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



Cal es...
el moder...

verificació

Realitat

validació
...redictiva

validaci...

Cal emprar el model per...
predir/avaluar...

**FIB-MASTER IN DATA SCIENCE-SIM**

**LECTURE NOTES:**

**UNIT 3: Tools for the Tactical Level - Statistical Modeling (I)**

# CONTENTS

# 3.1-1   READING LIST

**Basic references:**

📖 Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Edition 2015.
📖 Fox and Weisberg ***An R Companion to Applied Regression***. Sage Publications, Edition 2011.
📖 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, 2009.

# 3.1-2   INTRODUCTION TO LINEAR MODELS

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

In linear models, the **random component** $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ is assumed to be normally distributed $Y_i | X_i \sim N(\mu_i, \sigma)$ with constant variance $\sigma^2$, $V(Y_i | X_i) = \sigma^2$ and expectation $E(Y_i | X_i) = \mu_i$

➡ Therefore, the response variable is modeled as normally distributed; thus, negative or positive values, which may be arbitrarily small or large, may be encountered as data for the response and prediction.

➡ **The systematic component** of the model consists in specifying a vector called the linear predictor, denoted $\eta$, of the same length as the response, dimension $n$, obtained from the linear combination of regressors (explanatory variables). In vector notation, the parameters are $\boxed{\beta^T = (\beta_1, \ldots, \beta_p)}$ and the regressors are $\mathbf{X} = (X_1, \ldots, X_p)$ and, thus, $\boxed{\eta = \mathbf{X}\beta}$ where $\eta$ is $nx1$, $\mathbf{X}$ is $nxp$ and $\beta$ is $px1$.

➡ Vector $\mu$ is the direct linear predictor $\eta$; therefore, the **link function** is $\eta = \mu$ or $\mu_i = X_i\beta$

# 3.1-2 INTRODUCTION TO LINEAR MODELS

**Empirical problem: What do the data say about class sizes and test scores according to 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. TARGET (Y)
- Student-teacher ratio (STR) = number of students in the district divided by number of full-time equivalent teachers. (X)

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 | 10% | 25% | 40% | 50% (median) | 60% | 75% | 90% |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Percentile | | | |
| 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 |

# 3.1-2 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.
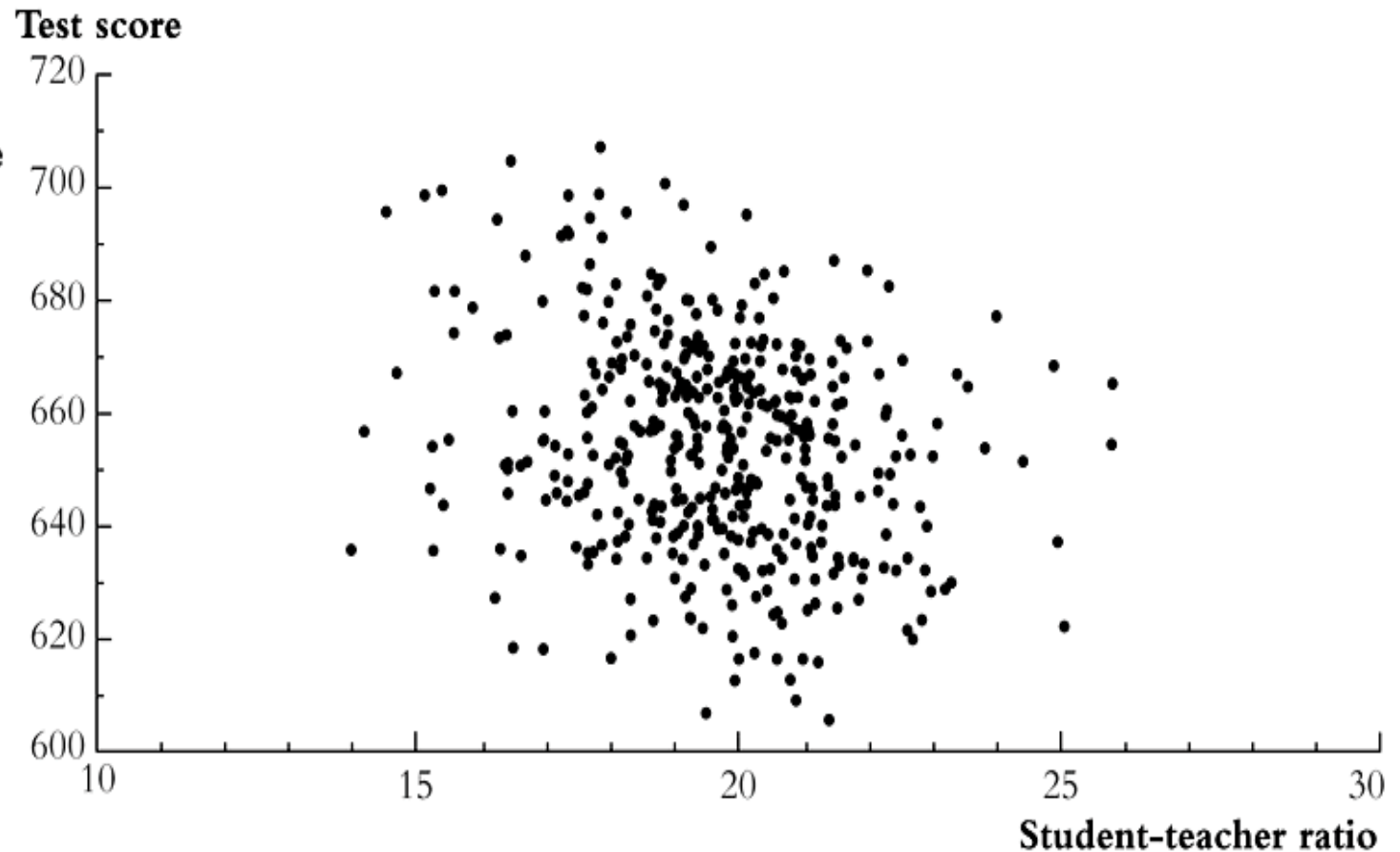


Stock and Watson (2007)

# 3.1-2 SOME NOTATION AND TERMINOLOGY

- The population regression line is

  **Test score** = $\beta_1$ + $\beta_2$**STR**

  $\beta_1$ is the intercept

  $\beta_2$ is the slope of the population regression line

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

- Why are $\beta_1$ and $\beta_2$ "population" parameters?
- We would like to know the population value of $\beta_2$.
- We don't know $\beta_2$, so we must estimate it using data.
- How can we **estimate** $\beta_1$ and $\beta_2$ from the data?

We will focus on the least squares ("ordinary least squares" or "OLS") estimator of the unknown parameters $\beta_1$ and $\beta_2$, which solves

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

# 3.1-2 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**



$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$
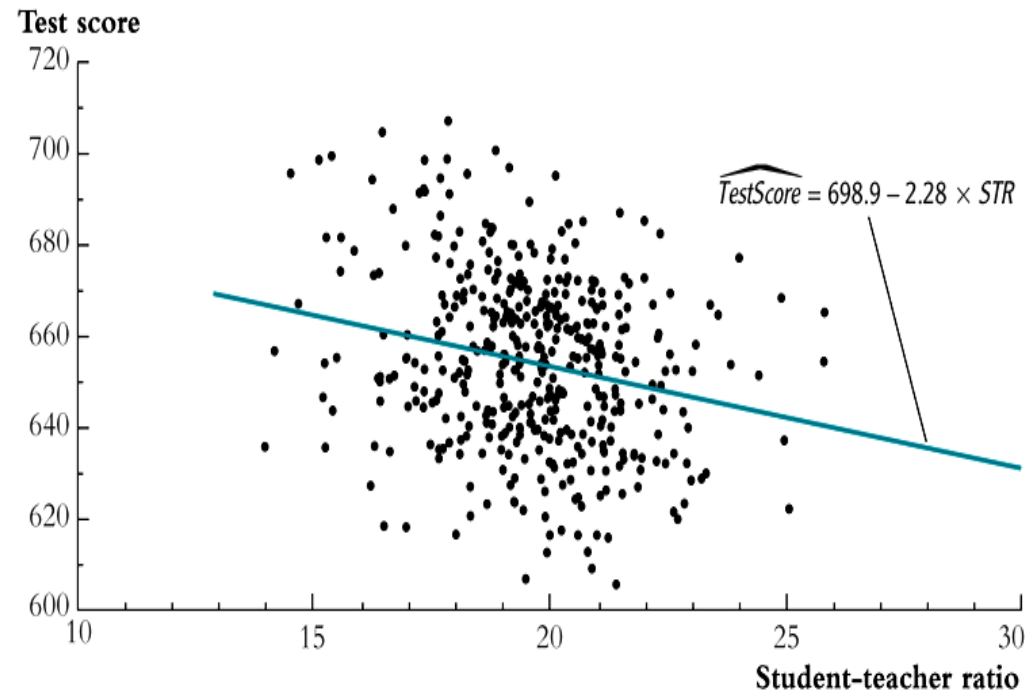
Estimated slope = – 2.28

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. *This makes no sense, since it extrapolates outside the range of the data.*

# 3.1-2 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 $\hat{\beta}_2$.
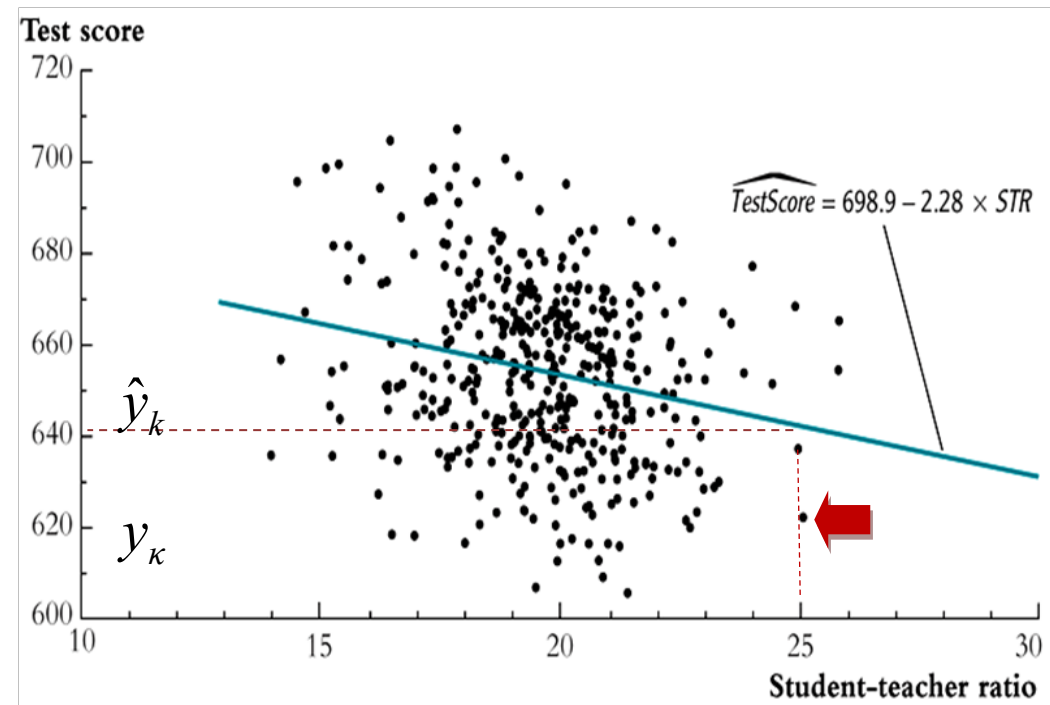
# 3.1-2 SOME NOTATION AND TERMINOLOGY

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 $\hat{\beta}_2$?

Let us proceed in four steps:

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

Vector-matrix notation for regression elements will be used since it simplifies the mathematical framework when dealing with several explanatory variables (regressors).

# 3.1-2 INTRODUCTION TO LINEAR MODELS

## Classification of statistical tools for analysis and modeling

| Explanatory variables | Response variable- Target | | | | |
|---|---|---|---|---|---|
| | Dichotomous or binary | Polytomous | Counts (discrete) | Continuous | |
| | | | | *Normal* | Time between events |
| Dichotomous | Contingency tables Logistic regression Log-linear models | Contingency tables Log-linear models | Log-linear models | Tests for 2 subpopulation means: t-test | Survival analysis |
| Polytomous | Contingency tables Logistic regression Log-linear models | Contingency tables Log-linear models | Log-linear models | ONE-WAY, 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 |

# 3.1-2 INTRODUCTION TO LINEAR MODELS

Normal regression analysis is a particular case of generalized linear models where the random component is normally distributed and thus a member of the exponential family.

An exponential family distribution has the following canonical form:

$$f_Y(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

Where $a(.)$, $b(.)$ and $c(.)$ are specific functions with known $\phi$ parameter and depending of a single parameter $\theta$, named the canonic parameter.

➡ Let us verify that the density function of a normal distribution variable can be written in the canonical form of exponential family members :

$$f_Y(y, \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right)$$

where $a(\phi) = \phi = \sigma^2$, $b(\theta) = \dfrac{\theta^2}{2} = \dfrac{\mu^2}{2}$ ( and $\theta = \mu$ ) $c(y, \phi) = -\dfrac{1}{2}\left(\dfrac{y^2}{\phi} + \log(2\pi\phi)\right)$.

# 3.1-2 INTRODUCTION TO LINEAR MODELS

➡ And thus:

$$\ell\big(\theta,\phi,y\big)=\log f_Y\big(y,\theta,\phi\big)=\frac{y\theta-b\big(\theta\big)}{a\big(\phi\big)}+c\big(y,\phi\big)=\frac{y\theta-\theta^2/2}{\phi}-\frac{1}{2}\left(\frac{y^2}{\phi}+\log\big(2\pi\phi\big)\right)$$

➡ For n observations belonging to the normal distribution, the scaled deviance statistic is

$$D'\big(\mathbf{y},\hat{\boldsymbol{\mu}}\big)=2\,\ell(\mathbf{y},\phi,\mathbf{y})-2\,\ell\big(\hat{\boldsymbol{\mu}},\phi,\mathbf{y}\big)=\sum_{i=1}^{n}\frac{\big(y_i-\hat{\mu}_i\big)^2}{\phi}$$

➡ And the deviance statistic is the well known residual sum of squares:

$$D\big(\mathbf{y},\hat{\boldsymbol{\mu}}\big)=D'\big(\mathbf{y},\hat{\boldsymbol{\mu}}\big)\phi=\sum_{i=1}^{n}\big(y_i-\hat{\mu}_i\big)^2$$

## 3.1-3    LEAST SQUARES ESTIMATION IN MULTIPLE REGRESSION

Assume we have a linear model with no distribution hypothesis,

$$\boxed{\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$ , where $\mathbf{Y}$ is *nx1*, $\mathbf{X}$ design matrix *nxp* and $\beta$ the vector parameters *px1*

**Let Y be a** numeric response variable and $\varepsilon$ the model error

$$
\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} == 
\begin{pmatrix}
1 & \overbrace{x_{12}}^{j=2} & \overbrace{x_{13}}^{j=3} & \overbrace{\cdots}^{\cdots} & \overbrace{x_{1p}}^{j=p} \\
1 & x_{22} & x_{23} & \cdots & x_{2p} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{n-1\ldots2} & x_{n-13} & \cdots & x_{n-1p} \\
1 & x_{n2} & x_{n3} & \cdots & x_{np}
\end{pmatrix}
\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} +
\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}
$$

The ordinary least squares estimation of model parameters $\beta$ can be written in the general case as

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

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

# 3.1-3 LEAST SQUARES ESTIMATION

➡️ First-order condition for a minimum 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:

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

- The solution of the normal equations is the least square estimator $\hat{\beta}$ of the $\beta$ parameters.

- If the design matrix is not singular, i.e., the column range of $X$ 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 $X$ is not full-rank, 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 $\boxed{\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}}$.

# 3.1-3 LEAST SQUARES ESTIMATION

➡ ANOVA and ANCOVA models have singular design matrices and reparametrization is required, and solutions to the normal equations can be computed using generalized inverses or pseudo inverses

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X^T X}\right)^{-} \mathbf{X^T Y}$$

➡ $\mathbf{A}^{-} = \left(\mathbf{X}^T \mathbf{X}\right)^{-}$ is a g-invers or generalized invers if satisfies $\mathbf{A A}^{-} \mathbf{A} = \mathbf{A}$.

$\mathbf{A}^{-}$ always exists, but it not uniqu. When it is also satisfied:

1. $\mathbf{A}^{-} \mathbf{A A}^{-} = \mathbf{A}^{-}$    2. $\left(\mathbf{A A}^{-}\right)^T = \mathbf{A A}^{-}$    3. $\left(\mathbf{A}^{-} \mathbf{A}\right)^T = \mathbf{A}^{-} \mathbf{A}$

Then $\mathbf{A}^{-}$ is unique and it is called the Moore-Penrose inverse or p-inversa noted as $\mathbf{A}^{+}$.

➡ Second order conditions of minima request that second order derivatives of $S(\boldsymbol{\beta})$, the hessian matrix of the quadratic form, be positive definite or ,

$$\nabla_{\boldsymbol{\beta}}^{2} S(\boldsymbol{\beta}) > \mathbf{0} \leftrightarrow \left[\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}\right] > 0 \quad i, j = 1, \dots, p$$

# 3.1-3 LEAST SQUARES ESTIMATION

The range of $\mathbf{X^T X}$ is equal to the range of $\mathbf{X}$. This attribute leads to two criteria that must be met in order to ensure $\mathbf{X^T X}$ is nonsingular 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 can not be perfectly linearly related, but even near collinearity can cause statistical problems.
3. Use **vif(model)** in R to check collinearity (thresholds will be discussed in lab sessions).

## 3.1-3.1 Geometric properties

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

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

And $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ the predictions once the least squares estimator of model parameters is calculated.
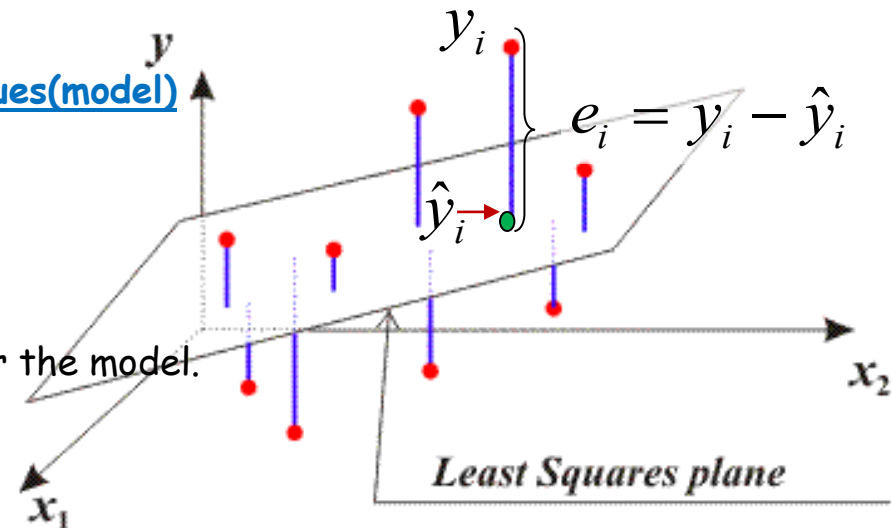
# 3.1-3 LEAST SQUARES ESTIMATION

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

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}} \text{ puts a ^ on } \mathbf{Y}: \boxed{\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}\mathbf{Y}}.$$

Graphically,

**Properties of the hat matrix: Diagonal H – hatvalues(model)**

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



*Least Squares plane*

# 3.1-4 LEAST SQUARES ESTIMATION: INFERENCE

## 3.1-4.1 Basic inference properties

➡ A continuous response linear model is assumed where

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$ or $Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \varepsilon_i, \; i = 1,\dots,n$

where $\mathbf{Y}, \mu$ are nx1, $\mathbf{X}$ nxp of p rank and $\beta$ px1 and the conditional distribution of unbiased errors are i.i.d. of constant variance and normally distributed: $\boxed{\varepsilon \,|\, \mathbf{X} \approx \mathrm{N}\left(0, \sigma^2\right)}$ or $\mathbf{Y}|\mathbf{X} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n})$.

➡ Then, the **minimum-variance unbiased estimator** of $\beta$ is $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, the ordinary least squares estimator is equal to the maximum likelihood estimator $\hat{\beta}_{MV}$.

➡ If normal assumptions are not met, then OLS estimators are not efficient (they do not have minimum variance as ML estimator).

➡ If the hypothesis holds true, then the unbiased estimator of $\sigma^2$, denoted $s^2$, is efficient (minimum variance) and its square root, $s$, is called the standard error of the model.

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

Residual sum of squares

# 3.1-4 LEAST SQUARES ESTIMATION: INFERENCE

➡ **More detailed (Th 3.5 Seber (77) pag. 54) :**

1. $\hat{\beta} \approx N_p\left(\beta, \sigma^2\left(\mathbf{X^T X}\right)^{-1}\right)$ is the sampling distribution of the OLS estimator.

2. $\left(\hat{\beta} - \beta\right)^T \mathbf{V}\left[\hat{\beta}\right]^{-1} \left(\hat{\beta} - \beta\right) = \left(\hat{\beta} - \beta\right)^T \mathbf{X^T X}\left(\hat{\beta} - \beta\right)/\sigma^2 \approx \chi^2_p$ .

3. $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{Y}$ is independent of $s^2$.

4. $RSS/\sigma^2 = (n-p)s^2/\sigma^2 \approx \chi^2_{n-p}$ .

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 $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{Y}$, an estimator of something, specifically the population parameter vector $\beta$ 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.

# 3.1-4 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 $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ an estimator of something, specifically the population parameter vector $\beta$, and the former theorem tells us that the OLS estimator is linear and unbiased. It is better and more 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 $\beta_j$ and sampling variance $\mathbf{V}\left(\hat{\beta}_{\mathbf{j}}\right) = \sigma^2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)_{jj}^T$ and we can test the simple hypothesis (i.e., make some inference):

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

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

## 3.1-4 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 $SE\left(\hat{\beta}_{\mathbf{j}}\right)$ (standard error of

jth parameter) is distributed as a Student t with n-p degrees of freedom (since $\hat{\beta}_j$ and $s^2$ are independent).

$$t_0=\frac{\beta_j-\beta^0_j}{s\sqrt{\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^T_{jj}}}\approx Student\ t_{n-p}$$ can be defined and thus $P\left(H_0\right)=P\left(t_{n-p}>t_0\right)$ or a bilateral

confidence interval at $100\left(1-\alpha\right)\%$ for $\beta_j\in\hat{\beta}_j\pm t^{\alpha/2}_{n-p}SE\left(\hat{\beta}_j\right)$ can be calculated.

Inference for multiple coefficients will be further presented by the F-test.

# 3.1-4 GENERALIZED LEAST SQUARES ESTIMATION

## 3.1-4.2 Generalized least squares

Let $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ be a multiple regression model and as usual $\mathbf{Y}$ *nx1*, $\mathbf{X}$ *nxp* of range *p* and $\boldsymbol{\beta}$ *px1*. Let the model errors be unbiased and correlated with normal distribution

$$\boldsymbol{\varepsilon}\,|\mathbf{X} \approx \mathrm{N}_n\left(0,\ \sigma^2\mathbf{W}\right)$$ or equivalently say $Y|\mathbf{X} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{W})$

,

where $\mathbf{W}$ is a symmetric and positive defined matrix of dimension *nxn* .

$\mathbf{W}$ in the model $Y|\mathbf{X} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{W})$ can be Cholesky factorized uniquely and transformed in a matrix product of a lower triangular **K** matrix of dimensión *nxn and its transpose; i.e.,* $\mathbf{W} = \mathbf{K}\mathbf{K}^\mathbf{T}$. Cholesky factorization is usually computed as a triangular decomposition,

$$\mathbf{W} = \mathbf{L}\mathbf{U} = \mathbf{L}\left(\mathbf{D}\mathbf{L}^\mathbf{T}\right) = \mathbf{L}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{L}^\mathbf{T} = \left(\mathbf{L}\mathbf{D}^{\frac{1}{2}}\right)\left(\mathbf{D}^{\frac{1}{2}}\mathbf{L}^\mathbf{T}\right) = \mathbf{K}\mathbf{K}^\mathbf{T}.$$

## 3.1-4 GENERALIZED LEAST SQUARES ESTIMATION

$\mathbf{K}$ matrix appearing in the factorization of $\mathbf{W}$ is the tool to compute linear transformations for the response variable $\left(\text{noted } \widetilde{\mathbf{Y}}\right)$, and for the design matrix $\widetilde{\mathbf{X}}$ and the error term $\widetilde{\varepsilon}$ :

1. $\widetilde{\mathbf{Y}} = \mathbf{K}^{-1}\mathbf{Y}$     2. $\widetilde{\mathbf{X}} = \mathbf{K}^{-1}\mathbf{X}$     3. $\widetilde{\varepsilon} = \mathbf{K}^{-1}\varepsilon$

4. $\mathbf{K}^{-1}\mathbf{Y} = \mathbf{K}^{-1}\eta + \mathbf{K}^{-1}\varepsilon = \mathbf{K}^{-1}\mathbf{X}\beta + \mathbf{K}^{-1}\varepsilon \leftrightarrow \widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\beta + \widetilde{\varepsilon}$ .

➡ Now, in the transformed linear model $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\beta + \widetilde{\varepsilon}$ errors are uncorrelated and the variance is common and equal to $\sigma^2$,

$$\mathbf{V}\left[\widetilde{\mathbf{Y}}\right] = \mathbf{V}\left[\widetilde{\mathbf{X}}\beta + \widetilde{\varepsilon}\right] = \mathbf{V}\left[\widetilde{\varepsilon}\right] = \mathbf{V}\left[\mathbf{K}^{-1}\varepsilon\right] = \mathbf{K}^{-1}\mathbf{V}\left[\varepsilon\right]\left(\mathbf{K}^{-1}\right)^{T} = \sigma^2\mathbf{K}^{-1}\mathbf{W}\mathbf{K}^{-T} = \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}^{T}\mathbf{K}^{-T} = \sigma^2\mathbf{I}_n$$

# 3.1-4 GENERALIZED LEAST SQUARES ESTIMATION

Sinthesis of the least square estimation for linear models with normal response :

| **MODEL** | $Y = X\beta + \varepsilon,\ V[\varepsilon] = \sigma^2 I$ | $Y = X\beta + \varepsilon,\ V[\varepsilon] = \sigma^2 D$ | $Y = X\beta + \varepsilon,\ V[\varepsilon] = \sigma^2 W = \sigma^2 KK^T$ |
|---|---|---|---|
| **Transformation** | $Y \to Y,\ X \to X$ | $Y \to D^{-\frac{1}{2}}Y,\ X \to D^{-\frac{1}{2}}X$ | $Y \to K^{-1}Y,\ X \to K^{-1}X$ |
| $S(\beta)$ Quadratic form to minimize | $(Y - X\beta)^T (Y - X\beta)$ | $(Y - X\beta)^T D^{-1} (Y - X\beta)$ | $(Y - X\beta)^T W^{-1} (Y - X\beta)$ |
| Normal Equations $\boxed{S\beta = Q}$ | $S = X^T X,\ Q = X^T Y$ | $S = X^T D^{-1} X,\ Q = X^T D^{-1} Y$ | $S = X^T W^{-1} X,\ Q = X^T W^{-1} Y$ |
| Estimator $\hat{\beta}$ | $(X^T X)^{-1} X^T Y$ | $(X^T D^{-1} X)^{-1} X^T D^{-1} Y$ | $(X^T W^{-1} X)^{-1} X^T W^{-1} Y$ |
| Variance of $\hat{\beta}$ | $\sigma^2 (X^T X)^{-1}$ | $\sigma^2 (X^T D^{-1} X)^{-1}$ | $\sigma^2 (X^T W^{-1} X)^{-1}$ |
| RSS | $Y^T Y - (X^T Y)^T \hat{\beta}$ | $Y^T D^{-1} Y - (X^T D^{-1} Y)^T \hat{\beta}$ | $Y^T W^{-1} Y - (X^T W^{-1} Y)^T \hat{\beta}$ |

## 3.1-5    HYPOTHESIS TESTING IN MULTIPLE REGRESSION

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

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$ , where $\mathbf{Y}$ _nx1_, $\mathbf{X}$ _nxp_ of range _p_ and $\boldsymbol{\beta}$ _px1_

where errors are unbiased and iid is normally distributed $\boldsymbol{\varepsilon}|\mathbf{X} \approx \mathrm{N}_n\left(0, \sigma^2\mathbf{W}\right)$. **Ordinary least square estimators are denoted** $\hat{\boldsymbol{\beta}}$ .

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

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$ , where $\mathbf{Y}$ _nx1_, $\mathbf{X}$ _nxp_ of range _p_ and $\boldsymbol{\beta}$ _px1_

is subject to linear constraints $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ that define a linear hypothesis to contrast (to make an inference) that we call H; $\mathbf{A}$ is a $qxp$ matrix of range $q < p$. **Ordinary least square estimates subject to constraints are** $\hat{\boldsymbol{\beta}}_H$. The residual sum of squares is denoted RSS_H.

# 3.1-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

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

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

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

$$\boldsymbol{\beta}^{\mathbf{T}} = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}$$

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

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

# 3.1-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

### 3.1-5.1 Testing in R

| anova() |
|---|

- For nested linear regression models, method anova(restrictedmodel, fullmodel) implements the F-test. Hypothesis testing in this subject will be suggested by this method.
- Net effects tests for a multiple regression model performed with package car Anova(model)
- Inconvenient: the two models have to be computed beforehand.

| 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)

# 3.1-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

## 3.1-5.2 Confidence interval for single model parameters

Individual confidence interval for $\beta_i$ in OLS is summarized as:

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

$$s = \hat{\sigma} = \sqrt{\frac{RSS}{n-p}}$$

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

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

There is much literature about how to build confidence regions or perform simultaneous tests on several hypotheses: this falls outside of the scope of this material. Some specific suggestions will be presented in the practical sessions.

# 3.1-6    MULTIPLE CORRELATION COEFFICIENT

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

$$R = \frac{\sum_{k}\left(y_k - \overline{y}\right)\left(\hat{y}_k - \overline{\hat{y}}\right)}{\left\{\sum_{k}\left(y_k - \overline{y}\right)^2 \sum_{k}\left(\hat{y}_k - \overline{\hat{y}}\right)^2\right\}^{\frac{1}{2}}}$$

➡️ The square of the multiple correlation coefficient **R²** 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= $\sum_{k}\left(y_k - \overline{y}\right)^2$  where  $\overline{y} = \frac{1}{n}\sum_{k} y_k$  is the mean of the observed response data.

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

# 3.1-6 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

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

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

# 3.1-6 MULTIPLE CORRELATION COEFFICIENT

## 3.1-6.1 Properties of the multiple correlation coefficient

1. $|R| \leq 1$ and if **|R|=1** this 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.

## 3.1-6.2 Adjusted R2

➡ Since this correlation can not go down when variables are added, the R2 must be adjusted in order to increase only when truly significant regressors are added to the model. This is the adjusted R-Squared:

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

➡ Adjusted R-Squared is always less than the original R-Squared and might be negative.

## 3.1-7    GLOBAL TEST FOR REGRESSION. ANOVA TABLE

➡️ The **global test for regression** is a particular case of multiple comparison of hypotheses in which all parameters related to the explanatory variables are tested to be simultaneously zero.

$$H: \quad \beta_2 = 0, \dots, \beta_p = 0 .$$

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

➡️ This Omnibus test sometimes is presented to clarify in form of a table, called the **ANOVA Table for a Regression Model**:

| ANOVA TABLE | Descomposition | Degrees of freedon | Variance | Contrast |
|---|---|---|---|---|
| **ESS** | $\sum_k \left( \hat{y}_k - \bar{y} \right)^2$ | p-1 | $s^2_{\exp} = ESS/(p-1)$ | |
| **RSS** | $\sum_k \left( y_k - \hat{y}_k \right)^2$ | n-p | $s^2 = RSS/(n-p)$ | $F = s^2_{\exp} / s^2$ |
| **TSS** | $\sum_k \left( y_k - \bar{y} \right)^2$ | n-1 | $s^2_Y = TSS/(n-1)$ | |

## 3.1-8  PREDICTIONS AND INFERENCE

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

---

1. $E\left[\hat{Y}_k\right] = E\left[\mathbf{x}_k^T \hat{\boldsymbol{\beta}}\right] = \mathbf{x}_k^T \boldsymbol{\beta}$.

2. $V\left[\hat{Y}_k\right] = V\left[\mathbf{x}_k^T \hat{\boldsymbol{\beta}}\right] = \mathbf{x}_k^T V\left[\hat{\boldsymbol{\beta}}\right]\mathbf{x}_k = \sigma^2 \mathbf{x}_k^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_k = \sigma^2 h_{kk}$, where $h_{kk}$ is the k diagonal term of projection matrix appearing in OLS (H, hat matrix).

3. Fitted values are normally distributed.

4. In R: predict(model, interval = "confidence") for 95% CI for fitted values.

5. New data: predict(model, newdata=dataframe, interval = "prediction") for 95% CI for fitted values.

---

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

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

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

# 3.1-9    MODEL VALIDATION

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

Let a model for 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 $\beta$ is *px1*

and errors are unbiased, uncorrelated and normally distributed with constant variance, i.e.,

$\boldsymbol{\varepsilon}\big|\mathbf{X} \approx \mathrm{N}_n\big(0,\ \sigma^2\mathbf{I}\big)$ or equivalently $Y|\mathbf{X} \sim N_n(\boldsymbol{\mu},\sigma^2\mathbf{I})$.

➡️ Residuals are the difference between observed response values and fitted values:

$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \big(\mathbf{I} - \mathbf{H}\big)\mathbf{Y}$.

➡️ Check:

- ○ Linearity: $E(Y_i|X_i) = \mu_i = X_i\beta$
- ○ Homokedasticity (constant variance): $V(Y_i|X_i) = \sigma^2$
- ○ Normality: $Y_i|X_i \sim Normal$
- ○ Independence: $Y_i|X_i \perp Y_j|X_j, i \neq j$

# 3.1-9 MODEL VALIDATION: RESIDUAL ANALYSIS

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

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

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

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

- 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.

# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

## Residual analysis - usual plots:



Histogram of rstudent(anscombe.lmA)

Empirical quantiles of Studentized Residuals(anscombe.lmA)
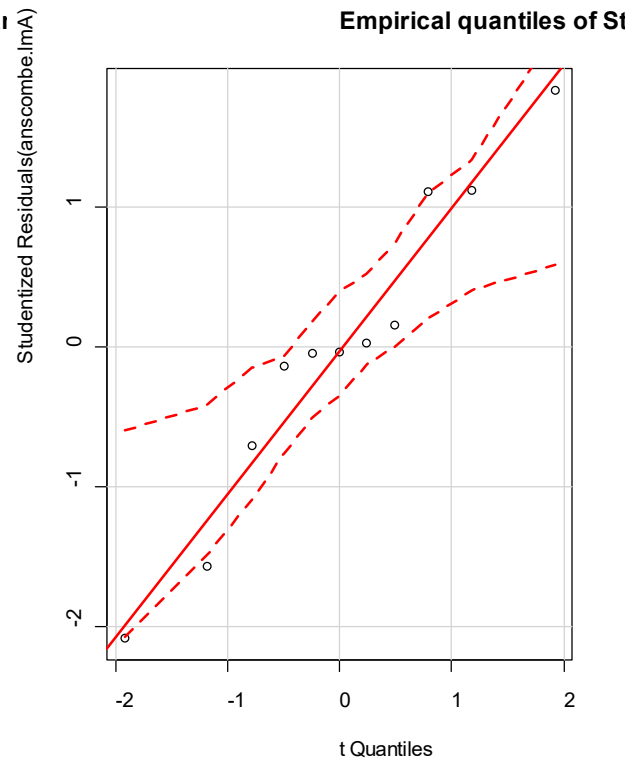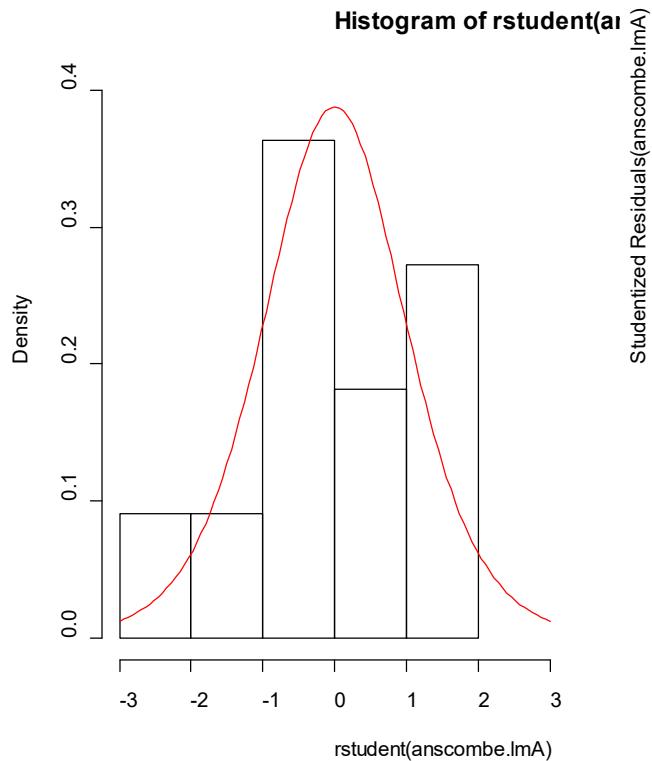
➡ Histogram residuals: normal density is checked.

```
hist(rstudent(model),
freq=F)

curve(dt(x, model
$df),col=2,add=T)
```

➡ Boxplot residuals to identify outliers of regression.

➡ Normal probability plot or quantile plot of standardized or studentized residuals.

In R, the confidence band for studentized residuals is shown loading car package and qqPlot method.

```
library(car)

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

# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

o  When the normality assumption is not fully met then the comparisons of the hypothesis based on the Student or F-test are approximate and estimates are not efficient.

o  Tests for normality assessment can be used. The Shapiro-Wilke test is one of my favorites, but there are lots of tests depending on sample size. Package nortest in R contains some common possibilities. Even in non-normal errors, residuals tend towards normality (due to central limit theorem) in medium and large samples.

o  Residuals are correlated with observations $Y_i$, but not with fitted values $\hat{Y}_i$, so scatterplots with fitted values on the X-axis are suggested.

o  Use residualPlots(model) method in package car for R.

➡  Scatterplots: $\boxed{e_i \ vs \ \hat{Y}_i}$, or $\boxed{d_i \ vs \ \hat{Y}_i}$ or $\boxed{r_i \ vs \ \hat{Y}_i}$. Failed linearity (regressor transformations or additions might be needed) or heteroscedasticity (transformations required) might be detected. Unusual observations might make interpretation difficult.

➡  Use default residual analysis of a linear model for R.

```
par(mfrow=c(2,2))
plot(model, id=list(n=Inf,labels=row.names(df)))
par(mfrow=c(1,1))
```
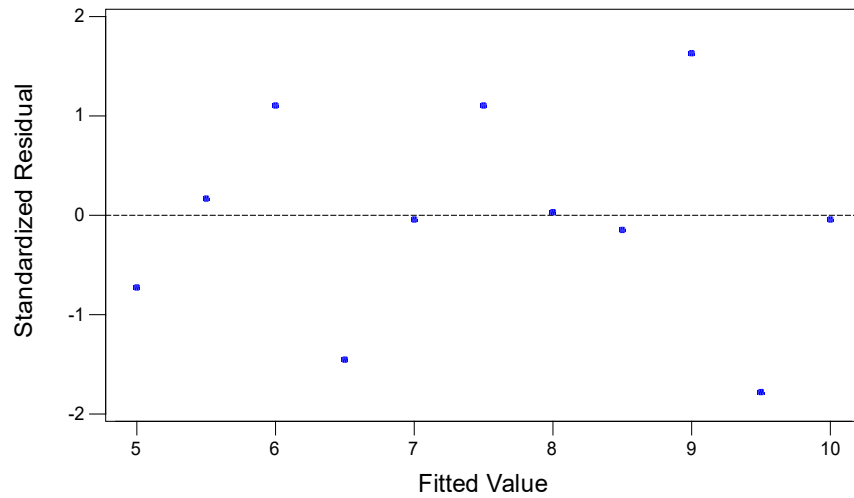
# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS
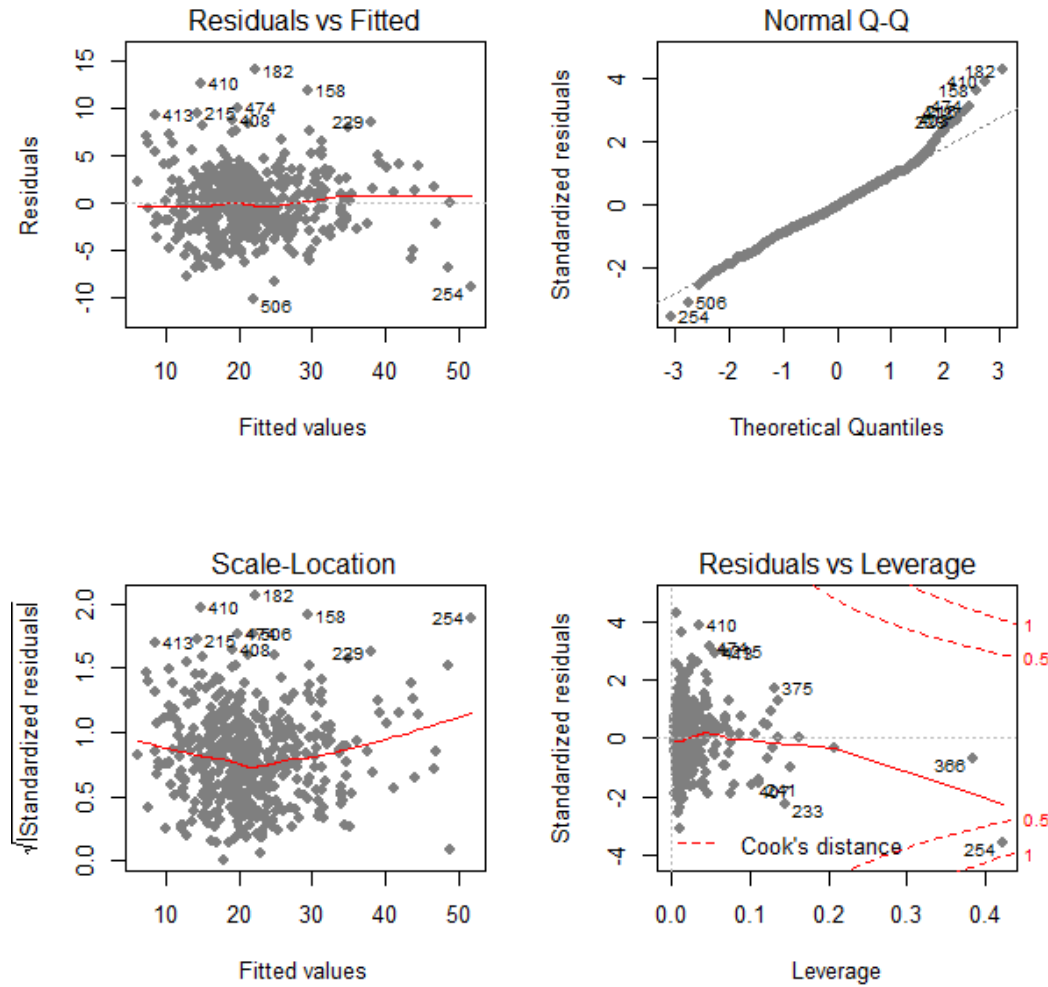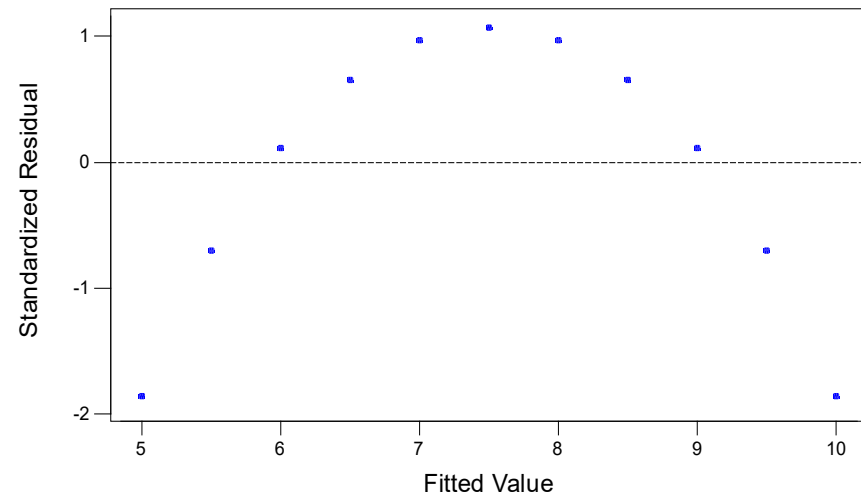
# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS
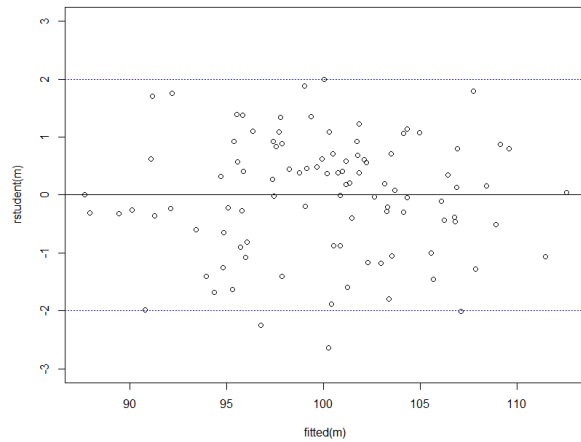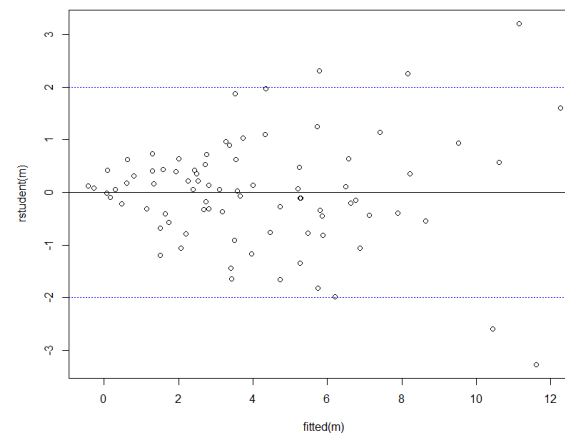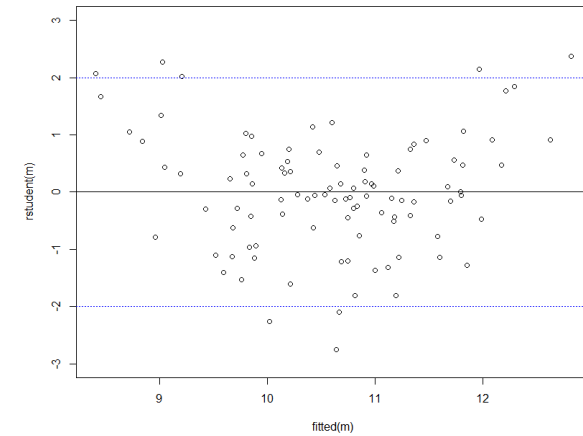
# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

Linear hypothesis and homoskedastic    Heteroskedastic    Linear hypothesis rejected

# 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

➡️ ***Scatterplots of residuals vs each regressor*** (except intercept)

The horizontal band indicates linearity satisfaction and homoskedastic hypothesis.

o Use residualPlots(model) method in package car for R. Example: Duncan data
`lm(prestige~education+income,data=Duncan)`



Lack of fit for education

OK for income
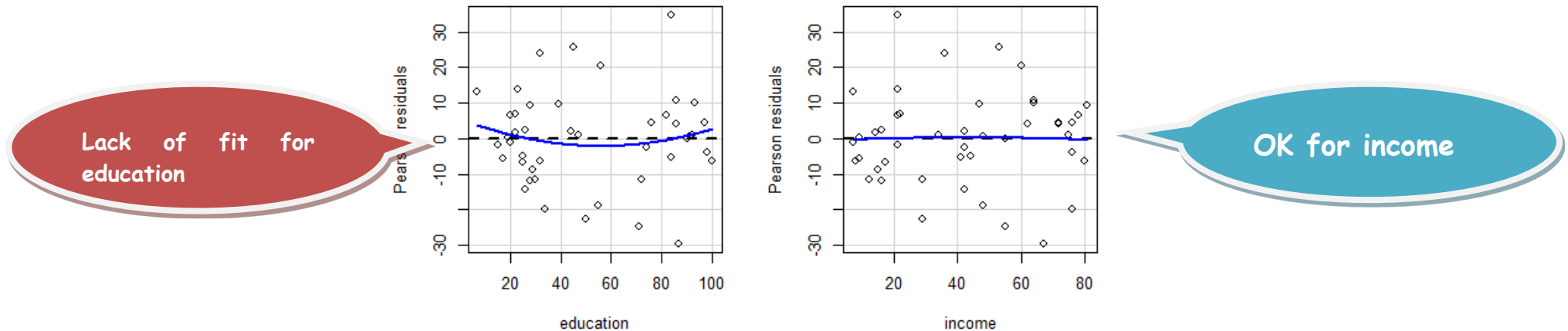
o Homoskedastic Hypothesis Test - Breusch-Pagan test in package lmtest might be of interest.

```
> library(lmtest)
> bptest(model)   # Null Hypothesis: Homoskedasticity holds
      studentized Breusch-Pagan test data:   model
BP = 0.57522, df = 2, p-value = 0.7501
```

H0 can't be rejected

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UPC

Departament d'Estadística
i Investigació Operativa

*SIM course. Master in Data Science – FIB- UPC*

## 3.1-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

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

Such as the 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 contrasts for simultaneous hypothesis $r(j)=0 \text{ for } j>k$.

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

No temporal dependency                    Temporal dependency
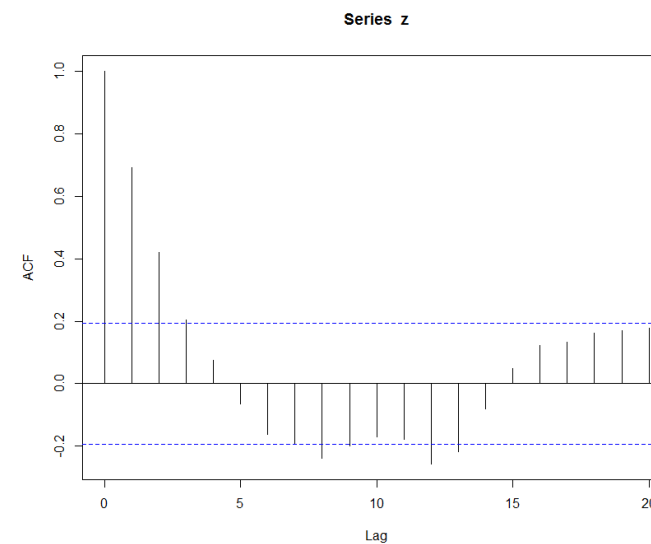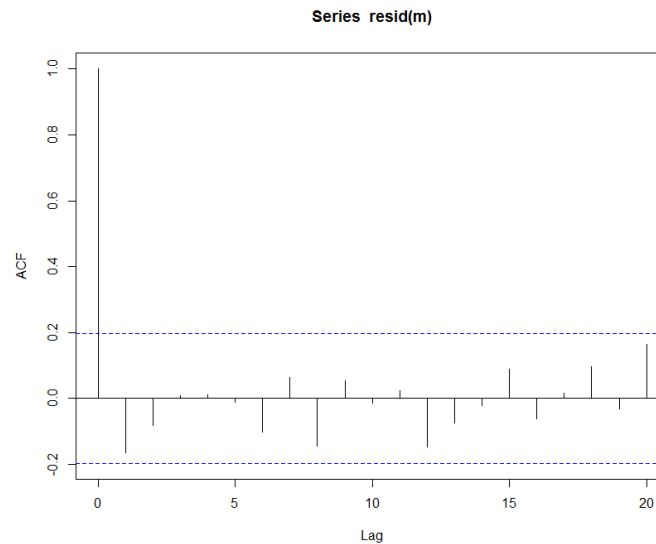
# 3.1-10 MODEL TRANSFORMATIONS ON Y OR X

o Use marginalModelPlots(model) method in package car for R or avPlots(model). Lack of fit between data smoother and current model behavior for one variable indicates that transformation on selected regressor is needed.

   o Use poly(varname,2) to model linear and quadratic terms on varname regressor.

   o Use poly(varname,3) to model linear, quadratic and cubic terms on varname regressor.



Marginal Model Plots
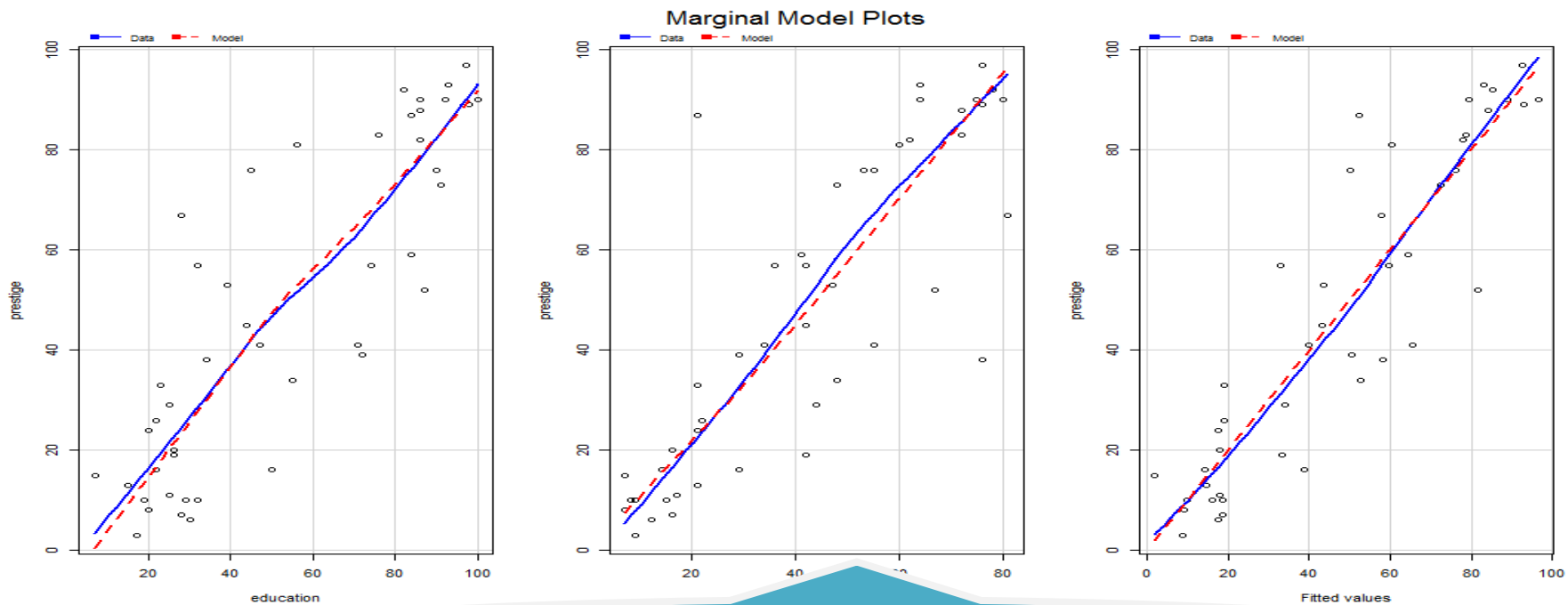
There seem to be no problems in any of them!

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS

## 3.1-10.1    Box-Cox transformation of Y

o The Box-Cox transformation of Y functions to normalize the error distribution, stabilize the error variance and straighten the relationship of Y to the Xs. **Basic transformations are log(Y), 1/Y, $\sqrt{Y}$ :**

   o If $\lambda$ =1, no transformation is necessary. Note that all of the $Y_i$ must be positive.

   o $\lambda$ may be by maximum likelihood in R: boxcox( formula, data ) method in MASS library.

   o Or boxCoxVariable(response) after loading car package with $\lambda$ =1-coefficient for the boxCoxVariable() ). The t-test for the coefficient of the constructed variable is an approximate score test for whether a transformation is required.

$$h(Y) = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ log\ Y & \lambda = 0 \end{cases}$$

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS

```
lm(formula = prestige ~ boxCoxVariable(prestige) + income + education, data = Duncan)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             13.7510     7.7283   1.779 0.082607 .
boxCoxVariable(prestige) 0.5954     0.2001   2.975 0.004888 **
income                   0.5860     0.1099   5.332 3.84e-06 ***
education                0.4165     0.1001   4.160 0.000158 **

Residual standard error: 12.27 on 41 degrees of freedom
Multiple R-squared: 0.8587,    Adjusted R-squared: 0.8483
```
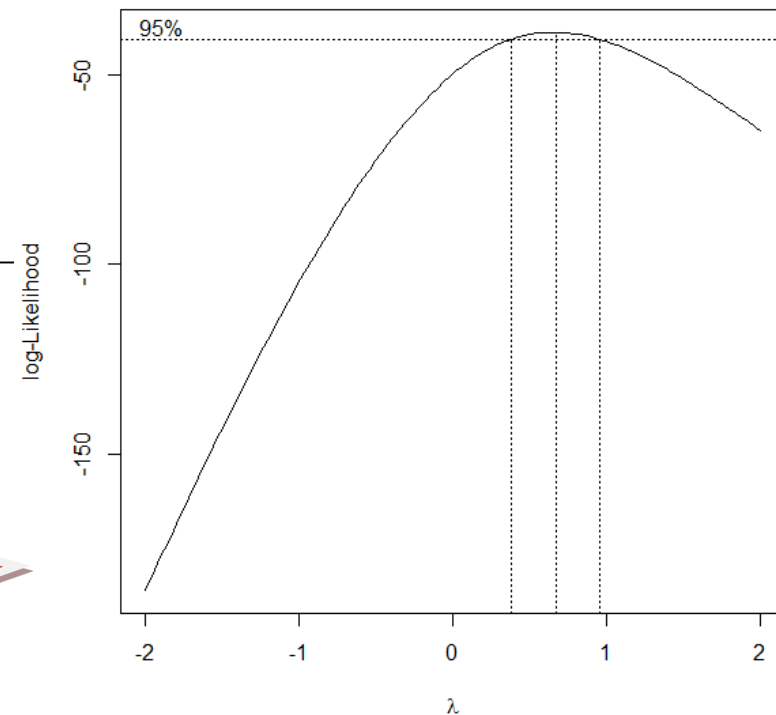
*Or*

```
boxcox(prestige ~ income + education, data = df)
```

Transformation is needed

95% CI for λ excludes 1

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS

### 3.1-10.2    Box-Tidwell Transformation of the X's

Maximum likelihood can also be used to find an appropriate linearizing transformation for the X variables

The Box-Tidwell model is a non-linear model that estimates transformation parameters for the X's simultaneously with the regular parameters,

$$Y_i = \beta_1 + \beta_2 X^{\gamma_2}_{2i} + ... + \beta_p X^{\gamma_p}_{pi} + \varepsilon_i, \quad i = 1,...,n$$ errors are iid normal and $X_i$ are positive

Explicit in this model is a power transformation of each of the X's

Of course, **it makes no sense to transform dummy variables** and the like, so we should not attempt to estimate transformation parameters for them.

➡ BoxTidwell method in R is available (for Duncan data)

```
> library(car)
> boxTidwell(prestige ~ income, ~ education,  data=duncan1)
 Score Statistic    p-value MLE of lambda
      -0.1323884 0.8946771     0.9275973
> boxTidwell(prestige ~ income + education,  data=duncan1)
          Score Statistic    p-value MLE of lambda
income          -0.6859686 0.4927329     0.6622138
education        1.0408104 0.2979635     1.7312407
```

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS
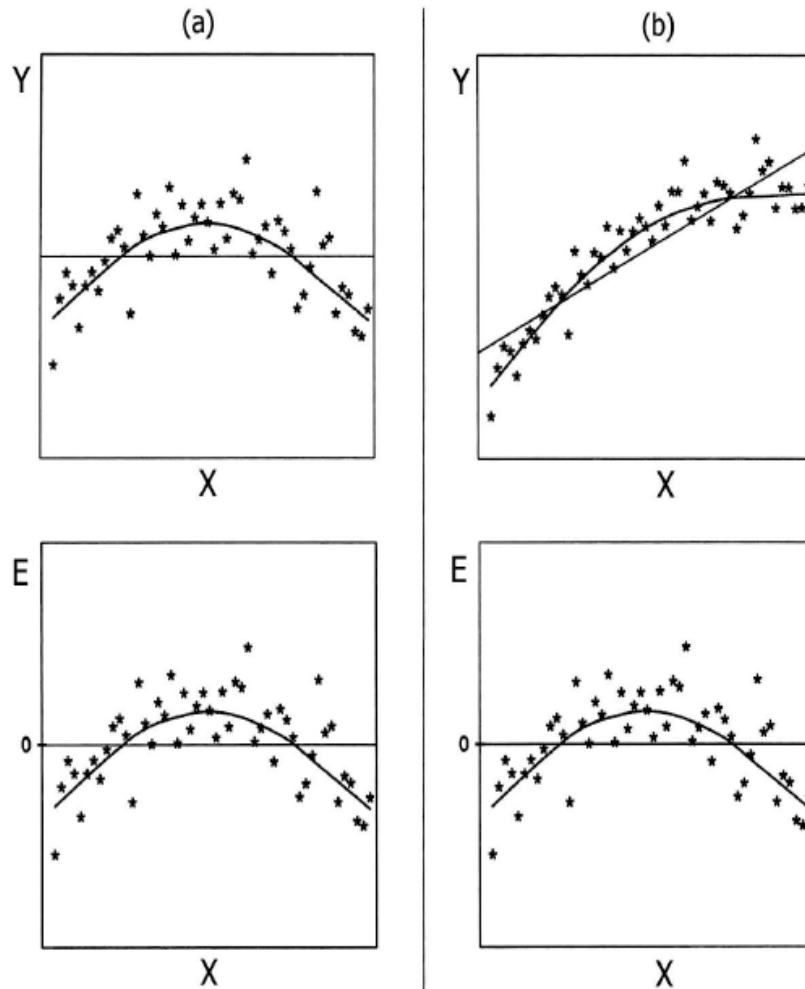


- Scatterplot matrices are useful for preliminary assessments of the relationship between several variables in a multiple regression model, but can be misleading because they plot the marginal rather than partial relationships between Y and each X (i.e., they do not control for the other X's)

- Partial-regression plots (or added-variable plots, avPlot()) are not very useful either because they are unable to distinguish **between monotone linearity (which can often be corrected with a simple transformation)** and **non-monotone non-linearity** (which cannot be corrected with a transformation)

- Partial-residual plots, however, can reveal both monotone and non-monotone linearity.

o Notice that although column (a) is characterized by non-monotone non-linearity.

o Only (b) can be transformed to satisfy the linearity requirement.

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS



Added-Variable Plots
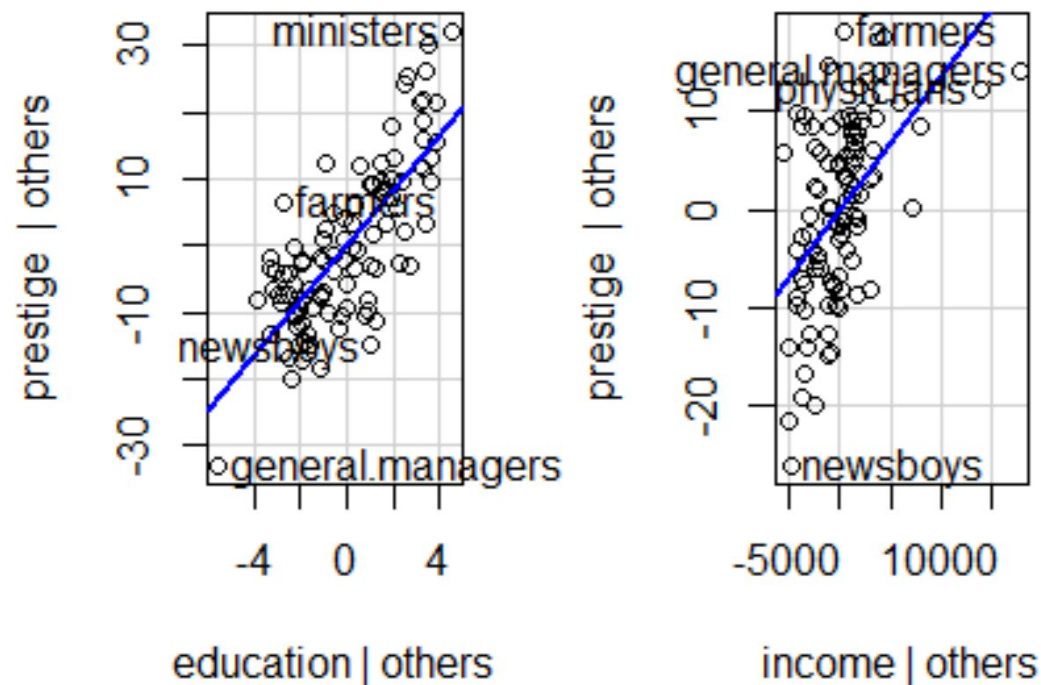
avPlots(model) : Partial-regression plots (or added-variable plots, avPlots()). Partial regression plots for a response variable (prestige) and predictor (income) are scatterplots of the residuals from 2 regressions - regressing the outcome prestige on all of the other predictors (education, in the example), and regressing that particular predictor (income) on all of the other predictors.

It is useful to address the contribution of the predictor to explain the outcome and highlights potentially influential observations
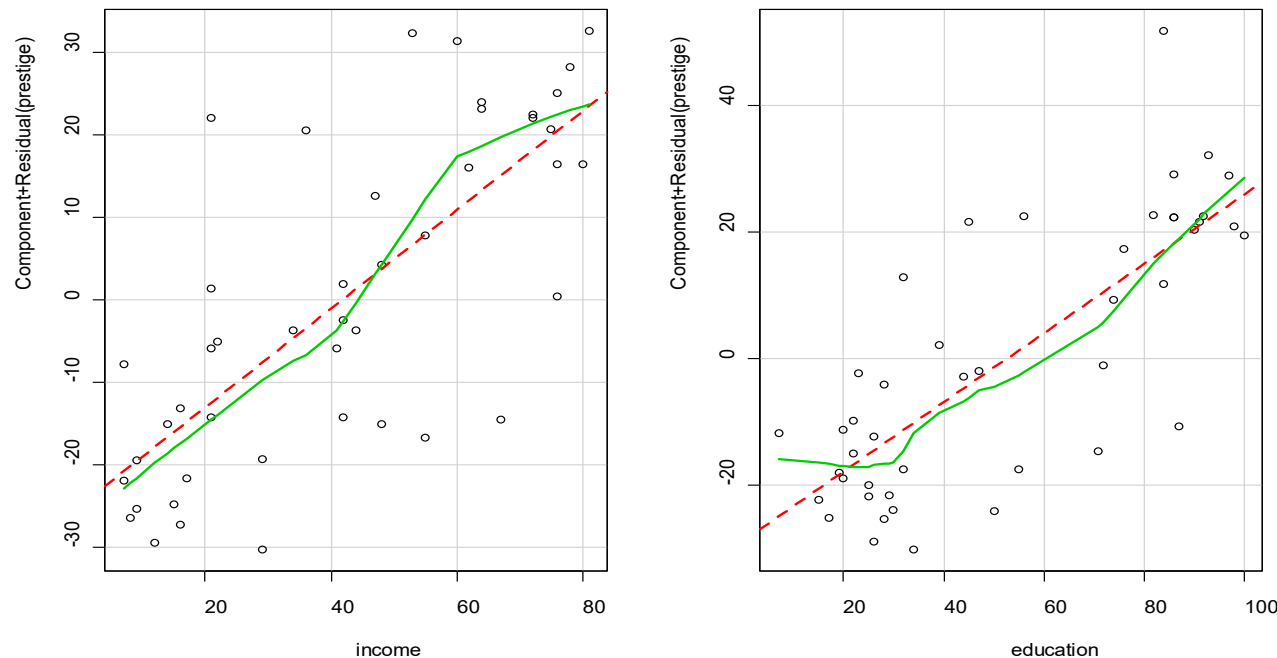
# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS

**Partial-residual plots** can reveal both monotone and non-monotone linearity. The partial residual for the $j_{th}$ explanatory variable from a multiple regression is,

$$pe_i^{(j)} = e_i + \beta^{(j)} x_{ij}$$

Where this simply adds the linear component of the **partial regression between Y and $X_j$** (which may be characterized by a non-linear component) to the least squares residuals.

The "partial residuals" $pe_i^{(j)}$ are plotted versus $X_j$, meaning that $\beta^{(j)}$ is the slope of the multiple simple regression of $pe_i^{(j)}$ on $X_{j..}$

*In R car package:*

```
crPlot(duncan1.lm,"income")

crPlot(duncan1.lm,"education")
```

# 3.1-10 MODEL VALIDATION: TRANSFORMATIONS

## 3.1-10.3      Transformations and Interpretation

Often the response variable or regressors have to be transformed to linearize its relationship with y.

However, it is essential putting model interpretations back in the metric of y and x. This is easy for log transformations:

- log(y) = a + bx: when x increases by one unit, on average y increases by 100 × b percent.
- y = a + b log(x): when x increases by 1 percent, on average y increases by b 100 units.
- log(y) = a + b log(x): when x increases by 1 percent, on average y increases by b percent.

# 3.1-11 MODEL VALIDATION: UNUSUAL AND INFLUENTIAL DATA

It is easy to find examples of regression analysis that show the existence of a few observations that strongly affect the estimation of model parameters in OLS: 1% of data might weigh more in parameter estimation than 99% of the data.



→ Influential data affects model prediction and it seems only natural that predicted values should be supported by 99% of data and not seriously affected by the remaining 1%.

→ Classifying an observation as *a priori influential* is related to robustness of the design of data collection.

→ We'll see a technical case study (Anscombe's data) in one of the lab sessions to further discuss this extremely important aspect of model validation if regression (in general linear) models are formulated and estimated with a predictive future scope of use.
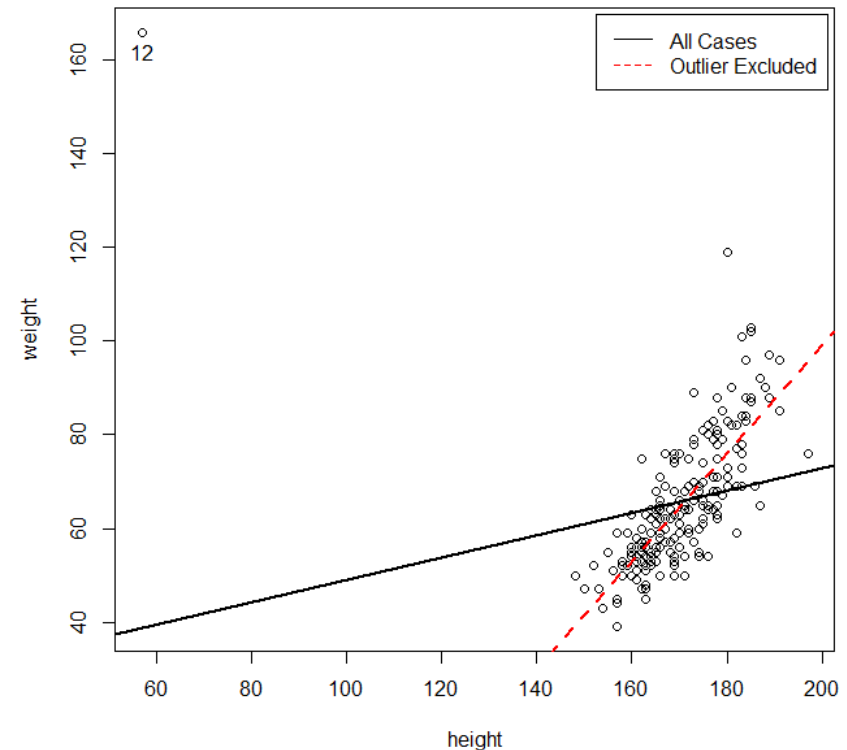
# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

Outlying observations can lead us to misinterpret patterns in plots.

- Temporarily removing them can sometimes help us see patterns that we otherwise would not.
- 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. A final model shouldn't contain influent data.
- Outliers may also indicate that our model fails to capture important characteristics of the data.

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

**Types of unusual observations**

- A regression outlier is an observation that has an unusual value of the outcome variable Y, conditioned by the value of the explanatory variable X.
  - An observation that is unconditionally unusual in either its Y or X value is called a univariate outlier; this 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. They are 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:

$$s^2 = \frac{\mathbf{e}^{\mathrm{T}} \cdot \mathbf{e}}{n-p} = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \cdot \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p} = \frac{RSS}{n-p}$$
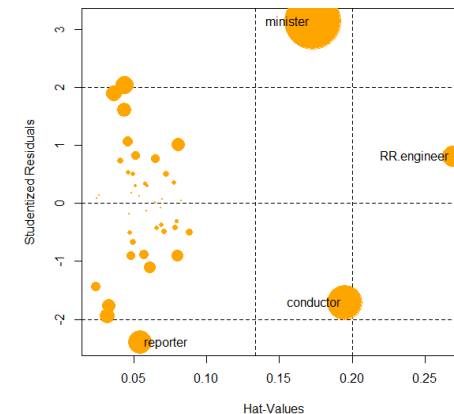
# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

- Observation with leverage (large diagonal element of hat matrix, $h_{ii}$)
    - 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 high leverage and yet fall straight in line with the pattern of the rest of the data.
- 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.
    - Both the intercept and slope are affected.

         **Discrepancy × leverage = influence**

- Use influencePlot(model) method in package car

```
> influencePlot(m1, col="orange",pch=19,id=list(method="noteworthy",n=3))
             StudRes        Hat       CookD
minister    3.1345186 0.17305816 0.56637974
reporter   -2.3970224 0.05439356 0.09898456
conductor  -1.7040324 0.19454165 0.22364122
RR.engineer 0.8089221 0.26908963 0.08096807
```

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

### 3.1-11.1  *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 is from the mean of X (either in a positive or negative direction), the more leverage it has.
- Multiple regression: We must imagine a cloud of points defined by regressors in **X** (each column on an axis) and a center of gravity for those points.

Heterogeneous points **x** ( $\mathbf{x} \in \Re^p$ ) 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 – value comes from their calculation based on the fitted values ( $\hat{y}_j$ ): leverages $h_i$ are computed for all observations and the i-th distance from point $\mathbf{x_i}$ to the center of gravity is measured for the whole set of observation data. Use the hatvalues(model) method in R.

➡ Thus, the average value for the leverage is $\overline{h} = \dfrac{\sum_i h_{ii}}{n} = \dfrac{p}{n}$ . Belsley et al. show that atypical values for leverage observations are those that meet the cut-off: $h_{ii} > 2\overline{h}$ .

➡ This cut-off is not useful when considering big data sets: the cut-off has to be increased to $h_{ii} > 3\overline{h}$ .

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA
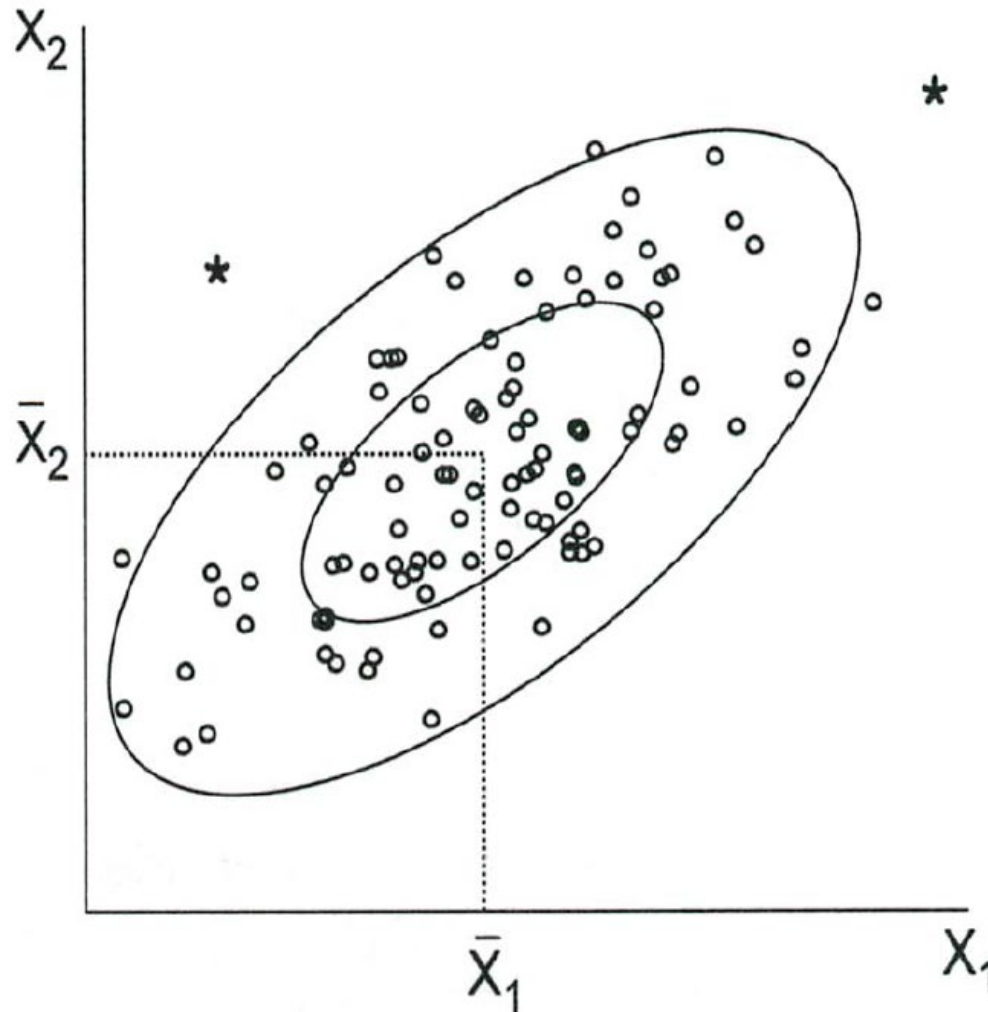


According to Fox (Figure 11.3 in Fox, 1997), the diagram to the left shows elliptical contours of hat values for two explanatory variables.

As the contours suggest, hat values in multiple regression take into consideration the correlational and variational structure of the Xs.

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

o Use aq.plot(dataframe) in package mvoutlier for R, where dataframe contains only numeric variables.

o Use chemometrics package: Moutlier method.

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

### 3.1-11.2    A posteriori influential data

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

1. Modifies the vector of estimated parameter $\hat{\beta}$.

2. Modifies the fitted values $\hat{\mathbf{Y}}$.

3. If the i-th observation is influential, then 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 and 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. Use dfbetas(model) in R.

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

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

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

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

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

where $\hat{\beta}_{(i)}$ are the coefficients for the same model with the $i_{th}$ observation removed. Use cooks.distance(model) method in R.

- Cook's D measures the 'squared distance' between $\hat{\beta}$, which are the coefficients for all the data, and $\hat{\beta}_{(-i)}$, which 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 one proposed by Chatterjee and Hadi (1988) that defines influential observations as those that satisfy Di >4/(n-p).

- For large samples, the Chatterjee and 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 considered influential (R criteria).

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

● In R it's possible to plot DFBETAS (using matplot) for each observation and regressor, as well as the indication of overall influence related to Cook's distance. To unify metrics, Cook's distance is a squared distance, but DFBETAS are ordinary units, so the squared root of Di (Cook's distance) can be added to the DFBETAS matplot. Standard cut-offs may be suggested and drawn.
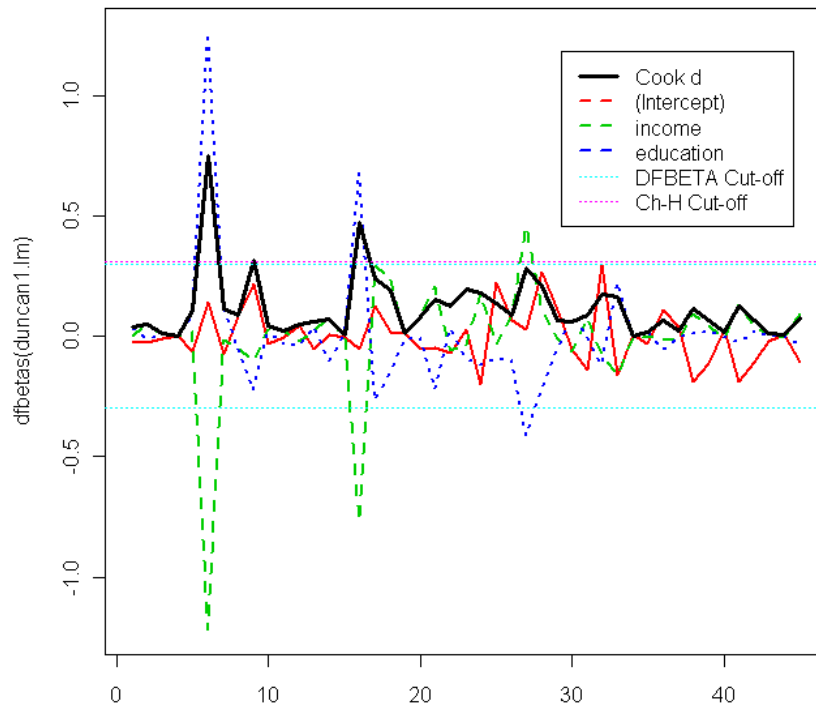


```
>matplot(dfbetas(duncan1.lm), type="l",
col=2:4,lwd=2)

>lines(sqrt(cooks.distance(duncan1.lm))
,col=1,lwd=3)

>abline(h=2/sqrt(dim(duncan1)[1]),
lty=3,lwd=1,col=5)

>abline(h=-2/sqrt(dim(duncan1)[1]),
lty=3,lwd=1,col=5)

> abline(h=sqrt(4/(dim(duncan1)[1]-
length(names(coef(duncan1.lm))))),
lty=3,lwd=1,col=6)

>llegenda<-c("Cook d",
names(coef(duncan1.lm)), "DFBETA Cut-
off", "Ch-H Cut-off")

> legend(locator(n=1), legend=llegenda,
col=1:length(llegenda),
lty=c(1,2,2,2,3,3), lwd=c(3,2,2,2,1,1))
```
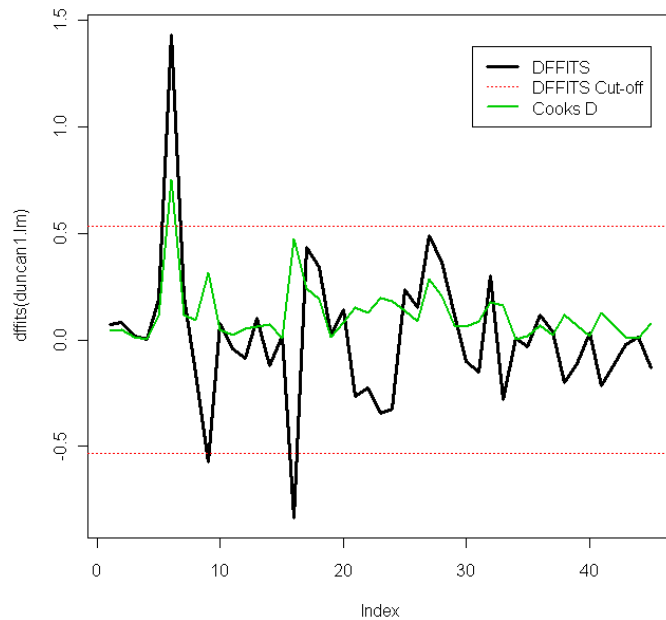
# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

- DFFITS$_i$ suggested by Belsley et al. (1980) are related to Cook's distance and combine studentized residuals and leverages. A usual cut-off for abnormal absolute value is $\left|DFFITS_i\right| > 2\sqrt{p/(n-p)}$. Use dffits(model) in R.

$$\boxed{DFFITS_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{X}\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{s_{(i)}^2} = \left(\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}\right)\sqrt{\frac{h_{ii}}{1-h_{ii}}} \rightarrow \left|DFFITS_i\right| = \sqrt{pD_i}}} \text{ or } DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_{ii}}}$$



```
influence(model)
plot(dffits(model),type="l",lwd=3)
pp=length(names(coef(model)))
lines(sqrt(cooks.distance(model)),col=3,lwd=2)
abline(h=2*(sqrt(pp/(nrow(duncan)-pp))),lty=3,lwd=1,col=2)
abline(h=-2*(sqrt(pp/(nrow(duncan1)-pp))),lty=3,lwd=1,col=2)
llegenda<-c("DFFITS","DFFITS Cut-off","Cooks D")
legend(locator(n=1),legend=llegenda,col=1:3,lty=c(1,3,1),lwd=c(3,1,2))
```

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

## 3.1-11.3    Joint influence

**Subsets of cases can jointly influence a regression line, or can offset each other's influence.**

- Depending on where the jointly influential cases lie, they can have different effects on the regression line.
- Cook's D helps to determine joint influence if there are relatively few influential cases.
- Cases can be deleted sequentially, updating the model each time and exploring the Cook's D's again. This approach is impractical if there are potentially a large number of subsets to explore.

Added-variable plots or partial-regression (leverage) plots provide a more useful method of assessing joint influence, since these plots essentially show the partial relationships between Y and each X:

For each regressor j in the design matrix X one plot is made:

1. Let $ey^{(j)}$ represent the residuals from the least-squares regression of Y on all of the X's except for $X_j$:

2. Similarly, $ex^{(j)}$ be the residuals from the regression of $X_j$ on all the other X's.

3. These two equations determine the residuals $ex^{(j)}$ and $ey^{(j)}$ as parts of $X_j$ and Y that remain when the effects of all other Xs are removed.

4. Plot of $ey^{(j)}$ versus $ex^{(j)}$ (in R, avPlots(model, variable) in car package is available).

# 3.1-11 UNUSUAL AND INFLUENTIAL DATA

**The residuals ex$^{(j)}$ and ey$^{(j)}$ have the following properties:**

- The slope of the regression of ey$^{(j)}$ on ex$^{(j)}$

  is the least-squares slope $\hat{\beta}_j$ from the full multiple regression.

- **The partial correlation coefficient of Y and Xj given all the other X's are in the model is the square root of the coefficient of determination of the simple regression of ey$^{(j)}$ on ex$^{(j)}$.**

- Residuals from the regression of ey$^{(j)}$ on ex$^{(j)}$ are the same as the residuals from the full regression.

- Variation of ex$^{(j)}$ is the conditional variance of X$_j$ holding the other X's constant.
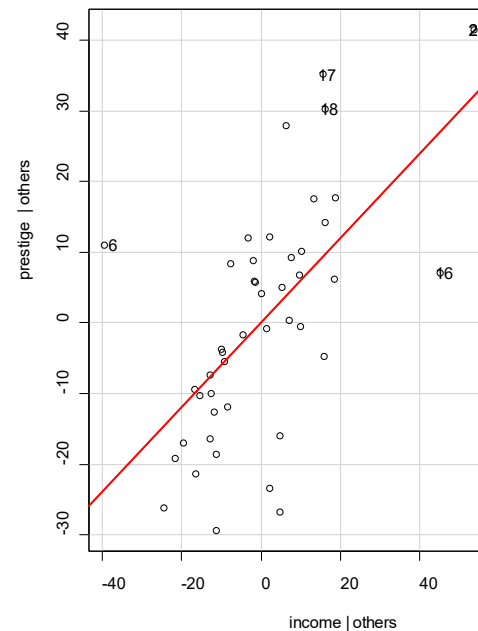
Added-Variable Plots
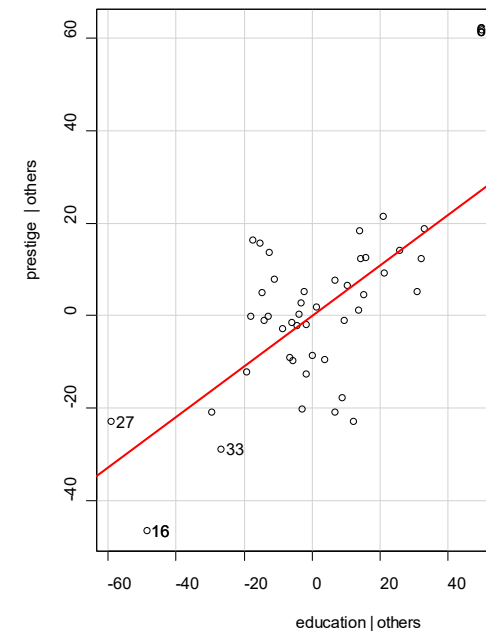


A plot of ey$^{(j)}$ versus ex$^{(j)}$ allows to examine the leverage and influence of cases on $\hat{\beta}_j$. These plots also give us an idea of the **precision of the slope** $\hat{\beta}_j$.

```
avPlots(duncan1.lm,labels = rownames(duncan1), id.method=list("x", "y"), id.n=3)
```

# 3.1-12 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 rather than linear for covariate variables) 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 precision in confidence interval.

2.  Many variables are expensive to obtain (data collection) and difficult to maintain.

➡ It is not practical to build all possible regression models and choose the best one according to some balance criteria.

➡ A good model should be consistent with theoretical properties in residual analysis. Neither influential nor unusual data should be included.

## 3.1-12 BEST MODEL SELECTION

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

1. Determination coefficient $R^2$. A marginal increase is expected when the number of regressors included in the model is consistent with the data available. Any added regressors would (marginally) increase the determination coefficient, so stability must be found. Sometimes the adjusted coefficient $R^2_a$ is useful.

2. Stability of the standard error of regression estimate. Estimation of $\sigma^2$ by $s^2$ of underfitting is biased and greater than the true value. The stability of $s^2$ confirms or at least points to goodness of fit.

3. Residual analysis.

4. Unusual and influential 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(nrow(data.frame)))` for BIC.

# 3.1-12 BEST MODEL SELECTION

## 3.1-12.1      Stepwise regression

➡️ Backward elimination is a heuristic strategy to select the best model given a number of regressors and a maximal model built from them. It is a robust method that suppresses insignificant terms from the maximal model to the point that all the terms maintained are statistically significant and cannot be removed. It has been proven to be very effective for polynomial regression.

➡️ Forward inclusion is a heuristic strategy to select the best model given a set of regressors from the null model by iteratively adding terms and regressors to the target set. It is not a robust procedure and it is not recommended as an automatic procedure to find the best model for a data set and regressor terms.

➡️ Stepwise regression is a forward strategy that builds on the starting model but, at each iteration, regressor terms are checked for statistical significance.

---

➡️ *Criteria for adding/removing regressor terms vary in 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 proven to be successful in the selection of regressors to increase the current model.*

---

# 3.1-12 BEST MODEL SELECTION

R software implements these heuristics in a sophisticated way 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

>
```