Quality Data Analysis

Linear models - part 2

Bianca Maria Colosimo – biancamaria.colosimo@polimi.it

Sources: * chapter 3 Alwan + chapter 10 “Applied Statistics and Probability for Engineers” –D.C. Montgomery and G.C. Runger - John Wiley & Sons 2nd edition

Non-linear trend processes

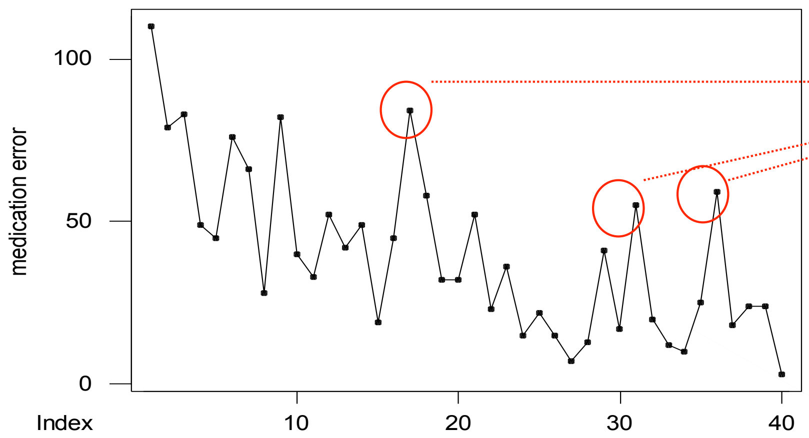
Mistakes in the medication of medical center. Target is continuous improvement!

- Missing medication
- Wrong medication
- Incorrect dosage

(mederror.dat)

Total number of errors in 40 weeks:

110	79	83	49	45	76	66	28	82	40
33	52	42	49	19	45	84	58	32	32
52	23	36	15	22	15	7	13	41	17
55	20	12	10	25	59	18	24	24	3



Unexpected events?

Non-linear trend processes

Curvilinear trend?

- Independent variables (trend variables): t^2 or $1/t$ (in principle an infinite set of possibilities)
- t, t^2, $1/t$: any possible combination of the three independent variables in the regression model: 2^p-1 possible regression models
- Pay attention: multiple linear regression (more than one regressors, e.g., t, t^2, $1/t$);

Explore all the possible models

Methods to search the "best" model:

- Forward selection
- Backward elimination
- Stepwise regression

Pay attention to overfitting !

Stepwise regression

Forward selection:

Sequential procedure – one variable is added at a time. At each step, the variable that provides the better contribution to the "fitting" is selected. Once the variable is added, it cannot be removed in the subsequent steps.

Backward elimination:

Sequential procedure – one variable is removed at each time. Starting from a model that contains all the possible variables, at each step we remove the variable that is "less useful" to explain the data variability. Once the variable is removed, it is never reincluded in the following steps.

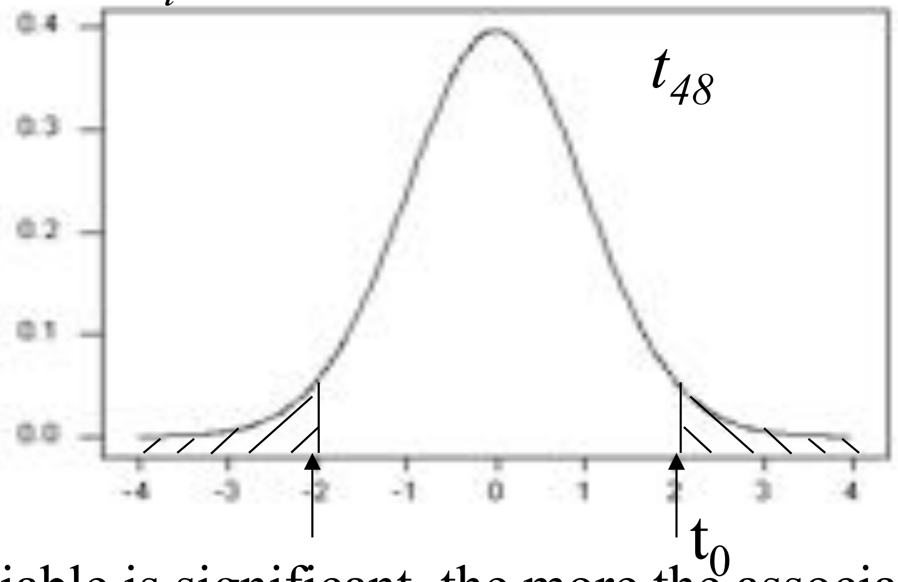
Stepwise selection:

It combines forward e backward. We start as a forward selection but each time that a variable is added a backward step is carried out to check wether a variable has to be removed. The procedure stops when no regressorhas to be included in the model and no regressor has to be removed.

We need to specifiy Alpha to enter- Alpha to remove (usually =)

Some “commens” on the stepwise regression

$H_0 : \beta_i = 0 \qquad t_0 = \frac{b_i - 0}{s_{b_i}} \sim t_{n\text{-}K}$ t di Student a n-K gradi di libertà



1. The more a variable is significant, the more the associate parameter has a small associated p-value ($|t_0|$ large) :
2. If β_i is significantly greater than zero, (p-value<alpha to enter): associated regressor should be included in the model
3. If β_i is not significantly greater than zero (p-value>alpha to remove): assoiated regressor should be removed from the model

Regression Analysis: medication error versus t

```
The regression equation is
medication error = 70.0 - 1.47 t
Predictor         Coef     SE Coef        T        P
Constant        70.042       6.069    11.54    0.000
t              -1.4716      0.2579    -5.71    0.000
```

Regression Analysis: medication error versus t2

```
The regression equation is
medication error = 56.9 - 0.0308 t2
Predictor         Coef     SE Coef        T        P
Constant        56.932       4.901    11.62    0.000
t2           -0.030816    0.006642    -4.64    0.000
```

Regression Analysis: medication error versus 1/t

```
The regression equation is
medication error = 29.5 + 97.4 1/t
Predictor         Coef     SE Coef        T        P
Constant        29.460       3.582     8.22    0.000
1/t              97.37       17.80     5.47    0.000
```

Step 1

Forward:
All the p-value's< alpha to enter (15%)
All the variables should be included:
The one with max |T| (min p-value) is chosen
Backward: the added regressor has not to be removed

Step 2

Regression Analysis: medication error versus t, t2

The regression equation is

medication error = 84.2 - 3.50 t + 0.0494 t2

Predictor	Coef	SE Coef	T	P
Constant	84.214	9.023	9.33	0.000
t	-3.496	1.015	-3.44	0.001
t2	0.04938	0.02401	2.06	$\boxed{0.047}$

Regression Analysis: medication error versus t, 1/t

The regression equation is

medication error = 53.1 - 0.948 t + 58.4 1/t

Predictor	Coef	SE Coef	T	P
Constant	53.066	8.047	6.59	0.000
t	-0.9484	0.2964	-3.20	0.003
1/t	58.45	20.07	$\boxed{2.91}$	$\boxed{0.006}$

Forward:
All the p-value's< alpha to enter
Both t2 and 1/t could be included: I choose the one with max |T| (min p-value)

Step 2b

Regression Analysis: medication error versus t, 1/t

```
The regression equation is
medication error = 53.1 - 0.948 t + 58.4 1/t
```

Predictor	Coef	SE Coef	T	P
Constant	53.066	8.047	6.59	0.000
t	-0.9484	0.2964	-3.20	0.003
1/t	58.45	20.07	2.91	0.006

Backward:
All the p-value’s< alpha to remove
I do not reject any regressor
(If we find a set of regressors whose p-value> alpha to remove, we can remove the one with max p-value)

PASSO 2c

Regression Analysis: medication error versus t, 1/t, t2

```
The regression equation is

medication error = 58.5 - 1.51 t + 51.7 1/t + 0.0122 t2
```

Predictor	Coef	SE Coef	T	P
Constant	58.53	15.65	3.74	0.001
t	-1.509	1.403	-1.08	0.289
1/t	51.68	26.19	1.97	0.056
t2	0.01221	0.02982	0.41	$\boxed{0.685}$

Forward:
p-value> alpha to enter: we do not add t^2

MODELLO FINALE:

STOP

Regression Analysis: medication error versus t, 1/t

```
The regression equation is

medication error = 53.1 - 0.948 t + 58.4 1/t
```

Predictor	Coef	SE Coef	T	P
Constant	53.066	8.047	6.59	0.000
t	-0.9484	0.2964	-3.20	0.003
1/t	58.45	20.07	2.91	0.006

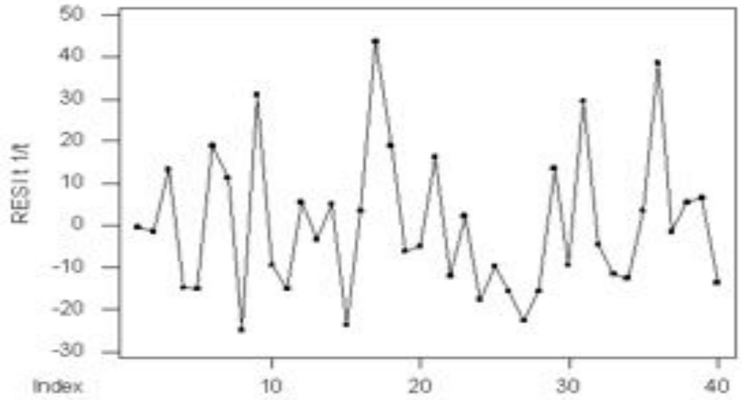
Stepwise

Stepwise Regression: medication error versus t, t2, 1/t

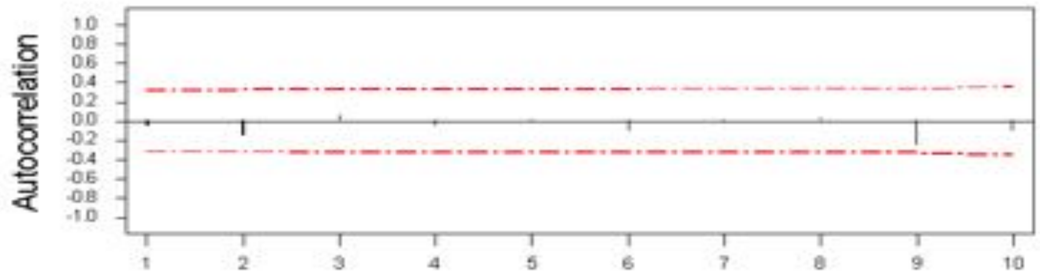
```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15
 Response is medicati on  3 predictors, with N =   40

    Step           1        2
Constant       70.04    53.07

t              -1.47    -0.95
T-Value        -5.71    -3.20
P-Value        0.000    0.003

1/t                        58
T-Value                  2.91
P-Value                 0.006

S               18.8     17.2
R-Sq           46.14    56.18
R-Sq(adj)      44.72    53.81
```

$$\boxed{\text{Fitted med err}_t = 53.07 - 0.95t + 58(1/t)}$$

Diagnostick check

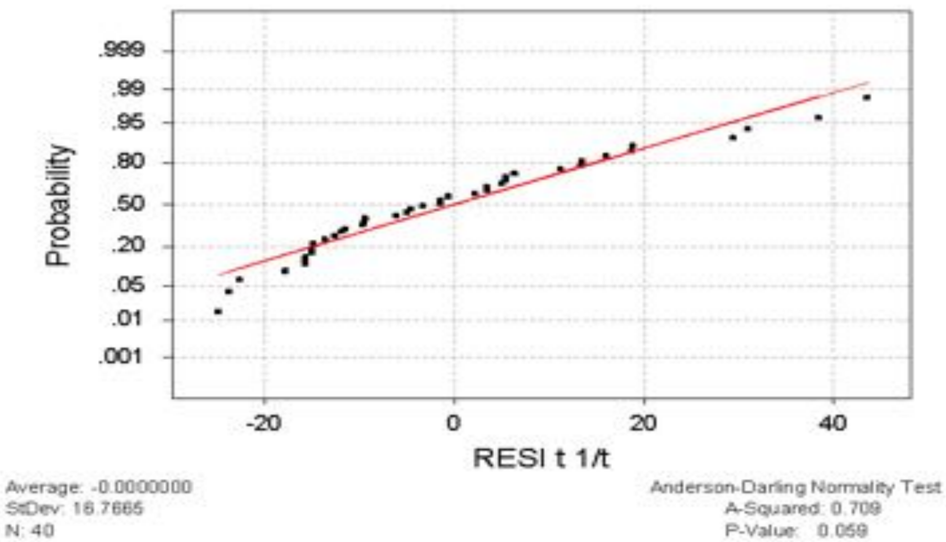


Autocorrelation Function for RESI t 1/t



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	-0.05	-0.34	0.12	8	0.03	0.17	2.00
2	-0.15	-0.96	1.15	9	-0.25	-1.51	5.37
3	0.06	0.36	1.31	10	-0.10	-0.57	5.93
4	-0.05	-0.31	1.43				
5	-0.01	-0.04	1.43				
6	-0.10	-0.63	1.96				
7	-0.01	-0.05	1.96				

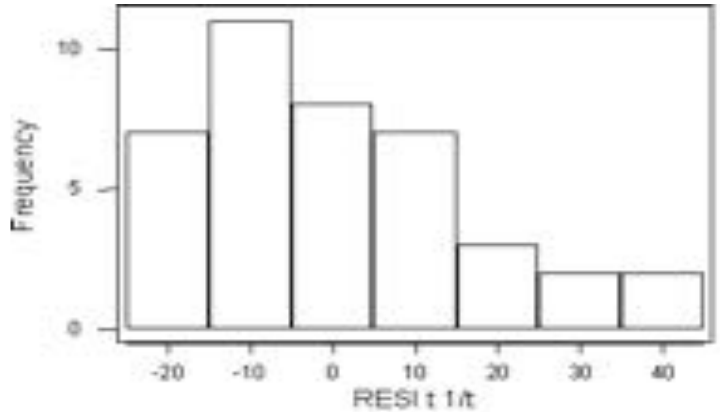
Normal Probability Plot

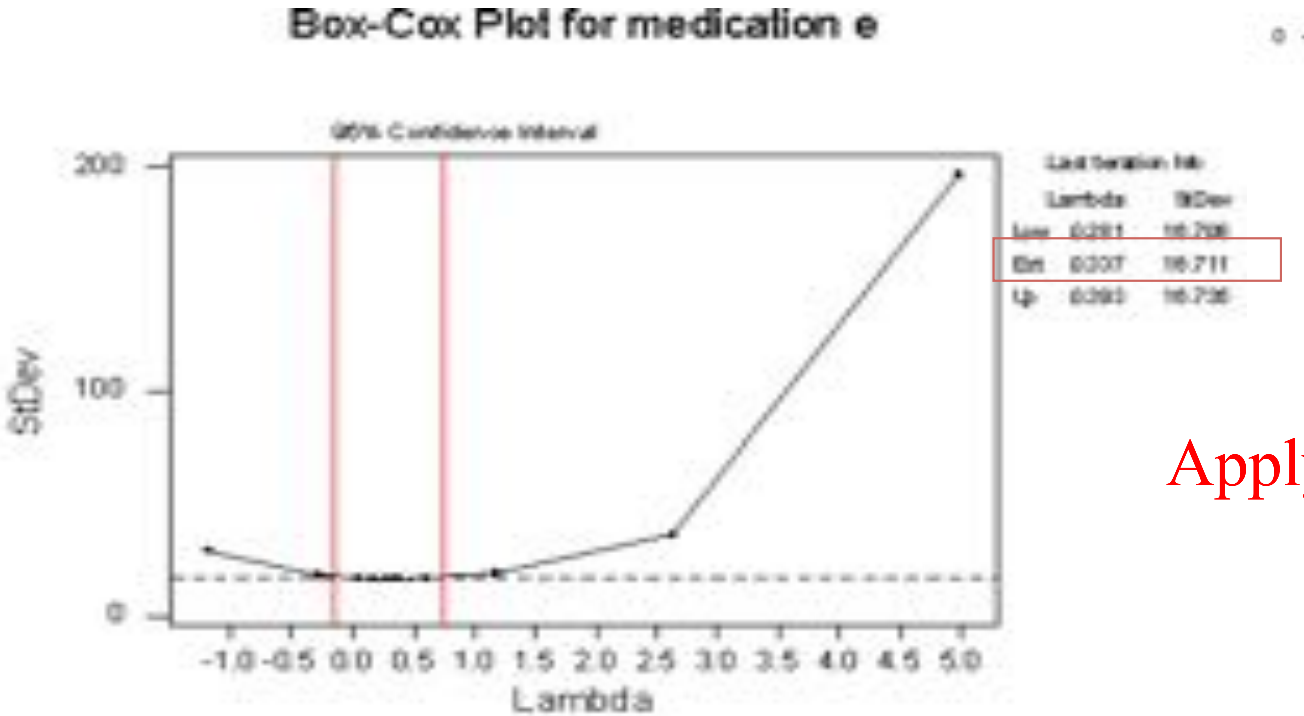


p-value=0.059 acceptable with α=5%
Just as teaching example, assume we used α=10% for A-D test:
TRASFORM THE DATA
(not only the residuals)

Looking for data transformation:

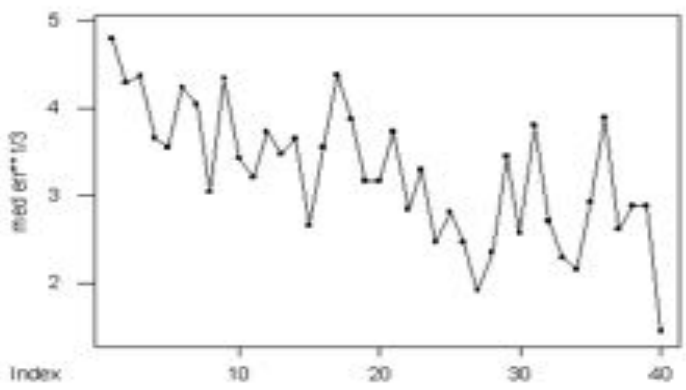
Positive skewness: suggested transformations (λ<1):
$x^{1/2}$, $x^{1/3}$, ln x, 1/x





Apply: $x^{1/3}$

Transformed data:



Linear Trend?

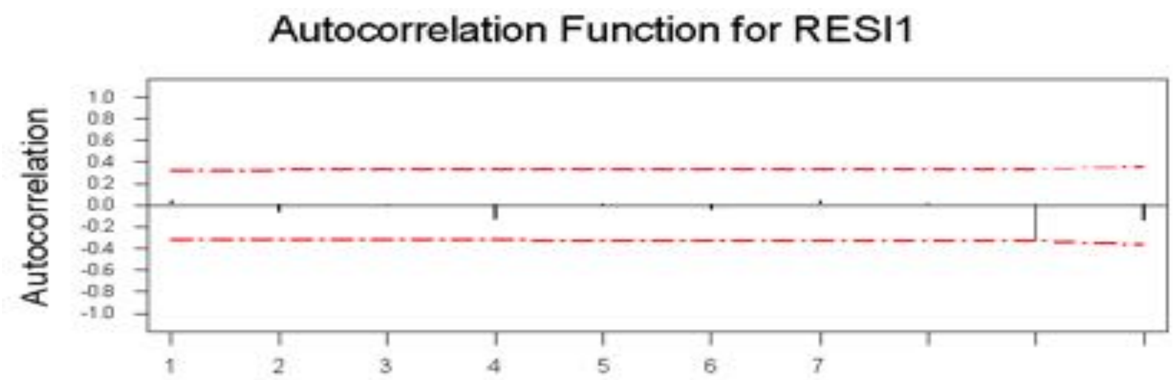
Stepwise Regression: med err1/3 versus t, t2, 1/t**

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is med err* on  3 predictors, with N =   40

  Step           1
Constant     4.172
t          -0.0448
T-Value      -5.87
P-Value      0.000
S            0.557
R-Sq         47.58
R-Sq(adj)    46.20
```

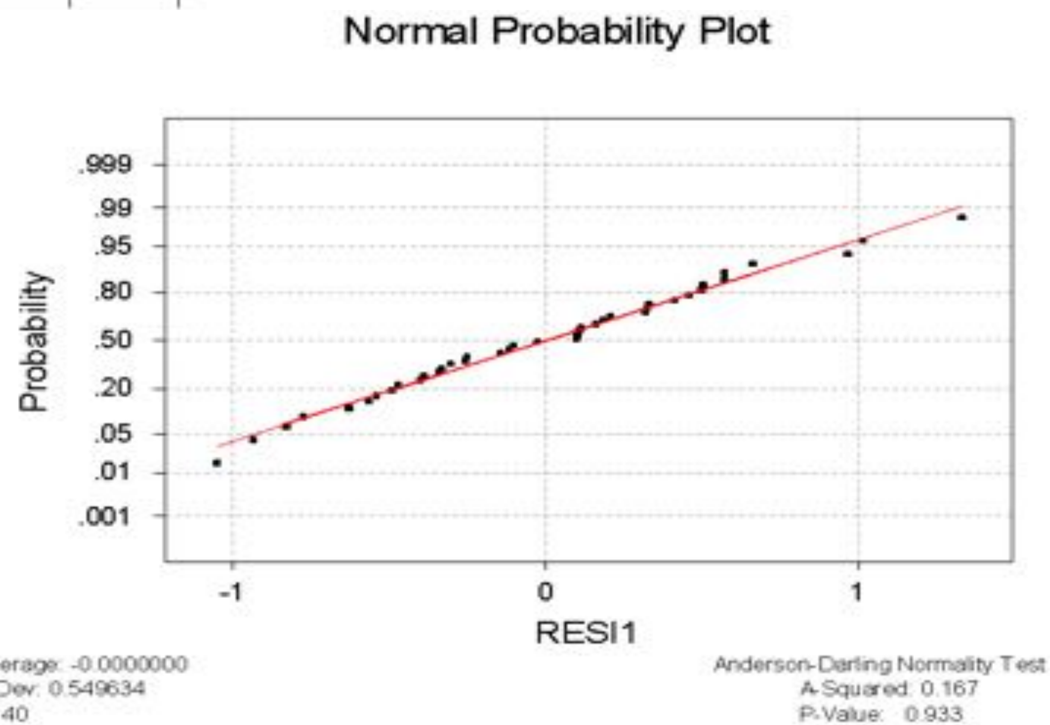
$$(\text{Fitted med err}_t)^{1/3}=4.172-0.044795\ t$$

$$\text{Fitted med err}_t=(4.172-0.044795\ t)^3$$

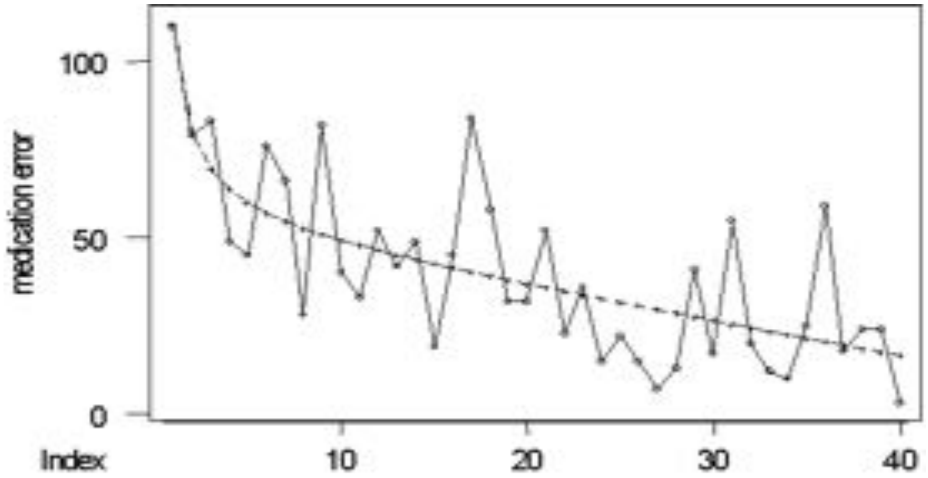
New residuals:



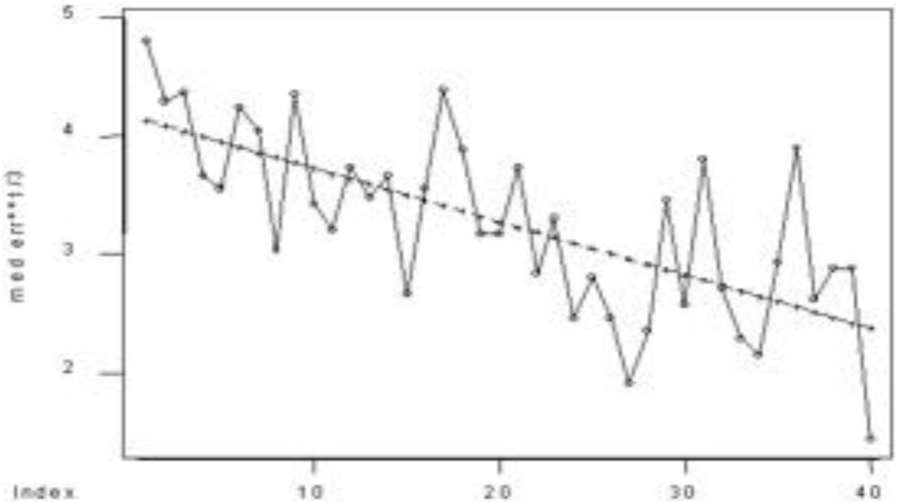
Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.05	0.30	0.10	8	0.01	0.07	1.40
2	-0.07	-0.45	0.32	9	-0.33	-2.03	7.34
3	-0.01	-0.08	0.33	10	-0.15	-0.82	8.54
4	-0.14	-0.86	1.21				
5	-0.01	-0.07	1.22				
6	-0.04	-0.25	1.30				
7	0.04	0.27	1.39				

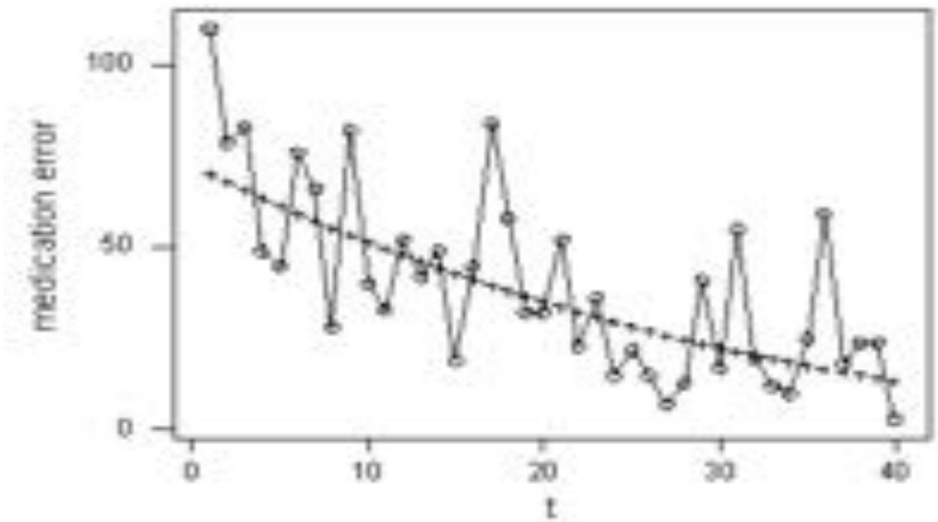


1st model



2° model





Model Checking*

More systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.

Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 and Cross-validation (CV).

*The Elements of Statistical Learning
Libro di Jerome H. Friedman, Robert Tibshirani e Trevor Hastie

Other Considerations in the Regression Model

Qualitative Predictors

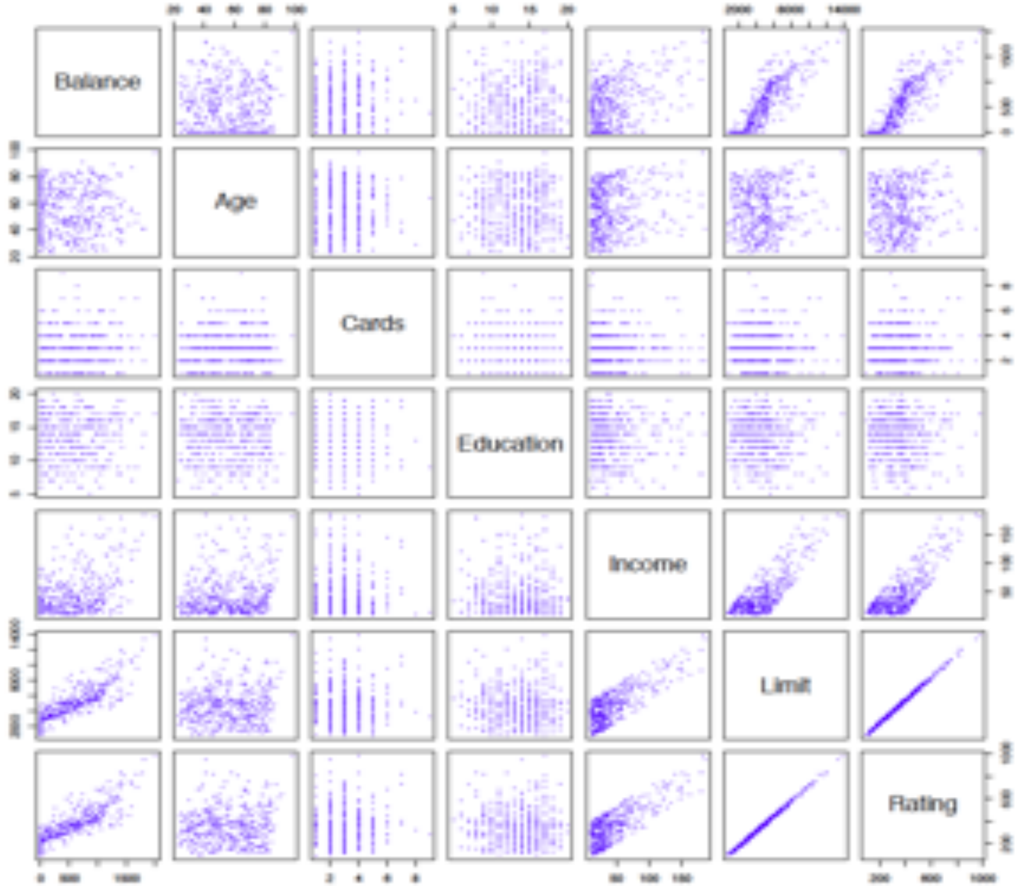
- Some predictors are not *quantitative* but are *qualitative*, taking a discrete set of values.
- These are also called *categorical* predictors or *factor variables*.
- See for example the scatterplot matrix of the credit card data in the next slide.

  In addition to the 7 quantitative variables shown, there are four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Slides taken from https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

*Source "The Elements of Statistical Learning" By Jerome H. Friedman, Robert Tibshirani e Trevor Hastie

Credit Card Data



https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Qualitative predictors with more than two levels — continued.

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Extensions of the Linear Model

Removing the additive assumption: *interactions* and *nonlinearity*

Interactions:

- In our previous analysis of the `Advertising` data, we assumed that the effect on `sales` of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

states that the average effect on `sales` of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.

Interactions — continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of $100,000, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- The results in this table suggests that interactions are important.
- The p-value for the interaction term `TV`×`radio` is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts `sales` using `TV` and `radio` without an interaction term.

- This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in `sales` that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, `TV` and `radio`) do not.
- The *hierarchy principle*:

  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

  - The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
  - Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Other interesting predictors

How to model a response that depends by the day of the week?

How to model autocorrelated processes?

Proofs

Properties of normal equations:

Let:

$$\hat{y}_t = b_0 + b_1 x_t$$

$$e_t = y_t - \hat{y}_t$$

$$P1) \quad \sum_{t=1}^{n} e_t = 0 \quad \text{(from the first normal equation)}$$

$$P2) \quad \sum_{t=1}^{n} x_t e_t = 0 \qquad \text{(from the second normal equation)}$$

$$P3) \quad \sum_{t=1}^{n} e_t \hat{y}_t = \sum_{t=1}^{n} e_t \left(b_0 + b_1 x_t\right) = 0$$

Back

Dim (result at p. 34):

$$\sum_{t=1}^{n}(y_t-\overline{y})^2=\sum_{t=1}^{n}(y_t-\hat{y}_t+\hat{y}_t-\overline{y})^2=$$

$$=\sum_{t=1}^{n}(y_t-\hat{y}_t)^2+\sum_{t=1}^{n}(\hat{y}_t-\overline{y})^2+2\underbrace{\sum_{t=1}^{n}(y_t-\hat{y}_t)(\hat{y}_t-\overline{y})}_{0}$$

because

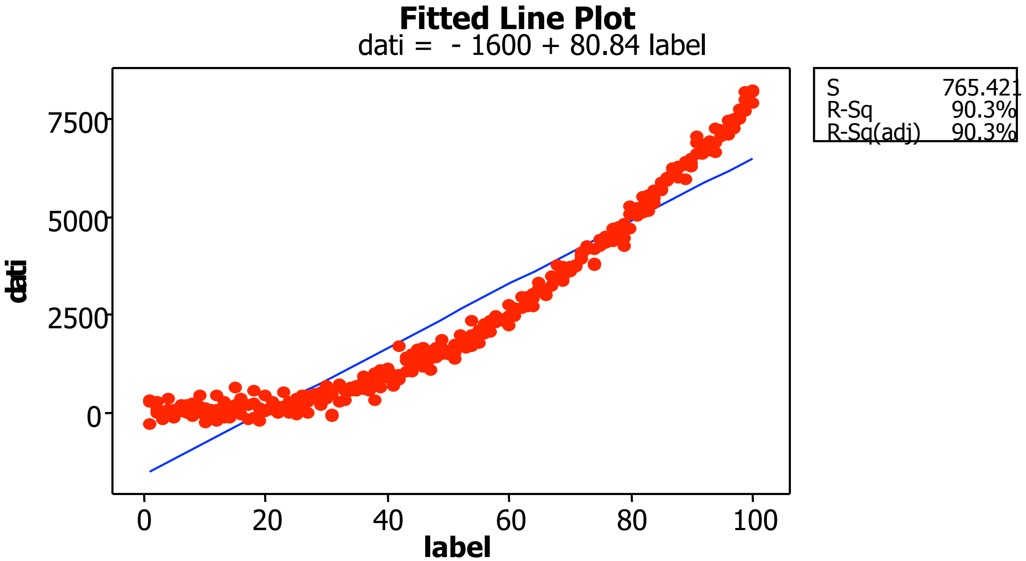
$$\sum_{t=1}^{n}(y_t-\hat{y}_t)(\hat{y}_t-\overline{y})=\sum_{t=1}^{n}e_t\hat{y}_t-\overline{y}\sum_{t=1}^{n}e_t=0$$

$P1$) e $P3$)
Normal equations

Back

Extra

Lack of fit test

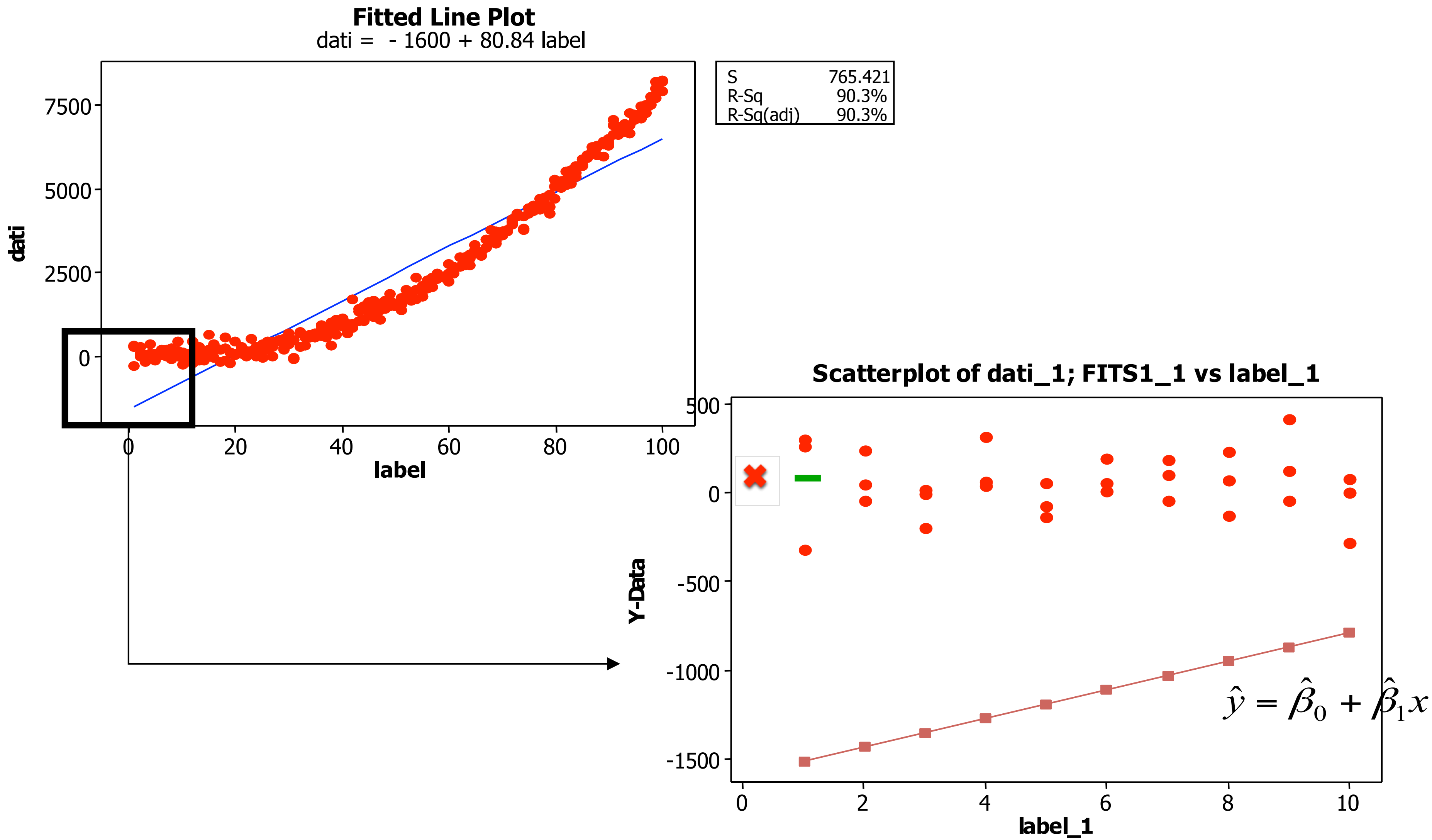


H_0: the simple linear regression model is correct

H_1: … is not correct

$$SS_E = SS_{PE} + SS_{LOF}$$

"pure error" "lack of fit"


Fitted Line Plot
dati = - 1600 + 80.84 label
S 765.421
R-Sq 90.3%
R-Sq(adj) 90.3%
dati
label
Scatterplot of dati_1; FITS1_1 vs label_1
Y-Data
label_1
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$SS_E = SS_{PE} + SS_{LOF}$$

In order to compute SS_{PE} we need to have repeated observations for at least one regressor

$y_{11}, y_{12}, \ldots, y_{1n_1}$ livello x_1 $\longrightarrow$ $\sum_{u=1}^{n_1} (y_{1u} - \bar{y}_1)^2$

Contributo alla somma dei quadrati PE relativa al livello1

$y_{21}, y_{22}, \ldots, y_{2n_2}$ livello x_2

$\ldots$ $\ldots$

$y_{m1}, y_{m2}, \ldots, y_{mn_m}$ livello x_m $\longrightarrow$ $\sum_{u=1}^{n_m} (y_{mu} - \bar{y}_m)^2$

$$SS_{PE} = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2$$

$$n_{pe} = \sum_{i=1}^{m} (n_i - 1) = n - m$$

$$SS_{LOF} = SS_E - SS_{PE}$$

$$n - K - n_{pe} = n - K - n + m = m - K$$

$$F_0 = \frac{SS_{LOF}/(m-K)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}}$$

Reject H0 (the assumed model is correct) if MS_{LOF} "large" with respect to MS_{PE}, or more precisely if:

$$F_0 > F_{\alpha,\, m\text{-}K,\, n\text{-}m}$$

Orthogonality in the design matrix

If the columns of the $\mathbf{X}$ matrix are orthogonal ($X_j'X_i=0$ for $i \neq j$)

The matrix $\mathbf{X}^T\mathbf{X}$ is diagonal and the estimated coefficients of β have null covariance (are not correlated)