Introduction to General and Generalized Linear Models

General Linear Models - part II

Henrik Madsen
Poul Thyregod

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kgs. Lyngby

February 27, 2012

Today

- Quick summary
- Test for model reduction
- Type I/III SSQ
- Collinearity
- Inference on individual parameters
- Confidence intervals
- Prediction intervals
- Residual analysis
- Examples

Summary: General Linear Model

- A general linear model is:

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

Consider the well known two way ANOVA:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{ i.i.d. } N(0, \sigma^2), \quad i = 1, 2, \quad j = 1, 2, 3.$$

An expanded view of this model is:

$$\begin{array}{lllllll}
y_{11} = \mu & + \alpha_1 & & + \beta_1 & & & + \varepsilon_{11} \\
y_{21} = \mu & & + \alpha_2 & + \beta_1 & & & + \varepsilon_{21} \\
y_{12} = \mu & + \alpha_1 & & & + \beta_2 & & + \varepsilon_{12} \\
y_{22} = \mu & & + \alpha_2 & & + \beta_2 & & + \varepsilon_{22} \\
y_{13} = \mu & + \alpha_1 & & & & + \beta_3 & + \varepsilon_{13} \\
y_{23} = \mu & & + \alpha_2 & & & + \beta_3 & + \varepsilon_{23}
\end{array} \tag{1}$$

The exact same in matrix notation:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}} \tag{2}$$

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

- $\mathbf{y}$ is the vector of all observations
- $\mathbf{X}$ is known as the *design matrix*
- $\boldsymbol{\beta}$ is the vector of parameters
- $\boldsymbol{\varepsilon}$ is a vector of independent $N(0, \sigma^2)$ "measurement noise"
  - The vector $\boldsymbol{\varepsilon}$ is said to follow a *multivariate normal distribution*
  - Mean vector $\mathbf{0}$
  - Covariance matrix $\sigma^2\mathbf{I}$
  - Written as: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ specifies the model, and everything can be calculated from $\mathbf{y}$ and $\mathbf{X}$.

In a general linear model (with both factors and covariates), it is surprisingly easy to construct the design matrix $\mathbf{X}$.

- For each factor: Add one column for each level, with ones in the rows where the corresponding observation is from that level, and zeros otherwise.
- For each covariate: Add one column with the measurements of the covariate.
- Remove linear dependencies (if necessary)

Example: linear regression:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon$$

In matrix notation:

$$\mathbf{y} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \boldsymbol{\varepsilon}$$

General Linear Model

- A general linear model is:

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

- $\boldsymbol{\beta}$ can be estimated by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}X^T\boldsymbol{y} \quad \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

- σ^2 can be estimated by:

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{f} \quad \sim \sigma^2\chi^2_f/f \quad \text{where } f = n - k.$$

- The deviance is:

$$D(\boldsymbol{y}, p(\boldsymbol{y})) = ||\boldsymbol{y} - p(\boldsymbol{y})||^2 = (\boldsymbol{y} - p(\boldsymbol{y}))^T(\boldsymbol{y} - p(\boldsymbol{y})) = \sum_i (\boldsymbol{y}_i - p(\boldsymbol{y})_i)^2$$

Test for model reduction

Theorem (A test for model reduction)

The likelihood ratio test statistic for testing

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \Omega_0 \quad \text{against the alternative} \quad \mathcal{H}_1 : \boldsymbol{\mu} \in \Omega_1 \setminus \Omega_0$$

is a monotone function of

$$F(\boldsymbol{y}) = \frac{\mathrm{D}(p_1(\boldsymbol{y}); p_0(\boldsymbol{y}))/(m_1 - m_0)}{\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))/(n - m_1)}$$

where $p_1(\boldsymbol{y})$ and $p_0(\boldsymbol{y})$ denote the projection of $\boldsymbol{y}$ on Ω_1 and Ω_0, respectively. Under $\mathcal{H}_0$ we have

$$F \sim F(m_1 - m_0, n - m_1)$$

i.e. large values of F reflects a conflict between the data and $\mathcal{H}_0$, and hence lead to rejection of $\mathcal{H}_0$. The p-value of the test is found as $p = P[F(m_1 - m_0, n - m_1) \geq F_{obs}]$, where F_{obs} is the observed value of F given the data.

Tests for model reduction

- Assume that a rather comprehensive model (a *sufficient model*) $\mathcal{H}_1$ has been formulated.
- Initial investigation has demonstrated that at least some of the terms in the model are needed to explain the variation in the response.
- The next step is to investigate whether the model may be reduced to a simpler model (corresponding to a smaller subspace),.
- That is we need to test whether all the terms are *necessary*.

Successive testing, type I partition

- Sometimes the practical problem to be solved by itself suggests a *chain* of hypothesis, one being a sub-hypothesis of the other.
- In other cases, the statistician will establish the chain using the general rule that more complicated terms (e.g. interactions) should be removed before simpler terms.
- In the case of a classical GLM, such a chain of hypotheses corresponds to a sequence of linear parameter-spaces, $\Omega_i \subset \mathbb{R}^n$, one being a subspace of the other.

$$\mathbb{R} \subseteq \Omega_M \subset \ldots \subset \Omega_2 \subset \Omega_1 \subset \mathbb{R}^n \ ,$$

where

$$\mathcal{H}_i : \boldsymbol{\mu} \in \Omega_i, \quad i = 2, \ldots, M$$

with the alternative

$$\mathcal{H}_{i-1} : \boldsymbol{\mu} \in \Omega_{i-1} \setminus \Omega_i \quad i = 2, \ldots, M$$

Partitioning of total model deviance

Theorem (Partitioning of total model deviance)

Given a chain of hypotheses that has been organised in a hierarchical manner then the model deviance $\mathrm{D}(p_1(\boldsymbol{y}); p_M(\boldsymbol{y}))$ *corresponding to the initial model* $\mathcal{H}_1$ *may be partitioned as a sum of contributions with each term*

$$\mathrm{D}(p_{i+1}(\boldsymbol{y}); p_i(\boldsymbol{y})) = \mathrm{D}(\boldsymbol{y}; p_{i+1}(\boldsymbol{y})) - \mathrm{D}(\boldsymbol{y}; p_i(\boldsymbol{y}))$$

representing the increase in residual deviance $\mathrm{D}(\boldsymbol{y}; p_i(\boldsymbol{y}))$ *when the model is reduced from* $\mathbb{H}_i$ *to the next lower model* $\mathbb{H}_{i+1}$.

Partitioning of total model deviance

- Assume that an initial (*sufficient*) model with the projection $p_1(\boldsymbol{y})$ has been found.
- By using the Theorem, and hence by partitioning corresponding to a chain of models we obtain:

$$||p_1(\boldsymbol{y}) - p_M(\boldsymbol{y})||^2 = ||p_1(\boldsymbol{y}) - p_2(\boldsymbol{y})||^2 + ||p_2(\boldsymbol{y}) - p_3(\boldsymbol{y})||^2 + \cdots + ||p_{M-1}(\boldsymbol{y}) - p_M(\boldsymbol{y})||^2$$

- It is common practice for statistical software to print a table showing this partitioning of the model deviance $\mathrm{D}(p_1(\boldsymbol{y}); p_M(\boldsymbol{y}))$.
- The partitioning of the model deviance is from this sufficient or total model to lower order models, and very often the most simple model is the null model with $\dim = 1$
- This is called *Type I partitioning*.

Type I partitioning

Source	f	Deviance	Test
$\mathcal{H}_M$	$m_{M-1} - m_M$	$\|\|p_{M-1}(\boldsymbol{y}) - p_M(\boldsymbol{y})\|\|^2$	$\frac{\|\|p_{M-1}(\boldsymbol{y}) - p_M(\boldsymbol{y})\|\|^2/(m_{M-1} - m_M)}{\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2/(n - m_1)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{H}_3$	$m_2 - m_3$	$\|\|p_2(\boldsymbol{y}) - p_3(\boldsymbol{y})\|\|^2$	$\frac{\|\|p_2(\boldsymbol{y}) - p_3(\boldsymbol{y})\|\|^2/(m_2 - m_3)}{\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2/(n - m_1)}$
$\mathcal{H}_2$	$m_1 - m_2$	$\|\|p_1(\boldsymbol{y}) - p_2(\boldsymbol{y})\|\|^2$	$\frac{\|\|p_1(\boldsymbol{y}) - p_2(\boldsymbol{y})\|\|^2/(m_1 - m_2)}{\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2/(n - m_1)}$
Residual under $\mathcal{H}_1$	$n - m_1$	$\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2$	

Table: Illustration of Type I partitioning of the total model deviance $||p_1(\boldsymbol{y}) - p_M(\boldsymbol{y})||^2$. In the table it is assumed that $\mathcal{H}_M$ corresponds to the null model where the dimension is $\dim = 1$.

Type I partitioning conclusions

- Corresponds to a successive projection corresponding to a chain of hypotheses reflecting a chain of linear parameter sub-spaces, such that spaces of lower dimensions are embedded in the higher dimensional spaces.
- The effect at any stage (typically the effect of a new variable) in the chain is evaluated after all previous variables in the model have been accounted for.
- The Type I deviance table depends on the order of which the variables enters the model.

Reduction of model using partial tests

- We have considered a fixed layout of a chain of models. However, a particular model can be formulated along a high number of different chains.
- Let us now consider some other types of test which by construction do not depend on the order of which the variables enters the model.
- Consider a given model $\mathcal{H}_i$. This model can be reduced along different chains. More particular we will consider the *partial likelihood ratio test*:

Partial likelihood ratio test

Definition (Partial likelihood ratio test)

Consider a sufficient model as represented by $\mathcal{H}_i$. Assume now that the hypothesis $\mathcal{H}_i$ allows the different sub-hypotheses $\mathcal{H}_{i+1}^A \subset \mathcal{H}_i$; $\mathcal{H}_{i+1}^B \subset \mathcal{H}_i$; $\ldots \mathcal{H}_{i+1}^S \subset \mathcal{H}_i$. A *partial likelihood ratio test* for $\mathcal{H}_{i+1}^J$ under $\mathcal{H}_i$ is the (conditional) test for the hypotheses $\mathcal{H}_{i+1}^J$ given $\mathcal{H}_i$.

The numerator in the F-test quantity for the partial test is found as the deviance between the two models, i.e. $\widehat{\boldsymbol{\mu}}$ under $\mathcal{H}_i$ and $\widehat{\widehat{\boldsymbol{\mu}}}$ under $\mathcal{H}_{i+1}^J$

$$F(\boldsymbol{y}) = \frac{\mathrm{D}(\widehat{\boldsymbol{\mu}}; \widehat{\widehat{\boldsymbol{\mu}}})/(m_i - m_{i+1})}{||\boldsymbol{y} - p_i(\boldsymbol{y})||^2/(n - m_i)} \ ,$$

where $||\boldsymbol{y} - p_i(\boldsymbol{y})||^2/(n - m_i)$, which for $\boldsymbol{\Sigma} = \boldsymbol{I}$ is the variance of the residuals under $\mathcal{H}_i$.

Simultaneous testing, Type III partition

Type III partition

The Type III partition is obtained as the partial test for all factors.

- The Type III partitioning gives the deviance that would be obtained for each variable if it were entered last into the model.
- That is, the effect of each variable is evaluated after all other factors have been accounted for.
- Therefore the result for each term is equivalent to what is obtained with Type I analysis when the term enters the model as the last one in the ordering.

Type I/III

- There is no consensus on which type should be used for unbalanced designs, but most statisticians generally recommend Type III.
- Type III is the default in most software packages such as SAS, SPSS, JMP, Minitab, Stata, Statista, Systat, and Unistat
- R, S-Plus, Genstat, and Mathematica use Type I.
- Type I SS also called *sequential sum of squares*, whereas Type III is called *marginal sum of squares*.
- Unlike the Type I SS, the Type III SS will NOT sum to the Sum of Squares for the model corrected only for the mean (Corrected Total SS).

Example: Detergent powder

The efficacy of detergent powder is often assessed by washing pieces of cloth that has been stained with specified amounts of various types of fat. In order to assess the staining process an experiment was performed where three technicians (α) applied the same amount of each of two types of oil (β) to pieces of cloth, and subsequently measured the area of the stained spot. The results are:

Type of fat	Technician		
	A	D	E
Lard	46.5	43.9	53.7
Olive	55.4		61.7

Table: The area of the stained spot after application of the same amount of two types of oil to pieces of cloth.

Example: Detergent powder

The data are analyzed by a two-way model $Y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$ with the additive mean value structure:

$$\mathcal{H}_1 : \mu_{ij} = \mu + \alpha_i + \beta_j; i = 1, 2; \;\; j = 1, 2, 3$$

with suitable restrictions on α_i and β_j to assure identifiability.
Now, we formulate the chain

$$\mathcal{H}_M \subset \mathcal{H}_2 \subset \mathcal{H}_1$$

with $\mathcal{H}_2 : \beta_j = 0$, $j = 1, 2, 3$

$$\mathcal{H}_2 : \mu_{ij} = \mu + \alpha_i \;\; i = 1, 2; \;\; j = 1, 2, 3$$

and

$$\mathcal{H}_M : \mu_{ij} = \mu; \;\; i = 1, 2; \;\; j = 1, 2, 3$$

Example: Detergent powder

```
> Area <- c(46.5, 43.9, 53.7, 55.4, 61.7)
> Fat <- as.factor(c("L", "L", "L", "O", "O"))
> Techn <- as.factor(c("A", "D", "E", "A", "E"))
> fit <- lm(Area ~ Fat + Techn)
> anova(fit)

Analysis of Variance Table

Response: Area
          Df  Sum Sq Mean Sq F value  Pr(>F)
Fat        1 132.720 132.720  655.41 0.02485 *
Techn      2  71.189  35.595  175.78 0.05326 .
Residuals  1   0.203   0.203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(fit, test = "F")

Single term deletions

Model:
Area ~ Fat + Techn
       Df Sum of Sq    RSS     AIC F value   Pr(F)
<none>               0.203 -8.0323
Fat     1    71.403 71.605 19.3086  352.60 0.03387 *
Techn   2    71.189 71.392 17.2937  175.78 0.05326 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Collinearity

Collinearity

When some predictors are linear combinations of others, then $\boldsymbol{X}^T\boldsymbol{X}$ is singular, and there is (exact) *collinearity*. In this case there is no unique estimate of $\boldsymbol{\beta}$. When $\boldsymbol{X}^T\boldsymbol{X}$ is close to singular, there is collinearity (some texts call it *multicollinearity*).

There are various ways to detect collinearity:

i Examination of the correlation matrix for the estimates may reveal strong pairwise col-linearities

ii Considering the change of the variance of the estimate of other parameters when removing a particular parameter.

Inference on individual parameters in parameterized models

Theorem (Test of individual parameters)

A hypotheses $\beta_j = \beta_j^0$ related to specific values of the parameters are evaluated using the test quantity

$$t_j = \frac{\widehat{\beta}_j - \beta_j^0}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}_{jj}}},$$

where $(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}_{jj}$ denotes the j'th diagonal element of $(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$. The test quantity is compared with the quantiles of a $r(n-m_0)$ distribution.
The hypotheses is rejected for large values t_j, i.e. for

$$|t_j| > t_{1-\alpha/2}(n-m_0)$$

and the p-value is found as $p = 2P[t(n-m_0) \geq |t_{obs}|]$
In the special case of the hypotheses $\beta_j = 0$ the test quantity is

$$t_j = \frac{\hat{\beta}_j}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}_{jj}}}$$

Confidence intervals and confidence regions

Confidence intervals for individual parameters

$100(1-\alpha)\%$ confidence interval for β_j is found as

$$\widehat{\beta}_j \pm t_{1-\alpha/2}(n-m_0)\widehat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}_{jj}}$$

Simultaneous confidence regions for model parameters

It follows from the distribution of $\widehat{\boldsymbol{\beta}}$ that

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq m_0\widehat{\sigma}^2 F_{1-\alpha}(m_0, n-m_0)$$

These regions are ellipsoidally shaped regions in $\mathbb{R}^{m_0}$. They may be visualised in two dimensions at a time.

Prediction - known parameters

We will assume that $\boldsymbol{\Sigma} = \boldsymbol{I}$. For predictions in the case of correlated observations we refer to the literature on time series analysis, Madsen (2008) [1].

Consider the linear model with *known parameters.*

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\theta} + \varepsilon_i$$

where $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}[\varepsilon_i] = \sigma^2$ (i.e. constant). The prediction for a future value $Y_{n+\ell}$ given the independent variable $\boldsymbol{X}_{n+\ell} = \boldsymbol{x}_{n+\ell}$ is

$$\widehat{Y}_{n+\ell} = \mathrm{E}[\, Y_{n+\ell} | \boldsymbol{X}_{n+\ell} = \boldsymbol{x}_{n+\ell}] = \boldsymbol{x}_{n+\ell}^T \boldsymbol{\theta}$$

$$\mathrm{Var}[\, Y_{n+\ell} - \widehat{Y}_{n+\ell}] = \mathrm{Var}[\varepsilon_{n+\ell}] = \sigma^2$$

[1] Madsen, H. (2008) Time Series Analysis. Chapman, Hall

Prediction - unknown parameters

Most often the parameters are unknown but assume that there exist some estimates of $\boldsymbol{\theta}$. Assume also that the estimates are found by the estimator

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

then the variance of the prediction error can be stated.

Theorem (Prediction in the general linear model)

Assume that the unknown parameters $\boldsymbol{\theta}$ in the linear model are estimated using the least squares method, then prediction is

$$\widehat{Y}_{n+\ell} = \mathrm{E}[\, Y_{n+\ell} | \boldsymbol{X}_{n+\ell} = \boldsymbol{x}_{n+\ell}] = \boldsymbol{x}_{n+\ell}^T \widehat{\boldsymbol{\theta}}$$

The variance of the prediction error $e_{n+\ell} = Y_{t+\ell} - \widehat{Y}_{n+\ell}$ becomes

$$\mathrm{Var}[e_{n+\ell}] = \mathrm{Var}[\, Y_{n+\ell} - \widehat{Y}_{n+\ell}] = \sigma^2 [1 + \boldsymbol{x}_{n+\ell}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{n+\ell}]$$

Prediction - unknown parameters

If we use an estimate for σ^2 and $\boldsymbol{\Sigma} = \boldsymbol{I}$ then a $100(1-\alpha)\%$ *confidence interval for the future* $Y_{n+\ell}$ is given as

$$\begin{aligned}
&\widehat{Y}_{n+\ell} \pm t_{\alpha/2}(n-k)\sqrt{\mathrm{Var}[e_{n+\ell}]} \\
&\quad = \widehat{Y}_{n+\ell} \pm t_{\alpha/2}(n-k)\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_{n+\ell}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_{n+\ell}}
\end{aligned}$$

where $t_{\alpha/2}(n-k)$ is the $\alpha/2$ quantile in the t distribution with $(n-k)$ degrees of freedom and n is the number of observations used in estimating the k unknown parameters.

A confidence interval for a future value is also called a *prediction interval*.

Residuals, standardization and studentization

- The residuals denote the difference between the observed value y_i and the value $\widehat{\mu}_i$ fitted by the model.
- The raw residuals do not have the same variance.
- The variance of r_i is $\sigma^2(1 - h_{ii})$ with h_{ii} denoting the i'th diagonal element in the hat-matrix.
- Therefore, in order to meaningful compare the residuals, it is usual practice to rescale them by dividing by the estimated standard deviation.

Standardized residual

Definition (Standardized residual)

The standardized residual is

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

Standardization does not imply that the variance is 1

It is a usual convention in statistics that *standardization* of a random variable means transforming to a variable with mean 0, and variance 1. Often standardization takes place by dividing the variable by its standard deviation. We have only divided by an estimate of the standard deviation, so although we have achieved equal variance for the standardized residuals, the variance is not 1.

Standardized residual

- Residuals are often used to identify *outliers*, i.e. observations that for some reasons do not fit into the general pattern, and possibly should be excluded.
- However, if a particular observation y_i is such a contaminating observation giving rise to a large residual r_i, then that observation would also inflate the estimate $\widehat{\sigma}$ of the standard deviation in the denominator of the standardized residual thereby masking the effect of the contamination.
- Therefore, it is advantageous to scale the residual with an estimate of the variance, σ^2, that does not include the i'th observation.

Studentized residual

Definition (Studentized residual)

The studentized residual is

$$r_i^{rt} = \frac{r_i}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}$$

where $\widehat{\sigma}^2_{(i)}$ denotes the estimate for σ^2 determined by deleting the i'th observation.

It may be shown that $\widehat{\sigma}^2_{(i)} \sim \sigma^2\chi^2(f)/f$-distribution with $f = n - m_0 - 1$ and therefore, as r_i is independent of $\widehat{\sigma}^2_{(i)}$ that the studentized residual follow a $t(f)$-distribution when $\mathcal{H}_0$ holds.

Residual analysis

- The residuals provide a valuable tool for model checking.
- The residuals should be checked for individual large values.
- A large standardized or studentized residual is an indication of poor model fit for that point, and the reason for this outlying observation should be investigated.
- For observations obtained in a time-sequence the residuals should be checked for possible autocorrelation, or seasonality.
- The distribution of the studentized residuals should be investigated and compared to the reference distribution (the t-distribution, or simply a normal distribution) by means of a qq-plot to identify possible anomalies.

Influential observations, leverage

It follows from

$$\frac{\partial \widehat{\boldsymbol{\mu}}}{\partial \boldsymbol{y}} = \boldsymbol{H}$$

that the i'th diagonal element h_{ii} of the hat-matrix $\boldsymbol{H}$ denotes the change in the fitted value $\widehat{\mu}_i$ induced by a change in the i'th observation y_i.

In other words, the diagonal elements in the hat-matrix $\boldsymbol{H}$ indicate the "weight" with which the individual observation contributes to the fitted value for that data point.

Definition (Leverage)

The i'th diagonal element in the hat-matrix is called the *leverage* of the i'th observation.

Leverage

- One should pay special attention to observations with large leverage.
- It is not necessarily is undesirable that an observation has a large leverage.
- If an observation with large leverage is in agreement with the other data, that point would just serve to reduce the uncertainty of the estimated parameters.
- When a model is reasonable, and data are in agreement with the model, then observations with large leverage will be an advantage.
- Observations with a large leverage might however be an indication that the model does not represent the data.

Cook's distance

- Observations with large **residuals** and/or high **leverage** may distort the outcome, reliability and/or accuracy of a regression.
- Cook's distance for observation No. i is defined by

$$D_i = \frac{r_i^2}{k\widehat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

or

$$D_i = \frac{(r_i^{rs})^2}{k}\left[\frac{h_{ii}}{(1-h_{ii})}\right]$$

where h_{ii} is the i'th diagonal element of the hat matrix, k the number of estimated parameters, and r^{rs} is the standardized residual,

Residual plots

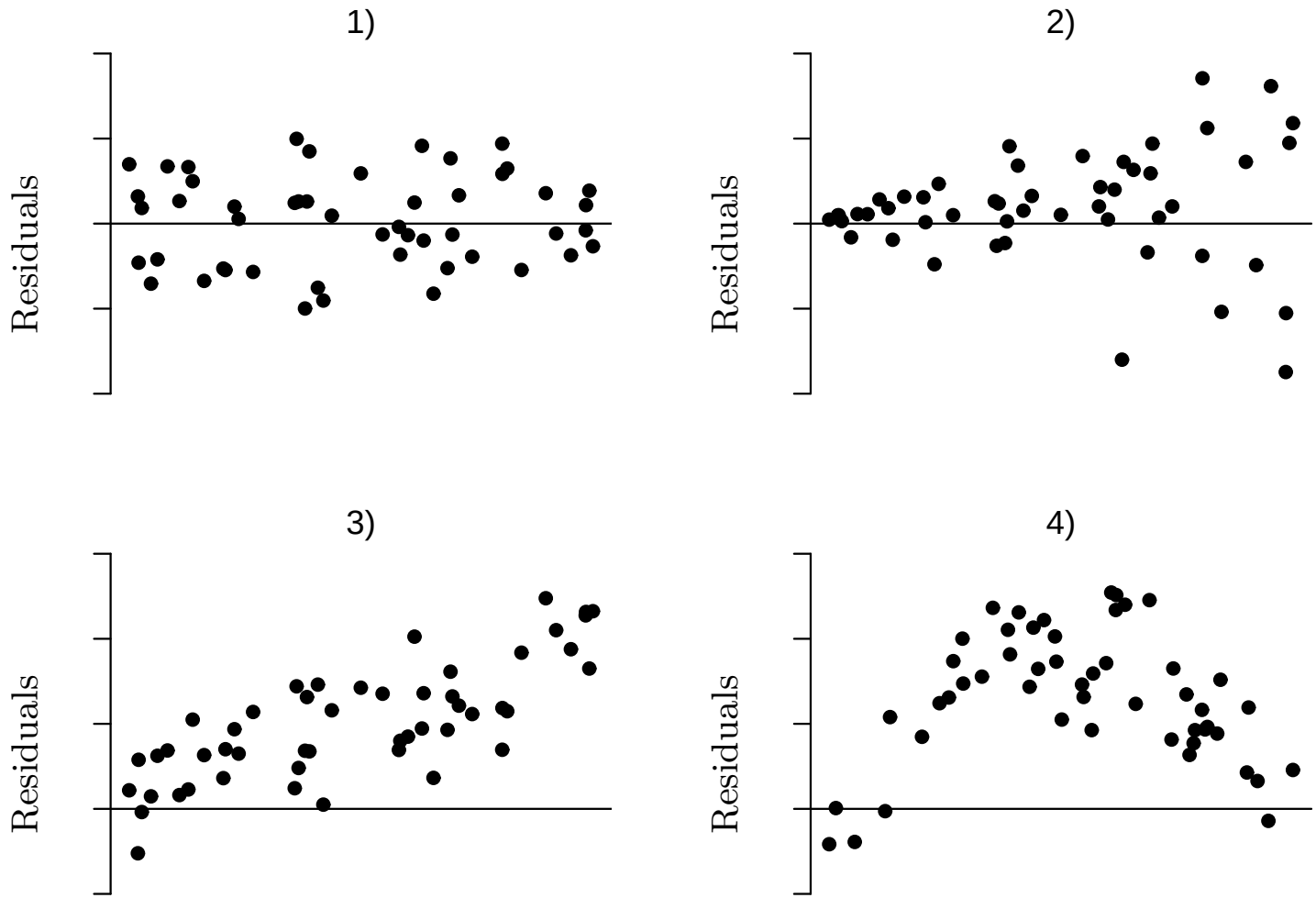


Figure: Residual plots.

Plot of residuals against time

1) Acceptable
2) The variance grows with time. Do a weighted analysis.
3) Lack of term of the form $\beta \cdot \texttt{time}$.
4) Lack of term of the form $\beta \cdot \texttt{time} + \beta_1 \cdot \texttt{time}^2$.

Plot of residuals against independent variables

1) Acceptable
2) The variance grows with x_i. Perform weighted analysis or transform the Y'es (e.g. with the logarithm or equivalent).
3) Error in the computations.
4) Lack of quadratic term in x_i.

Plot of residuals against fitted values

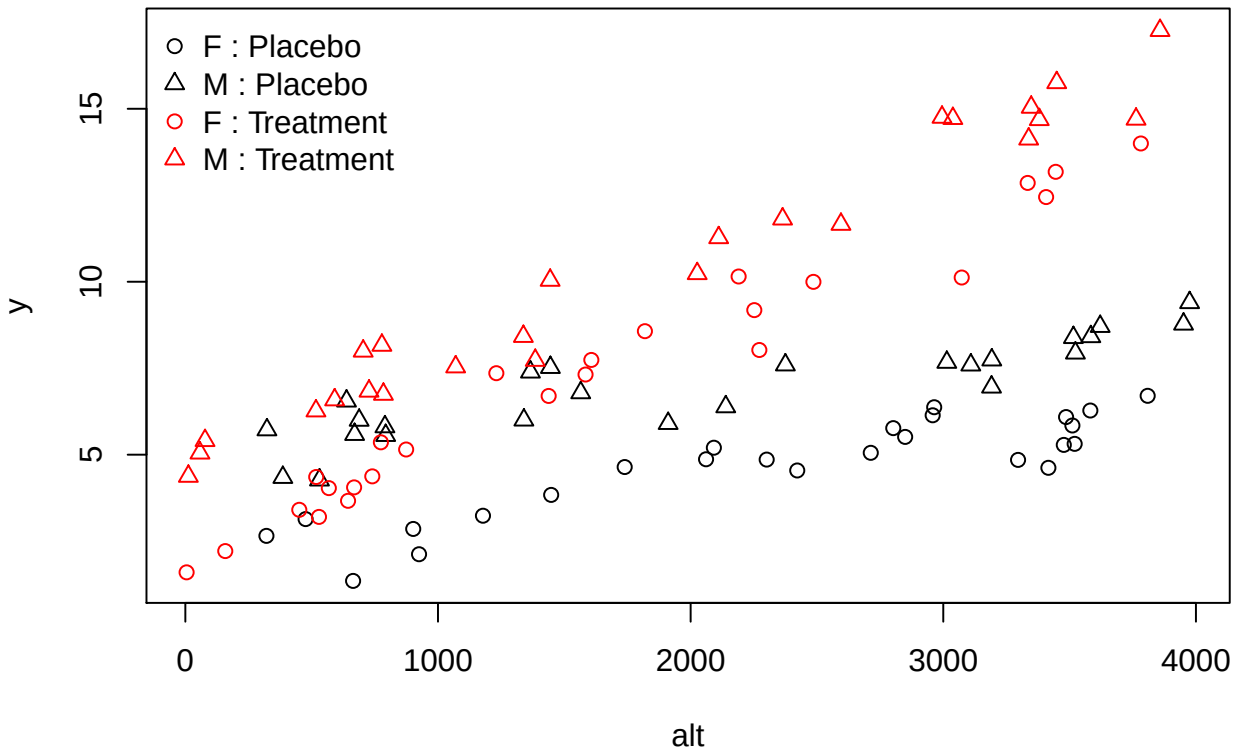
1) Acceptable
2) The variance grows with $\hat{\mu}_i$. Do a weighted analysis or transform the Y'es.
3) Lack of constant term. The regression is possibly erroneously forced trough zero.
4) Bad model, try transforming the Y'es.

Important R functions

```
> fit <- lm(y ~ A + B:x)
> plot(fit)
> anova(fit)
> drop1(fit, test = "F")
> coef(fit)
> predict(fit, newdata = predpoints)
> hatvalues(fit)
> residuals(fit)
> rstandard(fit)
> rstudent(fit)
```

R example

Consider this dataset:



R example

```
> fit0 <- lm(y ~ sex * tmt + sex * tmt * alt)
> drop1(fit0, test = "F")

Single term deletions

Model:
y ~ sex * tmt + sex * tmt * alt
            Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>                   42.983 -68.437
sex:tmt:alt  1  0.077585 43.060 -70.257  0.1661 0.6846
```

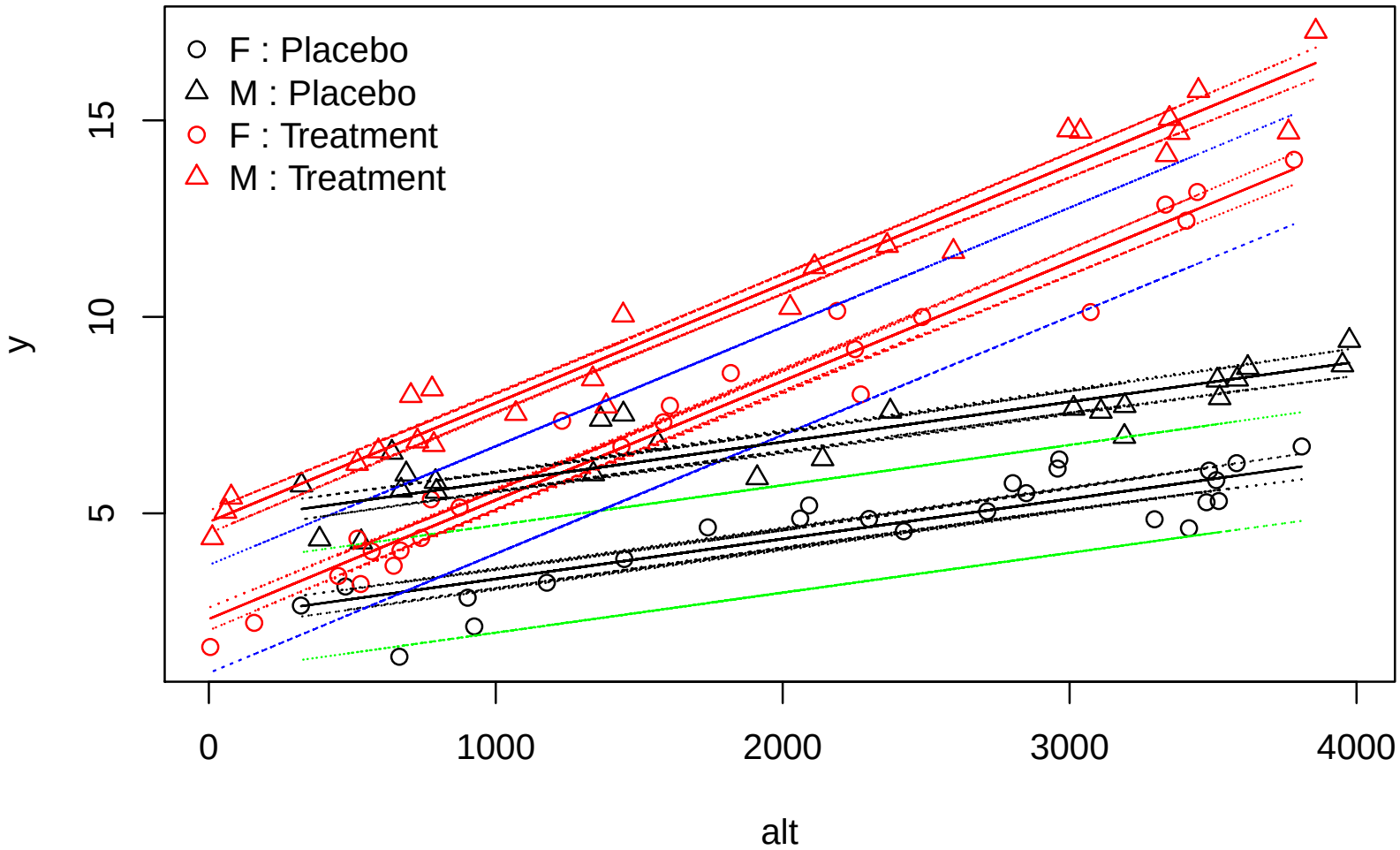
```
> fit1 <- lm(y ~ sex * tmt + (sex + tmt) * alt)
> drop1(fit1, test = "F")

Single term deletions

Model:
y ~ sex * tmt + (sex + tmt) * alt
        Df Sum of Sq     RSS     AIC  F value  Pr(F)
<none>                43.060 -70.257
sex:tmt  1     0.245  43.305 -71.690   0.5287 0.4690
sex:alt  1     0.848  43.909 -70.306   1.8324 0.1791
tmt:alt  1   143.386 186.446  74.297 309.6786 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> fit2 <- lm(y ~ sex + tmt + (sex + tmt) * alt)
> drop1(fit2, test = "F")

Single term deletions

Model:
y ~ sex + tmt + (sex + tmt) * alt
        Df Sum of Sq     RSS     AIC  F value  Pr(F)
<none>                  43.305 -71.690
sex:alt  1     0.694  43.999 -72.101   1.5054 0.2229
tmt:alt  1   143.628 186.933  72.558 311.7645 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> fit3 <- lm(y ~ sex + tmt * alt)
> fit4 <- lm(y ~ sex + tmt:alt)
> drop1(fit3, test = "F")

Single term deletions

Model:
y ~ sex + tmt * alt
        Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                43.999 -72.101
sex      1    150.34 194.338  74.443  324.61 < 2.2e-16 ***
tmt:alt  1    143.95 187.946  71.099  310.80 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(fit4, fit3)

Analysis of Variance Table

Model 1: y ~ sex + tmt:alt
Model 2: y ~ sex + tmt * alt
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     96 44.005
2     95 43.999  1 0.0061976 0.0134 0.9082
```

```
> fit4 <- lm(y ~ sex + tmt:alt)
> drop1(fit4, test = "F")

Single term deletions

Model:
y ~ sex + tmt:alt
        Df Sum of Sq    RSS     AIC F value     Pr(F)
<none>                44.00 -74.087
sex      1    151.90 195.90  73.245  331.38 < 2.2e-16 ***
tmt:alt  2    950.58 994.59 233.716 1036.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results


F : Placebo
M : Placebo
F : Treatment
M : Treatment
y
5
10
15
0
1000
2000
3000
4000
alt