stats545.5 545.5 Extending the linear model

Gunnar Stefansson

September 3, 2022

Contents

1 Introduction to nonlinear statistical models

1.1 The assumptions - and what goes wrong

$$\mathbf{y} \sim n\left(\mathbf{X}\beta, \sigma^2 I\right)$$

may be wrong.
Simple variance exceptions are easy to handle:

- $Vy_i = u_i\sigma^2$ where u_i are known
- $\Sigma_{\mathbf{y}} = \sigma^2 B$ where B is known
- $\Sigma_{\mathbf{y}}$ may contain "a few" unknown parameters

1.1.1 Details

The assumption

$$\mathbf{y} \sim n\left(\mathbf{X}\beta, \sigma^2 I\right)$$

consists of several components: Normality, independence, linearity in the mean and equal variances.

The following sections give some examples of cases where general violations of these assumptions are handled using general nonlinear models.

Consider first the simplest deviations, namely of the variance structure:

- If $Vy_i = u_i\sigma^2$ where u_i are known, then we can define $w_1 = 1/u_i$ and maximum likelihood is equivalent to $min_{\mathbf{b}} \sum_i w_i \left(\mathbf{y}_i - \mathbf{x}_i'\beta\right)^2$ where $\mathbf{x}_i'$ is the i'th row of $\mathbf{X}$. This is the traditional **weighted linear regression**.
- If $\Sigma_{\mathbf{y}} = \sigma^2\mathbf{B}$ where $\mathbf{B}$ is known, then we can write $\mathbf{B}^{-1} = \mathbf{L}\mathbf{L}'$ where $\mathbf{L}$ is lower triangular and define a new regression problem with $\tilde{\mathbf{y}} = \mathbf{L}'\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{L}'\mathbf{X}$ and it follows that $E\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta$, $V\tilde{\mathbf{y}} = \ldots = \sigma^2\mathbf{I}$ so ordinary least squares can be used to estimate the parameters and for uncertainty estimation in the revised regression problem.
- If $\Sigma_{\mathbf{y}}$ contains "a few" unknown parameters, then these can be estimated as a part of maximum likelihood estimation of all parameters.

1.2 Maximum likelihood

The MLE is usually a good estimator
Applies to very many estimation problems
Need to specify the complete likelihood function
Can take into account dependence, different variances, non-normality, nonlinear response etc

1.2.1 Details

the likelihood approach involves...

1.2.2 Examples

Example: Consider maximum likelihood estimation of the mean of the gamma density.

$$\ldots$$

$$\Rightarrow \hat{\mu} = \bar{y}$$

Example: Consider a model for the growth of fish.

The data set at http://notendur.hi.is/gunnar/kennsla/alsm/data/set121.dat contains measurements of individual fish, collected by the Marine Research Institute (http://www.hafro.is). The data include a column (aldur) containing the age of fish and the column (le) containing the length of the same fish.

The von Bertalanffy growth model can be fitted using the R commands

```
dat<-read.table("http://notendur.hi.is/~gunnar/kennsla/alsm/data/set121.dat",header=T)
le<-dat$le
a<-dat$aldur
fm<-nls(le~Linf*(1-exp(-K*(a-t0))),start=list(t0=0,Linf=80,K=0.25))
summary(fm)
```

Once the above commands have been issued, the summary command can be used:

```
> summary(fm)

Formula: le ~ Linf * (1 - exp(-K * (a - t0)))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
t0   -0.23160    0.23739  -0.976 0.331683
Linf 91.22292   14.47924   6.300 8.72e-09 ***
K     0.15672    0.04414   3.550 0.000595 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.788 on 97 degrees of freedom

Number of iterations to convergence: 3
Achieved convergence tolerance: 6.375e-07
```

A different test can also be used to investigate whether $t_0 = 0$:

```
> fmR<-nls(le~Linf*(1-exp(-K*(a))),start=list(Linf=80,K=0.25))
> anova(fm,fmR)
Analysis of Variance Table

Model 1: le ~ Linf * (1 - exp(-K * (a - t0)))
Model 2: le ~ Linf * (1 - exp(-K * (a)))
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     97     753.78
2     98     762.33 -1 -8.557  1.1012 0.2966
```

Note that the F-test and t-test are not the same in the nonlinear case. Both depend on linearity assumptions but in different ways.

1.3 Nonlinear least squares

Common model:

$$Ey_i = g(\mathbf{x_i}, \beta),\ 1 \leq i \leq n$$

Common estimation method:

$$min_\beta \sum_i (y_i = g(\mathbf{x_i}, \beta))^2$$

Also need to estimate uncertainty. Use Hessian or bootstrap.

1.3.1 Detail

A common nonlinear model to predict y_i-values is to assume some functional relationship:

$$Ey_i = g(\mathbf{x_i}, \beta),\ 1 \leq i \leq n$$

where β is a vector of unknown parameters.

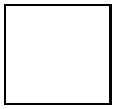
A very common estimation method is nonlinea least squares:

$$min_\beta \sum_i (y_i = g(\mathbf{x_i}, \beta))^2 .$$

This leaves the question of how to estimate the uncertainty in the parameters.

2 Generalized linear models

2.1 The linear model - different formulation



2.1.1 Details

The traditional linear model with independent $y_1, y_2, ..., y_n$ can be written as follows

1. $y_i \sim n(\mu_i, V[y_i])$
2. $\mu_i = \eta_i$
3. $V[y_i] = \sigma^2$

where $\eta_i = \underline{x}_i'\underline{\beta}$.

2.2 The generalized linear model

2.2.1 Details

The generalized linear model with independent $y_1, y_2, ..., y_n$ can be defined by

1. $y_1, y_2, ..., y_n$ come from a member of the exponential family of distributions.
2. $g(\mu_i) = \eta_i$ for some monotone function g, where $\eta_i = \underline{x}_i'\underline{\beta}$ and $V[y_i]$ may be a function of μ_i.

The function g is the **link function** and n_i is the **linear predictor**. (The link function links the mean to the linear predictor.)

The traditional linear model is a special case of this generalized linear model.

The **exponential family** consists of distributions with density (pdf or pnf) of the form

$$f(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)} \quad (*)$$

where a, b and c are functions which make this a density. Commonly, $a(\phi) = \frac{\phi}{w}$ where $w > 0$ is known and $\phi = V[y_i] = \sigma^2$.

2.2.2 Examples

Example 1 - The Gaussian Density

Consider the usual linear model

$$\mathbf{y} \sim n(\mathbf{X}\beta, \sigma^2 I)$$

Here, $y_i \sim n(\mu_i, \sigma^2)$ where $\mu_i = \mathbf{x}_i'\beta$ are independent and $g =$ identity. We need to verify that the Gaussian density is of the form (*), that is,

$$f(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

This Gaussian density is

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2 - 2\mu y + \mu^2}{2\sigma^2}} \\ &= \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2) \right) \right) \end{aligned}$$

which is of the form (*) if we set

$\phi = \sigma^2$,

$\theta = \mu$,

$a(\phi) = \phi$, $b(\theta) = \frac{1}{2}\theta^2$ and

$c(y,\phi) = -\left(\frac{y^2}{2\phi} + \frac{1}{2}\ln(2\pi\phi) \right)$.

Example 2 - The Γ-Density

First write the gamma density in the form

$$f(y;\mu,r) = \frac{y^{r-1} e^{\frac{-ry}{\mu}}}{\Gamma(r)\left(\frac{\mu}{r}\right)^r} \qquad \text{with } y > 0$$

If e.g. $y_1, y_2, ..., y_n$ are independent with this density and $\mu_i = E[y_i] = g(\mathbf{x}_i'\beta)$, then this is a GLM since the density is from the exponential family:

$$f(y;\mu,r) = \exp\left(\frac{y/\mu - \ln(\mu)}{1/r} + (r-1)\ln(y) - \ln(\Gamma(r)) + r\ln(r) \right)$$

which is of the form (*) with

$\theta = \frac{1}{\mu}$,

$\phi = r$,

$a(\phi) = \frac{1}{\phi}$,

$b(\theta) = -\ln(\theta)$ and

$c(y,\phi) = (\phi - 1)\ln(y) - \ln(\Gamma(\phi)) + \phi\ln(\phi)$.

Here $g(\mu) = \frac{1}{\mu}$ is called the **canonical link**, but it is much more common to use $g(\mu) = \ln(\mu)$, e.g. in fisheries or other ecological applications.

Example 3

Consider estimating the mean under the gamma assumption. So assume we have independent $y_1, y_2, ..., y_n$ with density

$$\frac{y^{r-1}e^{-ry/\mu}}{\Gamma(r)\left(\frac{\mu}{r}\right)^r} \qquad \text{with } y > 0$$

The maximum likelihood estimator for μ is obtained by maximizing the likelihood function:

$$L(\mu, r) = \prod_{i=1}^{n} \frac{y_i^{r-1}e^{-ry_i/\mu}}{\Gamma(r)\left(\frac{\mu}{r}\right)^r}$$

or maximizing

$$\ln(L(\mu, r)) = -n \cdot \ln\left(\Gamma(r)\frac{1}{r^r}\right) - nr \cdot \ln(\mu) + (r-1)\sum_{i=1}^{n} \ln(y_i) - \frac{r}{\mu}\sum_{i=1}^{n} y_i$$

and now we solve

$$0 = \frac{d}{d\mu}\ln(L(\mu, r)) = -\frac{nr}{\mu} + \frac{r\sum_{i=1}^{n} y_i}{\mu^2}$$

$$\Rightarrow \qquad n\mu = \sum_{i=1}^{n} y_i$$

and we obtain

$$\hat{\mu} = \overline{y}$$

Note that here $Ey_i = \mu(= \alpha\beta)$, so $\hat{\mu}$ is unbiased.

2.3 Estimation: MLEs and deviances

2.3.1 Details

Parameters in a GLM include both $\underline{\beta}$ and ϕ. The elements of $\underline{\beta}$ can be estimated using maximum likelihood, but several approaches exist for this. The likelihood function is in general

$$L(\underline{\mu}; \underline{y}) = \prod_{i=1}^{n} e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)},$$

where one needs to think of θ_i as an appropriate function of the expected value, μ_i.

If we define

$$l(\underline{\mu}, \underline{y}) := -2\ln\left(L(\underline{\mu}; \underline{y})\right)$$

as a natural quantity to be minimized, then we can also define the scaled deviance by

$$D^*(\underline{y}, \underline{\mu}) := l(\underline{\mu}, \underline{y}) - l(\underline{y}, \underline{y})$$

and the deviance is

$$D(\underline{y}; \mu) := \phi D^*(\underline{y}, \underline{\mu})$$

We note that

1. Minimizing D over $\underline{\beta}$ is equivalent to maximizing the likelihood function.
2. If $a(\phi) = \frac{\phi}{w}$ then D does not include ϕ.

For a given model, the deviance for the model is $D(\underline{y}, \underline{\hat{\mu}})$ where

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(\underline{x}_i'\underline{\hat{\beta}})$$

is the MLE, and this quantity only depends on the data - it does not involve unknown parameters $\left(\text{if } a(\phi) = \frac{ph}{w}\right)$.

2.3.2 Examples

Examples of distributions covered by the exponential family include the normal, gamma, Poisson and binomial. Common link functions include

Family	Link function
Gaussian	id
Gamma	ln (or inverse)
Binomial	logistic (or probit)
Poisson	ln

Example 4

Consider predicting a value between zero and one. Within a small interval the probability might be linear, but obviously an assumption of linearity will not hold across a wide range.

$$y_i \sim b(n = 1, p_i)$$

so $Ey_i = p_i$ and $Vy_i = p_i(1 - p_i)$.

A common link function is the logistic function

$$\eta = g(\mu) = \text{logit}(p) := \ln\left(\frac{p}{1-p}\right) = \mathbf{x}'\beta$$

with inverse

$$\frac{1}{1 - e^{-\mathbf{x}'\beta}} = \frac{1}{1 + e^{-\eta}}$$

Note that the data are NOT transformed.

The p.m.f. and likelihood

$$\prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i)$$

where each y_i is 0 or 1. Here

$$\begin{aligned} l(\mathbf{y},\mathbf{p}) &= -2\ln\big(L(\mathbf{p};\mathbf{y})\big) \\ &= -2\sum_{i=1}^{n}\left(y_i\ln(p_i)+(1-y_i)\ln(1-p_i)\right) \end{aligned}$$

and $\phi = 1$.

Numerical methods are available for finding these values.

By transforming the data one gets biased estimators, so the GLM estimation technique is generally preferred.

Example 5 - Poisson

$y_i \sim P(\lambda_i)$ are independent and usually $\ln(\lambda_i) = \mathbf{x}_i'\beta$. Here

$$L(\underline{\lambda}) = \prod_{i=1}^{n}\frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!}$$

and

$$\begin{aligned} l(\mathbf{y},\underline{\lambda}) &= -2\sum_{i=1}^{n}\ln\left(\frac{l_i^{y_i}e^{-\lambda_i}}{y_i!}\right) \\ &= -2\sum_{i=1}^{n}\left(y_i\ln(\lambda_i)-\lambda_i-\ln(y_i!)\right) \end{aligned}$$

and

$$\begin{aligned} D(\mathbf{y},\hat{\underline{\lambda}}) &= l(\hat{\underline{\lambda}},\mathbf{y}) - l(\mathbf{y},\mathbf{y}) \\ &= -2\sum(y_i\ln(\hat{\lambda}_i)-\hat{\lambda}_i-y_i\ln(y_i)+y_i) \\ &= -2\sum\left(y_i\ln\frac{\hat{\lambda}_i}{y_i}+(y_i-\hat{\lambda}_i)\right) \end{aligned}$$

and this needs to be minimized to find $\hat{\underline{\beta}}$, where $\hat{\lambda}_i = e^{\mathbf{x}_i'\beta}$.

3 Likelihood ratio tests

3.1 The LRT

The **likelihood ratio test** of $H_0 : \beta = \beta_0$ vs $H_a : \beta \neq \beta_0$

$$\lambda = \frac{L(\hat{\beta}_0)}{L(\hat{\beta})}.$$

The test rejects for large values of λ.

3.1.1 Handout

The **likelihood ratio test** of $H_0 : \beta = \beta_0$ vs $H_a : \beta \neq \beta_0$ proceeds by computing the ratio between the likelihood function values, at the two points, $L(\beta_0)$ and $L(\hat{\beta})$. This is the ratio

$$\lambda = \frac{L(\hat{\beta}_0)}{L(\hat{\beta})}.$$

The test rejects for large values of Λ. Note that $\lambda = \lambda(y_1, \ldots, y_n)$ is a function of the data so the rejection region translates into a statement about the data.

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta)}{\sup\limits_{\theta \in \Theta} L(\theta)} = \frac{L(\theta_0)}{L(\hat{\theta})}.$$

Theorem: (Asymptotic distribution of the LRT–simple H_0) For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, suppose $x_1, \ldots, X_n$ are i.i.d. $f(x|\theta)$, $\hat{\theta}$ is the MLE for θ (and $f(x|\theta)$ satisfies the regularity conditions found in Miscellanea 10.6.2. in Casella and Berger). Then under H_0, as $n \to \infty$,

$$-2\log\lambda(\mathbf{X}) \xrightarrow{\mathcal{D}} \chi^2_1$$

4 Case study: Fisheries

4.1 Biological systems are typically nonlinear

- Growth
- Mortality

⋮

•

so the models should be nonlinear

4.2 A relatively simple problem, ADAPT

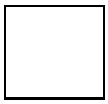
The ADAPT assessment model

$$min_{N_{0,y},N_{a,0},q_a} \sum_{ay} w_{ay} \left(ln(I_{ay}) - ln(q_a N_{ay})\right)^2$$

$$\text{w.r.t.} N_{a+1,y+1} = \left(N_{ay} e^{-M/2} - C_{ay}\right) e^{-M/2}$$

where M is fixed and the catches, C_{ay} are given as numbers by age and year.
But the weighting factors w_{ay} need to be specified.

4.3 Gadget biological components



4.3.1 Details

Core: Parametric forward simulation model

- Consumption: Suitability functions
- Mortality: Due to predation or other natural or fishing
- Growth: Can depend on consumption. Several growth update implementations
- Migration: Through migration matrices
- Maturation: Move from immature to mature stock component
- Spawning: Lose weight and generate yearclass
- Symmetric: All species implemented in same way - fleet is also a predator

4.4 Data are typically not Gaussian

- Length distributions
- Survey indices

⋮

-

Data from a normal distribution are actually very rare in fishery science.
Obvious modifications to assumptions such as the multinomial typically does not improve anything.

4.5 Nonlinearity is not an issue per se

- Use nonlinear minimisation algorithms for estimation
- Can handle a lot of unknown parameters
- Can in principle estimate variances using Hessian matrices or bootstrap

4.6 Consider each data set

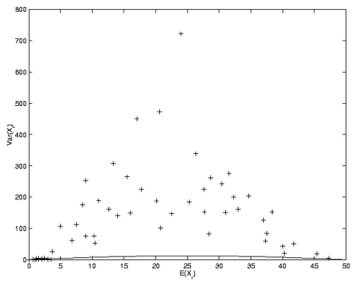
Look at single data sets and try to estimate true variances in each
Compare point estimates from each data set
Try to test formally whether results differ

4.7 Diagnostics for likelihood functions

Most likelihood functions can be verified, e.g. using Kolmogorov-Smirnov tests.
One should not be happy with a model which is rejected!

4.8 Likelihoods - Assumption

Take 50 fish from each station - compare with binomial



4.8.1 Details

Take 50 fish from each station - compare with binomial

4.9 Parsimony and flexibility

If data sources indicate different outcomes then the model is wrong!
Data are just data - they are not wrong.
Example: Catchability may vary in time and fleets may increase their catchability.
Need to add parameters until model is appropriately flexible. Notably add time series parameters...

5 Case study: Multispecies models for marine fish stocks

5.1 Combining data sets raises issues

- Weight given to each
- Do they all indicate the same model?

⋮

-

5.2 Several data sets means several likelihood components

In ADAPT

$$min_{N_{0,y},N_{a,0},q_a} \sum_{ay} w_{ay} \left(ln(I_{ay}) - ln(q_a N_{ay})\right)^2$$

$$\text{w.r.t.} N_{a+1,y+1} = \left(N_{ay} e^{-M/2} - C_{ay}\right) e^{-M/2}$$

the weighting factors w_{ay} need to be specified, since age groups are like data sets.
Complex data means means the components are not even of same form!

5.3 Length distributions

Multinomial?
Test assumptions using samples of survey stations, picking n fish from each.
Variance should be from binomial.
Covariance from multinomial.
Conclusion: Assumption fails very badly.

5.4 Effect of wrong variance assumptions

Linear model theory: Minor issue, just affects variance estimates, parameter estimates are still unbiassed.
But: If the base model is wrong for a small part of the data, may create havoc!

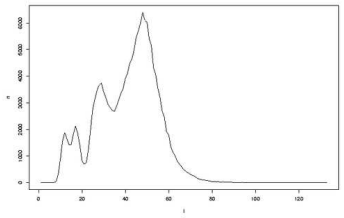
5.4.1 Examples

Example: Wrong weights in ADAPT

Weights on juveniles seem important - can drive entire assessment.

5.5 Likelihoods - Estimation procedure

Gadget is a statistical estimation model.
Internal dynamics are complex so deterministic forward projections are used.
Maximum likelihood estimation is used.



Length distributions are count data and are often assumed to come from a multinomial distribution, possibly with overdispersion.

5.5.1 Details

Estimation: (Negative log) likelihood functions

Gaussian, weighted

Multinomial

$$min_{\theta \in \mathbf{R}^n} \sum_k w_k l_k(\theta)$$

*N*ote: Given θ, can simulate. Now search for estimate giving best fit to weighted likelihood function.

5.6 Simple example of complexity problem

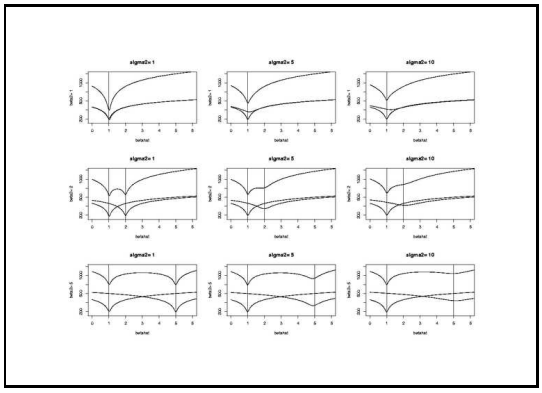
Take a simple problem

$$Y_{ij} \sim n(\alpha_i + \beta_i x_{ij}, \sigma_i^2), \quad j = 1, \ldots, n_i \;\; i = 1, 2,$$

but suppose we don't know the slopes are different, so fit

$$Y_{ij} \sim n(\alpha_i + \beta x_{ij}, \sigma_i^2), \quad j = 1, \ldots, n_i \;\; i = 1, 2,$$

5.7 Simple example of complex problem



6 Addding distributional assumptions: The multivariate normal and related distributions

6.1 A theorem from calculus

6.1.1 Handout

To find a multivariate density of a transformed variable, recall from calculus that if g is a $1-1$ function $g : \mathbf{R}^n \to \mathbf{R}^n$

$$\int f(\mathbf{z})d\mathbf{z} = \int f(g(\mathbf{y}))|J|d\mathbf{y} \ (*)$$

$where Jis the Jacobian of the transformation J = \left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| = \left|\frac{\partial g(\mathbf{y})}{\partial \mathbf{y}}\right|$ and the integrals are over corresponding regions.

If f is the density of $\mathbf{Z}$, then the left-hand integral over a set A is $P[\mathbf{Z} \in A]$, and if $\mathbf{Y} = g(\mathbf{Z})$ we also know that

$$P[\mathbf{Y} \in B] = P[g(\mathbf{Z}) \in g(A)],$$

but this left-hand side is the integral of the joint p.d.f. of $\mathbf{Y}$ over B, which must now be equal to the r.h.s. og (*).

It follows that the joint pdf of $\mathbf{Y}$ is h with $h(y) = f(g(y))|J|$.

6.2 The multivariate normal distribution

6.2.1 Handout

Suppose $Z_1, ..., Z_n$ are independent Gaussian with mean zero and variance one (i.e. $Z_1, ..., Z_n \sim n(0,1)$, i.i.d.) so their joint density is

$$f(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-z_i^2/2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-(1/2)\mathbf{z}^T\mathbf{z}\right)$$

and this is the density of the multivariate random variable $\mathbf{Z} = (Z_1, \ldots, Z_n)'$.

Let A be an invertible $n \times n$ matrix and $\mu \in \breve{}\mathbf{R}^n$ and define a new multivariate random variable, $\mathbf{Y} = A\mathbf{Z} + \mu$.

Some linear algebra gives

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where $\Sigma = AA^T$.

This leads to a natural definition of the multivariate normal distribution.

The n-dimensional random vector, $\mathbf{Y}$ is **defined** to have a multivariate normal distribution, denoted $\mathbf{Y} \sim n(\mu, \Sigma)$ if the density of $\mathbf{Y}$ is of the form

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where $\mu \in \mathbb{R}^n$ and Σ is a symmetric positive definite $n \times n$ matrix.

It is left to the reader to prove that if $\mathbf{Y} \sim n(\mu, \Sigma)$ and B is an $p \times n$ matrix of full rank p ($p < n$), then $B\mathbf{Y}$ also has a multivariate normal distribution.

6.3 Related distributions

6.3.1 Handout

If $Z \sim n(0,1)$ is standard normal, then we **define** the chi-squared distribution on one degree of freedom, χ^2_1 to be the distribution of $U := Z^2$ and write

$$U \sim \chi^2_1.$$

If $U_1, \ldots, U_p$ are i.i.d. χ_1, then we **define** χ^2_p to be the distribution of their sum and write

$$\sum_{i=i}^{p} U_i \sim \chi^2_p.$$

Finally, if $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ are independent, then we **define** the **F distribution on ν_1 and ν_2 degrees of freedom** to be the distribution of the ratio $\frac{U/\nu_1}{V/\nu_2}$ and write

$$\frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2}.$$

7 Orthogonal projections in multiple regression

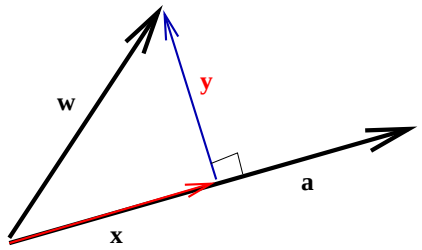
7.1 Background to projections

If $\mathbf{a}$ is a vector then we can write a general vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a}'\mathbf{y} = \mathbf{a} \cdot \mathbf{y} = 0$.
In the general case,

$$k = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2},$$

and for unit vectors $\mathbf{a}$ we obtain

$$k = \mathbf{w} \cdot \mathbf{a}.$$



7.1.1 Handout

If $\mathbf{a}$ is a vector then we can write a general vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a} \cdot \mathbf{y} = 0$.

With this

$$\begin{aligned} \mathbf{w} \cdot \mathbf{a} &= (\mathbf{x} + \mathbf{y}) \cdot \mathbf{a} \\ &= (k\mathbf{a} + \mathbf{y}) \cdot \mathbf{a} = k\mathbf{a} \cdot \mathbf{a} + \underbrace{\mathbf{y} \cdot \mathbf{a}}_{=0} \\ &= k \cdot ||\mathbf{a}||^2 \end{aligned}$$

i.e.

$$k = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2},$$

and therefore $\mathbf{w} = \mathbf{x} + \mathbf{y}$ with

$$\mathbf{x} = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2}\mathbf{a}$$

and residual:

$$\mathbf{y} = \mathbf{w} - \frac{\mathbf{u} \cdot \mathbf{a}}{||\mathbf{a}||^2}\mathbf{a}.$$

Note that we have shown that this is the **only possible** solution to writing $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a} \cdot \mathbf{y} = 0$ but not that it is indeed such a solution. Obviously $\mathbf{x}$ is of the stated form and it is not hard to see that $\mathbf{a} \cdot \mathbf{y} = 0$ is indeed true for this solution. This orthogonal decomposition therefore both exists and is unique.

7.2 Projections and bases

The Gram-Schmidt technique uses projections to iteratively build an orthonormal basis, $\mathbf{u}_1, \ldots \mathbf{u}_r$ which spans the same space as a sequence of arbitrary starting vectors, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_p$.
In linear regression, the starting vectors are typically the columns of the $\mathbf{X}$-matrix. r above is then the rank of the matrix.

7.2.1 Handout

The Gram-Schmidt technique uses the projections of the previous section to iteratively build an orthonormal basis, $\mathbf{u}_1, \ldots \mathbf{u}_r$ which spans the same space as a sequence of arbitrary starting vectors, $\mathbf{a}_1,\ \mathbf{a}_2, \ldots \mathbf{a}_p$:

$$\mathbf{u}_1 := \frac{1}{||\mathbf{a}||}\mathbf{a}_1$$

then for $i = 1, \ldots, p-1$

$$\mathbf{v}_{i+1} := (\mathbf{a}_{i+1} \cdot \mathbf{a}_1)\mathbf{a}_1 + \ldots + (\mathbf{a}_{i+1} \cdot \mathbf{a}_i)\mathbf{a}_i$$

with residual

$$\mathbf{e}_{i+1} := \mathbf{a}_{i+1} - \mathbf{v}_{i+1}$$

and next vector

$$\mathbf{u}_{i+1} := \frac{1}{||\mathbf{e}_{i+1}||}\mathbf{e}_{i+1}$$

If the starting vectors are linearly independent, then $r = p$, otherwise $r < p$ (some of the proposed basis vectors have turned out to be zero and are omitted).

It is often useful to expand the set of starting vectors, to $\mathbf{a}_1,\ \mathbf{a}_2, \ldots \mathbf{a}_p, \mathbf{e}_1,\ \mathbf{e}_2, \ldots \mathbf{e}_n$, where the $\mathbf{e}_i$ are the usual unit vectors. The method will then result in a full basis for $\mathbb{R}^n$, the first r of which span the same space as the starting vectors.

8 A basis for V=span(X)

8.1 Subspaces

In the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, assume $rank(\mathbf{X}) = r$ where $\mathbf{X}$ is $n \times p$ and $r \leq p$

Recall $\mathbf{X}\hat{\beta}$ is a project so $\mathbf{y} - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta}$ so that $\mathbf{y} - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$. If $r = p$, then:

$$\underbrace{\hat{\mathbf{e}}}_{n\times 1} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

$$= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')y = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - p$

8.1.1 Details

Assume $rank(\mathbf{X}) = r \leq p$ ($\mathbf{X}$ is $n \times p$).

Parameters in the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ are estimated with $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ if the inverse exists or in general with any $\hat{\beta}$ which is such that $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is a projection onto the subspace $sp(\mathbf{X})$.

By definition, a projection $\hat{\mathbf{y}}$ simply corresponds to a decomposition of the original vector into two orthogonal components, i.e. writing $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$. We have $\hat{\mathbf{e}} = y - \mathbf{X}\hat{\beta} \perp \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ so that $y - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$.

$$\underbrace{\hat{\mathbf{e}}}_{n\times 1} = y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

$$= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')y = (\mathbf{I} - \mathbf{H})y$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - r$

8.2 A basis for the span of X

Orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$:

Using Gram-Schmidt, first generate $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span $sp\{\mathbf{X}\}$, with $rank\{\mathbf{X}\} = r$ and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$.

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots + \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

8.2.1 Details

The probability distributions can best be viewed by defining a new orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$.

This basis is defined by first generating a set of r vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span the space defined by $sp\{\mathbf{X}\}$, and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$. This is obviously always possible using the method of Gram-Schmidt. This gives the following sequence of spaces and spans:

$$\begin{aligned} sp\{\mathbf{X}\} &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \\ \mathbf{R}^n &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots \mathbf{u}_n\} \end{aligned}$$

One can then write each of $\mathbf{X}\hat{\beta}$ and $\mathbf{y}$ in terms of the new basis as follows:

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

where it is well-known that $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$.

It is important to note that the same coefficients $\hat{\zeta}_i$ are obtained for $1 \leq i \leq r$. This follows from considering the coefficient of $\mathbf{y}$ in the basis and noting that $\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{e}}$ where the residual vector $\hat{\mathbf{e}}$ is orthogonal to all column vectors of $\mathbf{X}$ and therefore also to $\mathbf{u}_i$ for $1 \leq i \leq r$. Therefore,

$$\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y} = \mathbf{u}_i \cdot \mathbf{X}\hat{\beta}$$

8.3 Q-R decomposition

$$\mathbf{Q} := \left[\mathbf{u}_1 \vdots \mathbf{u}_2 \vdots \ldots \vdots \mathbf{u}_n\right]$$

is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X} = \mathbf{QR}$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y}$$

and hence

$$\begin{aligned} E[\mathbf{z}] &= \mathbf{Q}'\mathbf{X}\beta \\ V[\mathbf{z}] &= \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I} \end{aligned}$$

8.3.1 Details

$\mathbf{Q} := [\mathbf{u}_1\mathbf{u}_2 \ldots \mathbf{u}_n]$ is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X}$.

$\mathbf{Q}$ has important properties, e.g. $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ so $\mathbf{Q}^{-1} = \mathbf{Q}'$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y} \text{ and } \mathbf{y} = \mathbf{Q}\mathbf{z}$$

and hence

$$E[\mathbf{z}] = \mathbf{Q}'\mathbf{X}\beta$$
$$V[\mathbf{z}] = \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I}$$

8.4 Variances of coefficients

For each i we obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

8.4.1 Details

For each i we trivially obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

8.5 Expected values of coefficients

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = 0$$

8.5.1 Details

The expected values of the coefficients, $\hat{\zeta}_i$ depend on which space these correspond to. Define

$$\zeta_i = E\left[\hat{\zeta}_i\right]$$

and by linearity we obtain

$$\zeta_i = E[\mathbf{u}_i \cdot \mathbf{y}] = \mathbf{u}_i \cdot (\mathbf{X}\beta).$$

Now note that we have defined the basis vectors in three sets. The first is such that they span the same space as the columns of $\mathbf{Z}$. The second set complements the first to span the $\mathbf{X}$ and the last set complements the set to span all of $\mathbf{R}^n$. The basis vectors are of course all orthogonal and each basis vector is orthogonal to all vectors in spaces spanned by preceding vectors.

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta) = 0$$

since $\mathbf{X}\beta$ is trivially in the space spanned by the column vectors of $\mathbf{X}$ and is therefore a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{u}_i$ is orthogonal to all of these.

8.6 Sums of squares and norms

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=p+1}^{n} \hat{\zeta}_i^2$$

8.6.1 Details

It is now quite easy to see how to form sums of squared deviations based on the new orthonormal basis, since each set of deviations corresponds to a specific portion of the space.

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=r+1}^{n} \hat{\zeta}_i^2$$

8.7 Normality and independence of coeffients

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$.

When the y_i are independent Gaussian random variables, $\hat{\zeta}_i$ have zero covariance and are thus also independent.

8.7.1 Details

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$. The $\hat{\zeta}_i$ have zero covariance and when the y_i are independent Gaussian random variables, the $\hat{\zeta}_i$ are also independent.

This final result uses the fact that Gaussian random variables which have zero covariance are also independent. The fact that they have zero covariance is easy to establish, but the corrollary of independence is a result from multivariate normal theory.

The normal theory is fairly simple in this case:

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right) = \mathbf{Q}'\mathbf{y}$$

and

$$\mathbf{y} \sim n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

.

It follows that $\mathbf{z}$ is multivariate normal and from the earlier derivations of the mean and variance we have

$$\mathbf{z} \sim n\left(\mathbf{Q}'\mathbf{X}\beta, \sigma^2\mathbf{I}\right).$$

8.8 Degrees of freedom

$SSE(F)$has $n-r$ degrees of freedom.

8.8.1 Details

$SSE(F)$ has $n-r$ degrees of freedom.