Experimental Design and Data Analysis, Lecture 11

Eduard Belitser

VU Amsterdam

Lecture overview

1. course overview: Lectures 1-10
2. nonlinear regression

course overview: Lectures 1-10

Course overview (1)

In this course we have discussed:

- bootstrap methods
  - confidence intervals — when the distribution of the estimator is unknown
  - bootstrap test — when a statistics T under H_0 can be simulated
- 1 sample tests — for normal and non-normal data:
  - t-test
  - sign test, Wilcoxon rank test
- 2 sample tests — for paired or independent samples:
  - t-test, Pearson correlation, Spearman rank correlation, permutation
  - t-test, Mann-Whitney, Kolmogorov-Smirnov
- permutation tests — any data set in which labels can be permuted

Course overview (2)

- 1-way anova (for k independent normal samples)
- Kruskal-Wallis — nonparametric alternative to 1-way anova
- 2-way anova (k-factorial design, incomplete block designs)
- randomized block design (RBD) — like anova 2-way, without interactions
- repeated measures (RM) — each unit undergoes all treatment levels
- Friedman test — nonparametric alternative to RBD and RM
- cross-over design — incorporating random effects for the individuals
- split-plot design — incorporating random effects for block and block-outer factor interactions

Course overview (3)

- contingency tables for count data (chi-square test, Fisher test)
- linear regression — numerical explanatory variables
  - step-up, step-down, lasso strategies for finding a model
  - outliers, influence points
  - collinearity
- ancova — factor and numerical explanatory variables in one model
- Prediction, feature selection in linear regression (lasso, ridge, elastic net)
- Multiple testing procedures, FDR
- generalized linear models — the outcome Y depends via some link on a linear combination of explanatory variables
  - logistic regression — for 0-1 response variables
  - poisson regression — for count response variables

nonlinear regression

Example: ELISA assay for the recombinant protein

- 176 measurements $(C_1, D_1), \ldots, (C_{176}, D_{176})$ are obtained during development of an ELISA assay for the recombinant protein DNase.

```
> attach(DNase); DNase
     Run       conc  density
1      1  0.04882812    0.017
2      1  0.04882812    0.018
3      1  0.19531250    0.121
[ a lot of output deleted ]
```

- Researcher assumes that the (true) optical density d relates to the protein concentration C (column conc) as follows:

$$d = \frac{\theta_1}{1 + \exp\{[\theta_2 - \log(C)]/\theta_3\}}.$$

Measured optical density ($D_i's$) are in the column density of DNase.

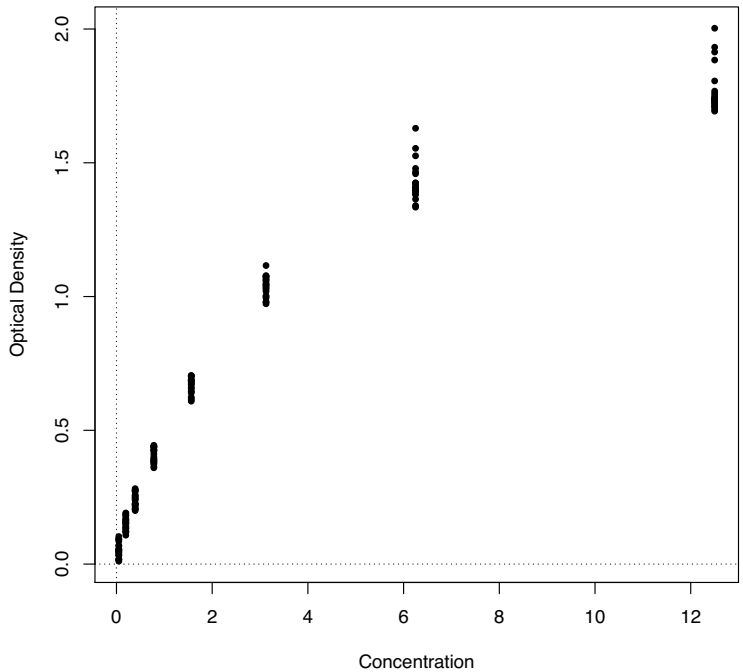
- This is a nonlinear expression: it cannot be manipulated into a linear function with respect to $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$.
- Researcher wishes to obtain an estimate of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, and possibly predict the optical density d for some non-observed protein concetration C.

Example

```
> attach(DNase); n=length(density)
> plot(conc,density,xlab="Concentration",ylab="Optical density")
```

A scatter plot of the data shows a characteristic logistic curve. A plot of the logistic curve using a good estimate $\hat{\theta}$ of θ should fit the data with minimal error.

To conduct this data analysis, we use nonlinear regression techniques.



Nonlinear regression model

- Independent observations $(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)$, Y_i is the response, $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{ki})$ is the vector of the covariates/predictors.
- Nonlinear parametric regression model:

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is the parameter vector, the errors $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent, $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$. (Note that the true distribution of ϵ_i's does not have to be normal.)

- The model is nonlinear if the regression function f is nonlinear, i.e., $f(\boldsymbol{x})$ cannot be written as $f(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots \theta_p x_p$.
- The form of f may be known based on scientific studies, or it may be a guess based on plots of the data.

Least-squares estimator

- Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{f}(\boldsymbol{X}, \theta) = (f(\boldsymbol{x}_1, \theta), \ldots, f(\boldsymbol{x}_n, \theta))^T$.
- The least squares estimator (LSE) of θ is the vector $\hat{\theta}$ that minimizes

$$S(\theta) = \sum_{i=1}^{n} [Y_i - f(\boldsymbol{x}_i, \theta)]^2 = \|\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{X}, \theta)\|^2, \quad \min_{\theta} S(\theta) = S(\hat{\theta}) = RSS(\hat{\theta}).$$

- The LSE $\hat{\theta}$ must satisfy the p normal equations

$$\sum_{i=1}^{n} \frac{\partial f}{\partial \theta_k}(\boldsymbol{x}_i, \theta)[Y_i - f(\boldsymbol{x}_i, \theta)] = 0, \quad k = 1, \ldots, p.$$

- An estimator of σ^2 is $\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n-p} = \frac{RSS(\hat{\theta})}{n-p}$ (similar to linear model).

n = # of measurements

- If f and distribution of errors meet certain conditions, then $\hat{\theta}$ and $\hat{\sigma}^2$ are consistent (as $n \to \infty$) estimators of θ and σ^2, respectively.
- The LSE $\hat{\theta}$ cannot in general be found analytically. Instead, we can use an iterative numerical procedure, such as the Gauss–Newton method.
- Starting with an initial θ^0, this method computes a sequence of estimates $\theta^0, \theta^1, \theta^2, \ldots$ until a convergence criterion is satisfied, to obtain $\hat{\theta}$.
- The initial guess θ^0 may be based on known values for related data, or based on plots of the data.

Example: ELISA assay for the recombinant protein

- Recall: in an assay of recombinant protein, the optical density d is assumed to relate to the protein concentration C according to

$$d = f(C, \theta_1, \theta_2, \theta_3) = \frac{\theta_1}{1 + \exp\{[\theta_2 - \log(C)]/\theta_3\}}.$$

- Given measurements $(C_1, D_1), \ldots, (C_{176}, D_{176})$, the LSE of $(\theta_1, \theta_2, \theta_3)$:

```
> form=as.formula(density~theta1/(1+exp((theta2-log(conc))/theta3)))
> nmodel=nls(form,DNase,start=c(theta1=3,theta2=0,theta3=1)); nmodel
 Nonlinear regression model
  model: density ~ theta1/(1 + exp((theta2 - log(conc))/theta3))
   data: DNase
theta1 theta2 theta3
 2.485  1.518  1.098
 residual sum-of-squares: 0.3801
Number of iterations to convergence: 6
Achieved convergence tolerance: 4.739e-07
```

$RSS(\hat{\theta})$

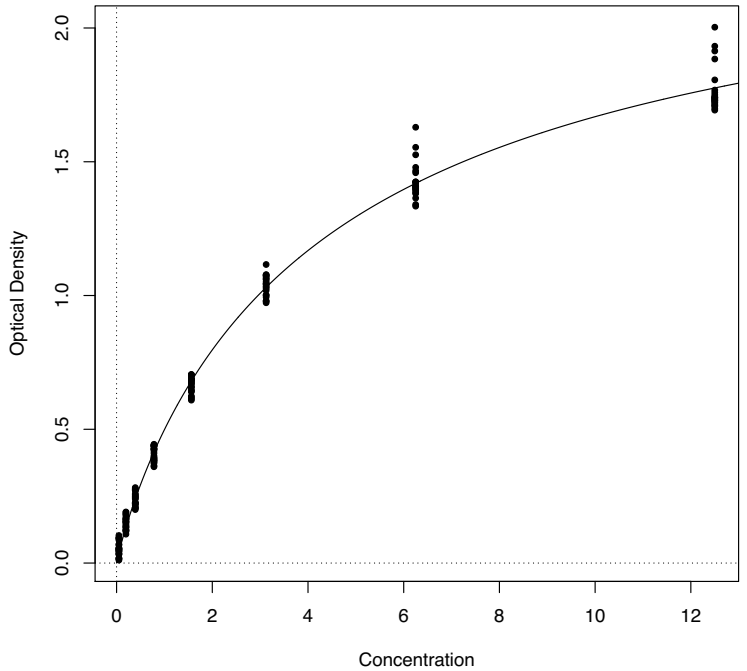
- The Gauss-Newton method with starting value $\theta^0 = (3, 0, 1)$, gives the optimal value $\hat{\theta} = (2.485, 1.518, 1.098)$ after only 6 iterations.
- The estimate of σ^2 is $\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n-p} = \frac{RSS(\hat{\theta})}{n-p} = \frac{0.3801}{176-3} = 0.0022.$

Example

```
> coef(nmodel) # estimates for theta
   theta1    theta2    theta3
2.485319  1.518117  1.098307
> plot(conc,density,xlab="Concentration",ylab="Optical density")
> f=function(x,theta)return(theta[1]/(1+exp((theta[2]-log(x))/theta[3])))
```

A plot of the curve defined by these parameter values fits the data well.

```
> x=seq(from=0.1,to=13,by=0.1)
> lines(x,f(x,coef(nmodel)))
```



Inference about θ: CI's and testing

- Under certain regularity assumptions on f, as $n \to \infty$,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx Z \sim N\big(\mathbf{0}, \sigma^2(\mathrm{V}^T\mathrm{V})^{-1}\big) = N\big(\mathbf{0}, \Sigma\big), \quad \Sigma = \sigma^2(\mathrm{V}^T\mathrm{V})^{-1},$$

the p-variate normal distribution, the $(n \times p)$ matrix $\mathrm{V} = (V_{ij})$,

$$V_{ij} = \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j}, \quad i = 1, \ldots, n, \;\; j = 1, \ldots p.$$

- We estimate V by $\hat{\mathrm{V}} = (\hat{V}_{ij})$, where $\hat{V}_{ij} = \partial f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})/\partial \theta_j$. Then the estimated covariance matrix is $\widehat{\Sigma} = \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\hat{\mathrm{V}}^T\hat{\mathrm{V}})^{-1}$.
- To test $H_0 : \theta_k = a$ (typically $a = 0$) against $H_1 : \theta_k \neq a$, $k = 1, \ldots, p$,

$$\text{test statistics} \;\; T = \frac{\hat{\theta}_k - a}{\sqrt{\widehat{\Sigma}_{kk}}} \approx Z \sim N(0,1) \;\; \text{under} \;\; H_0.$$

- By analogy with the lin. regr., t_{n-p}-distr. is used (instead of $N(0,1)$): reject H_0 if the p-value $p = P(|T| \geq |t|) = 2P(T \leq -|t|) < \alpha$, for a significance α, $T \sim t_{n-p}$ and the value t of the test statistics. (Or, test by checking whether $|T| > t_{n-p;1-\alpha/2}$, or, by checking whether a lies in the CI.
- Approx. $(1-\alpha)100\%$ CI for θ_k, $k = 1, \ldots, p$, is $\hat{\theta}_k \pm t_{n-p;1-\alpha/2}\sqrt{\widehat{\Sigma}_{kk}}$.

Example

- Returning to the example, the estimated covariance matrix $\widehat{\Sigma}$ is

```
> cov.est=vcov(nmodel);cov.est
            theta1      theta2       theta3
theta1 0.003952687 0.003976003 0.0013907552
theta2 0.003976003 0.004091625 0.0014269021
theta3 0.001390755 0.001426902 0.0005963426
```

- Tests for θ_i's (the output also contains the estimate $\hat{\sigma} = 0.04687$):

```
> p=length(coef(nmodel)); summary(nmodel)
Parameters:
       Estimate Std. Error t value Pr(>|t|)
theta1  2.48532    0.06287   39.53   <2e-16 ***
theta2  1.51812    0.06397   23.73   <2e-16 ***
theta3  1.09831    0.02442   44.98   <2e-16 ***
  ...
  Residual standard error: 0.04687 on 173 degrees of freedom
```

- Approximate 95% CI's for θ_1, θ_2, θ_3 are obtained by `confint(nmodel)`.
These CI's slightly differ from directly computed, for example, CI for θ_1 is $\hat{\theta}_1 \pm t_{n-p;1-\alpha/2}\sqrt{\widehat{\Sigma}_{11}}$, left and right ends of CI computed in R as

```
>coef(nmodel)[1]-qt(0.975,n-p)*sqrt(cov.est[1,1]) #left end
>coef(nmodel)[1]+qt(0.975,n-p)*sqrt(cov.est[1,1]) #right end
```

Mean response at any covariate

- Suppose $\boldsymbol{x} = (x_1, \ldots, x_p)$ is any covariate vector for which we would like to know the expected value of the response variable,

$$E(Y|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}).$$

- We can use the parameter estimates $\hat{\boldsymbol{\theta}}$ derived from nonlinear regression on the observed data $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$.
- Clearly, $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ is a good predictor of the mean response at value $\boldsymbol{x}$.
- $\boldsymbol{x}$ does not have to be one of the observed covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.
- Then by the Taylor expansion and asymptotic normality of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$,

$$f(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) - f(\boldsymbol{x}, \boldsymbol{\theta}) \approx v_{\boldsymbol{x}}^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx Z \sim N\left(0, v_{\boldsymbol{x}}^T \Sigma v_{\boldsymbol{x}}\right), \quad v_{\boldsymbol{x}} = \nabla f(\boldsymbol{x}, \boldsymbol{\theta}).$$

- But $\boldsymbol{\theta}$, σ are unknown, we substitute their estimates to obtain the estimate $\hat{v}_{\boldsymbol{x}} = \nabla f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ of $v_{\boldsymbol{x}}$ and use the estimate $\widehat{\Sigma}$ instead of true covariance matrix Σ. Then

$$\frac{f(\boldsymbol{x}, \hat{\theta}) - f(\boldsymbol{x}, \boldsymbol{\theta})}{\sqrt{\hat{v}_{\boldsymbol{x}}^T \widehat{\Sigma} \hat{v}_{\boldsymbol{x}}}} \approx Z \sim N(0, 1).$$

Testing and confidence interval for the mean response

- Test $H_0 : f(\boldsymbol{x}, \boldsymbol{\theta}) = a$ (typically $a = 0$) against $H_1 : f(\boldsymbol{x}, \boldsymbol{\theta}) \neq a$. Then

$$T = \frac{f(\boldsymbol{x}, \hat{\theta}) - a}{\sqrt{\hat{v}_{\boldsymbol{x}}^T \widehat{\Sigma} \hat{v}_{\boldsymbol{x}}}} \approx Z \sim N(0, 1) \quad \text{under} \quad H_0.$$

But, by analogy with the linear regression, t_{n-p}-distribution is used.

- If $|T| > t_{n-p;1-\alpha/2}$, reject H_0 (or, by computing the p-value, or check whether a lies in the corresponding confidence interval).
- When asymptotic normality holds, an approximate $(1 - \alpha)100\%$ confidence interval for the mean response $E(Y|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta})$ is then

$$f(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \pm t_{n-p;1-\alpha/2} \sqrt{\hat{v}_{\boldsymbol{x}}^T \widehat{\Sigma} \hat{v}_{\boldsymbol{x}}}.$$

Example

- Recall that the optical density d is assumed to relate to the protein concentration C as $d = f(C, \theta) = \frac{\theta_1}{1+\exp\{[\theta_2 - \log(C)]/\theta_3\}}$.
- By nonlinear regression we obtained $\hat{\theta} \approx (2.49, 1.52, 1.1)$ and $\widehat{\Sigma}$.
- Suppose we wish to estimate the expected value of D when $C = 4$, i.e., $f(4, \theta)$. We compute $f(4, \hat{\theta}) = 1.1682$, $t_{173;0.95} = 1.6537$,

$$\hat{v}_4 = \nabla f(4, \hat{\theta}) = (0.47003, -0.5637, 0.0677)^T, \quad \sqrt{\hat{v}_4^T \widehat{\Sigma} \hat{v}_4} = 0.007.$$

- Appr. 90% CI for $f(4, \hat{\theta})$ is $1.1682 \pm 1.6537 \cdot 0.007 = [1.1566, 1.1797]$.
- In R, one needs to program the function $f(\boldsymbol{x}, \theta)$ and the gradient $\nabla f(\boldsymbol{x}, \theta)$ as, say, functions `f(x,theta)` and `grad(x,theta)`. Then

```
> f4=f(4,coef(nmodel)); gradvec=grad(4,coef(nmodel))
> se=sqrt(t(gradvec)%*%vcov(nmodel)%*%gradvec)
> lb=f4-qt(0.95,n-p)*se; ub=f4+qt(0.95,n-p)*se; c(lb,ub) # CI
```

Check model assumptions

- To check the model assumptions, compute the residuals (as for lin. regr.):

$$\hat{\epsilon}_i = Y_i - f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}), \quad i = 1, \ldots, n,$$

or standardized residuals $[Y_i - f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})]/\hat{\sigma}$ against the fitted values

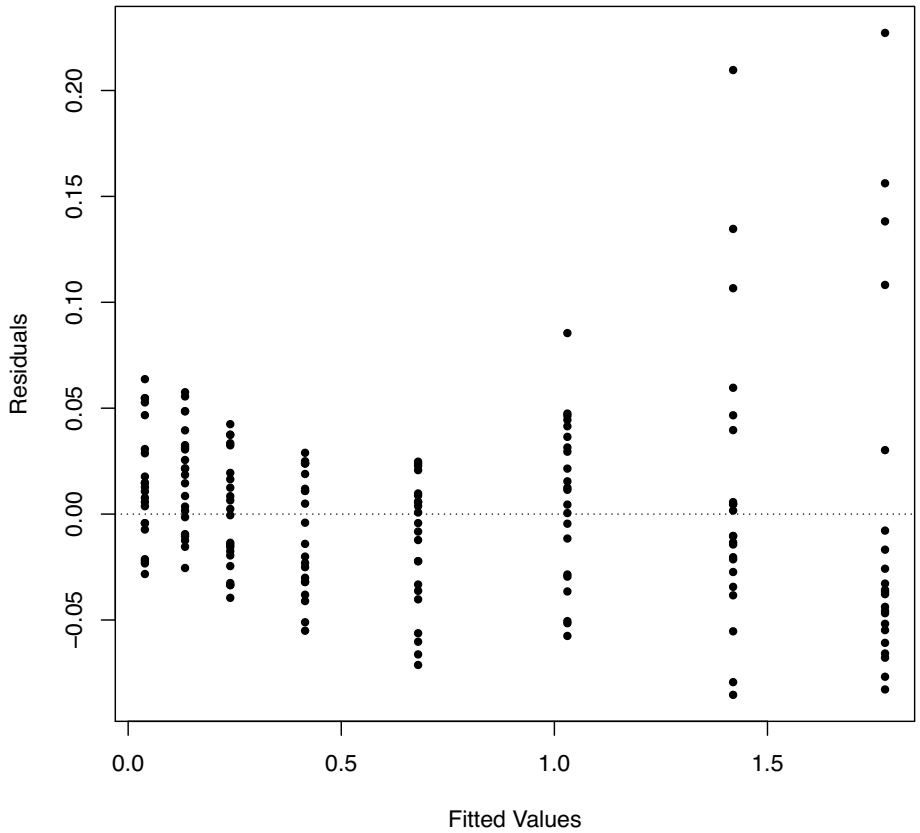
$$\hat{Y}_i = f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}), \quad i = 1, \ldots, n.$$

- The spread of the points should be fairly constant within a horizontal band symmetric about the horizontal axis.
- If not, some transformation of the response variable may be needed.
- To check the normality of the errors, examine a normal qq-plot of the residuals.

Example

A plot of the residuals against the fitted values indicates that the assumption of constant error variance may not be valid. Also, the plot shows a functional pattern, which suggests that the regression function $f(C, \theta)$ may not be appropriate for these data.

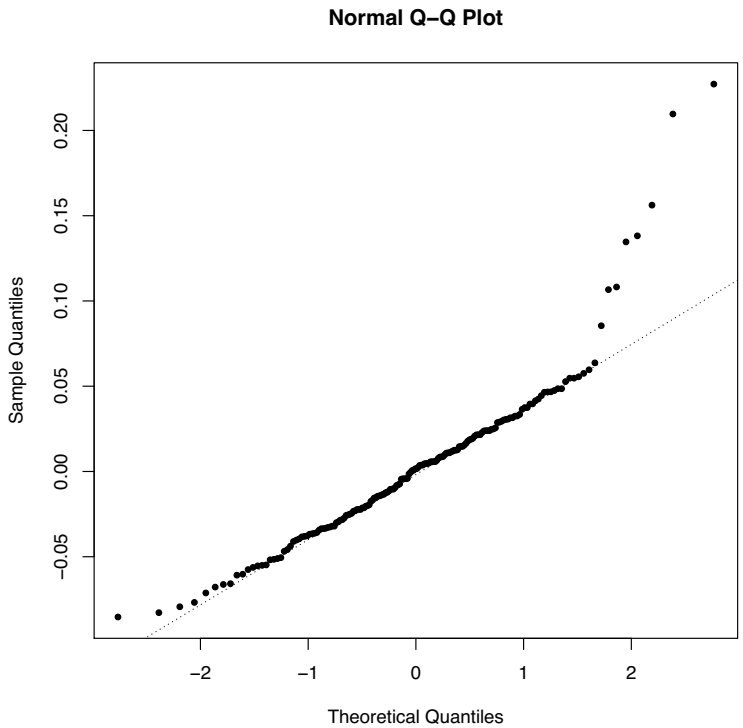
```
> plot(fitted(nmodel),resid(nmodel))
> abline(h=0,lty=3)
```

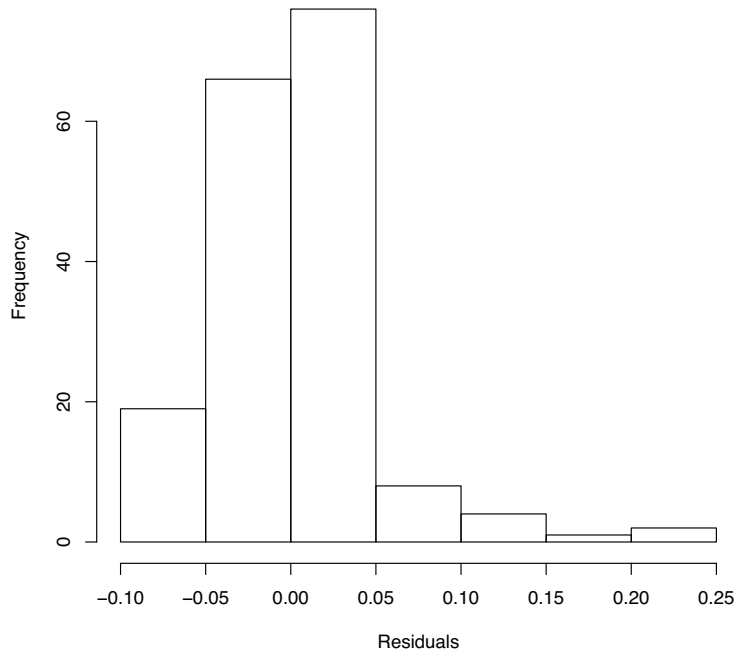


Example

A normal qq-plot of the residuals indicates that an assumption of normality for the errors may not be valid. The errors appear to be skewed. This is also confirmed by a histogram of the residuals.

```
> qqnorm(resid(nmodel)); qqline(resid(nmodel)); hist(resid(nmodel))
```





Comparing nested models

- Model ω with parameter $\boldsymbol{\theta}_p \in \mathbb{R}^p$ is nested within a global model Ω with parameter $\boldsymbol{\theta}_q \in \mathbb{R}^q$ $(q > p)$, e.g., by setting some coordinates of $\boldsymbol{\theta}_q$ to some constants. If $\boldsymbol{\theta}_q = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_{q-p})$, where $\boldsymbol{\theta}_q \in \mathbb{R}^q$ and $\boldsymbol{\theta}_{q-p} \in \mathbb{R}^{q-p}$,

$$\Omega : Y = f_\Omega(\boldsymbol{x}, \boldsymbol{\theta}_q) + \epsilon, \quad \omega : Y = f_\omega(\boldsymbol{x}, \boldsymbol{\theta}_p) + \epsilon.$$

  so that $f_\omega(\boldsymbol{x}, \boldsymbol{\theta}_p) = f_\Omega(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_q)$, where $\bar{\boldsymbol{\theta}}_q = (\boldsymbol{\theta}_p, \boldsymbol{a})$ for some $\boldsymbol{a} \in \mathbb{R}^{q-p}$.
- If the (asymptotic) normality assumption for the errors is valid, we can test whether the nested (reduced) model ω is adequate.
- We first fit the global model Ω: $S(\hat{\boldsymbol{\theta}}_q) = \min_{\boldsymbol{\theta}_q} \|\boldsymbol{Y} - f_\Omega(\boldsymbol{x}, \boldsymbol{\theta}_q)\|^2$, then we fit the reduced model ω: $S(\hat{\boldsymbol{\theta}}_p) = \min_{\boldsymbol{\theta}_p} \|\boldsymbol{Y} - f_\omega(\boldsymbol{x}, \boldsymbol{\theta}_p)\|^2$.
- Clearly, $S(\hat{\boldsymbol{\theta}}_q) \leq S(\hat{\boldsymbol{\theta}}_p)$. If $S(\hat{\boldsymbol{\theta}}_p) - S(\hat{\boldsymbol{\theta}}_q)$ is large, ω is not adequate.
- Exactly, the reduced model ω is not adequate for describing the data at significance level α if (idea: $(S(\hat{\boldsymbol{\theta}}_q) - S(\hat{\boldsymbol{\theta}}_p))/\sigma^2 \sim \chi^2_{q-p}$, $S(\hat{\boldsymbol{\theta}}_q)/\sigma^2 \sim \chi^2_{n-q}$.)

$$V = \frac{[S(\hat{\boldsymbol{\theta}}_p) - S(\hat{\boldsymbol{\theta}}_q)]/(q-p)}{S(\hat{\boldsymbol{\theta}}_q)/(n-q)} > F_{q-p,n-q;1-\alpha}.$$

- Equivalently, perform test by computing the p-value $P(F > v)$, where $F \sim F_{q-p,n-q}$ and v the realized value of statistic V.

Example

- All p-values are very small, so in principle it makes no sense to set any of them to zero. However, suppose we drop θ_2 (i.e., we set $\theta_2 = 0$) to obtain a nested submodel ω consisting only of θ_1 and θ_3: $D = f(C, \theta_1, \theta_3) = \frac{\theta_1}{1+\exp[-\log(C)/\theta_3]}$.
- For the reduced model ω, compute $S(\hat{\theta}_2) = 6.6503$. For the full model Ω, compute $S(\hat{\theta}_3) = 0.38$. We reject the submodel ω because the F statistic $\frac{(6.6503-0.38)/(3-2)}{0.38/(176-3)} = 2852.9 > F_{1,173,0.95} = 3.9$. The same analysis in R:

```
> form2=as.formula(density~theta1/(1+exp((-log(conc))/theta3)))
> nmodel2=nls(form2,DNase,start=c(theta1=3,theta3=1)) # submodel
> anova(nmodel,nmodel2)
Model 1: density ~ theta1/(1 + exp((-log(conc))/theta3))
Model 2: density ~ theta1/(1 + exp((theta2 - log(conc))/theta3))
  Res.Df Res.Sum Sq Df Sum Sq F value    Pr(>F)
1    174     6.6481
2    173     0.3801  1  6.268  2852.9 < 2.2e-16 ***
```

- Thus reject the submodel ω and retain the full model Ω.