Cross-validation

Miguel Angel Luque Fernandez
Faculty of Epidemiology and Population Health
Department of Non-communicable Disease.

August 25, 2015

Contents

1. Cross-validation
2. Cross-validation justification
3. Cross-validation methods
4. Examples: Model selection
5. References

Cross-validation

Definition

- Cross-validation is a **model validation technique** for assessing how the results of a statistical analysis will generalize to an independent data set.

- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice (note: performance = **model assessment**).

Cross-validation

Definition

- Cross-validation is a **model validation technique** for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice (note: performance = **model assessment**).

Cross-validation

Definition

- Cross-validation is a **model validation technique** for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice (note: performance = **model assessment**).

Cross-validation

Applications

- However, cross-validation can be used to compare the **performance** of different modeling specifications (i.e. models with and without interactions, inclusion of exclusion of polynomial terms, number of knots with restricted cubic splines, etc).

- Furthermore, cross-validation can be used in **variable selection** and select the suitable level of flexibility in the model (note: flexibility = **model selection**).

Cross-validation

Applications

- However, cross-validation can be used to compare the **performance** of different modeling specifications (i.e. models with and without interactions, inclusion of exclusion of polynomial terms, number of knots with restricted cubic splines, etc).
- Furthermore, cross-validation can be used in **variable selection** and select the suitable level of flexibility in the model (note: flexibility = **model selection**).

Cross-validation

Applications

- However, cross-validation can be used to compare the **performance** of different modeling specifications (i.e. models with and without interactions, inclusion of exclusion of polynomial terms, number of knots with restricted cubic splines, etc).
- Furthermore, cross-validation can be used in **variable selection** and select the suitable level of flexibility in the model (note: flexibility = **model selection**).

Cross-validation

Applications

- MODEL ASSESSMENT: To **compare** the performance of different modeling specifications.
- MODEL SELECTION: To **select** the suitable level of flexibility in the model.

Cross-validation

Applications

- MODEL ASSESSMENT: To **compare** the performance of different modeling specifications.
- MODEL SELECTION: To **select** the suitable level of flexibility in the model.

Cross-validation

Applications

- MODEL ASSESSMENT: To **compare** the performance of different modeling specifications.
- MODEL SELECTION: To **select** the suitable level of flexibility in the model.

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

$$Y = f(x) + \epsilon$$

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

$$Y = f(x) + \epsilon$$

MSE

Regression Model

$$f(x) = f(x_1 + x_2 + x_3)$$

$$Y = \beta x_1 + \beta x_2 + \beta x_3 + \epsilon$$

$$Y = f(x) + \epsilon$$

MSE

Expectation

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

MSE

$$E[(Y - \hat{f}(X))^2|X = x]$$

MSE

Expectation

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

MSE

$$E[(Y - \hat{f}(X))^2|X = x]$$

MSE

Expectation

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

MSE

$$E[(Y - \hat{f}(X))^2|X = x]$$

Bias-Variance Trade-off

Error descomposition

$$MSE = E[(Y - \hat{f}(X))^2 | X = x] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

Trade-off

As flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases.

BIAS-VARIANCE-TRADE-OFF

Bias-variance trade-off

Chossing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

And thus, this amounts to a bias-variance trade-off.

BIAS-VARIANCE-TRADE-OFF

Bias-variance trade-off

Chossing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

And thus, this amounts to a bias-variance trade-off.

Rule

- More flexibility increases variance but decreases bias.
- Less flexibility decreases variance but increases error.

BIAS-VARIANCE-TRADE-OFF

Bias-variance trade-off

Chossing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

And thus, this amounts to a bias-variance trade-off.

Rule

- More flexibility increases variance but decreases bias.
- Less flexibility decreases variance but increases error.

BIAS-VARIANCE-TRADE-OFF

Bias-variance trade-off

Chossing the model flexibility based on average test error

Average Test Error

$$E[(Y - \hat{f}(X))^2 | X = x]$$

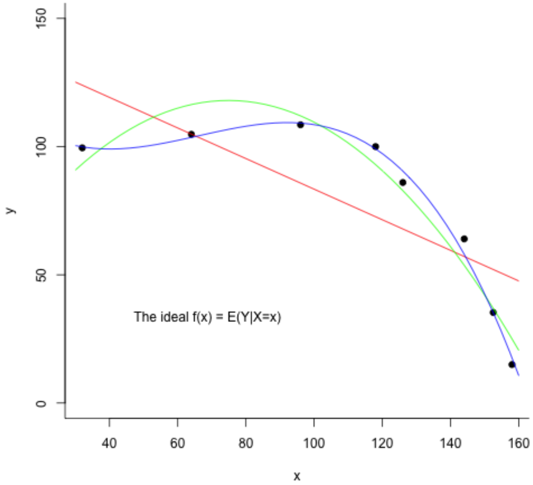
And thus, this amounts to a bias-variance trade-off.

Rule

- More flexibility increases variance but decreases bias.
- Less flexibility decreases variance but increases error.

Bias-Variance trade-off

Regression Function



Overparameterization

George E.P.Box,(1919-2013)

All models are wrong but some are useful

Quote, 1976

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration (...). Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

Overparameterization

George E.P.Box,(1919-2013)

All models are wrong but some are useful

Quote, 1976

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration (...). Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.

- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.

- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom

- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.
- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom
- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.
- There is some disagreement over the use of AIC and BIC with non-nested models.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.
- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom
- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.
- There is some disagreement over the use of AIC and BIC with non-nested models.

Justification

AIC and BIC

- AIC and BIC are both maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior.

- AIC = -2*ln(likelihood) + 2*k, k = model degrees of freedom

- BIC = -2*ln(likelihood) + ln(N)*k, k = model degrees of freedom and N = number of observations.

- There is some disagreement over the use of AIC and BIC with non-nested models.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.

- Cross-validation makes fewer assumptions about the true underlying model.

- Cross-validation can be used in a wider range of model selections tasks, even in cases where it is hard to pinpoint the number of predictors in the model.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.
- Cross-validation can be used in a wider range of model selections tasks, even in cases where it is hard to pinpoint the number of predictors in the model.

Justification

Fewer assumptions

- Cross-validation compared with AIC, BIC and adjusted R^2 provides a direct estimate of the **ERROR**.
- Cross-validation makes fewer assumptions about the true underlying model.
- Cross-validation can be used in a wider range of model selections tasks, even in cases where it is hard to pinpoint the number of predictors in the model.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.
- Bootstraping.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.
- Bootstraping.

Cross-validation strategies

Cross-validation options

- Leave-one-out cross-validation (LOOCV).
- k-fold cross validation.
- Bootstraping.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.
- The idea is to randmoly divide the data into k equal-sized parts. We leave out part k, fit the model to the other k-1 parts (combined), and then obtain predictions for the left-out kth part.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.
- The idea is to randmoly divide the data into k equal-sized parts. We leave out part k, fit the model to the other k-1 parts (combined), and then obtain predictions for the left-out kth part.

K-fold Cross-validation

K-fold

- Technique widely used for estimating the test error.
- Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model.
- The idea is to randmoly divide the data into k equal-sized parts. We leave out part k, fit the model to the other k-1 parts (combined), and then obtain predictions for the left-out kth part.

K-fold Cross-validation

K-fold

$$\mathrm{CV} = \sum_{k-1}^{k} \frac{n_k}{n} MSE_k$$

$$\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - (\hat{y}_i))/n_k$$

Seeting K = n yields n-fold or leave-one-out cross-validation (LOOCV)

K-fold Cross-validation

K-fold

$$\mathrm{CV} = \sum_{k-1}^{k} \frac{n_k}{n} MSE_k$$

$$\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - (\hat{y}_i))/n_k$$

Seeting K = n yields n-fold or leave-one-out cross-validation (LOOCV)

K-fold Cross-validation

K-fold

$$\text{CV} = \sum_{k-1}^{k} \frac{n_k}{n} MSE_k$$

$$\text{MSE}_k = \sum_{i \in C_k} (y_i - (\hat{y}_i))/n_k$$

Seeting K = n yields n-fold or leave-one-out cross-validation (LOOCV)

Particular case

Linear regression and polynomials

$$LOOCV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

h_i is the leverage coming from the geometrical interpretation of the residuals in the hat matrix. Where $h_{i,j} = \frac{cov(\hat{y}_i, y_j)}{var(y_j)}$

Correlation when K = n

Which is equal to the ordinary MSE, except the ith residual is divided by 1-h_i. However, with LOOCV the estimates from each fold are highly correlated and hence their average can have high variance. A better choice is a K-fold Cross-Validation with K = 5 or 10.

Particular case

Linear regression and polynomials

$$LOOCV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

h_i is the leverage coming from the geometrical interpretation of the residuals in the hat matrix. Where $h_{i,j} = \frac{cov(\hat{y}_i, y_j)}{var(y_j)}$

Correlation when K = n

Which is equal to the ordinary MSE, except the ith residual is divided by 1-h_i. However, with LOOCV the estimates from each fold are highly correlated and hence their average can have high variance. A better choice is a K-fold Cross-Validation with K = 5 or 10.

Particular case

Linear regression and polynomials

$$LOOCV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

h_i is the leverage coming from the geometrical interpretation of the residuals in the hat matrix. Where $h_{i,j} = \frac{cov(\hat{y}_i, y_j)}{var(y_j)}$

Correlation when K = n

Which is equal to the ordinary MSE, except the ith residual is divided by 1-h_i. However, with LOOCV the estimates from each fold are highly correlated and hence their average can have high variance. A better choice is a K-fold Cross-Validation with K = 5 or 10.

Motivation

Motivation

- To investigate the functional form of a continuous variable.
- In linear regression, and generalized linear models, partial residuals are used to assess whether continuous covariates are in the model in their correct form, or whether a transformation is needed.

Motivation

Motivation

- To investigate the functional form of a continuous variable.
- In linear regression, and generalized linear models, partial residuals are used to assess whether continuous covariates are in the model in their correct form, or whether a transformation is needed.
- In survival models we have to assess is whether the log-hazard is linear in x. We use a scatter plot of martingale residuals againts the variable in assesment. If the scatterplot is linear, this indicates that the log-hazard is linear in x otherwise a transformation of the variable is suitable.

Motivation

Motivation

- To investigate the functional form of a continuous variable.
- In linear regression, and generalized linear models, partial residuals are used to assess whether continuous covariates are in the model in their correct form, or whether a transformation is needed.
- In survival models we have to assess is whether the log-hazard is linear in x. We use a scatter plot of martingale residuals againts the variable in assesment. If the scatterplot is linear, this indicates that the log-hazard is linear in x otherwise a transformation of the variable is suitable.

Hosmer, D. W., Lemeshow S., and May, S. (2008). Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, 2nd edition, 392 pages.

Motivation

Motivation

- To investigate the functional form of a continuous variable.
- In linear regression, and generalized linear models, partial residuals are used to assess whether continuous covariates are in the model in their correct form, or whether a transformation is needed.
- In survival models we have to assess is whether the log-hazard is linear in x. We use a scatter plot of martingale residuals againts the variable in assesment. If the scatterplot is linear, this indicates that the log-hazard is linear in x otherwise a transformation of the variable is suitable.

Hosmer, D. W., Lemeshow S., and May, S. (2008). Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, 2nd edition, 392 pages.

Motivation

Martingale residuals

Given the ln h(t) = ln h_0(t) + xβ;
The martingale residuals are defined as:

$$\hat{M}_i = c_i - \hat{H}\left(t_i, x_i, \hat{\beta}\right)$$

Hosmer, D. W., Lemeshow S., and May, S. (2008). Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, 2nd edition, 392 pages.

The Data

```
[fontsize=\small]
variable name   type    format
----------------
sex             str1    %9s     labels: F(1), M(2)
age             byte    %8.0g
tt              float   %9.0g
site            str5    %9s     labels:Ear, Face, Neck, Scalp
censor          byte    %8.0g
survival        float   %9.0g
----------------
```

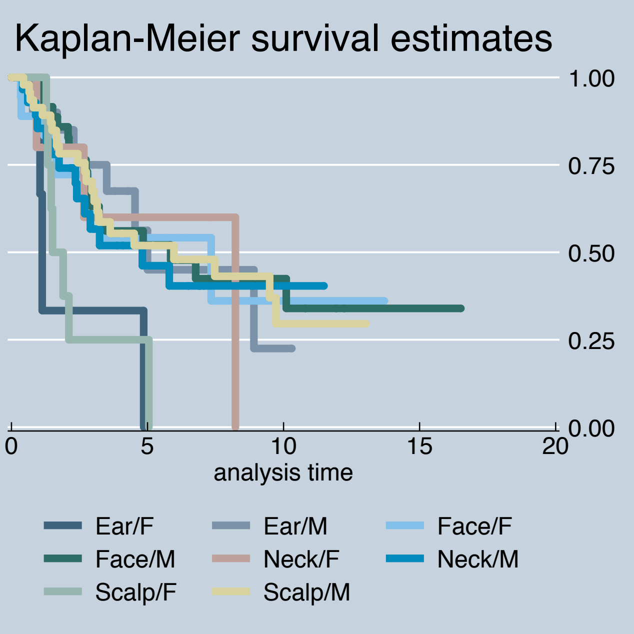
The Data

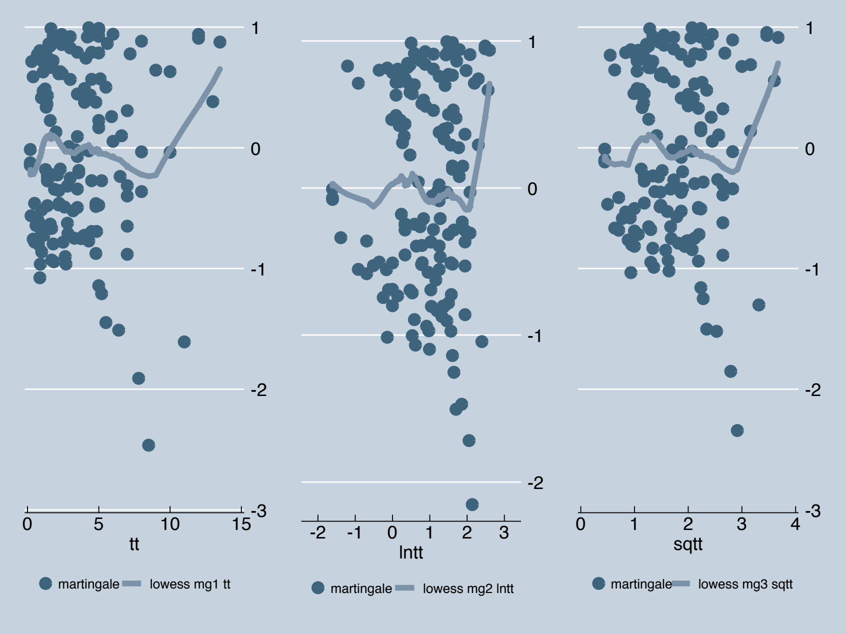
```
----------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+--------------------------------
           tt |    1.15688    .0477728     3.53   0.000     1.066936    1.254406
      _Isex_2 |   .0707031    .0600307    -3.12   0.002     .0133882    .3733826
     _Isite_2 |   .1051597    .0869398    -2.72   0.006      .020803    .5315848
     _Isite_3 |   .1203827    .1145662    -2.22   0.026     .0186419    .7773897
     _Isite_4 |   .4360958    .3629804    -1.00   0.319     .0853281    2.228804
_IsexXsit_2_2 |   13.54933    12.82601     2.75   0.006      2.11913      86.632
_IsexXsit_2_3 |    15.7232    16.72104     2.59   0.010     1.955778    126.4044
_IsexXsit_2_4 |   3.322081    3.126073     1.28   0.202     .5253284    21.00823
----------------------------------------
```

Male vs Female

```
lincom _Isite_2 + _IsexXsit_2_2
 ( 1)  _Isite_2 + _IsexXsit_2_2 = 0
---------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+--------------------------------
         (1) |   1.424844    .647053     0.78   0.436     .5850838    3.469898
---------------------------------------
Males with face melanomas do not have significantly different death rates
to females with face melanomas, of the same thickness.

. lincom _Isite_3 + _IsexXsit_2_3, hr
 ( 1)  _Isite_3 + _IsexXsit_2_3 = 0
---------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+--------------------------------
         (1) |   1.892801   .8879783     1.36   0.174     .7547045     4.74715
---------------------------------------
Males with scalp melanomas do not have significantly different death rates
to females with scalp melanomas, of the same thickness.
```


Kaplan-Meier survival estimates
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
analysis time
Ear/F
Ear/M
Face/F
Face/M
Neck/F
Neck/M
Scalp/F
Scalp/M


1
0
-1
-2
-3
0
5
10
15
tt
martingale
lowess mg1 tt
1
0
-1
-2
-2
-1
0
1
2
3
lntt
martingale
lowess mg2 lntt
1
0
-1
-2
-3
0
1
2
3
4
sqtt
martingale
lowess mg3 sqtt

Model assessment

```
---------------------------
    Variable |     a             b             c
-------+--------------------
          tt |  1.1568797
      _Isex_2 |  .07070305     .07496949     .07373205
     _Isite_2 |  .10515972     .10639845     .10642456
     _Isite_3 |  .12038267     .11606331     .11908885
     _Isite_4 |  .43609584     .42994272     .43798124
_IsexXsit_~2 |  13.549333     13.858472     13.546268
_IsexXsit_~3 |  15.723202     16.500689     15.947097
_IsexXsit_~4 |  3.3220807     3.2716299     3.2410989
        lntt |                1.5596707
        sqtt |                              1.7353908
-------+--------------------
         AIC |  699.66119     700.98922     700.37545
         BIC |  724.00859     725.33662     724.72285
---------------------------
```

K-fold cross-validation Stata; K=10

```
crossfold: xi: streg tt i.sex*i.site, dist(exp) mae k(10)
matrix list r(est)
matrix a = r(est)
matrix list a
svmat double a, name(modela)
mean modela1
gen modela = modela1

crossfold: xi: streg lntt i.sex*i.site, dist(exp) mae k(10)
matrix list r(est)
matrix b = r(est)
matrix list b
svmat double b, name(modelb)
mean modelb1
gen modelb = modelb1

crossfold: xi: streg sqtt i.sex*i.site, dist(exp) mae k(10)
matrix list r(est)
matrix c = r(est)
matrix list c
svmat double b, name(modelc)
mean modelc1
gen modelc = modelc1

mean modela modelb modelc
```

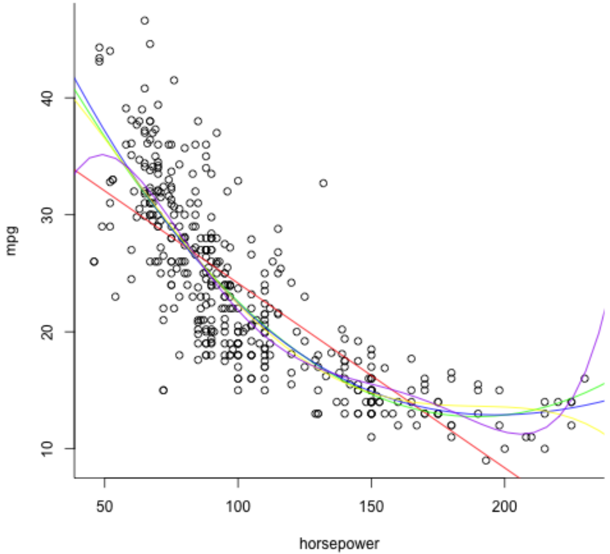
K-fold cross-validation k=10

```
-------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------+------------------------
      modela |   3.411219   .2175402      2.919109    3.903329
      modelb |   3.497022   .2103749      3.005649    3.927495
      modelc |   3.497522   .2121749      3.017549    3.977495
-------------------------------
```

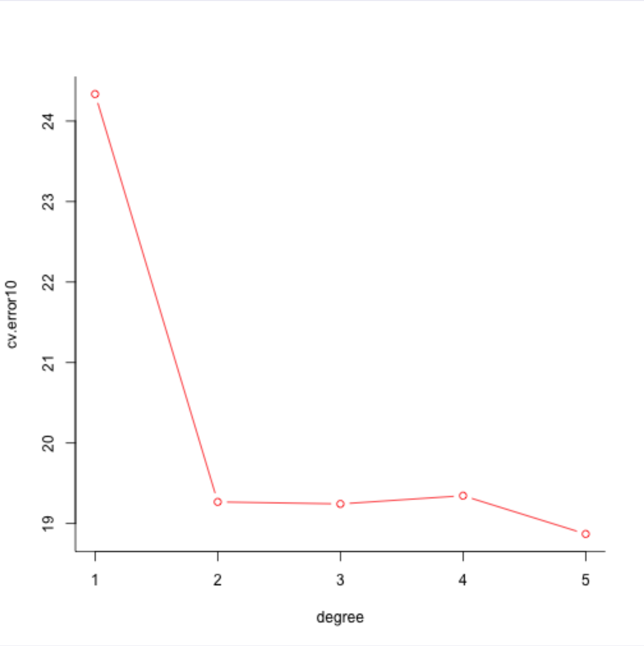
Models

```
#fit first degree polynomial equation:
fit  <- lm(mpg~horsepower,data=Auto)
#Polynomial degrees
fit2 <- lm(mpg~poly(horsepower,2,raw=TRUE), data=Auto)
fit3 <- lm(mpg~poly(horsepower,3,raw=TRUE), data=Auto)
fit4 <- lm(mpg~poly(horsepower,4,raw=TRUE), data=Auto)
fit5 <- lm(mpg~poly(horsepower,5,raw=TRUE), data=Auto)
#generate range of 50 numbers starting from 30 and ending at 160
plot(mpg~horsepower,data=Auto, bty="l")
xx <- seq(10,250, length=50)
lines(xx, predict(fit, data.frame(horsepower=xx)), col="red")
lines(xx, predict(fit2, data.frame(horsepower=xx)), col="green")
lines(xx, predict(fit3, data.frame(horsepower=xx)), col="blue")
lines(xx, predict(fit4, data.frame(horsepower=xx)), col="black")
lines(xx, predict(fit5, data.frame(horsepower=xx)), col="purple")
```

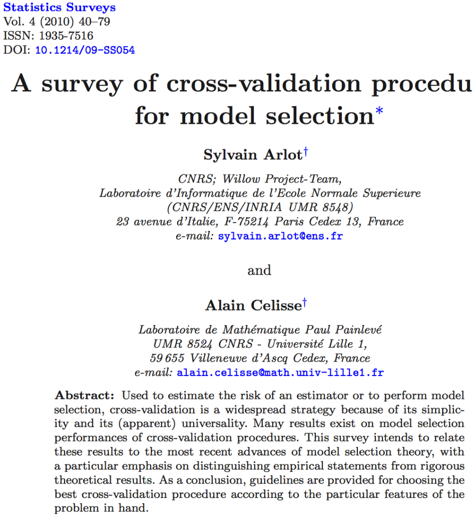
Gareth James: R Dataset Auto



10-fold CV



References



Statistics Surveys
Vol. 4 (2010) 40–79
ISSN: 1935-7516
DOI: 10.1214/09-SS054

A survey of cross-validation procedures for model selection*

Sylvain Arlot†

CNRS; Willow Project-Team,
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23 avenue d'Italie, F-75214 Paris Cedex 13, France
e-mail: sylvain.arlot@ens.fr

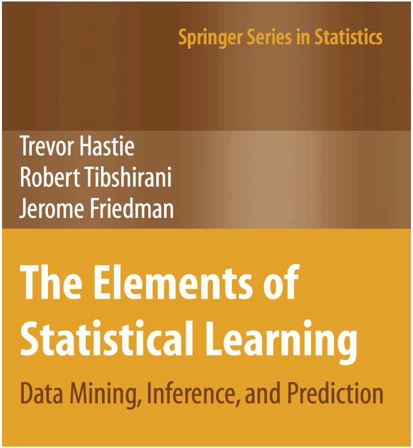
and

Alain Celisse†

Laboratoire de Mathématique Paul Painlevé
UMR 8524 CNRS - Université Lille 1,
59 655 Villeneuve d'Ascq Cedex, France
e-mail: alain.celisse@math.univ-lille1.fr

Abstract: Used to estimate the risk of an estimator or to perform model selection, cross-validation is a widespread strategy because of its simplicity and its (apparent) universality. Many results exist on model selection performances of cross-validation procedures. This survey intends to relate these results to the most recent advances of model selection theory, with a particular emphasis on distinguishing empirical statements from rigorous theoretical results. As a conclusion, guidelines are provided for choosing the best cross-validation procedure according to the particular features of the problem in hand.

References


Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Thank you

THANK YOU FOR YOUR TIME