

Linear models

Data Mining course

Master in Information Technologies

Enginyeria Informàtica

Tomàs Aluja

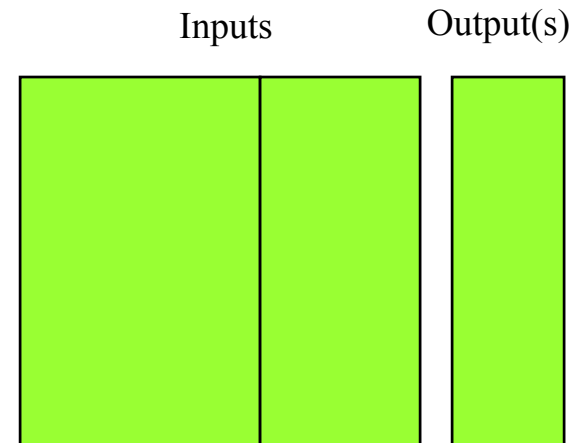
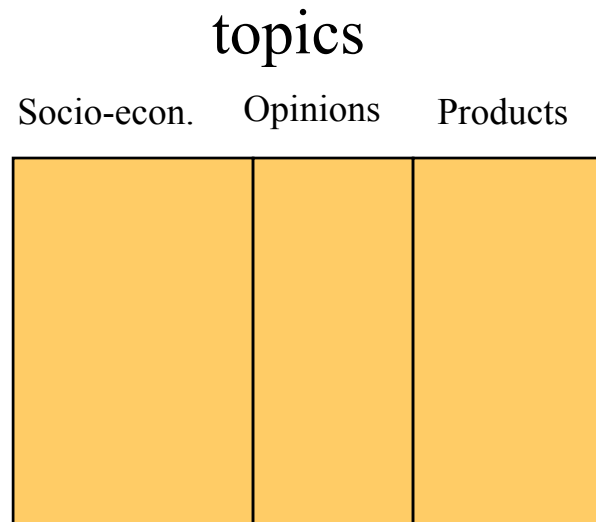


UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Two types of datasets to analyze

Data in *Data Mining*:

massive, secondary, not random, with errors and missing values



Data to explore

Data to modelize

A statistical model

$$\text{Data} = \text{Fit} + \text{Error}$$

$$y = r(x_1, \dots, x_p) + \varepsilon = \hat{y} + \varepsilon$$

Data: Response variable (output)

- continuous $\rightarrow$ **Regression model**
- categorical $\rightarrow$ **Classification model**

$$\hat{y} = E[y | x_1, \dots, x_p]$$

$$\hat{y} = P[C_k | x_1, \dots, x_p]$$

Fit: function of the explanatory variables (inputs)
they can be either continuous or categorical

Error: express the random fluctuation of the phenomena, it can be interpreted as the effect of all other (unknown) causes intervening in the phenomena

Some problems

Household valuation
Economic capacity prediction
Time to failure prediction
....

Type of the response variable:

- If *categorical* (supervised) **classification** $P[C_k|x_1, \dots, x_p]$
- If *continuous* **REGRESSION** $E[y|x_1, \dots, x_p]$

Boston housing.data

2. Sources:

Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Date: July 7, 1993

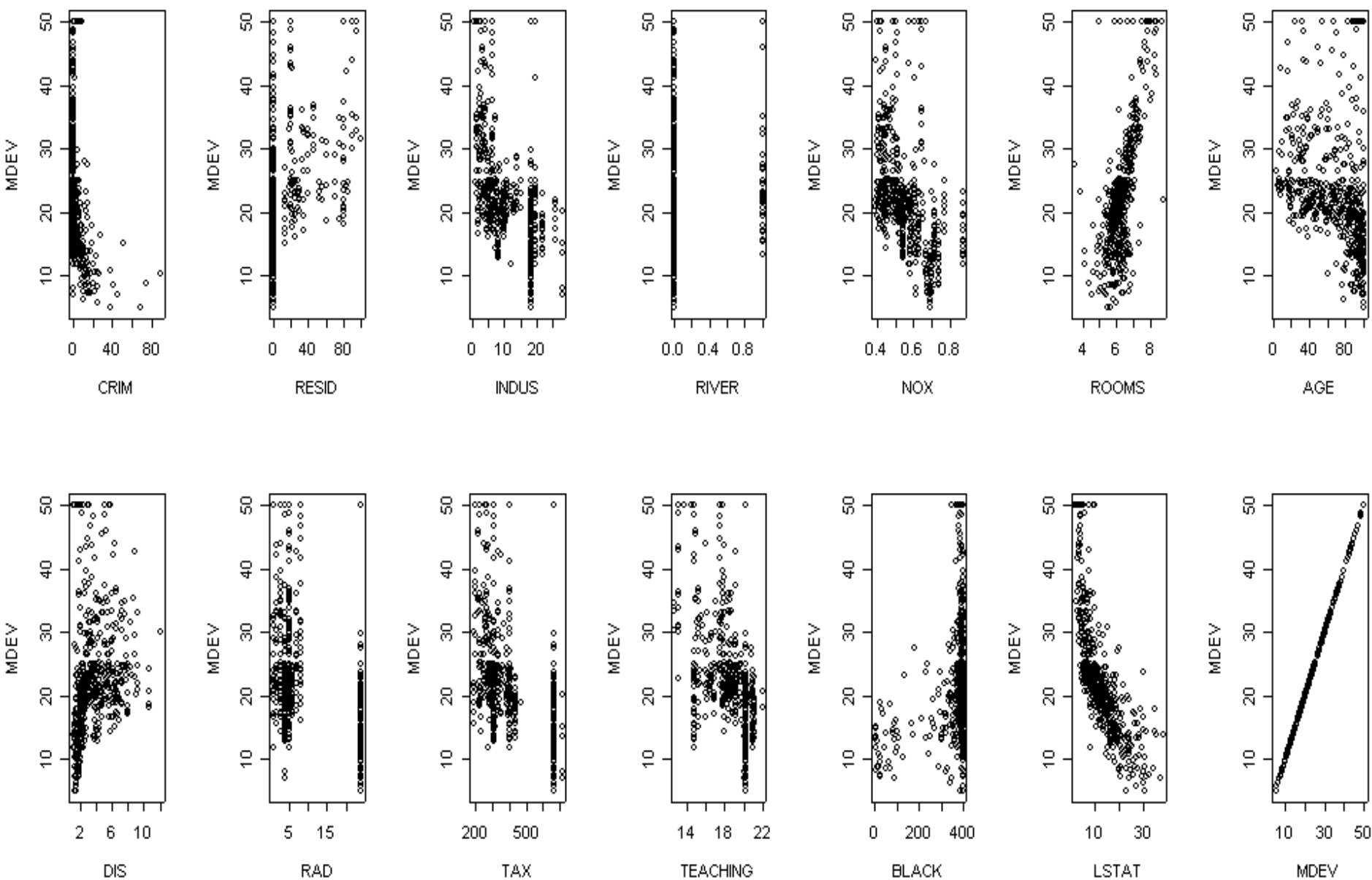
4. Relevant Information: Concerns housing values in suburbs of Boston.

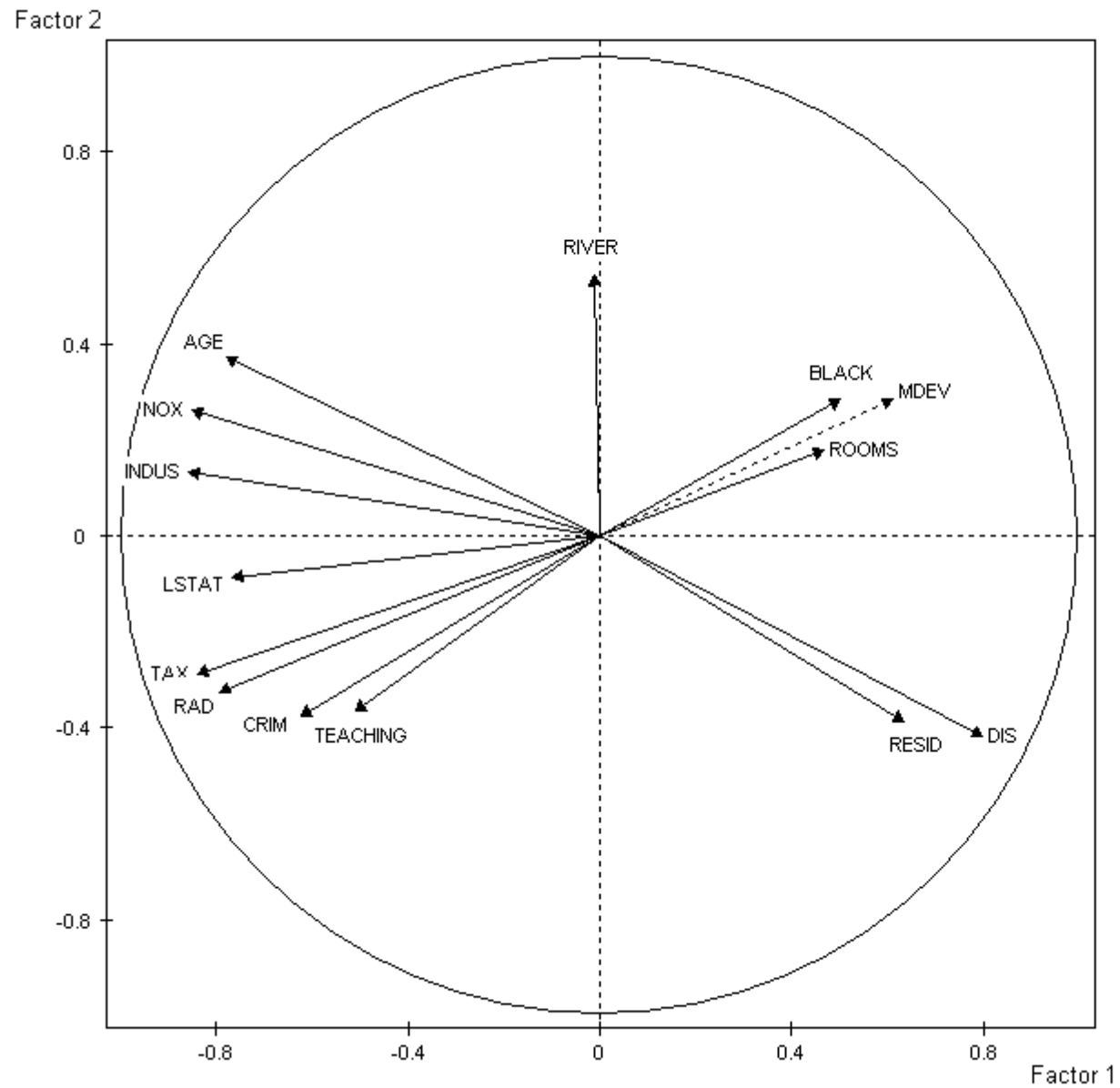
5. Number of Instances: 506

6. Number of Attributes: 13 continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute.

7. Attribute Information:

- | | |
|--------------|---|
| 1. CRIM | per capita crime rate by town |
| 2. RESID | proportion of residential land zoned for lots over 25,000 sq.ft. |
| 3. INDUS | proportion of non-retail business acres per town |
| 4. RIVER | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| 5. NOX | nitric oxides concentration (parts per 10 million) |
| 6. ROOMS | average number of rooms per dwelling |
| 7. AGE | proportion of owner-occupied units built prior to 1940 |
| 8. DIS | weighted distances to five Boston employment centres |
| 9. RAD | index of accessibility to radial highways |
| 10. TAX | full-value property-tax rate per \$10,000 |
| 11. TEACHING | pupil-teacher ratio by town |
| 12. BLACK | $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town |
| 13. LSTAT | % lower status of the population |
| 14. MEDV | Median value of owner-occupied homes in \$1000's |





Statistical modeling of a continuous variable

“to explain” the behavior of the response knowing the inputs

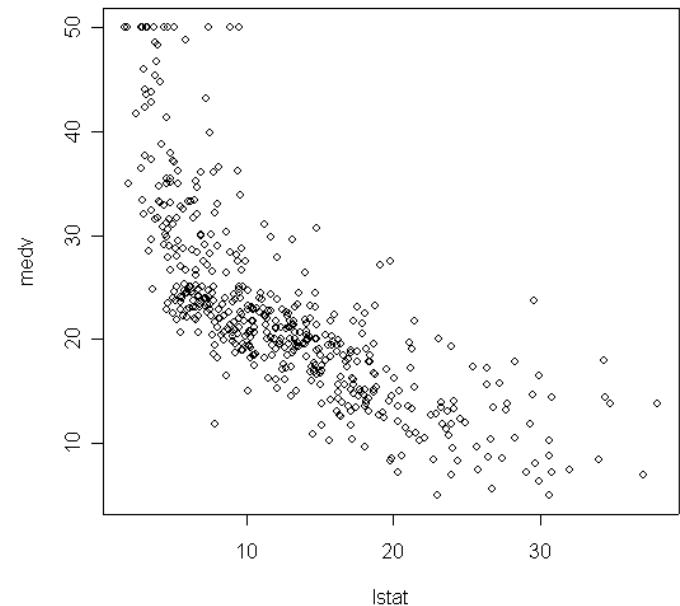
$$y_i = r(x_{i1}, \dots, x_{ip}) + \varepsilon_i$$

y **response variable** (output)

x_j **explanatory variables** (inputs)

$$E[\varepsilon_i] = 0, \text{ var}[\varepsilon_i] = \sigma^2, \varepsilon_i \propto \text{Normal}$$

- Two possible goals of modeling:
 - Prediction (*data mining*)
 - Intervention



Linear Regression

$$\text{Regression fit criterion: } \min_r E \left[\left(y_i - r(x_{i1}, \dots, x_{ip}) \right)^2 \right]$$
$$r(x_{i1}, \dots, x_{ip}) = E \left[y_i | x_{i1}, \dots, x_{ip} \right]$$

$$E \left[y_i | x_{i1}, \dots, x_{ip} \right] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Estimation of coefficients

$$y_i = b_0 + b_1 x_{i1} + \dots + b_p x_{ip} + e_i$$

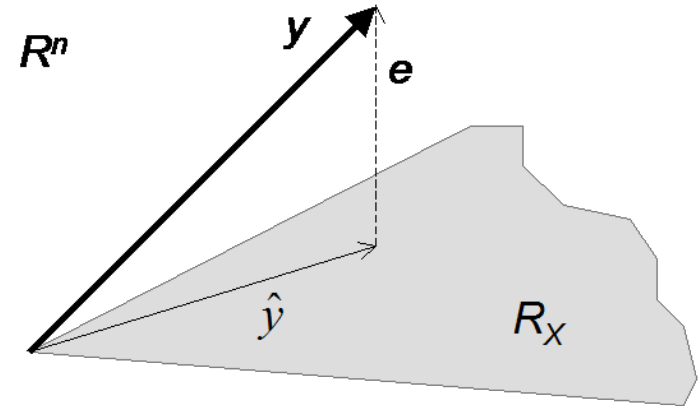
In matrix notation

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \equiv y = Xb + e = \hat{y} + e$$

Geometric interpretation

$$y_i = \hat{y}_i + e$$

$$\begin{array}{ccc} y & \hat{y} & e \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} & = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \end{array}$$



$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$$

Criterion:
$$\min_{b_0, \dots, b_p} \sum_{i=1}^n (e_i)^2 = \|e\|^2$$

$$\langle \hat{y}, e \rangle = \langle \hat{y}, y - \hat{y} \rangle = 0$$

$$\hat{y} = Xb, \quad b'X'y - b'X'Xb = 0$$

$$b = (X'X)^{-1} X'y$$

Linear regression in R

```
> plot(lstat,medv)
> l1 = lm(medv~lstat)
> lines(lstat,l1$fitted.values, col="red")
> summary(l1)
```

Call:

```
lm(formula = medv ~ lstat)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.167	-3.990	-1.318	2.034	24.500

Coefficients:

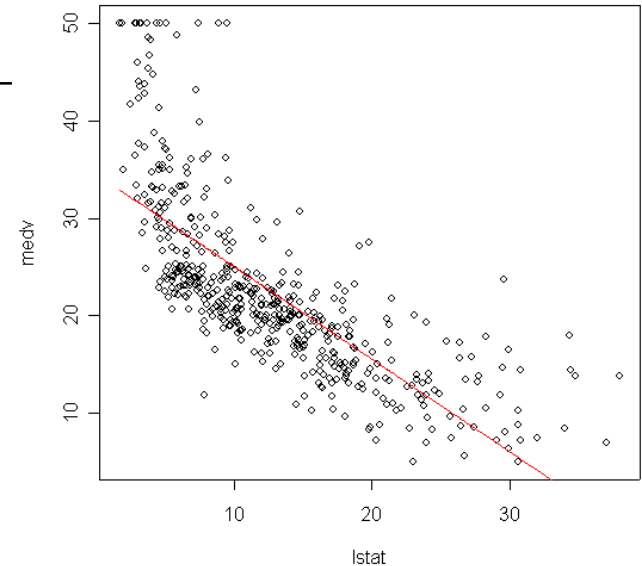
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	34.55384	0.56263	61.41	<2e-16	***
lstat	-0.95005	0.03873	-24.53	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

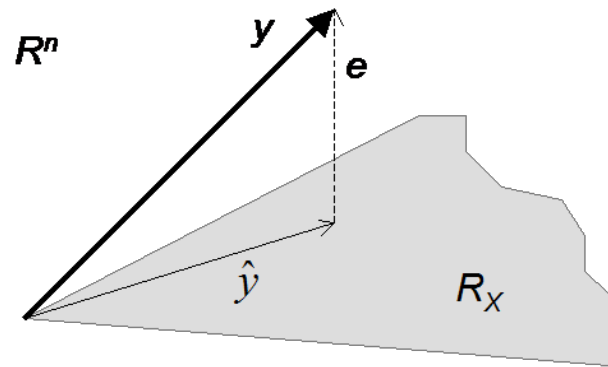
Residual standard error: 6.216 on 504 degrees of freedom

Multiple R-Squared: 0.5441, Adjusted R-squared: 0.5432

F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16



How can we assess the quality of the fit?



$$\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2 \quad \equiv \quad \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2$$

$$R^2 = \frac{\|\hat{y}\|^2}{\|y\|^2} = \cos^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n e_i^2}{(n-1) \text{var}(y)} \quad R^2 : \text{Proportion of the } y \text{ variability explained by the model}$$

```
> anova(l1)
Analysis of Variance Table
Response: medv
      Df Sum Sq Mean Sq F value    Pr(>F)
lstat    1 23243.9  23243.9   601.62 < 2.2e-16 ***
Residuals 504 19472.4    38.6
> var(medv)*(length(medv)-1)    # TOTAL VARIABILITY
[1] 42716.30
```

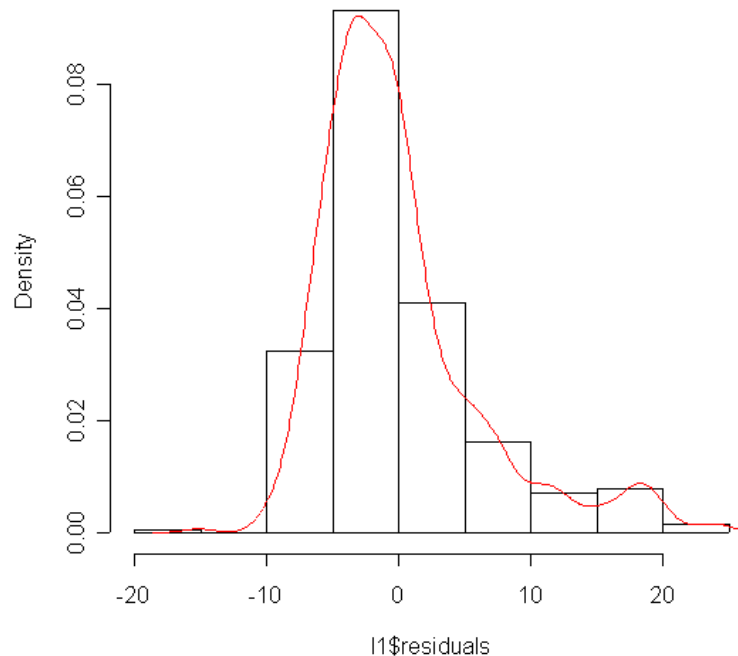
Residual analysis

Residuals shouldn't contain information, just noise

```
> dens=density(l1$residuals)
> hist(l1$residuals, prob=T)
> lines(dens,col="red")

> par(mfrow=c(2,2))
> plot(l1)
```

Histogram of l1\$residuals



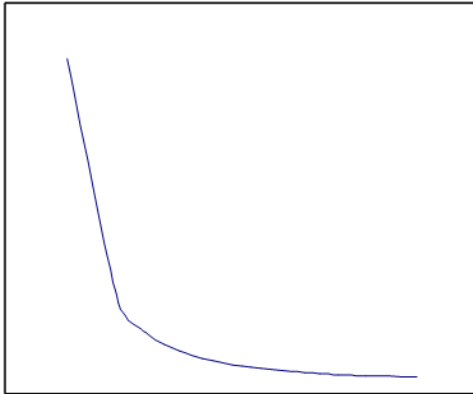
Improving the model

- Linearization
- Including more regressors
- Transforming the regressors (expanding the feature space)

→ making more complex the model

Linearisation

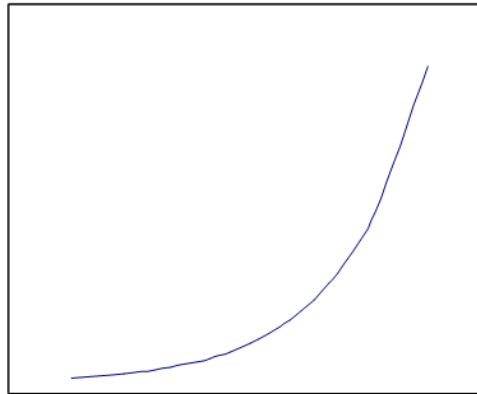
*Model of constant elasticity
(double logarithmic)*



$$y = \alpha x^{\beta}$$

$$\dot{y} = \ln(y) \quad \dot{x} = \ln(x)$$

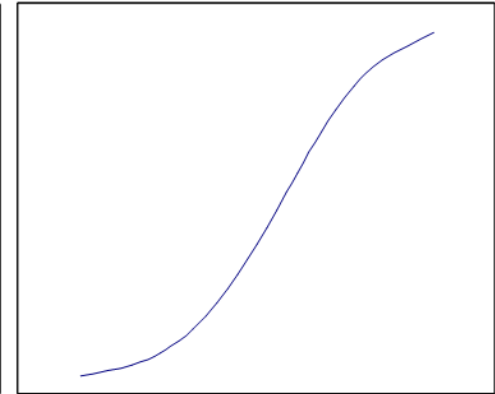
Exponential growth



$$y = \alpha \exp\{\beta x\}$$

$$\dot{y} = \ln(\alpha) + \beta x$$

Logistic model



$$y = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

$$\dot{y} = \ln \frac{y}{1-y}$$

$$\dot{y} = \alpha + \beta x$$

Polynomial regression

$$\hat{y}_i = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \dots$$

Model of constant elasticity

```
> l2=lm(log(medv)~log(lstat))
> summary(l2)
```

```
lm(formula = log(medv) ~ log(lstat))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.997564	-0.129355	-0.003148	0.140449	0.812280

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.36182	0.04210	103.60	<2e-16	***
log(lstat)	-0.55982	0.01721	-32.52	<2e-16	***

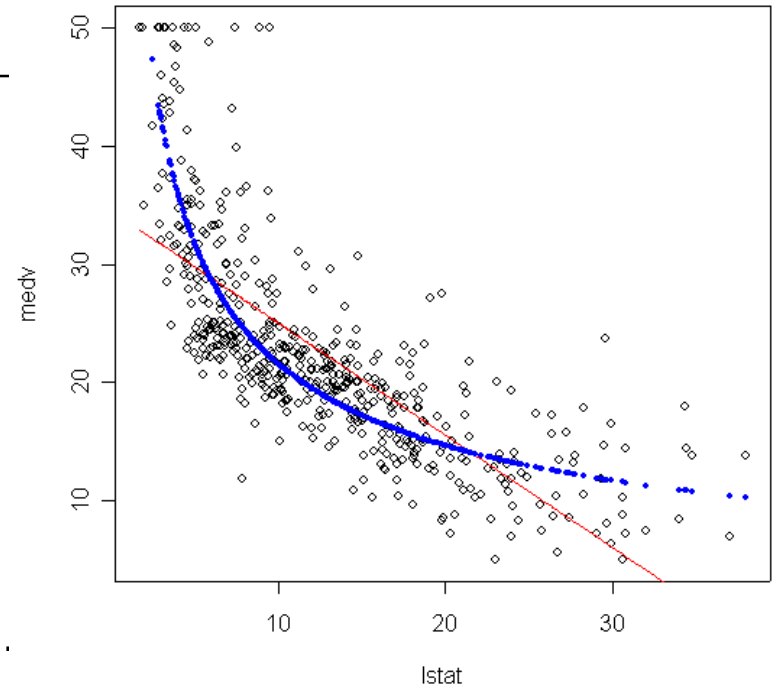
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2324 on 504 degrees of freedom

Multiple R-Squared: 0.6773, Adjusted R-squared: 0.6766

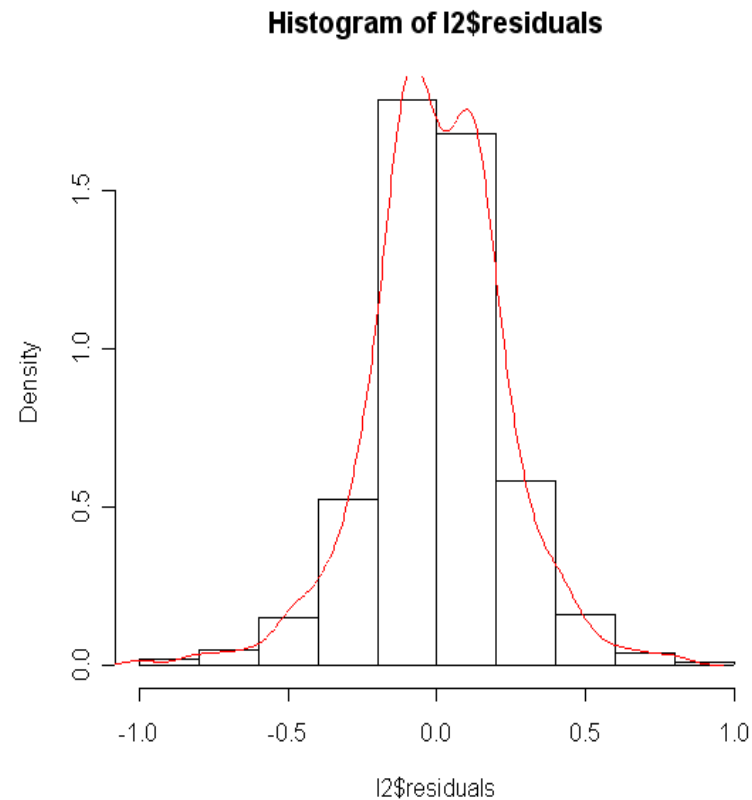
F-statistic: 1058 on 1 and 504 DF, p-value: < 2.2e-16

```
> ECMPcv = sum((l2$residuals/(1-ls.diag(l2)$hat))^2)/n
> R2cv = 1 - ECMPcv*n/(var(log(medv))*(n-1))
> R2cv
[1] 0.6742381
```



Residual analysis

```
> dens=density(l2$residuals)
> hist(l2$residuals, prob=T)
> lines(dens,col="red")
> par(mfrow=c(2,2))
> plot(l2)
```



Expansions of the X space

Transform the x variables to allow a more flexible predictors

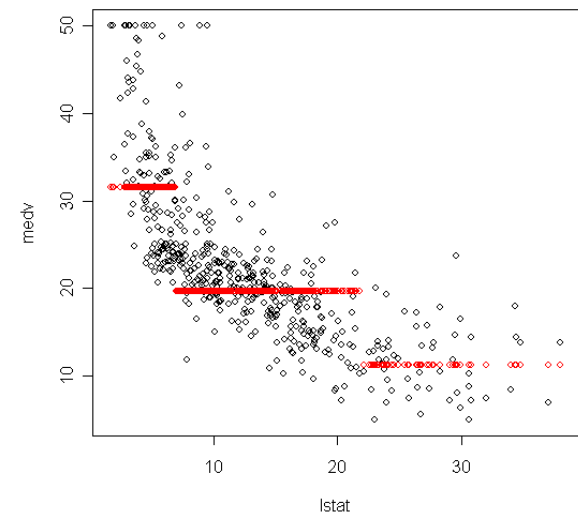
- $\mathbf{X} \rightarrow \mathbf{h}(\mathbf{X})$
 - Splines
- Splines of degree 0
 - Recode the continuous variables in intervals

$$h_1(x) = I(x < \xi_1) \quad h_2(x) = I(\xi_1 \leq x < \xi_2) \quad h_3(x) = I(\xi_2 \leq x)$$

$$y(x) = \bar{y}_1 h_1(x) + \bar{y}_2 h_2(x) + \bar{y}_3 h_3(x) + \varepsilon$$

Inconvenient: How to define the *knots*.

```
> rlstat = cut(lstat, breaks=c(0,7,22,99))  
> l7 = lm(log(medv) ~ rlstat)  
> plot(lstat, medv)  
> points(lstat, exp(l7$fitted), col="red")
```



Regression splines

Splines of degree 1:

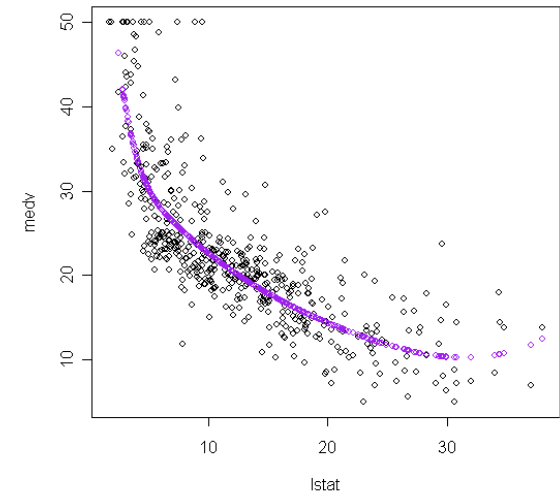
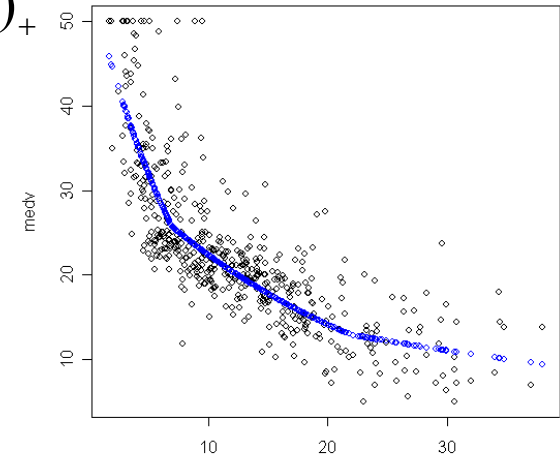
Fit a straight line in each interval imposing the coincidence in consecutive intervals in the limits (*knots*).

$$h_1(x) = 1 \quad h_2(x) = x \quad h_3(x) = (x - \xi_1)_+ \quad h_4(x) = (x - \xi_2)_+$$

```
> x1 = lstat-7
> x2 = lstat-22
> x1[x1 < 0] = 0
> x2[x2 < 0] = 0
> l8 = lm(log(medv) ~ lstat+x1+x2)
> plot(lstat, medv)
> points(lstat, exp(l8$fitted), col="blue")
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.016266	0.065010	61.779	< 2e-16	***
lstat	-0.110101	0.010761	-10.231	< 2e-16	***
x1	0.064230	0.012226	5.254	2.21e-07	***
x2	0.026388	0.007225	3.652	0.000287	***



Cubic splines

$$h_1(x) = 1 \quad h_2(x) = x \quad h_3(x) = x^2 \quad h_4 = x^3$$

$$h_5 = (x - \xi_1)_+^3 \quad h_5(x) = (x - \xi_2)_+^3$$

Non parametric regression (local regression)

$$y = r(x) + \varepsilon,$$

- In general $r(x_1, \dots, x_p) = E[y/x_1, \dots, x_p]$ is not linear neither additive. The linear assumption is often an approximation to the reality but this is the only thing that we can do for a lower n or a large p .
- We approach $r(x_1, \dots, x_p)$ directly from data without any functional assumption for r .

Local weighted fit by polinomials (*loess*)

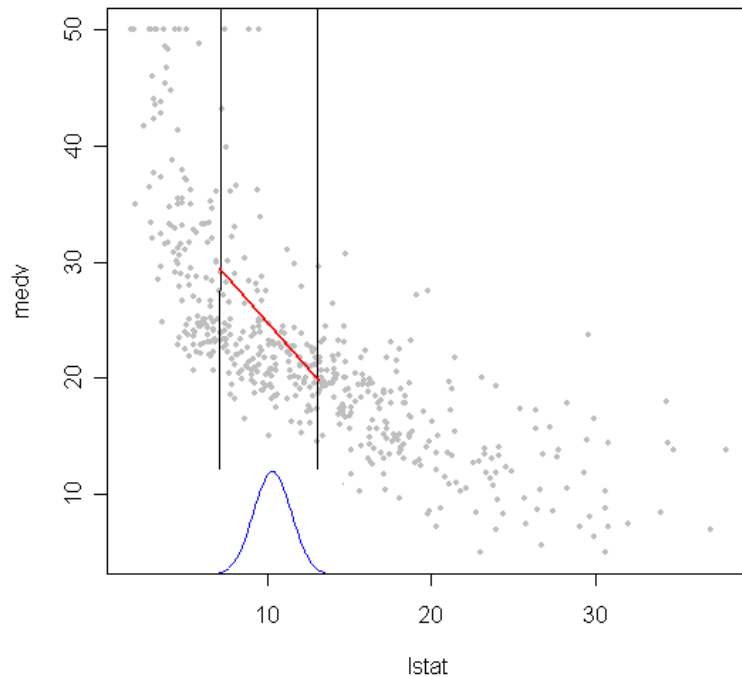
Definition of a (sliding) window h with t center

Local weights

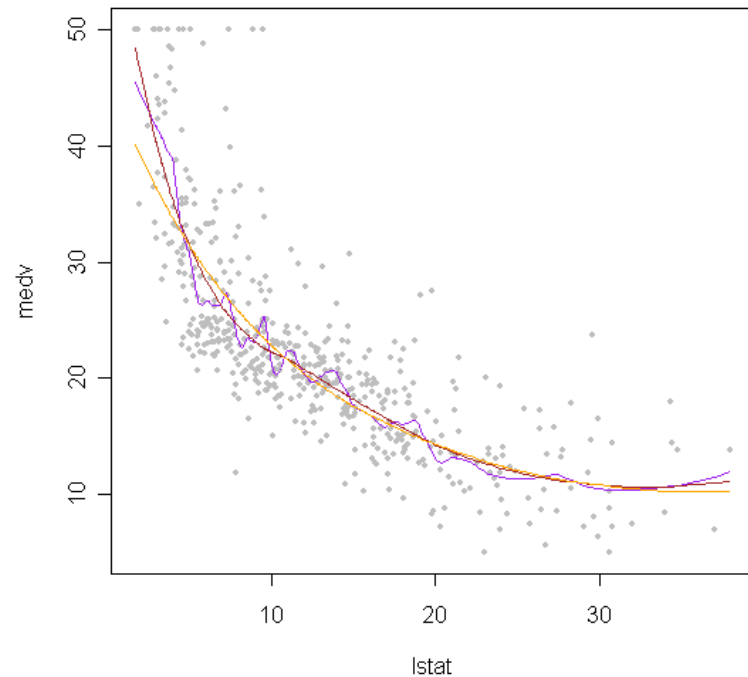
$$w_i = w(t, x_i) = \frac{K\left(\frac{x_i - t}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - t}{h}\right)}$$

$$\min_{\beta_0, \dots, \beta_q} \sum_{i=1}^n w_i \left(y_i - (\beta_0 + \beta_1(x_i - t) + \dots + \beta_q(x_i - t)^q) \right)^2$$

$$\hat{r}_q(t) = \hat{\beta}_0(t)$$



Local fit with gaussian kernel
and first degree fit



Local fit with gaussian kernel
by polinomials of degree 2 with different *span*

```
> l9 = loess(log(medv) ~ lstat, span=0.1)
> l10 = loess(log(medv) ~ lstat)
> l11 = loess(log(medv) ~ lstat, span=1)
> plot(lstat, medv, pch=20, col="gray")
> lines(lstat, exp(l9$fitted), col="purple")
> lines(lstat, exp(l10$fitted), col="brown")
> lines(lstat, exp(l11$fitted), col="orange")
```

LECTURA DE LES DADES HOUSING

```
> dd = read.table("housing.txt")
```

IDENTIFIQUEM LES VARIABLES DE HOUSING

```
> nom
```

```
=c("crim", "zn", "indus", "chas", "nox", "rm", "age", "dis", "rad", "tax", "pbrat", "b", "lstat",  
  "medv")
```

```
> names(dd) = nom
```

PLOT DEL VALOR DELS HABITATGES I %CLASSE BAIXA

```
> plot(lstat, medv)
```

INCORPOREM LA RECTA DE REGRESSIÓ LINEAL

```
> l1=lm(medv~lstat)
```

```
> lines(lstat, l1$fitted.values, col="red")
```

INCORPOREM LA REGRESSIÓ DOBLE-LOGARITMICA

```
> l2=lm(log(medv)~log(lstat))
```

```
> points(lstat, exp(l2$fitted.values), col="blue", pch=20)
```

INCORPOREM LA FUNCIO DE REGRESSIÓ PONDERADA LOCAL LOESS

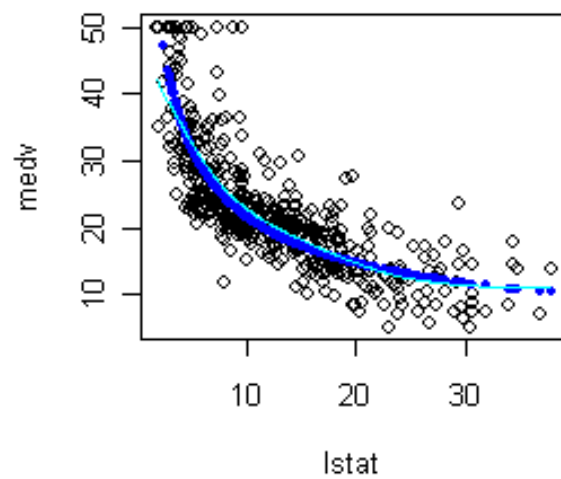
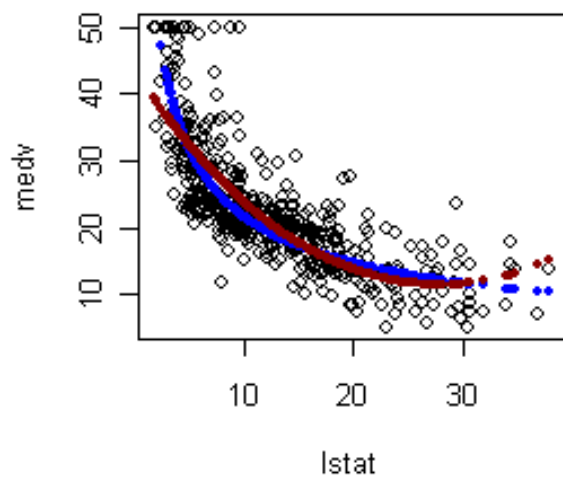
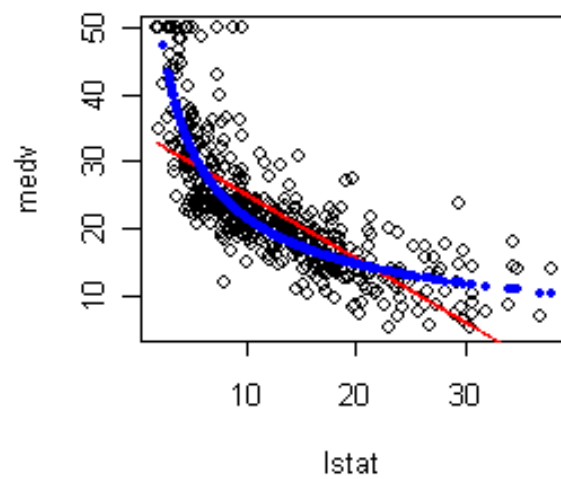
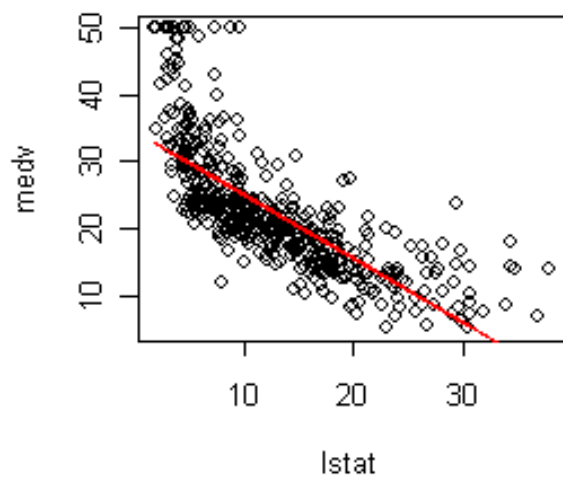
```
> l4=loess(medv~lstat, span=2/3)
```

```
> points(lstat, l4$fitted, col="dark red", pch=20)
```

INCORPOREM LA FUNCIO DE REGRESSIÓ PER SPLINES CUBICS

```
> l5 = smooth.spline(lstat, medv, df=5)
```

```
> lines(l5$x, l5$y, col="cyan")
```



Introducing categorical variables as predictors

We introduce the binary codification of the categorical variables

Ex. Profession: manual, clerk, executive

	u_1	u_2	u_3
clerk	0	1	0
manual	1	0	0
executive	0	0	1

We introduce all modalities in the model minus one (since one modality is redundant; it defines the baseline).

Coefficients of modalities defines the differential effect of the modality respect of the baseline one.

To measure the significance of a categorical variable we need to look at the **Anova** table (the t statistics gives the significance of the difference respect to the baseline but the Fisher's F gives the significance of the variable).

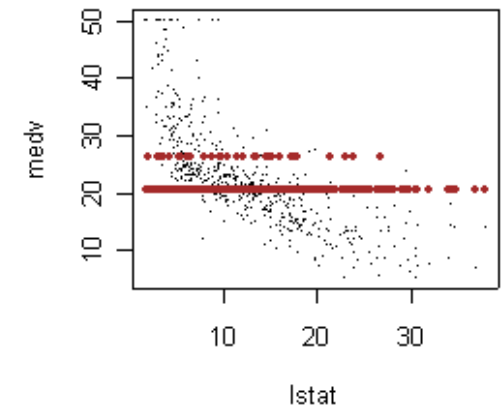
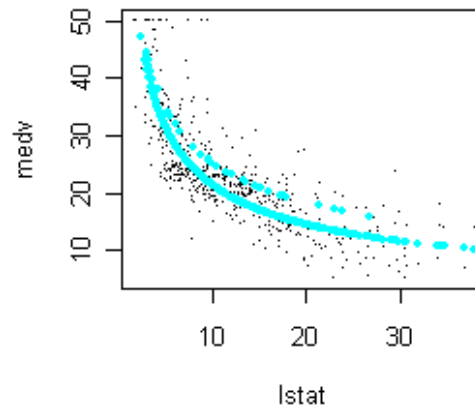
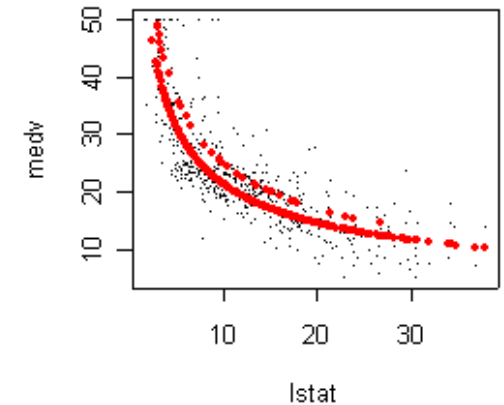
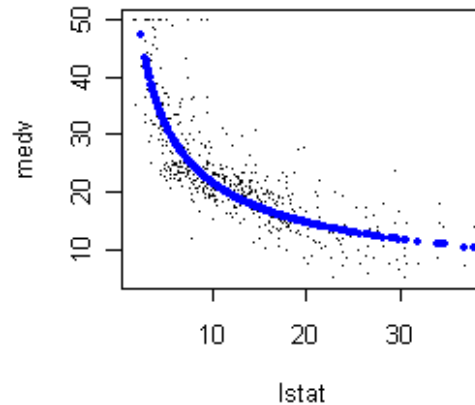
Value of housing depending of RIVER

Example: Value of households depending they are in the Charles river or not

Models:

1. $y \sim x$
2. $y \sim x + A$
3. $y \sim x + A + x * A$
4. $y \sim A$

$$\begin{aligned} y &= \log(mdev) \\ x &= \log(lstat) \\ A &= RIVER \end{aligned}$$



Multiple regression

improving the prediction

```
> l3=lm(log(medv) ~ log(lstat)+crim+rooms)
> summary(l3)

lm(formula = log(medv) ~ log(lstat) + crim + rooms)

Residuals:
    Min       1Q   Median       3Q      Max
-0.70755 -0.11746 -0.01710  0.12014  0.90245

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.433037   0.149664  22.938 < 2e-16 ***
log(lstat)   -0.415149   0.021956 -18.909 < 2e-16 ***
crim         -0.011826   0.001175 -10.067 < 2e-16 ***
rooms         0.100008   0.017676   5.658 2.58e-08 ***
---
Residual standard error: 0.2081 on 502 degrees of freedom
Multiple R-Squared:  0.7423,    Adjusted R-squared:  0.7407
F-statistic: 481.9 on 3 and 502 DF,  p-value: < 2.2e-16

> ECMPcv = sum((l3$residuals/(1-ls.diag(l3)$hat))^2)/n
> R2cv = 1 - ECMPcv*n/(var(log(medv))*(n-1))
> R2cv
[1] 0.7344521
```

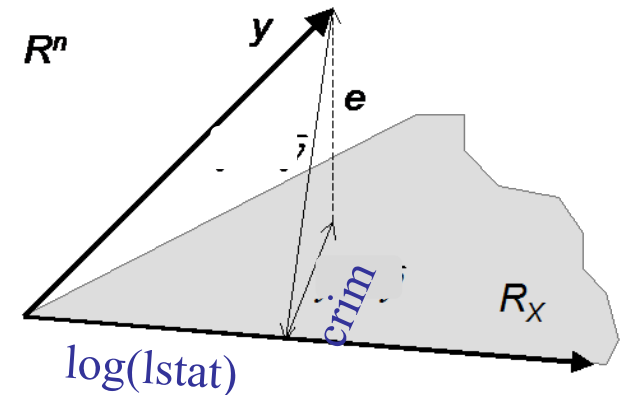
Importance of variables

```
> l3 = lm(log(medv) ~ log(lstat)+crim+rooms)
> anova(l3)
Analysis of Variance Table

Response: log(medv)
          Df Sum Sq Mean Sq  F value    Pr(>F)
log(lstat)  1  57.145   57.145 1319.190 < 2.2e-16 ***
crim        1   4.099    4.099   94.622 < 2.2e-16 ***
rooms       1   1.387    1.387   32.013 2.577e-08 ***
Residuals 502  21.746    0.043
---
```

```
> l3 = lm(log(medv) ~ rooms+crim+log(lstat))
> anova(l3)
Analysis of Variance Table

Response: log(medv)
          Df Sum Sq Mean Sq  F value    Pr(>F)
rooms       1  33.704   33.704   778.06 < 2.2e-16 ***
crim        1  13.439   13.439   310.23 < 2.2e-16 ***
log(lstat)  1  15.488   15.488   357.53 < 2.2e-16 ***
Residuals 502  21.746    0.043
---
```



Model selection

Data Mining goal of modeling: To minimize the Mean Prediction Error (Generalization Error)

- Let be a model: $\hat{y} = f(X, w)$
 - f represents a class of functions to predict y
 - X is the matrix of explanatory variables
 - w represents the coefficients of the model
- Loss function: $L(y, f(X, w))$
 - If regression: $\sum (y - f(X, w))^2$
 - If classification: *percentage of misclassified*

Model selection

- Risk (=Generalization error)
Expected loss with independent data

$$Risk = E[L] = \int L(y, f(X, w)) dP(X, y)$$

(if regression) $Risk = GE(\hat{y}) = E[y_0 - \hat{y}_0]^2$

- Goal: Minimize the *Risk* respect w
- Risk estimation: Empirically (on observed data)

$$Risk_{learn} = \frac{1}{n} \sum L(y, f(X, w))$$

If computed in the learning data we call it $Risk_{learn}$

(What we expect from $Risk_{learn}$?
if $n \rightarrow \infty$)

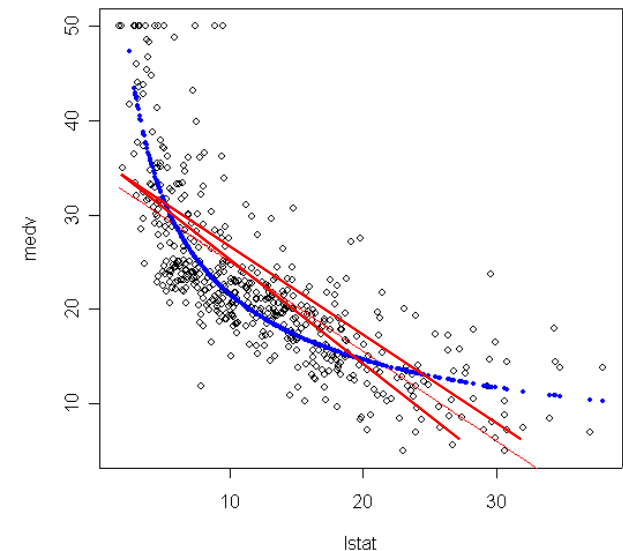
Model selection

Components of the Risk (Generalization Error) in regression:

$$GE(\hat{y}) = E[y_0 - \hat{y}_0]^2 = E[\hat{y}_0 - r(x_0)]^2 + \sigma_0^2 = \sigma_0^2 + \text{var}(\hat{y}_0) + (E[\hat{y}_0] - r(x_0))^2$$

Risk = Generalization (Prediction) Error

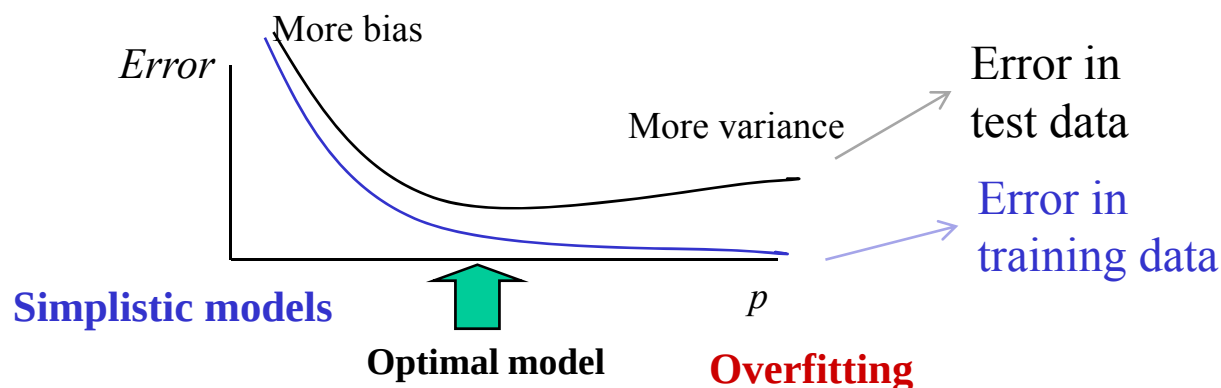
1. **Random fluctuation** of the phenomenon (deviation of the actual data respect to the true model due to multiple small and unknown causes)
2. **Variance** of fit (deviation of the prediction trained model respect to its average, i.e. with another sample we would obtained a different fit)
3. **Bias** (deviation of the true model respect to the average prediction of the trained model)



Model selection

With p predictors we have $2^p - 1$ possible models

Trade-off between bias and variance



Components of bias:

- Bad specification of the regression function r

- Not including all the relevant predictors in the model

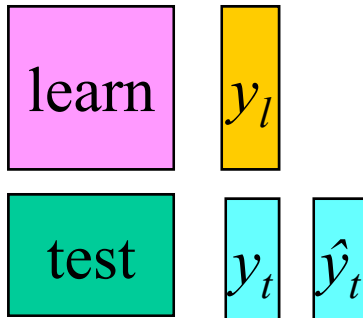
Components of prediction variance:

- Complexity of the model

How can we evaluate honest estimates of the generalization error?

To obtain reliable estimations we need to estimate it applying the model in an **independent** sample (from the one used to estimate the model)
Instead of the *MQPE*, we can compute the R^2 .

Test sample



$$GE_{test}(\hat{y}) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i^{test} - \hat{y}_i^{test})^2$$
$$R_{test}^2 = 1 - \frac{GE_{test}(\hat{y})}{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i^{test} - \bar{y}^{test})^2}$$

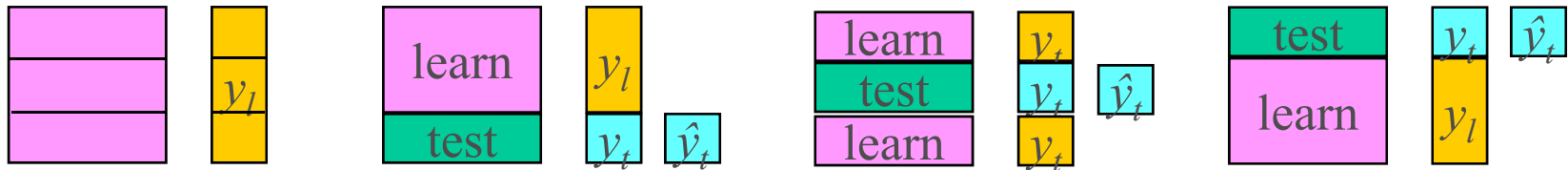
We split the available data in two parts at random, one used for learning (2/3) the other used to test the model (1/3). Although it would be better to use as test sample a complete independent sample (i.e. obtained in a posterior moment).

Evaluation of the generalization error in **R** with a test sample

```
> ypred = predict(l1, data=dd.test)
> 1 - (sum((ypred-ytest)^2)/sum((ytest-mean(ytest))^2))
```

Estimation of the generalization error by crossvalidation

We split the available data in k parts at random



$$GE_{cv}(\hat{y}) = \frac{1}{n} \sum_{k=1}^R \sum_{i \in k} (y_i^{test} - \hat{y}_i^{test})^2$$

$$R_{cv}^2 = 1 - \frac{GE_{cv}(\hat{y})}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

$$GE_{cv}(\hat{y}) = \frac{1}{n} PRESS = \frac{1}{n} \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

one leave out crossvalidation
(each individual is a part, $k=n$)

```
> n      = length(medv)
> ECMPcv = sum((l1$residuals/(1-ls.diag(l1)$hat))^2)/n
> R2cv   = 1 - ECMPcv*n/(var(medv)*(n-1))
> R2cv
[1] 0.5393236
```

Model selection

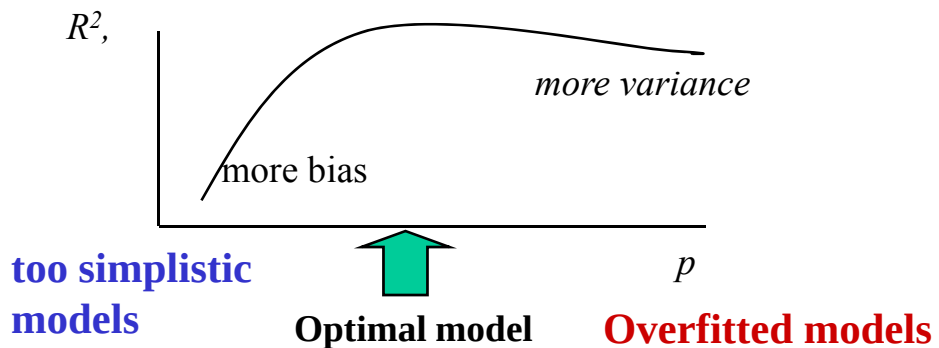
- Calculate the GE in a different sample of the one used to estimate the parameters, with a model with increasing complexity and select the model giving the minimum generalization error (or maximum R^2).

$$GE_{test} = \frac{1}{n_{test}} \sum_i^{n_{test}} (y_{i_{test}} - \hat{y}_{i_{test}})^2$$

$$R_{test}^2 = 1 - \frac{GE_{test}(\hat{y})}{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_{i_{test}} - \bar{y}_{test})^2}$$

Trade-off between bias and variance

R_{test}^2 computed on the test sample }
 R_{cv}^2 by crossvalidation }



Model selection by penalized likelihood

Instead of computing iteratively models of increasing complexity, we optimize a mix of the likelihood and the complexity of the model

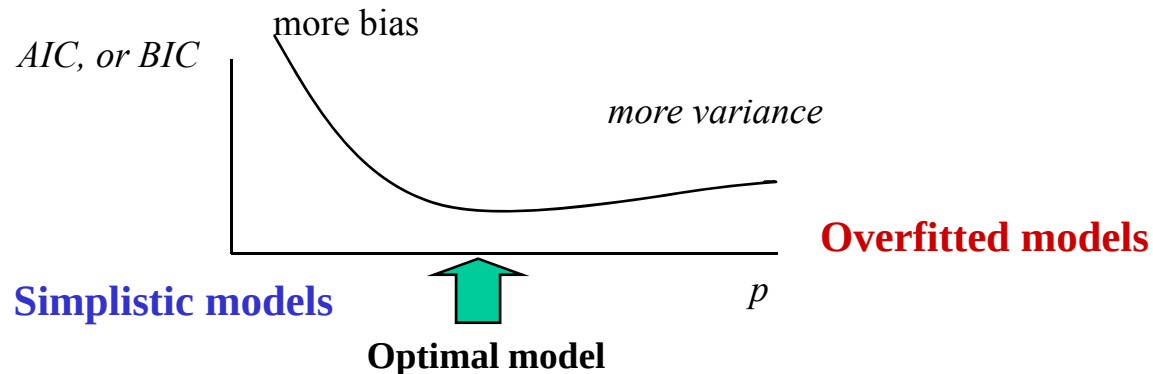
$$\text{Min } \text{Error}_{\text{learn}} + \text{Complexity Penalization } (= p)$$

AIC Akaike criterion

$$AIC = n \log \frac{\sum_i^n e_i^2}{n} + 2p$$

BIC

$$BIC = n \log \frac{\sum_i^n e_i^2}{n} + \ln(n)p$$



```
> l4 = lm(log(medv) ~
log(lstat)+crim+resid+indus+river+nox+rooms+age+dis+rad+tax+teaching+black)
> step(l4)
```

Start: AIC= -1697.69

```
log(medv) ~ log(lstat) + crim + resid + indus + river + nox + rooms +
age + dis + rad + tax + teaching + black
```

	Df	Sum of Sq	RSS	AIC
- resid	1	0.002231	16.71	-1699.62
- indus	1	0.02	16.73	-1699.19
<none>			16.71	-1697.69
- age	1	0.07	16.78	-1697.63
- river	1	0.21	16.93	-1693.25
- rooms	1	0.43	17.14	-1686.76
- black	1	0.49	17.21	-1684.92
- tax	1	0.53	17.24	-1683.86
- nox	1	0.86	17.57	-1674.38
- dis	1	0.89	17.60	-1673.45
- rad	1	0.97	17.68	-1671.15
- teaching	1	1.62	18.33	-1652.92
- crim	1	2.95	19.66	-1617.50
- log(lstat)	1	8.42	25.13	-1493.25

Step: AIC= -1699.62

```
log(medv) ~ log(lstat) + crim + indus + river +
nox + rooms + age + dis + rad + tax + teaching + black
```

	Df	Sum of Sq	RSS	AIC
- indus	1	0.02	16.73	-1701.15
<none>			16.71	-1699.62
- age	1	0.07	16.78	-1699.61
- river	1	0.21	16.93	-1695.18
- rooms	1	0.44	17.15	-1688.47
- black	1	0.49	17.21	-1686.87
- tax	1	0.54	17.26	-1685.46
- nox	1	0.86	17.57	-1676.24
- rad	1	0.97	17.68	-1673.06
- dis	1	1.03	17.74	-1671.48
- teaching	1	1.82	18.53	-1649.31
- crim	1	2.96	19.67	-1619.08
- log(lstat)	1	8.49	25.20	-1493.87

Model Selection in R using AIC

Step: AIC= -1701.15

```
log(medv) ~ log(lstat) + crim + river + nox + rooms + age +
dis + rad + tax + teaching + black
```

	Df	Sum of Sq	RSS	AIC
<none>			16.73	-1701.15
- age	1	0.07	16.80	-1701.11
- river	1	0.23	16.96	-1696.29
- rooms	1	0.43	17.16	-1690.39
- black	1	0.49	17.22	-1688.59
- tax	1	0.57	17.30	-1686.15
- nox	1	0.86	17.59	-1677.75
- rad	1	0.98	17.71	-1674.45
- dis	1	1.19	17.92	-1668.38
- teaching	1	1.82	18.54	-1651.02
- crim	1	2.98	19.71	-1620.13
- log(lstat)	1	8.47	25.20	-1495.84

Multiple R-Squared: 0.8017
R2cv: 0.785853

Model Selection

1. The training data will be used to choose the model. Hence we will apply all the potential models to this data and we will see which model perform better. We take as best model the one with lower GE .
2. To have a reliable GE we need to compute it independently of the data used for training the model. Hence we need to split the available data in two parts at random, one used to estimate the coefficients and the other to calculate its GE , or alternatively, we calculate the GE by cross validation on the training data. (Instead of the GE , we can evaluate the AUC of the ROC or *concentration curve*).
3. Once we have chosen the best model, we have to evaluate its quality on a test (holdout) sample.
4. We split the total available data in two parts at random, one will be the training data and the other will be the test data (holdout).
5. We distinguish from selecting the model (functional form of $r(x_1, \dots, x_p)$ with specification of the intervening p explanatory variables), from tuning (estimating) the coefficients of these variables. The former is done in 2, and the latter in 6.
6. We estimate the tuning parameters of the chosen model in the training data and estimate its GE with the test data. In order to have a reliable measure of the GE we can perform this operation several times.
7. We give as final model the average model obtained and the mean GE obtained in these simulations with its standard error.