Linear models

(Some basic results)

R. Bellio & N. Torelli

Spring 2018

University of Udine & University of Trieste

Matrix notation

Inference in Linear models

Model validation and model selection

Matrix notation

The linear model

- Linear models (LM) is appropriate when analyzing the relationship between a quantitative *response variable* Y and a set of *covariates* $x_1, x_2, \ldots, x_{p-1}$ $(p \geq 2)$.
  It is assumed that a sample of n values of the response variable Y is observed as well as n values of each covariate.
- The aim is to evaluate the impact of covariates on the mean μ_i of the response variable Y_i for the i-th unit. In a linear model this is represented by the equation

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} \tag{1}$$

  The value y_i for the i-th unit of the sample can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i \ , \tag{2}$$

  the model above can be also written for the set of all the n units in matrix notation $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $X\boldsymbol{\beta}$ is the so called systematic component
- $\boldsymbol{\epsilon}$ is the stochastic component.

Matrix notation

The content of each component in matrix notation is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

con:

- $\mapsto$ $\boldsymbol{y} =$ is the vector of the values of the response variable $(n \times 1)$;
- $\mapsto$ $X =$ is a matrix $(n \times p)$ which contains the values of the covariates. This matrix is also called the *design matrix*;
- $\mapsto$ $\boldsymbol{\beta} =$ is the vector $(p \times 1)$ of the regression coefficients;
- $\mapsto$ $\boldsymbol{\epsilon} =$ is the vector $(n \times 1)$ of the stochastic components.

The model written for the i-th unit can be also written as $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$ where $\boldsymbol{x}_i^T$ is the i-th row of the design matrix.

Matrix notation: main assumptions

In the linear model

- The response variable Y is a quantitative variable
- The covariates x could be either:
  - $\mapsto$ quantitative (numeric) variables or
  - $\mapsto$ categorical variables (factors).
- it is usually assumed that the values in the matrix X are fixed constant (non stochastic). X it is also called the **design matrix**
- the design matrix X it is assumed to be of full rank. Since usually $n >> p$ this means that the rank of X is p (that is $min(p, n)$). The columns of X are linearly independent vectors.

Matrix notation: main assumptions

The linear model is completely specified by assumptions on the stochastic components, the random variables ϵ_i

1. $E(\epsilon_i = 0)$ or equivalently $E(\boldsymbol{\epsilon}) = \mathbf{0}$
2. $Var(\epsilon_i) = \sigma^2$ homoscedasticity
3. $E(\epsilon_i, \epsilon_j) = 0$ per $i \neq j$ uncorrelation.
   The last two conditions can be more concisely expressed in matrix form as
   $Cov(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \boldsymbol{I}_n$
   where $Cov(\boldsymbol{\epsilon})$ denotes variance-covariance matrix of the random vector $\boldsymbol{\epsilon}$.
   The assumption 1-3 are called the second order assumptions (they refer only to the first two moments of the variables).
   A distributional assumption is then often added
4. $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \Sigma)$
   In matrix notation
   $E(\boldsymbol{y}) = X\beta$ and $\boldsymbol{y} \sim N(X\beta, \sigma^2 \boldsymbol{I})$

Discussion of the assumptions

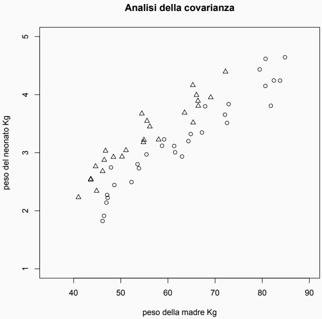
- *Linearity.* The assumption about the linear effect of the covariates is actually not very restrictive. Non linear relationships can be introduced by appropriate transformation of the xs. For instance:
    - $y_i = \beta_0 + \beta_1 log(z_i) + \epsilon_i$ introduces a logarithmic effect of z_i. But if one simple redefines $x_i = log(z_i)$ then we are back to a standard linear model for the transformed variable x.
    - x^2 introduces a parabolic effect.
- *Homoscedasticity of the random components*. This is the standard assumption. The use of diagnostic checks can help verifying it. If possible departures from omoschedasticity are ignored it can impair quality of estimates. Possible remedies can be introduced
- *Uncorrelated random components.*
  It is assumed that all random components are mutually uncorrelated. In some context this assumption is questionable (for instance with data that are temporally or spatially ordered). Also this assumption can be verified with diagnostic tools and remedies are available.

Continuous covariates, factors, interactions

- When the covariate is a quantitative one, under the linearity assumption, the the value of the parameter associated to it represents simply the derivative of Y wrt to x.
- The effect of a categorical variable (a factor) measure the difference in the expected value of the response variable for each value of the factor wrt the reference category for the factor itself (all the other variables being equal). There will be then as many parameters as the number of levels of the factor minus one.
- usually the interactions between two (or more) variables is introduced. Interpretation of interaction is easier when it refers to two factors or to a factor and a numeric variable.

Another example: one factor and a quantitative variable

`Weight`, Y, in kilograms, for a sample of newborn babies, from smoker mothers `smokers` (F) (in the graph different symbols are used for smokers - triangles - and non smokers - circles). For each women the pre-pregnancy weight `mother weight` x is observed)



As expected the weights of newborn babies is greater on the average when the mothers are non smokers.It seems that a systematic difference exists between the two groups though the relationship between weight of the mother and weight of the babies does not change.

A model with a continuos covariate and a factor

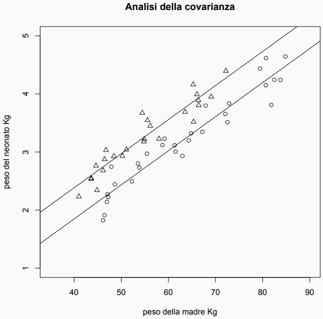
The two variables y="weight of the babies" and X="weight of the mother"" are both continuous numeric variable while the variable F is a factor with two levels ($F = 1$ if smoker, $F = 2$ if non smoker). The model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \epsilon_i \ .$$

with the corresponding matrices:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \vdots & & \vdots \\ 1 & x_k & 1 \\ \vdots & & \vdots \\ 1 & x_n & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \ .$$

Intrpretation of the parameters



For the given specification

- parameter β_1 measures the (linear) effect of the weight of the mother on the weight of the baby and represents the (common) slope of the two lines in the graph;
- β_0 measures the intercept of the smoker's line and β_2 is, for a given weight of the mother, the vertical distance between the two lines.

An alternative parametrization

Any model, particularly when factors are involved, can have alternative parametrization.

The model introduced above can be also written in the following form:

$$Y_i = \gamma_1 x_i + \gamma_2 I_{(F_i=1)} + \gamma_3 I_{(F_i=2)} + \epsilon_i \ , \tag{3}$$

where $I_{(F_i=j)}$ is a indicator variable which takes on 1 if $(F_i = j)$, $j = 1, 2$, and 0 otherwise.

The model is equivalent but interpretation of parameters changes.

In this case the matrix form is:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 & 0 \\ x_2 & 1 & 0 \\ \vdots & & \vdots \\ x_n & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} .$$

Interpretation of the parameters

In this new parametrization

- parameter β_1 measures the effect of the weight of the mother on the weight of the baby
- β_2 and β_3 estimate the mean of the y the weight of the babies, for a given weight of the mother, for smokers and non smokers respectively.
- The columns of the design matrix X can be obtained from those of the previous specification by using a linear combination. The model is the same but interpretation of parameters changes.
- one could not add the intercept otherwise X will become rank deficient

A model with interaction

The independent variables enter the model additively: the effect of the variable X is the same for each level of the factor F. In many case this is a too simplistic model, and the effect can change for different values of the factor.

This effect can be caught by introducing interaction between the covariates. The following regression model includes an interaction

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \beta_3 I_{(F_i=2)} x_i + \epsilon_i \ .$$

Interaction implies the relationship between x (weight of the mother) and y (weight of the baby) can be different for smokers and non smokers.

Inference in Linear models

Least square estimation

- A sample $y_1, y_2, \ldots, y_n$ along with the values of the vector $\boldsymbol{x_i}$ for the covariates can be used for estimating the parameters of the models $(\beta_0, \beta_1, \ldots, \beta_{p-1}, \sigma^2)$
- An estimator of the parameters vector $\boldsymbol{\beta}$ can be obtained by using the **least square method**:
  1. The least square estimator (LSE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the vector for which the following quantity is minimized

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \ ;$$

$$LS(\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X \boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}$$

  2. Taking the derivative $\frac{\partial LS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T \boldsymbol{y} + 2X^T X \boldsymbol{\beta}$ and then equating to 0, the LSE is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \ .$$

     To invert $(X^T X)$ we have to assume that this matrix is non singular. This is always true if X is a full rank matrix

Properties of LS estimator

The LSE $\hat{\beta}$ has the following properties:

1. $E(\hat{\beta}) = \beta$ and $var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$;
2. asymptotically $\hat{\beta} \sim N_p(\beta, \sigma^2 V^{-1})$ where $V = \lim_{n\to\infty} X_n^T X_n$ and X_n is the sequence of design matrices and V is positive definite; in most of the cases then for large n, $\hat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$. This allows us to test significance of parameters and to build confidence intervals easily.
3. $\hat{\beta}$, the LSE is the *best estimator* in the sense that it has minimum variance among all linear estimators (BLUE Best Linear Umbiesed Estimator- Gauss-Markov theorem).
4. When $(X^T X)$ is not singular but its determinant is very close to 0 then estimates are very unstable. This happens if (multiple) correlation among the column of X is very close to 1. Regularization could be a solution.

ML estimation

- When normality is assumed for the random components and taking account of uncorrelation (also indipendence in this case) then β can be obtained by using maximum likelihood estimation:
- Choose the value
  β which minimizes

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta))$$

  it is easy to show (by evaluating the log-likelihood and by deriving with respect to β) that the likelihood is maximized if the following quantity is minimized $(\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta)$.
- **under normality assumption, MLE and LSE are equivalent**
- But to the properties already listed for the LSE now it can be added the one regarding the (exact) distribution of $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

Estimation of σ^2

- To estimate σ^2 the following estimator is usually considered

$$
\begin{aligned}
S^2 &= \frac{SSE}{n-p} \quad \text{where} \\
SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i \\
&= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\
&= (\boldsymbol{y} - X\hat{\beta})^T(\boldsymbol{y} - X\hat{\beta})
\end{aligned}
$$

  is the *Sum of Squares of residuals* e_i.
- Note also that, when normality holds

$$\hat{\beta} \sim \mathrm{N}_p(\beta, \sigma^2(X^TX)^{-1}) \quad \text{and} \quad \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}\ ,$$

  and the two estimator $\hat{\beta}$ and S^2 are independent.
- note that for small samples if σ^2 is unknown and S^2 is used, tests and confidence intervals for a single β_j are based on student t distribution with $n-p$ df.

Model validation and model selection

Testing a general linear hypotesis

The test of more common interest are:

- test of significance for a single element of $\boldsymbol{\beta}$
  $H_0 : \beta_j = 0 \quad$ against $\quad H_0 : \beta_j \neq 0$
- test on a subvector $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)$
  $H_0 : \boldsymbol{\beta}_1 = 0 \quad$ against $\quad H_0 : \boldsymbol{\beta}_1 \neq 0$
- test of equality
  $H_0 : \beta_j - \beta_r = 0 \quad$ against $\quad H_0 : \beta_j - \beta_r \neq 0$

All these hypoteses are special case of the *general linear hypotesis*
$H_0 : \boldsymbol{C}\beta = \boldsymbol{d} \quad$ against $\quad H_0 \boldsymbol{C}\beta \neq \boldsymbol{d}$
$\boldsymbol{C}$ is and $r \times p$ matrix with rank$= r \leq p$

If data are fit to the model under the restriction $\boldsymbol{C}\beta = \boldsymbol{d}$ the residuals of this model are ${}_{H_0}e_i$ and one can compute $SSE_{H_0}{=}\sum_i^n {}_{H_0}e_i^2$ and calculate the statistic

$$\frac{n-p}{r}\,\frac{SSE_{H_0} - SSE}{SSE}$$

that, when H_0 is true, under normality assumption is a $F_{r,n-p}$ random variable.

Test significance of a single coefficient

When testing significance for a single element of $\boldsymbol{\beta}$

$H_0 : \beta_j = 0$ against $H_0 : \beta_j \neq 0$

applying, general linear hypotesis, it can be shown that

$$\frac{\hat{\beta}_j^{\,2}}{\widehat{Var(\hat{\beta}_j)}} \sim F_{1,n-p}$$

This is the square of a t with $n-p$ df and equivalently the following test statisticis is usually considered

$$t_j = \frac{\hat{\beta}_j}{\widehat{Var(\hat{\beta}_j)}^{1/2}}$$

This result can be also used to obtain for *beta* the following confidence interval at level $1-\alpha$

$\beta_j \pm t_{n-p,1-\alpha/2}(Var(\hat{\beta}_j))^{1/2}$

Decomposition of Sum of Squares

- the following holds:

$$SST = SSR + SSE \; ; \tag{4}$$

  the Total Sum of Squares (*total deviance*) is the sum of the Regression Sum of Squares (*deviance explained by the model*) and the Residual Sum of squares (*deviance of the residuals*). Studying the components of (4) is of great relevance, the ratio between SSR e SST is clearly related to the quality of the model.
- Let $\mathcal{F}_1$ be the minimal model (the one which contains only the intercept, $p = 1$).
  Let $\mathcal{F}_p$ be the current model with p parameters and let $\mathcal{F}_{p_0}$ be a reduced model with $1 < p_0 < p$ nested in $\mathcal{F}_p$.
  Then the variance explained by the current model $\mathcal{F}_p$ can be partitioned as it is shown in the table that follows (Table 1) that is called Analysis of variance table.

The analysis of variance

Table 1: Analysis of Variance (Anova)

Source of variability	df	SS	testing models improvement
total	n	SST	
constant	1	$n\bar{Y}^2$	
total	$n-1$	SST_{cor}	
improvement with $\mathcal{F}_{p_o}$ with respect to $\mathcal{F}_1$	$p_o - 1$	SSR_{p_o}	$\frac{SSR_{p_o}/(p_o-1)}{SSE_{p_o}/(n-p_o)} \sim F_{p_o-1,n-p_o}$
improvement with $\mathcal{F}_p$ with respect to $\mathcal{F}_{p_o}$	$p - p_o$	$SSR_p - SSR_{p_o}$	$\frac{(SSR_p-SSR_{p_o})/(p-p_o)}{SSE_p/(n-p)} \sim F_{p-p_o,n-p}$
residuals $\mathcal{F}_p$	$n-p$	SSE_p	

- The fall in the fit from $\mathcal{F}_{p_0}$ to $\mathcal{F}_p$ can be evaluated using the statistic

$$F = \frac{(SQE_{p_0} - SQE_p)/(p-p_0)}{SQE_p/(n-p)} \sim F_{p-p_0,n-p} \ .$$

Coefficient of determination R^2

The coefficient of determination R^2 is defined as the proportion of of total variance explained by the regression model.
It can be used as a godness–of–fit measure for the models

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} = 1 - \frac{SSE}{SST}$$

and $0 \leq R^2 \leq 1$.
This decomposition is possible if the model includes the intercept.
For nested models R^2 always increases adding covariates.

When comparing nested models the *corrected coefficient of determination* $\bar{R}^2$ is instead used

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Since it penalizes inclusion of new variables that are non significant.

Residuals

- The mean of $\boldsymbol{y}$ can be predicted once the model is estimated by $\hat{\mu}_i = \hat{\boldsymbol{y}} = X\hat{\beta}$ and consequently $\hat{\boldsymbol{y}} = X(X^TX)^{-1}X^T\boldsymbol{y} = H\boldsymbol{y}$
- $H = X(X^TX)^{-1}X^T$ is a square matrix of size n and it is called the *hat matrix* or the projection matrix. It has the following properties:
  1. it is symmetric and idempotent
  2. rank(H)=trace(H)=p
  3. h_{ii} have values that range from $1/n$ and 1
  4. the matrix $I - H$ is also symmetrical and idempotent with rank equal to $(n - p)$
- The residuals of the model are then $\boldsymbol{e}_i = (I - H)\boldsymbol{y}$
- under normality assumption $\boldsymbol{e}_i \sim N(\boldsymbol{0}, \sigma^2(I - H))$

Analysis of the residuals

- The quality of the model and the validity of the assumptions can be judged by using some diagnostic tools that mainly rely upon analysis of residuals as defined above.
- In the linear models the residuals can be standardized (to take into account the fact that they have unequal variance)

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{S^2(1 - h_{ii})}} \;, \tag{5}$$

  where h_{ii} is the i-th element on the diagonal of $H = X(X^T X)^{-1} X^T$. These values are called leverages.
- to identify which values are outliers with respect the majority of the data points the (externally) studentized residual are introduced $e_i^* = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}}$ where $S_{(i)}^2$ is the variance of the residuals when the i-th observation is excluded. For a model correctly specified e_i^* follows a t with $n - p - 1$ df.

Analysis of the residuals

- Cook distances are defined as $\frac{1}{p} r_i^2 \frac{h_{ii}}{1-h-ii}$ observations that, when excluded from the analysis, will cause substantial modifications in the values of the parameters estimated have higher values of Cook distances.
- Classical graphical tools based on residuals are:
  - plot of residuals against the predicted values
  - plot of residuals against the explanatory variables
  - plot of residuals against variables not in the model (added variable plot)
  - Q-Q norm of residuals,
  - plot of leverages h_{ii} and of Cook distances
- formal test of normality, such as Shapiro-Wilks test or Jarque-Bera test (the latter based on estimated values of third and fourth standardized moments)

Dealing with non constant variance and residual correlation

- Residual analysis can reveal that some assumptions could be questionable. Critical assumptions are those on uncorrelation and heteroscedasticity of residuals
- The assumptions $Cov(\epsilon) = \sigma^2 I$ should be replaced by a more general $Cov(\epsilon) = \sigma^2 W^{-1}$ where W is assumed to be simply a positive definite matrix
- In this case LSE is still an unbiased estimator but the estimate of its variance covariance matrix is biased.
- If we ignore this the main consequence are that tests or confidence intervals based on assumption of uncorrelation and homoscedasticity assumption lead to wrong conclusions (tests tends to say a parameter is significant too often and confidence intervals appear shorter)

Heteroscedasticity

- We will consider here only the case of heteroscedasticity. In this case $Cov(\epsilon) = \sigma^2 W^{-1}$ with $W^{-1} = diag(1/w_1, 1/w_2, \ldots, 1/w_n)$
- If each ϵ_i is multiplied by $\sqrt{w_i}$ one obtains the transformed values $\epsilon_i^* = \sqrt{w_i}\epsilon_i$ which have constant variance
- $var(\epsilon_i^*) = var(\sqrt{w_i}\epsilon_i) = \sigma^2$ and then the random components are homoscedastic.
- The model does not change if we transform also the response variable and the covariates (including the intercept) accordingly.
- We then obtain

$y_i^* = \sqrt{w_i} y_i$ and

$x_{ij}^* = \sqrt{w_i} x_{ij}$

for each of the p covariates (including the intercept) and then the model

$$y_i = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_{p-1} x_{i(p-1)}^* + \epsilon_i^*$$

is homoscedastic and the same assumptions of a standard LM hold. - This transformations, in matrix notation, are equivalent to pre-multiply all component of the model $\boldsymbol{y} = X\beta + \epsilon$ by the matrix $W^{1/2}$

Weigthed Least Squares (WLS)

- If then

$\boldsymbol{y}^* = W^{1/2}\boldsymbol{y}$, $\boldsymbol{X}^* = W^{1/2}\boldsymbol{X}$ and $\boldsymbol{\epsilon}^* = W^{1/2}\boldsymbol{\epsilon}$

we get a new model for transformed data where homoscedasticy holds and parameters can again be estimated by LSE obtaining

$$
\begin{aligned}
\hat{\beta} &= (X^{*T}X^*)^{-1}X^{*T}\boldsymbol{y}^* \\
&= (X^T W^{1/2} W^{1/2} X)^{-1} X^T W^{1/2} W^{1/2}\boldsymbol{y} \\
&= (X^T W X)^{-1} X^T W \boldsymbol{y}
\end{aligned}
$$

- This estimator is the Weighted Least Square Estimator (WLS)
- Note that weights are inversely proportional to variances of ϵ (which are originally heteroscedastic)
- to units with a more erratic random component are given smaller weights.
- Application of this strategy requires that the weights are known.

Model choice and variable selection

In many applications a large number of candidate predictors are available.
A naive approaches often used is the following: Estimate the most complex model that

includes all the covariates (and possibly all the interactions). Then , remove all insignificant variables from the model

This strategy is not advisable for many reasons. Let us list some of them:

- the resulting model can overfit the data and then its predictive performance for new data can decrease
- the larger the number of covariate the higher the risk of multi-collinearity (correlated regressors)
- There are many models with equivalent performances but different substantial interpretation. You are not sure that the variable which remains in you model after such backward selection strategy should be really considered the most relevant.

Other naive strategy for variable selection are:

- All subset selection (chose the best among $\sum_j^p \binom{p}{j}$ possible models)
- Forward selection
- Stepwise selection

Model choice criteria

Since one of the principles to consider when building a model is the *Occam's razor*, criteria to select the model that has good performances and at the same time is less complex should be introduced.

when considering alternative LM we have already seen some criteria

- R^2 and corrected R^2
- F test (for nested models)
- Mallows's C_p

$$C_p = \frac{\sum_i^n (y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + M$$

  where M is the number of covariates in the model and $\hat{y}_{iM}$ are the predicted values with those M covariates. The "best" model is the one with lowest C_p.
- Akaike Information Criteria (AIC)

$$AIC = -2l(\hat{\beta}_M, \hat{\sigma}^2) + 2(M+1)$$

  Better fit corresponds to smaller smaller AIC values.
  For a linear model with gaussian components and p β_j parameters
  $AIC = n\log(\hat{\sigma}^2) + 2(p+1)$.
  Note that $\hat{\sigma}^2$ is SSE divided by n.
- Bayesian Information Criteria (BIC)

$$AIC = -2l(\hat{\beta}_M, \hat{\sigma}^2) + \log(n)(M+1)$$

Avoiding collinearity

A diagnosis of collinearity is obtained by computing the *variance inflation factor* (VIF) associated to the *j*-th predictor

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination when is regressed on all the remaining covariates. (VIF_J>10$ is usually taken as a symptom that the variable can cause collinearity).

Typical solutions are:

- omission of covariates
- using principal components extracted from regressors (or other combination of the regressors)
- Ridge regression: it is an alternative to LSE and

  $$\hat{\beta} = (X^T X + \lambda \boldsymbol{I})^{-1} X^T \boldsymbol{y}$$

  and λ is a chosen tuning parameter

Regularization Techniques

It has been assumed so far that X has full rank, and this gives a unique solution to equations

$$(X^T X)\beta = X^T \boldsymbol{y}$$

We have already discussed that $X^T X$ could be close to singularity. Regularization consists in changing the objective function by penalizing it:

$$PLS(\beta) = (\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta) + \lambda \text{pen}(\beta)$$

where $\text{pen}(\beta)$ is a term that measure the complexity of the model and $\lambda \geq 0$ is a smoothing paarmeter that reflect the weight given to the penalty. Penalty can be large when many β are large or can panalize a regerssion vector with too many non zero coefficients.

Ridge regression is an example of penalized least square. It corresponds to introducing the following penalty

$$\text{pen}(\beta) = \sum_j^p \beta_j^2 = \beta^T \beta$$

With large λ the penalty term dominates and all coefficient are shrunk to 0.

LASSO (Least Absolute Shrinkage and Selection Operator)

Also LASSO corresponds to a penalized least square criterion and

$$\hat{\beta} = \arg\min(\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta) + \lambda \sum_{j}^{p} |\beta_j|$$

The penalization chosen with LASSO tend to shrink some of the values of the coefficients to 0.
Note that no closed explicit solution of the minimization problem exists. Numerical optimization must be used.