pick the one with the best performance criterion.

- **Issue:** In dimension p, 2^p different subsets.
- Combinatorial problem too expensive when p is not small.
- **Rk:** Classical performance criterion may be too optimistic when there are too many models.

Clever Exploration

Clever Exploration

- Minimization of the criterion but without an exhaustive exploration of the subsets.
- Generic strategy:
    - Start with a pool of subsets of size P
    - Create a larger pool of size PC by adding and/or removing variables from the previous subset
    - Keep only the best P subset according to the criterion and iterate
- Variations on the size of the subsets, the initial subsets, the rule to add and remove variables, the criterion...
- Forward, Backward, Forward/Backward, Stochastic (Genetic) Algorithm...

Clever Exploration

Forward strategy

- Start with an empty model
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

- **Rk:** Fisher test can also be used to decide.

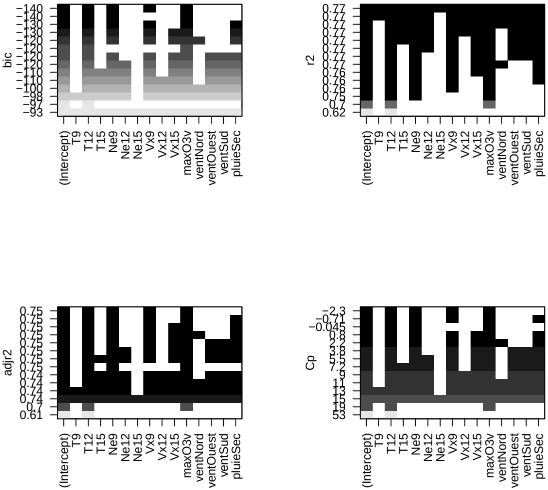
Clever Exploration

Forward/Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time) and to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

- Various Stochastic (Genetic) Algorithm.
- Stability issue...

Exhaustive Search



Forward-Backward


```
Start:  AIC=612.99
maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
maxO3v + vent + pluie

Step:  AIC=608.61
maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
maxO3v + pluie

.........

Step:  AIC=596.02
maxO3 ~ T12 + Ne9 + Vx9 + maxO3v
```

Outline

Natural questions

Effect of an explanatory variable

- Is the variable $\underline{X}^{(j)}$ useful ?
- One need a hypothesis test to answer this question.

Model

- Is the model suitable?
- One need a hypothesis test to answer this question.

Fisher's Approach

- Hypothesis H_0 (Null hypothesis) to be **nullified**
- Construction of a random variable T, the test statistic, under H_0 and whose law is known (at least approximately) under H_0.

p value

- Measure of a realization t on the dataset and decision according to the p value (pivotal value):

$$p = \mathbb{P}_{H_0}(T > t)$$

- **Rational:** If p is small, this means that we have observed a rare event for T if H_0 were true and thus that we have clues against H_0...
- Fisher example: reject H_0 if $p < 0.05$ where 0.05 (5%) is a arbitrary (nice) value that remains used!

Linear Regression and Student Test

- Hypothesis H_0: $\beta^{(k)} = b$ and $\underline{Y} \sim \mathcal{N}(\mathbb{X}\beta, \sigma^2 I_n)$
- Property: Under H_0,

$$\frac{\widehat{\beta^{(k)}} - b}{\widehat{\sigma}\sqrt{[(\mathbb{X}^\top \mathbb{X})^{-1}]_{k,k}}} \sim T(n-p)$$

where $T(n-p)$ is a Student law of degree $n-p$.

Student t-test

- Test statistic: $T = \left| \dfrac{\widehat{\beta^{(k)}} - b}{\widehat{\sigma}\sqrt{[(\mathbb{X}^\top \mathbb{X})^{-1}]_{k,k}}} \right|$
  of known law under H_0.
- Fisher's approach: T is *small* under H_0

- Link with confidence intervals under the Gaussian i.i.d. error assumption.

Global Test

- Assume $\underline{Y} \sim \mathcal{N}(\mathbb{X}\beta, \sigma^2 I_n)$

Global test for the model

- **Global test:** Is the model better than a constant model?
- Assumption to nullify:
$$H_0 : \beta^{(j)} = 0 \text{ pour tout } j \in \{1, \dots, p\},$$
- Alternative hypothesis H_1: there is at least one $j \in \{1, \dots, p\}$ such that $\beta^{(j)} \neq 0$.

- How to do that?

Linear Regression and Fisher Test

- Two nested hypothesis:
  - H_0: $\underline{Y}_{(n)} \sim \mathcal{N}(\mathbb{X}_{(n)}\beta^\star, \sigma^2 I_n)$ with $\beta^\star \in \mathbb{R}^p$
  - H_1: $\underline{Y}_{(n)} \sim \mathcal{N}(\mathbb{Z}_{(n)}\gamma^\star, \sigma^2 I_n)$ with $\gamma^\star \in \mathbb{R}^q$ and $\mathrm{Im}X \subset \mathrm{Im}Z$
- Special case: $\gamma^\star = W\beta^\star$...
- Property:
  - Under H_0 or H_1, $\mathbb{X}\widehat{\beta}$, $\mathbb{Z}\widehat{\gamma} - \mathbb{X}\widehat{\beta}$ and $\underline{Y} - \mathbb{Z}\widehat{\gamma}$ are independent
  - Under H_0, $\|\mathbb{Z}\widehat{\gamma} - \mathbb{X}\widehat{\beta}\|^2 \sim \sigma^2\chi^2(q-p)$
  - Under H_0 or H_1, $\|\underline{Y} - \mathbb{Z}\widehat{\gamma}\|^2 \sim \sigma^2\chi^2(n-q)$

Fisher statistic

- Test statistic: $T = \dfrac{\|\mathbb{Z}\widehat{\gamma} - \mathbb{X}\widehat{\beta}\|^2/(q-p)}{\|\underline{Y} - \mathbb{Z}\widehat{\gamma}\|^2/(n-q)}$

  is of known law under H_0: Fisher law $F(q-p, n-q)$ of degrees $q-p$ and $n-q$.

Ozone Example

- **Gaussian** Linear Model
$$\texttt{O3}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \epsilon_i$$
with ϵ_i i.i.d. Gaussian variables.
- $n = 112$ observations.

R Summary

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035    0.003 **
T12           5.4687     0.4125  13.258   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared: 0.6151,     Adjusted R-squared: 0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

Regression on Qualitative Variables

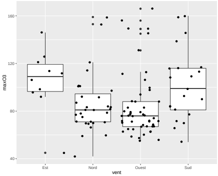
- So far $\underline{X} \in \mathbb{R}^d$...
- How to deal with a qualitative variable Z with k modalities $A_1, A_2, \ldots, A_k$?

Coding

- How to **code** such a qualitative variable with k modalities as a vector in $\mathbb{R}^d$?
- **Dummy Coding:** Code Z by k dummy variables
$$\underline{X} = (\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \ldots, \mathbf{1}_{A_k})$$
- **Disjunctive Coding:** Code Z by $k-1$ dummy variables
$$\underline{X} = (\mathbf{1}_{A_2}, \ldots, \mathbf{1}_{A_k})$$
- Disjunctive Coding is prefered for linear models (colinearity issue with the intercept column)

- **Dummy variable**: quantitative variable equals either 0 or 1

One Factor ANOVA



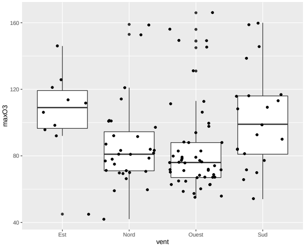
vent variable: A_1: Est, A_2 : Nord, A_3 : Ouest and A_4: Sud

Est	Nord	Ouest	Sud
10	31	50	21

Model

- **One factor ANOVA model:**
  $$\texttt{maxO3}_{ij} = \beta^{(0)} + \beta^{(j)} + \epsilon_{ij} \qquad i = 1, \ldots, n_j \qquad j = A_1, \ldots, A_k$$
- **Equivalent Linear Model** with a dummy code:
  $$\texttt{maxO3} = \beta^{(0)}\mathbf{1}_{\texttt{Est}} + \beta^{(1)}\mathbf{1}_{\texttt{Nord}} + \beta^{(2)}\mathbf{1}_{\texttt{Ouest}} + \beta^{(3)}\mathbf{1}_{\texttt{Sud}} + \epsilon$$
- **Equivalent Linear Model** with a disjunctive code:
  $$\texttt{maxO3} = \beta^{(0)} + \beta^{(1)}\mathbf{1}_{\texttt{Nord}} + \beta^{(2)}\mathbf{1}_{\texttt{Ouest}} + \beta^{(3)}\mathbf{1}_{\texttt{Sud}} + \epsilon$$

One Factor ANOVA



R Result

- Model with Intercept

(Intercept)	ventNord	ventOuest	ventSud
105.60	-19.47	-20.90	-3.08

- Model without Intercept

ventEst	ventNord	ventOuest	ventSud
105.60	86.13	84.70	102.52

- **What can you notice?**

One Factor ANOVA

$$\texttt{maxO3} = \beta^{(0)} + \beta^{(1)}\texttt{ventNord} + \beta^{(2)}\texttt{ventOuest} + \beta^{(3)}\texttt{ventSud} + \epsilon$$

One obtains the following summary:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  105.600      8.639  12.223   <2e-16 ***
ventNord     -19.471      9.935  -1.960   0.0526 .
ventOuest    -20.900      9.464  -2.208   0.0293 *
ventSud       -3.076     10.496  -0.293   0.7700
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 27.32 on 108 degrees of freedom
Multiple R-squared: 0.08602,        Adjusted R-squared: 0.06063
F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

Embedded Model Comparison

MLG1 $\texttt{maxO3}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \beta^{(2)}\texttt{Vx12}_i + \epsilon_i$

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4242     9.3943  -1.535  0.12758
T12           5.0202     0.4140  12.125  < 2e-16 ***
Vx12          2.0742     0.5987   3.465  0.00076 ***
Residual standard error: 16.75 on 109 degrees of freedom
Multiple R-squared: 0.6533,     Adjusted R-squared: 0.6469
F-statistic: 102.7 on 2 and 109 DF,  p-value: < 2.2e-16
```

MLG3 $\texttt{O3}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \beta^{(2)}\texttt{Vx12}_i + \beta^{(3)}\texttt{Ne12}_i + \epsilon_i$

```
lm(formula = maxO3 ~ T12 + Vx12 + Ne12)
Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8958    14.8243   0.263   0.7932
T12           4.5132     0.5203   8.674 4.71e-14 ***
Vx12          1.6290     0.6571   2.479   0.0147 *
Ne12         -1.6189     1.0181  -1.590   0.1147
Residual standard error: 16.63 on 108 degrees of freedom
Multiple R-squared: 0.6612,     Adjusted R-squared: 0.6518
F-statistic: 70.25 on 3 and 108 DF,  p-value: < 2.2e-16
```

Embedded Model

Fisher Test (Anova)

- One test the nullity of a number q of parameters in a model with p parameters.
  - H_0: reduced model with $p - q$ parameters
  - H_1 : full model with p parameters.

MLG1 $\mathtt{O3}_i = \beta^{(0)} + \beta^{(1)}\mathtt{T12}_i + \beta^{(2)}\mathtt{Vx12}_i + \epsilon_i$

MLG3 $\mathtt{O3}_i = \beta^{(0)} + \beta^{(1)}\mathtt{T12}_i + \beta^{(2)}\mathtt{Vx12}_i + \beta^{(3)}\mathtt{Ne12}_i + \epsilon_i$

```
Model 1: O3 ~ T12 + Vx12
Model 2: O3 ~ T12 + Vx12 + Ne12
Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     109 30580
2     108 29881  1    699.61 2.5286 0.1147
```

Fisher Test and Nullity Test

- **Here: Equivalent to the T test of nullity of the coefficient of the variable Ne12 in the model MLG3.**

Arbitrary Model Comparison

MLG1 $\texttt{max03}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \beta^{(2)}\texttt{Vx12}_i + \epsilon_i$

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4242     9.3943  -1.535  0.12758
T12           5.0202     0.4140  12.125  < 2e-16 ***
Vx12          2.0742     0.5987   3.465  0.00076 ***
Residual standard error: 16.75 on 109 degrees of freedom
Multiple R-squared: 0.6533, Adjusted R-squared: 0.6469
F-statistic: 102.7 on 2 and 109 DF, p-value: < 2.2e-16
```

MLG2 $\texttt{max03}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \beta^{(2)}\texttt{Ne12}_i + \epsilon_i$

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7077    15.0884   0.511  0.61050
T12           4.4649     0.5321   8.392 1.92e-13 ***
Ne12         -2.6940     0.9426  -2.858  0.00511 **
Residual standard error: 17.02 on 109 degrees of freedom
Multiple R-squared: 0.6419, Adjusted R-squared: 0.6353
F-statistic: 97.69 on 2 and 109 DF, p-value: < 2.2e-16
```

- **Fisher test can't be used!**
- Need to use AIC, BIC, CV...

Outline

Residuals

- **Residual:** $\widehat{\mathcal{E}} = \underline{Y} - \mathbb{X}\widehat{\beta} = \underline{Y} - \Pi\underline{Y}$ where $\Pi = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ is a projection matrix.
- **Proxy** for $\mathcal{E} = \underline{Y} - \mathbb{X}\beta^\star$

Properties

- If $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_n)$ then $\widehat{\mathcal{E}} \sim \mathcal{N}(0, \sigma^2(I_n - \Pi))$
- $\dfrac{\|\underline{Y} - \Pi Y\|^2}{\sigma^2} = \dfrac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$.
- Natural unbiased estimate of σ^2:
$$\hat{\sigma}^2 = \frac{1}{n-p}\|\underline{Y} - \Pi\underline{Y}\|^2.$$
- $\hat{\beta} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\underline{Y} = \Pi\underline{Y}$ and σ^2 are independent.

Residual

- Residual: $\widehat{\mathcal{E}} = \underline{Y} - \mathbb{X}\widehat{\beta}$ estimates $\mathcal{E} = \underline{Y} - \mathbb{X}\beta^{\dagger}$
- Property: If $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_n)$ then $\widehat{\mathcal{E}} \sim \mathcal{N}(0, \sigma^2(I_n - \Pi_{\mathbb{X}}))$

Residual standardization

- Formulas:
  - $\widetilde{t}_i = \frac{\widehat{\mathcal{E}}_i}{\sigma\sqrt{1-\Pi_{i,i}}}$
  - $t_i = \frac{\widehat{\mathcal{E}}_i}{\widehat{\sigma}\sqrt{1-\Pi_{i,i}}}$
  - $t_i^{\star} = \frac{\widehat{\mathcal{E}}_i}{\widehat{\sigma}_{(i)}\sqrt{1-\Pi_{i,i}}}$ where $\widehat{\sigma}_{(i)}$ is obtained without using the ith observation.
- Magic formula: $\widehat{\sigma^2}_{(i)} = \widehat{\sigma^2}\frac{n-p-1-t_i^2}{n-p}$

Residual Testing

- Prop: If $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_n)$ then $t_i^\star \sim T(n-1-p)$
- Beware: the normalized residuals are correlated.

Normality testing

- Symmetry (Wilcoxon)
- Symmetry and Independence (Wilcoxon-Wolfowitz)
- Independence (Durbin-Watson)
- Normality (Pearson, Shapiro, Lillieford...)

Outliers?

Source: Unknown

Observation *Testing*

Leverage ($\Pi_{i,i}$)

- $\widehat{y}_i = \Pi_{i,i} y_i + \sum_{j \neq u} \Pi_{i,j} y_j$
- Prop: $\sum_{i=1}^{n} \Pi_{i,i} = p$
- Test: $\Pi_{i,i} \geq \kappa p/n \to$ observation to study ($\kappa \simeq 2 - 3$).

Outlier

- Test: $|t_i| > \kappa \to$ badly predicted observation or atypical observation?
- Same test with $t_i^\star$...

Outline

Model Validation

Validation Tools

- Quality of the fit of the obtained model
- Residual plot (simples, standardized or studentized residuals)
- QQ-plot
- Gaussiannity tests (e.g. Shapiro-Wilks, Kolmogorov-Smirnov)

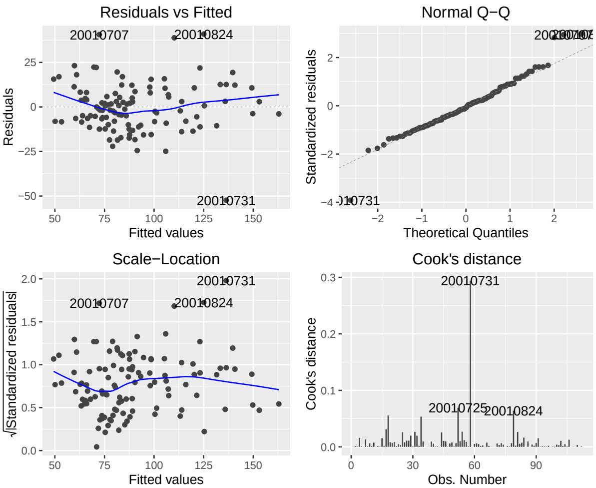
- **Rk:** Gaussiannity not required for asymptotic tests on coefficients...

Ozone

$$\texttt{maxO3}_i = \beta^{(0)} + \beta^{(1)}\texttt{T12}_i + \beta^{(2)}\texttt{Vx9}_i + \beta^{(3)}\texttt{Ne9}_i + \beta^{(4)}\texttt{maxO3v}_i + \epsilon_i$$

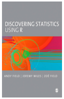
```
---
Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
maxO3v       0.35483    0.05789   6.130 1.50e-08 ***

Residual standard error: 14 on 107 degrees of freedom
Multiple R-squared: 0.7622,     Adjusted R-squared: 0.7533
F-statistic: 85.75 on 4 and 107 DF,  p-value: < 2.2e-16

---
Shapiro-Wilk normality test
W = 0.9659, p-value = 0.005817
```

Residual Analysis

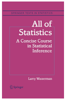


Outline

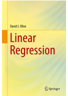
References



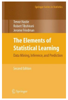
A. Field, J. Miles, and Field Z.
Discovering Statistics Using R.
Sage, 2012



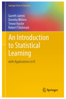
L. Wasserman.
All of Statistics.
Springer, 2004



D. Olive.
Linear Regression.
Springer, 2017



T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning.
Springer Series in Statistics, 2009



G. James, D. Witten, T. Hastie, and Tibshirani R.
An Introduction to Statistical Learning with Applications in R.
Springer, 2014