NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (17)

ANOVA is also MLR [4.2]
Part 5: Model assessment and transformation [7.1-7.4]

Mette Langaas

To be lectured: March 3, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Outline

- ANOVA is also MLR [4.2]
- MLR model assessment using residuals [7.1]
- Tranforming data to achieve better MLR fit [7.2-7.3]
  - Box-Cox transformation
  - Taylor expansion as basis for variance stabilizing transformation.
- Orthogonality - to multicollinearity [7.4]

Ch2: Concrete aggregates data

Aggregate:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Table 13.1 of WMMY.

Ch2: Age and memory

— Why do older people often seem not to remember things as well as younger people? Do they not pay attention? Do they just not process the material as thoroughly?

— One theory regarding memory is that verbal material is remembered as a function of the degree to which is was processed when it was initially presented.

— Eysenck (1974) randomly assigned 50 younger subjects and 50 older (between 55 and 65 years old) to one of five learning groups.

— After the subjects had gone through a list of 27 items three times they were asked to write down all the words they could remember.

Eysenck study of recall of older and younger subjects under conditions of differential processing, Eysenck (1974) and presented in Howell (1999).

Ch2: Two factors and interaction

Model:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$\text{for } i = 1, 2, ..., r \text{ and } j = 1, 2, ..., n \text{ and } k = 1, ..., m$$
$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

TMA 4267 : lecture 17

Do the ANOVA-models of Ch2 [2.6-2.8] fit into the MLR framework ?

Ex: one-way ANOVA with $k=3$ groups and $n_i = 3$

$i = 1,2,3$, $n = \sum_{i=1}^{3} n_i = 9$

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad , \; \varepsilon_{ij} \text{ i.i.d } N(0,\sigma^2)$$

Relabel Y and ε

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ \vdots \\ Y_{33} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \\ \\ \\ Y_9 \end{bmatrix}$$

and the same for ε_{ij}

"$\mu + \alpha_i$"

MLR: $\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$

dummy variable encoding

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \qquad X = \begin{array}{c} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} \leftarrow \text{gr } 1 \\ \\ \\ \\ \text{gr } 2 \\ \\ \\ \\ \text{gr } 3 \end{array}$$

Observe: $X[,1] = X[,2] + X[,3] + X[,4]$

so, X doesn't have full rank $\Rightarrow$ can't find $(X^TX)^{-1}$.

Many solutions:

1) drop the intercept (drop col 1)

2) Set $\alpha_1 = 0$ so $\mu = \mu_1$ (drop col 2)

= treatment contrast

3) Set $\alpha_k = -\sum_{i=1}^{k-1} \alpha_i$ and drop the $(k+1)$th column from X

= sum-zero constraint

$\Rightarrow$ ESP3

- Assume that we do 1). Then $(X^TX) = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}$

is a diagonal matrix

$$\hat{\beta} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = (X^T X)^{-1} X^T Y = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{n_3} \end{bmatrix} \begin{bmatrix} 111 & 000 & 000 \\ 000 & 111 & 000 \\ 000 & 000 & 111 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$y_1 \; y_2 \; y_3$

$$= \begin{bmatrix} \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} \\ \frac{1}{n_3} \sum_{j=1}^{n_3} y_{3j} \end{bmatrix}$$

$\Rightarrow$ The ANOVA model can be seen as a MLR.

This also holds for two-way.

R: see ESP

PART 5: Model check and transformations

$$Y = \underbrace{E(Y)}_{X\beta} + \varepsilon \quad , \varepsilon \sim N_n(0, \sigma^2 I)$$

MLR assumptions

i) $E(Y) = X\beta$ is a linear function in parameters (and covariates)

ii) Errors (ε) are additive.

iii) Errors are independent.

iv) Errors are normal

v) Errors have equal variance

All of these are fulfilled if we start with $\underbrace{(X_1, Y_1)^T}_{1 \times p}, (X_2, Y_2)^T, \dots, (X_n, Y_n)^T$ independent and mvN, and look at $E(Y \mid X = x)$ and $Var(Y \mid X = x)$.

$E(Y \mid X = x)$: linear in x

$Var(Y \mid X = x)$: independent of x

Modelling $(Y \mid X = x)$ using MLR fulfills (i)-(v) above

How may we assess/check assumptions i-v, and what can we do if the assumptions are violated?

i) can be assessed by plotting x_2 vs y, x_3 vs y, ...

R: pairs

i–v) can be assessed by studying various types of residual plots.

Let $\underset{n\times 1}{e} = Y - \hat{Y} = (I-H)Y$ be our raw residuals

What do we know about e when the MLR assumptions hold?

a) $\underset{n\times 1}{e} \sim N_n(0, (I-H)\sigma^2)$ $\quad$ $Cov((I-H)Y)$

b) e and $\hat{Y}$ are independent

Model assessment with residuals

- All the sample information on lack of fit is contained in the residuals: $e_i = y_i - \hat{y}_i$, $i = 1, ..., n$.
  $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
- If we assume $\varepsilon_i \sim N(0, \sigma^2)$ and independent, then
  $e_i \sim N(0, \sigma^2(1 - h_{ii}))$.
  $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\boldsymbol{e} \sim N_n(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{H})\sigma^2)$.
- The diagonal elements of $\boldsymbol{H}$ are called *leverages* h_{ii}.
- Thus, the residuals have unequal variances and nonzero correlation.
- However, often the correlations are small and the variances are nearly equal.

Due to a) we often instead look at "standardized" residuals:

$$H = X(X^TX)^{-1}X$$

$$h_{ii} = [H]_{ii} = h_i$$

$$S_i = \frac{e_i}{\sqrt{(1-h_{ii})}\,\hat{\sigma}} \approx t_{n-p}$$

R: rstandard(lmobj)

or "standardized deleted" residuals

$$e_{-i} = y_i - x_i^T \hat{\beta}_{-i}$$

estimated without obs i in the data set

$$S_{-i} = \frac{e_{-i}}{\sqrt{1-h_{ii}}\;\hat{\sigma}} \quad [\sim t_{n-p-1} \text{ exact}]$$

R: rstudent(lmobj)

Residuals

- Using the estimated variance for residual i, $\hat{\text{Var}}(e_i) = \hat{\sigma}^2(1 - h_{ii})$
- we define *standardized residuals*, or internally studentized residuals (R: rstandard)

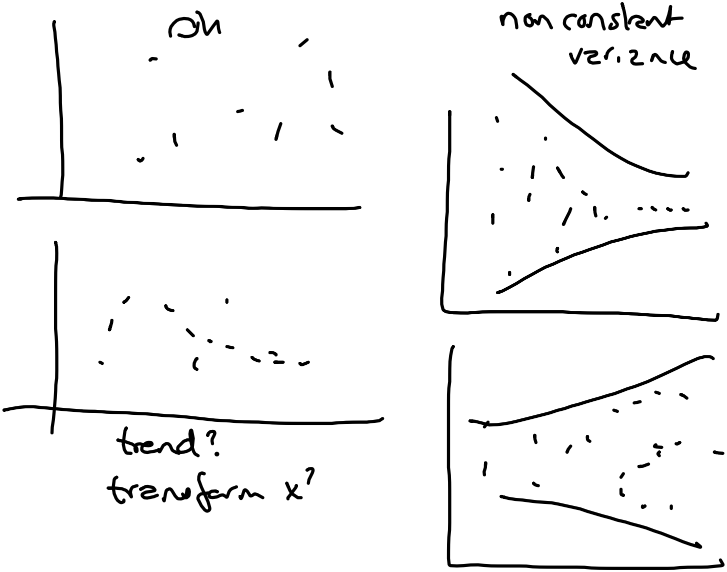
$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

- *Externally studentized residuals* are even better, base $\hat{\sigma}^2$ and $\hat{y}_i$ on all observations except nr i. (R: rstudent).

Plotting residuals

1. Plot the residuals, r_i against the predicted values, $\hat{y}_i$.
    - Dependence of the residuals on the predicted value: wrong regression model?
    - Nonconstant variance: transformation or weighted least squares is needed?
2. Plot the residuals, r_i, against predictor variable or functions of predictor variables. Trend suggest that transformation of the predictors or more terms are needed in the regression.
3. QQ-plots and histograms of residuals. Normality?
4. Plot the residuals, r_i, versus time. Dependence or autocorrelation?

1) $\hat{y}_i$ vs S_{-i}



2) If 1 is strange → proceed to plot each x_{ij} vs S_{-i}

$j = 2, \ldots, p$

3) QQplot of S_{-i}

4) If data in time or space: look for trend.

Effect of wrong model

Correct model:

$$Y_i = 1 + 3 \cdot \log x_i + 1 \cdot x_2 + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

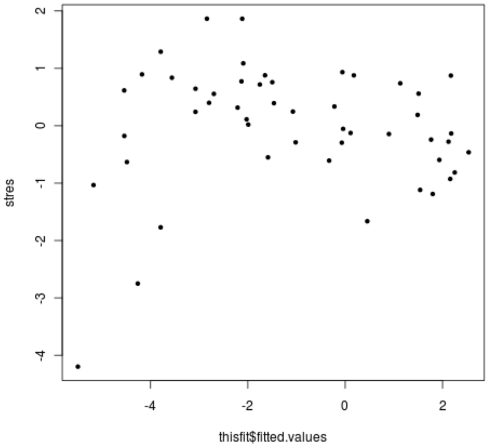
where x_{1i} and x_{2i} both generated from uniform[0,1].
Fitted model:

$$Y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

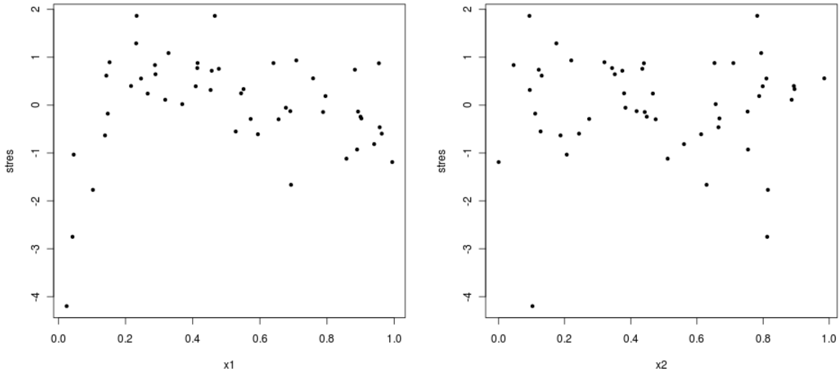
Data analysis based on $n = 50$ observations.

```
lm(formula = y ~ x1 + x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.8404     0.5117 -11.413 3.80e-15 ***
x1            7.6819     0.6773  11.343 4.71e-15 ***
x2            1.5452     0.7204   2.145   0.0372 *
---
Residual standard error: 1.391 on 47 degrees of freedom
Multiple R-squared: 0.7458,Adjusted R-squared: 0.735
F-statistic: 68.96 on 2 and 47 DF,  p-value: 1.048e-14
```

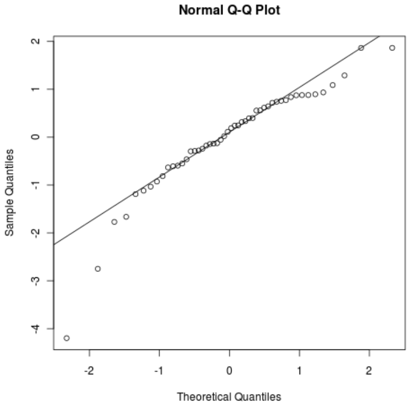
Wrong model: Studentized residuals vs. fitted

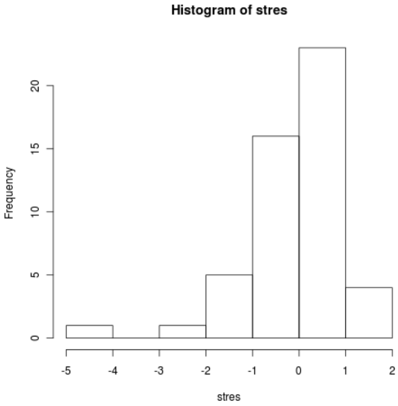


Wrong model: Studentized residuals vs. covariates



Wrong model: Normal qq-plot from studentized residuals





Corrected: Effect of wrong model

$$Y_i = 1 + 3 \cdot \log x_i + 1 \cdot x_2 + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

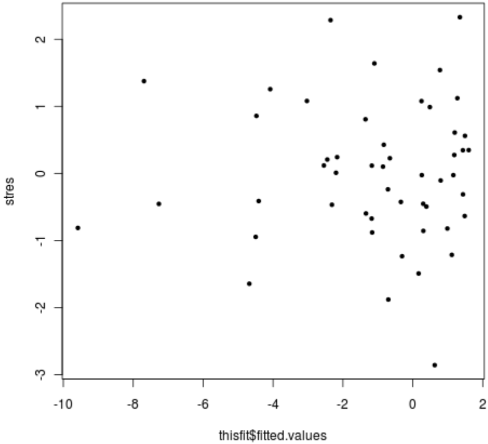
where x_{1i} and x_{2i} both generated from uniform[0,1].
Fitted model:

$$Y_i = \beta_0 + \beta_1 \cdot \log x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

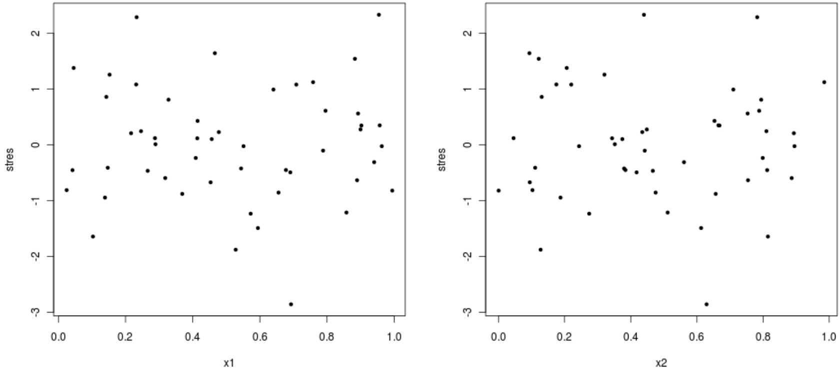
```
lm(formula = y ~ log(x1) + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0039     0.3101   3.237  0.00222 **
log(x1)       2.8879     0.1532  18.846  < 2e-16 ***
x2            1.0848     0.4782   2.269  0.02793 *
---
Residual standard error: 0.9192 on 47 degrees of freedom
Multiple R-squared: 0.889,Adjusted R-squared: 0.8843
F-statistic: 188.2 on 2 and 47 DF,  p-value: < 2.2e-16
```

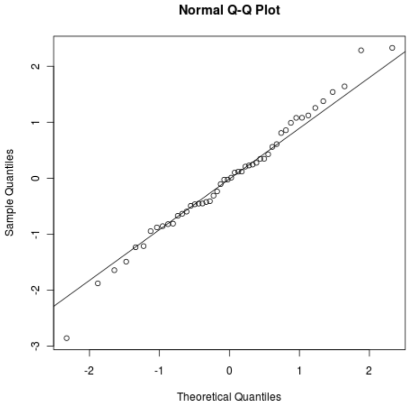
Correct model: Studentized residuals vs. fitted

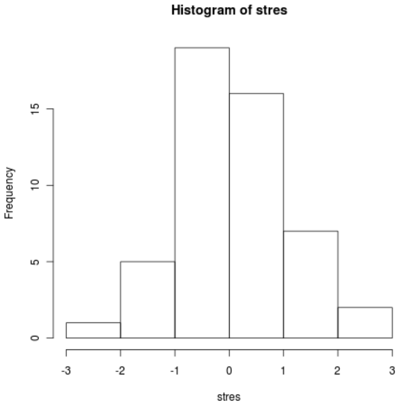


Correct model: Studentized residuals vs. covariates



Correct model: Normal qq-plot from studentized residuals





Influential observations

- Observations that significantly affect inferences drawn from the data are said to be influential.
- The leverage, h_{ii}, associated with the ith datapoint measures “how far the ith observation is from the other $n-1$ observations”.
- Methods for assessing influential observations may be be based on change in β estimate when observations are deleted.
- Cook’s distance can be used to identify influential observations.

Influential observation

Remember $e \sim N_n(0, \sigma^2(I-H))$, when $H = X(X^TX)^{-1}X^T$

$h_i = [H]_{ii}$: if h_i is large, then $Var(e_i)$ is small.

What would be an average value of h_i?

$$\sum_{i=1}^{n} h_i = tr(H) = tr\left(X(X^TX)^{-1}X^T\right) = tr\big(\underbrace{X^TX\,(X^TX)^{-1}}_{\underset{p\times p}{I}}\big) = p$$

so p/n would be an average value of h_i.

Rule of thumb: look at obs when $h_i > 2 \cdot p/n$

Cook statistic

$$D_i = \frac{(\hat{y}_i - \hat{y}_{(i)})^T(\hat{y}_i - \hat{y}_{(i)})}{p\,\hat{\sigma}^2}$$

$$= \frac{(\hat{\beta} - \hat{\beta}_{-i})^T X^TX(\hat{\beta} - \hat{\beta}_{-i})}{p\,\hat{\sigma}^2} = \frac{1}{p} S_i^2 \frac{h_i}{1-h_i}$$

stand res.

Used to identify influential observations

R: cooks.distance(lmobj)

For further reading: Faraway: "Linear models with R"

Transformations [7.2-7.3]

of response and predictors may improve the fit and correct violations of model assumptions

N, add errors, noncont error

a) Multiplicative or additive errors?

$$\log y = \beta_0 + \beta_1 x + \varepsilon$$

$$y = \exp(\beta_0 + \beta_1 x) \cdot \exp(\varepsilon)$$

By taking $\log y$ I'm assuming a multiplicative model.

Tomorrow

b) Box Cox

c) Variance stab