

FIB-Bachelor's
degree in
Informatics
Engineering

LECTURE NOTES:




UNIT 4: Tools for the Tactical Level - Statistical Modeling (I)

CONTENTS

4-1	READING LIST	3
4-2	INTRODUCTION TO LINEAR MODELS	4
4-3	LEAST SQUARES ESTIMATION IN MULTIPLE REGRESSION	12
4-3.1	GEOMETRIC PROPERTIES	14
4-4	LEAST SQUARES ESTIMATION: INFERENCE	16
4-4.1	BASIC INFERENCE PROPERTIES	16
4-5	HYPOTHESIS TESTING IN MULTIPLE REGRESSION	19
4-5.1	TESTING IN R	21
4-5.2	CONFIDENCE INTERVAL FOR MODEL PARAMETERS	22
4-6	MULTIPLE CORRELATION COEFFICIENT	23
4-6.1	PROPERTIES OF THE MULTIPLE CORRELATION COEFFICIENT	25
4-6.2	ADJUSTED R²	25
4-7	GLOBAL TEST FOR REGRESSION. ANOVA TABLE	26
4-8	PREDICTIONS AND INFERENCE	27
4-9	MODEL VALIDATION	28
4-10	MODEL TRANSFORMATIONS ON Y OR X	37
4-10.1	BOX-COX TRANSFORMATION OF Y	38
4-11	MODEL VALIDATION: UNUSUAL AND INFLUENTIAL DATA	40
4-11.1	A PRIORI INFLUENTIAL OBSERVATIONS	44
4-11.2	A POSTERIORI INFLUENTIAL DATA	46
4-12	BEST MODEL SELECTION	50
4-12.1	STEPWISE REGRESSION	52

4-1 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.

4-2 INTRODUCTION TO LINEAR MODELS

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

- ➔ In linear models, the **random component** distribution for $\mathbf{Y}^T = (Y_1, \dots, Y_n)$ is assumed to be normal with constant variance σ^2 and expectation $E[\mathbf{Y}] = \boldsymbol{\mu}$.
- ➔ 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 $\boldsymbol{\eta}$, of the same length as the response, dimension n , obtained from the linear combination of regressors (explanatory variables). In vector notation, the parameters are $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$ and the regressors are $\mathbf{X} = (X_1, \dots, X_p)$ and, thus, $\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta}$ where $\boldsymbol{\eta}$ is $n \times 1$, \mathbf{X} is $n \times p$ and $\boldsymbol{\beta}$ is $p \times 1$.
- ➔ Vector $\boldsymbol{\mu}$ is the direct linear predictor $\boldsymbol{\eta}$; therefore, the **link function** is $\boldsymbol{\eta} = \boldsymbol{\mu}$.

4-2 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)

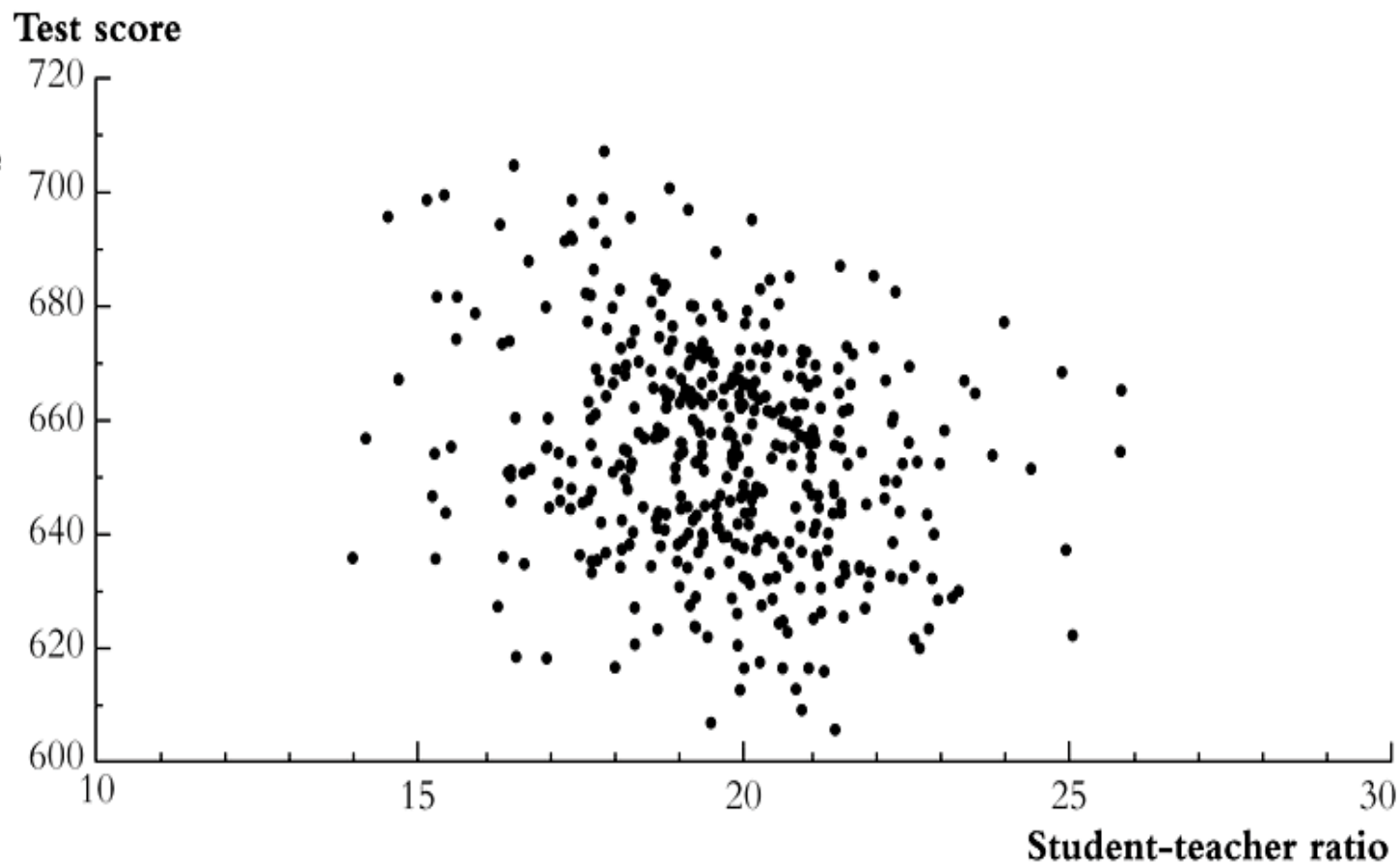
Policy question: What is the effect of reducing class size by one student per class? By 8 students/class? Do districts with smaller classes (lower STR) have higher test scores?

An initial look at the California test score data:

	Average	Standard Deviation	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

$$\text{Test score} = \beta_1 + \beta_2 \text{STR}$$

β_1 is the intercept

β_2 is the slope of the population regression line

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

- 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 we must estimate it using data.
- How can we **estimate** β_1 and β_2 from the data?

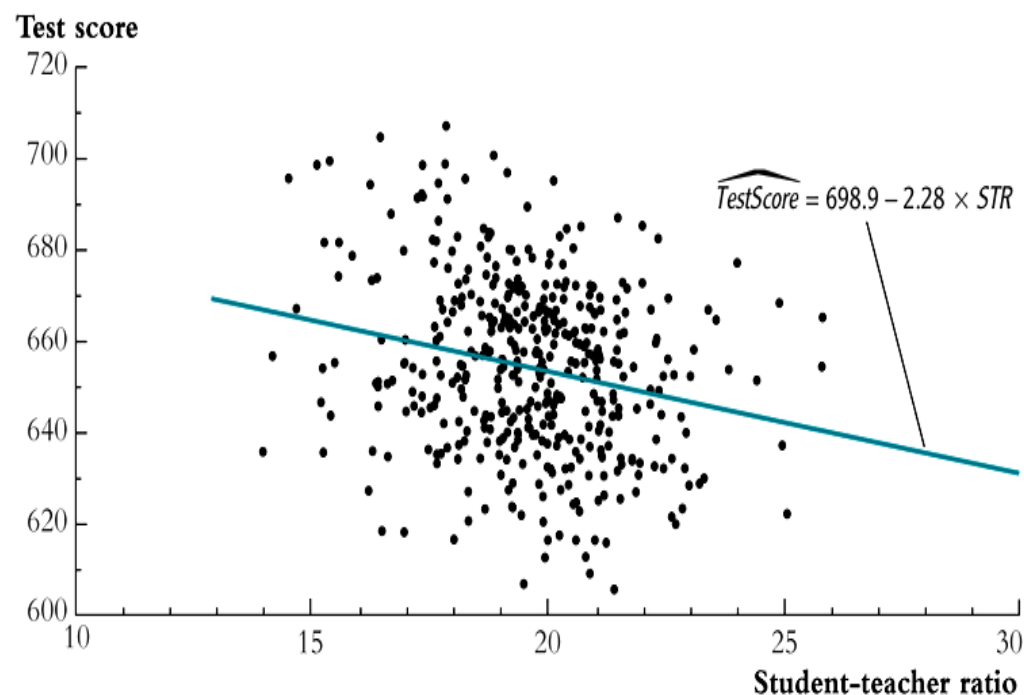
We will focus on the least squares ("ordinary least squares" or "OLS") estimator of the unknown parameters β_1 and β_2 , which solves

$$\text{Min}_{\beta=(\beta_1, \beta_2)} S(\beta) = (Y - X\beta)^T \cdot (Y - X\beta) = \sum_k (Y_k - \beta_1 - \beta_2 x_k)^2$$

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

4-2 SOME NOTATION AND TERMINOLOGY: EXAMPLE

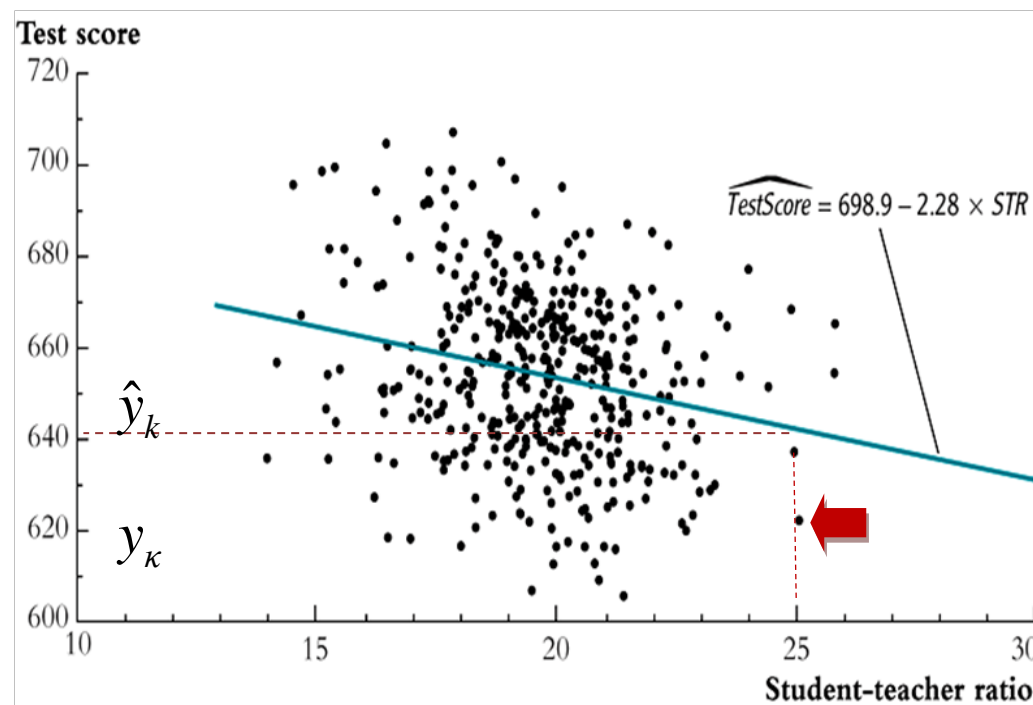
Predicted values & residuals:

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

predicted value: $= 698.9 - 2.28 \times 25 = 641.9$

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 $\hat{\beta}_2$?

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).

4-2 INTRODUCTION TO LINEAR MODELS

Classification of statistical tools for analysis and modeling

Explanatory variables	Response variable				
	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

4-3 LEAST SQUARES ESTIMATION IN MULTIPLE REGRESSION

Assume we have a linear model with no distribution hypothesis,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{Y} \text{ is } n \times 1, \mathbf{X} \text{ design matrix } n \times p \text{ and } \boldsymbol{\beta} \text{ the vector parameters } p \times 1$$

Let Y be a numeric response variable and $\boldsymbol{\varepsilon}$ the model error

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

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

$$\text{Min}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \cdot (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1 \dots n} (Y_k - \mathbf{x}_k^T \boldsymbol{\beta})^2 = \mathbf{Y}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y}$$

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. Matrix-vector notation is used in these notes.

4-3 LEAST SQUARES ESTIMATION

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

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

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

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

- 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., 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 $\hat{\mathbf{y}} = \hat{\mu} = X\hat{\beta}$.

4-3 LEAST SQUARES ESTIMATION

The range of $\mathbf{X}^T \mathbf{X}$ is equal to the range of \mathbf{X} . This attribute leads to two criteria that must be met in order to ensure $\mathbf{X}^T \mathbf{X}$ 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 \mathbf{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).

4-3.1 Geometric properties

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

$$\mathcal{R}(\mathbf{X}) = \left\{ \boldsymbol{\mu} \mid \boldsymbol{\mu} = \mathbf{X} \boldsymbol{\beta} \quad \boldsymbol{\beta} \in \mathcal{R}^p \right\} \subset \mathcal{R}^n.$$

And $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X} \hat{\boldsymbol{\beta}}$ the predictions once the least squares estimator of model parameters is calculated.

4-3 LEAST SQUARES ESTIMATION

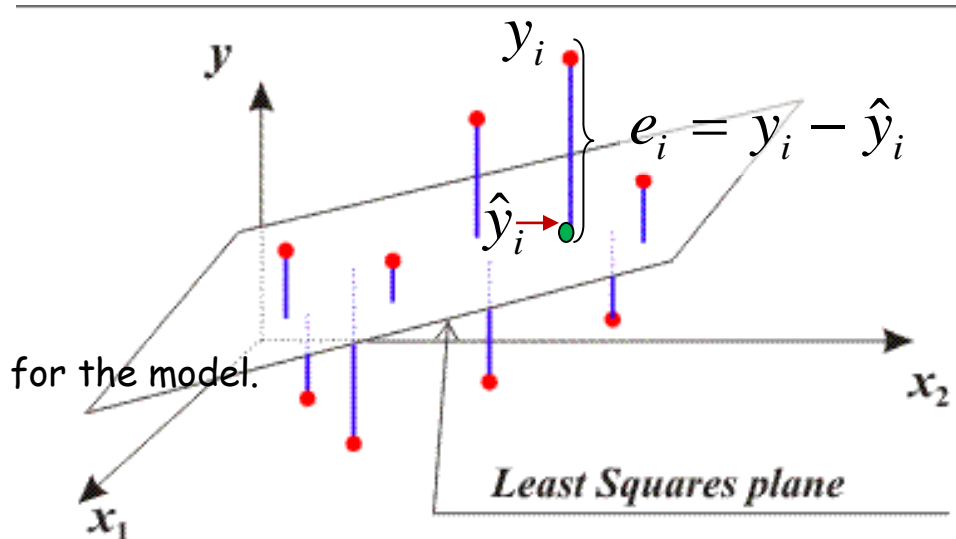
➔ Then it can be shown that $\hat{\boldsymbol{\mu}}$ is the orthogonal projection of \mathbf{Y} , it is unique, the projection operator being defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-} \mathbf{X}^T$, and it is called the hat matrix since, applied to \mathbf{Y} , it provides the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

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

Graphically,

Properties of the hat matrix:

- It depends solely on the X predictor variables.
- It is square, symmetric and idempotent: $\mathbf{H}\mathbf{H} = \mathbf{H}$.
- 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 assumed where

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

where $\mathbf{Y}, \boldsymbol{\mu}$ are $n \times 1$, \mathbf{X} $n \times p$ of p rank and $\boldsymbol{\beta}$ $p \times 1$ and the conditional distribution of unbiased errors are i.i.d. of constant variance and normally distributed: $\boldsymbol{\varepsilon} | \mathbf{X} \approx N(0, \sigma^2 \mathbf{I}_n)$ or $\mathbf{Y} | \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$.

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

➔ 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 σ^2 , denoted s^2 , is efficient (minimum variance) and its square root, s , is called the standard error of the model.

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

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 $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ an estimator of something, specifically the population parameter vector β , 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 β_j and sampling variance $V(\hat{\beta}_j) = \sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1}$ and we can test the simple hypothesis (i.e., make some inference):

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

But since it does not help that 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 $\hat{\beta}_j$, $\hat{V}(\hat{\beta}_j)$.

4-4 LEAST SQUARES ESTIMATION: INFERENCE

$\hat{V}(\hat{\beta}_j) = s^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1} \rightarrow SE(\hat{\beta}_j) = s \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$ and the ratio of $\hat{\beta}_j$ and $SE(\hat{\beta}_j)$ (standard error of j th parameter) is distributed as a Student t with $n-p$ degrees of freedom (since $\hat{\beta}_j$ and s^2 are independent).

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

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

4-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

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

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{Y}_{n \times 1}, \mathbf{X}_{n \times p} \text{ of range } p \text{ and } \boldsymbol{\beta}_{p \times 1}$$

where errors are unbiased and iid is normally distributed $\boldsymbol{\varepsilon} \approx N_n(0, \sigma^2 \mathbf{I}_n)$. **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.,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{Y}_{n \times 1}, \mathbf{X}_{n \times p} \text{ of range } p \text{ and } \boldsymbol{\beta}_{p \times 1}$$

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 $q \times p$ 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 .

4-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

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

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

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

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

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,

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

4-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

4-5.1 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)

4-5 HYPOTHESIS TESTING IN MULTIPLE REGRESSION

4-5.2 Confidence interval for model parameters

Individual confidence interval for β_i in OLS is summarized as:

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

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

$t_{n-p}^{\alpha/2}$ is the *Student t* for the bilateral confidence interval $1-\alpha$. The degrees of freedom 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 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.

4-6 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 :

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

➔ The square of the multiple correlation coefficient 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. $TSS = \sum_k (y_k - \bar{y})^2$ where $\bar{y} = \frac{1}{n} \sum_k y_k$ is the mean of the observed response data.

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

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:

$$\begin{aligned} \sum_k (y_k - \bar{y})^2 &= \sum_k ((y_k - \hat{y}_k) + (\hat{y}_k - \bar{y}))^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2 + 2 \sum_k (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) = \\ &\quad \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2 \end{aligned}$$

where

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

And:

➔ Or equivalently

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

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

4-6 MULTIPLE CORRELATION COEFFICIENT

4-6.1 Properties of the multiple correlation coefficient

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

4-6.2 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:

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

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

4-7 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.

$$H_0: \beta_2 = 0, \dots, \beta_p = 0.$$

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

4-8 PREDICTIONS AND INFERENCE

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

1. $E[\hat{Y}_k] = E[\mathbf{x}_k^T \hat{\boldsymbol{\beta}}] = \mathbf{x}_k^T \boldsymbol{\beta}$.
2. $V[\hat{Y}_k] = V[\mathbf{x}_k^T \hat{\boldsymbol{\beta}}] = \mathbf{x}_k^T V[\hat{\boldsymbol{\beta}}] \mathbf{x}_k = \sigma^2 \mathbf{x}_k^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_k = \sigma^2 h_{kk}$, 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 :

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

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

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 for a continuous response be

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{Y} \text{ is } n \times 1, \mathbf{X} \text{ is } n \times p \text{ of column rank } p \text{ and } \boldsymbol{\beta} \text{ is } p \times 1$$

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

$$\boldsymbol{\varepsilon} \approx \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \text{ or equivalently } \mathbf{Y} \approx \mathbf{N}_n(\mathbf{X} \boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

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

4-9 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 , $c_i = \frac{e_i}{s}$. It is not too serious when leverages show no big changes, since $V[e_i] = \sigma^2(1 - h_{ii})$.

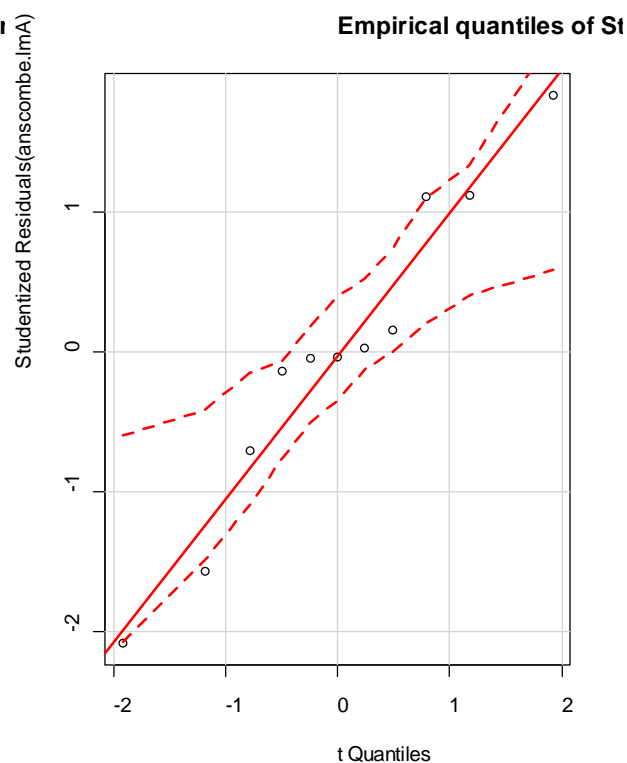
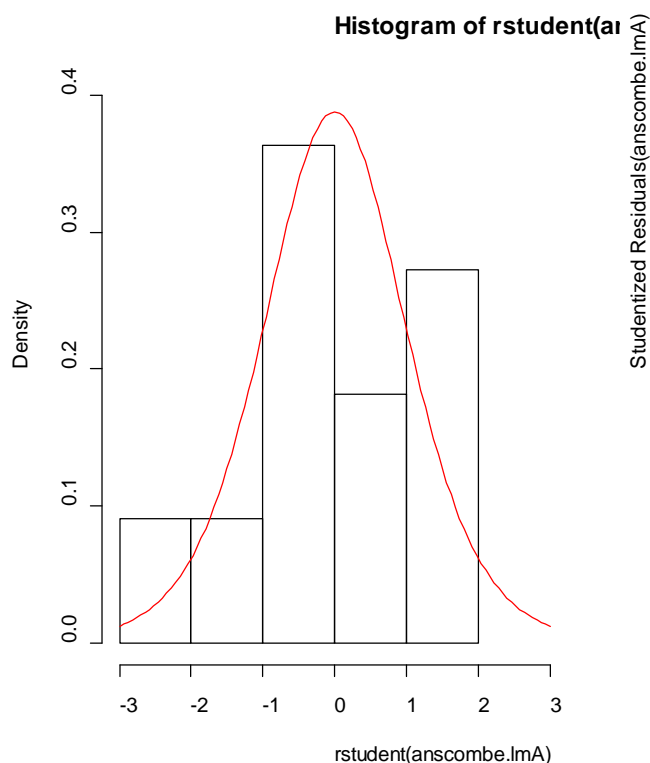
2. Standardized residual d_i is defined as the residual divided by its standard error: $d_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$.

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

- 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:



➡ 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)
```

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

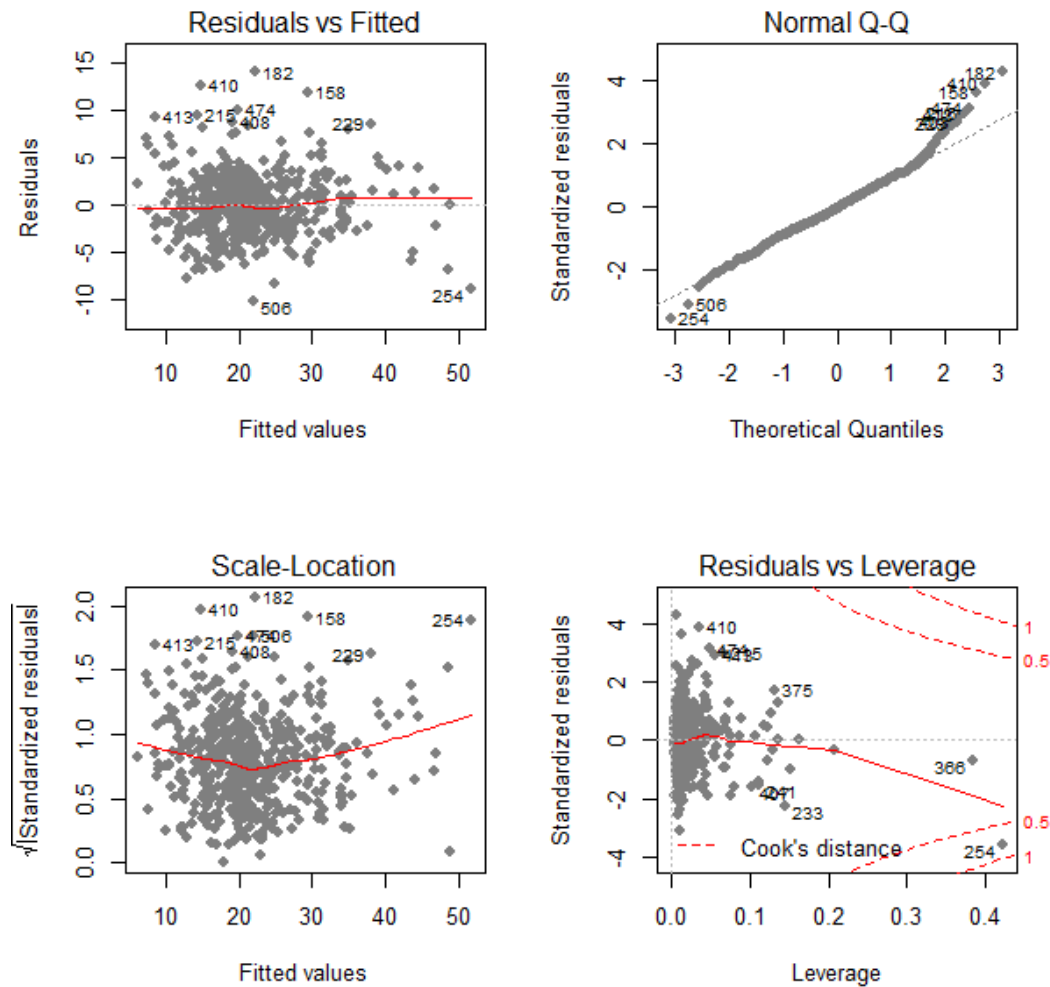
- When the normality assumption is not fully met then the comparisons of the hypothesis based on the Student or F-test are approximate and estimates are not efficient.
- Tests for normality assessment can be used. The Shapiro-Wilke test is one of my favorites, but there are lots of tests depending on 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.
- 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.
- Use `residualPlot(model)` method in package `car` for R.

➡ Scatterplots: e_i vs \hat{Y}_i , or d_i vs \hat{Y}_i or 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)  
par(mfrow=c(1,1))
```

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

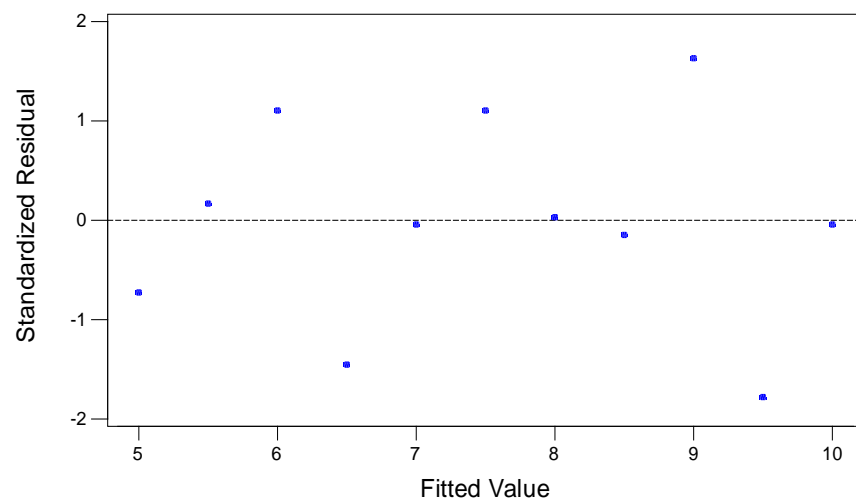


4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

Anscombe's data

Residuals Versus the Fitted Values

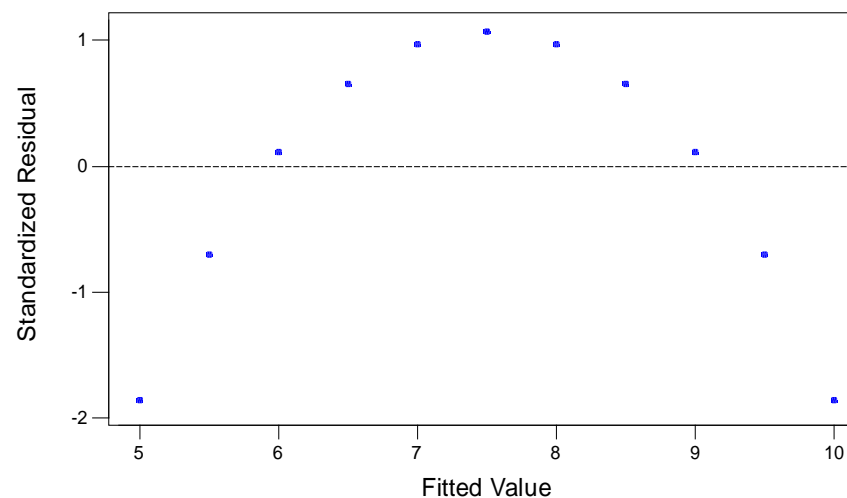
(response is YA)



Linear Hypothesis valid

Residuals Versus the Fitted Values

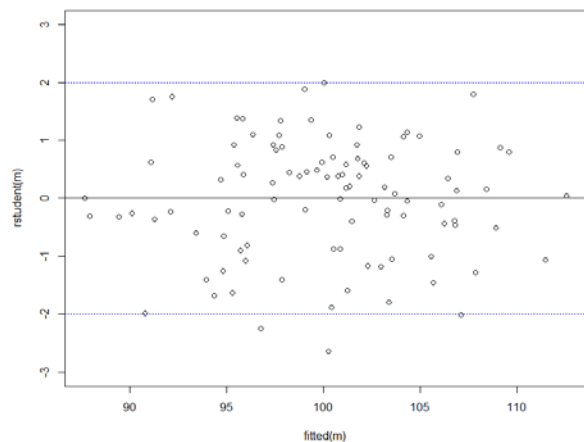
(response is YB)



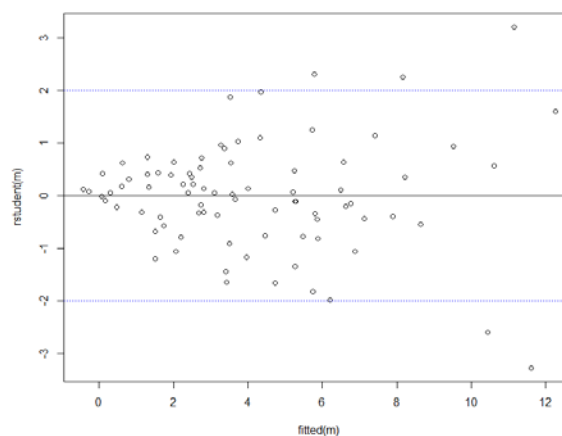
Linear Hypothesis fails

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

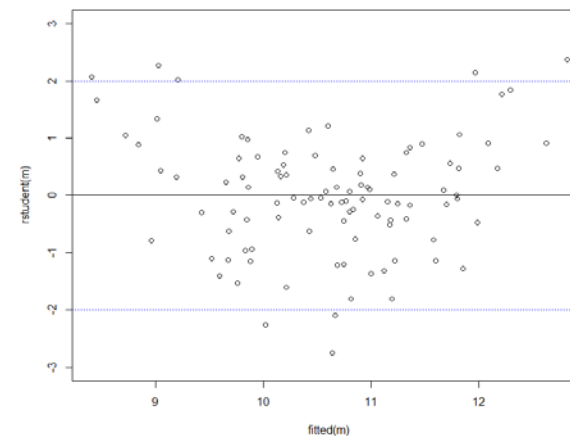
Linear hypothesis and homoskedastic



Heteroskedastic



Linear hypothesis rejected



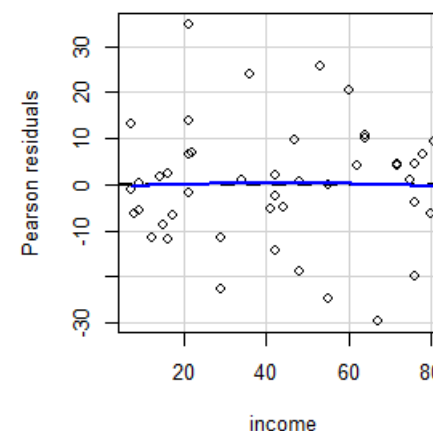
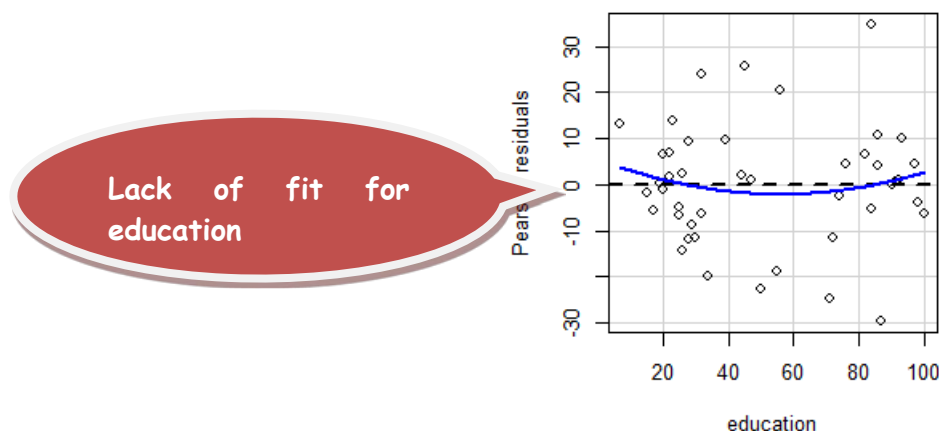
4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

➔ Scatterplots of residuals vs each regressor (except intercept)

The horizontal band indicates linearity satisfaction and homoskedastic hypothesis.

- Use `residualPlots(model)` method in package `car` for R. Example: Duncan data `lm

```
prestige~education+income,data=Duncan
````



- 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

4-9 RESIDUAL ANALYSIS: DIAGNOSTIC PLOTS

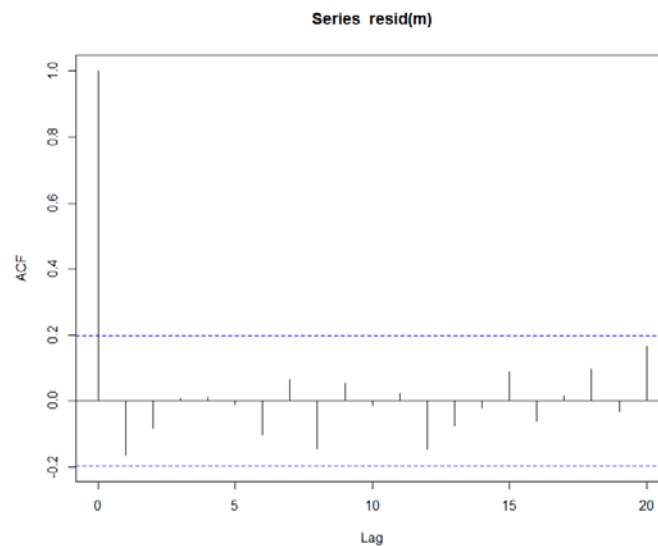
➔ Residual vs time/order or any omitted variable in the model suspected to affect hypothesis

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

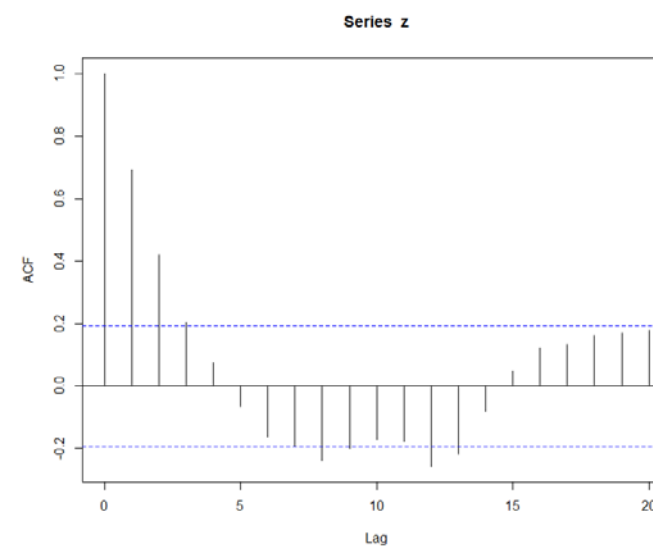
$$r(k) = \frac{\sum_i e_i e_{i+k}}{\sum_i e_k^2}, \text{ and contrasts for simultaneous hypothesis } r(j)=0 \text{ for } j>k.$$

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

No temporal dependency

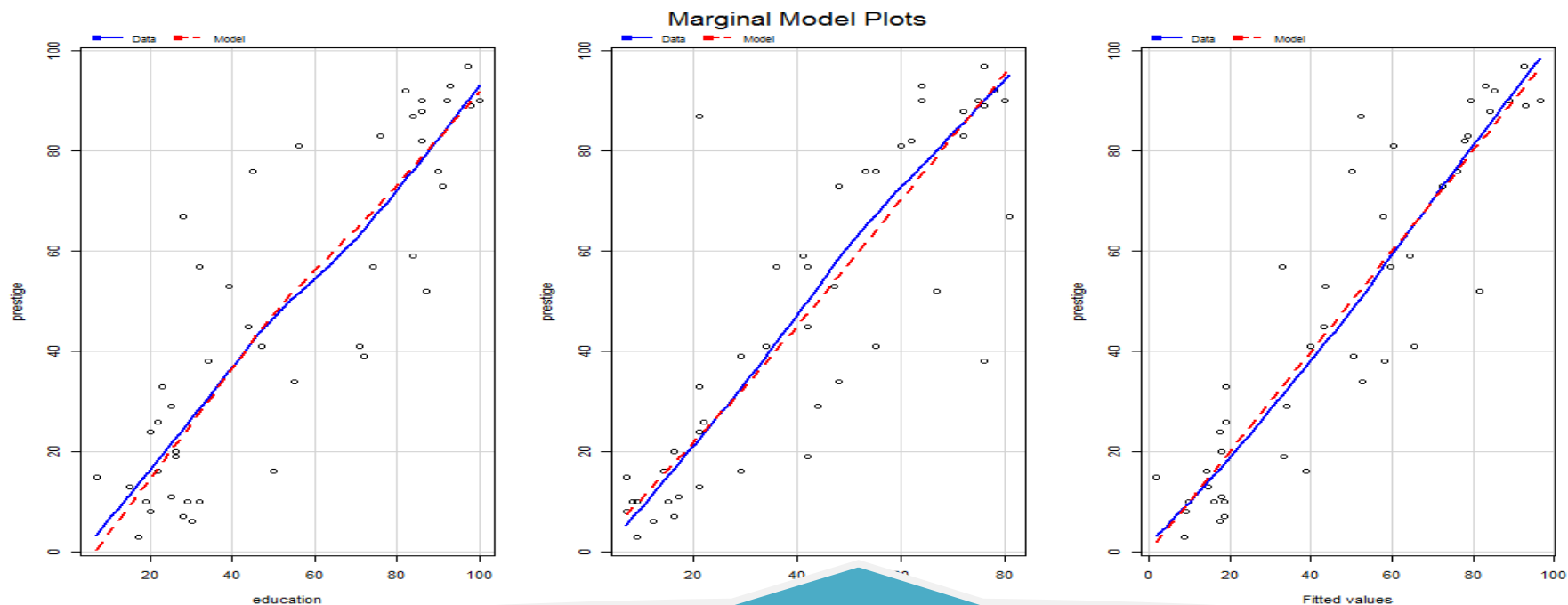


Temporal dependency



4-10 MODEL TRANSFORMATIONS ON Y OR X

- Use `marginalModelPlots(model)` method in package `car` for R. Lack of fit between data smoother and current model behavior for one variable indicates that transformation on selected regressor is needed.
- Use `poly(varname,2)` to model linear and quadratic terms on varname regressor.
- Use `poly(varname,3)` to model linear, quadratic and cubic terms on varname regressor.



There seem to be no problems in any of them!

4-10 MODEL VALIDATION: TRANSFORMATIONS

4-10.1 Box-Cox transformation of Y

- 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} :**
 - If $\lambda=1$, no transformation is necessary. Note that all of the Y_i must be positive.
 - λ may be by maximum likelihood in R: `boxcox(formula, data)` method in MASS library.
 - 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.

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

4-10 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 **

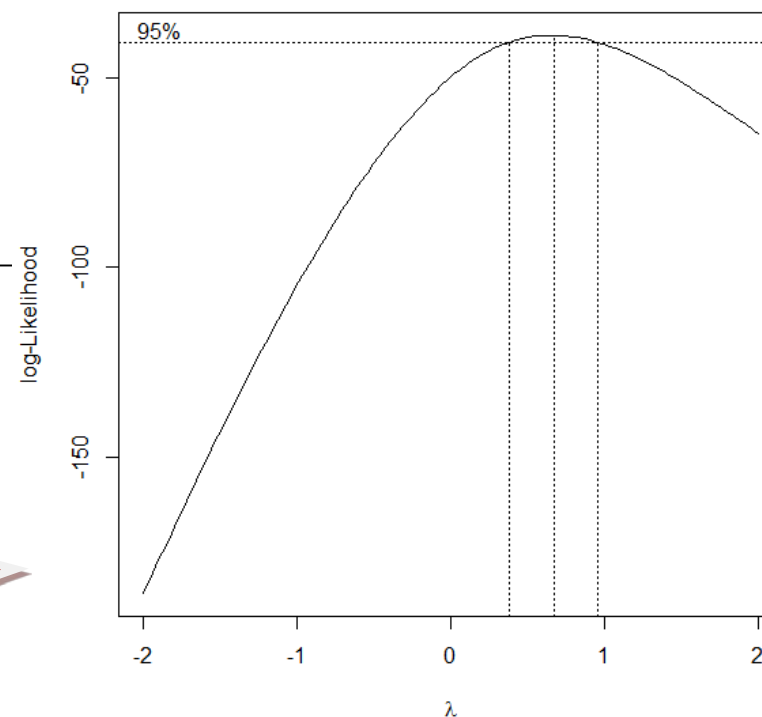
Transformation is
needed

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)
```

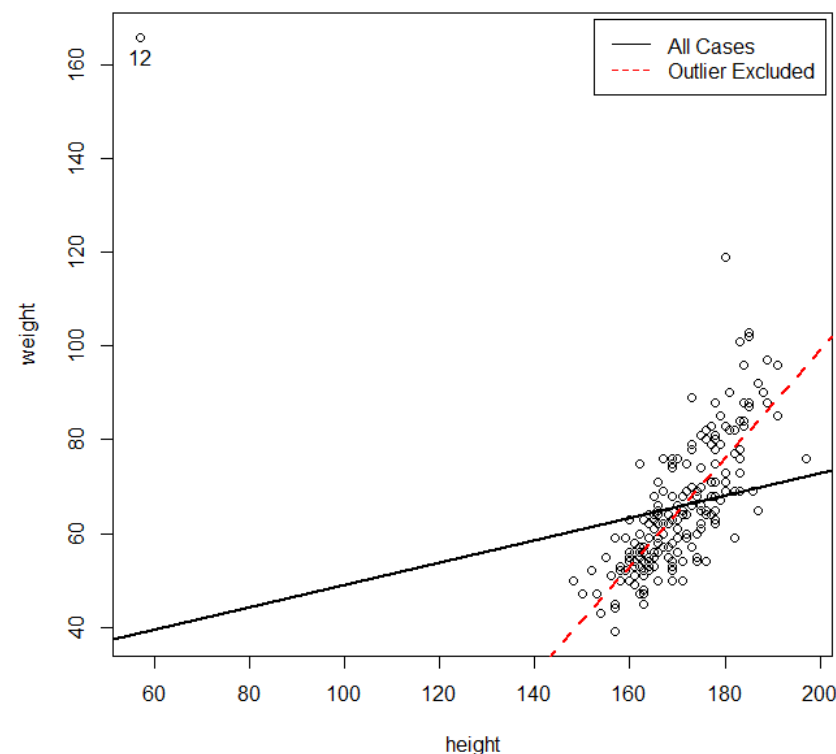


95% CI for λ excludes 1

4-11 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.



4-10 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.
- 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 , 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:

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

4-10 UNUSUAL AND INFLUENTIAL DATA

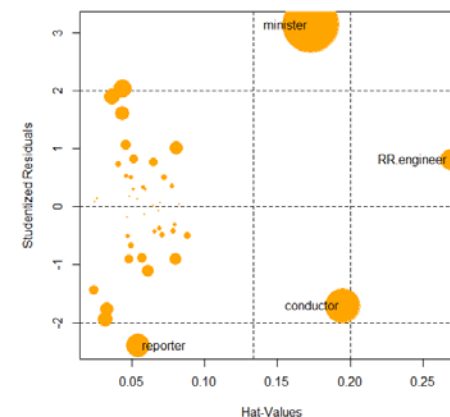
- Observation with leverage (large diagonal element of hat matrix, h_{ii})
 - An observation that has an unusual X value (i.e., it is far from the mean of X) has leverage on the regression line.
 - The further the outlier sits from the mean of X (either in a positive or negative direction), the more 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



4-10 UNUSUAL AND INFLUENTIAL DATA

4-11.1 *A priori* influential observations

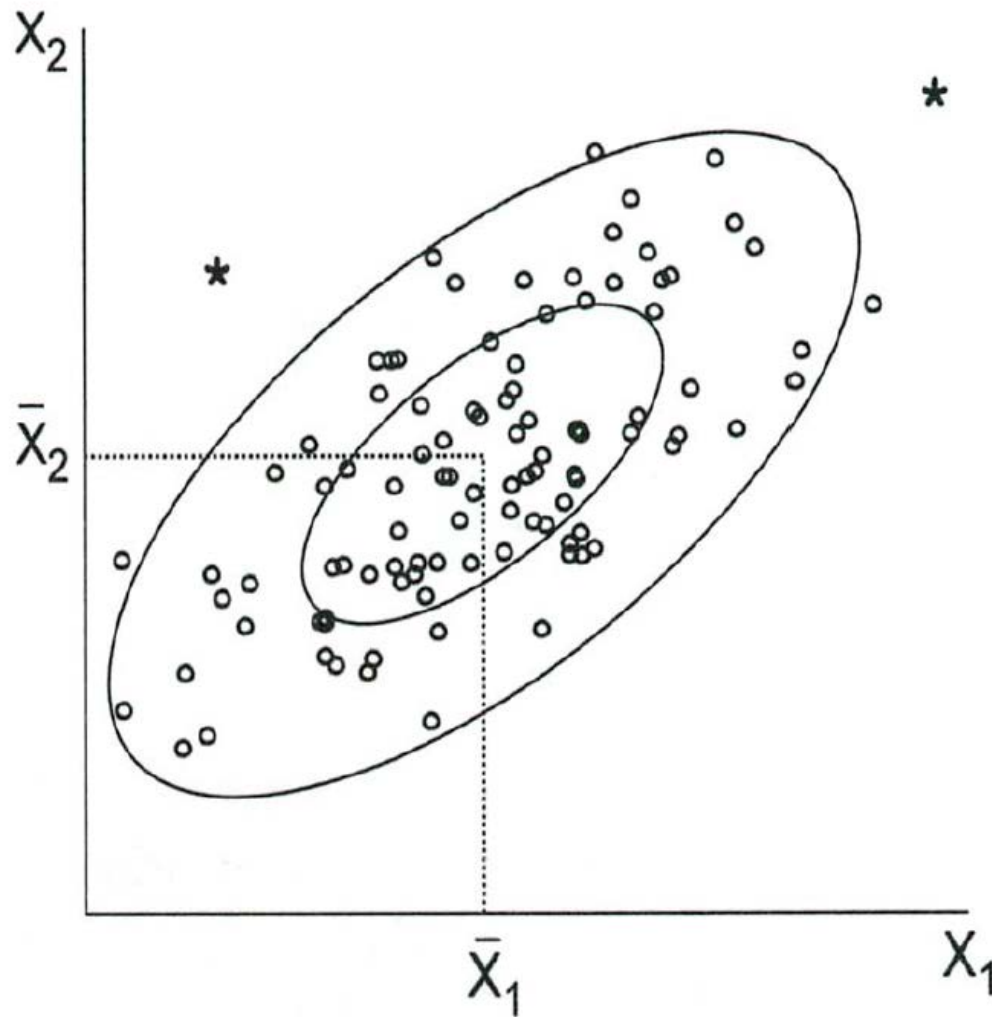
- Simple regression: An observation that has an unusual X value (i.e., it is far from the mean of \mathbf{X}) has leverage on the regression line. The further the outlier is from the mean of X (either in a positive or negative direction), the more leverage it has.
- Multiple regression: We must imagine a cloud of points defined by regressors in \mathbf{X} (each column on an axis) and a center of gravity for those points.

Heterogeneous points \mathbf{x} ($\mathbf{x} \in \mathbb{R}^p$) 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 $\bar{h} = \frac{\sum_i h_{ii}}{n} = \frac{p}{n}$. Belsley et al. show that atypical values for leverage observations are those that meet the cut-off: $h_{ii} > 2\bar{h}$.
- ➡ This cut-off is not useful when considering big data sets: the cut-off has to be increased to $h_{ii} > 3\bar{h}$.

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 h_i).

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

4-10 UNUSUAL AND INFLUENTIAL DATA

4-11.2 *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{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:

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

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!!!)

4-10 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 :

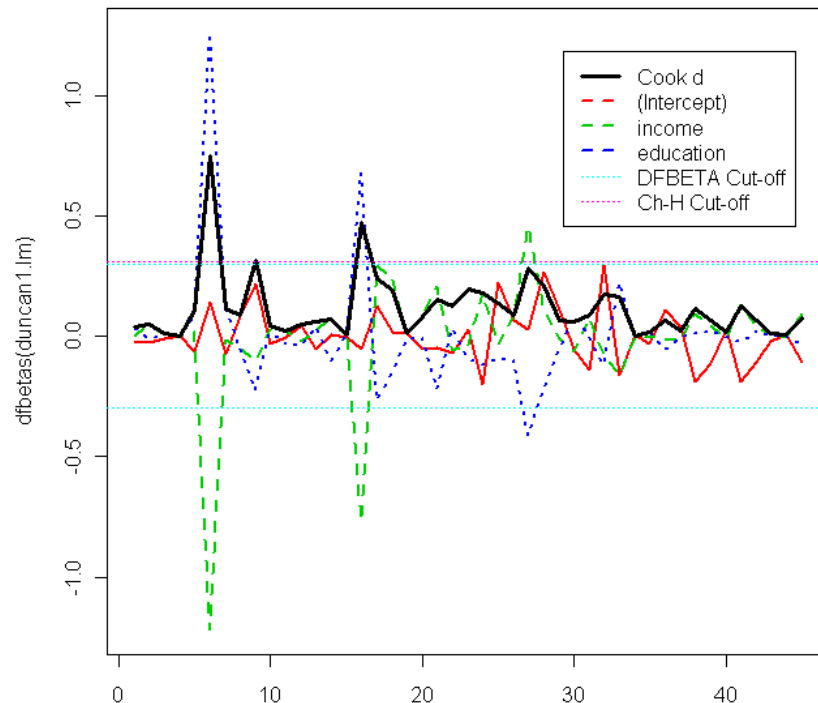
$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{p s^2} = \left(\frac{e_i}{s \sqrt{1 - h_{ii}}} \frac{1}{\sqrt{p}} \right)^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

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 $H_i : \beta = \hat{\beta}_{(-i)}$.
- There is no significance test for D_i but a commonly used cut-off is the one proposed by Chatterjee and Hadi (1988) that defines influential observations as those that satisfy $D_i > 4/(n-p)$.
- For large samples, the Chatterjee and Hadi cut-off does not work and, as a rule of thumb, $D_i > 0.5$ are suspected to be influential data and $D_i > 1$ are considered influential (R criteria).

4-10 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 D_i (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)

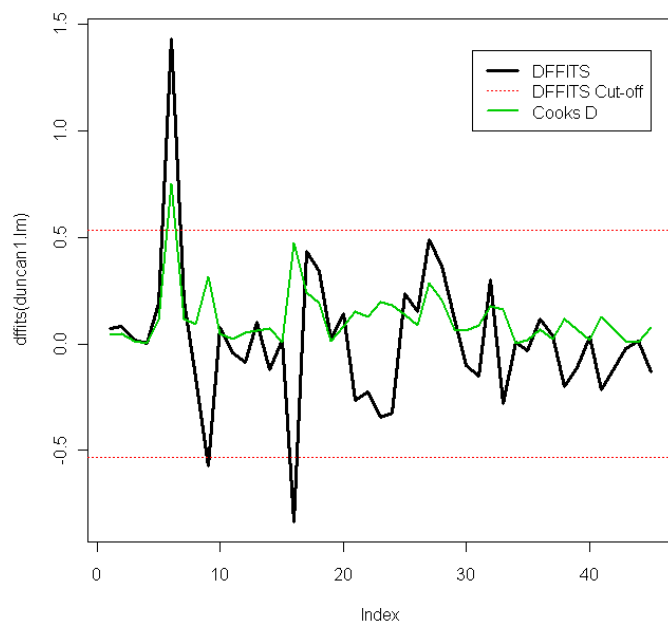
>llegenda<-c("Cook d",
names(coef(duncan1.lm)), "DFBETA Cut-
off", "Ch-H Cut-off")

> 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))
```


4-10 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 $|DFFITS_i| > 2\sqrt{\frac{p}{n-p}}$.
Use `dffits(model)` in R.

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



```
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(duncan1)-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))
```

4-12 BEST MODEL SELECTION

The best regression equation for Y given the regressors (X_1, \dots, 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, \dots, 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.

4-11 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_a^2 is useful.
2. Stability of the standard error of regression estimate. Estimation of σ^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) $AIC = 2(-\ell(\hat{\beta}, \mathbf{y}) + p)$. Models with lower values of C_p or AIC indicator are preferred.
 - ➡ Some authors strongly recommend BIC (Bayesian Information Criteria) Schwartz criteria $BIC = -2\ell(\hat{\beta}, \mathbf{y}) + p \log n$ 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.

4-11 BEST MODEL SELECTION

4-12.1 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 X_j , 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.*

4-11 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
```

```
>
```