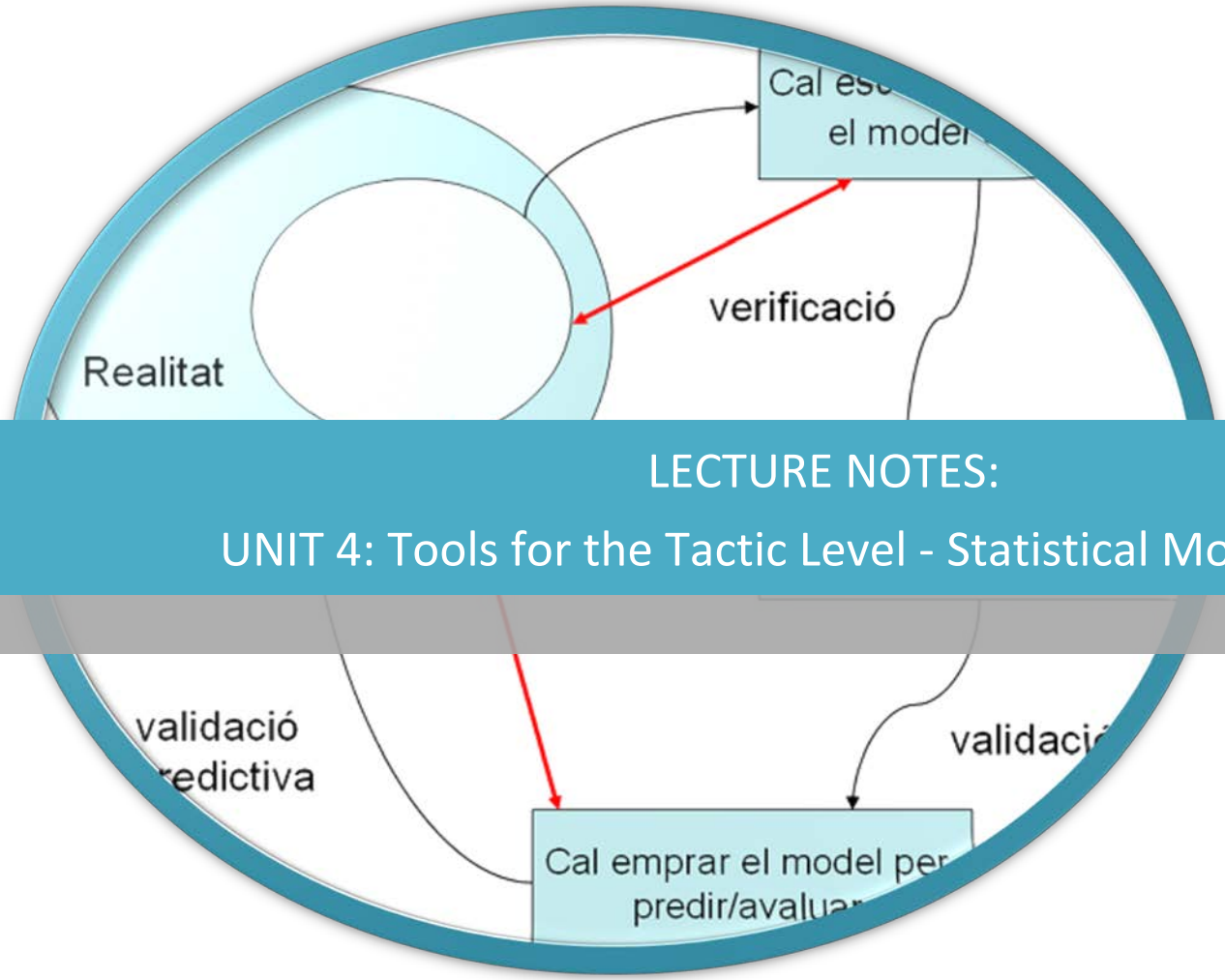
Course 2014-15

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa



FIB-Degree in Informatics Engineering

LECTURE NOTES:

UNIT 4: Tools for the Tactic Level - Statistical Modelling (I)

CONTENTS

4-1 READINGS

Basic References:

- Fox, J. *Applied Regression Analysis and Generalized Linear Models.* Sage Publications, 2nd Edition 2008.
- James H. Stock and Mark W. Watson, ***Introduction to Econometrics,*** Prentice Hall, 2007
- Seber, G.A.F.: *Linear Regression Analysis*. Wiley, 1977.
- Fox and Weisberg ***An R Companion to Applied Regression.*** Sage Publications, 2nd Edition 2010.
- Peña, D.: *Estadística. Modelos y métodos. Vol. 2, Modelos lineales y series temporales.* Alianza Universidad Textos, 1989.

4-2 INTRODUCTION TO LINEAR MODELS

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a vector of n *observations* considered a draw of the random vector $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, whose variables are statistically independent and distributed with expectation $\mu^T = (\mu_1, \ldots, \mu_n)$:

- In linear models, the **Random Component** distribution for $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ is assumed normal with constant variance σ^2 and expectation $\mathrm{E}[\mathbf{Y}] = \mu$.
- So, the *response variable* is modeled as normal distributed, thus negative or positive values, arbitrary small or large can be encountered as data for the response and prediction.
- **The Systematic component** of the model consists on specifying a vector called the linear predictor, notated η, of the same length as the response, dimension *n*, obtained from the linear combination of regressors (explicative variables) . In vector notation, parameters are $\beta^T = (\beta_1, \ldots, \beta_p)$ and regressors are $\mathbf{X}^T = (X_1, \ldots, X_p)$ and thus $\eta = \mathbf{X}\beta$ where η is *nx1,* $\mathbf{X}$ is *nxp* and β is *px1.*
- Vector μ is directly the linear predictor η, thus the **link function** is $\eta = \mu$.

4-2 INTRODUCTION TO LINEAR MODELS

Empirical problem: What do data say about class sizes and test scores according in The California Test Score Data Set

All K-6 and K-8 California school districts (n = 420)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

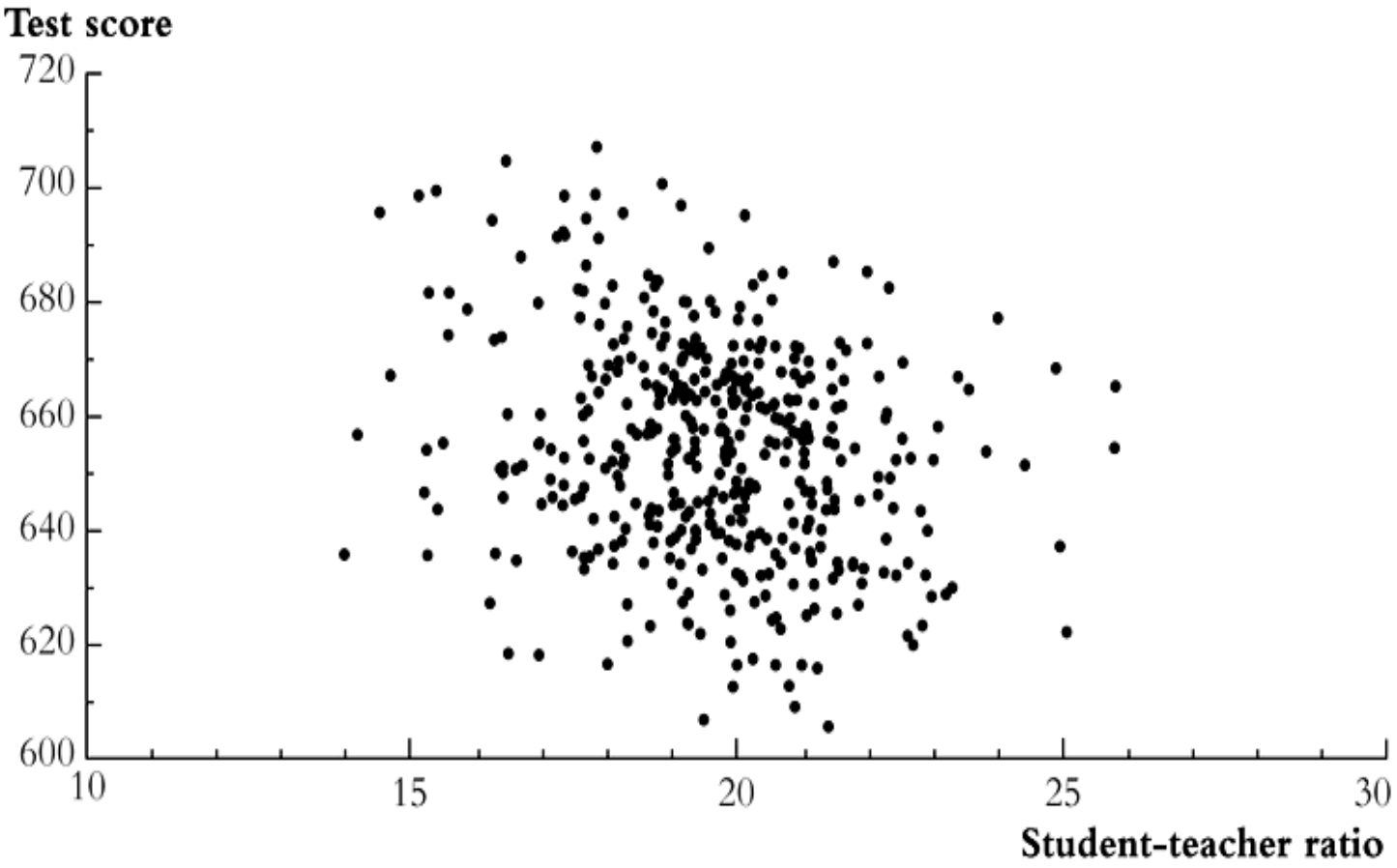
Policy question: What is the effect of reducing class size by one student per class? by 8 students/class? Do districts with smaller classes (lower STR) have higher test scores?

An initial look at the California test score data:

	Average	Standard Deviation	Percentile						
			10%	25%	40%	50% (median)	60%	75%	90%
Student-teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

4-2 INTRODUCTION TO LINEAR MODELS

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.



Stock and Watson (2007)

4-2 SOME NOTATION AND TERMINOLOGY

- The population regression line is

  Test Score = $\beta_1 + \beta_2 STR$

  β_1 is the intercept

  β_2 is the slope of population regression line

  $= \dfrac{\Delta \text{Test score}}{\Delta STR}$ = change in test score for a unit change in STR

- Why are β_1 and β_2 "population" parameters?
- We would like to know the population value of β_2.
- We don't know β_2, so must estimate it using data.
- How can we **estimate** β_1 and β_2 from data?

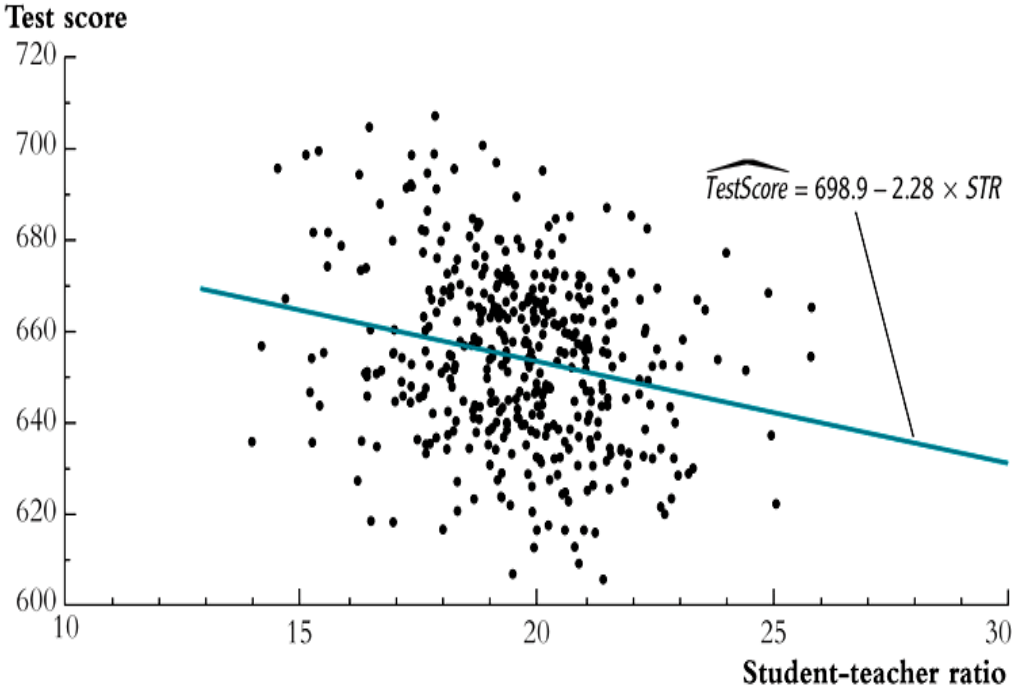
We will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters β_1 and β_2, which solves,

$$Min_{\boldsymbol{\beta}=(\beta_1,\beta_2)} S(\boldsymbol{\beta}) = (\mathrm{Y} - \mathrm{X}\beta)^T \cdot (\mathrm{Y} - \mathrm{X}\beta) = \sum_k \left(Y_k - \beta_1 - \beta_2 x_k\right)^2$$

4-2 SOME NOTATION AND TERMINOLOGY: EXAMPLE

The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.

Application to the California Test Score – Class Size data



Estimated slope = = – 2.28

Estimated intercept = = 698.9

Estimated regression line: = 698.9 – 2.28 STR

Interpretation of the estimated slope and intercept:

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. *It makes no sense, since it extrapolates outside the range of the data.*

4-2 SOME NOTATION AND TERMINOLOGY: EXAMPLE

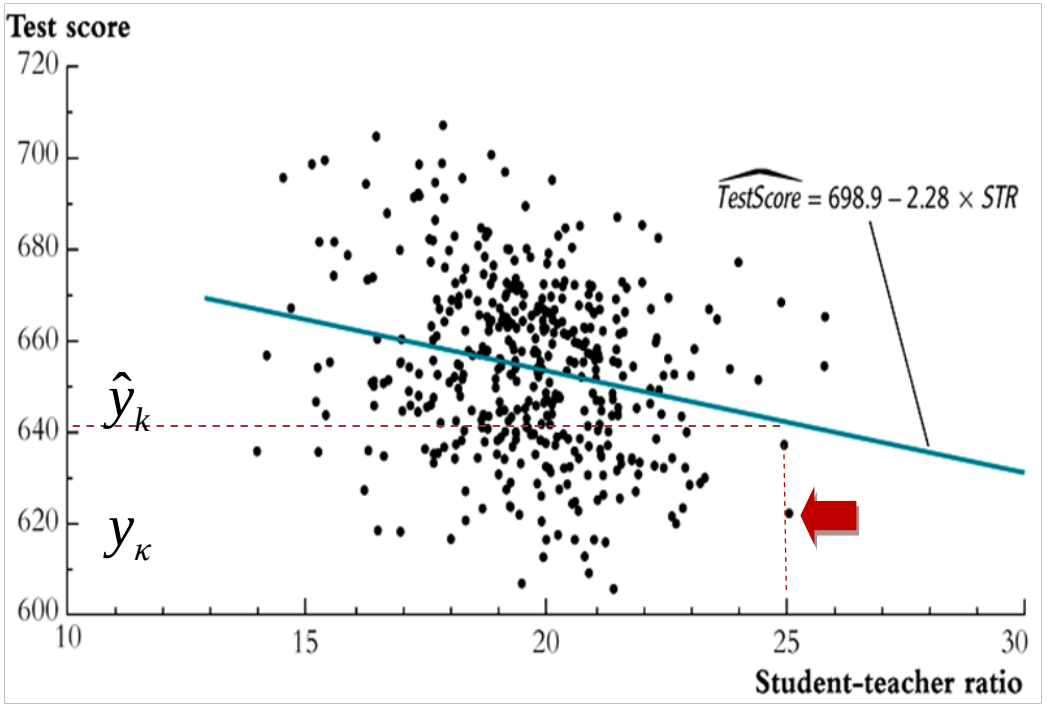
Predicted values & residuals:

One of the districts in the data set for which STR = 25 and Test Score = 621

predicted value: = 698.9 – 2.28*25 = 641.9

residual: = 621 – 641.9 = -20.9

The OLS regression line is an estimate, computed using our sample of data; a different sample would have given a different value of $\hat{\beta}_2$.



4-2 SOME NOTATION AND TERMINOLOGY

How can we:

- quantify the sampling uncertainty associated with $\hat{\beta}_2$?
- use $\hat{\beta}_2$ to test hypotheses such as $\beta_2 = 0$?
- construct a confidence interval for $\hat{\beta}_2$?

We are going to proceed in four steps:

- The probability framework for linear regression
- Estimation
- Hypothesis Testing
- Confidence intervals

A vector-matrix notation for regression elements will be considered since it simplifies the mathematical framework when dealing with several explicative variables (regressors) .

4-2 INTRODUCTION TO LINEAR MODELS

Classification of statistical tools for analysis and modeling

Explicative Variables	Response Variable				
	Dicothomic or Binary	*Polythomic*	*Counts (discrete)*	*Continuous*	
				Normal	*Time between events*
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polythomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival Analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival Analysis
Factors and covariates	Logistic regression	*	Log-linear models	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

4-3 LEAST SQUARES ESTIMATION IN MULTIPLE REGRESSION

Assume a linear model without any distribution hypothesis,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is *nx1,* $\mathbf{X}$ is the design matrix *nxp* and $\boldsymbol{\beta}$ is the vector parameters *px1* **Let Y be a** numeric response variable, and $\mathcal{E}$ be the model error

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} == \begin{pmatrix} \overset{1}{\mathbf{1}} & \overbrace{x_{12}}^{j=2} & \overbrace{x_{13}}^{j=3} & \overbrace{\cdots}^{\cdots} & \overbrace{x_{1p}}^{j=p} \\ \mathbf{1} & x_{12} & x_{13} & \cdots & x_{1p} \\ \mathbf{1} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & x_{n-1\ldots 2} & x_{n-13} & \cdots & x_{n-1\,p} \\ \mathbf{1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

The Ordinary Least Squares estimation of the model parameters $\boldsymbol{\beta}$ can be written in the general case as,

$$Min_{\beta}\, S(\beta) = (\mathrm{Y} - \mathrm{X}\beta)^T \cdot (\mathrm{Y} - \mathrm{X}\beta) = \sum_{i=1\ldots n} \left(Y_k - \mathrm{x}_k^T \beta\right)^2 = \mathrm{Y}^T\mathrm{Y} + \beta^T\, \mathrm{X}^T\mathrm{X}\,\beta - 2\beta^T\, \mathrm{X}^T\mathrm{Y}$$

The OLS estimator minimizes the average squared difference between the actual values of Yi and the prediction (predicted value) based on the estimated line. The matrix-vector notation is used along these notes.

4-3 LEAST SQUARES ESTIMATION

- First order condition for a mínimum of $S(\beta)$ are:

$$\nabla_\beta S(\beta) = 0 \leftrightarrow \frac{\partial S(\beta)}{\partial \beta_j} = 0 \quad j = 1, \ldots, p$$

Once derivatives are computed, the so-called **normal equations** appear,

$$\nabla_\beta S(\beta) = 0 \leftrightarrow \frac{\partial S(\beta)}{\partial \beta} = 2\,\mathrm{X}^T\mathrm{X}\,\beta - 2\,\mathrm{X}^T\mathrm{Y} = 0 \rightarrow \mathrm{b} = \hat{\beta} = \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{Y}$$

- The solution of the normal equations is the least square estimator $\hat{\beta}$ of the β parameters.
- If the design matrix is not singular, i.e, column rang of $\mathbf{X}$ is *p*, then $\hat{\beta}$ the solution to normal equations is unique.
- Perfect **multicollinearity** is when one of the regressors is an exact linear function of the other regressors.
- If $\mathbf{X}$ is not full-rang and this happens where there is perfect multicollinearity, then infinite solutions to normal equations exist, but all of them give the same vector of prediction values $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

4-3 LEAST SQUARES ESTIMATION

The rang of $\mathbf{X}^T\mathbf{X}$ is equal to the rang of $\mathbf{X}$. This attribute leads to two criteria that must be met in order to ensure $\mathbf{X}^T\mathbf{X}$ is non singular and thus obtain a unique solution:

1. At least as many observations as there are coefficients in the model are needed.
2. The columns of X must not be perfectly linearly related, but even near collinearity can cause statistical problems.

4-3.1 Geometric properties

Let $\Re(\mathbf{X})$ be the linear variation expanded by the columns of $\mathbf{X}$,

$$\Re(\mathrm{X}) = \left\{ \boldsymbol{\mu} \,\middle|\, \boldsymbol{\mu} = \mathrm{X}\boldsymbol{\beta} \quad \boldsymbol{\beta} \in \Re^p \right\} \subset \Re^n .$$

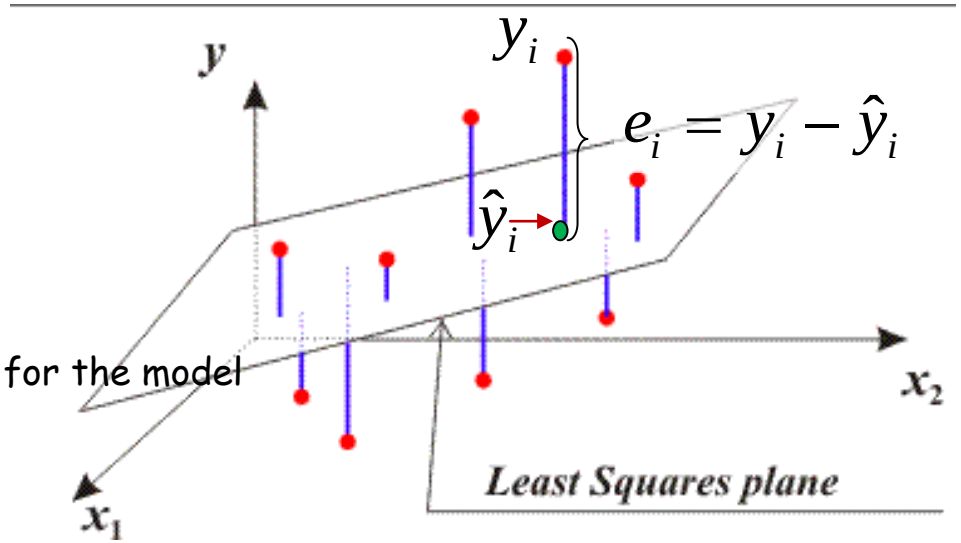
And $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ the predictions once computed the least squared estimator of model parameters,

4-3 LEAST SQUARES ESTIMATION

- Then it can be shown that $\hat{\boldsymbol{\mu}}$ is the ortogonal projection of **Y** and it is unique being the projection operator defined as, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}$, and called the hat matrix since applied to **Y** provide the fitted values or predictions tipically noted as $\hat{\mathbf{Y}}$,

$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}$ puts a ^ to **Y**: $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}\mathbf{Y}$.

Graphically,



Properties of the Hat Matrix:

- It depends solely on the predictor variables X
- It is square, symmetric and idempotent: HH = H
- Finally, the trace of H is the degrees of freedom for the model

4-4 LEAST SQUARES ESTIMATION: INFERENCE

4-4.1 Basic Inference Properties

- A continuous response linear model is assummed where,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{or} \quad y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1,\ldots,n$$

where $\mathbf{Y}$, μ are *nx1*, $\mathbf{X}$ *nxp* of rang *p* and β *px1* and the conditional distribution of unbiased errors are i.i.d. of constant variance and normal $\varepsilon \mid X \approx \mathrm{N}\left(0, \sigma^2\right)$ - equivalent to $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n\right)$.

- …Then, the **minimum variance unbiased estimator** of β is $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, is the ordinary least square estimator and is equal to the maximum likelihood estimator $\hat{\beta}_{MV}$.
- If normal assumption is not met then OLS estimators are not efficient (do not have mínimum variance as ML estimator).

- If the hypothesis hold then the unbiased estimator of σ^2, noted s^2 is efficient (mínimum variance),

$$s^2 = \frac{\mathbf{e}^{\mathrm{T}} \cdot \mathbf{e}}{n-p} = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \cdot \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p} = \frac{RSS}{n-p}$$

Residual Sum of Squares

4-4 LEAST SQUARES ESTIMATION: INFERENCE

What we knew before presenting the statistical properties for OLS was about a sample (in particular, our sample) but now we know something about a larger set of observations from which this sample is drawn.

We consider the statistical properties of $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, an estimator of something, specifically the population parameter vector $\boldsymbol{\beta}$ and the former theorem says to us that OLS estimator is linear and unbiased. It is the best, the most efficient (with smaller variance) than any other estimator and has a normal sampling distribution.

Any individual coefficient $\hat{\beta}_j$ is distributed normally with expectation β_j and sampling variance $\mathbf{V}\left(\hat{\beta}_\mathbf{j}\right) = \sigma^2\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^T_{jj}$ and we can test the simple hypothesis (i.e., *make some inference*):

$$H_0: \quad \beta_j = \beta_j^0 \text{ with } Z_0 = \frac{\hat{\beta}_j - \beta_j^0}{\sigma\sqrt{\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^T_{jj}}} \approx N(0,1)$$

But since it does not help so much since β_j and σ^2 are unknown an unbiased estimator of σ^2 is proposed based on the standard error of regression s^2 and to estimate the sample variance of $\hat{\beta}_j$, $\hat{\mathbf{V}}\left(\hat{\beta}_\mathbf{j}\right)$.

4-4 LEAST SQUARES ESTIMATION: INFERENCE

$\hat{\mathbf{V}}\left(\hat{\beta}_{\mathbf{j}}\right)= s^2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^T_{jj} \rightarrow SE\left(\hat{\beta}_{\mathbf{j}}\right)= s\sqrt{\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^T_{jj}}$ and the ratio of $\hat{\beta}_j$ and $SE\left(\hat{\beta}_{\mathbf{j}}\right)$ is distributed as a Student t with n-p degrees of freedom (since $\hat{\beta}_j$ and s^2 are independent)

$t_0 = \dfrac{\beta_j - \beta_j^0}{s\sqrt{\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^T_{jj}}} \approx Student\ t_{n-p}$ can be defined and thus $P\left(H_0\right)= P\left(t_{n-p} > t_0\right)$ computed or a bilateral confidence interval at $100\left(1-\alpha\right)\%$ for $\beta_j \in \hat{\beta}_j \pm t^{\alpha/2}_{n-p} SE\left(\hat{\beta}_j\right)$

Inference for Multiple Coefficient will be presented further by F-test

4-5 HYPOTHESIS TESTS IN MULTIPLE REGRESSION

➡ Let a statistical model for multiple regression for a given data set be,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$, where **Y** *nx1,* **X** *nxp* of rang *p* and $\boldsymbol{\beta}$ *px1*

Where errors are unbiased and iid normally distributed $\varepsilon \approx \mathrm{N}_n\left(0,\, \boldsymbol{\sigma}^2\mathbf{I}_n\right)$. **Ordinary least square estimators are noted as** $\hat{\beta}$.

➡ Let a simpler model of multiple regression for the same set of data be the former one with a set of linear restrictions applying to the model parameters (a **nested model**); i.e,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$, where **Y** *nx1,* **X** *nxp* of rang *p* and $\boldsymbol{\beta}$ *px1*

subject to linear constraints $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ that define a linear hypothesis to contrast (to make inference) that we call H, **A** is a qxp matrix of rang $q < p$. **Ordinary least square estimates subject to constraints are** $\hat{\boldsymbol{\beta}}_H$. Residual Sum of Squares is noted as RSS_H.

4-5 HYPOTHESIS TESTS IN MULTIPLE REGRESSION

- If the hypothesis H is true than it can be shown that,

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{\left(\mathbf{A}\hat{\beta} - \mathbf{c}\right)^T \left(\mathbf{A}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\hat{\beta} - \mathbf{c}\right)}{q\, s^2} \rightarrow F_{q,n-p}$$

Example: Suppose we have a model with a set of parameters

$$\boldsymbol{\beta}^{\mathbf{T}} = (\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4)$$

and we want to test the hypothesis $H: \begin{cases} \beta_1 + \beta_2 - 4\beta_4 = 2 \\ \beta_1 - \beta_2 = 0 \end{cases}$ then

$$\mathbf{A}\beta = \mathbf{c} \rightarrow \begin{bmatrix} 1 & 1 & 0 & -4 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

4-5 HYPOTHESIS TESTS IN MULTIPLE REGRESSION

4-5.1 Testing in R

Anova()

- For nested linear regression models, method anova(fullmodel, restrictedmodel) implements F-test. Hypothesis testing in this subject will be suggested by this method.

- Inconvenient: The two models have to be previously computed

Linear.hypothesis() in **car package**: following the previous generic example

```
library(car)
linearHypothesis(model,
    hypothesis.matrix=matrix(c(1,1,0,-4,1,-1,0,0),nrow=2,ncol=4,byrow=TRUE),
    rhs=as.vector(c(2,0)) )
```

For glm() object performs a Wald test

It requires the estimation of one model only (compared to anova() method)

4-5 HYPOTHESIS TESTS IN MULTIPLE REGRESSION

4-5.2 Confidence interval for model parameters

Individual confidence interval for β_i in OLS resumes:

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \approx t_{n-p} \quad \rightarrow \quad \hat{\beta}_i \pm t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_i} \quad \text{donde} \quad \hat{\sigma}_{\hat{\beta}_i} = s\sqrt{\left(X^T X\right)^{-1}_{ii}} \quad \text{y} \quad s = \hat{\sigma} = \sqrt{\frac{RSS}{n-p}}$$

$t_{n-p}^{\alpha/2}$ is the *t de Student* for bilateral confidence interval 1-α . Degrees of freedon are (*n-p*) and correspond to the standard error of regression.

Estimates of β_i parameters are statistically dependent and individual confidence intervals might give a wrong idea of the jointly distributed values.

There is a large literature about how to build confidence regions or perform simultaneous test of several hypothesis: it is out of the scope of this material, some particular suggestions will be presented in the practical sessions.

4-6 MULTIPLE CORRELATION COEFFICIENT

- ***Multiple correlation coefficient R,*** is a goodness of fit measured of a regression model defined as the Pearson correlation coefficient between fitted values $\hat{y}_k$ and observations y_k :

$$R = \frac{\sum_k \left(y_k - \bar{y}\right)\left(\hat{y}_k - \bar{\hat{y}}\right)}{\left\{\sum_k \left(y_k - \bar{y}\right)^2 \sum_k \left(\hat{y}_k - \bar{\hat{y}}\right)^2\right\}^{1/2}}$$

- The squared of the multiple correlation coefficient $\mathbf{R^2}$ is called the coefficient of determination. The multiple correlation coefficient generalizes the standard coefficient of correlation. It is used in multiple regression analysis to assess the quality of the prediction of the dependent variable. It corresponds to the squared correlation between the predicted and the actual values of the response variable.

1. According to the decomposition of the Total Sum of Squares (TSS) in the Residual Sum of Squares plus the Explained Sum of Squares for a given model (valid for models including an intercept), R-Squared can be rewritten.
2. TSS= $\sum_k \left(y_k - \bar{y}\right)^2$ where $\bar{y} = \frac{1}{n}\sum_k y_k$ is the mean of the observed response data.

3. ESS= $\sum_k \left(\hat{y}_k - \bar{y}\right)^2$ and RSS= $\sum_k \left(y_k - \hat{y}_k\right)^2$.

4-6 MULTIPLE CORRELATION COEFFICIENT …

4. **TSS=ESS+RSS**, $\sum_k (y_k - \bar{y})^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2$,

Proof:

$$\sum_k (y_k - \bar{y})^2 = \sum_k ((y_k - \hat{y}_k) + (\hat{y}_k - \bar{y}))^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2 + 2\sum_k (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) =$$
$$\sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2$$

where,

$$\sum_k (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) = \sum_k (y_k - \hat{y}_k)\hat{y}_k - \bar{y}\sum_k (y_k - \hat{y}_k) = \sum_k (y_k - \hat{y}_k)\hat{y}_k = (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathbf{T}} \hat{\mathbf{Y}} =$$
$$= (\mathbf{Y} - \mathbf{HY})^{\mathbf{T}} \mathbf{HY} = \mathbf{Y}^{\mathbf{T}} (\mathbf{I} - \mathbf{H}) \mathbf{HY} = \mathbf{0}$$

And:

➡ Or equivalently

$$RSS = (1 - R^2) TSS \, .$$

$$R^2 = \frac{\sum_k (\hat{y}_k - \bar{\hat{y}})^2}{\sum_k (y_k - \bar{y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

4-6 MULTIPLE CORRELATION COEFFICIENT …

4-6.1 Properties of the Multiple Correlation Coefficient

1. $|R| \leq 1$ and if $|R| = 1$ indicates a perfect linear relation between response data and regressors.
2. $100(1-R^2)$ represents the fraction of response data variability not explained by the current model.
3. $100R^2$ represents the fraction of response data variability explained by the current model.

4-6.2 R2-adjusted

- Since this correlation cannot go down when variables are added, an adjustment must be made to the R2 in order to increase only when really significative regressors are added to the model. This is the adjusted R-Squared:

$$R_a^2 = 1 - \frac{SCR/(n-p)}{SCT/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right)$$

  - Adjusted R-Square is always less to the original R-Square and might be negative.

4-7 GLOBAL TEST FOR REGRESSION. ANOVA TABLE

- The **global test of regression** is a particular case of a multiple contrast of hypothesis where all parameters related to explicative variables are tested to be simultaneously zero.

$$\mathbf{H}: \beta_2 = 0, \ldots, \beta_p = 0 .$$

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} = \frac{ESS/(p-1)}{TSS/(n-p)} = \frac{ESS}{(p-1)s^2} \approx F_{p-1,n-p,}$$

4-8 PREDICTIONS AND INFERENCE

- Let $\hat{Y}_k$ be the fitted value (prediction) for observation data k where the values for explicative variables are $\mathbf{x_k^T} = \begin{pmatrix} 1 & x_2 & \dots & x_p \end{pmatrix}$: $\hat{Y}_k = \mathbf{x_k^T}\hat{\boldsymbol{\beta}}$.

1. $\mathrm{E}\left[\hat{Y}_k\right] = \mathrm{E}\left[\mathbf{x_k^T}\hat{\boldsymbol{\beta}}\right] = \mathbf{x_k^T}\boldsymbol{\beta}$.
2. $V\left[\hat{Y}_k\right] = V\left[\mathbf{x_k^T}\hat{\beta}\right] = \mathbf{x_k^T}V\left[\hat{\beta}\right]\mathbf{x_k} = \sigma^2\mathbf{x_k^T}\left(\mathbf{X^TX}\right)^{-1}\mathbf{x_k} = \sigma^2 h_{kk}$, where h_{kk} is the k diagonal term of projection matrix appearing in OLS (H, hat matrix) whose diagonal elements range between *1/n* and *1*. The variance of the fitted value is minimum when the observation lies in the center of gravity of regressor values.
3. Fitted values are normally distributed and correlated.

- It is called the point and the variance of a mean prediction at $\mathbf{x_k}$ and the confidence interval should be computed based on *Student t with n-p df* using the standard error of regression *s*:

$$\frac{\hat{Y} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma_{\hat{Y}}} \approx N(0,1) \rightarrow \; t = \frac{\hat{Y} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}_{\hat{Y}}} \approx t_{n-p} \quad \text{donde} \quad \hat{\sigma}_{\hat{Y}} = \hat{\sigma}\sqrt{\mathbf{x}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}} = s\sqrt{\mathbf{x}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}}$$

And at a confidence level 100(1-α)% the *true mean value* lies in : $\boxed{\hat{Y} \pm t_{n-p}^{\alpha/2}\hat{\sigma}_{\hat{Y}}}$.

4-9 MODEL VALIDATION

- Residual analysis constitutes a practical tool for graphically assessing model fitting and satisfaction of optimal hypothesis for OLS estimates:

Let a model a continuous response be,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$, where $\mathbf{Y}$ *nx1,* $\mathbf{X}$ is *nxp* of column range *p* and $\boldsymbol{\beta}$ is *px1*

And errors are unbiased, uncorrelated and normally distributed with constant variance; i.e.: $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ or equivalently $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$.

- Residuals are the difference between observed response values and fitted values : $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}$.
- Do not confuse errors and residuals; residuals are the observed errors when the proposed model is correct.

4-9 MODEL VALITION: RESIDUAL ANALYSIS

In order to interpret residuals, some transformed residuals are defined:

1. Scaled residual c_i defined as i residual divided by the standard error of regression estimate for the model, s, $c_i = \frac{e_i}{s}$. It is not so bad when leverages show no big changes, since $V[e_i] = \sigma^2(1-h_{ii})$.

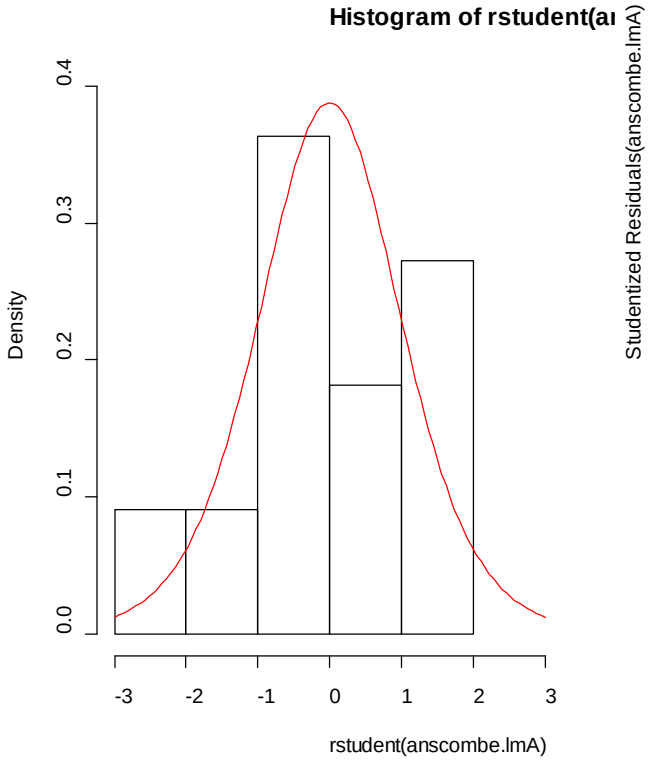
2. Standarized residual d_i is defined as the residual divided by its standard error: $d_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$.

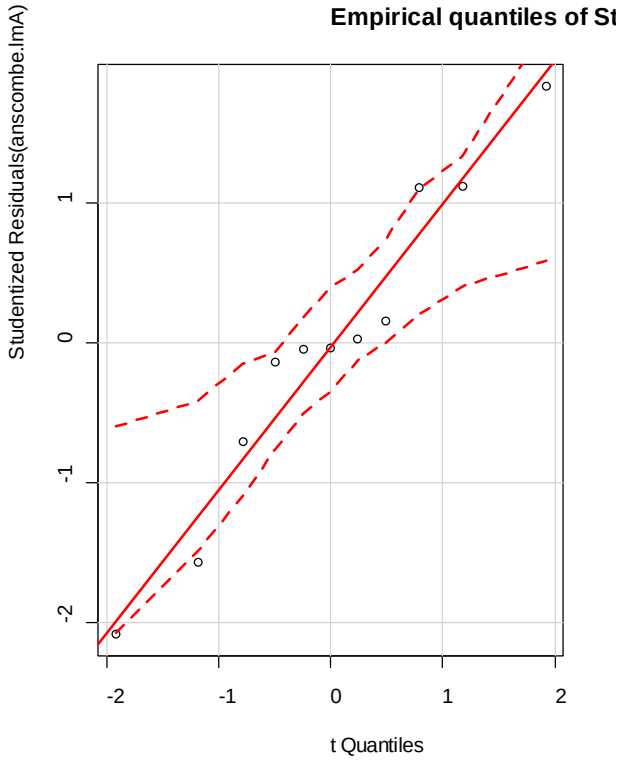
3. Studentized residual r_i is defined as $r_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$ where $s^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}$.

- Outliers for r_i can be detected using t.Student lower and upper bounds for sample size or by univariate descriptive graphical tools as a boxplot.

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

Residual analysis - usual plots:





1. Histogram residuals: a normal density is checked.

```
hist(rstudent(model), freq=F)
curve(dt(x, model $df),col=2,add=T)
```

2. Boxplot residuals to identify outliers of regression.

3. *Normal Probability Plot* or Quantile Plot of standardized or studentized residuals.

In R the confidence envelope for studentized residuals is shown loading car package and method,

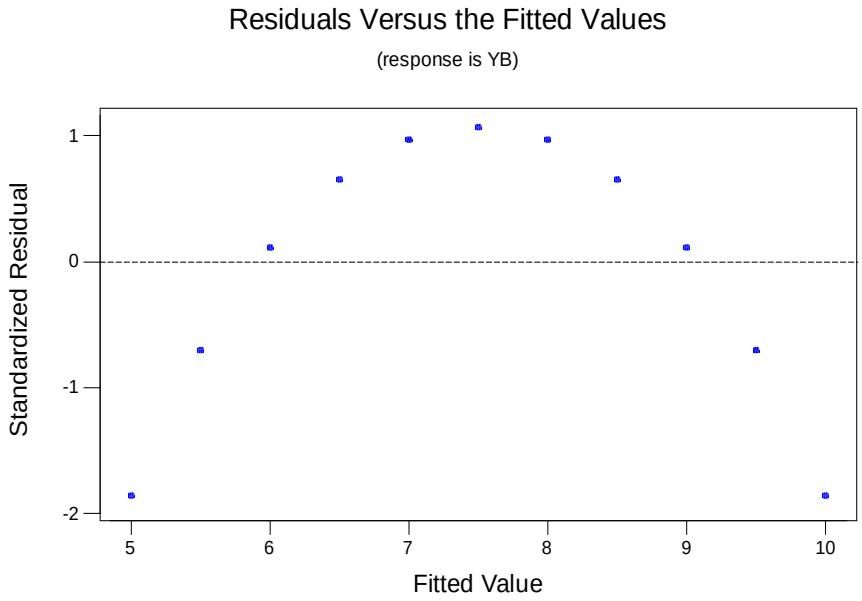
```
library(car)
qqPlot( model, simulate=T, labels=F)
```

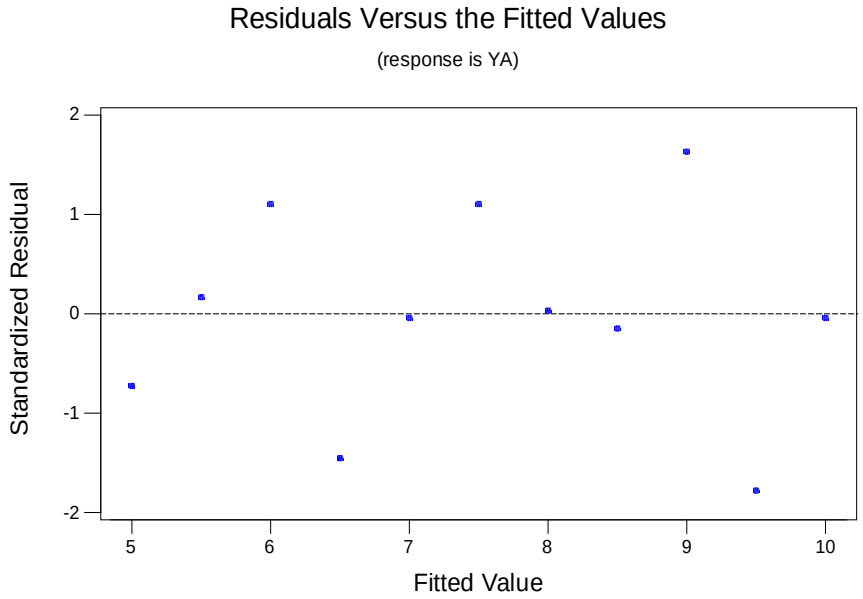
4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

- When normality assumption is not fully met than the constrast of hypothesis based on Student o F Test are approximate and estimates are not efficient.
- Tests for normality assessment can be used. Shapiro-Wilk is one of my favorites, but there are a lot of tests depending on sample size. Package nortest in R contains some common possibilities. Even on non normal errors, residuals tend to normality (due to Central Limit Theorem) on medium and large samples.
- Residuals are correlated with observations Y_i, but not with fitted values $\hat{Y}_i$, then scatterplots with fitted values on X axis are suggested.

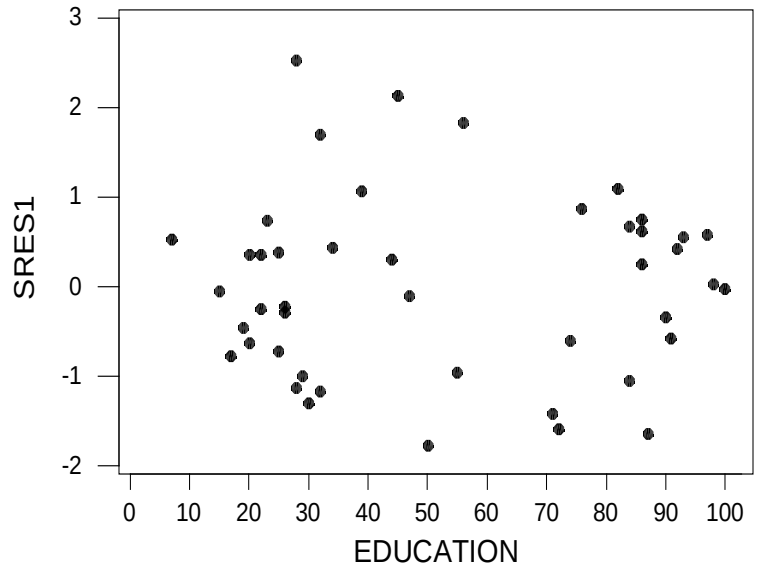
4. Scatterplots: $\boxed{e_i \ vs \ \hat{Y}_i}$, or $\boxed{d_i \ vs \ \hat{Y}_i}$ o $\boxed{r_i \ vs \ \hat{Y}_i}$. Failing of linearity can be detected (transformations or addition of regressors might be needed) or heterocedasticity (transformations required). Unusual observations might difficult the interpretation.

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS





4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS



5. ***Scatterplots of Residuals vs each regressor*** (except intercept).

Horizontal band indicates linearity satisfaction and homocedasticity.

Homocedasticity Breusch-Pagan test in package lmtest might be of interest.

```
> library(lmtest)
> bptest(model)
```

6. Residual vs time/order or any omitted variable in the model suspected to affect hypothesis

Such as, autocorrelation function for residuals (method in R, acf(rstudent(model)) of order *k*,

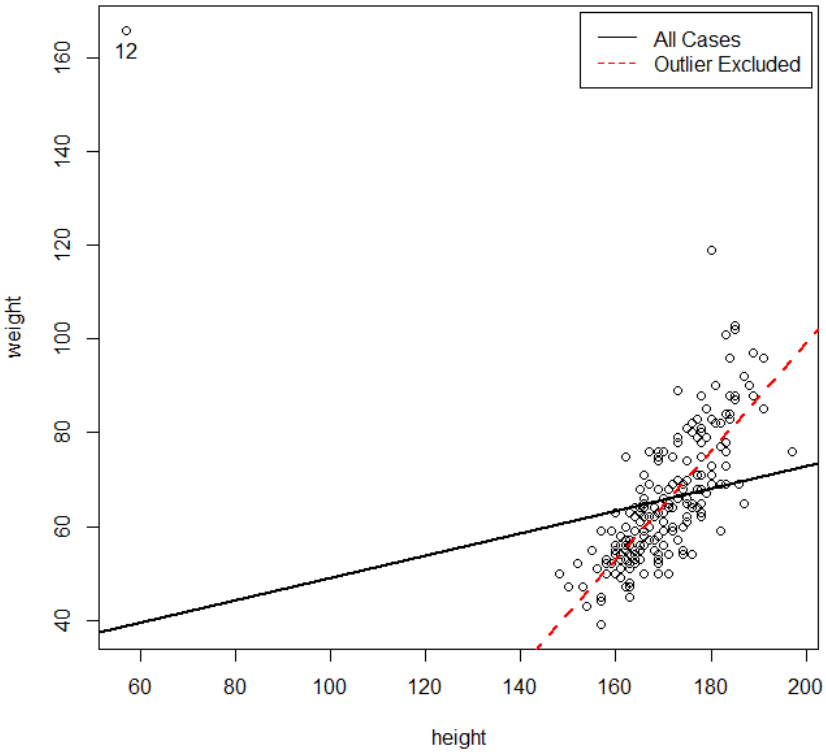
$$r(k)=\frac{\sum_i e_i e_{i+k}}{\sum_i e_k^2}$$ and contrats for simultaneous hypothesis r(j)=0 for j>k.

Or Durbin-Watson test (tables very difficult to interpret) also for autocorrelation testing.

4-10 MODEL VALIDATION: UNUSUAL AND INFLUENTIAL DATA

It is easy to find examples for regression analysis that show the existence of a few observations that strongly affect the estimation of model parameters in OLS: one percent of data might have a weight in parameter estimation greater than 99% data.

- Influential data affects model prediction and it seems natural that predicted values should be supported by 99% of data and not seriously affected by the 1% left.
- Classifying an observation as ***a priori influential*** is related to robustness of the design of data collection.
- We have a technical case study (Anscombe data) presented in one of the lab sessions to further discuss this extremely important aspect of model validation, mainly if regression (in general linear) models are formulated and estimated with a predictive scope of use in the future.



4-10 UNUSUAL AND INFLUENTIAL DATA

Outlying Observations can cause us to misinterpret patterns in plots

- Temporarily removing them can sometimes help see patterns that we otherwise would not have
- Transformations can also spread out clustered observations and bring in the outliers
- More importantly, separated points can have a strong influence on statistical models - removing outliers from a regression model can sometimes give completely different results

Unusual cases can substantially influence the fit of the OLS model

- Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model
- Outliers may also indicate that our model fails to capture important characteristics of the data

4-10 UNUSUAL AND INFLUENTIAL DATA

Types of Unusual Observations

- A Regression Outlier is an observation that has an unusual value of the outcome variable Y, conditional on its value of the explanatory variable X
  - An observation that is unconditionally unusual in either its Y or X value is called a univariate outlier, but it is not necessarily a regression outlier
  - In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
  - Regression outliers often have large residuals but do not necessarily affect the regression slope coefficient. Also sometimes referred to as vertical outliers.
  - Outliers could increase standard errors of the estimated parameters (or fitted values) since the standard error of regression is computed from residuals:

$$s^2 = \frac{\mathbf{e}^T \cdot \mathbf{e}}{n-p} = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \cdot \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p} = \frac{RSS}{n-p}$$

4-10 UNUSUAL AND INFLUENTIAL DATA

- Observation with Leverage (diagonal element of the hat matrix)
    - An observation that has an unusual X value - i.e., it is far from the mean of X - has leverage on the regression line
    - The further the outlier sits from the mean of X (either in a positive or negative direction), the more leverage it has
    - High leverage does not necessarily mean that it influences the regression coefficients, it is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data. Such cases are sometimes called "good" leverage points because they help the precision of the estimates.

- Influential Observations
    - An observation with high leverage that is also a regression outlier will strongly influence the regression line
    - In other words, it must have an unusual X-value with an unusual Y-value given its X-value
    - In such cases both the intercept and slope are affected, as the line chases the observation

Discrepancy × Leverage = Influence

4-10 UNUSUAL AND INFLUENTIAL DATA

4-10.1 *A priori* influential observations

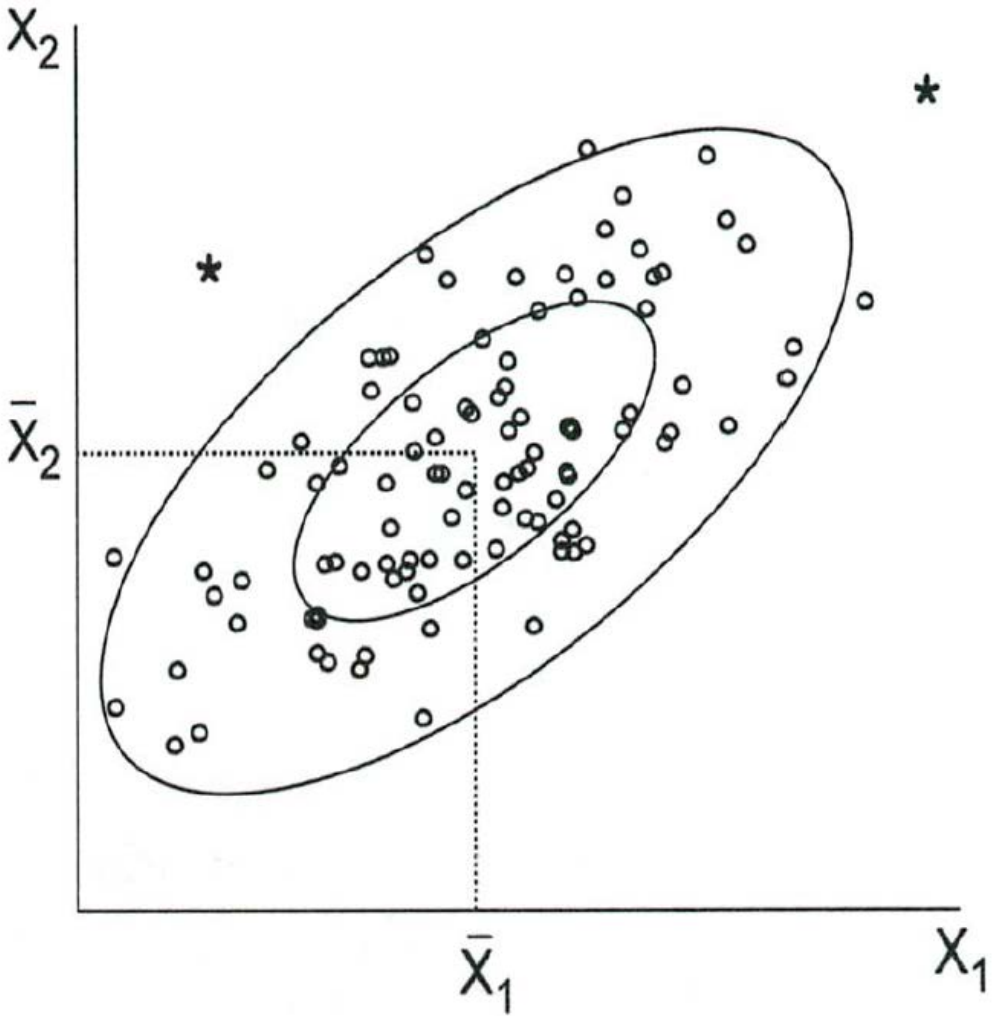
Simple regression: An observation that has an unusual X value - i.e., it is far from the mean of **X** - has leverage on the regression line The further the outlier lays from the mean of X (either in a positive or negative direction), the more leverage it has.

Multiple regression: we have to think in a cloud of points defined by regressors in **X** (each column in an axis) and center of gravity of those points.

Points **x** ($\mathbf{x} \in \Re^p$) heterogenous regarding the cloud of X points and their center of gravity identify *a priori* influential data.

- The most common measure of leverage is the hat – value, h_i, the name hat – values results from their calculation based on the fitted values ($\hat{y}_j$): Leverages h_i are computed for all observations and it is a measured of the i-th distance from the point $\mathbf{x_i}$ to the center of gravity of the whole set of observation data.
- And thus the average value for the leverage is $\overline{h} = \frac{\sum_i h_{ii}}{n} = \frac{p}{n}$ Belsley *et al.*, shown that atypical values for leverage observations are those that satisfy the cut-off: $h_{ii} > 2\overline{h}$.
- This cut-off is not useful when big data set are considered (cut off has to be increased to $h_{ii} > 3\overline{h}$

4-10 UNUSUAL AND INFLUENTIAL DATA



According to Fox (figure 11.3 in Fox, 1997), the diagram to the left shows elliptical contours of hat values for two explanatory variables.

As the contours suggest, hat values in multiple regression take into consideration the correlational and variational structure of the X's

As a result, outliers in multi-dimensional X-space are high leverage observations - i.e., the outcome variable values are irrelevant in calculating hi.

4-10 UNUSUAL AND INFLUENTIAL DATA

4-10.2 A posteriori influential data

An influential observation implies that the inclussion of the data in OLS:

1. Modifies the vector of estimated parameter $\hat{\boldsymbol{\beta}}$.
2. Modifies the fitted values $\hat{\mathbf{Y}}$.
3. If the i-th observation is influent that its fitted value is very good when i-th data is included in the data set for OLS estimation, but when it is removed its fitted value is bad, leading to a high value of the absolute residual.

An influential observation is one that combines discrepancy with leverage.

4. The most direct approach to assessing influence is to assess how the regression coefficients change if outliers are omitted from the model. We can use D_{ij} (often termed DFBetas$_{ij}$) to do so:

$$D_{ij} = \left(\hat{\beta}_j - \hat{\beta}_{j(i)}\right) \Big/ \hat{\sigma}_{\hat{\beta}_j} \qquad i = 1, \ldots, n;\ j = 1, \ldots, p$$

Where the $\hat{\beta}$ are the coefficients for all the data and the $\hat{\beta}_{(i)}$ are the coefficients for the same model with the i_{th} observation removed.

A standard cut-off for an influential observation is: $D_{ij} \geq 2/\sqrt{n}$ (be careful in small samples !!!)

4-10 UNUSUAL AND INFLUENTIAL DATA

A problem with DFBetas is that each observation has several measures of influence -one for each coefficient np different measures.

Cook's D overcomes the problem by presenting a single summary measure for each bservation D_i :

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^{\mathbf{T}} \mathbf{X}^{\mathbf{T}}\mathbf{X}\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{p\, s^2} = \left(\frac{e_i}{s\sqrt{1-p_{ii}}}\frac{1}{\sqrt{p}}\right)^2 \left(\frac{p_{ii}}{1-p_{ii}}\right) \approx F_{p,n-p}$$

where $\hat{\beta}_{(i)}$ are the coefficients for the same model with the i_{th} observation removed.

- Cook's D measures the 'squared distance' between $\hat{\beta}$ are the coefficients for all the data and the $\hat{\beta}_{(i)}$ are the coefficients for the same model with the i_{th} observation removed by calculating an F-test for the hypothesis that $H_i : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{(\mathbf{i})}$
- There is no significance test for Di but a commonly used cut-off is the Chatterjee y Hadi (88) that justifies an influential observation for those that satisfy Di >4/(n-p).
-
- For large samples, Chatterjee-Hadi cut-off does not work and as a rule of thumb Di >0.5 are suspected to be influential data and Di>1 are qualified as being influential (R criteria).

4-11 BEST MODEL SELECTION

The best regression equation for Y given the regressors $(X_1,\ldots,X_p)$ might contain dummy variables, transformations of the original variables and terms related to polynomial regression (higher order tan linear for covariate variable) for the original variables $(Z_1,\ldots,Z_q)$. Model selection should satisfy trade-off between simplicity and goodness of fit, often called parsimony criteria.

1. As many regressors as necessary to make good predictions, on average and with the highest precission on confidence interval.
2. Many variables are expensive to obtain data collection and difficult to mantain.

- It is not practical to built all possible regression models and choose the best one according to some balance criteria.
- A good model should show consistence to theoretical properties in residual analysis. Neither influential nor unsual data should be included.

4-11 BEST MODEL SELECTION

Available elements to assess the quality of a particular multiple regression **(goodness of fit)** model are:

1. Determination coefficient, R^2. Marginal increase is expected when the number of included regressors in the model gives consistency with available data. Any added regressors would increase (marginally) the determination coefficient, so stability is what has to found. Sometimes the adjusted coefficient is useful R_a^2.
2. Stabilibity on the standard error of regression estímate. Estimation of σ^2 by s^2 on underfitting is biased and greater than the true value. Stability on s^2confirms or at least points to goodness of fit.
3. Residual analysis.
4. Unusual and influent data analysis.
5. And a new element, Mallows C_p. Related to Akaike Information Criteria (AIC) $AIC = 2\left(-\ell\left(\hat{\boldsymbol{\beta}}, \mathbf{y}\right) + p\right)$. Models with lower values of Cp or AIC indicator are preferred.
   - Some authors strongly recommend BIC (*Bayesian Information Criteria*) Schwartz criteria $BIC = -2\ell\left(\hat{\boldsymbol{\beta}}, \mathbf{y}\right) + p\log n$ where extra parameters are penalized.
   - In R, `AIC(model)` for AIC on model objects for which a log-likelihood value can be obtained and `AIC(model, k=log(dim(data.frame)[1]) )` for BIC.

4-11 BEST MODEL SELECTION

4-11.1 Stepwise Regression

- Backward Elimination is an heuristic strategy to select the best model given a number of regressor and a maximal model built from them. It is a robust method that supressed non significative terms from the maximal model to the point that all mantained terms are statistically significative and can not be removed. It has been proven to be very effective for polynomial regression.
- F*orward inclusión* is an heuristic strategy to select the best model given a set of regressor from the null model is iteratively adding terms and regressor in the target set. It is not a robust procedure and it is not recommended as an authomatic procedure to find the best model to a data set and regressor terms.
- Stepwise Regression is a strategy that is forward increasing the starting model, but at each iteration regressor terms are checked for statistical significancy.

- *Criteria for adding/removing regressor terms varies in the different statistical packages, but F tests or AIC are commonly used. Partial correlation between Y and each Xj once some subset of regressors is already in the model has been proved succesful in the selection of regressors to increase the current model.*

4-11 BEST MODEL SELECTION

R software has a sophisticated implementation of these heuristics in the method step(model, target model) based on AIC criteria for model selection at each step.

```
> step(duncan1.lm0, ~income+education, direction="forward",data=duncan1)
#AIC direction "forward"
Start:  AIC=311.52
prestige ~ 1

           Df Sum of Sq   RSS    AIC
+ education  1     31707 11981 255.30
+ income     1     30665 13023 259.05
<none>                   43688 311.52

Step:  AIC=255.3
prestige ~ education

         Df Sum of Sq     RSS    AIC
+ income  1    4474.2  7506.7 236.26
<none>               11980.9 255.30

Step:  AIC=236.26
prestige ~ education + income

Call:
lm(formula = prestige ~ education + income, data = duncan1)
```

```
Coefficients:
(Intercept)    education       income
    -6.0647       0.5458       0.5987

> step(duncan1.lm2,data=duncan1)  # Without scope direction is "backward" using AIC
Start:  AIC=236.26
prestige ~ income + education

            Df Sum of Sq     RSS    AIC
<none>                    7506.7 236.26
- income     1    4474.2 11980.9 255.30
- education  1    5516.1 13022.8 259.05

Call:
lm(formula = prestige ~ income + education, data = duncan1)

Coefficients:
(Intercept)       income    education
    -6.0647       0.5987       0.5458

> step(duncan1.lm2,k=log(dim(duncan1)[1]),data=duncan1)
# Without scope direction is "backward" using BIC
Start:  AIC=241.68
prestige ~ income + education

            Df Sum of Sq     RSS    AIC
<none>                    7506.7 241.68
- income     1    4474.2 11980.9 258.91
- education  1    5516.1 13022.8 262.66
```

```
Call:
lm(formula = prestige ~ income + education, data = duncan1)

Coefficients:
(Intercept)       income    education
    -6.0647       0.5987       0.5458

> step(duncan1.lm0, ~income+education, data=duncan1)
#AIC direction "both"
Start:  AIC=311.52
prestige ~ 1

            Df Sum of Sq   RSS    AIC
+ education  1     31707 11981 255.30
+ income     1     30665 13023 259.05
<none>                   43688 311.52

Step:  AIC=255.3
prestige ~ education

            Df Sum of Sq   RSS    AIC
+ income     1      4474  7507 236.26
<none>                   11981 255.30
- education  1     31707 43688 311.52

Step:  AIC=236.26
prestige ~ education + income

            Df Sum of Sq     RSS    AIC
<none>                    7506.7 236.26
```

```
- income      1    4474.2 11980.9 255.30
- education   1    5516.1 13022.8 259.05

Call:
lm(formula = prestige ~ education + income, data = duncan1)

Coefficients:
(Intercept)    education       income
    -6.0647       0.5458       0.5987

>
```