Linear models

Data Mining course
Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja
Two types of datasets to analyze

Data in *Data Mining*:
massive, secondary, not random, with errors and missing values

**topics**

<table>
<thead>
<tr>
<th>Socio-econ.</th>
<th>Opinions</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data to explore</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Output(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data to modelize</strong></td>
<td></td>
</tr>
</tbody>
</table>
A statistical model

\[ y = r(x_1, \ldots, x_p) + \varepsilon = \hat{y} + \varepsilon \]

**Data**: Response variable (output)
- continuous → **Regression model**
- categorical → **Classification model**

**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**  \( P[C_k|x_1, \ldots x_p] \)
- If continuous **REGRESSION**  \( E[y|x_1, \ldots x_p] \)
### 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:**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIM</td>
<td>per capita crime rate by town</td>
</tr>
<tr>
<td>RESID</td>
<td>proportion of residential land zoned for lots over 25,000 sq.ft.</td>
</tr>
<tr>
<td>INDUS</td>
<td>proportion of non-retail business acres per town</td>
</tr>
<tr>
<td>RIVER</td>
<td>Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)</td>
</tr>
<tr>
<td>NOX</td>
<td>nitric oxides concentration (parts per 10 million)</td>
</tr>
<tr>
<td>ROOMS</td>
<td>average number of rooms per dwelling</td>
</tr>
<tr>
<td>AGE</td>
<td>proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>DIS</td>
<td>weighted distances to five Boston employment centres</td>
</tr>
<tr>
<td>RAD</td>
<td>index of accessibility to radial highways</td>
</tr>
<tr>
<td>TAX</td>
<td>full-value property-tax rate per $10,000</td>
</tr>
<tr>
<td>TEACHING</td>
<td>pupil-teacher ratio by town</td>
</tr>
<tr>
<td>BLACK</td>
<td>1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town</td>
</tr>
<tr>
<td>LSTAT</td>
<td>% lower status of the population</td>
</tr>
<tr>
<td>MEDV</td>
<td>Median value of owner-occupied homes in $1000's</td>
</tr>
</tbody>
</table>
DM Linear models, T. Aluja
DM Linear models, T. Aluja
Statistical modeling of a continuous variable

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

\[ y_i = r(x_{i1}, \cdots, x_{ip}) + \varepsilon_i \]

- \( y \) response variable (output)
- \( x_j \) explanatory variables (inputs)

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

- Two possible goals of modeling:
  - Prediction \((data \ mining)\)
  - Intervention
Linear Regression

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

\[
r(x_{i1}, \ldots, x_{ip}) = E \left[ y_i | x_{i1}, \ldots, x_{ip} \right]
\]

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

\[
y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i
\]

Estimation of coefficients

\[
y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i
\]

In matrix notation

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix} =
\begin{bmatrix}
  1 & x_{i1} & \cdots & x_{ip} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n1} & \cdots & x_{np}
\end{bmatrix}
\begin{bmatrix}
  b_0 \\
  b_1 \\
  \vdots \\
  b_p
\end{bmatrix} +
\begin{bmatrix}
  e_1 \\
  \vdots \\
  e_n
\end{bmatrix}
\]

\[
y = Xb + e = \hat{y} + e
\]
Geometric interpretation

\[ y_i = \hat{y}_i + e \]

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix}
= 
\begin{bmatrix}
  \hat{y}_1 \\
  \vdots \\
  \hat{y}_n
\end{bmatrix}
+
\begin{bmatrix}
  e_1 \\
  \vdots \\
  e_n
\end{bmatrix}
\]

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

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

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

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

\[ b = (X'X)^{-1} X'y \]
Linear regression in R

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

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?

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

$$R^2 = \frac{\|\hat{y}\|^2}{\|y\|^2} = \cos^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} e_i^2}{(n-1) \text{var}(y)}$$  

$R^2$ : Proportion of the $y$ variability explained by the model

> anova(l1)
Analysis of Variance Table
Response: medv
   Df Sum Sq Mean Sq F value    Pr(>F)
 1stat   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

```r
> 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
**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**

\[ \hat{y}_i = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots \]
Model of constant elasticity

```r
> 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) 4.361822   0.042100 103.600   <2e-16 ***
log(lstat)  -0.559820   0.017209  -32.520   <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

> ECMPerf = sum((l2$residuals/(1 - ls.diag(l2)$hat))^2)/n
> R2cv = 1 - ECMPerf*n/(var(log(medv))*(n-1))
> R2cv

[1] 0.6742381
```
Residual analysis

```r
> 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

- \( X \rightarrow h(X) \)
  - Splines

- Splines of degree 0
  - Recode the continuous variables in intervals

\[
\begin{align*}
  h_1(x) &= I(x < \xi_1) \\
  h_2(x) &= I(\xi_1 \leq x < \xi_2) \\
  h_3(x) &= I(\xi_2 \leq x)
\end{align*}
\]

\[
y(x) = \bar{y}_1 h_1(x) + \bar{y}_2 h_2(x) + \bar{y}_3 h_3(x) + \epsilon
\]

Inconvenient: How to define the knots.

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

DM Linear models, T. Aluja
Regression splines

Splines of degree 1:
Fit a straight line in each interval imposing the coincidence in consecutive intervals in the limits (knots).

\[ h_1(x) = 1 \quad h_2(x) = x \quad h_3(x) = (x - \xi_1) \quad h_4(x) = (x - \xi_2) \]

Cubic splines

\[ h_1(x) = 1 \quad h_2(x) = x \quad h_3(x) = x^2 \quad h_4(x) = x^3 \]
\[ h_5(x) = (x - \xi_1)^3 \quad h_5(x) = (x - \xi_2)^3 \]

R code:

```r
> 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 *** |

DM Linear models, T. Aluja
Non parametric regression (local regression)

\[ y = r(x) + \varepsilon, \]

- In general \( r(x_1, \ldots, x_p) = E[y/x_1, \ldots, x_p] \) 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 \( r(x_1, \ldots, x_p) \) directly from data without any functional assumption for \( r \).

**Local weighted fit by polynomials (loess)**

Definition of a (sliding) window \( h \) with \( t \) center

Local weights

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

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

\[ \hat{r}_q(t) = \hat{\beta}_0(t) \]
Local fit with gaussian kernel and first degree fit

Local fit with gaussian kernel by polynomials of degree 2 with different span

```r
> 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 FUNCIÓ DE REgressió PONDERADA LOCAL LOESS
> l4 = loess(medv~lstat,span=2/3)
> points(lstat,l4$fitted, col="dark red",pch=20)

INCORPOREM LA FUNCIÓ 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

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>clerk</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>manual</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>executive</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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. \( y \sim x + A + x \times A \)
4. \( y \sim A \)

\( y = \log(mdev) \)
\( x = \log(lstat) \)
\( A = \text{RIVER} \)
Multiple regression
improving the prediction

```r
> 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)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(lstat)</td>
<td>1</td>
<td>57.145</td>
<td>57.145</td>
<td>1319.19</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>crim</td>
<td>1</td>
<td>4.099</td>
<td>4.099</td>
<td>94.62</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>rooms</td>
<td>1</td>
<td>1.387</td>
<td>1.387</td>
<td>32.013</td>
<td>2.577e-08 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>502</td>
<td>21.746</td>
<td>0.043</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
---

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

Response: log(medv)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rooms</td>
<td>1</td>
<td>33.704</td>
<td>33.704</td>
<td>778.06</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>crim</td>
<td>1</td>
<td>13.439</td>
<td>13.439</td>
<td>310.23</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>log(lstat)</td>
<td>1</td>
<td>15.488</td>
<td>15.488</td>
<td>357.53</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>502</td>
<td>21.746</td>
<td>0.043</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
---
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)

\[
Risk_{\text{learn}} = \frac{1}{n} \sum L(y, f(X, w))
\]
If computed in the learning data we call it \( Risk_{\text{learn}} \)

(What we expect from \( Risk_{\text{learn}} \)?
if \( n \to \infty \)
Model selection

Components of the Risk (Generalization Error) in regression:

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

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

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

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

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 independende sample (i.e. obtained in a posterior moment).

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

```r
> 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

\[
GE_{cv}(\hat{y}) = \frac{1}{n} \sum_{k=1}^{R} \sum_{i \in k} (y_{i}^{\text{test}} - \hat{y}_{i}^{\text{test}})^2
\]

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

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

```r
> 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 $R^2$).

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

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

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

Trade-off between bias and variance

$R^2$, $P$

too simplistic models
Optimal model
Overfitted models

more variance
more bias
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**

\[ BIC = n \log \frac{\sum_i^n e_i^2}{n} + \ln(n)p \]
```r
> 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
- 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
- dis         1   0.98    17.71 -1674.45
- rad         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 -1493.84

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
- dis         1   0.98    17.71 -1674.45
- rad         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 in **R** using AIC
Model Selection

1. The training data will be used to chose 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 $r(x_1,\ldots,x_p)$ 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.