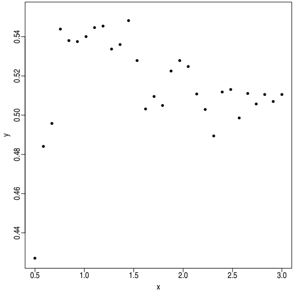
Bias-variance trade-off

A simple prototype problem

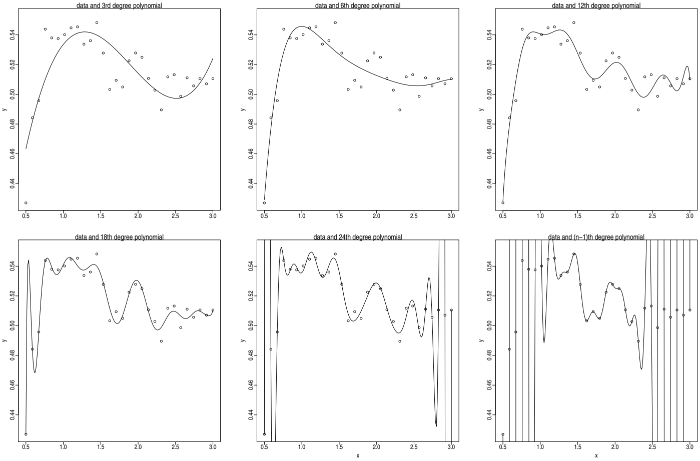


- Yesterday we observed n couples (x_i, y_i), for $i = 1, \ldots, n$, of data $(n = 30)$.
- These data are artificially generated by the law $y = f(x) + \text{error}$ where $f(x)$ is a unspecified smooth and regular function.
- We wish to obtain a rule (model), like $\hat{y} = \hat{f}(x)$, that enables us to predict y once we know x; a rule that allows us to predict y as new observations of x become available, i.e. *tomorrow*.

A simple prototype problem

- A simple possibility is to interpolate data with a polynomial
- Of which degree? $0, 1, 2, \ldots, 29$?
- Let's try to use polynomials of degree p (with $p = 0, 1, \ldots, n-1 = 29$).
  We need to estimate p parameters.

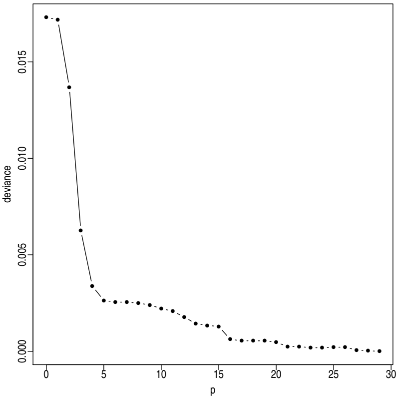
A simple prototype problem

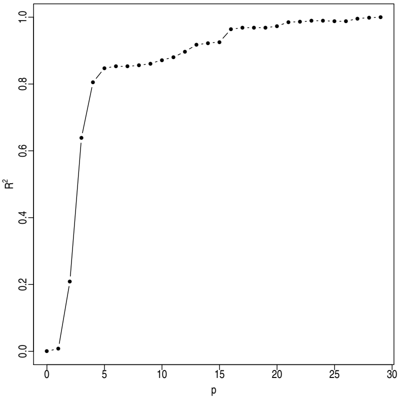


By growing p the fitting of the polynomials is getting better.

A simple prototype problem

We measure the goodness of fit by obtaining, for each p the residual deviance and the coefficient of determination R^2.

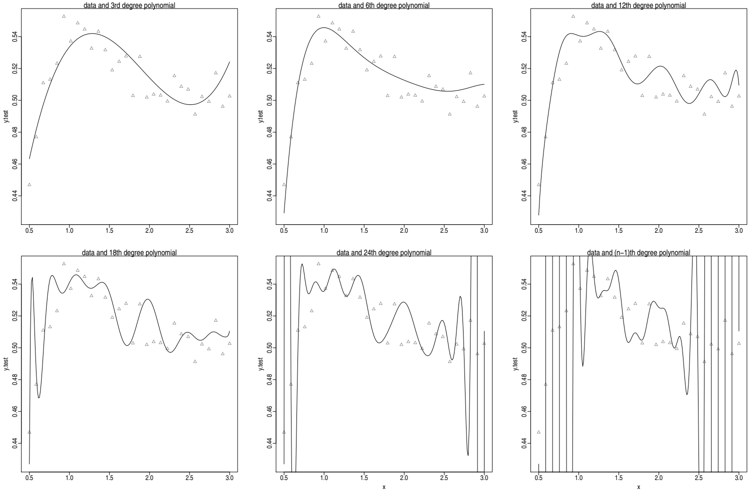




A simple prototype problem

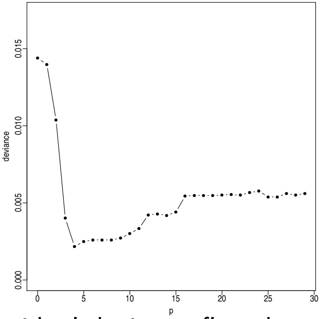
- *Tomorrow* we will receive a new set of n data $\{y_i, i = 1, \ldots, n\}$, generated by the *same* phenomenon of the yesterday data, that is, the same function $f(x)$
- We want to predict these new observations, by assuming, for simplicity, that the new y_i are associated to the same x_i of the yesterday data.
- We compare our predictions (one for each polynomial) with the new data observed tomorrow.

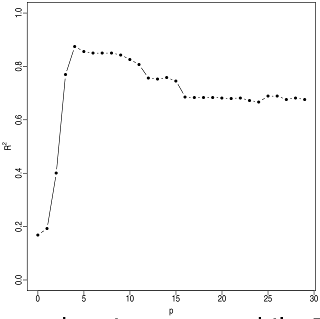
A simple prototype problem



A simple prototype problem

- Goodness of fit for each p: residual deviance and coefficient of determination R^2 on the new data (*tomorrow*).





- Residual deviance first decreases, then increases, while R^2 reaches a maximum value and then decreases.

A simple prototype problem

If we knew $f(x)$...

- We want to estimate $f(x)$ using a generic estimator $\hat{y} = \hat{f}(x)$ (in our example, can be one of the 30 fitted polynomials)
- We start by considering a specific value x' of x, among the n observed.
- If we knew the mechanism used to generate the data precisely, we knew also $f(x')$, and we could calculate some quantities of interest to evaluate the estimator $\hat{y}$.
- For example, an important goodness-of-fit indicator is the mean squared error (with respect to the random variable y)

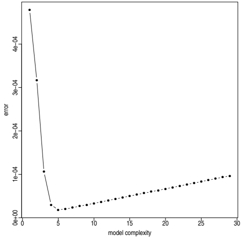
$$\mathbb{E}_y\{[\hat{y} - f(x')]^2\}$$

A simple prototype problem

- Since we are not interested only on the single point x', we consider the sum of the mean squared errors for all the n values of x,

$$\sum_{i=1}^{n} \mathbb{E}_y\left\{\left[\hat{y} - f(x_i)\right]^2\right\}$$

- If we do it for all the possible choices of p, which is an indicator of the model complexity, we may obtain the plot



Even if the true $f(x)$ is not a polynomial, there exists a degree p which is better than the others

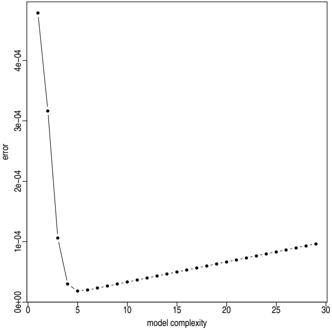
A simple prototype problem

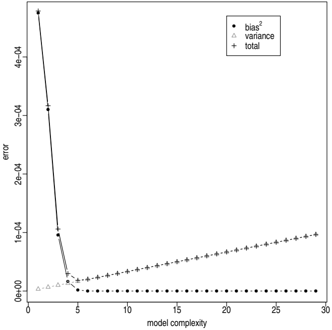
The *mean squared error* may be divided in two components

$$\begin{aligned}\mathbb{E}\big\{[\hat{y} - f(x')]^2\big\} &= \mathbb{E}\big\{[\hat{y} \pm \mathbb{E}\{\hat{y}\} - f(x')]^2\big\} \\ &= \big[\mathbb{E}\{\hat{y}\} - f(x')\big]^2 + \mathrm{var}\{\hat{y}\} \\ &= \text{bias}^2 + \text{variance}\end{aligned}$$

A simple prototype problem

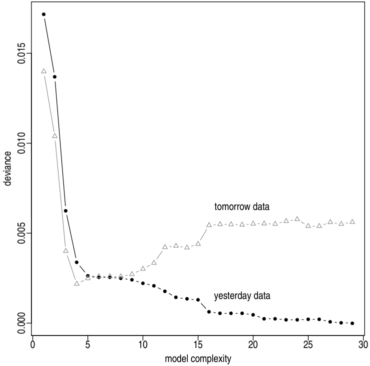
If we knew $f(x)$, we could plot separately bias and variance





A simple prototype problem

- But as we do not know $f(x)$, we only may compute the *residual variance* for the new (*tomorrow*) data:



This plot gives the residual deviance as function of the degree p, by using the model obtained with the *yesterday* data to predict the *tomorrow* data

A simple prototype problem

- When p (the model complexity indicator) increases, the fit improves on the *yesterday* data, but this is not true for the *tomorrow* data.
- goodness-of-fit measure is not a good indicator of the quality of the model
- When p increases too much, we 'overfit' the data and this indicates an excess of *optimism*!

- This happens because the model (the polynomial in the example) follows random fluctuations in yesterday's data not observed in the new sample (and not characteristic of the studied phenomenon), and it mistakes local (random) regularity with a systematic pattern.
- Bias and variance are conflicting entities, and we cannot minimize both simultaneously.
- We must therefore choose a trade-off between bias and variance.

A simple prototype problem

- So that... do not evaluate a model by using the same data used to fit it (the *yesterday* ones).
- If we want a more reliable evaluation, we need to use other data (the *tomorrow* ones)
- How?

A simple prototype problem

- We need tools in order to select models:
    1. Training set and Test set
    2. cross-validation
    3. information criteria

Training set, test set

- If we have n data, and n is *large*, we can divide it in two groups randomly chosen:
  a training set used to fit the various candidate models and
  a test set (sometime called *evaluation set*) used to evaluate the performance of the available models and to choose the most accurate one.
- We compare results obtained with different models on the test set.
- This scheme reduces the sample size used for fitting the model, but this is not a problem when n is huge.
- training and test sets are somehow similar to what was done with *yesterday* and *tomorrow* data.

Information criteria

- The residual variance (or the deviance) is an unreliable indicator of the quality of the model, because it is too optimistic in evaluating the prediction error.
- We can penalize the *deviance* $D = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- ...or a monotonic transform: $-2 \log L = n \log(D/n) +$ (costant)
- with a suitable quantity quantifying the model complexity
- The $\log L$ has an interpretation as log-likelihood.
- Criteria that follow this logic can be traced back to objective functions such as

$$IC(p) = -2 \log L + \text{penalty}(p)$$

- The choice of the specific penalty function identifies a particular criterion.

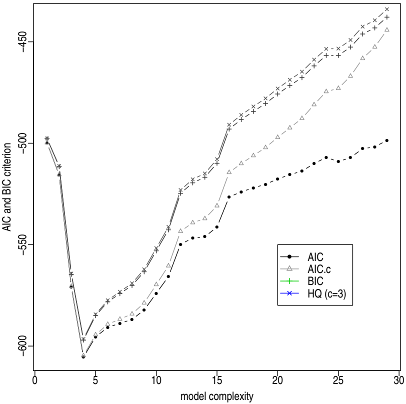
Information criteria

- Some possibile penalty are in the following table

criterion	author	penalty(p)
AIC	Akaike	$2p$
AIC$_c$	Sugiura, Hurvich-Tsay	$2p + \frac{2p\,(p+1)}{n-(p+1)}$
BIC/SIC	Akaike, Schwarz	$p\,\log n$
HQ	Hannan-Quinn	$c\,p\,\log\log n, \quad (c > 2)$

- These criteria are applied also to *not nested* models.

Information criteria – example



We choose p minimising $IC(p)$ using some criteria in the previous table; in our example all choices for penalty suggest $p = 4$.

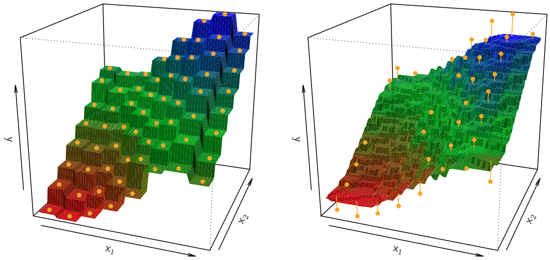
Non parametric regression

KNN: regression

Given a value k and a prediction point x_0, the KNN regression identifies in the training set the k nearest observations, N_0

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_0} y_i$$

KNN: regression



KNN with $p = 2$, $k = 1$ (left) and $k = 9$ (right). With small k high variance and low bias, since prediction is performed on a single observation.

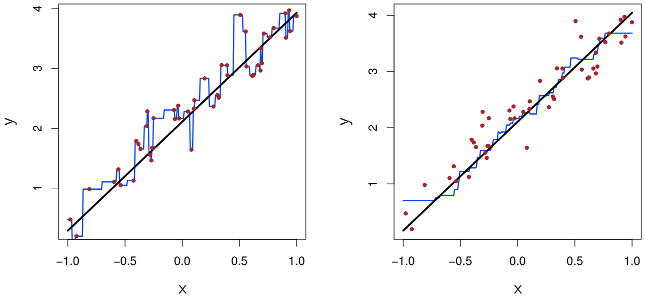
KNN: regression

The optimal value of k is related to the trade-off bias-variance.

- small $k \to$ high variance and low bias
- big $k \to$ low variance (smoother prediction) and high bias - local structure of $f(X)$ may not be captured-

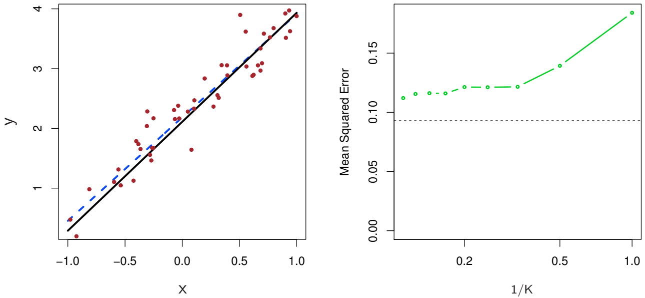
KNN: regression

Parametric approach may be preferred to the non parametric if the parametric form is close to the 'real' f.



Comparison between KNN with $k = 1$ (left) e $k = 9$ (right).
Since the true relationship is linear the non parametric approach will have a worse performance.

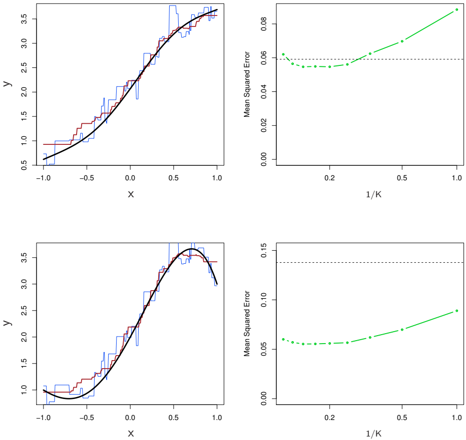
KNN: regression



Regression line (dashed line) Test MSE for regression line (dashed) and KNN (green) as function of $1/k$.

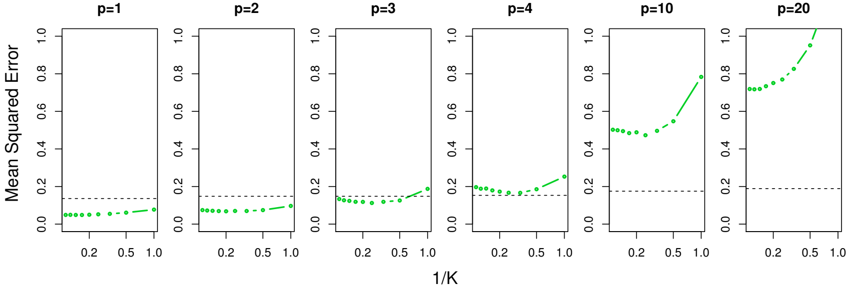
Best results for KNN are with high value of k.

KNN: regression



Nonlinear relationships and KNN with $k = 1$ (blue) and $k = 9$ (red). Conditional to nonlinearity of f the KNN performance changes with respect to LM. As the nonlinearity becomes more evident, the performance of KNN with high k will increase.

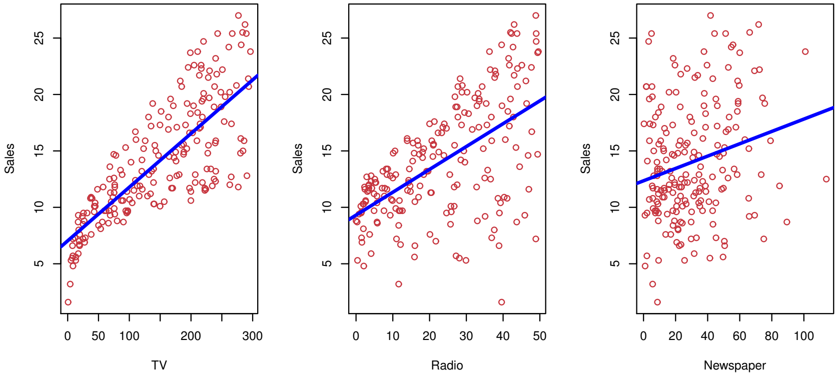
KNN: regression



By increasing the number of variables p, the KNN performance will rapidly decrease in terms of MSE test.
It is more difficult to find the 'nearest neighbours' ... curse of dimensionality

Example

Sales of a product in thousands of units as function of budget in `tv`, `radio`, `newspapers` for 200 different markets.



Regression line for tv, radio, newspapers.

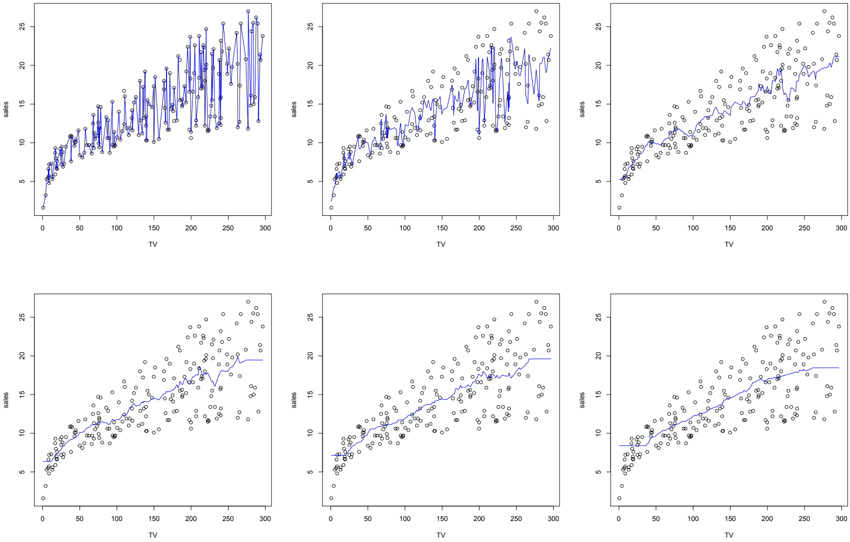
Example

We wish to study the performance of KNN for some values of k with the only variable `tv`

Example

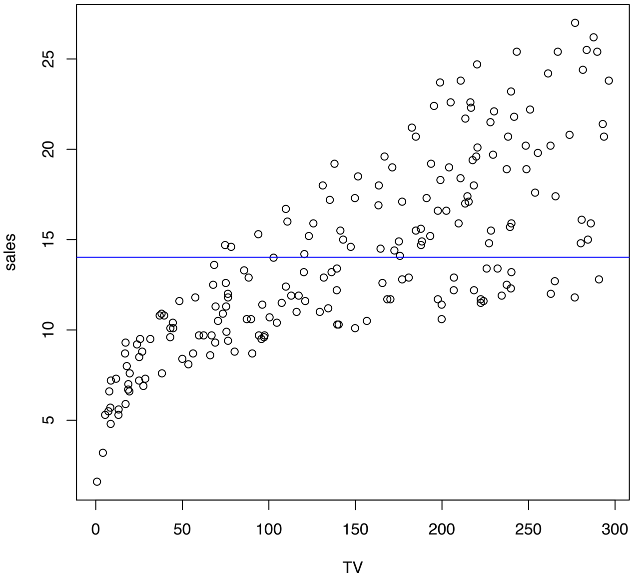
all data

$$k = 1, 2, 10, 20, 30, 50$$

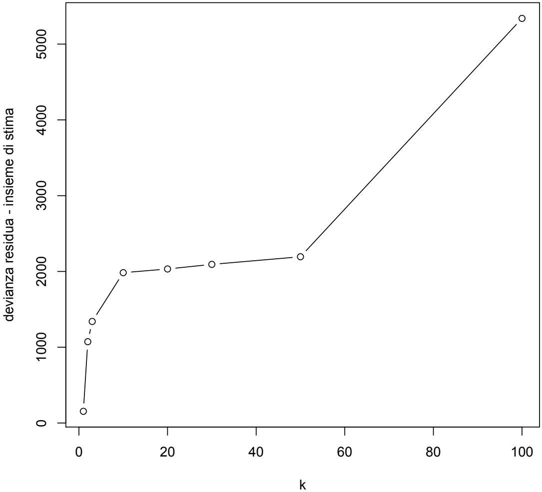


Example

$k = 200$



Example

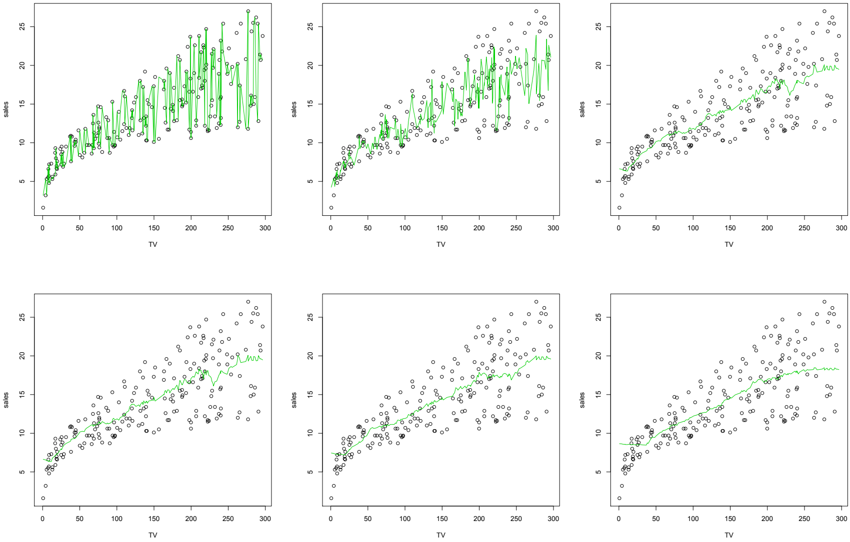


KNN performance decreases as k increases.

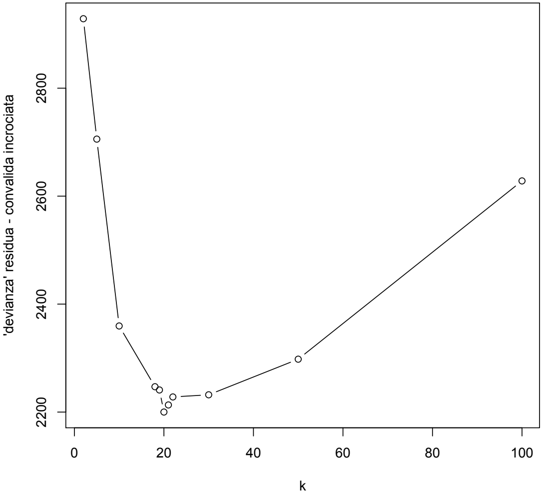
Example

Cross validation leave-one-out

$$k = 1, 2, 10, 20, 30, 50$$



Example

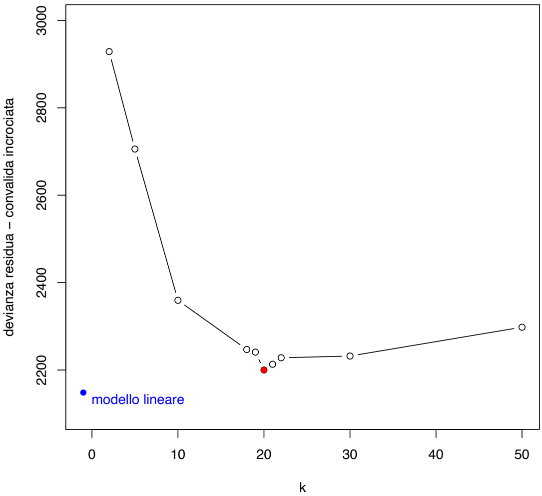


A minimum has been reached . . .

trade-off between variance and bias

Example

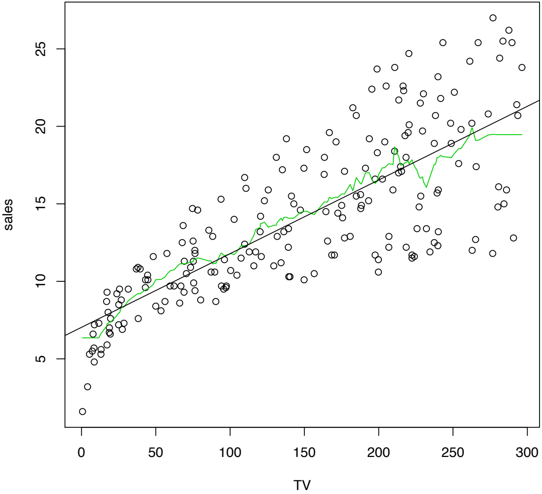
Performance of linear model and KNN with variable `tv`



In the case of `tv` the linear model performs better than the KNN for each value of k.

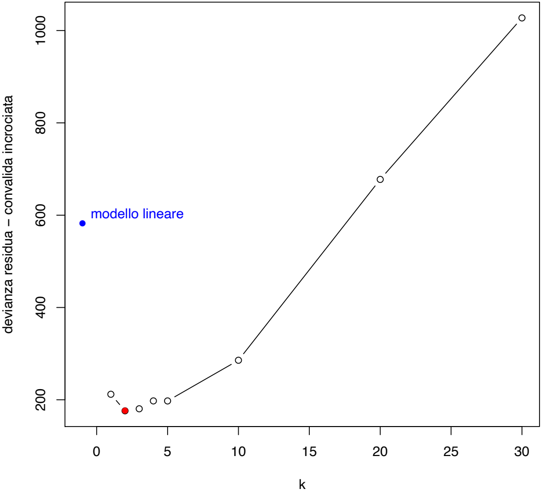
Example

Linear model and KNN-20 with variable `tv`



Example

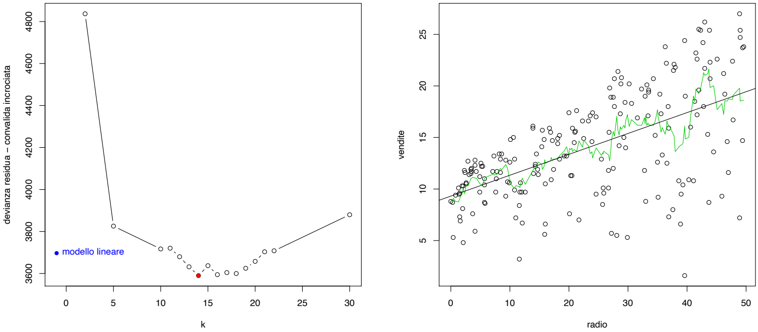
Performance of linear model and KNN with variable `tv` and `radio`



Adding the variable `radio` highly increases the performance of KNN. The minimum is reached for $k = 2$

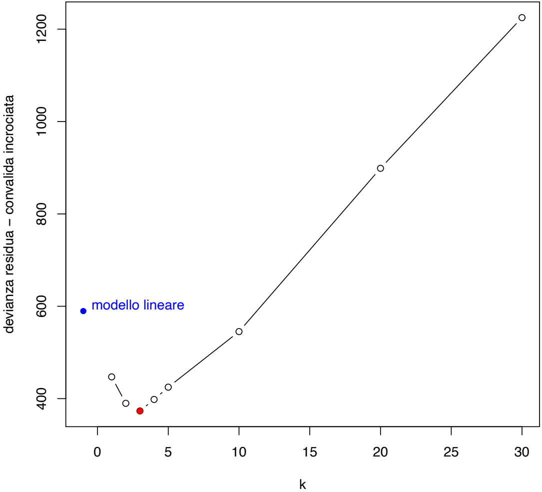
Example

Performance of linear model and KNN with variable `radio`



Example

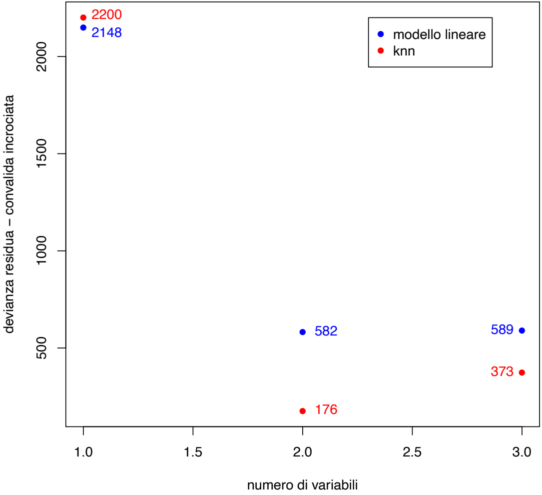
Let us add the variable `newspapers`



Adding the variable `newspapers` does not increase the performance of the model.

The KNN is better than the linear model in any case.

Example



modello lineare
knn
2200
2148
582
589
373
176
devianza residua – convalida incrociata
numero di variabili
500
1000
1500
2000
1.0
1.5
2.0
2.5
3.0